UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos

por

STANLEY LOH

Tese submetida à avaliação, como requisito parcial
para a obtenção do grau de Doutor em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
orientador

Porto Alegre, Julho de 2001.

Agradecimentos

Às Universidades Luterana do Brasil (ULBRA) e Católica de Peloas (UCPEL), pelo apoio e incentivo que me foram dados como professor destas instituições;

Aos colegas do PPGC e aos colegas professores da ULBRA e UCPEL, pelo incentivo;

À Clínica Olivé Leite, por ter cedido os prontuários médicos, os quais são objetos de pesquisa apoiada pelo Fundo de Incentivo ao Desenvolvimento do Ensino e da Pesquisa em Saúde (FIDEPS, Ministério da Saúde);

Ao pessoal da Clínica Olivé Leite, Dr. Fábio Leite Gastal, Dr. Sérgio Olivé Leite e Dr. Sérgio Andreolli, pelos comentários e avaliações feitas;

Aos companheiros Leandro Krug Wives e Maurício Almeida Gameiro, pelas valiosas discussões e pelo excelente trabalho cooperativo;

Ao Prof. Mário Ulyssea Capanema da UCPEL, por ter fornecido a coleção textual sobre estratégias agroalimentares;

A meu orientador, Prof. Dr. José Palazzo Moreira de Oliveira, pela inestimável contribuição neste trabalho, pelo incentivo pessoal e pela dedicação profissional;

Ao saudoso Prof. Dr. José Mauro Volkmer de Castilho, o qual foi meu primeiro orientador neste curso de doutorado, mas a quem o destino não permitiu ver os frutos deste trabalho;

A minha querida mãe Hilária, pela paciência e incentivo;

A minha esposa, Simone, que se negou para que eu alcançasse esta realização (esta conquista também é tua);

A minha filhinha Ana Luíza, que me deu mais motivação para esta realização e iluminou ainda mais minha vida;

A Jesus Cristo, meu senhor, que dirigiu minha vida até aqui (toda glória e todo louvor).

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Esta tese apresenta uma abordagem baseada em conceitos para realizar descoberta de conhecimento em textos (*KDT*). A proposta é identificar características de alto nível em textos na forma de conceitos, para depois realizar a mineração de padrões sobre estes conceitos.

Ao invés de aplicar técnicas de mineração sobre palavras ou dados estruturados extraídos de textos, a abordagem explora conceitos identificados nos textos. A idéia é analisar o conhecimento codificado em textos num nível acima das palavras, ou seja, não analisando somente os termos e expressões presentes nos textos, mas seu significado em relação aos fenômenos da realidade (pessoas, objetos, entidades, eventos e situações do mundo real).

Conceitos identificam melhor o conteúdo dos textos e servem melhor que palavras para representar os fenômenos. Assim, os conceitos agem como recursos meta-lingüísticos para análise de textos e descoberta de conhecimento. Por exemplo, no caso de textos de psiquiatria, os conceitos permitiram investigar características importantes dos pacientes, tais como sintomas, sinais e comportamentos. Isto permite explorar o conhecimento disponível em textos num nível mais próximo da realidade, minimizando o problema do vocabulário e facilitando o processo de aquisição de conhecimento.

O principal objetivo desta tese é demonstrar a adequação de uma abordagem baseada em conceitos para descobrir conhecimento em textos e confirmar a hipótese de que este tipo de abordagem tem vantagens sobre abordagens baseadas em palavras.

Para tanto, foram definidas estratégias para **identificação dos conceitos** nos textos e para **mineração de padrões** sobre estes conceitos. Diferentes métodos foram avaliados para estes dois processos. Ferramentas automatizadas foram empregadas para aplicar a abordagem proposta em estudos de casos.

Diferentes experimentos foram realizados para demonstrar que a abordagem é viável e apresenta vantagens sobre os métodos baseados em palavras. Avaliações objetivas e subjetivas foram conduzidas para confirmar que o conhecimento descoberto era de qualidade. Também foi investigada a possibilidade de se realizar descobertas proativas, quando não se tem hipóteses iniciais.

Os casos estudados apontam as várias aplicações práticas desta abordagem. Pode-se concluir que a principal aplicação da abordagem é permitir análises qualitativa e quantitativa de coleções textuais. Conceitos podem ser identificados nos textos e suas distribuições e relações podem ser analisadas para um melhor entendimento do conteúdo presente nos textos e, conseqüentemente, um melhor entendimento do conhecimento do domínio.

Abstract

This thesis presents a concept-based approach for performing knowledge discovery in texts (*KDT*). The purpose is to identify high-level textual characteristics and after that to perform a mining process for discovering patterns in the textual collection.

Instead of applying mining techniques over keywords or structured data extracted from texts, the approach explores concepts. Concepts represent text content better than words, minimizing the vocabulary problem. Performing the analysis over concepts allows to understand the meaning of terms and expressions and thus the phenomena referenced by the texts (people, objects, events, entities, situations, etc.). That allows to analyze the knowledge codified in texts in a way closer to the reality. Thus, the knowledge acquisition process requires less effort.

The main goal of this thesis is to demonstrate that a concept-based approach is suitable for knowledge discovery in texts and has advantages over keyword-based approaches.

The work defined a strategy for identifying concepts in texts and for mining patterns over the concepts. Different methods were analyzed and the best ones were embedded in automated tools. Some experiments were carried out and subjective and objective evaluations were conducted to validate the quality of the discovered knowledge and the advantage of the concept-based approach. The thesis also investigated the proactive discovery, when there is no initial hypothesis.

The approach was applied in different domains to show its practical benefits. The main conclusion is that the approach is useful to perform qualitative and quantitative analyses of textual collections. Concepts identified in the texts and patterns over concepts help to understand the content of the collection and thus imply in a better understanding of the domain knowledge.

1 INTRODUÇÃO

Com o crescente uso de computadores, cada vez mais documentos eletrônicos estão sendo armazenados e colocados à disposição das pessoas. Em sua grande maioria, estes documentos contêm informações codificadas em forma textual, tais como dicionários, manuais, enciclopédias, guias e mensagens de correio eletrônico. Estudos recentes afirmam que 80% da informação de uma companhia estão contidos em documentos textuais [TAN99].

Toda esta documentação textual possui muito conhecimento implícito, que pode ser explorado de alguma forma [BOW96]. Davies [DAV89] afirma que muito conhecimento que nunca foi formalizado explicitamente ou mesmo implicitamente pode ser inferido do que já foi publicado. O mesmo autor compara a biblioteca de hoje (um armazém de objetos passivos) com a biblioteca do futuro, a qual poderá fornecer ao usuário conexões desconhecidas, fazer associações e analogias, sugerir conceitos remotos ou novos, descobrir novos métodos, teorias, medidas, etc.

Entretanto, encontrar tal conhecimento é uma tarefa árdua. Existem técnicas e ferramentas para Recuperação de Informação (RI), as quais auxiliam as pessoas a encontrar documentos que contenham informações relevantes [SPA97]. Entretanto, é necessário examinar os documentos resultantes para encontrar a informação desejada. A dificuldade vem do fato de que documentos são insatisfatórios como respostas, por serem grandes e difusos em geral [WIL94]. Além disto, ferramentas de RI costumam produzir como resposta uma quantidade muito grande de documentos, causando a chamada "sobrecarga de informações" (*information overload)*, que acontece quando o usuário tem muita informação ao seu alcance, mas não tem condições de tratá-la ou de encontrar o que realmente deseja ou lhe interessa [CHE94].

A evolução da área de Recuperação de Informação teve como conseqüência o surgimento da área de Descoberta de Conhecimento em Textos (*Knowledge Discovery from Text- KDT*). O termo foi utilizado pela primeira vez por Feldman e Dagan [FEL95] para designar o processo de encontrar algo interessante em coleções de textos (artigos, histórias de revistas e jornais, mensagens de *e-mail*, páginas *Web*, etc.). Hoje em dia, sinônimos como *Text Mining* ou *Text Data Mining* também são utilizados para o mesmo fim [TAN99].

Pode-se então definir **Descoberta de Conhecimento em Textos (*KDT*)** ou ***Text Mining*** como sendo o processo de extrair padrões ou conhecimento, interessantes e não-triviais, a partir de documentos textuais [TAN99].

Assim, ao invés de encontrar os textos que contenham informações e deixar que o usuário mesmo procure o que lhe interessa, a nova área se preocupa em encontrar informações dentro dos textos e tratá-las de forma a apresentar ao usuário algum tipo de conhecimento útil e novo. Mesmo que tal conhecimento novo não seja a resposta direta às indagações do usuário, ele deve contribuir para satisfazer as necessidades de informação do usuário.

Segundo Tan [TAN99] e Feldman e Dagan [FEL95], o processo de *KDT* pode ser realizado aplicando-se técnicas de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discobery in Databases - KDD*) sobre dados extraídos de textos (não necessariamente valores numéricos, mas podendo ser também valores nominais, como

palavras do texto). Entretanto, *KDT* não inclui somente a aplicação das técnicas tradicionais de *KDD*, mas também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto.

Outros autores pesquisam o mesmo problema há tempos, mas usando termos diferentes. Oard [OAR96], por exemplo, utiliza um termo mais amplo Busca de Informação (*Information Seeking*) para descrever qualquer processo pelo qual usuários procuram obter informação a partir de sistemas de informação automatizados, incluindo a busca de informações em textos. Chen [CHE93] cita alguns autores que usam o termo Conhecimento Público Não-Descoberto (*Undiscovered Public Knowledge*). Lewis [LEW96] utiliza o termo Recuperação de Conhecimento (*Knowledge Retrieval*), que difere de "recuperação de documentos" e de "recuperação de dados", porque usa menos pré-codificação e requer mais poder de inferência. Na seção seguinte, as técnicas para *KDT* serão discutidas.

As aplicações de sistemas de *KDT* são inúmeras. Qualquer domínio que utilize intensivamente textos poderão beneficiar-se destes sistemas, tal como as áreas jurídicas e policiais, os cartórios e órgãos de registros, empresas em geral, etc.

1.1 Estratégias e Técnicas para Descoberta de Conhecimento em Textos

Tan [TAN99] apresenta um esquema para realizar descoberta de conhecimento em textos. Segundo este esquema, os textos seriam transformados para formas intermediárias (etapa denominada de *text refining*), as quais posteriormente seriam analisadas para a mineração de padrões (etapa de *knowledge distillation*). Tan sugere que há dois tipos de formas intermediárias possíveis: um tipo é baseado no documento (*document-based*) e outro em conceitos (*concept-based*). No primeiro caso, cada entidade presente no formato intermediário representa um documento, independente do domínio. Já no segundo caso, entidades representam objetos ou conceitos de um domínio e conforme interesses específicos. Segundo Tan, é possível também transformar formatos baseados em documentos para formatos baseados em conceitos (um processo chamado de extração, segundo o referido autor).

Segundo o esquema de Tan, o processo de mineração envolve descobrir padrões e relacionamentos entre (a) documentos, no primeiro caso, e entre (b) objetos ou conceitos, no segundo caso. Para cada tipo de mineração, Tan sugere algumas técnicas básicas: (a) agrupamento (*clustering*), categorização e visualização e (b) modelagem preditiva e descoberta associativa.

A seguir, serão apresentadas rapidamente várias abordagens ou técnicas para descoberta de conhecimento em textos (estas abordagens são discutidas com mais detalhes em [LOH99]). Os nomes utilizados servem apenas para diferenciar as abordagens, portanto nem sempre correspondem aos termos utilizados na literatura, nem seguem uma classificação previamente estabelecida pela comunidade científica.

A técnica mais básica e mais usada para *KDT* é a recuperação de informação (RI), que se limita a encontrar documentos ou textos onde informações relevantes possam estar. Sparck-Jones e Willet [SPA97] apresentam artigos clássicos sobre este tema. Conforme Chen [CHE93], a RI é parte de um processo maior de exploração, correlação e síntese de informação. As técnicas de RI podem ajudar apresentando documentos com visão

geral das informações ou assuntos (RI tradicional) ou apresentando partes de documentos com detalhes de informações (recuperação por passagens). Também as ferramentas de RI por filtragem contribuem garimpando documentos interessantes para o usuário, sem que este precise formular consultas.

Outra técnica básica é a categorização de textos, cujo objetivo é associar categorias (assuntos, classes ou temas) pré-definidas a textos livres [YAN99]. Há muitos trabalhos neste área, apresentando diversos métodos para categorização de textos [APT94] [COH96] [YAN94] [LID94]. Yang e outros [YAN97] [YAN99] fazem análises de vários métodos de categorização. Mas também há a preocupação com a escolha das características textuais que serão usadas no método. Yang e Pedersen [YAN97] comparam métodos para fazer esta seleção. Em geral, os trabalhos de categorização de textos procuram encontrar o tema central de um texto (ou temas, se houver mais de um).

Também clássica é a técnica de Extração de Informação (EI), cujo objetivo é encontrar informações específicas dentro dos textos (conforme Sparck-Jones e Willet [SPA97]). Riloff e Lehnert [RIL94] afirmam que o objetivo da área de EI é diferente do objetivo da área de processamento de linguagem natural (PLN), porque é mais focado e mais bem definido, visando extrair tipos específicos de informação. A técnica de EI procura converter dados não-estruturados em informações explícitas, geralmente armazenadas em bancos de dados estruturados. Isto pode ser feito isolando-se partes relevantes do texto, extraindo informação destas partes e transformando-as em informações mais digeridas e melhor analisadas (conforme Cowie e Lehnert [COW96]). Em geral, os métodos utilizados são direcionadas para extrair características do domínio (objetos, entidades, relações), servindo apenas para aplicações específicas (conforme Croft [CRO95]).

A abordagem de Descoberta Tradicional após Extração é a mais simples, pois utiliza técnicas já testadas e consagradas. Nesta abordagem, os dados são extraídos dos textos e formatados em bases de dados estruturadas, com o auxílio de técnicas de Extração de Informações (EI). Depois, são aplicadas técnicas e algoritmos de *KDD* (mineração de dados estruturados), para descobrir conhecimento útil para o usuário .

A abordagem de descoberta por Extração de Passagens designa um tipo de descoberta situado entre a recuperação de informações por passagens e a extração de informações. Esta nova abordagem visa encontrar informações específicas, mas de forma um pouco mais independente de domínio do que as ferramentas tradicionais de extração. Esta abordagem difere da Extração de Informação pois permite ao usuário levantar hipóteses e formas de procura de informações em tempo de execução, não sendo necessário um grande esforço de engenharia do conhecimento (para definir as formas de procura, por exemplo os "*tags*"), nem um profundo conhecimento prévio do texto e de sua estrutura. A descoberta por extração de passagens auxilia usuários a encontrar detalhes de informação, sem que este precise ler todo texto. Entretanto, ainda assim, é necessário que o usuário leia e interprete as partes do texto que forem recuperadas para extrair a informação desejada. Grobelnik e outros [GRO00] citam a ferramenta automatizada (*workbench*) de Caruana e Hodor para auxiliar especialistas humanos na extração de informações. A ferramenta permite combinar a precisão do trabalho humano com tarefas em larga escala. No trabalho citado, especialistas humanos revisaram 5.000 documentos em uma tarde e extraíram informações com precisão e abrangência maiores que 99,9%.

A abordagem por Análise Lingüística procura descobrir informações e regras analisando sentenças da linguagem a nível léxico, morfológico, sintático e semântico.

Ambrosio e outros [AMB97], por exemplo, descobrem generalizações escondidas, analisando padrões sintáticos (*tags*). Lascarides e outros [LAS92] e Hobbs [HOB79] relatam pesquisas sobre inferências de relações de coerência em textos (por exemplo, causa e efeito), também utilizando *tags*. Os trabalhos apresentados em [LAS92], [HWA92], [KAM93] e [WEB88] inferem relações de tempo analisando textos. Bowden e outros [BOW96] descobrem relações conceituais (definições, exemplos, partições e composição) através de *tags* no texto.

A descoberta por Análise de Conteúdo é semelhante aos dois tipos anteriores, pois investiga lingüisticamente os textos e apresenta ao usuário informações sobre o seu conteúdo. Entretanto, a diferença para a descoberta por análise lingüística é que, na análise de conteúdo, há maior esforço no tratamento semântico dos textos, passando o limite léxico-sintático. Em relação à extração de passagens, a diferença é que, aqui, o objetivo é encontrar o significado do texto pretendido pelo autor ao invés de partes ou informações específicas. Por exemplo, Saggion e Carvalho [SAG95] utilizam técnicas que analisam a estrutura de resumos ou sumários, identificando informações por palavras-chave, tais como hipóteses, conclusões, experimentos, etc. Em [WIE94], há estudos sobre descoberta de crenças e intenções em diálogos, por inferências sobre palavras-chave ("*tags*").

A abordagem de descoberta por *Sumarização* ou resumos utiliza as técnicas dos tipos anteriores, mas com ênfase maior na produção do resumo ou sumário. Segundo Sparck-Jones e Willet [SPA97], *sumarização* é a abstração das partes mais importantes do conteúdo do texto. Miike e outros [MII94] apresentam um trabalho de geração automática de resumos em tempo de execução através de interações com o usuário. Já McKeown e Radev [MCK95] apresentam técnicas e ferramentas para analisar diversos artigos sobre um mesmo evento e criar um resumo em linguagem natural. Em [HER95] é apresentada uma ferramenta para *sumarização* com dois componentes principais: um planejador de conteúdo (que seleciona informações de uma base de *slots*) e um componente lingüístico (para gerar as frases de saída em linguagem natural).

Já a descoberta por Associação entre Passagens busca encontrar automaticamente conhecimento e informações relacionadas no mesmo texto ou em textos diferentes. Sua aplicação imediata está na definição automática de *links* nos sistemas de hipertexto. Entretanto, a vantagem deste tipo de descoberta é apresentar ao usuário partes de textos que tratam do mesmo assunto específico (detalhe de informação e não conteúdo geral).

Na descoberta por Listas de Conceitos-Chave, a idéia é apresentar uma lista com os conceitos principais de um único texto (geralmente, os conceitos são termos ou expressões extraídos por análises estatísticas). Moscarola e outros [MOS98], por exemplo, sugerem uma lista de termos próximos (antes e depois), os quais permitem a análise do conteúdo por quase-frases. Outros exemplos são a técnica de afinidades léxicas discutida em [MAA92] e a técnica dos relacionamentos semânticos apresentada em [SPA97].

A descoberta de Estruturas de Textos segue a premissa da coesão léxica, segundo a qual, determinar a estrutura de um texto ajuda a entender seu significado [MOR91]. Um texto não é um conjunto aleatório de frases, mas deve haver uma unidade e também coesão, com as frases funcionando juntas para a função do todo. A coesão pode ser analisada pelas referências, conjunções e relações semânticas presentes no texto.

A descoberta por Agrupamento (*clustering*) procura separar automaticamente elementos em grupos por afinidade ou similaridade (não há classes pré-definidas). A técnica de agrupamento é diferente da classificação, pois a primeira visa criar as classes através da

organização dos elementos, enquanto que a segunda procura alocar elementos em classes já pré-definidas (conforme Willet [WIL88]). O agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes.

Geralmente, a técnica de agrupamento vem associada com alguma técnica de descrição de conceitos, para identificar os atributos de cada classe. Esta posterior identificação das classes através de suas características é chamada de "análise da classe" (*cluster analysis*), conforme Willet [WIL88], e gera uma nova abordagem de descoberta: a descoberta por Descrição de Classes de Textos. Dada uma classe de documentos textuais (já previamente agrupados) e uma categoria associada a esta classe (por exemplo, tema ou assunto dos textos), a descoberta por descrição procura encontrar as características principais desta classe. Estas características permitem identificar os elementos que pertencem à classe e distingui-los dos elementos de outras classes. Esta abordagem segue geralmente as técnicas para construção do centróide de classes. Ela é diferente da abordagem por listas de conceitos-chave, porque descobre características comuns em vários textos e não em um único texto.

A abordagem de descoberta por Associação entre Textos procura relacionar descobertas presentes em vários textos diferentes. As descobertas estão presentes no conteúdo ou no significado dos textos. Esta abordagem é diferente do que acontece na descoberta por associação entre passagens, cujo objetivo é somente relacionar partes de textos sobre o mesmo assunto. Na associação entre textos, a interpretação semântica é fundamental. Swanson [SWA97] comenta que o conhecimento novo pode emergir de inúmeros fragmentos individualmente não-importantes, sem relação no momento em que foram elaborados ou adquiridos. Por exemplo, Swanson e Smalheiser [SWA97b] fizeram descobertas na área médica relacionando textos que não se referenciam e que aparentemente não continham assuntos comuns. Em [MCK95], é apresentada uma ferramenta que analisa diversos artigos sobre um mesmo evento e cria um resumo único em linguagem natural. São extraídas informações de partes dos textos e analisadas para encontrar similaridades e diferenças de informações. Davies [DAV89] acredita que existe muita informação publicada e conhecida, mas que algumas conclusões a partir destas informações só poderão ser descobertas recuperando estes documentos e notando as conexões lógicas entre eles.

A descoberta por Associação entre Características procura relacionar tipos de informação (atributos) presentes em textos, aplicando a técnica de correlação ou associação tradicional em *KDD* diretamente sobre partes do texto. Uma das diferenças é que os valores para os atributos são partes do texto e não necessariamente dados extraídos por técnicas de extração de informações. Feldman e Dagan [FEL98], por exemplo, marcam documentos textuais com palavras-chave tomadas de um vocabulário controlado, organizado em estruturas hierárquicas de tópicos. Ferramentas de descoberta procuram encontrar padrões na coleção de documentos por análise de distribuições de palavras-chave. Feldman e Hirsh [FEL97] também discutem a descoberta de associações (padrões de co-ocorrência) entre termos que marcam textos.

A abordagem de descoberta por Hipertextos é um caso especial de descoberta utilizando técnicas de recuperação de informações (no caso, o modelo de hipertextos). Nesta abordagem, a descoberta é exploratória e experimental, feita através de mecanismos de navegação (*browsing*), conforme comentam Marchionini e Shneiderman [MAR88]. Com tais ferramentas, é possível expandir e comparar o conhecimento através dos *links* que

relacionam as informações, funcionando de modo análogo à mente humana (memória associativa). A aprendizagem pode ocorrer acidentalmente e de forma cumulativa, não exigindo estratégias cognitivas. A criatividade e a curiosidade guiam tal processo. Segundo Morgado [MOR98], "*hipertextos possibilitam a criação de ambientes onde o utilizador pode experimentar um certo grau de autonomia enquanto navega na informação, o que contribui sem dúvida para que se expressem estratégias individuais de aprendizagem, sendo o sujeito responsável pelo seu próprio processo de aprendizagem*".

A abordagem por Manipulação de Formalismos procura representar o conteúdo dos textos em formalismos (como a lógica de predicados, por exemplo). Assim, mecanismos de manipulação simbólica podem inferir novos conhecimentos, simplesmente por transformações na forma. As representações resultantes podem ser depois transformadas para estruturas na linguagem natural, facilitando a compreensão de usuários leigos no formalismo. As técnicas de dedução, comuns na área de Inteligência Artificial, executam bem este trabalho.

No mesmo sentido, a abordagem de descoberta por Combinação de Representações faz uso de representações de textos. Os formalismos internos podem ser modelos conceituais ou tradicionais (por exemplo, o modelo relacional) ou ontologias, linguagens baseadas em lógica, etc. Um exemplo é o trabalho de Croft e Turtle [CRO92], que compara grafos (representando, por exemplo, estruturas sintáticas de textos ou conteúdos mais complexos), os quais são combinados por elementos comuns, gerando um grafo novo, hipótese de novos conhecimentos.

Por fim, as técnicas de visualização e navegação permitem analisar grupos de textos usando representações gráficas. Tan [TAN99] discute ferramentas que apresentam documentos e relacionamentos através de agrupamentos visuais, mapas bi ou tridimensionais, conexões gráficas e outras estruturas complexas.

1.2 Estrutura desta Tese

Este documento está estruturado como descrito a seguir. O capítulo 2 apresenta os principais trabalhos relacionados ao tema da tese e discute alguns problemas ainda não adequadamente resolvidos. O capítulo 3 apresenta os objetivos da tese, descrevendo também as atividades propostas e os resultados esperados para cada objetivo. No capítulo 4, é descrita a proposta de abordagem baseada em conceitos, bem como sua fundamentação teórica através de trabalhos relacionados e soluções similares de outras áreas.

Após, no capítulo 5, descreve-se como foram feitos os experimentos para avaliação dos objetivos propostos, incluindo a descrição do domínio de aplicação, a caracterização da coleção de textos usada, os conceitos empregados na descoberta, o detalhamento do processo padrão de descoberta, exemplos de conhecimento descoberto e o ambiente computacional utilizado para os experimentos.

No capítulo 6, são apresentados e discutidos os resultados das avaliações feitas, bem como observações sobre os experimentos e algumas conclusões iniciais. O capítulo 7 apresenta diversas aplicações da abordagem proposta, enfatizando sua utilidade prática e benefícios. Finalmente, o capítulo 8 (considerações finais) discute os objetivos alcançados, as contribuições desta tese, as vantagens e limitações da proposta e também dá início a discussões sobre trabalhos futuros.

2 TRABALHOS CORRELATOS E PROBLEMAS EM ABERTO

Feldman e outros [FEL95] [FEL97] [FEL98] aplicam técnicas de mineração de dados (*KDD*) sobre palavras-chave que identificam textos. São utilizadas técnicas estatísticas para descobrir padrões tais como palavras-chave mais comuns e correlações entre palavras. O problema é que as palavras-chave devem ter sido previamente associadas e somente indicam o assunto principal dos textos.

Lin e outros [LIN98] também descobrem associações em textos, mas utilizam termos extraídos automaticamente dos textos. Os termos mais freqüentes são associados aos textos como palavras-chave. Feldman e outros [FEL98b] também sugerem extrair termos diretamente dos textos, mas somente os considerados mais significativos. Para selecionar os termos, devem ser analisadas as seqüências sintáticas que podem indicar padrões interessantes, como por exemplo "*substantivo substantivo*", "*substantivo preposição substantivo*" e "*adjetivo substantivo*".

Já Davies [DAV89] sugere estratégias mais complexas para descoberta de conhecimento em textos. Uma delas procura identificar analogias em diferentes textos através da análise de termos comuns. Sua sugestão é omitir termos relacionados à área para reunir documentos de diversas áreas que possam estar relacionados. Entretanto, a relação deve ser feita por atividades humanas. Para este problema, Chen [CHE93] sugere a construção automática de resumos combinando partes de distintos textos, usando para isto estruturas internas (redes semânticas) e termos comuns aos textos.

Outro tipo de descoberta sugerida por Davies [DAV89] são as correlações escondidas (combinações de conceitos através de relações estatísticas). Para tanto, o referido autor sugere que sejam analisadas as distribuições de termos numa coleção. Assim, por exemplo, foi possível identificar uma hipótese de relação entre um certo tipo de falha num sistema e alguns itens mais freqüentes (possíveis causas das falhas).

Davies [DAV89] afirma que o todo é mais que a mera soma das partes, o que permite que conhecimentos novos não explicitamente presentes nos textos possam ser descobertos analisando relações semânticas entre os textos. Em [SWA97b], é apresentada uma estratégia para descobrir relações entre temas presentes em textos diferentes e sem conexões. No caso, as relações são identificadas através da análise dos termos presentes nos textos.

Um dos problemas das estratégias acima citadas é que elas se baseiam em termos ou palavras e não em conceitos da realidade. Estratégias baseadas em palavras dificultam o entendimento do conhecimento descoberto. Por exemplo, resultados de um processo de descoberta em textos médicos concluíram que o termo "*visual*" é muito comum em prontuários de pacientes com determinada doença mental. Entretanto, não se pode concluir se o termo se refere a "*deficiência visual*" ou a "*ilusão visual*". A dificuldade se dá, neste caso, porque os termos não estão relacionados a conceitos da realidade, mas somente a documentos (forma intermediária baseada no documento, segundo o esquema de Tan [TAN99]).

Além disto, o conhecimento descoberto não pode ser embutido em sistemas baseados em conhecimento, mas somente em sistemas automáticos que utilizem regras ou

técnicas baseadas em termos. Neste último caso, o raciocínio de decisão usando tal conhecimento se torna difícil de ser verificado ou explicado.

Outra limitação das técnicas baseadas em termos ou palavras é o chamado "problema do vocabulário" (*vocabulary problem*), discutido em [FUR87], [CHE94] e [CHE97]. Este problema ocorre porque a linguagem pode ocasionar erros semânticos devido aos sinônimos (palavras diferentes com o mesmo significado), à polisemia (a mesma palavra com diferentes significados), às variações léxicas (uso de radicais, conjugações verbais, variações de gênero e número) e aos chamados quase-sinônimos (palavras correlatas, como "bomba" e "explosão").

Se as análises dos textos forem feitas sem considerar estes problemas, o processo de *KDT* pode levar a resultados incorretos. Por exemplo, em textos sobre crimes, aparecem os termos sinônimos "*homicídio*" e "*assassinato*". O processo de *KDT* deve saber identificar que ambos os termos se referem ao mesmo conceito para poder descobrir conhecimento sobre homicídios.

Para minimizar o problema, alguns autores sugerem o uso ou a análise dos termos sinônimos. Chen [CHE94] argumenta que as pessoas, em geral, usam termos diferentes para descrever conceitos similares (caso dos sinônimos). Furnas e outros [FUR87] discutem a efetividade do uso de sinônimos em uma estratégia chamada de "*unlimited aliasing*", onde objetos podem ser representados por inúmeros sinônimos. Jensen e Martinez [JEN00] e Wilcox e outros [WIL00] obtiveram melhoras de desempenho em processos de categorização de textos usando sinônimos.

Outro caso de problema do vocabulário é a forma diferente como conceitos da realidade podem ser expressos em textos. Por exemplo, em textos médicos, para indicar que o paciente tem sintomas de alcoolismo, podem ser usadas expressões como "*faz uso de álcool*", "*tem hálito etílico*" e "*bebe destilados*".

Uma das maneiras de minimizar tal problema é utilizar um vocabulário controlado. Lima e outros [LIM97] propõem um modelo para assinalar códigos do CID (Classificação Internacional de Doenças) [CEN89] a prontuários médicos contendo informações sobre pacientes. Os termos presentes nos textos dos prontuários são analisados em relação a termos usados no CID (um *thesaurus* da área médica) e em relação a sinônimos da área para associar diagnósticos aos prontuários. Entretanto, o citado vocabulário foi utilizado somente para melhorar a recuperação de documentos e não para realizar *KDT*.

Além disto, vocabulários controlados não necessariamente ajudam a resolver erros semânticos. Por exemplo, em prontuários médicos, se a expressão "*paciente nega dor de cabeça*" fosse encontrada, um vocabulário controlado só permitiria descobrir que o sintoma "dor de cabeça" teria sido citado no texto (análise a nível de documento). O correto, contudo, seria descobrir que o tal sintoma não está presente no paciente (nível de conceitos, mais próximo da realidade).

Apesar das estratégias acima terem minimizado o problema do vocabulário, elas utilizam formas intermediárias baseadas no documento e não em conceitos do domínio, segundo o esquema proposto por Tan [TAN99]. Isto quer dizer que o conhecimento descoberto somente diz respeito aos documentos e não a objetos ou entes da realidade. Tal afirmação é corroborada pelas técnicas de descoberta empregadas nestes trabalhos: classificação e recuperação.

Tan [TAN99] analisou várias ferramentas de *Text Mining* e conclui que os produtos existentes somente conseguem trabalhar bem com a forma intermediária baseada no documento, não existindo ferramentas de *KDT* eficientes para analisar formas intermediárias

baseadas em conceitos. Segundo Tan, faltam estratégias de mineração para destilar conhecimento analisando estruturas de mais alto nível, mais próximas de conceitos ou objetos da realidade do domínio.

A hipótese deste trabalho é que processos de *KDT* podem ser realizados sobre conceitos do domínio, de uma maneira mais próxima do raciocínio humano. Assim, os conceitos agem como recursos meta-lingüísticos para análise de textos e descoberta de conhecimento. A idéia é realizar o processo de descoberta num nível acima das palavras, ou seja, não analisando somente os termos e expressões presentes nos textos, mas seu significado em relação aos fenômenos da realidade (pessoas, objetos, entidades, eventos e situações do mundo real). Pressupõe-se também que conceitos servem melhor que palavras para representar e explicar o conhecimento usado em processos intelectuais.

Alguns trabalhos seguem esta tendência. Subasic e Huettner [SUB00] analisam textos sobre filmes para extrair atributos qualitativos como "horror", "justiça" e "dor". O objetivo é classificar os filmes em gêneros dependendo das características identificadas nos textos.

Wilcox e outros [WIL00], por sua vez, convertem textos médicos narrativos para códigos padronizados que representam observações de alto nível sobre pacientes (sinais e sintomas). O objetivo da estratégia é caracterizar classes para um processo posterior de classificação de textos médicos.

Entretanto, tais trabalhos não realizam nenhum tipo de mineração sobre os conceitos extraídos dos textos para destilar conhecimento novo, confirmando as conclusões de Tan [TAN99].

A proposta desta tese é analisar características de alto nível em textos para realizar descoberta de conhecimento. Ao invés de aplicar as técnicas sobre termos ou palavras-chave presentes nos textos, a proposta é identificar conceitos presentes nos textos e depois aplicar técnicas de mineração sobre estes conceitos. Assim, seria possível diminuir o problema do vocabulário e permitir descobertas a nível de conceitos (nível meta-lingüístico, mais próximo da realidade) e não a nível de palavras (nível lingüístico).

A abordagem proposta combina um processo de categorização, para identificar conceitos presentes nos textos, com a posterior aplicação de técnicas de mineração sobre estes conceitos, para descobrir padrões interessantes através de análises estatísticas.

3 OBJETIVOS DA TESE

O principal objetivo desta tese é demonstrar a adequação de uma abordagem baseada em conceitos para descobrir conhecimento em textos e confirmar a hipótese de que este tipo de abordagem tem vantagens sobre abordagens baseadas em palavras.

Para tanto, serão definidas estratégias para **identificação dos conceitos** nos textos e para **mineração de padrões** sobre estes conceitos. Diferentes métodos serão avaliados para estes dois processos. Ferramentas automatizadas serão empregadas para aplicar a abordagem proposta em estudos de casos. A hipótese será comprovada por avaliações específicas.

A seguir, são listados os objetivos específicos deste trabalho.

1) Avaliar alternativas para definição de conceitos:
   Conceitos são identificados nos textos pela análise de palavras. Faz-se necessário estudar o processo de identificação dos conceitos nos textos. Uma parte deste processo se refere ao modo como os conceitos são selecionados e caracterizados, para que possam ser identificados. Para tanto, deve-se estudar: como os conceitos são escolhidos, que tipos de termos permitem identificar os conceitos, como deve ser a escolha destes termos e as relações possíveis entre os termos. Espera-se que sejam identificados fatores que influenciam tal processo. Esta avaliação servirá para fundamentar a estratégia a ser usada no processo geral de *KDT*.

2) Avaliar métodos de categorização (identificação de conceitos):
   A outra parte do processo de identificação de conceitos nos textos é um processo de classificação ou categorização, pois procura identificar que classes ou categorias estão relacionadas a um certo texto. Devem ser estudados diferentes métodos para tal processo. Espera-se poder identificar em que situações cada tipo de métodos é melhor aplicado.

3) Definir um processo padrão de identificação de conceitos:
   Usando os melhores métodos de definição e categorização, deve-se definir um processo padrão para identificação de conceitos nos textos. Espera-se ter ferramentas automatizadas que realizem o processo de forma quase-automática.

4) Definir um processo padrão de mineração sobre conceitos:
   Deve-se decidir como os conceitos serão analisados para gerar novo conhecimento. Duas técnicas de mineração sobre conceitos serão estudadas: distribuição de conceitos (lista de conceitos chave) e associação. Espera-se ter ferramentas automatizadas que realizem o processo de forma quase-automática.

5) Realizar *KDT* em algum domínio usando o processo padrão definido:
   Escolher um domínio e aplicar o processo padrão definido.

6) Avaliar o grau de acerto na identificação de conceitos:

Sobre o experimento do passo anterior, deve-se avaliar a margem de erro através de critérios de teste.

7) Avaliar a qualidade do conhecimento descoberto:
   O conhecimento descoberto no experimento com o processo padrão será avaliado
   a) subjetivamente: através da validação de especialistas do domínio;
   b) objetivamente: através da construção de sistemas automáticos que utilizem o conhecimento descoberto.

8) Comparar métodos baseados em palavras com métodos baseados em conceitos:
   Devem ser avaliadas as vantagens da abordagem baseada em conceitos sobre a abordagem baseada em palavras.

9) Avaliar a abordagem baseada em conceitos com outras técnicas de mineração:
   Deve-se verificar se é possível usar os conceitos extraídos dos textos com outras técnicas de mineração. No caso, a técnica escolhida foi o agrupamento (*clustering*).

10) Avaliar a possibilidade de descoberta proativa:
    Avaliar o quanto o processo de descoberta de conhecimento pode ser feito de forma automática. Espera-se conhecer que partes do processo podem ser feitas de forma automática e o quanto de intervenção humana é necessária (se é possível iniciar o processo sem hipóteses sobre o que descobrir).

11) Avaliar diferentes aplicações da abordagem:
    O processo de *KDT* proposto deve ser aplicado em diferentes situações e domínios para que se possa avaliar sua utilidade e os tipos de problemas onde pode ser aplicado.

TABELA 3.1 - Objetivos deste trabalho

Objetivo	Atividades Previstas	Resultados Esperados
1) avaliar alternativas para definição de conceitos	- estudo de mecanismos de apoio à definição de conceitos - avaliação e comparação de diferentes métodos de definição de conceitos	- identificação de fatores que influenciam o processo - métodos com melhores desempenhos
2) avaliar métodos de categorização (identificação de conceitos)	- estudo, avaliação e comparação de métodos de classificação aplicados para identificação de conceitos	- identificação de situações onde utilizar os tipos de métodos existentes
3) definir um processo padrão de identificação de conceitos	- definir um processo com ferramentas automatizadas e com os melhores métodos	- processo padrão para identificação de conceitos nos textos
4) definir um processo padrão de mineração sobre conceitos	- definir um processo com ferramentas automatizadas e com os melhores métodos	- processo padrão para mineração sobre conceitos
5) realizar um processo de	- escolher um domínio de aplicação	- conhecimento resultante do

KDT com o processo padrão definido	- realizar um processo de descoberta com o processo padrão	processo
Objetivo	**Atividades Previstas**	**Resultados Esperados**
6) avaliar grau de acerto na identificação de conceitos	- definir medidas de avaliação - avaliar o processo pelas medidas	- margem de erro pelas medidas definidas
7a) avaliar subjetivamente a qualidade do conhecimento descoberto	- apresentar resultados (conhecimento descoberto) para especialistas do domínio	- parecer de especialistas sobre conhecimento descoberto
7b) avaliar objetivamente a qualidade do conhecimento descoberto	- construir um sistema automatizado de decisão usando o conhecimento descoberto - avaliar os resultados do sistema	- nível de acerto do sistema automatizado
8) comparar métodos baseados em palavras com métodos baseados em conceitos	- construir sistemas automatizados com ambas as abordagens - comparar resultados dos sistemas - comparar raciocínio usando palavras X conceitos	- graus de acerto dos sistemas construídos - comparação entre regras de raciocínio
9) avaliar a abordagem baseada em conceitos com outras técnicas de mineração	- selecionar ferramentas que implementem outras técnicas de mineração - definir critérios de avaliação - realizar mineração sobre conceitos e sobre palavras - comparar abordagens (palavra X conceitos) - avaliar resultados para descoberta de conhecimento	- comparação dos resultados da técnica de mineração sobre conceitos X sobre palavras - conhecimento descoberto com a nova técnica de mineração aplicada sobre conceitos
10) avaliar a possibilidade de descoberta proativa	- realizar processo de descoberta sem hipóteses iniciais - investigar intervenção humana - investigar necessidade de conhecimentos prévios	- estratégia para descoberta proativa - estudo da necessidade de intervenção humana e conhecimentos prévios
11) avaliar aplicações da abordagem	- definir aplicações para a abordagem proposta - realizar estudos de casos	- aplicações da abordagem - benefícios práticos

4 DESCOBERTA BASEADA EM CONCEITOS

A proposta desta tese é estudar uma abordagem baseada em conceitos para realizar Descoberta de Conhecimento em Textos (*KDT*). O fundamento básico é aplicar as técnicas tradicionais de mineração de dados (da área de *KDD*) sobre conceitos extraídos de textos, ao invés de trabalhar com palavras (presentes nos textos ou associadas a estes) ou trabalhar sobre valores de atributos.

A abordagem proposta utiliza a **forma intermediária baseada em conceitos**, segundo o esquema de Tan [TAN99]. Neste caso, as representações correspondem a objetos ou conceitos de um domínio e não aos documentos. Para realizar a descoberta de conhecimento, o processo de mineração analisa padrões e relacionamentos entre objetos ou conceitos e não entre documentos. Portanto, o processo geral de descoberta é dependente do domínio.

De acordo com Sowa [SOW00], conceitos são expressos por linguagens (palavras e gramáticas), mas pertencem ao conhecimento extra-lingüístico sobre o mundo. Por isto, a definição de um conceito é determinada pelo ambiente, atividades e cultura das pessoas que falam a língua [SOW00]. Por exemplo, analisando discursos de políticos, alguém pode querer identificar conceitos como "progresso", "problemas", "investimentos", etc. Por outro lado, num ambiente psiquiátrico, conceitos podem ser "violência", "alcoolismo", "suicídio", etc. Soderland [SOD97] utilizou conceitos na área de previsão do tempo. Cada condição climática era um conceito com sua própria definição.

Abordagens baseadas em conceitos (*concept-based approaches*) já são usadas com sucesso na área de Recuperação de Informação (RI). Lin e Chen [LIN96] comentam os benefícios deste tipo de abordagem em relação à busca por palavras-chave. Sua principal vantagem é minimizar o problema do vocabulário. Conceitos representam melhor que palavras os objetos, eventos, sentimentos, ações, etc. do mundo real. Em geral, são usados em áreas como análise de discurso para identificar idéias e ideologias presentes em textos. Chen e outros [CHE94b], por exemplo, usaram com sucesso a identificação de conceitos para organizar idéias discutidas num processo de *brainstorming* eletrônico.

Apesar de o termo "conceito" ser muito usado, é difícil encontrar uma definição formal. Os dicionários apontam sinônimos tais como "*idéia, opinião, pensamento*". Isto confirma a idéia geral e intuitiva de que conceitos são usados para explorar e examinar o conteúdo de palestras, textos, documentos, livros, mensagens, etc.

A estrutura básica do processo de *KDT* proposto neste trabalho aparece na figura 4.1. O primeiro passo é escolher que conceitos são interessantes de serem analisados e definir cada conceito. O segundo passo é a categorização, onde se procura identificar a presença dos conceitos nos textos da coleção em estudo. Após, é possível realizar a mineração, ou seja, a aplicação das técnicas de KDD sobre os conceitos identificados. Esta proposta pode ser considerada dentro do paradigma probabilístico e estatístico, de acordo com a classificação de Mannila [MAN00].

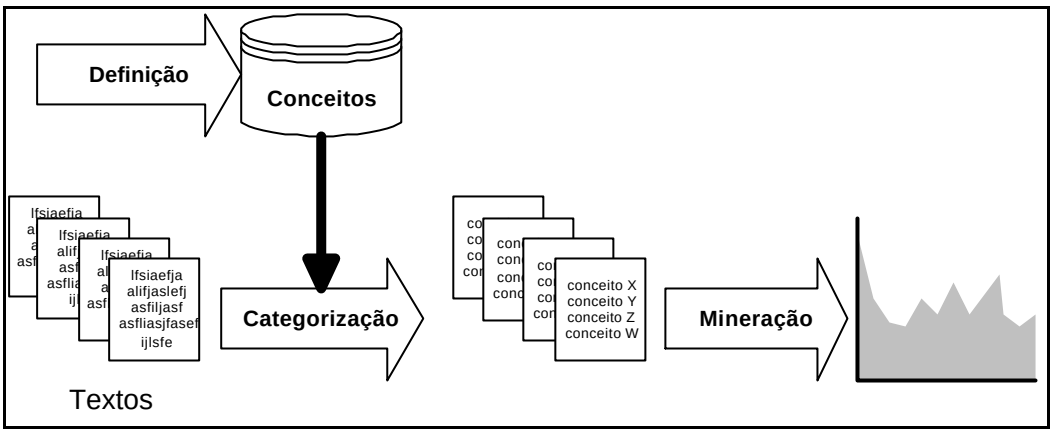

FIGURA 4.1 - Estrutura geral do processo de KDT

De forma resumida, pode-se comparar a abordagem proposta com as etapas de descoberta de conhecimento sugeridas em [GOE99]:

a) entendimento do domínio de aplicação e definição de objetivos para o processo de descoberta: o usuário deve definir que tipo de conhecimento é interessante de ser descoberto;
b) aquisição e seleção de um conjunto de dados: os textos a serem analisados devem ser reunidos em uma coleção;
c) integração e verificação do conjunto de dados: cada texto deve estar contido em um único arquivo; sub-coleções podem ser formadas de acordo com critérios definidos pelo usuário;
d) limpeza, pré-processamento e transformação dos dados: cada texto e seu conteúdo devem ser transformados para representações internas; uma lista de *stopwords* deve ser definida e as mesmas devem ser desconsideradas para análise;
e) desenvolvimento de modelos e construção de hipóteses: definição dos conceitos e identificação dos mesmos nos textos (transformação para a forma intermediária baseada em conceitos);
f) escolha e aplicação de técnicas e métodos de mineração: processo de mineração sobre conceitos;
g) visualização e interpretação dos resultados: pessoas devem interpretar os padrões identificados, usando conhecimento sobre o domínio;
h) teste e verificação dos resultados: avaliação e validação do conhecimento descoberto (de forma subjetiva ou objetiva);
i) uso e manutenção do conhecimento descoberto: aplicação do conhecimento para solução de problemas do domínio (por humanos ou em sistemas automáticos).

Nas seções a seguir, cada etapa da abordagem proposta será discutida através da apresentação das alternativas estudadas para realização da etapa e dos métodos escolhidos, bem como de sua justificativa.

4.1 Representação de Conceitos

O modo como conceitos são representados depende de pontos-de-vista particulares. Nesta proposta, optou-se por uma estrutura simples que permitisse representar objetos, eventos, pensamentos, opiniões e idéias do mundo real de forma fácil e com um grau de fidelidade adequado. Foram usados e testados dois modelos para representar internamente os conceitos:

1) o modelo espaço de vetores (*vector space*), seguindo sugestões de [CHA00], [CHE94b] e [SAL83]; e
2) o modelo contextual, seguindo sugestões de [COH96] e [CHE94b].

Em ambos os casos, parte-se do pressuposto de que conceitos são expressos por palavras, mas que as palavras sozinhas não são adequadas para identificar um conceito [CHE94] [CHE94b]. Isto devido ao problema do vocabulário, discutido anteriormente. Ao examinar estratégias para Recuperação de Informação (RI), Bates [BAT86] concluiu que, para obter sucesso, o usuário que procura informação deve usar uma variedade de termos tão grande quanto a variedade produzida no momento da indexação. Este tipo de redundância permite identificar termos comuns usados pelo autor ou indexador e pelo usuário, no momento de expressar idéias e conceitos. Assim, a eficiência na identificação dos conceitos é maior porque mais termos foram cobertos. Este processo é conhecido como "expansão semântica" e seu sucesso na área de RI foi demonstrado em [BUC94], [IIV95] e [SPA92].

Então, um conjunto suficiente de termos ou palavras deve ser utilizado para representar cada conceito. Em ambos os modelos, os termos descritores de um conceito podem incluir sinônimos, quase-sinônimos (palavras semanticamente relacionadas), variações léxicas (conjugações verbais, verbos e substantivos correlatos, variações em grau e gênero) e outros. Os termos funcionam como "*tokens*", então não é necessário que o termo tenha um significado universal. Assim, podem ser usados nomes próprios, abreviações e siglas específicas do domínio.

4.1.1 Modelo espaço de vetores

No modelo espaço de vetores, cada conceito é representado por um vetor de termos simples. Neste caso, não há relação direta entre os termos e todos são considerados do mesmo nível (vetor não-ordenado e sem conexões entre os termos). A razão desta escolha é que este modelo é o mais simples e facilita as tarefas de definição e identificação dos conceitos. Associado a cada termo no vetor deve haver um peso, descrevendo o grau de importância do termo para descrever ou identificar o conceito. De acordo com Chakrabarti [CHA00], o vetor com pesos é melhor que o modelo binário (sem pesos) porque aumenta a precisão.

A definição do peso de um termo descritor pode seguir a estratégia proposta por Morris [MOR76], na área de Semiótica. Este autor faz distinção entre signos indicadores e signos caracterizadores. Os primeiros apontam para um objeto ou elemento específico, enquanto que os últimos restringem elementos em um conjunto. No caso deste trabalho, o objeto ou elemento é o conceito que se quer descobrir. Assim, termos indicadores devem receber um peso maior, pois possuem maior força para indicar a presença do conceito (nomes próprios, por exemplo). Enquanto que os termos caracterizadores devem receber

pesos relativos menores, pois, apesar de ajudarem a identificar um conceito, não dão certeza de tal.

Lagus e Kaski [LAG99] afirmam que um bom descritor deve caracterizar alguma propriedade importante. Salton e McGill [SAL83] defendem que bons termos descritores são os mais freqüentes dentro de um texto mas infreqüentes na coleção toda (freqüência inversa pequena). Os pesos devem ser normalizados para uma escala entre um e zero, para indicar a força relativa do termo descritor.

Por exemplo, para representar o conceito "*futebol*", o termo "*futebol*" pode receber um grau maior que "*jogador*", uma vez que a presença deste termo indica fortemente a presença do conceito "*futebol*". Já o segundo termo pode aparecer em outros conceitos semelhantes, como "*vôlei*" e "*basquete*", e portanto deve receber um peso menor.

Feldman e Dagan [FEL95] defendem o uso de estruturas simples porque permitem que as tarefas sejam apoiadas por ferramentas automatizadas e porque geram menos esforço.

Entretanto, o problema do modelo espaço de vetores é que o contexto dos termos não é analisado e isto pode levar a interpretações erradas. Por exemplo, o termo "*não*" pode alterar completamente o significado de uma expressão.

Cada conceito deve ter somente um conjunto de descritores, mas um termo pode aparecer em mais de um conceito. No momento, somente termos simples são permitidos neste modelo, devido a limitações computacionais. Entretanto, sabe-se que o uso de pares de termos e expressões complexas melhoram os métodos [APT94]. O uso de termos simples não deve influenciar demais nos resultados, pois o uso exclusivo de termos simples é relativamente eficiente em contrapartida ao uso exclusivamente de pares de termos que implica em resultados mais pobres [APT94].

4.1.2 Modelo contextual

Para minimizar o problema de interpretações erradas, outro tipo de representação foi testada também: o modelo contextual. Neste caso, a relação entre os termos influencia na representação do conceito. A idéia é permitir analisar o contexto em que os termos aparecem no texto para poder entender melhor o significado dos termos e assim poder decidir se um conceito está ou não presente.

Segundo Cohen e Singer [COH96], o contexto pode ser entendido pela análise dos termos próximos (aparecendo antes ou depois). Para representações contextuais, Chen e outros [CHE94b] sugerem uma rede de termos e Cohen e Singer [COH96] sugerem uma lista ordenada de termos.

No modelo contextual proposto nesta tese, a representação de um conceito deve ser feita através de uma ou mais regras, nas quais devem ser indicados termos positivos e termos negativos. Para um conceito estar presente, todos os termos positivos devem estar presentes na frase e nenhum termo negativo pode aparecer. Se uma das regras for verdadeira para a frase sendo analisada, então o conceito está presente na frase e, conseqüentemente, no texto. Por exemplo, no domínio médico, o conceito "alcoolismo" pode ser definido pelas regras (o símbolo "–" indica um termo negativo):

(i) álcool –nega

(ii) hálito etílico

O termo negativo "*nega*" aparece para eliminar frases como "*o paciente nega uso de álcool*".

No modelo contextual, não foram utilizados pesos para as regras, ou seja, não há prioridade ou prevalência de uma regra sobre as outras.

4.2 Definição dos Conceitos

Para criar as representações dos conceitos, é necessário escolher os conceitos que serão empregados no processo de descoberta e descrever cada um de acordo com o modelo escolhido (espaço de vetores ou contextual). Este é um processo de aprendizado e pode ser feito manualmente ou com ajuda de ferramentas de software, segundo Chakrabarti [CHA00].

Uma das hipóteses é que mecanismos de apoio podem facilitar o processo de definição e podem melhorar os resultados finais da descoberta. Entre estes meios de apoio estão incluídos: dicionários técnicos, dicionários gerais, *thesauri*, a intervenção humana e casos-exemplo, para métodos de aprendizado supervisionado (*supervised learning*).

Chen e outros [CHE97] sugerem o uso de vocabulários controlados, tais como dicionários, *thesauri* ou ontologias. Se não existir previamente um vocabulário para o domínio, o mesmo pode ser gerado automaticamente [CHE97].

Uma limitação dos *thesauri* é que eles são estruturas muito rígidas e não suportam variações mesmo que pequenas para apoiar subdomínios específicos (falta de especificidade ou de cobertura de conceitos). Yang e Chute [YAN94] relatam problemas com um *thesaurus* médico porque, na prática diária, eram usados termos específicos que não estavam no *thesaurus*.

Quanto às ontologias, como a WordNet [MIL95], elas possuem a mesma limitação, a qual se faz mais visível quando é necessário utilizar nomes próprios.

Em relação aos dicionários genéricos, Liddy e outros [LID94] já demonstraram seus benefícios. Entretanto, por usarem termos muito genéricos, algumas relações importantes não são encontradas neste tipo de apoio. Por exemplo, o conceito "*futebol*" aparece como "*jogo de bola disputado entre dois times com 11 jogadores cada …*" Neste caso, não aparecem termos importantes relacionados ao conceito, tais como "*campeonato*", "*atacante*", etc.

A geração automática de um vocabulário controlado pode ser feita através de mecanismos de aprendizagem de máquina (processos supervisionados). Estes utilizam exemplos para extrair definições. Entretanto, há a dificuldade de se obter uma amostra de qualidade, com casos-exemplo apropriados e representativos [APT94]. Já os processos não-supervisionados podem ser feitos com a técnica de agrupamento (*clustering*) [ETZ96]. Segundo Fisher [FIS87], o processo de agrupamento recebe descrições de objetos e produz um esquema de classificação a partir de observações sobre relações entre os objetos (aprendizado por observação). O problema dos processos não-supervisionados é que as classes geradas podem não ser de interesse ou apropriadas para o objetivo dos usuários.

Por fim, há a possibilidade de especialistas humanos auxiliarem no processo de definição dos conceitos.

Um dos objetivos deste trabalho é comparar meios de apoio ao processo de definição dos conceitos. Foram escolhidos para avaliação os seguintes mecanismos de apoio: dicionários, *thesaurus*, a intervenção humana e a aprendizagem supervisionada (análise

automática de casos de treino). Por serem muito semelhantes aos *thesauri*, as ontologias não foram investigadas.

No momento da escolha dos termos para definição dos conceitos, é sugerida a remoção dos termos classificados como "*stopwords*", que são termos muito freqüentes e pouco significativos, tais como preposições, artigos, alguns tipos de pronomes, etc [SAL83].

4.3 Identificação dos Conceitos (Categorização)

O objetivo deste processo é identificar os conceitos presentes nos textos. Como os textos não possuem explicitamente conceitos, mas sim palavras, a análise deve partir daí [APT94] [SOW00]. O processo também pode ser chamado de *categorização*, uma vez que é feita a classificação de unidades de textos escritos em língua natural em classes pré-definidas (conforme a definição de categorização de Lewis e Hayes [LEW94]).

Riloff e Lehnert [RIL94] avaliaram 3 métodos de categorização. Dois deles consideram que um conceito está presente se e somente se existe uma palavra ou expressão-chave no texto. Entretanto, estes métodos estão sujeitos a erros por não considerarem o contexto (problema do vocabulário). O terceiro método avaliado por eles analisa o contexto, usando um grau de relevância para decidir se o conceito está ou não presente no texto. Os referidos autores concluíram que a escolha do método depende das características da coleção de textos e da linguagem em que são escritos.

Wiener e outros [WIE95] utilizam redes neurais para realizar a categorização de textos. A estratégia é chamada de "*topic spotting*" por permitir descobrir vários temas presentes nos textos. A desvantagem das redes neurais é que são necessários muitos e bons casos de exemplo, por se tratar de um processo de aprendizado supervisionado.

Ragas e Koster [RAG98] realizaram experimentos usando 4 métodos para categorização: Rocchio, Bayes, *Sleeping Experts* e Winnow. O método Rocchio [ROC66] utiliza um vetor protótipo (um centróide) para representar cada classe ou categoria (nesta proposta, conceitos). O vetor é composto de termos e pesos associados. A avaliação de pertinência de um elemento na classe é feita usando uma função de similaridade (ou de distância) entre os dois vetores representativos. Dependendo do grau de similaridade, o elemento sendo testado será ou não considerado pertencente à classe. Apesar de não ser considerado o melhor método, é o mais simples [YAN99].

Já o método Bayes utiliza uma estratégia semelhante, mas baseada em cálculos probabilísticos [LEW98] [YAN99]. A probabilidade de o elemento pertencer a uma classe é avaliada pela comparação entre os vetores representativos. Neste caso, o centróide (ou vetor protótipo) da classe define os termos que provavelmente aparecem num texto da classe. O peso associado é a probabilidade de o termo aparecer em documentos da classe. Quanto mais termos da classe o texto contiver, maior a probabilidade de ele pertencer àquela classe. O método Naive Bayes assume que não há dependência entre os termos, isto é, a probabilidade de um termo não é condicionada por outro [LEW98].

O método *Sleeping Experts* [RAG98] é semelhante ao Rocchio mas ajusta os pesos dos termos em sessões de treino. Este método funciona melhor com pares e trios de palavras, do que com termos únicos. O método Winnow [RAG98] é semelhante ao *Sleeping Experts*, com a diferença de que os pesos somente são ajustados se forem produzir algum tipo de erro. O resultado final é conseguido após várias iterações, quando os pesos permanecem estáveis.

Nos experimentos de Ragas e Koster [RAG98], os métodos Rocchio e Bayes atingiram melhores resultados. A conclusão destes autores é que estes métodos devem ser utilizados em conjunto.

Entretanto, apesar da sua relativa eficiência e simplicidade, há a desvantagem de que estes métodos não consideram o contexto semântico [COH96], podendo levar a interpretações erradas, como discutido anteriormente.

Um método semelhante e que tem conseguido bons resultados em processos de categorização é o *Latent Semantic Indexing* (LSI) [DEE90] [DUM96]. O método é útil para encontrar termos que caracterizam uma classe (por exemplo, para encontrar o centróide), minimizando assim o problema de sinônimos. Entretanto, há dúvidas de que a polisemia possa ser resolvida [PAP98]. Deerwester e outros [DEE90] afirmam que há uma solução parcial através da análise contextual, mas a conseqüência pode ser enganos ("*false hits*"). Isto ocorre porque o LSI precisa de uma boa amostra de textos para treino (como um método supervisionado) e isto nem sempre é possível. Além disto, as amostras de cada classe devem ser "puras" (cada texto exemplo deve estar associado a somente uma classe) e "separáveis" (deve haver poucos termos comuns a mais de uma classe) [PAP98]. O método LSI funciona bem quando a coleção de textos não muda muito, ou seja, quando o método é treinado com os mesmos textos que serão avaliados para categorização (as coleções de treino e teste são as mesmas).

Yang e Liu [YAN99] analisaram vários métodos de categorização de textos. O método *Support Vector Machines* (SVM) encontra a fronteira ótima para separar os elementos da coleção em dois conjuntos. A limitação é que só trabalha com duas classes. O método *Linear Least Squares Fit* (LLSF) cria um modelo de regressão a partir de casos de treino para caracterizar cada classe, utilizando computações complexas.

O método *k-Nearest Neighbor* (k-NN) decide a categoria de um caso de teste pelas categorias associadas aos seus *k* vizinhos mais próximos. Ou seja, deve-se encontrar, nos casos de treino, os *k* casos que são mais semelhantes e então usar a categoria mais forte entre eles (é feito um cálculo sobre os graus de relacionamento ou utilizada a categoria mais freqüente). Sua aplicação não é recomendada para encontrar temas muito específicos, pois a avaliação de similaridade (geralmente usando uma medida de distância) é feita sobre todo o texto.

Outro método testado em [YAN99] foi o Naive Bayes, que usa probabilidades de categorias e termos para decidir a categoria final. Este é o método mais simples entre os avaliados, exigindo portanto pouca computação. Também foi testada uma Rede Neural (RN), observando-se a desvantagem de exigir muito tempo de treino para alcançar bons resultados.

Os métodos estudados por Yang e Liu necessitam ser treinados para cada domínio específico usando casos-exemplo. Esta é uma desvantagem quando não existem casos de treino em número suficiente e representativos do domínio. Yang e Liu concluíram que, quando há poucos casos de treino por classe (menos de 10), os métodos SVM, LLSF e k-NN atingem melhores resultados no processo de categorização. Já quando há casos de treino em número suficiente (mais de 300), o desempenho dos métodos avaliados é semelhante.

Quando não é possível dispor de casos de treino de boa qualidade, outros métodos podem ser usados. Uma categoria de métodos que não necessitam ser treinados são os métodos baseados em processamento de língua natural (PLN). Nestes métodos as regras de categorização devem ser definidas manualmente. Isto não quer dizer que não seja necessário

analisar exemplos. Casos de exemplo servem para entender como as informações são codificadas na linguagem e no contexto do domínio.

Apesar de eficientes, tais métodos são dispendiosos porque realizam uma análise completa do texto [RIL94] e porque necessitam de muito conhecimento formalmente codificado, na forma de modelos e regras de extração [KNI99]. Chinchor e outros [CHI93] comentam que o esforço (custo) para adaptar os sistemas MUC-3 para um novo domínio são da ordem de 10 a 11 homem/mês, por sistema.

Uma tentativa de mesclar métodos de aprendizado supervisionado com PLN são os "*wrappers*" [MAT99] [ETZ96]. Estes sistemas de extração de informação analisam páginas da Web à procura de padrões lingüísticos (estruturas, palavras-chave, relações entre palavras). Cada padrão determina o tipo de informação que está codificado. Páginas-exemplo são utilizadas para associar automaticamente padrões aos tipos de informação. Apesar de serem usados para extração de informações (para identificar, por exemplo, valores de atributos de um banco de dados), as mesmas técnicas podem ser empregadas para categorização de textos. Assim, quando novas páginas estão sendo analisadas, se certos padrões forem reconhecidos, o documento textual se enquadraria numa determinada categoria.

Contudo, sistemas tipo "*wrappers*" são ainda muito dependentes do domínio e aplicados somente a certos tipos de documentos [GAR99]. Para classificar textos, isto exigiria quase que um sistema específico para cada categoria de texto. Mattox e outros [MAT99] afirmam que o processo de construção de um *wrapper* eficiente exige muito conhecimento semântico sobre o domínio. Isto implica em algum tipo de análise semântica e não somente utilizar técnicas para reconhecimento de padrões.

Assim, ao invés de usar métodos complexos para identificar conceitos nos textos, a proposta desta tese é utilizar técnicas simples mas que permitam algum tipo de análise semântica.

Baseando-se em que os conceitos podem ser identificados por sinais nos textos (termos e relações), segundo os estudos de Riloff e Lehnert [RIL94], as regras de identificação de conceitos podem ser simples, não necessitando realizar análise sintática ou morfológica.

Neste trabalho, foram definidos dois métodos de identificação de conceitos, correspondendo aos dois modelos para representar internamente os conceitos: o modelo espaço de vetores (*vector space*) e o modelo contextual. Ambos os métodos usam uma técnica simples de reconhecimento de padrões e um mecanismo de raciocínio. O reconhecimento de padrões procura identificar termos-chave no texto. Se isto ocorrer, admite-se a hipótese da presença do conceito. Já o mecanismo de raciocínio serve para dar semântica ao processo, permitindo entender o significado dos padrões reconhecidos e, conseqüentemente, possibilitando inferir a presença ou não dos conceitos nos textos. Para esta análise semântica, os métodos levam em consideração o contexto dos padrões reconhecidos.

Nestes métodos, a identificação de conceitos é feita num nível intermediário entre a classificação e a extração de informações (EI). Isto porque as abordagens de classificação, em geral, procuram identificar o tema ou assunto principal de um texto. Nesta tese, pretende-se identificar várias características (no caso, conceitos), numa relação muitos-para-muitos entre textos e categorias. Esta estratégia é semelhante ao problema de "*topic spotting*" discutido em [WIE95]; a diferença reside no objetivo: em [WIE95], deseja-se encontrar os

temas principais no conteúdo de um texto e nesta proposta, o objetivo é identificar a presença de conceitos, os quais podem não ser tão centrais ou importantes no conteúdo.

Por outro lado, a diferença em relação à EI é que as abordagens existentes de EI procuram descobrir valores para atributos, enquanto que, na abordagem proposta, basta identificar a presença de conceitos. Uma das vantagens é amenizar o trabalho de Engenharia de Conhecimento necessário para definir as regras de extração, seja manualmente ou por métodos supervisionados. A aplicação dos métodos propostos neste trabalho também é vantajosa quando não importa a informação exata. Por exemplo, a abordagem serviria bem para se saber se o paciente veio acompanhado de "familiares" mas não interessando saber quem exatamente. [CRO94] discute este problema de incerteza, sugerindo a recuperação de documentos com base em informações *fuzzy*.

Foram testados dois métodos de identificação de conceitos, mas a abordagem geral proposta para *KDT* poderia admitir outro tipo de método de categorização (por exemplo, os analisados em [YAN99]).

Para a definição dos métodos de identificação de conceitos, foram levados em conta os seguintes objetivos:

1) baixa complexidade com desempenho regular:
   Para facilitar a implementação dos métodos, já que não é objetivo da tese usar o melhor método existente, mas mostrar que com certa precisão se pode descobrir conhecimento útil e usá-lo posteriormente com segurança.

2) capacidade de reconhecimento de padrões, mesmo que simples:
   Um mecanismo simples para identificação de termos-chave nos textos.

3) capacidade de análise semântica, mesmo que simples:
   Deve haver um mecanismo de raciocínio com capacidade para analisar os padrões identificados e inferir significado, para avaliar se o conceito está ou não presente.

4) capacidade de análise do contexto, mesmo que simples:
   Considera-se que o contexto de um termo-chave são outros termos presentes no texto todo ou os termos próximos na mesma frase. A análise do contexto permite limitar os significados possíveis dos termos [COH96] ou apontar com maior segurança para determinado objeto ou elemento (no caso, um conceito) [MOR76].

A seguir, serão detalhados os dois métodos propostos e avaliados.

4.3.1 Método baseado no espaço de vetores

Neste método, utiliza-se o modelo espaço de vetores para representar os conceitos. Como cada conceito é definido por um conjunto de termos, o processo de categorização busca encontrar a presença destes termos (sinais) nos textos. Depois, usando um processo de raciocínio *fuzzy*, os pesos dos sinais (termos) encontrados são computados para avaliar a possibilidade de presença do conceito no texto.

Para o reconhecimento de padrões, baseou-se nos métodos Rocchio e Naive Bayes, que são simples e alcançam uma eficiência adequada. O método proposto compara características entre os vetores representativos da classe e do texto.

O tipo de raciocínio utilizado sobre os padrões reconhecidos baseia-se nas sugestões apresentadas em [RIL94]. A hipótese é de que algumas características, quando juntas, indicam a descrição de um evento, com certo grau de confiabilidade (segundo Riloff e Lehnert, um índice de relevância - *relevance index*). Cabe salientar que o contexto de análise é o texto todo e não suas partes.

Assim, o método baseado no espaço de vetores avalia os pesos definidos para os termos identificados e a freqüência destes termos no texto para calcular a possibilidade da presença de um conceito. Isto segue a sugestão de McCarthy [MCC00] quanto ao uso de conceitos aproximados. Segundo o referido autor, existem condições suficientes e condições necessárias para certificar a presença de um conceito numa base de conhecimento. As condições suficientes (CS) implicam na presença obrigatória do conceito (CS → CONCEITO), enquanto que as condições necessárias (CN) são conseqüências da presença do conceito (CONCEITO → CN).

O método aqui proposto para categorização apenas considera as condições necessárias. A função *fuzzy* realiza então um raciocínio abductivo. De acordo com Gulla e outros [GUL97], na dedução, se "A → B" e "A é verdadeiro", então pode-se inferir que "B é verdadeiro". Já na abdução, se "A → B" e "B é verdadeiro", então "A é uma provável causa de B ser verdadeiro". Isto significa que, se termos que definem um conceito aparecem em um texto, então há uma certa possibilidade de que o conceito esteja presente. A idéia é avaliar o contexto (conjunto de termos) para obter a decisão final. Entretanto, a proposta não impede que um conceito seja implicado por único termo (por exemplo, um nome próprio indicando a referência a uma pessoa ou companhia).

Os pesos associados aos termos na definição dos conceitos ajudarão a aumentar ou diminuir o grau com que um termo indica a presença de um conceito. O fundamento é de que cada termo contribui com uma relativa força para a presença do conceito no texto, aumentando a possibilidade desta presença. Indicadores mais fortes devem receber pesos maiores na definição do conceito. Isto funciona como o índice de relevância (*relevancy index*) de Riloff e Lehnert [RIL94]. A decisão se o conceito está ou não presente no texto depende do limiar utilizado para podar valores indesejáveis. A configuração do limiar permitirá que mesmo um termo único indique a presença de um conceito. Esta configuração poderá ser feita numa sessão de treino sobre amostras de textos.

A abordagem aqui proposta está sob o paradigma estatístico de PLN, de acordo com as definições de Knight [KNI99], uma vez que baseia-se em freqüências. Chakrabarti [CHA00] defende o uso de métodos estatísticos afirmando que estes tornam as regras independentes da presença ou ausência de palavras específicas, podendo assim minimizar o problema do vocabulário.

Para facilitar o processo de categorização, os textos devem ser representados em formatos intermediários. Seguindo a sugestão de Salton e McGill [SAL83], a proposta está utilizando o mesmo modelo espaço de vetores. Assim, os textos devem ser previamente analisados para gerar vetores de termos com pesos associados. O peso associado a cada termo é a freqüência relativa do termo no texto, sendo esta calculada pelo número de aparições do termo no texto dividido pelo número total de termos do texto [SAL83].

O método compara todos os textos com cada conceito definido, assumindo que textos e conceitos foram previamente representados por vetores como descrito acima. Esta comparação é feita através de um processo de raciocínio *fuzzy*, seguindo estratégias apresentadas em [ZAD73] e [NAK93]. Os pesos de termos comuns aos dois vetores são multiplicados. A soma total destes produtos, limitada a 1, é o resultado do raciocínio. Este valor indica o grau de relação entre o texto e o conceito, significando o grau de possibilidade de o conceito estar presente no texto ou o grau de importância com que o texto referencia o conceito. Termos não comuns não são contados, não influenciando assim o resultado, uma vez que podem estar sendo utilizados sinônimos.

A fórmula a seguir define precisamente tal processo de raciocínio:

[conceitos **X** textos**] = [**conceitos **X** termos**]** **O** **[**termos **X** textos**]**

sendo que:
- o símbolo **O** representa uma combinação entre relações *fuzzy*, utilizada para realizar a inferência (regra de inferência composicional, conforme Nakanishi e outros [NAK93]);
- os símbolos **[ ]** delimitam uma relação *fuzzy* (que pode ser associada a uma matriz).

A relação *fuzzy* resultante da combinação **O** segue o raciocínio *fuzzy* de Nakanishi e outros [NAK93], onde

$R \circ S: \mu_{R \circ S}(x,z) = \bigvee \{ \mu_R(x,y) \wedge \mu_S(y,z) \}$

Na combinação **O**, os operadores utilizados para as disjunções e conjunções das relações *fuzzy* são:

Ú $\Rightarrow$ soma limitada = min(1, x + y), já que os termos de um conceito que não aparecem em um texto e os termos de um texto que não fazem parte do conceito não devem diminuir o grau da relação entre o conceito e o texto, pois podem estar sendo usados sinônimos para estes termos;

Ù $\Rightarrow$ produto algébrico = (x * y), para que ambos os pesos sejam computados (o do termo no conceito e a freqüência relativa do termo no texto), uma vez que ambos são importantes para o resultado final.

4.3.2 Método contextual

O método contextual é semelhante ao anterior por também apresentar reconhecimento de padrões baseado na identificação de termos-chave. Só que neste caso, não há vetores representando texto e conceito. Os textos são analisados em sua forma original. Já os conceitos são definidos por uma ou mais regras de identificação.

O mecanismo de raciocínio é diferente. Primeiro, porque o contexto não é mais o texto todo mas cada frase individual. Isto quer dizer que somente os termos encontrados na frase serão analisados em conjunto para decidir se o conceito está ou não presente na frase e, conseqüentemente, no texto. Isto ajuda a limitar as possibilidades de significado, uma vez que

o texto todo é um contexto muito grande, podendo ainda gerar problemas de vocabulário (um contexto mais restrito melhora a precisão da interpretação semântica).

A segunda diferença é que as regras são definidas com termos positivos e negativos. Assim, o método avalia a combinação dos termos positivos e negativos definidos em cada regra em relação a cada uma das frases do texto. Para um conceito estar presente, todos os termos positivos devem estar presentes na frase e nenhum termo negativo pode aparecer. Se uma das regras for verdadeira para a frase sendo analisada, então o conceito está presente na frase e, conseqüentemente, no texto.

A presença de mais de um termo positivo na mesma frase, melhora a precisão. Por exemplo, os termos "*sistemas*" e "*operacionais*" na mesma frase geram um significado bem diferente de quando os mesmos termos aparecem no mesmo texto mas de forma independente (em locais distantes, por exemplo).

Também se consegue melhor precisão com a definição de termos negativos, tais como "*não*", "*nega*" e "*ao contrário de*", cuja presença inverte o significado de uma expressão. Este mecanismo de raciocínio com termos positivos e negativos reduz os problemas de vocabulário. Cabe salientar que não é feita nenhuma análise sintática nas frases, portanto a ordem ou relação entre os termos positivos e negativos não é considerada para efeito da categorização.

Outra diferença para o método anterior é que não há um grau de certeza (ou incerteza) quanto à presença do conceito (como o índice de relevância); ou o conceito está presente ou não está (valor binário para cada frase). Entretanto, como todas as frases são comparadas contra todos os conceitos (e todas as suas regras), se um conceito estiver presente mais de uma vez no texto, este valor poderia ser usado para indicar o quanto um conceito é referenciado no texto.

No momento, os textos são analisados no seu estado original, sem que uma representação intermediária seja criada. Entretanto, seria possível criar uma representação mais econômica usando uma lista ou vetor de termos, desde que houvesse um separador de frases nesta representação.

4.4 Mineração sobre Conceitos

O processo de mineração aplica técnicas estatísticas sobre os conceitos extraídos na etapa de categorização (independente do método usado). As técnicas aplicadas nesta abordagem são as mesmas já existentes na área de *KDD*, somente que elas devem ser utilizadas sobre os conceitos identificados nos textos e não sobre itens de um banco de dados, como faz a área de *KDD*.

No momento, a mineração não está utilizando o grau de relacionamento entre conceitos e textos, mas sim trabalhando com valores binários. Ou seja, basta saber se o conceito está presente ou não, e não o quanto (para ambos os métodos de categorização).

Duas técnicas específicas estão sendo usadas nesta proposta: a análise de distribuição e a técnica associativa.

A **técnica de análise de distribuição** (ou lista de conceitos-chave) verifica a freqüência com que ocorrem os conceitos num conjunto de textos (pode ser a coleção toda ou parte dela). O resultado é um tipo de centróide (um vetor de conceitos e suas freqüências).

Isto permite analisar que temas são mais dominantes e quais aparecem menos. Também é possível comparar um centróide com outro (por exemplo, centróides de duas subcoleções diferentes). Assim, podem ser encontrados temas comuns em duas coleções ou temas exclusivos e também disparidades ou similaridades nas freqüências dos conceitos.

Já a **técnica associativa** descobre relações ou associações entre conceitos, expressando os resultados na forma de regras X ➔ Y (X pode ser um ou mais conceitos e Y somente um conceito). A regra significa que "*se* **X** *está presente em um texto, então* **Y** *também está presente com um certo grau de certeza*".

O grau de certeza é dado por valores de confiança (*confidence*) e suporte (*support*). De acordo com a analogia proposta em [LIN98] e [GAR99], os textos (ou documentos) são tratados como transações e os conceitos como os itens do banco de dados. Assim, a interpretação do grau de *confiança* (*confidence*) para uma regra associativa do tipo X ➔ Y é a proporção de textos que possuem X e Y em relação ao número de textos que possuem somente X. Da mesma forma, o *suporte* da mesma regra (*support*) é interpretado como o número de documentos onde X e Y estão presentes (ou a proporção em relação à coleção toda). O grau de confiança funciona como uma probabilidade condicional. Isto permite predizer a presença de um conceito em função da presença de outro.

Nem todas as regras resultantes são importantes, novas ou úteis. Para realizar este filtro, devem ser definidos limiares para os valores de confiança e suporte. Como medidas de "*interestingness*" (que mede o quanto uma descoberta é interessante), Feldman e Dagan [FEL98] sugerem analisar as distribuições (freqüências) que diferem significativamente de valores da coleção toda ou de outra coleções (partições por características ou por períodos de tempo). As comparações entre as subcoleções ou entres estas e a coleção toda servem também para encontrar regras comuns e regras exclusivas. Feldman e Dagan [FEL95] também medem o grau de "*interestingness*" comparando as distribuições com um algum modelo esperado (avaliando diferenças ou semelhanças). Para as regras associativas, Feldman e Hirsh [FEL97] utilizam o suporte mínimo de 5 documentos e como confiança mínima o valor de 10%.

A estratégia de mineração utilizada pode ser considerada dentro do paradigma probabilístico e estatístico de acordo com Mannila [MAN00], uma vez que é baseada na distribuição de variáveis na coleção.

5 EXPERIMENTOS

Para realizar a avaliação dos métodos e da abordagem proposta, foram desenvolvidos experimentos em um domínio específico: a área de psiquiatria.

A seguir, serão explicados o domínio de aplicação e suas características, a coleção de textos usada nos experimentos e os conceitos usados no processo de mineração.

5.1 Domínio de Aplicação

O domínio de psiquiatria tem características especiais. Primeiro, o processo de diagnóstico é mais complexo do que em outras especialidades médicas. Sintomas e sinais podem estar presentes em diferentes doenças e não há sintomas exclusivos de uma única doença.

Isto quer dizer que é necessário analisar mais que simplesmente sintomas e sinais. Deve-se levar em conta o contexto histórico, social e comportamental do paciente.

Para os experimentos foram contatados profissionais de uma clínica psiquiátrica de relativo porte, os quais ajudaram nos processos de coleta de textos, definição de conceitos, mineração e validação dos resultados.

5.2 Coleção de Textos Usada

Nos experimentos, foram utilizados os prontuários de internação dos pacientes. Estes prontuários são textos escritos pelos médicos com informações colhidas na entrevista de internação do paciente. Desta entrevista, pode também participar algum familiar ou a pessoa que acompanha o paciente.

Estes prontuários podem ser considerados documentos semi-estruturados, pois os médicos possuem uma certa orientação do que registrar e, em alguns casos, podem mesmo utilizar de códigos para identificar o conteúdo das partes do documento. Entretanto, não há rigidez e alguns médicos podem mesmo não seguir esta orientação. Apesar de os documentos serem semi-estruturados, os textos registrados são livres, isto é, escritos em linguagem natural irrestrita, sem formatos ou padrões pré-definidos e sem um vocabulário controlado. Nos experimentos, a estrutura dos documentos não foi usada, pois o objetivo era trabalhar com textos livres.

Estes textos formam parte do registro do paciente e contém informações sobre o comportamento social e familiar do paciente, história pregressa, remédios que toma ou que foram prescritos, além de sinais e sintomas identificados pelo médico durante a entrevista. Exemplos de prontuários encontram-se no anexo 1.

Foram coletados 400 prontuários, correspondendo a internações feitas na clínica durante aproximadamente 4 meses. Cada prontuário é único, contendo um texto diferente dos demais. Entretanto, pode haver mais de um prontuário referente ao mesmo paciente, uma vez que são admitidas reinternações.

A coleção foi dividida em duas partes, com 200 textos cada. A primeira parte foi usada para treino dos métodos (e é portanto assim chamada). A segunda subcoleção foi utilizada para teste dos métodos.

Para cada prontuário havia associado um diagnóstico, para representar a doença mental do paciente e decidido por um médico da clínica em um processo real e prévio de diagnóstico. Entretanto, a indicação do diagnóstico não estava explicitamente expressa no texto.

A classificação usada para o diagnóstico segue as regras da Classificação Internacional de Doenças, décima revisão (CID-10) [CEN89]. Foram usadas somente as classes de primeiro nível do CID-10, as quais correspondem aos diagnósticos mais freqüentes na clínica estudada, a saber (os nomes em negrito servirão para identificar futuramente cada uma das classes):

a) transtornos mentais **orgânicos**, incluindo diagnósticos de códigos F00 a F09 do CID-10 e os sub-níveis;
b) transtornos mentais e comportamentais devidos ao uso de **substâncias** psicoativas, incluindo códigos de F10 a F19 e sub-níveis;
c) **esquizofrenia**, transtornos esquizotípicos e transtornos delirantes, incluindo códigos de F20 a F29 e sub-níveis;
d) transtornos do humor (**afetivos**), incluindo códigos de F30 a F39 e sub-níveis.

A distribuição das classes nas duas subcoleções é semelhante. A primeira subcoleção (para treino) era composta de: 27 textos da classe "*afetivos*" (13.5%), 103 textos de "*esquizofrenia*" (51.5%), 18 texto da classe "*orgânicos*" (9%) e 52 textos de "*substâncias*" (26%). A segunda subcoleção (para teste) continha: 25 textos de "*afetivos*" (12,5%), 105 textos de "*esquizofrenia*" (52,5%), 17 textos de "*orgânicos*" (8,5%) e 53 textos de "*substâncias*" (26,5%).

Todos os documentos tinham entre 1 e 4 Kbytes de tamanho, com um mínimo de 22 termos e um máximo de 413 termos (incluindo códigos, números e *stopwords*).

Os textos foram analisados sem nenhum tipo de correção. Isto quer dizer que erros ortográficos ou de digitação foram mantidos e fizeram parte do processo.

5.3 Conceitos Usados

Nos experimentos, foram usados conceitos de dois tipos. Um tipo referenciava conceitos mais gerais e outro, conceitos mais específicos. Os conceitos gerais correspondiam às classes dos documentos, ou seja, eram 4 conceitos referentes aos 4 grandes diagnósticos do CID-10 (orgânicos, esquizofrenia, afetivos e substâncias).

No segundo tipo, foram utilizados 65 conceitos referentes a características do paciente e 32 conceitos referentes a remédios.

Os conceitos mais específicos (características dos pacientes) incluíam referências a:

- sintomas e sinais, tais como inapetência, agressividade, insônia, tabagismo, uso de álcool, tentativa de homicídio, ideação suicida, lesões, dor de cabeça;
- pessoas relacionadas ao paciente, como marido, esposa, mãe, pai, tios, outros familiares, amigos, vizinhos, etc.;
- objetos, tais como faca, arma;

- características de comportamento, como “morar sozinho”, falar sozinho (solilóquio), andar de um lado para outro (dromomania);
- eventos, como enforcamento, morte, fuga, envenenamento;
- outras referências, como trabalho, religião, fogo.

Estas referências podiam ser do paciente ou de pessoas relacionadas, bem como podiam ter sido relatadas no momento mas acontecido há muito tempo atrás. A interpretação do significado dos conceitos depende de como foram definidos e de como podiam ser identificados nos textos. Por exemplo, o conceito “alcoolismo” (uso de bebidas alcoólicas) podia referir-se ao paciente ou a alguém citado no texto. Para restringir o significando, somente permitindo a identificação do conceito no texto se fosse uma característica do paciente, foi preciso refinar as regras de identificação. Conseguiu-se uma melhor precisão, mas mesmo assim não se pode ter completa certeza de que a referida característica é do paciente, se o conceito for encontrado no texto.

Os conceitos específicos foram selecionados entre as características que apareciam em dicionários técnicos de psiquiatria e no documento CID-10 [CEN89]. Também foram criados conceitos para informações encontradas nos textos que não estavam referenciadas nos dicionários ou no CID. Para tanto, foram avaliadas amostras de textos, analisando termos mais freqüentes e seu significado. Médicos especialistas ajudaram no refinamento final.

O objetivo da seleção de conceitos era ter referências a eventos, objetos, pessoas ou características que pudessem caracterizar pacientes e doenças.

5.4 Processo Padrão para Descoberta de Conhecimento

Foi definido um processo padrão para realizar a descoberta de conhecimento na coleção de textos selecionada. O processo foi realizado sobre os conceitos específicos.

Para a definição e identificação de conceitos específicos nos textos, foram utilizados os métodos que obtiveram melhores resultados (conforme avaliações a serem discutidas no próximo capítulo). Este processo está sob o paradigma probabilístico e sob o modelo bayesiano, conforme a classificação proposta em [CHA00].

Foram utilizados os seguintes mecanismos de apoio para definição dos conceitos específicos: *thesaurus*, dicionários e a intervenção humana. A aprendizagem supervisionada, apesar do bom desempenho, não foi empregada por falta de casos de treino, uma vez que foi difícil selecionar expressões lingüísticas referentes aos conceitos específicos. Entretanto, o fundamento básico desta técnica, que é aprender com casos de treino, foi usada juntamente com a intervenção humana (pessoas realizaram a análise dos casos de treino).

Nos experimentos, os conceitos específicos (características dos pacientes) foram definidos manualmente com auxílio de ferramentas de software. O *thesaurus* médico CID-10 foi tomado como base para as definições. Dicionários técnicos da área de psiquiatria e dicionários da língua portuguesa ajudaram na definição de termos sinônimos. As ferramentas de software auxiliaram a encontrar termos sinônimos ou correlatos que não apareciam nos dicionários mas eram usados nos textos. Ferramentas de software também permitiram analisar o contexto em que os termos eram usados. Alarmes falsos (“*false hits*”) também foram analisados com as ferramentas. Isto permitiu refinar as definições (por exemplo, utilizar termos negativos), melhorando assim o desempenho do processo de identificação. Dois profissionais

ligados à área (não especialistas) ajudaram no processo de definição, o qual levou aproximadamente 30 horas, no total, durante 2 meses.

Para a identificação dos conceitos nos textos, foi empregado o método baseado no modelo contextual. Este processo de categorização não levou mais que 40 minutos no ambiente computacional descrito na seção 5.6 (lembrando ser uma comparação de 97 conceitos contra 400 textos).

Depois, as técnicas de análise de distribuição e associação foram aplicadas sobre os conceitos associados aos textos. Para a filtragem de regras interessantes, foi utilizado um limiar de 80% para o grau de confiança e de 37% para o suporte.

A mineração se assemelha ao processo de aprendizagem supervisionada, já que procura descobrir algo de novo em documentos existentes. Entretanto, não se podia tomar como verdadeiro o resultado do processo de mineração sobre a coleção de treino. Era preciso validar o conhecimento descoberto.

Deste modo, dividiu-se a mineração em duas etapas. Primeiro, as técnicas de mineração foram aplicadas sobre a coleção de treino, resultando num conjunto de padrões interessantes, mas ainda não considerados verdadeiros. Depois, as mesmas técnicas foram aplicadas na coleção de teste e estes resultados comparados com os anteriores. Aqueles padrões que apareciam em ambos os resultados ou que eram semelhantes nas duas coleções foram considerados verdadeiros (conhecimento descoberto).

Com a técnica de análise de distribuição, foram aceitas como semelhantes as distribuições com variação menor que 10 pontos.

Com a técnica associativa, o processo foi um pouco mais complexo. Se a técnica fosse aplicada sobre toda a coleção, as regras resultantes poderiam ser influenciadas pelos casos de **esquizofrenia**, que eram metade da coleção. Assim, a técnica foi aplicada a cada classe separadamente, ou seja, procurando descobrir regras associativas sobre os 4 grupos de textos de forma independente. Depois, as regras comuns a todas as classes foram tomadas como representantes da coleção toda.

O processo de mineração levou menos de 2 minutos.

5.5 Conhecimento Descoberto

A seguir, são apresentados os resultados do processo de descoberta de conhecimento sobre a coleção de textos selecionada.

5.5.1 Técnica de análise de distribuição

A figura 5.1 a seguir apresenta os conceitos com distribuições mais freqüentes (acima de 50%), para a coleção toda. Tal conhecimento permite traçar o perfil do paciente típico que se interna na clínica estudada.

Um grupo especial de casos chamou a atenção: o dos pacientes reinternados (33% dos casos). Procurou-se comparar os padrões destes casos com o padrão da coleção toda. Foram consideradas interessantes as variações superiores a 10 pontos. Descobriu-se que:

- "*insônia*" aumentou de 71% para 83,3% (coleção toda para reinternados);
- "*homicida*" aumentou de 36,5% para 45,5%;
- "*fala doente*" diminuiu de 35% para 25,8%.

familiares (84,5%) agressividade (77,0%) inapetência (76,0%) remédios (74,5%)	insônia (71,0%) alteração de pensamento (70,5%)	nervosismo (68,5%) alteração de atenção (54,5%)

FIGURA 5.1 - Conceitos mais freqüentes na coleção toda

Analisando as distribuições dos conceitos entre as classes, descobriu-se que:

1- "*atenção doente*" é mais freqüente nos **afetivos;**
2- "*suicida*" aparece mais nos **afetivos** (81,5% contra 38.8%, 16.7% e 30.8%);
3- "*depressão*" aparece mais nos *afetivos* (74.1% contra 11.7%, 11.1% e 25%);
4- **afetivos** e **substâncias** tem distribuições semelhantes para alguns conceitos ("*insônia*", "*inapetência*", "*nervoso*", "*agressividade*"), mas diferem por "*alcoolismo*" alto no segundo e baixo no primeiro e por "*depressão*", "*suicida*" e "*choro*", o contrário;
5- "*autismo*" aparece em **esquizofrenia** (37.9%) e um pouco em **orgânicos** (16.7%);
6- "*alcoolismo*" aparece mais em **substâncias** (94.2%), mas também aparece nos demais (25.9%, 16.5% e 11.1%);
7- "*consciência normal*" e "*consciência doente*" tiveram a mesma distribuição em **substâncias** (17.3%) e em **orgânicos** (11.1%), enquanto que em **afetivos** e **esquizofrenia**, o primeiro foi mais freqüente que o segundo (40.7% x 7.4%; 32% x 13.6%);
8- referências a "*morte*" aparecem com freqüência média em **afetivos**, **esquizofrenia** e **orgânicos** (40.7%, 35% e 33.3%) e baixo em **substâncias** (17.3%);
9- "*negativismo*" é baixo em **substâncias** (17.3%) e médio nos demais (29.6%, 38.8% e 38.9%);
10- "*alucinações visuais*" e "*ver bichos*" (*zoopsias*) não aparecem em **orgânicos**;
11- "*lesões*" são mais altas em **orgânicos** (38.9%) contra os demais (18.5%, 13.6% e 21.2%);
12- "*morar sozinho*" não aparece em **orgânicos** e é baixo nos demais (14.8%, 8.7% e 3.8%);
13- "*casamento*" não aparece em **orgânicos**, nem "*marido*" e "*esposa*";
14- "*pueril*" não aparece em **substâncias**;
15- "*mania*" é média em **esquizofrenia** (29.1%), baixa em **afetivos** e **substâncias** (7.4% e 5.8%) e não aparece em **orgânicos**;
16- "*dromomania*" é média em **esquizofrenia** (25.2%), baixa em **afetivos** (7.4%) e não aparece em **orgânicos** e **substâncias**;
17- "*tremores*" aparecem bastante em **substâncias** (40.4%), pouco em **afetivos** (7.4%) e não aparecem em **esquizofrenia** e **orgânicos**;
18- "*fumar*" não é relatado em **orgânicos**;
19- "*delírios*" não aparece em **afetivos**.

Um estudo especial foi feito sobre cada remédio específico, procurando-se encontrar os conceitos associados a cada remédio. Como exemplo, a seguir são apresentados os conceitos

mais freqüentes associados ao remédio Dienpax e suas respectivas distribuições na subcoleção (37 textos):"inapetência" (91,8%), "*agressividade*" (83,7%), "*alteração de pensamento*" (78,3%), "*nervosismo*" (75,6%), "*insônia*" (64,8%), "*alcoolismo*" (62,1%), "*ouvir vozes*" (59,4%).

5.5.2 Técnica associativa

A figura 5.2 a seguir apresenta as regras associativas comuns aos 4 diagnósticos.

alteração de atenção ➔ familiares
alteração de atenção ➔ alteração de pensamento
inapetência ➔ familiares
insônia ➔ familiares
alteração de pensamento ➔ familiares
remédios ➔ familiares

FIGURA 5.2 - Regras associativas comuns aos 4 diagnósticos

As figuras 5.3, 5.4, 5.5 e 5.6 apresentam as regras exclusivas de cada classe.

agressividade ➔ inapetência
alteração de pensamento ➔ inapetência
trabalho ➔ inapetência

FIGURA 5.3 - Regras associativas exclusivas da classe substâncias

ouvir vozes ➔ agressividade
mania de perseguição ➔ insônia
ouvir vozes ➔ insônia
ouvir vozes ➔ nervosismo
mania de perseguição ➔ alteração de pensamento
ouvir vozes ➔ alteração de pensamento

FIGURA 5.4 - Regras associativas exclusivas da classe esquizofrenia

lesões ➔ agressividade
lesões ➔ nervosismo

FIGURA 5.5 - Regras associativas exclusivas da classe orgânicos

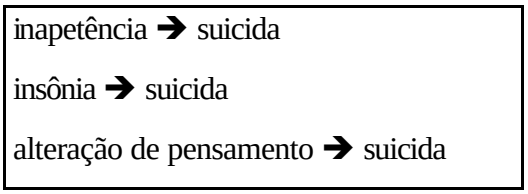

FIGURA 5.6 - Regras associativas exclusivas da classe afetivos

5.6 Ambiente Computacional Usado nos Experimentos

Para realização dos experimentos, foram utilizadas ferramentas de software implementadas em Delphi 4 e um microcomputador PC equipado com processador Pentium II de 400 MHz, 64 Mbytes de memória RAM e sistema operacional Windows 98 SE.

6 RESULTADOS DAS AVALIAÇÕES FEITAS

Neste capítulo, serão apresentados os resultados dos experimentos, bem como observações e conclusões iniciais, divididos por objetivos.

Os resultados foram medidos utilizando os graus de abrangência (*recall*) e precisão (*precision*) [LEW91]. Admitindo que ambas as medidas são igualmente importantes, utilizou-se ainda a medida F (*F-measure*), calculada por 2 * Pr * Rc / (Pr + Rc) [LEW94b].

Já que havia 4 classes, foram também utilizadas as medidas de *microaveraging* e *macroaveraging* sugeridas por Lewis [LEW91] para se obter o resultado geral de cada método para a coleção toda. *Microaveraging* considera a coleção toda como uma única classe e então avalia os graus de abrangência e precisão sem distinção de classes. Já a medida de *Macroaveraging* primeiro calcula a precisão e abrangência em cada classe e então extrai os valores médios para a coleção toda.

Para comparação entre os métodos, uma medida adicional foi usada para encontrar o melhor desempenho: a média (MED) entre os valores de *Microaveraging F-measure* e *Macroaveraging F-measure*.

6.1 Avaliação dos Métodos para Definição de Conceitos

Para este experimento, foram considerados como conceitos as 4 grandes classes de diagnósticos da área de psiquiatria (afetivos, esquizofrenia, orgânicos e substâncias), conforme o CID-10.

Foram avaliados os seguintes mecanismos de apoio para o processo de definição de conceitos:

- dicionários: servem para a identificação de sinônimos; foram usados dicionários técnicos da especialidade de psiquiatria e dicionários gerais da língua portuguesa;
- *thesaurus*: foi utilizado o CID-10, somente a parte referente à especialidade de psiquiatria; neste documento, cada doença é descrita através de sinais, sintomas e outras características importantes para o diagnóstico;
- a intervenção humana: através da análise de amostras de textos para encontrar sinônimos; a intervenção humana também foi testada para melhorar os métodos através da análise de exemplos positivos e negativos; e
- a aprendizagem supervisionada: foi usada para definição automática de conceitos pela análise de casos de treino.

Como no processo é necessário associar pesos aos termos definidos para um conceito, foi avaliada também a influência destes pesos. Pretende-se investigar a possibilidade de que os pesos sejam gerados de forma automática, com base nas freqüências dos termos dentro da coleção ou nos vocabulários controlados.

Os mecanismos de apoio não foram avaliados sobre os conceitos específicos (características do paciente) por não haver conceitos correspondentes no CID e por ser difícil encontrar casos de treino para cada conceito específico.

A seguir, serão descritos os métodos avaliados para definição dos conceitos gerais (classes do CID). A numeração principal indica o tipo de apoio principal utilizado e as letras servem para indicar variações.

1) Métodos tipo 1: o objetivo era testar um vocabulário controlado como mecanismo de apoio; nestes métodos, o *thesaurus* CID foi usado como base para geração das descrições dos conceitos;

a) CID automático: todos os termos presentes nas descrições dos conceitos foram usados (menos as *stopwords*), com o peso sendo a freqüência relativa do termo na descrição;
b) CID automático com poda: neste método, somente os termos mais freqüentes nas descrições foram usados (entre 70 e 85 mais freqüentes);
c) CID com termos selecionados por humanos: pessoas leigas intervieram no processo, selecionando termos para descrever cada conceito; foram eliminados termos muito genéricos e aqueles que levavam a erros, segundo amostras de casos; todos os termos receberam peso 1;
d) CID usando só as diferenças: foram usados somente os termos que aparecem em uma das classes;
e) CID com peso igual: usando os mesmos termos do método (b) mas todos com peso 1;
f) CID diferenças com peso igual: usando os mesmos termos do método (d) mas todos com peso 1;
g) CID pesos diferenciados: usando os termos do método (d) com peso igual a 1 e os termos do método (b) com peso original (freqüência relativa); assemelha-se à soma dos métodos (b) e (f).

2) Métodos tipo 2: o objetivo era testar a influência de sinônimos e da intervenção humana; nestes métodos, o *thesaurus* CID foi usado como base, mas também foram usados termos sinônimos que não apareciam no CID, os quais foram extraídos de dicionários técnicos de psiquiatria e dicionários da língua portuguesa; a intervenção humana se caracteriza pela seleção de termos:

a) CID com sinônimos: sinônimos e variações léxicas dos termos presentes no CID foram usados para completar as descrições dos conceitos; peso 1 para todos os termos;
b) CID com sinônimos e intervenção humana: pessoas leigas analisaram amostras de textos e alguns erros de classificação e refizeram as definições de (a); termos muito genéricos foram eliminados e sinônimos identificados nas amostras foram acrescentados; peso 1 para todos os termos;
c) CID com sinônimos e pesos negativos: partindo das definições de (b), foram associados pesos negativos aos termos do conceito que estavam gerando erros na classificação.

3) Métodos tipo 3: o objetivo era testar um mecanismo de aprendizagem supervisionada; nestes métodos, os conceitos foram definidos automaticamente através da análise de casos positivos; os termos a serem usados nas definições dos conceitos foram extraídos por ferramentas automatizadas analisando os textos referentes a cada classe na coleção de treino:

a) automático todos: foram usados todos os termos (menos as *stopwords*) que apareciam em dois ou mais textos nos casos de treino;
b) automático diferenças: foram usados somente os termos que apareciam numa única classe.

A tabela 6.1 a seguir, apresenta o tempo aproximado de categorização para cada método (sobre a coleção de treino) e o número de termos usados para descrever cada conceito (classe).

TABELA 6.1 - Tempo de categorização e número de termos por método

Método	Tempo (em min.)	Orgânicos	Substâncias	Esquizofr.	Afetivos
1a	30	313	304	440	246
1b	20	85	75	70	73
1c	15	89	68	116	70
1d	30	148	165	252	125
1e	16	85	75	70	73
1f	30	148	165	252	125
1g	50	233	240	322	198
2a	6	25	24	38	17
2b	5	21	24	36	12
2c	7	21	24	36	12
3a	120	403	993	1673	668
3b	60	28	241	773	64

A avaliação dos métodos de definição de conceitos foi feita sobre os resultados de classificação usando estas definições. O método de classificação usado foi o baseado no espaço de vetores, como explicado anteriormente. Não foi usado o método contextual por ser difícil definir conceitos de modo automático com o modelo contextual de representação de conceitos.

A classificação foi feita sobre as duas coleções de prontuários, separadamente (treino e teste). Isto permite avaliar o potencial de cada método de definição em relação ao uso de casos de treino ou não. A coleção de treino foi utilizada para análise de amostras de textos (com a intervenção humana) e para a geração automática das definições (nos métodos do tipo 3). A análise de amostras permitiu descobrir novos sinônimos, que não apareciam em dicionários ou no CID, mas que eram utilizados na prática pelos médicos, quando gerando o prontuário. Também foi possível corrigir erros analisando textos classificados erroneamente (os chamados "*false hits*").

Foram testados vários limiares de poda na decisão final de classificação. Nas tabelas, apenas aparecem os limiares que geraram melhores resultados. Nem todas as faixas de valores foram testadas nos limiares, mas somente as que mais se aproximavam do número ideal de documentos recuperados. Quando limiares foram usados, era permitido associar mais de uma classe (conceito) a cada texto.

Os métodos também foram testados com classificações por maior peso. Ou seja, tomava-se como decisão final somente a classe com maior grau de associação com o texto. Neste caso, cada texto só estava relacionado a uma classe ou categoria.

As tabelas referentes a “textos associados a nenhuma classe” só dizem respeito a classificações por maior peso. Neste caso, há ocorrência de textos que não se relacionavam a nenhuma classe (grau zero para todas). No caso dos limiares, o resultado associa uma classe pelo menos a cada texto.

TABELA 6.2 - Resultados com a coleção de treino

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
1a limiar 0,0012	0,46	0,51	0,46	0,38	0,46	0,44	0,45
1a limiar 0,0014	0,51	0,52	0,35	0,28	0,42	0,36	0,39
1a maior peso	0,49	0,50	0,49	0,36	0,49	0,42	0,45
1b limiar 0,0010	0,44	0,45	0,38	0,33	0,41	0,38	0,39
1b limiar 0,0012	0,50	0,52	0,29	0,22	0,37	0,31	0,34
1b limiar 0,0014	0,57	0,55	0,23	0,19	0,33	0,28	0,30
1b maior peso	0,47	0,48	0,47	0,39	0,47	0,43	0,45
1c limiar 0,025	0,36	0,31	0,47	0,39	0,41	0,35	0,38
1c maior peso	0,45	0,47	0,44	0,37	0,44	0,41	0,42
1d limiar 0,00015	0,36	0,34	0,49	0,45	0,42	0,39	0,40
1d maior peso	0,37	0,35	0,37	0,34	0,37	0,34	0,35
1e limiar 0,03	0,28	0,31	0,62	0,61	0,39	0,41	0,40
1e limiar 0,05	0,34	0,43	0,25	0,26	0,29	0,32	0,30
1e limiar 0,07	0,31	0,50	0,07	0,06	0,11	0,11	0,11
1e maior peso	0,32	0,38	0,32	0,35	0,32	0,36	0,34
1f limiar 0,02	0,32	0,26	0,62	0,51	0,42	0,34	0,38
1f limiar 0,025	0,36	0,28	0,47	0,36	0,41	0,32	0,36
1f limiar 0,03	0,36	0,27	0,30	0,24	0,33	0,25	0,29
1f maior peso	0,37	0,27	0,37	0,27	0,37	0,27	0,32
1g limiar 0,02	0,31	0,26	0,64	0,53	0,42	0,35	0,38
1g limiar 0,025	0,36	0,28	0,49	0,37	0,42	0,32	0,37
1g limiar	0,36	0,27	0,32	0,25	0,34	0,26	0,30

0,03							
1g maior peso	0,39	0,28	0,39	0,28	0,39	0,28	0,33
Método	**Microavg Precisão**	**Macroavg Precisão**	**Microavg Abrang.**	**Macroavg Abrang.**	**Microavg F-meas.**	**Macroavg F-meas.**	**MED**
2a limiar 0,005	0,51	0,42	0,79	0,65	0,62	0,51	0,56
2a limiar 0,008	0,61	0,48	0,67	0,51	0,64	0,49	0,56
2a maior peso	0,73	0,62	0,70	0,53	0,71	0,57	0,64
2b limiar ZERO	0,57	0,54	0,80	0,70	0,67	0,61	0,64
2b limiar 0,008	0,64	0,62	0,60	0,48	0,62	0,54	0,58
2b limiar 0,006	0,60	0,56	0,72	0,61	0,65	0,58	0,61
2b limiar 0,004	0,57	0,56	0,76	0,66	0,65	0,61	0,63
2b maior peso	0,72	0,73	0,65	0,53	0,68	0,61	0,64
2c limiar ZERO	0,65	0,59	0,77	0,68	0,70	0,63	0,66
2c limiar 0,005	0,66	0,61	0,69	0,61	0,67	0,61	0,64
2c limiar 0,007	0,71	0,66	0,59	0,50	0,64	0,57	0,60
2c maior peso	0,71	0,74	0,64	0,54	0,67	0,62	0,64
3a limiar 0,002	0,30	0,29	0,70	0,71	0,42	0,41	0,41
3a limiar 0,0025	0,40	0,39	0,36	0,38	0,38	0,38	0,38
3a limiar 0,003	0,43	0,42	0,10	0,13	0,16	0,20	0,18
3a maior peso	0,60	0,73	0,60	0,67	0,60	0,70	0,65
3b limiar 0,00001	0,84	0,87	0,92	0,84	0,88	0,85	0,86
3b limiar 0,00002	0,95	0,94	0,82	0,72	0,88	0,82	0,85
3b maior peso	0,93	0,95	0,93	0,86	0,93	0,90	**0,91**

6.1.1 Observações sobre os resultados

- Comparação entre limiares usados:

- comparando a classificação com limiar com a classificação pelo maior peso, pode-se notar na coleção de treino que a primeira conseguiu 4 melhores resultados e a segunda somente 3 (1 empate); conclui-se que é possível melhorar o desempenho usando um

limiar ao invés do maior peso; entretanto, não foi possível identificar um limiar ótimo para qualquer método, levando a crer que a escolha do limiar dependerá do método empregado; além disto, a escolha do limiar ótimo parece ser um problema complexo; pode-se fazer uma análise por amostras antes do processo final de classificação, mas nenhum estudo foi feito para garantir que isto resultará eficaz;

- a hipótese de que conceitos definidos com mais termos obtém maior abrangência ou maior precisão não se confirmou nos experimentos, como se pode notar comparando as medidas entre os métodos (1a) e (1b), pelo maior peso (tabelas 6.6, 6.7, 6.8 e 6.9), lembrando que o primeiro método tinha mais termos nas definições;
- pôde-se notar que, pelo mesmo limiar (tabelas 6.10, 6.11, 6.12 e 6.13), o método (1a), com mais termos, gera mais casos recuperados que o método (1b); esta pode ser a razão de o método (1a) ter obtido melhor resultado em abrangência, pelo mesmo limiar;
- ainda comparando (1a) com (1b), notou-se que o primeiro foi melhor na medida *Microaveraging F-measure*, perdendo na media *Macroaveraging F-measure*, em ambas as coleções (tabelas 6.2 e 6.4); não foi possível encontrar uma explicação para este fato;
- pelo maior peso, não se pode afirmar que um número maior de termos gera um número maior de casos associados ou recuperados, como se pode conferir nos resultados entre (1a) e (1b) (tabelas 6.6, 6.7, 6.8 e 6.9);
- pode-se notar que há uma forte tendência para se obter maior número de casos associados quando o limiar é mais baixo;

TABELA 6.3 - Textos associados a nenhuma categoria (coleção de treino)

Método	Número de Textos sem Categoria
1a	0
1b	1
1c	7
1d	2
1e	1
1f	2
1g	0
2a	10
2b	20
2c	20
3a	0
3b	1

TABELA 6.4 - Resultados com a coleção de teste

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
1a limiar 0,0012	0,50	0,29	0,42	0,35	0,46	0,32	0,39
1a maior peso	0,45	0,42	0,44	0,35	0,44	0,38	0,41
1b limiar 0,001	0,46	0,29	0,34	0,32	0,39	0,30	0,34
1b maior peso	0,43	0,44	0,42	0,36	0,42	0,40	0,41
1c limiar 0,025	0,37	0,30	0,44	0,38	0,40	0,34	0,37
1c maior peso	0,47	0,44	0,44	0,35	0,45	0,39	0,42
1d limiar 0,00015	0,37	0,39	0,50	0,47	0,43	0,43	0,43
1d maior peso	0,39	0,43	0,38	0,39	0,38	0,41	0,39
1e limiar 0,03	0,28	0,31	0,56	0,55	0,37	0,40	0,38
1e maior peso	0,32	0,35	0,22	0,19	0,26	0,25	0,25
1f limiar 0,02	0,32	0,27	0,61	0,54	0,42	0,36	0,39
1f maior peso	0,36	0,28	0,35	0,26	0,35	0,27	0,31
1g limiar 0,02	0,31	0,27	0,62	0,54	0,41	0,36	0,38
1g maior peso	0,37	0,28	0,36	0,26	0,36	0,27	0,31
2a limiar 0,005	0,49	0,43	0,76	0,66	0,60	0,52	0,56
2a limiar 0,008	0,62	0,52	0,66	0,52	0,64	0,52	0,58
2a maior peso	0,68	0,59	0,63	0,49	0,65	0,54	0,59
2b limiar ZERO	0,54	0,56	0,76	0,67	0,63	0,61	0,62
2b maior peso	0,69	0,83	0,62	0,46	0,65	0,59	0,62
2c limiar ZERO	0,60	0,59	0,72	0,65	0,65	0,62	**0,63**
2c maior peso	0,69	0,73	0,61	0,47	0,65	0,57	0,61
3a limiar 0,002	0,29	0,27	0,60	0,57	0,39	0,37	0,38
3a maior peso	0,41	0,54	0,41	0,44	0,41	0,48	0,44

3b limiar 0,00001	0,66	0,57	0,73	0,54	0,69	0,55	0,62
3b maior peso	0,73	0,60	0,72	0,49	0,72	0,54	**0,63**

TABELA 6.5 - Textos associados a nenhuma categoria (coleção de teste)

Método	Número de Textos sem Categoria
1a	6
1b	9
1c	15
1d	9
1e	9
1f	9
1g	6
2a	16
2b	22
2c	23
3a	2
3b	5

- Comparação entre os métodos tipo 1:

- os métodos (a) e (b) tiveram desempenho igual pelo maior peso, nas duas coleções; entretanto, pelos limiares, o método (a) foi melhor nas duas coleções; conclui-se que é possível usar um conjunto menor de termos (os mais freqüentes) para descrever os conceitos, sem perder desempenho, quando a classificação for pelo maior peso; quando algum limiar for utilizado, o melhor é manter um número maior de termos (nenhum estudo foi feito quanto ao número ideal);
- analisando o número de casos sem categoria pelos métodos (a) e (b) em ambas as coleções (0 x 1, 6 x 9, respectivamente), conclui-se que pode não valer a pena fazer algum tipo de seleção nos termos, sendo o método (a) eficiente e exigindo menor esforço para definição dos conceitos;
- o método (c) teve um desempenho pouco pior que os métodos (a) e (b) na coleção de treino (0,42 x 0,45), mas obteve um desempenho ligeiramente melhor na coleção de teste (0,42 x 0,41); conclui-se que a intervenção humana selecionando termos a partir de um *thesaurus* pode melhorar um pouco o desempenho final, mas o esforço talvez não seja compensador;
- o método (d) obteve desempenhos piores que os métodos (a) e (b) em ambas as coleções, levando a crer que as diferenças identificadas a partir de um *thesaurus* (termos exclusivos de classes) não são eficazes;
- os métodos (e), (f) e (g) obtiveram desempenhos piores que os métodos (a) e (b) em ambas as coleções, permitindo concluir que não é eficaz alterar os pesos dos termos identificados em um *thesaurus*;
- os métodos (a) e (g) foram melhores na avaliação dos casos não classificados (sem categoria); conclui-se que, no caso deste *thesaurus*, não é vantagem nenhum tipo de

intervenção para definir os termos de cada conceito, sendo razoável utilizar todos os termos que aparecem nas descrições, com a freqüência relativa como peso associado.

- Avaliação dos métodos tipo 2 (somente classificação pelo maior peso):
- os métodos do tipo 2 obtiveram resultados bem melhores que o melhor método do tipo 1 (1a), em ambas as coleções; conclui-se que é importante utilizar sinônimos para melhorar o desempenho;
- entretanto, os métodos do tipo 2 geraram um número maior de casos sem categoria, em ambas as coleções; a escolha do método pode depender do objetivo (obter melhor desempenho ou menor indecisão);
- analisando os resultados entre os métodos (a) e (b), nota-se que o segundo obtém melhor desempenho na coleção de teste (com empate na coleção de treino); pode-se concluir que há melhora de desempenho quando pessoas intervém no processo para corrigir erros e selecionar termos sinônimos; entretanto, há uma piora nos casos sem categoria;
- o método (2c) não resultou em melhor desempenho e ainda manteve o alto índice de casos sem categoria, levando à conclusão de que seu uso não é recomendado.
- Avaliação dos métodos tipo 3 (pelo maior peso):
- na coleção de treino, o método (3a) foi um pouco melhor que os métodos (2b) e (2c) e bem melhor que o método (1a); entretanto, na coleção de teste, o método (3a) foi bem pior que os citados; em ambas as coleções, o método (3a) gerou bem menos casos sem categoria; isto leva a crer que o método (3a) seria vantajoso em coleções que não mudam muito (coleção de treino é a mesma que a de teste ou uso);
- já o método (3b) teve desempenho um pouco melhor que os métodos (2b) e (2c) na coleção de teste mas obteve um desempenho bem superior na coleção de treino, sendo o melhor resultado em todos os métodos estudados;
- em relação ao método (3a), o método (3b) também foi bem melhor, mas com um pouco mais de casos sem categoria, sendo portanto mais recomendado;
- isto leva a crer que os métodos do tipo 3 são mais indicados quando há casos para treino, pois atingem os melhores resultados minimizando o esforço para definição dos conceitos (não exigem um *thesaurus* nem a intervenção humana ou uso de sinônimos).
- Comparação dos resultados nas duas coleções:
- pode-se observar que a maioria dos métodos estudados obtêm melhores resultados na coleção de treino do que na de teste;
- somente o método (1d) foi melhor na coleção de teste que na de treino, levando à hipótese de que as diferenças nas descrições usadas nos *thesauri* podem melhorar o processo de classificação;
- o método (1c) obteve desempenhos iguais nas duas coleções;
- na coleção de teste, houve sempre maior indecisão (mais casos sem categoria), para todos os métodos avaliados.

Pode-se concluir que todos os mecanismos de apoio avaliados são importantes para melhorar a definição de conceitos. Assim, recomenda-se utilizar termos extraídos de um *thesaurus* da área e termos sinônimos extraídos de dicionários ou por análise de amostras. A intervenção humana é importante para selecionar termos do *thesaurus*, dos dicionários e das amostras e para eliminar termos genéricos ou que levam a erros. Apesar de ter proporcionado os melhores resultados, a aprendizagem supervisionada só é possível quando existem casos de treino de boa qualidade.

Tais conclusões confirmam as decisões tomadas em [LIM97], onde o CID foi usado para identificar diagnósticos, em [BAT86], onde um dicionário específico do domínio permitiu expandir o vocabulário, e em [YAN94], onde empregou-se um dicionário técnico aumentado com termos sinônimos usados pelas pessoas do local da aplicação. As conclusões também são confirmadas pelos achados de Knight [KNI99], que afirma que pequenas amostras podem trazer bons resultados, e de Chen e outros [CHE97], que afirmam que ferramentas automatizadas minimizam o esforço de aquisição de conhecimento, auxiliando na criação de vocabulários controlados.

TABELA 6.6 - Resultado 1a por maior peso (coleção de teste)

Conceitos 1a			por maior peso				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	31	7	38	0,18	7	17	0,41
substânc.	1	2	3	0,67	2	53	0,04
esquizofr.	53	72	125	0,58	72	105	0,69
afetivos	22	7	29	0,24	7	25	0,28

TABELA 6.7 - Resultado 1b por maior peso (coleção de teste)

Conceitos 1b			por maior peso				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	43	8	51	0,16	8	17	0,47
substânc.	2	6	8	0,75	6	53	0,11
esquizofr.	35	62	97	0,64	62	105	0,59
afetivos	29	7	36	0,19	7	25	0,28

TABELA 6.8 - Resultado 1a por maior peso (coleção de treino)

Conceitos 1a			por maior peso				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	25	7	32	0,22	7	18	0,39
substânc.	0	3	3	1,00	3	52	0,06
esquizofr.	54	82	136	0,60	82	103	0,80
afetivos	24	5	29	0,17	5	27	0,19

TABELA 6.9 - Resultado 1b por maior peso (coleção de treino)

Conceitos 1b			por maior peso				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	33	7	40	0,18	7	18	0,39
substânc.	1	7	8	0,88	7	52	0,13
esquizofr.	38	71	109	0,65	71	103	0,69
afetivos	33	9	42	0,21	9	27	0,33

TABELA 6.10 - Resultado 1a por limiar 0,0012 (coleção de treino)

Conceitos 1a			limiar 0,0012				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	35	8	43	0,19	8	18	0,44
substânc.	0	1	1	1,00	1	52	0,02
esquizofr.	42	74	116	0,64	74	103	0,72
afetivos	33	9	42	0,21	9	27	0,33

TABELA 6.11 - Resultado 1b por limiar 0,0012 (coleção de treino)

Conceitos 1b			limiar 0,0012				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	22	6	28	0,21	6	18	0,33
substânc.	0	1	1	1,00	1	52	0,02
esquizofr.	14	48	62	0,77	48	103	0,47
afetivos	20	2	22	0,09	2	27	0,07

TABELA 6.12 - Resultado 1a por limiar 0,0014 (coleção de treino)

Conceitos 1a			limiar 0,0014				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	23	8	31	0,26	8	18	0,44
substânc.	0	1	1	1,00	1	52	0,02
esquizofr.	22	59	81	0,73	59	103	0,57
afetivos	23	2	25	0,08	2	27	0,07

TABELA 6.13 - Resultado 1b por limiar 0,0014 (coleção de treino)

Conceitos 1b			limiar 0,0014				
	errados	certos	recup.	Precisão	certos	ideal	Abrang.
orgânicos	12	6	18	0,33	6	18	0,33
substânc.	0	1	1	1,00	1	52	0,02
esquizofr.	9	38	47	0,81	38	103	0,37
afetivos	14	1	15	0,07	1	27	0,04

6.2 Avaliação dos Métodos de Categorização

Os métodos de categorização baseados no espaço de vetores e no contextual foram comparados para conceitos gerais (classes do CID) e para conceitos específicos (características dos pacientes), em ambas as coleções.

6.2.1 Conceitos gerais (classes do CID)

Foram utilizados os seguintes métodos contextuais:

4a) método contextual, utilizando as mesmas definições do método (2c), mas permitindo pares de termos; os termos com pesos negativos em (2c) não foram usados;

4b) método contextual, como o anterior mas ainda permitindo termos negativos e corrigindo erros (eliminando termos muito genéricos e acrescentando outros termos mais específicos).

Como representante dos métodos baseados no espaço de vetores, foi utilizado o método (2c), que usa os mesmos recursos de apoio.

As figuras 6.1, 6.3 e 6.3 apresentam as definições do conceito geral **substâncias** segundo os métodos avaliados (termos e pesos separados por '|'). No método (2c) aparecem temos com pesos negativos. Vale lembrar que os métodos contextuais (4a) e (4b) são definidos com regras (uma por linha), nas quais não são usados pesos, mas podem ser utilizados radicais de termos e termos negativos (o símbolo '–' serve para indicar um termo negativo).

alcool | 1,0000	alcoólica | 1,0000	toxicomano | 1,0000
álcool | 1,0000	embriaguez | 1,0000	toxicômano | 1,0000
bebe | 1,0000	cocaina | 1,0000	depressão | -1,0000
bebia | 1,0000	cocaína | 1,0000	depressivo | -1,0000
beber | 1,0000	maconha | 1,0000	depressiva | -1,0000
bebida | 1,0000	droga | 1,0000	deprimido | -1,0000
bebidas | 1,0000	drogas | 1,0000	deprimida | -1,0000
bebendo | 1,0000	alcoolismo | 1,0000	antidepressivo | -1,0000
alcoolista | 1,0000	tóxico | 1,0000	antidepressivos | -1,0000
alcoolico | 1,0000	tóxicos | 1,0000	depressao | -1,0000
alcoólico | 1,0000	toxicomania | 1,0000	

FIGURA 6.1 - Conceito "substâncias" segundo o método 2c

alcool	bebidas	embriaguez	alcoolismo
álcool	bebendo	cocaina	tóxico
bebe	alcoolista	cocaína	tóxicos
bebia	alcoolico	maconha	toxicomania
beber	alcoólico	droga	toxicomano
bebida	alcoólica	drogas	toxicômano

FIGURA 6.2 - Conceito "substâncias" segundo o método 4a

alco -nega -não	embriaguez	tóxico
álco -nega -não	cocaina	toxicoman
alcó -nega -não	cocaína	toxicôman
bebe -nega -não	maconha	etílic
bebi -nega -não	droga	etilic

FIGURA 6.3 - Conceito "substâncias" segundo o método 4b

TABELA 6.14 - Tempo aproximado de categorização na coleção de treino e número de termos por método

Método	Tempo (em min.)	Orgânicos	Substâncias	Esquizofr.	Afetivos
2c	7	21	24	36	12
4a	6	21	24	25	12
4b	3	10	15	12	7

TABELA 6.15 - Método espaço de vetores X contextual (coleção de treino)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
2c limiar ZERO	0,65	0,59	0,77	0,68	0,70	0,63	**0,66**
2c maior peso	0,71	0,74	0,64	0,54	0,67	0,62	0,64
4a limiar ZERO	0,48	0,46	0,55	0,59	0,51	0,52	0,51
4a maior peso	0,59	0,63	0,46	0,49	0,52	0,55	0,53
4b limiar ZERO	0,66	0,66	0,56	0,58	0,61	0,62	0,61
4b maior peso	0,72	0,78	0,49	0,50	0,58	0,61	0,59

TABELA 6.16 - Textos associados a nenhuma categoria (coleção de treino)

Método	Número de Textos sem Categoria
2c	20
4a	45
4b	66

TABELA 6.17 - Método espaço de vetores X contextual (coleção de teste)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
2c limiar ZERO	0,60	0,59	0,72	0,65	0,65	0,62	**0,63**
2c maior peso	0,69	0,73	0,61	0,47	0,65	0,57	0,61
4a limiar ZERO	0,48	0,50	0,55	0,59	0,51	0,54	0,52
4a maior peso	0,56	0,61	0,43	0,42	0,49	0,50	0,49
4b limiar ZERO	0,52	0,58	0,53	0,58	0,52	0,58	0,55

4b maior peso	0,62	0,72	0,44	0,45	0,51	0,55	0,53

TABELA 6.18 - Textos associados a nenhuma categoria (coleção de teste)

Método	Número de Textos sem Categoria
2c	23
4a	49
4b	61

Analisando-se os resultados, pode-se notar que:

- a classificação pelo limiar zero foi melhor que pelo maior peso (5 a 1), só perdendo no método (4a) pela coleção de treino;
- nos métodos do tipo 4, a classificação pelo limiar foi melhor que pelo maior peso (3 a 1);
- o método contextual foi mais rápido (teve menor tempo no processo de classificação);
- o método (4b) é melhor que (4a), em ambas as coleções, tanto por limiar quanto por maior peso; entretanto (4b) apresentou mais casos sem categoria (maior indecisão); pode-se concluir que é vantajoso corrigir erros e usar termos negativos;
- o método (2c) é melhor que os métodos do tipo 4, em ambas as coleções, por limiar e por maior peso, além de resultar em menos casos sem categoria (menor indecisão);
- o método (4b) é melhor que o método (2c) em precisão, pelo maior peso (na coleção de treino mas perde na de teste); entretanto (4b) perde em abrangência em ambas as coleções.

6.2.2 Conceitos específicos (características do paciente)

Para comparar os métodos espaço de vetores e contextual sobre conceitos específicos, foi necessário utilizar uma amostra da coleção de treino, composta de 50 textos. Neste caso, a distribuição dos diagnósticos não foi investigada por não interferir na avaliação. Especialistas do domínio analisaram os 50 textos e indicaram os conceitos específicos, relativos às características dos pacientes, que apareciam em cada texto.

Não foram avaliados todos os conceitos, mas somente os 12 mais complexos, cujas definições poderiam levar a confusões. Os demais conceitos foram considerados simples por terem como descritores poucos termos simples.

O método espaço de vetores utilizou o limiar zero para decidir a presença do conceito no texto. Neste caso, foram usados os mesmos termos presentes no método contextual, todos com peso 1. A categorização com o método espaço de vetores levou 3 minutos, enquanto que com o método contextual somente 1 minuto.

As figuras 6.4 e 6.5 apresentam as definições do conceito "*alcoolismo*" em ambos os métodos.

alcool | 1,0000	etílico | 1,0000	cana | 1,0000

alcoolismo | 1,0000	etilico | 1,0000	destilado | 1,0000
bebe | 1,0000	cachaça | 1,0000	destilados | 1,0000
álcool | 1,0000	alcoólico | 1,0000	garrafa | 1,0000
bebi | 1,0000	cerveja | 1,0000	garrafas | 1,0000
bebia | 1,0000	embriagado | 1,0000	bares | 1,0000
beber | 1,0000	embriagada | 1,0000	etilista | 1,0000

FIGURA 6.4 - Conceito "alcoolismo" segundo o método espaço de vetores

alcool –nega -não	etilic	cana
bebe –nega -não	cachaça	destilad -nega -não
álcool –nega -não	alcoólic -nega -não	garrafa
bebi –nega -não	cerveja	bares
etílic	embriagad	etilista

FIGURA 6.5 - Conceito "alcoolismo" segundo o método contextual

TABELA 6.19 - Comparação de métodos sobre conceitos específicos (espaço de vetores X contextual)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Espaço de vetores	0,80	0,78	0,83	0,83	0,81	0,80	0,81
Contextual	0,90	0,89	0,93	0,92	0,91	0,90	**0,91**

TABELA 6.20 - Resultados do método espaço de vetores, limiar zero, para cada conceito

Conceito	errados	certos	recup.	Precisão	certos	ideal	Abrang.
agressividade	4	30	34	0,88	30	35	0,86
alcoolismo	3	15	18	0,83	15	20	0,75
bichos	0	4	4	1,00	4	4	1,00
depressão	3	7	10	0,70	7	8	0,88
homicida	11	4	15	0,27	4	5	0,80
inapetência	3	22	25	0,88	22	28	0,79
insônia	0	32	32	1,00	32	37	0,86
medo	3	6	9	0,67	6	6	1,00
morte	2	6	8	0,75	6	11	0,55
nervosismo	6	24	30	0,80	24	30	0,80
reinternação	9	29	38	0,76	29	34	0,85
suicida	4	15	19	0,79	15	17	0,88

TABELA 6.21 - Resultados do método contextual para cada conceito

Conceito	errados	certos	recup.	Precisão	certos	ideal	Abrang.
agressividade	5	35	40	0,88	35	35	1,00
alcoolismo	3	16	19	0,84	16	20	0,80
bichos	0	4	4	1,00	4	4	1,00

depressão	3	8	11	0,73	8	8	1,00
homicida	0	4	4	1,00	4	5	0,80
inapetência	4	25	29	0,86	25	28	0,89
insônia	0	36	36	1,00	36	37	0,97
medo	2	6	8	0,75	6	6	1,00
morte	2	8	10	0,80	8	11	0,73
nervosismo	2	29	31	0,94	29	30	0,97
reinternação	2	31	33	0,94	31	34	0,91
suicida	0	17	17	1,00	17	17	1,00

Analisando-se os resultados, pode-se notar que:

- o método contextual obteve melhores resultados em todos os conceitos, pela medida de média final;
- somente na medida de precisão do conceito "*inapetência*", o método contextual foi pior que o método espaço de vetores (baixou de 0,88 para 0,86); isto porque recuperou mais textos (acertou mais e errou mais também);
- não houve indecisão, ou seja, todos os textos foram associados a alguma categoria em ambos os métodos (não houve nenhum texto sem categoria).

Avaliando em detalhe as definições dos conceitos, pôde-se observar que alguns conceitos só podem ser reconhecidos pele método contextual, devido às limitações do método espaço de vetores. Por exemplo, o conceito "*alcoolismo*" não deve ser reconhecido quando o termo "*nega*" aparecer na frase. A falta de um mecanismo de análise do contexto mais eficiente no método espaço de vetores levou a erros no resultado do processo (conceitos erroneamente reconhecidos). Outro problema ocorreu com o conceito "*reinternação*", o qual poderia ser reconhecido por expressões como "*paciente conhecido*", "*já internou*", "*várias internações*". Só que neste caso, o conceito deixou de ser reconhecido em alguns textos porque o método espaço de vetores não pode analisar a presença dos termos numa mesma frase e o reconhecimento dos termos no mesmo texto (independente da frase) pode levar a erros.

Problema semelhante houve no caso do termo "*matar*". Se este aparecesse com o pronome "*se*" indiciaria o conceito "*suicida*", caso contrário, o conceito "*homicida*". Somente o método contextual pode minimizar este problema de vocabulário.

6.2.3 Conclusão das avaliações

Pela análise dos resultados, pode-se verificar que o método baseado no modelo espaço de vetores é melhor no caso de conceitos mais gerais ou genéricos (nestes experimentos, referentes às classes de diagnósticos ou doenças). Enquanto que, no caso de conceitos mais específicos (como as características dos pacientes), o método contextual obteve melhor desempenho.

Conclui-se que a escolha do método (espaço de vetores X contextual) depende da granularidade dos conceitos (tipo de conceito usado). Assim, se os conceitos são mais específicos, deve-se utilizar o método contextual. Caso os conceitos sejam mais gerais ou genéricos, o método de identificação baseado no modelo espaço de vetores é mais eficaz.

A razão é que conceitos muito específicos são melhor identificados com o método contextual, o qual por sua vez não funciona bem quando os conceitos são muito gerais e devem ser reconhecidos através de uma análise mais abrangente sobre o conteúdo do texto. Neste último caso, quando é preciso avaliar um número maior de informações presentes no texto, o método baseado no espaço de vetores tende a dar melhores resultados pois o contexto a ser analisado são todos os termos presentes no texto.

6.3 Avaliação do Processo Padrão de Identificação dos Conceitos

Utilizando o processo padrão explicado na seção 5.4, foram avaliados os resultados da etapa de identificação dos conceitos (tabela 6.22).

TABELA 6.22 - Medidas de avaliação do processo padrão de identificação de conceitos específicos

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Contextual	0,90	0,89	0,93	0,92	0,91	0,90	0,91

Como explicado na seção anterior, especialistas do domínio analisaram 50 textos extraídos da coleção de teste, tomados como amostra representativa desta coleção. Os especialistas indicaram os conceitos específicos, relativos às características dos pacientes, que apareciam em cada texto. Não foram avaliados todos os conceitos, mas somente os 12 mais complexos, cujas definições poderiam levar a confusões. Os demais conceitos foram considerados simples por terem como descritores poucos termos simples.

Como pode ser visto na tabela 6.22, uma margem de erro menor que 10% pode ser considerada satisfatória por não influenciar muito o processo final de descoberta.

Vale lembrar que os resultados podem ser melhores se a definição dos conceitos for melhor feita. Na primeira rodada de identificação, os valores de *microaveraging* precisão, *macroaveraging* precisão, *microaveraging* abrangência e *macroaveraging* abrangência foram, respectivamente, 75%, 87%, 71% e 87%. Isto prova que é possível melhorar o processo refinando as definições dos conceitos.

Entretanto, uma atenção especial deve ser dada aos conceitos que tiveram baixos valores em alguma medida (ver tabela 6.21). Por exemplo, o conceito "*depressão*" obteve precisão de 73%, e o conceito "*morte*" obteve abrangência de 73%. Estes valores podem colocar em dúvida o conhecimento descoberto referente a estes conceitos.

6.4 Avaliação Subjetiva do Conhecimento Descoberto

O conhecimento descoberto, resultante da realização do processo padrão explicado na seção 5.4, foi avaliado subjetivamente por 2 médicos psiquiatras com bastante experiência em atividades de diagnóstico (ambos diretores de uma clínica psiquiátrica e um professor da disciplina de psiquiatria em curso de Medicina).

As distribuições de conceitos e as regras associativas referentes a cada classe (diagnóstico do CID) foram apresentadas e discutidas.

A resposta destes especialistas do domínio foi de que o conhecimento descoberto é muito similar ao utilizado pelos médicos para realizar o diagnóstico ou para treinar estudantes de medicina ou médicos residentes.

6.5 Avaliação Objetiva do Conhecimento Descoberto

Foi feita uma avaliação objetiva do conhecimento descoberto. Esta avaliação consistiu em utilizar o conhecimento descoberto através do processo padrão apresentado em 5.4 embutido em um sistema automatizado de suporte à decisão.

O objetivo do experimento era tentar descobrir o diagnóstico de cada caso correspondente aos 200 textos da coleção de teste de forma automática, usando o sistema. O processo padrão de descoberta de conhecimento foi aplicado sobre a coleção de treino, para identificar características das 4 classes do CID. Depois, o conhecimento resultante foi embutido num sistema implementado com o método baseado no espaço de vetores para categorização de textos (lembrando que este método segue o algoritmo de Rocchio para classificação). Neste processo, o sistema automatizado, além de identificar a presença de características (conceitos) nos textos (pelo método contextual), deveria comparar estas características com as que descrevem cada classe pré-definida.

Trabalhos semelhantes foram feitos por Subasic and Huettner [SUB00], que identificavam atributos qualitativos em textos referentes a filmes (tais como "*horror*", "*justiça*" e "*dor*"), com o objetivo de classificar os filmes em gêneros (por exemplo, em filmes de ação, "*horror*" é mais freqüente que "*humor*"). Wilcox e outros [WIL00] avaliaram métodos de categorização de textos médicos que analisavam características em alto nível. As características, identificadas nos textos, eram observações médicas representadas por códigos padrões. A estratégia era utilizar um método de aprendizado indutivo para extrair automaticamente regras de decisão em coleções de treino.

Assim, considerando que cada classe de diagnóstico pode ser representada por uma subcoleção de textos (aqueles referentes a pacientes do diagnóstico), as distribuições de conceitos e a regras associativas, resultantes do processo padrão de descoberta de conhecimento, foram utilizadas como características das 4 classes do CID.

O método baseado no espaço de vetores foi escolhido por ser o mais simples e por apresentar um desempenho razoável. Deste modo, as decisões resultantes do sistema automatizado podiam ser associadas mais às características de cada classe do que ao desempenho do método de classificação, conforme sugestão de Jensen e Martinez [JEN00].

No sistema, cada classe era representada por um vetor de características, cada uma destas com um peso associado. As características podiam ser conceitos ou pares de conceitos. Conceitos simples eram os presentes na classe, e os pares de conceitos foram extraídos das regras associativas, ambos resultantes do processo padrão de descoberta de conhecimento. Os pesos associados aos conceitos simples foram extraídos das distribuições descobertas para cada classe e serviram para definir a força da característica para indicar a presença da classe.

Pesos negativos foram usados para representar conceitos que nunca aparecem na classe (identificados na descoberta com o processo padrão), numa tentativa de eliminar falsos candidatos. Isto seguiu a sugestão de Galavotti e outros [GAL00], que obtiveram sucesso usando pesos iguais a –1, ao invés de zero, para evidências tidas como negativas em casos de treino.

Os diferentes métodos usados nesta avaliação são descritos a seguir:

- **Ca**: utiliza todos os conceitos e as respectivas distribuições originais, para cada classe, conforme descoberto no processo padrão de mineração;
- **Ca+n**: utiliza todos os conceitos (mesmos de **Ca**) e mais conceitos negativos com peso -1;
- **Clf+n**: usa os conceitos menos freqüentes de cada classe (distribuição menor que 50%), extraídos de **Ca**, e mais os conceitos negativos de **Ca+n**;
- **Cp**: utiliza pares de conceitos como descritores das classes, extraídos das regras associativas descobertas no processo de mineração com confiança maior ou igual a 80%;
- **Cp2**: utiliza pares de conceitos referentes às regras associativas extraídas de um processo de mineração utilizando como limiar de confiança 50%;
- **Cpd**: utiliza pares exclusivos de conceitos, ou seja, aqueles extraídos das regras associativas exclusivas da classe (confiança maior ou igual a 80%);
- **Ca+p**: utiliza todos os conceitos (como em **Ca**) mais os pares de conceitos de **Cp**.

O método **Cd**, tendo como descritores da classe os conceitos presentes somente na classe (diferenças), não foi usado porque o processo padrão de descoberta não encontrou conceitos exclusivos nas classes.

Estes métodos foram escolhidos para demonstrar os diferentes modos como conceitos podem ser usados em processos de classificação. A decisão do sistema segue a classificação por maior peso, isto é, quando havia mais de uma classe associada ao texto sendo avaliado, a classe com maior grau de relacionamento era tomada como resultado para a decisão final do sistema automático.

TABELA 6.23 - Resultados dos métodos baseados em conceitos (coleção de teste)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Ca	0,44	0,50	0,44	0,39	0,44	0,44	0,44
Ca+n	0,51	0,55	0,51	0,42	0,51	0,48	0,50
Clf+n	0,65	0,73	0,61	0,53	0,63	0,61	**0,62**
Cp	0,57	0,51	0,54	0,45	0,55	0,48	0,51
Cp2	0,43	0,47	0,41	0,42	0,42	0,44	0,43
Cpd	0,64	0,54	0,60	0,51	0,62	0,52	0,57
Ca+p	0,56	0,51	0,56	0,47	0,56	0,49	0,52

Como se pode notar na tabela 6.23, o melhor resultado foi obtido pelo método **Clf+n** (com os conceitos menos freqüentes mais conceitos negativos), com uma média de 62% (uma margem de erro próxima de 38%). Apresentando estes resultados para os médicos especialistas que participaram do experimento, eles consideraram os resultados numéricos

ótimos, uma vez que desempenhos maiores que 60% são melhores que algumas decisões de especialistas humanos.

E é bom lembrar que o processo foi realizado quase todo de forma automática, desde a identificação de conceitos nos textos, a descoberta de conhecimento através da mineração sobre conceitos e a classificação dos textos pelo sistema automático. A intervenção humana só apareceu na definição dos conceitos e na separação de casos referentes a cada classe de diagnóstico. Além disto, o sistema automático só analisava informações presentes no prontuário de internação e, no caso dos especialistas humanos, estes podiam coletar mais informações sobre o paciente ao longo do tempo e mesmo assim podiam mudar o diagnóstico durante o tratamento do paciente (a classe associada aos textos e utilizada como referência correta era a final, no momento da alta do paciente).

Por outro lado, cabe salientar que os conceitos descobertos em cada classe, sem uma prévia organização, não levam a bons resultados. Este foi o caso do resultado da técnica de análise de distribuições (método **Ca**) e do resultado da técnica associativa (método **Cp**), com desempenhos de 44% e 51%, respectivamente.

Entretanto, quando há uma certa organização destes resultados, o desempenho pode melhorar. Foi o caso utilizando ambos os resultados juntos (método **Ca+p**) que melhorou um pouco o desempenho (52%). Também houve melhora utilizando conceitos negativos, já que o método **Ca+n** foi melhor que o método **Ca** (50% contra 44%). A seleção de conceitos, tomando somente os mais freqüentes, também melhorou o desempenho e em muito, como se pode notar comparando os métodos **Clf+n** e **Ca+n** (62% contra 50%).

O uso de características exclusivas também melhora o desempenho, como se pode notar no caso do método **Cpd**, que utiliza pares de conceitos extraídos das regras exclusivas de cada classe. Este método obteve melhores resultados que o método **Cp** (usando todos os pares, inclusive os comuns a outras regras). Acredita-se que um método que utilize somente conceitos exclusivos (tipo **Cd**) possa trazer melhores resultados também. Entretanto, como não foi possível descobrir conceitos exclusivos às classes (todos os conceitos apareciam em todas as classes), este método não pôde ser avaliado nos experimentos.

Já o uso de conceitos extraídos de regras associativas com menor grau de confiança não ajudou a melhorar o desempenho, como se pode notar no resultado do método **Cp2** (43%).

Por fim, comparando o resultado das duas técnicas de mineração usadas, pode-se notar que pares de conceitos alcançam melhores resultados que conceitos simples (51% do método **Cp** contra 44% do método **Ca**). Isto confirma a conclusão de Apté e outros [APT94] com relação a métodos baseados em palavras: pares de palavras trazem melhores resultados do que palavras únicas.

6.6 Comparação das Abordagens Baseada em Conceitos e Baseada em Palavras

Em [YAN97], são comparados vários métodos existentes para classificação de textos. Entretanto, todos eles usam como características representantes dos textos as palavras ou termos presentes nos textos. Yang e Pedersen [YAN97] falam da necessidade de serem avaliados métodos que utilizem características de mais alto nível. Acredita-se que a

abordagem baseada em conceitos, proposta nesta tese, possa permitir representações de textos em mais alto nível que palavras e, assim portanto, representações mais próximas da realidade.

Com este intuito, procurou-se comparar o desempenho de métodos baseados em conceitos contra métodos baseados em palavras em processos de classificação de textos. A hipótese é que conceitos representam melhor que termos o conteúdo dos textos e, portanto, levariam a um melhor desempenho nos resultados de classificação.

Para comparar os dois tipos de representações, foram feitos experimentos de classificação de textos usando como representações dos textos duas alternativas:

a) os conceitos identificados pelo processo padrão de descoberta, e
b) os termos presentes nos textos.

Para implementar a classificação, foram usados dois tipos diferentes de métodos: um baseado no Rocchio e outro baseado no k-NN (*Nearest Neighbour*). Cada um destes representa uma abordagem diferente de classificação. O método Rocchio utiliza um vetor protótipo para descrever uma classe. Assim, palavras ou conceitos podem ser usados como elementos deste vetor e, portanto, pode-se avaliar a força de ambos como representantes do conteúdo dos textos e como descritores de classes.

Por outro lado, o método k-NN avalia a classe de um texto candidato pelas classes associadas aos textos mais semelhantes a ele. Não há descritores para as classes, mas somente casos passados. Assim, a função de similaridade avalia a semelhança entre o conteúdo dos textos tomando palavras ou conceitos como características. Assim, pode-se avaliar a força destas características para descrever casos e permitir comparações entre casos por similaridade.

Os resultados permitirão dizer que tipo de característica (palavras ou conceitos) representa melhor o conteúdo dos textos, seja comparando o texto candidato a descritores de classes ou comparando-o a outros textos similares.

A escolha destes métodos de classificação também se deu por serem simples e, portanto, por não influenciarem o processo. Ou seja, o resultado final do processo de classificação pode ser associado mais à força das características para representar o conteúdo dos textos do que à capacidade do método de classificação.

Também procurou-se comparar os dois tipos de representações (conceitos ou palavras) quanto ao valor para explicar o raciocínio usado no processo de classificação. As características descritas em ambas as abordagens foram analisadas por especialistas humanos.

6.6.1 Classificação com método Rocchio

Como representantes da abordagem baseada em conceitos foram usados os métodos **Ca, Ca+n, Clf+n, Cp, Cp2, Cpd** e **Ca+p**, descritos na seção anterior.

Para representar a abordagem baseada em palavras, foram definidos dois métodos, a saber:

- **Wa**: que utiliza como descritores de classes todas as palavras presentes em mais de um texto na coleção de treino (para cada classe); os pesos associados às palavras são calculados pela média das freqüências relativas da palavra nos textos de treino da classe (a freqüência relativa é o número de vezes em que a palavra aparece no texto dividido pelo número total de termos no mesmo texto, conforme Salton e McGill [SAL83]);

- **Wd**: que utiliza somente palavras exclusivas das classes (diferenças), ou seja, os descritores de cada classe são as palavras que aparecem em mais de um texto, mas somente nos textos de treino da classe.

TABELA 6.24 - Resultados dos métodos baseados em conceitos X palavras usando Rocchio (coleção de teste)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Wa	0,41	0,54	0,41	0,44	0,41	0,48	0,45
Wd	0,73	0,60	0,72	0,49	0,72	0,54	**0,63**
Ca	0,44	0,50	0,44	0,39	0,44	0,44	0,44
Ca+n	0,51	0,55	0,51	0,42	0,51	0,48	0,50
Clf+n	0,65	0,73	0,61	0,53	0,63	0,61	**0,62**
Cp	0,57	0,51	0,54	0,45	0,55	0,48	0,51
Cp2	0,43	0,47	0,41	0,42	0,42	0,44	0,43
Cpd	0,64	0,54	0,60	0,51	0,62	0,52	0,57
Ca+p	0,56	0,51	0,56	0,47	0,56	0,49	0,52

TABELA 6.25 - Textos associados a nenhuma categoria usando Rocchio (coleção de teste)

Método	Número de textos sem categoria
Wa	2
Wd	5
Ca	2
Ca+n	2
Clf+n	15
Cp	14
Cp2	9
Cpd	14
Ca+p	1

Os métodos foram aplicados nas duas coleções (treino e teste), tendo como decisão a classe resultante com maior peso associado, isto é, quando havia mais de uma classe associada ao texto sendo avaliado, a com maior grau de relacionamento era tomada como resultado para a decisão final.

Como se pode notar na tabela 6.24, na coleção de teste, o método **Wd** (baseado em palavras) obteve o melhor resultado (média de 63%) mas com pouca diferença para o método **Clf+n** (média de 62%), que conseguiu o melhor desempenho entre os métodos baseados em conceitos. Entretanto, o método **Wd** resultou em menos casos de textos sem categoria (menor indecisão).

Comparando os métodos mais simples em ambas as abordagens (**Wa** e **Ca**), os quais utilizam como descritores da classe todas as características sem seleção ou organização (e não aos pares), pode-se notar que também não houve grande diferença (45% contra 44%).

Como a diferença foi muito pequena, pode-se dizer que não há vantagem efetiva de uma abordagem sobre outra.

Em ambas as abordagens, as características exclusivas (diferenças em cada classe) resultaram em melhores resultados (**Wd** e **Cpd**).

Com a intenção de analisar o desempenho dos métodos quando a coleção de textos não muda muito (por exemplo, para processos de recuperação), foram feitos experimentos também com a coleção de treino, mas usando somente os métodos que apresentaram melhor desempenho em ambas as abordagens.

TABELA 6.26 - Resultados dos métodos baseados em conceitos X palavras usando Rocchio (coleção de treino)

Método	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Wd	0.93	0.95	0.93	0.86	0.93	0.90	**0.92**
Clf+n	0.69	0.70	0.66	0.61	0.67	0.65	0.66

Como se pode notar na tabela 6.26, o método **Wd** alcançou um ótimo resultado, bem melhor que o método **Clf+n**. Disto pode-se concluir que o método baseado em palavras tende a ter melhor desempenho quando a coleção não muda muito. A razão pode ser a grande especificidade dos termos, ou seja, a capacidade de melhor distinguir classes, principalmente porque os conceitos apareciam em mais de uma classe na coleção avaliada.

6.6.2 Classificação com método k-NN

Existem diversas variações do método k-NN. Por exemplo, a classe resultante pode ser a que tiver maior valor na soma dos pesos associados ao texto candidato (sendo testado), entre os *k* mais semelhantes. Neste experimento, a classe resultante é aquela mais freqüente entre os *k* textos mais semelhantes.

Da mesma forma, é necessário estabelecer qual o valor ideal de *k*. Testes preliminares com 17 textos da coleção de treino concluíram que o *k* ideal, neste experimento, é igual a 1.

Também existem diferentes funções de similaridade que podem ser empregadas, sendo as mais comuns a distância Euclidiana [WIL88], que mede em linha reta num espaço imaginário a distância entre os vetores representativos dos textos, e o cosseno [CRO92], que mede o ângulo entre estes mesmos vetores. Quanto maior a distância ou o ângulo, menor a similaridade. Entretanto, este tipo de função precisa normalizar valores e pode trazer problemas no caso de elementos similares mas sem atributos em comum [WIL88].

$$gs(X,Y) = \frac{\sum_{h=1}^{k} gi_h(a,b)}{n}$$

onde:

gs é o grau de similaridade entre os documentos ***X*** e ***Y***;

gi é o grau de igualdade entre os pesos do termo ***h*** (peso ***a*** no documento ***X*** e peso ***b*** no documento ***Y***);

h é um índice para os termos comuns aos dois documentos;

k é o número total de termos comuns aos dois documentos;

n é o número total de termos nos dois documentos (sem contagem repetida).

FIGURA 6.6 – Fórmula da função de similaridade entre dois textos

Então, para este experimento foi utilizada outra função de similaridade, definida na figura 6.6. O grau de similaridade entre dois textos é dado pela soma dos graus de igualdade dos termos comuns (aos dois textos sendo comparados) dividido pelo número total de termos nos dois documentos (sem contagem repetida).

Para o cálculo do grau de igualdade entre pesos associados aos termos comuns, foi utilizada a fórmula da figura 6.7, definida por [PED93]. Esta fórmula é necessária, porque, apesar do termo aparecer em ambos os textos, ele pode ter graus de importância diferentes em cada texto. Ao invés de calcular a média dos pesos, a função determina o grau de igualdade entre os pesos, isto porque o uso da média pode acarretar distorções. Ou seja, se um termo aparece no primeiro texto com peso 0,9 e no segundo com peso 0,3, a média usada é 0,6. Se outro termo aparece no primeiro texto com peso 0,6 e no segundo com peso 0,6, a média usada também é 0,6. Entretanto, no segundo caso, os pesos são iguais (mais semelhantes entre si que no primeiro caso), o que indica maior semelhança entre os textos. Para textos idênticos, isto é, com os mesmos termos e na mesma freqüência (não precisa ser na mesma ordem), a função de similaridade dá como resultado o valor 1.

$$gi(a,b) = \frac{1}{2}\left[(a \to b) \wedge (b \to a) + (\bar{a} \to \bar{b}) \wedge (\bar{b} \to \bar{a})\right]$$

onde,

$\bar{x} = 1 - x$

$a \to b = \max\{c \in [0,1] \mid atc \leq b\}, t = produto$

$\wedge = \min$

FIGURA 6.7 - Cálculo do grau de igualdade entre pesos de termos comuns

Para este experimento, foram avaliados dois métodos baseados em conceitos. Um utiliza os conceitos identificados no texto e seu grau de presença neste texto (o grau de associação ou referência entre texto e conceito). O outro método não utiliza estes pesos, assumindo que os conceitos possuem igual importância no texto ou porque não importa o quanto cada conceito é referenciado no texto.

TABELA 6.27 - Resultados dos métodos baseados em conceitos X palavras usando k-NN (coleção de teste)

	Microavg Precisão	Macroavg Precisão	Microavg Abrang.	Macroavg Abrang.	Microavg F-meas.	Macroavg F-meas.	MED
Conceitos sem peso	0,71	0,74	0,65	0,43	0,68	0,54	**0,61**
Conceitos com peso	0,66	0,56	0,64	0,47	0,65	0,51	0,58
Palavras com peso	0,65	0,55	0,64	0,54	0,64	0,54	0,59

TABELA 6.28 - Textos associados a nenhuma categoria usando k-NN (coleção de teste)

Método	Número de textos sem categoria
Conceito sem peso	18
Conceito com peso	7
Palavra com peso	2

Pode-se notar, na tabela 6.27, que o método baseado nos conceitos, sem considerar pesos, obteve o melhor resultado, pela média final. Entretanto, este método resultou em maior

número de casos sem categoria. Isto porque o método baseado em palavras tende a encontrar algum termo comum, mesmo que pouco significativo.

O método baseado em conceitos (em ambos os tipos) perde em abrangência (na medida *macroaveraging*) para o método baseado em palavras. Isto também pode estar associado ao fato de haver mais termos que conceitos e, portanto, o método de classificação baseado em palavras tender a recuperar mais casos para cada classe.

Ao se comparar o tempo despendido nos experimentos (tabela 6.29), nota-se que a classificação com os métodos baseados em conceitos é bem mais rápida, em razão de haver menor número de características (menos conceitos que palavras para representar textos).

Entretanto, vale lembrar que os métodos baseados em conceitos acrescentam o tempo de aproximadamente 1 hora para identificação dos conceitos nos textos (nos 200 textos da coleção de teste) e necessitam das definições dos conceitos (30 horas, para o domínio e conceitos deste experimento).

TABELA 6.29 - Tempo de classificação usando k-NN (coleção de teste)

Método	Tempo de Classificação
Conceito sem peso	2 horas e 24 minutos
Conceito com peso	3 horas e 21 minutos
Palavra com peso	19 horas e 12 minutos

6.6.3 Regras e explanação

Para avaliar o valor dos dois tipos de representações (conceitos ou palavras) para explicar o raciocínio usado no processo de classificação, foram apresentados a dois médicos especialistas os descritores de classes em ambas as abordagens. Num caso, os descritores eram os conceitos e suas distribuições e noutro, termos e pesos correspondentes.

Os especialistas consideraram os descritores com conceitos mais significativos para entendimento do raciocínio de classificação. Isto porque os termos podem causar confusão devido ao problema do vocabulário. Por exemplo, o termo "*visual*" poderia referenciar "*deficiência visual*" ou o sintoma "*ilusão visual*". Da mesma forma, o termo "*sozinho*" poderia levar ao conceito "*morar sozinho*" ou ao conceito "*solilóquio*" (ato de falar sozinho).

6.6.4 Conclusões

A diferença de desempenho entre as abordagens baseada em conceitos e baseada em palavras não foi muita, tanto na classificação usando Rocchio quanto usando k-NN. Pode-se concluir que não há vantagem de um método de representação sobre outro em processos de classificação de textos, podendo-se usar palavras ou conceitos como características representantes do conteúdo dos textos.

Entretanto, as regras utilizadas pelo método baseado em conceitos são mais fáceis de serem entendidas. Isto facilita a explicação ou mesmo a validação do raciocínio usado no processo de classificação.

Também ganha-se com a redução na dimensionalidade das características usadas para representar os textos. No método baseado em conceitos, o número é menor e isto pode ter a vantagem do menor tempo de processamento.

A vantagem da abordagem baseada em palavras é que os métodos podem ser totalmente automáticos, seja para descrever classes, seja para representar ou classificar textos. Já a abordagem baseada em conceitos precisa da intervenção humana para a definição dos conceitos.

Entretanto, a definição de conceitos é algo que não deve mudar muito e, uma vez feita, poderá ser reaproveitada em processos futuros. Isto significa que, se for possível obter definições de conceitos e estas não mudarem muito (não necessitarem refinamentos com o tempo), é vantajoso utilizar métodos baseados em conceitos pelo seu bom desempenho, pelo menor tempo de processamento e pela facilidade em entender o raciocínio usado.

6.7 Avaliação de Agrupamento Baseado em Conceitos

O objetivo desta avaliação foi de testar a abordagem baseada em conceitos com outra técnica de mineração. Foi escolhida a técnica de agrupamento (*clustering*), a qual separa automaticamente (sem intervenção humana) elementos de um conjunto em grupos de afinidades, avaliando a similaridade entre os elementos. A meta do agrupamento é colocar no mesmo grupo os elementos mais similares entre si.

A técnica de agrupamento avalia a similaridade dos elementos por suas características. No caso de textos, as características podem ser relativas ao documento (autor, data de publicação, tamanho, língua, tipo, etc) ou relativas a seu conteúdo. Neste experimento, foram consideradas somente as características referentes ao conteúdo.

Também foi objetivo desta avaliação comparar o processo de agrupamento utilizando conceitos como características dos textos em contrapartida ao agrupamento baseado em palavras.

Foram utilizados dois tipos de características para representar os textos: (a) as palavras presentes nos textos e (b) os conceitos identificados nos textos pelo processo padrão

É importante salientar que a abordagem conceitual, mesmo utilizando-se de um processo de classificação, não altera o fundamento da técnica de agrupamento, que é separar automaticamente, sem intervenção humana, elementos em classes (não pré-definidas mas que serão identificadas durante o processo). Isto porque a abordagem conceitual foi utilizada apenas na identificação das características de cada texto e não para definir as classes a que os textos pertenceriam (isto o processo de agrupamento é que faz).

Os passos desta avaliação foram os seguintes:

1) extração de características dos textos (palavras ou conceitos);
2) criação de uma tabela de similaridades entre os textos;
3) aplicação de um algoritmo de agrupamento;
4) avaliação objetiva do resultado (grupos gerados e seus elementos);
5) avaliação subjetiva do conhecimento descoberto.

Para a extração de características, foram avaliadas 3 abordagens: uma que utiliza as palavras presentes nos textos como características e duas que usam os conceitos identificados pelo processo padrão. Neste último caso, foram avaliados os conceitos sem e com pesos associados, conforme discutido na seção 6.6.2.

A tabela de similaridade entre os textos foi criada utilizando-se a função de similaridade definida na seção 6.6.2. Foram calculados os graus de similaridade aos pares, para todos os textos da coleção, segundo as 3 abordagens avaliadas (palavras, conceitos sem pesos e conceitos com pesos).

O algoritmo de agrupamento utilizado foi o "*best-star*", tido como melhor entre os analisados em [WIV99]. Este algoritmo seleciona um texto por vez (chamado central) e localiza os textos mais similares a este, para colocá-los todos juntos num mesmo grupo. Os textos sendo comparados ao central somente serão colocados no grupo se não houver outro grupo com o qual possua maior similaridade.

A avaliação objetiva dos resultados foi feita quantitativamente usando a medida de ganho de informação (*information gain*) proposta em [BRA98]. O ganho de informação mede a "pureza" dos grupos resultantes do processo de agrupamento. O objetivo é ter grupos "puros", com pouca mistura de classes (com pouca entropia ou com maior ganho de informação). Quanto maior for a mistura de classes num grupo, maior será a entropia do grupo (a entropia na física, mede a quantidade de desordem num sistema). A seguir são detalhadas as fórmulas utilizadas:

- **Ganho de Informação** = (Entropia total) – (Entropia ponderada do processo)
- **Entropia total** = $-(p_1 \log_2 p_1) - \ldots - (p_n \log_2 p_n)$
  onde,
  $\mathbf{p_x}$ é o número de documentos da classe **x** na coleção dividido pelo número total de documentos na coleção;
  n é número de classes existentes na coleção;
- **Entropia ponderada do processo** = somatório das entropias ponderadas de cada grupo;
- **Entropia ponderada do grupo** = (número total de documentos no grupo / número total de documentos na coleção) * (Entropia do rupo);
- **Entropia do grupo** = $-(p_1 \log_2 p_1) - \ldots - (p_n \log_2 p_n)$
  onde,
  $\mathbf{p_x}$ é o número de documentos da classe **x** no grupo dividido pelo número total de documentos no grupo;
  n é o número de classes presentes no grupo.

Para se utilizar esta medida, é necessário saber qual o agrupamento ideal (quais grupos e seus elementos). No caso deste experimento, foram tomados como ideais 4 grupos referentes às 4 grandes classes do CID (afetivos, esquizofrenia, orgânicos e substâncias).

Foram realizadas 5 avaliações diferentes:

- Avaliação 1: foram selecionados aleatoriamente 12 textos de cada classe (na coleção de teste) para avaliar o agrupamento com um número parelho de textos;

- Avaliação 2: para confirmar o resultado anterior, foram escolhidos 12 textos diferentes para cada classe (da coleção de treino);
- Avaliação 3: foi utilizado um conjunto maior, mas também com número semelhante de textos para cada classe (entre 35 e 40), num total de 155 textos (extraídos da coleção de teste);
- Avaliação 4: foi feita a avaliação do agrupamento sobre os 200 textos da coleção de treino;
- Avaliação 5: o agrupamento foi feito sobre os 200 textos da coleção de teste.

Para as avaliações 1, 2, 4 e 5, foram utilizados os 65 conceitos correspondentes às características dos pacientes. Para a avaliação 3, foram utilizados estes 65 conceitos mais 32 conceitos relativos a remédios.

Maiores detalhes deste experimento podem ser obtidos em [SAR00].

6.7.1 Comparação entre características (palavras e conceitos)

TABELA 6.30 - Comparação de métodos de agrupamento (extraído de [SAR00])

	Nº de Grupos	Tempo de Agrupamento	Entropia Total	Entropia Ponderada Total	Ganho de Informação	Diferença (%) em relação ao maior G.I.
Avaliação 1 (48 prontuários)						
Palavras	17	03:37:22	2,000	0,944	1,056	12,51%
Conceitos com peso	17	00:01:40	2,000	0,793	**1,207**	
Conceitos sem peso	17	00:01:39	2,000	0,862	1,138	5,72%
Avaliação 2 (48 prontuários)						
Palavras	14	03:45:11	2,000	1,111	0,889	37,31%
Conceitos com peso	15	00:01:29	2,000	0,792	1,208	14,81%
Conceitos sem peso	17	00:01:28	2,000	0,582	**1,418**	
Avaliação 3 (155 prontuários)						
Palavras	38	21:22:52	1,998	1,067	0,931	34,53%
Conceitos com peso	48 + 3 avulsos	00:12:00	1,998	0,576	**1,422**	
Conceitos sem peso	42 + 3 avulsos	00:12:11	1,998	0,749	1,249	12,17%
Avaliação 4 (200 prontuários)						
Palavras	60	28:39:50	1,701	0,722	0,979	12,20%
Conceitos com peso	56 + 4 avulsos	00:18:52	1,701	0,588	1,113	0,18%
Conceitos sem peso	59 + 4 avulsos	00:18:52	1,701	0,586	**1,115**	
Avaliação 5 (200 prontuários)						
Palavras	56 + 2 avulsos	30:36:59	1,669	0,816	0,853	23,63%
Conceitos com peso	63 + 7 avulsos	00:18:24	1,669	0,552	**1,117**	
Conceitos sem peso	61 + 7 avulsos	00:17:54	1,669	0,752	0,917	17,91%

Pela tabela 6.30, pode-se notar que a abordagem baseada em conceitos resultou em agrupamentos com maior ganho de informação. A diferença da abordagem baseada em palavras para o melhor resultado sempre foi maior que a diferença do segundo melhor desempenho.

Entre os métodos conceituais (com ou sem pesos), não há um que possa ser considerado melhor. Entretanto, pode-se notar que o método que considera os pesos foi

melhor nas avaliações que usava textos da coleção de teste (avaliações 1, 3 e 5), enquanto que o método que não considera pesos foi melhor com textos da coleção de treino (avaliações 2 e 4). Não foi possível identificar uma razão para estes resultados.

Fez-se ainda uma última avaliação. Como havia associada a cada texto (na coleção de teste) a indicação do médico que gerou o prontuário, foram avaliados os resultados utilizando como agrupamento ideal os grupos de textos escritos pelo mesmo médico. Esta avaliação foi feita sobre o mesmo conjunto de textos da avaliação 1 (médico X = 2 textos; médico M = 3; médico I = 2; médico D = 9; médico C = 8; médico L = 16; médico A = 3; médico Q = 3 e médico S = 2).

A entropia total do conjunto é de 2,735. As medidas de ganho de informação para cada abordagem foram as seguintes:

- palavras: 1,901
- conceitos com peso: 1,547
- conceitos sem peso: 1,523

Por estes resultados numéricos, pode-se constatar que a abordagem por palavras tende a agrupar melhor os textos por autor, pois os médicos tendem a usar termos próprios.

6.7.2 Avaliação da descoberta de conhecimento

Também procurou-se descobrir conhecimento novo a partir dos resultados do processo de mineração usando a técnica de agrupamento. Para tanto, foi necessário reutilizar a técnica de análise de distribuições mas desta vez em separado para cada grupo resultante do agrupamento.

Foi possível descobrir características (conceitos) com 100% de presença em determinados grupos (isto é, apareciam em todos os textos do grupo). Isto pode levantar a hipótese de que aquele grupo de pacientes (ou casos) tenha sido reunido pela técnica de agrupamento por justamente terem estas características. Esta lista de características pode ser usada para descrever o grupo de pacientes e distingui-los de outros grupos (modelos de descrição).

A importância de tal descoberta se dá porque em muitos casos não é possível descobrir as características comuns de um grupo grande de elementos, e processos manuais de separação podem interferir no processo. Por exemplo, buscando-se entender um grupo de pacientes com determinado diagnóstico que tiveram alta em pouco tempo, procurou-se por características comuns neste grupo. Entretanto, o resultado foi um conjunto vazio (nenhuma característica em comum, seja tipo de tratamento, remédio usado ou mesmo características dos pacientes). Disto conclui-se que não há um fator único que possa levar à cura. Se separados manualmente, estes pacientes poderiam gerar grupos relativos à idade, profissão, camada social, altura, etc., o que não é bom em razão da pré-discriminação de grupos. Assim, a técnica de agrupamento separa sem interferências os elementos por grupos de afinidades. A análise das características comuns a cada grupo resultante permite revelar que fatores em conjunto determinaram a cura destes pacientes (sejam características do paciente ou do tratamento).

O conhecimento descoberto por este experimento está sendo validado por médicos especialistas (detalhes podem ser encontrados em [SAR00]). Entretanto, é possível afirmar que a abordagem baseada em conceitos quando utilizada com a técnica de agrupamento para

a etapa de mineração também pode gerar conhecimento novo e útil. Conclui-se que é possível utilizar a abordagem com outras técnicas de mineração.

6.8 Avaliação da Descoberta Proativa

Uma discussão importante é se ferramentas de software poderão extrair automaticamente conhecimento a partir de coleções textuais. Foi feita uma investigação sobre a possibilidade de o processo de descoberta sobre conceitos ser feito quase que automaticamente (sem intervenção humana).

A maioria dos pesquisadores concorda que o processo de descoberta é cíclico, tendo como passos principais: [PAR89] [AGR93] [ING96]

a) a formulação de hipóteses;
b) o teste das hipóteses;
c) a observação dos resultados (para refutar ou confirmar as hipóteses); e
d) a revisão das hipóteses e a sua modificação (reiniciando o processo), até que o usuário se dê por satisfeito.

Entretanto, esta estratégia só pode ser aplicada quando o usuário consegue formular hipóteses iniciais, ou seja, quando ele tem idéia de qual é o seu objetivo ou necessidade e sabe do que precisa.

De acordo com Choudhury e Sampler [CHO97], existem dois modos para aquisição de informação: o modo reativo e o modo proativo. No primeiro caso, a informação é adquirida para resolver um problema específico do usuário (uma necessidade resultante de um estado anômalo de conhecimento). Nestes casos, o usuário sabe o que quer e poderá identificar a solução para o problema quando a encontrar.

Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório, para detectar problemas potenciais ou oportunidades. Neste segundo caso, o usuário não tem um objetivo específico.

Oard e Marchionini [OAR96] classificam as necessidades de informação em estáveis ou dinâmicas e em específicas ou abrangentes (gerais). Taylor (citado em [OAR96]) define 4 tipos de necessidades, os quais formam uma escala crescente para a solução do problema:

- necessidades viscerais: quando existe uma necessidade ou interesse, mas esta não é percebida de forma consciente;
- necessidades conscientes: quando o usuário percebe sua necessidade e sabe do que precisa;
- necessidades formalizadas: quando o usuário expressa sua necessidade de alguma forma;
- necessidades comprometidas: quando a necessidade é representada no sistema.

As necessidades tratadas pela abordagem de descoberta reativa poderiam ser classificadas como estáveis e específicas, segundo a classificação de Oard e Marchioninni, e como conscientes (no mínimo), segundo Taylor. Isto porque o usuário sabe o que quer, mesmo que não consiga formalizar.

No modo reativo, o usuário tem uma idéia, mesmo que vaga, do que pode ser a solução ou, pelo menos, de onde se pode encontrá-la. Pode-se dizer então que o usuário possui algumas hipóteses iniciais, que ajudarão a direcionar o processo de descoberta. Neste

caso, é necessário algum tipo de pré-processamento, por exemplo para selecionar atributos ou valores de atributos. Isto exige entender o interesse ou objetivo do usuário para limitar o espaço de busca (na entrada) ou filtrar os resultados (na saída). É o caso típico de quando se deseja encontrar uma informação específica, por exemplo, um valor para um atributo ou um processo (conjunto de passos) para resolver um problema.

Já as necessidades da abordagem proativa poderiam ser classificadas como dinâmicas e abrangentes, segundo a classificação de Oard e Marchioninni. São dinâmicas porque podem mudar durante o processo, já que o objetivo não está bem claro, e são abrangentes porque o usuário não sabe exatamente o que está procurando. Pela taxonomia de Taylor, as necessidades do modo proativo são viscerais. Isto quer dizer que há uma necessidade ou objetivo, mas o usuário não consegue definir o que precisa para resolver o problema. A necessidade típica do modo proativo poderia ser representada pela expressão: "*diga-me o que há de interessante nesta coleção*". Neste caso, o usuário não tem de forma definida o que lhe seja de interesse (o que precisa), podendo tal interesse mudar durante o processo. Pode-se dizer que é um processo exploratório, sendo, em geral, iterativo (com retroalimentação) e interativo (com participação e intervenção do usuário).

Na abordagem proativa, não há hipóteses iniciais ou elas são muito vagas. O usuário deverá descobrir hipóteses para a solução do seu problema e explorá-las, investigá-las e testá-las durante o processo. Em geral, acontece porque o usuário não sabe exatamente o que está procurando. É o caso típico de quando se quer monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo, talvez sendo necessário avaliar as hipóteses, para verificar se são verdadeiras ou não.

O objetivo desta avaliação foi o de analisar três aspectos no processo de descoberta:

a) a viabilidade de se fazer descoberta proativa de conhecimento sobre conceitos;
b) que tipos de estratégias podem ou devem ser usadas na descoberta proativa (por exemplo, por onde começar e que passos seguir depois);
c) qual a importância da intervenção humana no processo e como o conhecimento prévio sobre o assunto pode influenciar o processo de descoberta.

6.8.1 Viabilidade da Descoberta Proativa

Os experimentos realizados mostram que é possível automatizar partes do processo de descoberta, minimizando a dependência a pessoas. Foi possível observar que a descoberta proativa (aquela que inicia sem hipóteses) é viável. Isto é possível depois que os conceitos estão definidos. Neste caso, o processo de mineração pode ser realizado de forma automática sem necessidade de condução por pessoas. Pode-se verificar isto analisando os resultados das duas técnicas de mineração empregadas no processo padrão, quando aplicadas à coleção toda. O conhecimento resultante diz respeito a todos os casos e serve para entender o perfil do paciente típico da clínica.

As diferenças da abordagem proativa em relação ao modelo de descoberta de conhecimento proposto por Goebel e Gruenwald [GOE99] (apresentado no capítulo 4 e considerado reativo) são que:

1) o passo (a) pode não ter uma definição precisa do objetivo do processo (há um objetivo ou problema, mas a solução não pode ser prevista);

2) o passo (e) não existe na abordagem proativa, ou seja, o usuário não sabe ou não deseja formular hipóteses para a solução de seu problema.

Quanto à definição dos conceitos, é possível usar, no processo padrão de descoberta de conhecimento, conceitos previamente definidos, eliminando assim a necessidade de intervenção humana nesta etapa. Definições de conceitos podem ser conseguidas reutilizando definições feitas no passado ou utilizando vocabulários controlados (tal como um *thesaurus*) adaptáveis ao formato requerido pelo processo padrão.

Mesmo que seja difícil ou que não se queira eliminar a intervenção humana na definição dos conceitos, esta etapa pode ser parcialmente automatizada com auxílio de ferramentas de software, diminuindo assim a dependência a humanos. Em outro experimento (descrito no artigo 6 dos anexos), feito sobre uma coleção de textos políticos (total de 358 textos), não havia conhecimento prévio sobre os conceitos que apareciam nos textos. Esta situação é diferente do caso de psiquiatria, onde os conceitos foram escolhidos para representar características dos pacientes ou remédios. No caso dos textos políticos, foi necessário investigar a coleção de textos para se conhecer que conceitos estavam sendo referenciados. Ferramentas de software ajudaram a analisar cada termo presente na coleção (mais de dois mil termos), selecionando os mais significativos e categorizando-os. No final, havia 104 conceitos definidos.

6.8.2 Estratégias para descoberta proativa

Um dos problemas da abordagem proativa de descoberta é definir uma estratégia para investigar a coleção textual a fim de serem descobertas hipóteses iniciais (já que o processo inicia sem hipóteses).

Kuhlthau [KUH91] determinou seis fases em um processo de descoberta de informação: iniciação, seleção, exploração, formulação, coleção e apresentação. Cada fase é caracterizada por atitudes diferentes do usuário (por exemplo, em relação a sentimentos, pensamento, ações e tarefas). Uma das descobertas mais interessantes desta pesquisadora é que o usuário inicia procurando algum tipo de conhecimento mais geral, depois ele procura informação relevante em grupos mais restritos e termina procurando informações mais focadas ou específicas. Durante este processo, o usuário reconhece, identifica, investiga, formula, reúne e complementa o conhecimento.

Watts e Porter [WAT97] propõem um esboço de metodologia (*framework*) sobre algumas ferramentas de descoberta. A estratégia proposta é apropriada para encontrar nomes de pessoas e companhias, através da análise do vocabulário empregado. Entretanto, a estratégia é sugerida somente para resolver problemas da área de Inteligência Competitiva.

Seguindo as sugestões destes trabalhos e com base nas observações feitas durante os experimentos, foi possível identificar alguns passos comuns. Sugere-se então uma estratégia para descoberta proativa de conhecimento em textos. Não se pode considerar esta estratégia uma metodologia, mas sim um referencial, que poderá conduzir os usuários no processo, indicando os passos principais (técnicas ou ferramentas a serem usadas). Os passos são resumidamente descritos a seguir:

1) seleção de textos: o primeiro passo é selecionar uma coleção de textos sobre os quais serão aplicadas as técnicas de descoberta; as técnicas automáticas mais

indicadas são a recuperação de informação (que encontra textos procurando por palavras-chave ou termos presentes nos textos) e a classificação (que separa textos por assunto); outra possibilidade, é o usuário mesmo encontrar ou selecionar manualmente os textos;

2) definição de conceitos: podem ser usados conceitos previamente definidos ou os termos presentes na coleção podem ser analisados e categorizados para gerar conceitos novos;
3) análise da coleção toda ou de partes: neste ponto, o usuário deve decidir se irá aplicar as técnicas de descoberta sobre todos os textos ou sobre partes; a sugestão é que se comece analisando toda a coleção e depois se examine subcoleções; em alguns casos, nada de interessante é encontrado na coleção toda, o que leva o usuário necessariamente a investigar pequenos grupos; a separação em grupos pode ser feita de forma automática, com a técnica de agrupamento (*clustering*), ou sob algum critério estabelecido pelo usuário, como por exemplo selecionando partes de interesse com as técnicas de recuperação ou classificação;
4) análise de grupos de textos (toda a coleção ou partes): pode-se aplicar a técnica de análise de distribuição ou a técnica associativa;
5) comparação de subcoleções entre si ou em relação à coleção toda: os resultados para cada grupo (subcoleção) podem ser comparados entre si ou com os resultados obtidos com a coleção toda; características comuns ou exclusivas aos grupos podem levantar hipóteses interessantes; também é importante investigar variações nas distribuições;
6) validação de hipóteses: as hipóteses levantadas devem ser avaliadas com outras técnicas de descoberta;
7) retroalimentação: como o processo é cíclico, alguns passos ou o processo todo podem ser refeitos.

6.8.3 Importância da intervenção humana e de conhecimentos prévios

Apesar de que o processo de descoberta possa ser começado sem que o usuário defina hipóteses (abordagem proativa), a intervenção humana ainda se faz necessária. Para levantar ou validar hipóteses é necessário interpretar os resultados sob o contexto do domínio. Somente pessoas ligadas ao domínio podem filtrar os resultados para extrair conhecimento interessante. Em alguns casos, o conhecimento descoberto pode ser útil mas não novo e em outros ele pode ser novo mas não útil. Somente as pessoas do domínio podem avaliar o que é útil e novo e isto sim se torna conhecimento interessante. Neste caso, a intervenção humana serve para filtrar o que é interessante (útil e novo). Segundo Aamodt and Nygard [AAM95], o conhecimento é imprescindível para que os dados possam ser interpretados e se tornem informação. E o conhecimento é subjetivo e depende das pessoas.

Além de necessária, a intervenção humana também pode ser muito útil no processo de descoberta. Por exemplo, nos experimentos com a coleção de psiquiatria, a mineração foi aplicada aos 4 grupos previamente selecionados por especialistas do domínio. Isto permitiu descobrir conhecimento sobre cada classe de diagnóstico, o que seria difícil sem que especialistas do domínio tivessem associado diagnósticos aos textos.

Quanto à definição dos conceitos, esta pode ser melhorada com a intervenção humana, como se pôde concluir das avaliações feitas anteriormente sobre os diferentes métodos testados. Neste caso, as pessoas puderam intervir no processo escolhendo limiares, selecionando sinônimos, eliminando termos muito genéricos e analisando alarmes falsos para corrigir erros.

Da mesma forma, o conhecimento prévio (*background knowledge*) sobre o domínio também é importante no processo de descoberta de conhecimento em textos. Este conhecimento pode ser usado pelo usuário para limitar o espaço de pesquisa ou análise (no caso de descoberta proativa) ou para definir hipóteses (caso reativo).

Também pode-se usar conhecimento prévio na definição dos conceitos ou na interpretação dos resultados. Neste caso, ele é útil para eliminar ambigüidades.

Feldman e Hirsh [FEL97] aplicam técnicas de descoberta sobre palavras, permitindo que o usuário intervenha no processo, fazendo uso de seus conhecimentos prévios sobre o domínio ou assunto. Isto acelera o processo e permite filtrar os resultados de acordo com o interesse do usuário.

Notou-se, nos experimentos, que a falta de conhecimento sobre o domínio pode resultar em descobertas que não podem ser aproveitadas e, conseqüentemente, também em desperdício de esforços. Pelos experimentos, descobriu-se que uma boa maneira de obter um pouco mais de conhecimento sobre o domínio é examinando os termos mais freqüentes. Isto permite ao usuário conhecer o estilo dos textos ou o escopo do conteúdo (do que se fala e do que não se fala nos textos). Para maiores discussões teóricas sobre o assunto, ver Choudhury and Sampler [CHO97], que discutem tipos de conhecimento prévio em processos de aquisição de conhecimento.

7 APLICAÇÕES DA ABORDAGEM PROPOSTA

Neste capítulo, serão descritas algumas aplicações da abordagem baseada em conceitos proposta nesta tese.

7.1 Análise Qualitativa e Quantitativa de Documentação Textual

A abordagem baseada em conceitos permite que textos sejam analisados em um nível mais alto do que as palavras. Ou seja, o conteúdo dos textos é avaliado em relação ao domínio e não simplesmente de forma estatística e independente da realidade. Esta análise busca entender como os conceitos de um domínio estão sendo utilizados nos textos. Assim, pode-se dizer que a análise é feita mais próxima da realidade.

Nos experimentos com a coleção de textos de psiquiatria, foi possível analisar características dos pacientes, tais como sintomas, sinais e seu comportamento social e familiar. Se a análise fosse feita somente ao nível das palavras, seria difícil relacionar os padrões encontrados com os fenômenos da realidade (eventos, pessoas, objetos, sentimentos).

A identificação de conceitos permite a análise qualitativa de uma coleção textual, isto é, que características (temas, conceitos, assuntos, fenômenos da realidade) estão sendo referenciados nos textos. Já a etapa de mineração permite a análise quantitativa, uma vez que possibilita identificar padrões numéricos nas distribuições de conceitos e nas relações entre estes. A combinação destes dois tipos de análise é que gera conhecimento novo e útil sobre o domínio.

Os resultados do processo de descoberta podem ser usados para entender um domínio, para treinar pessoas ou para apoiar e validar decisões.

Um exemplo de como o conhecimento descoberto pode ajudar a entender um domínio é o caso dos textos da clínica psiquiátrica. Na figura 7.1, são apresentados alguns padrões interessantes encontrados na coleção toda. Na coluna da esquerda aparecem os conceitos mais freqüentes e na coluna da direita as regras associativas com maior grau de confiança.

Distribuições de Conceitos	Regras Associativas
familiares – 84,5% agressividade – 77,0% inapetência – 76,0% remédios – 74,5% insônia – 71,0% alteração de pensamento – 70,5% nervosismo – 68,5% alteração de atenção – 54,5%	alcoolismo ➔ inapetência (84%) autismo ➔ alteração de pensamento (95,3%) agressividade ➔ familiares (92,8%) depressão ➔ insônia (85,1%) religião ➔ remédios (85,1%)

FIGURA 7.1 - Alguns padrões descobertos na coleção toda

Analisando a figura 7.1, pode-se identificar o perfil do paciente típico que é atendido na clínica. Isto quer dizer que, pela figura 7.1: 84,5% dos pacientes têm familiares ou fazem

algum tipo de referência a estes; 77% apresentam sinais de agressividade; 76% citaram sofrer de algum tipo de inapetência (falta de apetite), 74,5% já fazem uso de algum remédio, etc.

Analisando-se as regras associativas da figura 7.1, é possível predizer características pela presença de outras ou inferir relações de dependência entre as características. Por exemplo, pode-se notar que: em 84% dos casos, pacientes com sintomas de alcoolismo também apresentam inapetência (falta de apetite); quase sempre, sintomas de autismo são acompanhados de alteração de pensamento; 85,1% dos pacientes com sintomas de depressão apresentam também insônia; e 85,1% dos pacientes que citaram algo relacionado a religiões tomam remédios. Pode-se também levantar a hipótese de que a agressividade esteja relacionada à família, já que 92,8% dos pacientes com sintomas de agressividade citaram familiares.

Estes resultados podem servir para estudos epidemiológicos e assim ajudar a melhorar o tratamento ou a prevenção de doenças.

Na figura 7.2, são apresentados padrões (conceitos mais freqüentes e regras associativas) referentes aos pacientes com o diagnóstico de esquizofrenia. Analisando tais padrões, é possível verificar que os pacientes esquizofrênicos apresentam maior incidência de "*alteração de pensamento*" (83,5%) do que a média geral (70,5% na figura 7.1). Pelas regras de associação, é possível inferir que pacientes que "*ouvem vozes*" têm predisposição para agressividade, e pacientes homicidas (com alguma história envolvendo ameaças de morte) também apresentam sintomas de agressividade.

Distribuições de Conceitos	Regras Associativas
familiares – 84,5%	agressividade ➔ familiares (92,94%)
alteração de pensamento – 83,5%	alteração de atenção ➔ agressividade (88,89%)
agressividade – 82,5%	
nervosismo – 75,7%	homicida ➔ agressividade (97,67%)
insônia – 75,7%	homicida ➔ familiares (95,35%)
remédios – 73,8%	ouvir vozes ➔ agressividade (90,91%)
inapetência – 68,9%	perseguição ➔ insônia (84,85%)
perseguição – 64,1%	insônia ➔ remédios (80,77%)
ouvir vozes – 64,1%	
alteração de atenção – 52,4%	

FIGURA 7.2 - Padrões para o diagnóstico de esquizofrenia

Os padrões de esquizofrenia podem ser comparados com os de outras doenças. Na figura 7.3, por exemplo, é apresentado o conhecimento descoberto sobre os pacientes com distúrbios afetivos. Pode-se notar que a incidência de insônia é bem maior nos pacientes com distúrbios afetivos que nos esquizofrênicos (85,2% contra 75,7%). Nesta segunda classe, aparecem novos sintomas entre os mais freqüentes: "*suicida*" (81,5%), "*depressão*" (74,1%) e "*choro*" (63%). As regras associativas também apresentam diferenças.

Os padrões descobertos em cada classe de paciente permitem traçar o perfil de pacientes para cada diagnóstico.

Também foi possível descobrir padrões referentes ao uso de remédios através das análises qualitativa e quantitativa. A seguir, são apresentados os conceitos mais freqüentes identificados na sub-coleção de pacientes que utilizaram o remédio Dienpax: "*inapetência*" (91.8%), "*agressividade*" (83.7%), "*alteração de pensamento*" (78.3%), "*nervosismo*" (75.6%), "*insônia*" (64.8%), "*alcoolismo*" (62.1%), "*ouvir vozes*" (59.4%).

Distribuições	Regras Associativas
insônia – 85,2% familiares – 85,2% suicida – 81,5% inapetência – 81,5% remédios – 77,8% depressão – 74,1% pensamento – 70,4% alteração de atenção – 66,7% choro – 63,0% nervosismo – 63,0% agressividade – 59,3%	agressividade ➔ familiares (87,50%) homicida ➔ suicida (92,31%) remédios ➔ suicida (85,71%) morte ➔ inapetência (90,91%) inapetência ➔ insônia (86,36%) inapetência ➔ suicida (81,82%)

FIGURA 7.3 - Padrões para o diagnóstico de distúrbios afetivos

Os conhecimentos resultantes do processo de descoberta podem ser utilizados no treinamento de estudantes e médicos assistentes. Também é possível validar as decisões tomadas pelos médicos comparando os resultados encontrados na coleção com o conhecimento tido como correto. O interessante desta abordagem é que ela permite entender parte do processo de raciocínio utilizado por especialistas do domínio.

Maiores detalhes sobre estes experimentos podem ser obtidos nos artigos 1 e 3 dos anexos.

7.2 Formalização e Exploração de Conhecimento Tácito

O conhecimento pode ser classificado em tácito ou explícito [NON97]. O primeiro é aquele conhecimento que não está formalizado. Em geral, este tipo de conhecimento encontra-se com as pessoas, sem ainda ter sido transformado em representações rigorosas. Já o segundo tipo de conhecimento é justamente aquele que foi formalizado em documentos, bancos de dados, gráficos, desenhos, etc.

Em uma organização, os dois tipos de conhecimento devem coexistir harmonicamente e de alguma forma interagir para que todo o potencial de utilização possa ser aproveitado. Nonaka e Takeuchi [NON97] identificaram 4 modos de conversão e interação entre os conhecimentos tácito e explícito. O processo de **externalização** é a transformação do conhecimento tácito em explícito. A **internalização** é o processo inverso. Já a **combinação** é o processo de interação entre conhecimentos explícitos para geração de novos conhecimentos. Por sua vez, a **socialização** é a interação entre conhecimentos tácitos.

Foram feitos experimentos para analisar o conhecimento tácito disponível com clientes de uma empresa e com colaboradores de uma clínica médica (maiores detalhes sobre estes experimentos podem ser encontrados no artigo 1 dos anexos).

Em ambos os casos, o conhecimento tácito foi externalizado através de textos sem restrições (as pessoas, clientes ou médicos, escreviam textos em linguagem livre). Depois, utilizou-se a abordagem de *KDT* proposta nesta tese para formalizar e explorar o conhecimento disponível nos textos de forma implícita.

No primeiro caso, os conceitos foram identificados e definidos analisando-se a linguagem utilizada (termos e significados). No segundo caso, os conceitos foram definidos como no processo padrão, explicado na seção 5.4.

O conhecimento descoberto em ambos os casos foi útil para entender como as pessoas pensam sobre determinado assunto. No caso dos clientes, o conteúdo dos textos dizia respeito a reclamações e sugestões dos clientes sobre processos e produtos da empresa. Foi possível identificar os temas mais preocupantes para os clientes, possivelmente pontos fracos da empresa. Em alguns textos, havia referências a concorrentes. Isto permitiu saber o que os clientes estavam pensando sobre os concorrentes ou sobre a empresa em relação à concorrência.

No caso médico, pôde-se entender parte do raciocínio utilizado pelos especialistas no processo de diagnóstico ou de prescrição de remédios. O conhecimento adquirido pode ser utilizado para validar ou apoiar decisões e para treinamento de novatos.

Os resultados destes dois experimentos demonstram que é viável formalizar conhecimento tácito em textos livres e depois explorá-lo com a abordagem de descoberta proposta nesta tese.

7.3 Construção de Sistemas Automatizados de Apoio à Decisão

Outro benefício do processo de descoberta baseado em conceitos é que os resultados podem ser usados para a criação de sistemas automatizados para validar ou apoiar decisões.

Neste caso, a abordagem de descoberta funciona como um mecanismo de aprendizagem supervisionada, onde textos são selecionados como exemplos de casos, e os padrões descobertos servem para descrever as características destes casos.

Como se pôde verificar pelas avaliações feitas e apresentadas nesta tese, um sistema automatizado criado com o conhecimento descoberto pelo processo padrão chegou a mais de 60% de acertos no diagnóstico de novos casos de psiquiatria, descritos textualmente.

Outra vantagem da abordagem é a redução no esforço para aquisição de conhecimento e modelagem de sistemas inteligentes. Wilcox e outros [WIL00] afirmam que processos manuais de inferência são difíceis e demandam muito tempo e esforço. A abordagem proposta auxilia na aquisição de conhecimento através da descoberta de conhecimento novo codificado em textos. Isto não implica em eliminar a intervenção humana, mas sim em aproveitar ferramentas automatizadas para fazer um melhor uso do conhecimento disponível e, muitas vezes, implícito em coleções textuais.

Maiores detalhes sobre a construção de sistemas automáticos com a abordagem proposta são discutidos no artigo 2 dos anexos.

7.4 Classificação e Recuperação de Documentos Textuais

O sistema automático implementado com o conhecimento descoberto pela abordagem permitiu classificar casos psiquiátricos em classes de diagnósticos. Vale lembrar que os casos estavam descritos de forma textual (documentos semi-estruturados mas com textos livres). Este sistema é do tipo classificador de textos (*text classifier*), uma vez que analisa textos novos, compara-os ao conhecimento embutido e determina a classe do texto.

Pode-se considerar que sistemas semelhantes podem ser construídos da mesma forma para outros tipos de aplicação que exijam classificação de textos, por exemplo:

classificação de notícias de jornais em assuntos, seleção de casos jurídicos por tipo, encaminhamento de mensagens de correio eletrônico para pessoas interessadas, organização de sugestões e reclamações de clientes por tipo, busca e filtragem de páginas Web, recuperação de bibliografias por resumos, etc.

7.5 Inteligência Competitiva

Experimentos foram feitos com a abordagem de *KDT* proposta nesta tese para suportar atividades relativas à Inteligência Competitiva (IC). A Inteligência Competitiva (*Competitive Intelligence*) é a área que procura suprir uma empresa com informações estratégicas sobre sua posição no mercado, em relação aos concorrentes e frente a seus clientes [ZAN98].

No nível atual de globalização e de competitividade, as informações do ambiente interno e externo de uma empresa tornam-se extremamente valiosas. As informações pertinentes ao ambiente interno de uma empresa são todas aquelas relacionadas com os seus produtos, vendas, serviços, estoques, empregados e fornecedores. Já o ambiente externo pode ser compreendido como o mercado em si e suas tendências, novas tecnologias de produção, a opinião e a satisfação dos clientes em relação à empresa e às ações da concorrência.

Muitas informações sobre mercados, produtos, serviços e empresas atuantes estão disponíveis publicamente, podendo ser analisadas por qualquer pessoa de forma lícita. Coletar e analisar essas informações é extremamente importante para que uma empresa adquira um diferencial competitivo.

A abordagem baseada em conceitos foi aplicada a 3 casos com o objetivo de descobrir conhecimento que pudesse ajudar em processos de IC.

7.5.1 Estratégias agroalimentares no MERCOSUL

Neste experimento, o objetivo era descobrir algo de interessante numa coleção de textos sobre potencialidades e oportunidades agroalimentares no MERCOSUL. A coleção era formada por 6 textos, cada um contendo informações sobre potencialidades, desafios, demandas e oportunidades do setor agroalimentar de um país latino-americano (maiores detalhes no artigo 4 dos anexos). Os conceitos foram definidos por grupos de termos relacionados e referenciavam características ou produtos de países.

Os resultados do processo de descoberta permitiram levantar hipóteses sobre a existência de temas comuns aos países (por exemplo, preocupações ou estratégias comuns) ou então de temas exclusivos de um único país (por exemplo, demandas ou estratégias particulares de um país). Algumas destas descobertas foram validadas positivamente por especialistas no assunto.

7.5.2 Marketing Político

Neste caso, um processo de descoberta foi conduzido sobre uma coleção de textos extraídos de um jornal *online*, os quais falavam de um determinado governante público. A coleção foi dividida em duas sub-coleções de acordo com o ano de publicação das notícias

(1997 e 1999). Maiores detalhes sobre o experimento são discutidos no artigo 4 dos anexos. Os conceitos foram definidos analisando-se todos os termos presentes na coleção.

As distribuições dos temas foram comparadas nas duas sub-coleções. Analisando temas que apareciam numa coleção e não na outra e temas que tinham distribuições muito diferentes, pôde-se levantar hipóteses sobre o domínio, as quais foram investigadas com ajuda da intervenção humana e de conhecimentos prévios sobre o domínio.

Esta aplicação demonstra que é possível descobrir conhecimento novo e útil sobre pessoas ou entidades, através da análise de conceitos presentes em textos públicos. Os resultados poderiam ser usados para traçar estratégias de ação ou de publicidade por parte de pessoas ou entidades. Também pode-se ter idéia de como as pessoas e entidades e suas atuações estão sendo vistas pela mídia. Isto pode sugerir mudanças no comportamento da pessoa ou da entidade.

7.5.3 Benchmarking de ferramentas de *KDD* e de *KDT*

Seguindo-se o trabalho de Goebel e Gruenwald [GOE99], procurou-se fazer uma comparação (*benchmarking*) de ferramentas de descoberta de conhecimento divulgadas na Internet. Neste caso, o objetivo era encontrar as técnicas utilizadas nas ferramentas (maiores detalhes nos artigos 4 e 6 dos anexos).

Analisando a linguagem utilizada na coleção, foi possível identificar que técnicas estavam sendo apresentadas e assim definir conceitos para cada uma destas técnicas. Também foram definidos conceitos para representar os benefícios que as empresas ofereciam com suas ferramentas.

Através da mineração pela análise de distribuições, pôde-se saber que técnicas eram mais comuns e quais estavam sendo menos utilizadas. Também foi possível identificar que apenas duas ferramentas alegavam ter todas as técnicas definidas. As regras associativas resultantes da mineração permitiram identificar os benefícios correspondentes a cada técnica, conforme alegado nos textos.

Tais descobertas poderão ser úteis em estratégias de implementação de futuras ferramentas, cabendo aos projetistas decidir se irão privilegiar as técnicas mais usadas ou as menos empregadas e se irão ou não implementar todas as técnicas disponíveis. Também poderão ser usadas as mesmas estratégias de divulgação de benefícios (benefícios X técnicas) ou a empresa poderá alegar benefícios diferentes dos concorrentes.

7.6 Inteligência do Negócio

O ramo da ciência que estuda e aplica métodos e técnicas de análise de informações para geração de inteligência com o intuito de oferecer vantagem competitiva a uma empresa é chamado de Inteligência do Negócio (*business intelligence*) [WAN99]. Alguns autores fazem uma pequena distinção entre os processos de inteligência: Inteligência do Negócio fica responsável pela análise de dados relativos à organização e Inteligência Competitiva é mais voltada para a análise dos dados do mercado e dos concorrentes.

Muitas organizações possuem informações importantes para se obter inteligência do negócio, disponíveis em formato textual. Entretanto, há uma enorme dificuldade em tratar este tipo de informação, pela falta de ferramentas e estratégias adequadas.

A abordagem de *KDT* proposta nesta tese pode auxiliar as organizações na análise de informações textuais para suportar atividades de inteligência do negócio.

Foram feitos alguns experimentos neste sentido, descritos em mais detalhes nos artigos 1 e 4 dos anexos.

Num dos experimentos, a abordagem de *KDT* foi aplicada a textos gerados a partir de uma pesquisa com clientes de uma empresa de TV por assinatura. Os textos continham sugestões e reclamações dos clientes sobre produtos e serviços da empresa. Cada texto (total de 225) correspondia a um cliente e fora escrito em formato livre.

Os conceitos para este experimento foram definidos de forma proativa, ou seja, pela análise da linguagem utilizada na coleção (termos e significados).

Na figura 7.4, são apresentados alguns exemplos de padrões interessantes descobertos nesta aplicação. Na primeira coluna, aparecem padrões referentes à distribuição dos conceitos na coleção toda, e na segunda coluna, as regras associativas derivadas da coleção toda com o seu grau de confiança.

Algumas conclusões podem ser obtidas dos resultados apresentados na figura 7.4. Metade dos clientes tem alguma sugestão ou reclamação sobre filmes. Em geral, uma sugestão vem de uma insatisfação e também pode ser considerada uma reclamação, só que não explícita. Destas reclamações (sobre filmes), 39,5% falam também de repetição, como pode ser notado nas regras associativas. Infere-se que esta é uma insatisfação dos clientes. Este então é um ponto fraco da empresa e seu negócio pode ser melhorado diminuindo-se a repetição de filmes. Ainda pode-se notar que alguns poucos clientes citaram a concorrência (5,3%), mas este pode ser um valor alto para a empresa (deve-se analisar a proporção de perdas de clientes). Destes, segundo as regras associativas, 33,3% citaram o custo. Infere-se que estes clientes estão dizendo que a concorrência tem custo menor.

Distribuições de Conceitos	**Regras Associativas**
filmes – 50,7%	imagem ➔ qualidade (80,00%)
custo – 20,4%	pacote *A* ➔ custo (66,67%)
programação – 19,6%	concorrência ➔ filmes (58,3%)
pacote – 15,6%	filmes ➔ repetição (39,5%)
revista – 10,7%	atendimento ➔ demora (37,5%)
pay per view – 6,2%	concorrência ➔ custo (33,3%)
esportes – 5,3%	filmes ➔ qualidade (18,4%)
concorrente – 5,3%	filmes ➔ concorrência (6,1%)
imagem – 4,4%	filmes ➔ pay per view (6,1%)
som – 4,4%	filmes ➔ lançamento (4,4%)
documentários – 3,1%	
seriados – 3,1%	
futebol – 2,7%	

FIGURA 7.4 - Padrões descobertos na coleção toda

Nesta pesquisa, havia também informações estruturadas, como tipo de plano (pacote) do cliente e seu canal preferido. Estes dados foram usados para separar a coleção de textos por classes (sub-coleções) com o intuito de explorar as sugestões e reclamações referentes a cada tipo ou perfil de cliente. Assim, pôde-se combinar o conhecimento explícito, presente na forma estruturada, com os conceitos extraídos dos textos não-estruturados.

A figura 7.5 apresenta os conceitos mais freqüentes nas reclamações dos clientes do pacote A (mais caro) e dos clientes do pacote D (mais barato).

Pacote A	*Pacote D*
filmes – 36,4% custo – 24,2% repetição – 22,7% programação – 22,7%	custo – 38,5% atendimento – 23,1% filmes – 15,4%

FIGURA 7.5 - Padrões por tipo de pacote

Pode-se observar, na figura 7.5, que os clientes do pacote *A* reclamam menos do custo que os clientes do pacote *D* e que os primeiros estão mais insatisfeitos com os filmes da programação geral do que os segundos. Este conhecimento permite entender melhor os interesses dos clientes de cada pacote, podendo-se gerar um perfil de clientes por tipo de pacote.

Como havia o registro explícito do canal preferido de cada cliente, a coleção de textos foi dividida em 3 partes referentes ao tipo de canal preferido (esportes, filmes e notícias). Na figura 7.6, são apresentados os conceitos mais freqüentes por tipo de canal preferido.

Esportes	*Filmes*	*Notícias*
filmes – 39,4% custo – 30,3% pay per view – 15,2% concorrência – 15,2% atendimento – 6,1% clube – 0 ponto extra – 0	filmes – 60,9% custo – 17,2% pay per view – 4,7% concorrente – 3,1% atendimento – 7,8% clube – 3,1% ponto extra – 3,1%	filmes – 65,4% custo – 19,2% pay per view – 7,7% concorrente – 0 atendimento – 11,5% clube – 7,7% ponto extra – 0

FIGURA 7.6 - Padrões por canal preferido

O quadro comparativo da figura 7.6 permite traçar um perfil do cliente por interesse. Uma das constatações é 15,2% dos clientes que preferem canais de esportes citaram também o conceito "*pay per view*", talvez estando mais suscetíveis a fazer aquisições deste tipo do que os demais (4,7% em filmes e 7,7% em notícias). Pode-se também notar que os clientes que citaram os canais de filmes como preferidos também citaram o conceito "*ponto extra*" (os outros não). Disto pode-se inferir que estes clientes estão mais interessados em ter um ponto extra. Dos clientes que escolheram um canal de esporte como favorito, 15,2% citaram a concorrência (bem mais que os demais clientes). Isto levanta a hipótese de que a concorrência possa estar oferecendo algo melhor em termos de esportes. Analisando-se as regras associativas desta sub-coleção (não apresentadas neste trabalho), não foi detectado nenhum padrão associativo entre "*pay per view*" e "*concorrente*". Assim, pode-se inferir que estes 15,2% referentes aos dois conceitos não são os mesmos clientes, ou seja, quem cita um destes dois conceitos provavelmente não cita o outro (nesta sub-coleção).

Em outro experimento, os chamados para atendimento do setor de suporte de informática eram registrados como campos textuais de um banco de dados (*memos*), no sistema de *help desk* de uma empresa. Nestes registros, havia informações sobre o setor que chamou, o tipo de problema apresentado (queixa) e a solução aplicada, indicando o recurso defeituoso (hardware ou software). Um processo de descoberta foi desempenhado sobre estes registros textuais (gravados como textos separados para cada chamado). Após a mineração, foi possível saber que setores ou recursos mais apresentavam queixas. Tal

descoberta permitiu lançar ações preventivas, seja através de melhores controles sobre equipamentos de hardware ou através de treinamento do pessoal usuário de informática.

7.7 Descoberta em Documentos da Web

A Web é uma grande coleção de textos, que vem-se tornando uma fonte valiosa de informação. Garofalakis e outros [GAR99] prevêem que a maior parte do conhecimento humano estará disponível na Web em 10 anos.

Entretanto, extrair informação útil destes textos não é uma tarefa fácil. A heterogeneidade e o grande volume de documentos levam à chamada "sobrecarga de informações", que acontece quando se tem muita informação disponível mas não se pode gerenciá-la.

Com base nos experimentos realizados, pode-se dizer a abordagem proposta minimiza o problema da sobrecarga, através da descoberta de conhecimento útil e novo em textos extraídos da Web. No artigo 6 dos anexos, é descrito em detalhes um experimento com textos de jornal, que tratavam de um determinado político, publicados eletronicamente na Web. Os resultados mostraram que uma pessoa poderia inferir acontecimentos mesmo sem ter conhecimentos prévios sobre o assunto. Por outro lado, alguém com certo conhecimento sobre o assunto poderia descobrir algo de novo, sem precisar ler todos os artigos.

No mesmo artigo, é descrito outro experimento onde a abordagem foi utilizada para avaliar ferramentas através da análise de páginas Web que as descreviam.

Uma limitação da aplicação da abordagem em textos da Web é a definição dos conceitos. Uma vez que a Web é caótica e dinâmica [ETZ96], a linguagem utilizada neste contexto tende a ser muito variada. Isto dificulta cobrir todas as possibilidades de expressão de um conceito. Entretanto, pode-se imaginar que as definições dos conceitos podem evoluir com o tempo, à medida que se conhece melhor cada estilo utilizado.

Apesar destas limitações, o conhecimento descoberto pode ser útil para um propósito determinado. Pressupondo que as informações disponíveis na Web não tem rigor científico (nem devem ser usadas desta forma), pode-se analisar os resultados como tendências. Neste caso, McCarthy [MCC00] defende o uso de conceitos aproximados (quando há falta de precisão nas definições), afirmando que mesmo assim é possível raciocinar sobre os conceitos. O único cuidado é interpretar os resultados sob este ponto de vista.

A abordagem é adequada para descobertas interativas na Web, porque o usuário não precisa despender muito tempo para definir, identificar e minerar conceitos. Além disto, as definições podem ser refinadas durante o processo de descoberta, com pouco conhecimento sobre o domínio, sem ser necessário utilizar modelos formais (como *thesauri* e ontologias). Isto é particularmente útil em descobertas proativas, quando o usuário não tem hipóteses iniciais ou quando não tem idéia de como é a solução que está procurando (descoberta exploratória).

7.8 Outras Aplicações Possíveis

As possibilidades de aplicação da abordagem baseada em conceitos são muitas. A princípio, qualquer coleção de documentos textuais pode ser examinada para se descobrir conhecimento novo e útil. Mas não só isto; bases de dados que fazem uso de campos textuais, poderão criar rotinas específicas para examinar este tipo de informação, que atualmente deve ser analisada manualmente por pessoas.

Os benefícios tendem a ser maiores com o aumento de documentos eletrônicos, sejam criados como imagens de documentos em papel ou diretamente em computadores. Também, à medida que as informações vão sendo colocadas na Web, a demanda por processos de *KDT* tende a aumentar muito.

Em especial, algumas áreas estão registrando suas informações em meios eletrônicos por recomendações de órgãos superiores. Este é o caso de hospitais, que deverão armazenar os prontuários médicos sobre pacientes em meios eletrônicos. A recomendação não se dá apenas para fins de padronização, mas principalmente para que as informações possam servir melhor aos processos de tomada de decisão, sejam administrativos ou dos profissionais especializados. Assim, no caso de hospitais, as administrações poderão prever melhor a demanda e então planejar melhor seus recursos (materiais, médicos, funcionários, leitos, remédios, etc). Da mesma forma, os médicos poderão tomar melhores decisões sobre diagnósticos e tratamentos. Também poderão ser feitos com maior rapidez e precisão estudos epidemiológicos, considerando conceitos como região de origem, condições de habitação, categorias de idade, sintomas, atividades profissionais e sociais, permitindo que as doenças e os doentes possam ser tratados de forma mais ampla e contextual.

Nas empresas, as aplicações de *KDT* sobre conceitos também trarão muitos benefícios. Setores de produção poderão examinar melhor problemas com máquinas ou processos, relacionando causas, conseqüências e soluções aplicadas.

Em áreas científicas, as aplicações também são muitas. Em geral, os pesquisadores quando procurando informações em bibliotecas digitais fazem uso apenas de sistemas de recuperação de informação. Após terem reunidos os textos possivelmente relevantes, as pessoas têm de ler os textos ou examiná-los de forma manual para encontrar o que realmente precisam. Num futuro próximo, será possível descobrir automaticamente relações escondidas entre textos, sem que seja necessário selecionar os textos antes, como fazem Swanson e Smalheiser [SWA97b]. Estes autores utilizam ferramentas de software para analisar títulos de artigos da base de dados Medline e encontrar possíveis relações entre trabalhos médicos. O trabalho dos referidos autores têm demonstrado a viabilidade desta estratégia, uma vez que suas descobertas estão sendo apresentadas em revistas e conferências da área médica e consideradas novas e úteis pela comunidade específica (as relações descobertas não haviam sido consideradas como hipóteses por especialistas humanos). Entretanto, as ferramentas utilizadas analisam os textos somente no nível de palavras. Acredita-se que a estratégia poderia trazer maiores benefícios se feita sobre conceitos.

Já na área de psicologia, podem ser descobertas relações entre comportamentos, sentimentos, intenções ou pensamentos de pessoas, descritos como conceitos. Da mesma forma, a área de sociologia pode investigar ideologias e idéias de pessoas ou entidades. Em especial, os processos de análise de discurso poderão examinar conceitos presentes em textos, avaliando temas centrais versus periféricos, além de estruturas de discursos. Já os

setores pedagógicos de escolas poderão avaliar melhor os motivos de evasão ou baixo rendimento dos alunos.

8 CONSIDERAÇÕES FINAIS

A figura 8.1 relaciona os objetivos propostos no início desta tese com os resultados alcançados e a produção científica correspondente (detalhes da produção científica nos anexos).

Esta proposta de tese apresentou uma abordagem baseada em conceitos para realizar descoberta de conhecimento em textos (*KDT*). Ao invés de aplicar técnicas de mineração sobre palavras extraídas de ou associadas aos textos ou sobre valores de atributos em bancos de dados, a abordagem explora conceitos identificados nos textos. Conceitos identificam melhor o conteúdo dos textos e servem melhor que palavras para representar os fenômenos do mundo real (eventos, pessoas, entidades, pensamentos, características, etc.). Por exemplo, no caso dos textos de psiquiatria, os conceitos permitiram investigar características importantes dos pacientes, tais como sintomas, sinais e comportamentos.

Isto permite explorar o conhecimento disponível em textos num nível mais próximo da realidade (como sugerido por Tan [TAN99]) e ao mesmo tempo minimizar o problema do vocabulário. Também minimiza o esforço humano para codificar e depois identificar informações em textos, uma vez que as pessoas podem gerar textos livres, sem preocupações com formatos ou estilos de linguagem, textos estes que depois serão analisados com auxílio de ferramentas automatizadas.

Este trabalho definiu um processo padrão para *KDT* a partir de avaliações de diferentes métodos para definição e identificação de conceitos. Diferentes aplicações da abordagem foram experimentadas, e o conhecimento descoberto foi avaliado subjetivamente por especialistas ou de forma objetiva embutido em sistemas automáticos baseados em conhecimento.

Os experimentos e avaliações demonstraram que a abordagem é viável e apresenta vantagens sobre os métodos baseados em palavras.

A principal aplicação da abordagem é permitir análises qualitativa e quantitativa de coleções textuais. Conceitos podem ser identificados nos textos e suas distribuições e relações podem ser analisadas para um melhor entendimento do conteúdo de textos individuais, da coleção toda ou de partes desta.

Conseqüentemente, o próprio domínio de aplicação pode ser melhor entendido, bem como o conhecimento e o tipo de raciocínio utilizado pelas pessoas do domínio. Pôde-se notar, nos experimentos, que o modo como as pessoas se expressam através da linguagem escrita pode revelar conhecimento novo e útil, muitas vezes implícito nos textos e em outras não percebido pelas próprias pessoas que utilizam o conhecimento.

Este novo tipo de descoberta sugere a possibilidade de serem analisadas idéias, ideologias, tendências e intenções presentes em textos.

FIGURA 8.1 - Objetivos alcançados e resultados

Objetivo	Resultados Alcançados	Produção Científica
1) avaliar alternativas para definição de conceitos	- identificação de fatores que influenciam o processo - identificação de estratégias que levam a melhores desempenhos	Artigo a elaborar
2) avaliar métodos de categorização (identificação de conceitos)	- identificação de situações onde utilizar os tipos de métodos existentes	Artigo a elaborar
3) definir um processo padrão de identificação de conceitos	- processo padrão para identificação de conceitos nos textos	Artigo 2 – Applied Intelligence Artigo 6 – SIGKDD
4) definir um processo padrão de mineração sobre conceitos	- processo padrão para mineração sobre conceitos	Artigo 2 – Applied Intelligence Artigo 6 – SIGKDD
5) realizar um processo de *KDT* com o processo padrão definido	- conhecimento resultante do processo	Artigo 6 – SIGKDD Artigo 3 - J. Documentation Artigo 4 - ISKMDM 2000 Artigo 1 - ISKMDM 2001
6) avaliar grau de acerto na identificação de conceitos	- margem de erro pelas medidas definidas	Artigo 2 – Applied Intelligence Artigo 3 - J. Documentation
7a) avaliar subjetivamente a qualidade do conhecimento descoberto	- parecer de especialistas sobre conhecimento descoberto	Artigo 3 - J. Documentation
7b) avaliar objetivamente a qualidade do conhecimento descoberto	- nível de acerto do sistema automático	Artigo 2 – Applied Intelligence
8) comparar métodos baseados em palavras com métodos baseados em conceitos	- comparação entre regras de raciocínio - graus de acerto dos sistemas construídos	Artigo 2 – Applied Intelligence Artigo a elaborar sobre categorização (seleção de características)
9) avaliar a abordagem baseada em conceitos com outras técnicas de mineração	- comparação dos resultados da técnica de mineração sobre conceitos X sobre palavras - conhecimento descoberto com a nova técnica de mineração	Artigo a elaborar
10) avaliar possibilidade de descoberta proativa	- estratégia para descoberta proativa - estudo da necessidade de intervenção humana e conhecimentos prévios	Artigo 5 – OIA 2000 Artigo 4 - ISKMDM 2000
11) avaliar aplicações da abordagem	- aplicações da abordagem - benefícios práticos	Artigo 6 – SIGKDD Artigo 3 - J. Documentation Artigo 4 - ISKMDM 2000 Artigo 1 - ISKMDM 2001

Os problemas estudados nesta tese se assemelham aos desafios propostos para a área de *Text Mining* por Tan [TAN99], a saber:

- análise semântica para obter representações ricas que capturem relacionamentos entre conceitos ou objetos do domínio: os experimentos e avaliações permitem concluir que este desafio foi resolvido neste trabalho;
- análises multilinguais: apesar de a abordagem não ter sido avaliada em ambientes multilinguais, algumas aplicações foram feitas com textos em outras línguas (inglês e espanho); acredita-se ser possível aplicar a abordagem em textos escritos em várias línguas, uma vez que o processo empregado permite identificar o mesmo conceito em linguagens diferentes, bastando acrescentar as regras necessários nas definições dos conceitos; por exemplo, o conceito "alcoolismo" poderia ser identificado em português através da regra "*álcool –nega*" ou em inglês por "*alcohol –denies*";
- uso de conhecimento do domínio na mineração: isto é possível uma vez que os conceitos representam fenômenos da realidade e não documentos ou os textos em si; o caso médico exemplifica bem que este desafio foi atacado, já que o conhecimento do especialista do domínio pôde ser adquirido pela análise dos textos;
- mineração autônoma e personalizada, para leigos, de forma quase automática: os estudos sobre a descoberta proativa revelam que é possível minimizar a intervenção humana, principalmente depois que os conceitos estão definidos; os experimentos demonstraram que conhecimento interessante pode ser descoberto de forma exploratória, sem que o usuário precise identificar hipóteses ou definir objetivos no início do processo; entretanto, pôde-se observar que a intervenção humana é importante e imprescindível para interpretar os resultados.

8.1 Contribuições

Esta tese demonstrou ser viável analisar o conteúdo de coleções textuais, através da identificação e mineração de conceitos.

Também foi demonstrado que o conhecimento tácito, disponível com pessoas, pode ser formalizado e explorado através da aplicação da abordagem proposta, se o conhecimento puder ser capturado em textos livres.

Os experimentos e avaliações demonstraram a viabilidade de adquirir conhecimento especialista através da análise de documentos textuais. Isto implica em menos esforço e tempo nos processos de aquisição de conhecimento e num melhor uso do grande volume de textos, os quais contêm muito conhecimento que em geral não pode ser aproveitado por falta de ferramentas de análise.

O conhecimento descoberto pela abordagem proposta pode ser facilmente incorporado (quase que automaticamente) em sistemas automatizados. Os experimentos demonstraram ser viável construir sistemas baseados em conhecimento a partir dos resultados encontrados pelo processo padrão de *KDT*. O desempenho satisfatório dos sistemas construídos (acima de 60% na média de acertos) leva a crer que o conhecimento descoberto é consistente com a realidade e, portanto, o processo padrão possui qualidade.

Além disto, os experimentos permitiram ainda investigar como o conhecimento descoberto pode ser embutido nestes sistemas automáticos. Pôde-se notar que é importante utilizar pares de conceitos e evidências negativas. Mesmo sem prévia seleção ou organização,

o conhecimento resultante da descoberta também pode ser utilizado. O resultado numérico de 44% na média de acertos, utilizando conceitos simples e as distribuições originais do resultado, pode ser considerado regular numa área tão complexa como a psiquiatria. Em áreas que utilizem conhecimento mais estruturado, talvez os resultados possam ser melhores. Ficou claro, nos experimentos, que desempenhos melhores podem ser obtidos com métodos melhores de classificação e com a adequada seleção e organização do conhecimento descoberto.

Já que é possível descobrir conhecimento com qualidade, também se pode aproveitar os resultados da descoberta para treinamento de pessoas ou para avaliar resultados de decisões humanas.

Quanto a aplicações, pode-se dizer que qualquer domínio que faça uso de textos pode-se beneficiar com a abordagem proposta. Os experimentos evidenciaram benefícios em nas áreas de inteligência (competitiva e de negócios), de classificação e recuperação de documentos por conteúdo e em áreas de alta especialidade e complexidade, como o domínio de psiquiatria.

Pode-se concluir que a abordagem é adequada para casos onde há conhecimento disponível de forma não-estruturada em grandes volumes de textos (que dificilmente poderiam ser analisados sem auxílios automatizados). Este é o caso da Web, a qual se torna candidata natural para aplicação da abordagem.

8.2 Vantagens da Proposta

Uma das principais vantagens da abordagem proposta é minimizar o problema do vocabulário, já que os conceitos podem ser expressos (nos textos) e definidos (na classificação) com diferentes palavras, como num processo de expansão semântica.

Na comparação entre as abordagens baseada em palavras e baseada em conceitos, não houve vantagem de uma sobre outra pelos resultados numéricos obtidos com processos de classificação. Entretanto, os conceitos permitem entender e explicar melhor o raciocínio de classificação que as palavras (como discutido na seção 6.6.3). Já na comparação das abordagens pelo processo de agrupamento, a abordagem baseada em conceitos proporcionou melhores resultados (grupos mais puros).

Outra vantagem é que a simplicidade da abordagem permite a definição e a identificação de conceitos sem que seja necessário gastar muito tempo ou esforço. Feldman e Dagan [FEL95] defendem o uso de estruturas simples porque os processos podem ser realizados com baixo custo e com ajuda de ferramentas de software. Isto permite realizar a descoberta mesmo sem ter conceitos previamente definidos (os conceitos podem ser definidos e refinados durante o processo – ver artigo 1 dos anexos). É bom salientar que conceitos pré-definidos como em *thesauri* e ontologias facilitam o processo, mas a vantagem da abordagem é que não se fica preso aos conceitos e definições existentes. Isto quer dizer que descobertas podem ser feitas em novas áreas, bastando definir novos conceitos.

A facilidade na definição de conceitos também permite que o processo seja especializado a uma comunidade. Ou seja, os conceitos gerais de um domínio de conhecimento (como a psiquiatria, por exemplo) podem ser refinados para incorporar as variações e especializações da linguagem utilizada por um grupo específico (por exemplo, médicos de uma clínica psiquiátrica). Esta facilidade também permite incorporar às definições

dos conceitos erros ortográficos comuns na comunidade, como foi feito nos experimentos com os textos psiquiátricos.

Cabe salientar que a abordagem baseada em conceitos pode ser realizada com outros métodos de definição, identificação e mineração de conceitos. Isto ficou evidente nos testes feitos com diferentes métodos de definição e identificação e com a técnica de agrupamento para mineração. Esta flexibilidade permite melhorar os métodos sem perder o fundamento que é utilizar conceitos para entender o conteúdo dos textos e representar fenômenos do mundo real.

8.3 Limitações e Cuidados no Uso da Abordagem

Os sistemas automáticos criados com o conhecimento descoberto com a abordagem obtiveram resultados que foram considerados satisfatórios pelos especialistas do domínio. Entretanto, este nível de acerto (mais de 60%) pode ser insuficiente para alguns propósitos.

Analisando os experimentos do caso médico, conclui-se que os resultados não foram melhores devido a circunstâncias extrínsecas aos métodos. Primeiro, os textos utilizados correspondiam a prontuários de internação, enquanto que o diagnóstico associado poderia ter sido decidido com informações adicionais. Ou seja, o diagnóstico associado ao texto era o final (o da alta do paciente), mas os sistemas automáticos deveriam descobrir o diagnóstico somente sobre as informações disponíveis neste prontuário de internação. Além disto, é comum que os médicos discordem entre si sobre alguns diagnósticos e, em muitos casos, no próprio prontuário havia a informação de que o diagnóstico deveria ser melhor avaliado durante o período de internação do paciente.

Outro fator que pode ter influenciado nos resultados é que os textos usados para descoberta de conhecimento (treino) não foram selecionados por critérios de qualidade, mas somente por período de tempo. Isto poderia introduzir algum tipo de ruído no processo. Acredita-se que, com uma seleção melhor destes textos, pode-se descobrir conhecimento mais consistente e assim melhorar os resultados. Além disto, acredita-se ser possível melhorar o processo utilizando um número maior de casos de treino (um trabalho futuro deverá avaliar esta hipótese).

Uma maneira de melhorar a construção dos sistemas automáticos é através da intervenção humana para refinar o conhecimento descoberto antes de este ser embutido no sistema. Nos experimentos feitos, o conhecimento descoberto, apesar de ter sido validado subjetivamente por especialistas (e portanto ser considerado verdadeiro), não sofreu nenhum tipo de refinamento para ser incorporado ao sistema automático. Isto quer dizer que ruídos podem ter passado despercebidos. Somente foram utilizados mecanismos simples de seleção tais como “conceitos menos freqüentes” e “conceitos exclusivos”, os quais não podem ser considerados intervenção de especialistas. Acredita-se que é possível melhorar a qualidade dos resultados refinando o conhecimento descoberto antes de ser incorporado à base de conhecimento. Por outro lado, a aplicação quase que automática do conhecimento descoberto, mesmo que com um desempenho pior, pode reduzir o tempo de construção do sistema (se o nível de acerto for satisfatório para o propósito, esta é uma vantagem).

Um fator que limita seriamente a aplicação da abordagem proposta são os erros ortográficos. Nos experimentos com prontuários, havia muitos destes erros nos textos (lembrando que os textos não sofreram alterações para a descoberta). Alguns puderam ser

corrigidos na definição dos conceitos, acrescentando-se as variações léxicas mais comuns. Contudo, outros erros não foram detectados e certamente diminuíram os níveis de precisão e abrangência na identificação dos conceitos. Apesar disto, a média de 90% de acertos, averiguada na identificação de conceitos, confirma a qualidade do conhecimento descoberto.

Com relação à margem de erro na identificação dos conceitos, cabe salientar que este é um ponto crucial no uso da abordagem. Se esta margem for grande, os resultados podem não ser confiáveis. Será difícil conseguir um processo sem erros, mas estes níveis podem ser controlados, para que o conhecimento descoberto possa ser interpretado corretamente.

A propósito, a interpretação dos resultados deve estar condicionada à definição dos conceitos. Por exemplo, no caso médico, o conceito "*alcoolismo*" podia aparecer no texto associado ao paciente ou a alguém da família. As distribuições e as regras associativas relativas a este conceito devem ser interpretadas sob este contexto. É possível eliminar tais ambigüidades refinando as regras para identificação deste conceito.

Outro cuidado que se deve ter é com a evolução da linguagem. Não se pode esperar que as mesmas definições possam ser usadas com o mesmo grau de acerto, em domínios semelhantes mas com pessoas diferentes. Da mesma forma, como a linguagem muda com o passar do tempo [CHE94], as definições precisarão ser atualizadas para incorporar novos termos, sinônimos ou variações.

Durante as aplicações feitas, notou-se outro cuidado a ser tomado: a representatividade da coleção. Por exemplo, no caso do experimento com textos políticos, os textos foram extraídos de um único jornal. Assim, as informações contidas nesta coleção não podem ser consideradas como um conjunto completo. Os resultados precisam ser interpretados dentro destes limites.

Além disto, neste experimento político, os textos foram publicados nos anos de 1997 e 1999. Então, os resultados do processo de descoberta estão condicionados aos eventos e fenômenos que ocorreram próximos a estes períodos. Decisões ou ações tomadas durante o ano de 1998 certamente influenciaram os eventos descritos na sub-coleção de 1999. Desta forma, a análise conjunta das duas sub-coleções (sem separá-las) poderia levar a hipóteses distorcidas.

Além disto tudo, há ainda o problema da confiabilidade das fontes de informação. No caso da Web, tal problema é ainda mais preocupante, já que os documentos mudam rapidamente [CHE93]. Alguns trabalhos sugerem estratégias para avaliação da qualidade da informação disponível na Web [SCH96] [OWE97] [SMI97].

Mesmo que as fontes geradoras dos textos sejam confiáveis, a informação pode não ser. Parsons [PAR96] sugere que problemas podem ocorrer devido a 5 tipos de imperfeições:

- informação incompleta: quando faltam detalhes de informação (por exemplo, atributos sem valores);
- informação imprecisa: devido à diferença de granularidade (por exemplo, datas sem o dia ou somente referenciando o ano);
- informação incerta: quando não pode ser provada;
- informação vaga: devido a imprecisões do vocabulário;
- informação inconsistente: por exemplo, quando há valores contraditórios.

8.4 Trabalhos Futuros

Entre os planos para o futuro está a aplicação da abordagem em uma coleção multilingual, isto é, com textos em línguas diferentes. A hipótese é de que a abordagem pode ser usada nestas situações.

Também pretende-se utilizar uma técnica diferente para mineração. Nos experimentos, foram usadas as técnicas de análise de distribuição, associação e agrupamento. Entretanto, acredita-se que outras técnicas podem ser empregadas sobre os conceitos identificados nos textos. Por exemplo, a técnica de resumos e a técnica de correlações escondidas, sugeridas em [CHE93], permitem identificar partes de textos que tratam do mesmo assunto. Em [SWA97b], estas técnicas foram utilizadas com êxito. Entretanto, as técnicas analisam o conteúdo somente a nível de palavras (termos comuns aos textos), o que pode deixar de fora importantes relações semânticas. A sugestão é utilizar estas técnicas sobre conceitos identificados nos textos (nível meta-lingüístico) e não sobre os termos presentes nos textos. Isto pode revelar associações entre fenômenos da realidade, tal como é feito em analogias.

Uma das técnicas que está planejada para ser avaliada é a seqüência de tempo. Nos experimentos realizados, os textos formam um único conjunto sem tempo associado. Nos próximos experimentos, a técnica de seqüência de tempo será usada para investigar a dependência entre conceitos ao longo do tempo, isto é, se um conceito condiciona a aparição de outro no passado ou no futuro. No caso de prontuários médicos, isto significa examinar a relação de conceitos durante a evolução do paciente, por exemplo, o surgimento de certas características após alguma medicação ou algum tipo de tratamento.

Também está prevista a avaliação de técnicas de mineração que combinem conceitos com outros tipos de informação. Por exemplo, pode-se explorar a estrutura do documento (exemplo: se um conceito aparece no título, outro conceito pode estar condicionado a aparecer na conclusão). Ou então utilizar informações mais específicas extraídas de bancos de dados ou por técnicas de extração de informação (exemplo: se idade é tal ou tempo de internação é de tanto, certo conceito sempre aparece)

Anexos

Anexo 1: Exemplos de Prontuários

Foram omitidos dos textos nomes de pessoas e lugares.

Prontuário 54546_F32

ATENDIMENTO INICIAL - INTERNACAO
A= (Anamnese):

Impressão sobre a paciente: trata-se de uma mulher com idade aparente superior a real, em regular estado de higiene, tem os cabelos curtos, presos e pintados de cobre.

Motivo da internação: Pcte com tentativa de suicidio, por ingestão de diazepan, insone, agressiva com companheiro.

Pcte é trazida ao plantão pela polícia, por ter ingerido vários comprimidos de diazepan, vem acompanhada do companheiro.

22:30h
Entrvista com a Pcte: Ficou chorando o tempo todo, diz estar sofrendo demais e não querer viver mais, conta que tomou uma caixa de diazepan e alguns comprimidos de ampicilina. Esta com as mãos e os pés cianóticos e gelados, apresenta tremores e calafrios. Esta com a TA 90/60 mmHg, encaminhamos ao PS.

01:30 Pcte retorna do PS após realizar lavagem gastrica. Encontra-se melhor.
Entrevista com a pcte: Conta que esta morando com o atual companheiro ha 3 meses, e este a iludiu, pois a tirou da casa dos pais e agora não quer mai viver com ela. Diz estar consanda de cuidar as pessoas e ninguém cuidar dela. Relata que apanhava do pai em casa e quando saiu de casa o pai falou que não poderia voltar, esta com medo pois não tem para onde ir, se terminar o relacionamento.
Pcte tem dois filhos menino de 6 anos e menina de 3 anos, sendo cada um dos filhos de pai diferente, diz nunca ter sorte com seus relacionamentos.
Conta que vinha fazendo uso de fluoxetina, mas no momento esta sem condições de comprar. Relata tentativa de suicidio anterior, ha 8 anos quando o pai do primeiro filho a abandonou. Conta que foi pior, pois tomou muitos remédios, inclusive veneno, ficando internada no hospital clinico por 4 dias.
Conta que um primo se matou ha + ou - 2 anos, e isso a marcou muito, pois gostava muito dele.

Entrevista com o companheiro: Relata que esta com a paciente ha 3 meses, é separado da 1º esposa, tem 40 anos e começaram a viver juntos achando que iam se dar bem. Diz que a pcte é muito imprevissivel, por vezes agressiva, sendo muito dificil viver com ela. Hoje falou que desejava terminar o relacionamento, e a pcte se trancou no banheiro e quando ele arronbou a porta ela estava tomando os comprimidos, não sabe precisar o número que a pcte tomou.
Relata que a pcte é muito agressiva com os filhos, que estão morando com os avós. O companheiro tira do bolso um bilhete que a pcte escreveu para ele se despedindo e o avisando que ia se matar, encontrou em cima da mesa da cozinha.
Conta que o primo da pcte matou a mulher com um tiro e depois se matou ha 2 anos atraz.

EP= (Exame Psiquico):
C - lúcida
A - hipovigil e hipertenaz
S - nada relatado
O - parcialmente orientada auto e alo
M - não avaliada
I - não avaliada
A - deprimido e ansioso
P - mágico, com idéias delirantes de desvalia, e ideação suicida
C - tentativa de suicidio, insone, inapetente, agressiva com familiares, impulsiva
L - normolalia

EF= (Exame Fisico):
BEG, hidratada, corada
Ta : 140/100 mmHg
Fc 80 bpm Fr 20 mrpm
Demais sp
Pcte tem historia de angina, após o segundo parto, familiares ficaram de trazer exames.

Pcte foi atendida no Ps pelo médico XYZ, que atestou que a pcte se encontrava estável.

PMI= (Presc. Medica): Haldol 5mg/dia
Nitrazepan 1 cp vo

PTI= (Plano Terapeutico): ACMA
ADTP= (Avaliacao de Tratamento e Prognostico): ACMA
AMF= (Atend. Familiar): ACMA

Prontuário 54810_F20.0_L

ATENDIMENTO INICIAL - INTERNACAO

A= (Anamnese): Paciente comparece acompanhado pelo Motorista da Prefeitura de sua cidade Sr. XYZ, com AIH assinada pelo Sec. Saúde de XYZ Sr. WKL.

Paciente encontra-se em péssimo estado de higiene, mau cheiroso, com estado negativista, não respondendo ao questionamento, mas aceitando certas indicações, como caminhar, colaborar no exame físico. Responde com muito custo seu nome. Sempre idagado diz que morava com os pais, mas estes não o querem mai. Alega que o "correram" de casa. Diz que o pai trabalha na roça, e ele também trabalhava. Fala que estava dormindo na "parage" no chão. Responde que por vêzes bebe. Não consegue estabelecer um diálogo continuo.

O acompanhante, diz que o paciente é conhecido de sua cidade, sendo que há muito tempo paciente é levado ao Hospital de Caridade de lá onde fica, é medicado e permanece sempre assim, em silêncio, negativista, aceita a alimentação, cuidados. Ao ser dado alta paciente passa andando na rua, caminha sem rumo, dorme na rua, sendo que passa a jogar pedras nas casas e e provocar pessoas e falar sozinho. Como estava novamente nestas condições resolveram trazê-lo.
EP= (Exame Psiquico): hipoprosexia, alucinações auditivas, pensamento mágico, curso bloqueado, ideação paranóide, anfeto indiferente, autismo, jogando pedras nas casa, negativista, andando sem rumo.
EF= (Exame Fisico):
TA140/80 FC80 AC RR 2T AP/MV DIMINUIDO. T 37,2
Edema de pés e MMII +/6, com lesòes crostosas.
PMI= (Presc. Medica): Haldol/Fenergan 01 amp IM na internação.
PTI= (Plano Terapeutico): Controle de níveis de sedação. Controlar Temperatura 3 vezes ao dia e anotar. Pen-v-Oral 01 comp VO 4 vezes ao dia.
Solicitado Revisão clínica.
Risco de Agitação e Cuidados comobjetso de risco.
ADTP= (Avaliacao de Tratamento e Prognostico): Observar evolução.

Anexo 2: Produção Científica

- Artigo 1 (aceito)

"Formalizando e explorando conhecimento tácito com a tecnologia de text mining para inteligência"
Co-autores: Eliseo Reategui, Leandro Krug Wives, José Palazzo M. de Oliveira, Maurício Almeida Gameiro
International Symposium on Knowledge Management / Document Management
PUC-PR - Curitiba, Brasil, 13-15 de Agosto de 2001.

- Artigo 2 (sendo impresso)

"Knowledge discovery in texts for constructing decision support systems"
Co-autores: José Palazzo M. de Oliveira, Maurício Almeida Gameiro
Journal of Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies
Special Issue on Text and Web Mining
Editores convidados: Ah-Hwee Tan e Philip S. Yu
http://textmining.krdl.org.sg/APIN/TWMcfp.html

- Artigo 3 (sendo impresso)

"Knowledge discovery in textual documentation: qualitative and quantitative analyses".
Co-autores: Jose Palazzo M. de Oliveira, Fábio Leite Gastal
Journal of Documentation, v.57, n.5, September 2001. pp.577-590.
Editor: R. T. Kimber
Publicado pela Aslib, The Association for Information Management (London) www.aslib.co.uk/jdoc

- Artigo 4

"Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva"
Co-autores: Leandro Krug Wives, José Palazzo M. de Oliveira
International Symposium on Knowledge Management / Document Management
Proceedings, pp.125-147
PUC-PR, Editora Universitária Champagnat
Eds: Edouard Lethelier, Flávio Bortolozzi, Kival Chaves Weber, Heitor Pereira
Curitiba, Brasil, 26-29 de Novembro de 2000

- Artigo 5

"Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses"
Co-autores: Leandro Krug Wives, José Palazzo M. de Oliveira
IV Oficina de Inteligência Artificial, pp.143-154
Universidade Católica de Pelotas.
Pelotas-RS, 25 de Agosto de 2000.
Editora EDUCAT. Organizador: Luiz Antônio Moro Palazzo

- Artigo 6

"Concept-based knowledge discovery in texts extracted from the web"
Co-autores: Leandro Krug Wives, José Palazzo Moreira de Oliveira
SIGKDD Explorations, v.2, n.1, July 2000, pp.29-39
Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. ACM Press.
Editor-Chefe: Usama Fayyad
Editor Assistente: Sunita Sarawagi
Editor Convidado: Paul Bradley
Disponível por WWW em http://www.acm.org/sigkdd/explorations

Anexo 3: Artigos completos

Artigo 1 – ISKMDM 2001

"Formalizando e explorando conhecimento tácito com a tecnologia de text mining para inteligência"
Co-autores: Eliseo Reategui, Leandro Krug Wives, José Palazzo M. de Oliveira, Maurício Almeida Gameiro
International Symposium on Knowledge Management / Document Management
PUC-PR - Curitiba, Brasil, 13-15 de Agosto de 2001.

FORMALIZANDO E EXPLORANDO CONHECIMENTO TÁCITO COM A TECNOLOGIA DE TEXT MINING PARA INTELIGÊNCIA

Stanley Loh[1,2], Eliseo Reategui[3], Leandro Krug Wives[1],
José Palazzo M. de Oliveira[1], Maurício Almeida Gameiro[4]

sloh@zaz.com.br
eliseo@godigital.com.br
wives@inf.ufrgs.br
palazzo@inf.ufrgs.br
magameiro@uol.com.br

1- Programa de Pós-Graduação em Computação (PPGC)
Instituto de Informática
Universidade Federal do Rio Grande do Sul (UFRGS)

2- Universidade Católica de Pelotas (UCPEL) e
Universidade Luterana do Brasil (ULBRA)

3- GoDigital
Marketing de Precisão

4- Clínica Psiquiátrica Olivé Leite
Pelotas, RS

Abstract

This work presents an approach for exploring knowledge available with people, using Text Mining technology. The knowledge may come from internal collaborators or from customers. To make the knowledge concrete in electronic ways, the approach acquires information through textual documents. Text Mining tools are then used to extract concepts present in the texts. Concepts represent real world events, people, objects, etc., and they help to understand what themes or subjects are referenced by the texts. After the extraction step, data mining tools are used to discover new knowledge through the analysis of concept distributions and relations. The approach is useful to obtain intelligence about the organization, in order to improve products, services, internal processes and the relationship with customers.

Two applications of the approach are discussed in this paper: one for exploring knowledge from customers of a cable television company and other to capture the knowledge used by physicians of a psychiatric hospital. In the first case, information was captured as suggestions of the customers. Results from the Text Mining approach enabled the understanding of how customers saw the offered services and products. In the second application, the approach was applied over medical records about patients, written by physicians. The discovered knowledge was useful to analyze patients' characteristics and to understand how the diagnosis process is made.

Keywords: business intelligence, text mining, tacit knowledge, knowledge discovery

1 INTRODUÇÃO

Hoje em dia, os clientes costumam selecionar produtos e serviços analisando a competência das organizações e procurando por algum diferencial. Logo, para que as organizações ganhem e mantenham seus clientes elas precisam conhecer cada um de seus clientes, identificando seus desejos e anseios. Além disso, as organizações necessitam saber se elas podem oferecer esse diferencial e, caso negativo, como elas poderiam fazê-lo. Isso significa que as organizações precisam de conhecimento sobre seus clientes e sobre os meios e processos capazes de atender às necessidades do cliente.

Este conhecimento pode estar em diferentes formas e provir de diferentes fontes. Uma destas fontes são as próprias pessoas ligadas à empresa. Por ser um dos recursos mais importantes, o conhecimento proveniente das pessoas passou a ser chamado de capital intelectual (Stewart, 1998). Rodrigues e outros (2000, p.72) propõem um modelo onde as pessoas são o foco das atenções da empresa. Neste caso, as pessoas são a fonte principal de conhecimento para que a organização atinja a competência e o diferencial desejados.

Este conhecimento pode estar disponível internamente com funcionários e colaboradores, mas também pode ser obtido dos clientes. Um modelo que começa a se popularizar é o de organizações centradas no cliente (Imhoff et al., 2001). Os clientes são fonte de conhecimento e inovação para a organização (Pereira e Angeloni, 2000). Muitas vezes, as organizações esquecem que existe mais de uma via na interação com clientes. As empresas fornecem produtos e serviços para satisfazer as necessidades dos clientes e esquecem que estes também têm algo para oferecer em troca, além do pagamento. Por exemplo, pode-se obter conhecimento sobre a área de atuação da empresa, sobre seus produtos ou sobre características dos concorrentes e do mercado..

Este conhecimento todo, seja de pessoas da própria organização ou de clientes, pode ser usado para entender e melhorar a organização, gerando o que se chama de Inteligência. O objetivo final da inteligência é criar diferencial e competência para a organização (inteligência do negócio - *Business Intelligence* – ou inteligência competitiva – *Competitive Intelligence*).

O conhecimento pode ser classificado em tácito ou explícito (Nonaka e Takeuchi, 1997). O primeiro é aquele conhecimento que não está formalizado. Em geral, este tipo de conhecimento encontra-se com as pessoas e não foi ou não pode ser transformado para representações rigorosas. Já o segundo tipo de conhecimento é justamente aquele que foi formalizado em documentos, bancos de dados, gráficos, desenhos, etc.

Em uma organização, os dois tipos de conhecimento devem coexistir harmonicamente e de alguma forma interagir para que todo o potencial de utilização possa ser aproveitado. Nonaka e Takeuchi (1997) identificaram 4 modos de conversão e interação entre os conhecimentos tácito e explícito. O processo de **externalização** é a transformação do conhecimento tácito em explícito. A **internalização** é o processo inverso. Já a **combinação** é o processo de interação entre conhecimentos explícitos para geração de novos conhecimentos. Por sua vez, a **socialização** é a interação entre conhecimentos tácitos.

A tecnologia da informação constitui-se num apoio importante para armazenar e explorar conhecimentos explícitos. Este trabalho apresenta uma abordagem para externalizar e explorar conhecimentos tácitos, disponíveis com clientes ou com pessoas internas à organização. O objetivo é gerar inteligência a partir da análise das informações capturadas e documentadas em textos livres. Para tanto, será utilizada a tecnologia de Text Mining (técnicas e ferramentas). A abordagem apoia os processos de:

- **formalização** de conhecimentos, transformando conhecimentos tácitos em explícitos (externalização); e
- **exploração** do conhecimento formalizado, analisando e integrando conhecimentos explícitos (combinação).

A etapa de **formalização** faz uma análise das informações contidas em textos livres. A tecnologia de Text Mining serve então para identificar os conceitos presentes nos textos. Conceitos representam "entes" do mundo real e permitem entender que temas estão presentes nos textos ou do que tratam os textos.

Em seguida, a **exploração** é feita através de um processo automático de mineração. Nesta etapa, são aplicadas ferramentas de data mining, não sobre dados estruturados de bancos de dados, mas sobre os conceitos extraídos dos textos livres, na etapa anterior. Esta mineração é feita analisando-se a distribuição dos conceitos em coleções (a freqüência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para descobrir associações e dependências.

Assim, o processo de formalização e exploração permite gerar novos conhecimentos (inteligência), com o objetivo de melhorar processos internos, serviços, produtos e relacionamento com clientes.

Duas aplicações da abordagem são apresentadas neste artigo. Uma para formalizar e explorar conhecimento de clientes adquiridos através de pesquisas. Neste caso, a pesquisa coletou sugestões e reclamações (textos livres) de clientes sobre produtos e serviços de uma empresa de TV por assinatura. Na segunda aplicação, a tecnologia foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica, com o objetivo de analisar as características de pacientes e entender o processo de diagnóstico. Neste segundo caso, foram usados prontuários (textos semi-estruturados), criados por médicos no momento da internação dos pacientes.

A seção 2 deste trabalho descreve como a tecnologia pode ser usada para gerar inteligência para a organização. Depois, na seção 3, é apresentada a tecnologia de Text Mining, de forma geral. Na seção seguinte (4), é detalhada a abordagem de Text Mining usada neste trabalho para gerar inteligência. A seção 5 apresenta as aplicações da abordagem, bem como discute as possibilidades de uso do conhecimento descoberto para melhorar a organização. Por fim, a seção 6 apresenta as conclusões do trabalho.

2 INTELIGÊNCIA E A TECNOLOGIA DE INFORMAÇÃO

O ramo da ciência que estuda e aplica métodos e técnicas de análise de informações para geração de inteligência com o intuito de oferecer vantagem competitiva a uma empresa é chamado de Inteligência do Negócio (Wanderley, 1999). Alguns autores fazem uma pequena distinção entre os processos de inteligência: Inteligência do Negócio fica responsável pela análise de dados relativos à organização e Inteligência Competitiva é mais voltada para a análise dos dados do mercado e dos concorrentes.

Para gerar inteligência, é necessário armazenar, analisar e disseminar conhecimento dentro da empresa. Pereira e Angeloni (2000) comentam estratégias para os processos de transformação e interação entre conhecimentos tácitos e explícitos. Por exemplo, o processo de externalização pode ser feito através de metáforas, modelos, analogias, conceitos e hipóteses. Já o processo de combinação deve trabalhar com conjuntos diferentes de conhecimentos explícitos e pode utilizar classificação, acréscimo e combinação.

Entretanto, tais processos não são fáceis de serem feitos, ainda mais quando o volume de informações é muito grande. Para apoiar estas atividades, pode ser utilizada a tecnologia da informação. Por exemplo, para os processos de socialização podem ser usadas tecnologias que apoiem o trabalho cooperativo entre pessoas, tais como sistemas de *groupware*, listas de discussão, fóruns na Web, Intranets, etc.

Já a internalização pode ser apoiada por Intranets (manuais, por exemplo), sistemas de busca na Web e sistemas de recuperação de informação (para encontrar documentos) e tecnologias para ensino à distância (EAD) ou treinamento auxiliado por computador (*Computer-Based Training* e *e-learning*) .

A tecnologia tem seu principal uso nos processos de combinação (explícito para explícito), já que é mais fácil trabalhar com conhecimento já formalizado. Neste caso, podem ser usados sistemas de data mining, sistemas de informações gerenciais (SIG), *executive information systems* (EIS), ferramentas OLAP e sistemas de informações geográficas (GIS).

Entretanto, uma das maiores dificuldades para o uso da tecnologia da informação é formalizar o conhecimento (externalização), ou seja, torná-lo disponível em algum meio eletrônico e em um formato que possa ser analisado. Este é o processo básico para que a tecnologia possa ser aplicada; sem ter como capturar o conhecimento, não se pode difundi-lo para outras pessoas (internalização) nem explorá-lo (combinação). Em geral, as organizações utilizam bancos de dados para formalizar o conhecimento, pois isto facilita o uso da tecnologia. Entretanto, as informações disponíveis em bancos de dados são codificadas de forma resumida e estruturada, após algum tipo de filtragem, o que certamente gera perdas. Além disto, 80% das informações de uma organização está disponível em forma textual, não estruturada (Tan, 1999).

Documentos textuais são mais fáceis de serem coletados e armazenados, pois permitem textos livres sem estruturas ou sem formatos limitadores. Isto gera uma riqueza de conhecimento maior que nos bancos de dados. Além disto, documentos textuais possuem conhecimento escondido, implícito nos textos ou em relações entre os documentos (Davies, 1989). Também com o crescente uso da Internet, o conhecimento está cada vez mais disponível em meios eletrônicos, e a forma mais utilizada são os textos. Garofalakis e outros (1999) estimam que a maior parte da informação humana estará disponível na Web em 10 anos.

Entretanto, na maioria dos casos, as pessoas e as organizações não sabem como analisar esta documentação textual para extrair informação nova e útil (combinação) e acabam desperdiçando importantes fontes de conhecimento. Para minimizar este tipo de problema, surgiu a tecnologia de Text Mining.

3 A TECNOLOGIA DE *TEXT MINING*

O meio mais simples de externalização é registrar, em textos livres, pensamentos, idéias, sentimentos e opiniões de pessoas. Nas organizações, há muito conhecimento deste tipo disponível na forma de:

- sugestões e reclamações de clientes em pesquisas, e-mails e serviços de atendimento;
- descrições de defeitos, causas e soluções aplicadas por funcionários;
- manuais, normas e procedimentos definidos como padrão;
- e-mails oriundos de listas de discussão;
- memorandos e comunicações formais, distribuídos através de meios eletrônicos; etc.

Entretanto, as organizações e as pessoas têm dificuldade para tratar adequadamente este tipo de informação por não estar estruturada. A área de Text Mining surgiu para minimizar este problema, ajudando a explorar conhecimento armazenado em meios textuais.

Tan (1999) define *Text Mining* (ou *KDT* – Descoberta de Conhecimento em Textos) como o processo de extrair padrões ou conhecimentos interessantes e não-triviais a partir de documentos textuais.

A tecnologia de Text Mining pode ser usada para formalizar e explorar conhecimento tácito. O conhecimento disponível com pessoas pode ser armazenado em textos, os quais serão analisados para se entender seu significado, ou seja, do que tratam os textos. Depois, pode-se explorar o conhecimento extraído dos textos para gerar novos conhecimentos. Também se pode combinar este conhecimento com o conhecimento explícito armazenado em bancos de dados estruturados.

Existem várias técnicas para Text Mining (Loh et al., 2000). Entretanto, por ser ainda uma área recente, as poucas ferramentas disponíveis são ainda ineficientes (Tan, 1999). Na maioria dos casos, as ferramentas apenas encontram textos que podem conter informações relevantes (ferramentas de recuperação de informação), deixando para os usuários a difícil tarefa de encontrar o conhecimento desejado. Ferramentas mais avançadas separam documentos em grupos por assunto ou afinidade (ferramentas de classificação e clusterização). Entretanto, não conseguem extrair conhecimento novo destes grupos. Também não existem ferramentas adequadas para combinar o conhecimento disponível em textos com conhecimentos formalizados de forma estruturada, por exemplo, em bancos de dados.

4 ABORDAGEM DE TEXT MINING PARA INTELIGÊNCIA

A abordagem apresentada neste artigo utiliza ferramentas de Text Mining para formalizar o conhecimento tácito, ou seja, para transformá-lo em conhecimento explícito (externalização), e para explorar este conhecimento depois de formalizado.

A etapa de **formalização** é feita através da análise de textos livres gerados por meios manuais. Nesta análise, ferramentas são utilizadas para identificar os conceitos presentes nos textos. Conceitos representam entes do mundo real e permitem entender que temas estão presentes nos textos ou do que tratam os textos.

Depois, a **exploração** é feita através de um processo automático de mineração. Nesta etapa, são aplicadas ferramentas de *data mining*, não sobre dados estruturados de bancos de dados, mas sobre os conceitos extraídos dos textos na etapa anterior. Esta mineração é feita analisando-se a distribuição dos conceitos em coleções (a freqüência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para descobrir associações e dependências.

A vantagem do uso de conceitos é que estes representam melhor que palavras os objetos, eventos, sentimentos e ações do mundo real. Abordagens baseadas em conceitos (*concept-based approaches*) já são usadas com sucesso para recuperação de informação. Lin e Chen (1996) comentam as vantagens deste tipo de abordagem em relação à busca por palavras-chave. Sua principal vantagem é minimizar o problema do vocabulário (uso de sinônimos, termos correlatos, palavras com vários significados). Uma área onde este tipo de abordagem está sendo usado de forma inovadora é a análise de discurso, para identificar idéias e ideologias presentes em textos. Por exemplo, Chen e outros (1994) usaram com sucesso a identificação de conceitos para organizar idéias discutidas num processo de *brainstorming* eletrônico.

A seguir, serão descritos os métodos e as ferramentas usadas em cada uma das etapas da abordagem (formalização e exploração).

4.1 A EXTRAÇÃO DE CONCEITOS (FORMALIZAÇÃO)

A extração de conceitos é feita através de um processo semi-automático. As regras para identificação dos conceitos são definidas manualmente com auxílio de ferramentas automatizadas. Depois, um processo de categorização identifica automaticamente os conceitos presentes nos textos usando as regras previamente definidas.

Textos não referenciam explicitamente conceitos, mas sim utilizam palavras para fazer referência a entes do mundo real (Apté et al., 1994). Então é possível identificar os conceitos através da análise de palavras e construções gramaticais (Sowa, 2000).

Nesta abordagem, cada conceito deve ser definido através de uma ou mais regras para identificação. Cada regra será verificada contra todas as frases de um texto. As regras combinam termos positivos e negativos. Para um conceito estar presente em uma frase, todos os termos positivos devem

estar presentes na frase e nenhum termo negativo pode aparecer. Se uma das regras for verdadeira para a frase sendo analisada, então o conceito está presente na frase e, conseqüentemente, no texto. Por exemplo, no domínio médico, o conceito *"alcoolismo"* pode ser definido pelas regras (o símbolo "–" indica um termo negativo):

(i) álcool –nega

(ii) hálito etílico

O termo negativo "*nega*" aparece para eliminar frases como "*o paciente nega uso de álcool*".

Todas as frases são comparadas contra todos os conceitos (e todas as suas regras). Se um conceito está presente mais de uma vez no texto, este valor pode ser usado para indicar o quanto um conceito é referenciado num texto. Por exemplo, se um cliente reclama três vezes de um certo problema na mesma interação, isto é diferente de um cliente citando apenas uma vez o mesmo problema. Por enquanto, a abordagem não está utilizando estes valores, mas sim trabalhando com valores binários (conceito presente ou não).

A definição dos conceitos (quais conceitos serão analisados e as regras de identificação de cada um deles) pode ser feita de várias formas. No momento, a abordagem combina tarefas manuais/intelectuais com ferramentas automatizadas. As ferramentas ajudam as pessoas a entenderem como os termos estão sendo utilizados nos textos (que termos estão sendo usados e como, em que contexto). As pessoas podem ainda aumentar este vocabulário usando sinônimos e palavras correlatas extraídas de dicionários. As ferramentas também são utilizadas para analisar amostras de frases onde os conceitos aparecem, para verificar se as regras funcionam corretamente. Alarmes falsos podem ser analisados para identificar termos negativos.

4.2 A MINERAÇÃO DOS CONCEITOS (EXPLORAÇÃO)

O processo de mineração aplica ferramentas de *data mining* sobre os conceitos extraídos na etapa anterior. As técnicas utilizadas são as mesmas existentes na área de mineração de dados ou descoberta de conhecimento em bancos de dados (*Data Mining* ou *KDD – Knowledge Discovery in Databases)* (Fayyad et al., 1996). A diferença é que as ferramentas devem ser aplicadas sobre os conceitos extraídos nos textos e não sobre itens de um banco de dados.

Duas ferramentas específicas estão sendo usadas: uma para análise de distribuições e outra para identificar associações. A primeira verifica a freqüência com que ocorrem os conceitos num conjunto de textos (pode ser a coleção toda ou parte dela). O resultado é o que se chama de centróide (um vetor de conceitos e suas freqüências). Isto permite analisar que temas são mais dominantes e quais aparecem menos. Também é possível comparar um centróide com outro (por exemplo, centróides de duas subcoleções diferentes). Assim, podem ser encontrados temas comuns em duas coleções ou temas exclusivos e também disparidades ou similaridades nas freqüências dos conceitos.

Já a segunda ferramenta descobre relações ou associações entre conceitos, expressando os resultados na forma de regras X ➔ Y (X pode ser um ou mais conceitos e Y somente um conceito). A regra significa que *"se* **X** *está presente em um texto, então* **Y** *também está presente com um certo grau de certeza"*.

O grau de certeza é dado por valores de confiança e suporte. De acordo com a analogia proposta por Lin e outros (1998) e Garofalakis e outros (1999), os textos (ou documentos) são tratados como transações e os conceitos como os itens do banco de dados. Assim, a interpretação do grau de *confiança* (*confidence*) para uma regra associativa do tipo X ➔ Y é a proporção de textos que possuem X e Y em relação ao número de textos que possuem somente X. Da mesma forma, o *suporte* da mesma regra (*support*) é interpretado como o número de documentos onde X e Y estão presentes (ou a proporção em relação à coleção toda). O grau de confiança funciona como uma probabilidade condicional. Isto permite predizer a presença de um conceito em função da presença de outro.

Nem todas as regras são importantes, novas ou úteis. Para filtrar regras interessantes, devem ser definidos limiares para os valores de confiança e suporte. Feldman e Dagan (1998) também sugerem fazer comparações entre subcoleções (regras comuns e exclusivas) ou comparar as regras das subcoleções com as regras da coleção toda. Também se pode separar a coleção por períodos de tempo e assim comparar as regras extraídas em cada período.

5 APLICAÇÕES (ESTUDO DE CASOS)

A aplicação da abordagem de Text Mining tem por objetivo gerar novos conhecimentos sobre a organização, para melhorar processos internos, serviços, produtos e relacionamento com clientes. Para tanto, o conhecimento tácito (de colaboradores ou de clientes) deve ser armazenado de forma livre em textos não estruturados.

Neste artigo, duas aplicações são apresentadas e discutidas. Uma para formalizar e explorar o conhecimento de clientes, adquiridos através de pesquisas e contendo reclamações ou sugestões sobre o negócio de uma empresa de TV por assinatura.

Na outra aplicação, a abordagem foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica, com o objetivo de analisar as características de pacientes e entender o processo de diagnóstico. Neste segundo caso, o conhecimento tácito foi capturado em textos escritos por médicos (prontuários) no momento da internação dos pacientes e contendo descrições de sinais, sintomas e comportamento social do paciente.

Dois métodos diferentes de definição dos conceitos foram usados (seleção de conceitos e definição das regras para identificação nos textos). Na primeira aplicação, os conceitos foram definidos por leigos, analisando os termos presentes nos textos da coleção e o seu contexto (como eram usados). Na segunda, especialistas da área ajudaram na definição dos conceitos e das regras.

5.1 PRIMEIRA APLICAÇÃO: CONHECIMENTO DE CLIENTES

Neste caso, o conhecimento tácito foi coletado através de uma pesquisa com clientes de uma empresa de TV por assinatura. Sugestões e reclamações dos clientes sobre produtos e serviços da empresa (num total de 225) foram registradas em formato de texto livre (um registro para cada cliente). Depois de coletados os textos, o processo de formalização seguiu com a identificação dos conceitos presentes.

Nesta pesquisa, havia também informações estruturadas, como tipo de plano ou pacote do cliente e seu canal preferido. Estes últimos dados foram usados no processo de exploração (mineração ou combinação) para separar a coleção de textos por classes.

Na tabela 1, são apresentados alguns exemplos de padrões interessantes descobertos nesta primeira aplicação. Na primeira coluna, aparecem padrões referentes à distribuição dos conceitos na coleção toda, e na segunda coluna, as regras associativas derivadas da coleção toda com o seu grau de confiança.

Algumas conclusões podem ser obtidas dos resultados apresentados na tabela 1. Metade dos clientes tem alguma sugestão ou reclamação sobre filmes. Em geral, uma sugestão vem de uma insatisfação e também pode ser considerada uma reclamação, só que não explícita. Destas reclamações (sobre filmes), 39,5% falam também de repetição, como pode ser notado nas regras associativas. Segundo o senso comum, infere-se que esta é uma insatisfação dos clientes. Este então é um ponto fraco da empresa e seu negócio pode ser melhorado diminuindo-se a repetição de filmes. Ainda pode-se notar que alguns poucos clientes citaram a concorrência (5,3%), mas este pode ser um valor alto para a empresa (deve-se analisar a proporção de perdas de clientes). Destes, segundo as regras associativas, 33,3% citaram o custo. Infere-se que estes clientes estão dizendo que a concorrência tem custo menor.

Tabela 1: padrões descobertos na coleção toda

Distribuições	Regras Associativas
filmes – 50,7%	imagem ➔ qualidade (80,00%)
custo – 20,4%	pacote *A* ➔ custo (66,67%)
programação – 19,6%	concorrência ➔ filmes (58,3%)
pacote – 15,6%	filmes ➔ repetição (39,5%)
revista – 10,7%	atendimento ➔ demora (37,5%)
pay per view – 6,2%	concorrência ➔ custo (33,3%)
esportes – 5,3%	filmes ➔ qualidade (18,4%)
concorrente – 5,3%	filmes ➔ concorrência (6,1%)
imagem – 4,4%	filmes ➔ pay per view (6,1%)
som – 4,4%	filmes ➔ lançamento (4,4%)
documentários – 3,1%	
seriados – 3,1%	
futebol – 2,7%	

Como discutido anteriormente, o conhecimento tácito formalizado em textos pode ser combinado com conhecimento explícito presente em bancos de dados. Nesta primeira aplicação, o tipo de plano ou pacote que o cliente assina bem como seu canal preferido foram explicitamente registrados. Assim, foi possível separar a coleção de textos em subcoleções, com o intuito de explorar as sugestões e reclamações referentes a cada tipo ou perfil de cliente.

Na tabela 2, são apresentados os conceitos mais freqüentes nas reclamações dos clientes do pacote A (mais caro) e dos clientes do pacote D (mais barato).

Tabela 2: padrões por tipo de pacote

Pacote A	***Pacote D***
filmes – 36,4%	custo – 38,5%
custo – 24,2%	atendimento – 23,1%
repetição – 22,7%	filmes – 15,4%
programação – 22,7%	

Pode-se observar na tabela 2 que os clientes do pacote *A* reclamam menos do custo que os clientes do pacote *D* e que os primeiros estão mais insatisfeitos com os filmes da programação geral do que os segundos. Este conhecimento permite entender melhor os interesses dos clientes de cada pacote, podendo-se gerar um perfil de clientes por tipo de pacote.

Como havia o registro explícito do canal preferido de cada cliente, a coleção de textos foi dividida em 3 partes referentes ao tipo de canal preferido (esportes, filmes e notícias). Na tabela 3, são apresentados os conceitos mais freqüentes por tipo de canal preferido.

Tabela 3: padrões por canal preferido

Esportes	***Filmes***	***Notícias***
filmes – 39,4%	filmes – 60,9%	filmes – 65,4%
custo – 30,3%	custo – 17,2%	custo – 19,2%
pay per view – 15,2%	pay per view – 4,7%	pay per view – 7,7%
concorrência – 15,2%	concorrente – 3,1%	concorrente – 0
atendimento – 6,1%	atendimento – 7,8%	atendimento – 11,5%
clube – 0	clube – 3,1%	clube – 7,7%
ponto extra – 0	ponto extra – 3,1%	ponto extra – 0

O quadro comparativo da tabela 3 permite traçar um perfil do cliente por interesse. Uma das constatações é 15,2% dos clientes que preferem canais de esportes citaram também o conceito "*pay per view*", talvez estando mais suscetíveis a fazer aquisições deste tipo do que os demais (4,7% em filmes e 7,7% em notícias). Pode-se também notar que os clientes que citaram os canais de filmes como preferidos também citaram o conceito *"ponto extra"* (os outros não). Disto pode-se inferir que estes clientes estão mais interessados em ter um ponto extra. Dos clientes que escolheram um canal de esporte como favorito, 15,2% citaram a concorrência (bem mais que os demais clientes). Isto levanta a hipótese de que a concorrência possa estar oferecendo algo melhor em termos de esportes. Analisando-se as regras associativas desta primeira subcoleção (não apresentadas neste trabalho), não foi detectado nenhum padrão associativo entre *"pay per view"* e *"concorrente"*. Assim, pode-se inferir que estes 15,2% referentes aos dois conceitos não são os mesmos clientes, ou seja, quem cita um destes dois conceitos provavelmente não cita o outro (nesta subcoleção).

5.2 SEGUNDA APLICAÇÃO: CONHECIMENTO INTERNO À ORGANIZAÇÃO

Na segunda aplicação, a abordagem foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica. O processo de formalização (externalização) iniciou com o registro em textos livres do resultado da entrevista do médico com o paciente e seus familiares, feita na internação. Estes textos formam parte do prontuário do paciente na clínica e contém informações sobre o comportamento social e familiar do paciente, história pregressa, remédios que toma ou que foram prescritos, além de sinais e sintomas identificados pelo médico durante a entrevista.

Durante 4 meses, foram coletados 400 textos. Para cada texto havia associado um diagnóstico, decidido por um médico da clínica para representar a doença mental do paciente. Entretanto, a indicação do diagnóstico não estava explicitamente expressa no texto. Os textos podem ser considerados semi-

estruturados, pois, apesar de serem escritos em linguagem livre, continham informações previamente planejadas. Isto é, o médico anotava somente informações relevantes para o diagnóstico.

A formalização foi completada com a identificação dos conceitos presentes nestes prontuários através das ferramentas de Text Mining. Os conceitos definidos para esta aplicação representavam sinais, sintomas, pessoas, objetos, eventos e referências ao comportamento do paciente, por exemplo: insônia, agressividade, familiares, mãe, irmãos, arma de fogo, choro, uso de álcool.

Para o processo de exploração (mineração), a coleção toda de textos foi analisada de forma conjunta e também de forma separada por diagnóstico, combinando-se o conhecimento tácito formalizado (conceitos) com o explícito previamente existente (doença do paciente). Também foi possível separar a coleção por remédio utilizado, já que os diferentes remédios foram definidos como conceitos e identificados nos textos, ou seja, não era necessário ter esta informação de maneira prévia e estruturada (os remédios foram inferidos dos textos livres).

Na tabela 4, são apresentados alguns padrões interessantes encontrados na coleção toda. Na coluna da esquerda aparecem os conceitos mais freqüentes e na coluna da direita as regras associativas com maior grau de confiança.

Tabela 4: padrões descobertos na coleção toda

Distribuições	**Regras Associativas**
familiares – 84,5%	alcoolismo ➔ inapetência (84%)
agressividade – 77,0%	autismo ➔ alteração de pensamento (95,3%)
inapetência – 76,0%	agressividade ➔ familiares (92,8%)
remédios – 74,5%	depressão ➔ insônia (85,1%)
insônia – 71,0%	religião ➔ remédios (85,1%)
alteração de pensamento – 70,5%	
nervosismo – 68,5%	
alteração de atenção – 54,5%	

Este conhecimento descoberto permite traçar um perfil do paciente típico que é atendido na clínica. Isto quer dizer que, pela tabela 4: 84,5% dos pacientes têm familiares ou fazem algum tipo de referência a estes; 77% apresentam sinais de agressividade; 76% citaram sofrer de algum tipo de inapetência (falta de apetite), 74,5% já fazem uso de algum remédio, etc.

Analisando-se as regras associativas da tabela 4, é possível predizer características pela presença de outras ou inferir relações de dependência entre as características. Por exemplo, pode-se notar que: em 84% dos casos, pacientes com sintomas de alcoolismo também apresentam inapetência (falta de apetite); quase sempre, sintomas de autismo são acompanhados de alteração de pensamento; 85,1% dos pacientes com sintomas de depressão apresentam também insônia; e 85,1% dos pacientes que citaram algo relacionado a religiões tomam remédios. Pode-se também levantar a hipótese de que a agressividade esteja relacionada à família, já que 92,8% dos pacientes com sintomas de agressividade citaram familiares.

Na tabela 5, são apresentados padrões (conceitos mais freqüentes e regras associativas) referentes aos pacientes com o diagnóstico de esquizofrenia. Analisando tais padrões, é possível verificar que os pacientes esquizofrênicos apresentam maior incidência de "*alteração de pensamento*" (83,5%) do que a média geral (70,5% na tabela 4). Pelas regras de associação, é possível inferir que pacientes que "*ouvem vozes*" têm predisposição para agressividade, e pacientes homicidas (com alguma história envolvendo ameaças de morte) também apresentam sintomas de agressividade.

Tabela 5: padrões para o diagnóstico de esquizofrenia

Distribuições	**Regras Associativas**
familiares – 84,5%	agressividade ➔ familiares (92,94%)
alteração de pensamento – 83,5%	alteração de atenção ➔ agressividade (88,89%)
agressividade – 82,5%	homicida ➔ agressividade (97,67%)
nervosismo – 75,7%	homicida ➔ familiares (95,35%)
insônia – 75,7%	ouvir vozes ➔ agressividade (90,91%)
remédios – 73,8%	perseguição ➔ insônia (84,85%)
inapetência – 68,9%	insônia ➔ remédios (80,77%)
perseguição – 64,1%	
ouvir vozes – 64,1%	
alteração de atenção – 52,4%	

Os padrões de esquizofrenia podem ser comparados com os de outras doenças. Na tabela 6, por exemplo, é apresentado o conhecimento descoberto sobre os pacientes com distúrbios afetivos. Pode-se notar que a incidência de insônia é bem maior nos pacientes com distúrbios afetivos que nos esquizofrênicos (85,2% contra 75,7%). Nesta segunda classe, aparecem novos sintomas entre os mais freqüentes: *"suicida"* (81,5%), *"depressão"* (74,1%) e *"choro"* (63%). As regras associativas também apresentam diferenças.

Os padrões descobertos em cada classe de paciente permitem traçar o perfil de doenças para estudos epidemiológicos. Também podem ser usados para treinamento de pessoas, para validação de decisões médicas ou para a construção de sistemas automáticos de classificação. Nestes casos, a abordagem de Text Mining funciona como um mecanismo de aprendizagem supervisionada, onde textos são selecionados como exemplos de uma determinada classe, e os padrões descobertos servem para descrever as características desta classe. Em um experimento anterior, um sistema automatizado criado com este conhecimento chegou a mais de 60% de acertos no diagnóstico de novos casos.

Tabela 6: padrões para o diagnóstico de distúrbios afetivos

Distribuições	**Regras Associativas**
insônia – 85,2%	agressividade ➔ familiares (87,50%)
familiares – 85,2%	homicida ➔ suicida (92,31%)
suicida – 81,5%	remédios ➔ suicida (85,71%)
inapetência – 81,5%	morte ➔ inapetência (90,91%)
remédios – 77,8%	inapetência ➔ insônia (86,36%)
depressão – 74,1%	inapetência ➔ suicida (81,82%)
pensamento – 70,4%	
alteração de atenção – 66,7%	
choro – 63,0%	
nervosismo – 63,0%	
agressividade – 59,3%	

Nesta segunda aplicação, foi possível também descobrir padrões referentes ao uso de remédios. A seguir, são apresentados os conceitos mais freqüentes identificados na subcoleção de pacientes que utilizaram o remédio Dienpax: "*inapetência*" (91.8%), "*agressividade*" (83.7%), "*alteração de pensamento*" (78.3%), "*nervosismo*" (75.6%), "*insônia*" (64.8%), "*alcoolismo*" (62.1%), "*ouvir vozes*" (59.4%). Este conhecimento permite avaliar o uso da medicação ou pode ser utilizado no treinamento de estudantes e médicos assistentes.

6 CONCLUSÃO

Este artigo mostrou que a tecnologia de Text Mining pode ser usada para formalizar e explorar conhecimento tácito, se este for capturado em textos.

Pelo que se pôde ver nas aplicações discutidas, a abordagem permitiu gerar novos conhecimentos sobre a organização.

As duas aplicações apresentadas demonstram que é possível analisar tanto o conhecimento interno das organizações quanto o conhecimento disponível com os clientes.

Na primeira aplicação, a abordagem foi utilizada para gerar inteligência do negócio a partir do conhecimento dos clientes de uma empresa de TV por assinatura. Foram apontados caminhos para melhorar processos, serviços, produtos e relacionamento com clientes.

Na segunda aplicação, o conhecimento interno disponível com colaboradores de uma clínica psiquiátrica pôde ser formalizado e explorado. Este conhecimento servirá para entender melhor o perfil dos pacientes da clínica e o processo de diagnóstico feito pelos médicos. Também poder-se-á utilizar o conhecimento descoberto para treinar novos colaboradores, para validar decisões e para a construção de sistemas automatizados (sistemas especialistas ou de suporte à decisão).

Alguns cuidados se fazem necessários no uso desta abordagem. Primeiro, o processo de mineração fica condicionado à qualidade da identificação dos conceitos nos textos. A avaliação do processo de extração conduzida no segundo experimento apontou um resultado de 90% de acertos na identificação dos conceitos. Também podem surgir erros quando o conhecimento tácito é registrado em textos livres. Erros ortográficos ou informações incorretas, imprecisas, ambíguas e incompletas podem distorcer os resultados finais. Por fim, a interpretação dos resultados também está condicionada à

interpretação dos conceitos. Por exemplo, na segunda aplicação, o conceito *"alcoolismo"* deve ser interpretado como uma referência a tal nos textos e não como a certeza de que o paciente tem este sintoma (a referência pode ser de que um familiar usa álcool).

7 AGRADECIMENTOS

Este trabalho tem o apoio parcial de: CNPq, CAPES, FIDEPS (Fundo de Incentivo ao Desenvolvimento do Ensino e da Pesquisa em Saúde).

8 REFERÊNCIAS BIBLIOGRÁFICAS

APTÉ, C. et al. (1994). "Automated learning of decision rules for text categorization". ACM Transactions on Information Systems, v.12, n.3, pp.233-251.

CHEN, H. et al. (1994). "Automatic concept classification of text from electronic meetings". Communications of the ACM, v.37, n.10, pp.56-73. Online at http://ai.bpa.arizona.edu/papers/ebs92/ebs92.html

DAVIES, Roy. (1989). "The creation of new knowledge by information retrieval and classification". Journal of Documentation, v.45, n.4, pp.273-301.

FAYYAD, Usama M. et al. (eds) (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park: The MIT Press.

FELDMAN, R. & DAGAN, I. (1998). "Mining text using keyword distributions". Journal of Intelligent Information Systems, v.10, n.3, pp. 281-300.

GAROFALAKIS, Minos N. et al. (1999). "Data mining and the web: past, present and future". In: ACM Workshop on Information and Data Management, Kansas City.

IMHOFF, C.; LOFTIS, L.; GEIGER, J. (2001). Building the customer centric enterprise. John Wiley & Sons.

LIN, C.H. & CHEN, H. (1996). "An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents". IEEE Transactions on Systems, Man and Cybernetics, v. 26, n.1, pp. 1-14. Disponível por WWW em http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

LIN, S.H. et al. (1998). "Extracting classification knowledge of Internet documents with mining term associations: a semantic approach". In Proc. 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98), Melbourne, August 1998, pp.241-249.

LOH, Stanley; WIVES, Leandro K.; OLIVEIRA, José Palazzo M. "Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva". In: LETHELIER, E. et al. (eds). Proceedings, International Symposium on Knowledge Management / Document Management, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.125-147.

NONAKA, I. & TAKEUCHI, H. (1997). Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Campus.

PEREIRA, Rita C. F. & ANGELONI, M. T. (2000). "O relacionamento com os clientes para transformação do conhecimento na organização". In: LETHELIER, E. et al. (eds). Proceedings, International Symposium on Knowledge Management / Document Management, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.89-104.

RODRIGUES, Hugo T. et al. (2000). "Arquitetura da gestão pelo conhecimento focada na inovação". In: LETHELIER, E. et al. (eds). Proceedings, International Symposium on Knowledge Management / Document Management, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.59-76.

SOWA, J.F. (2000). Knowledge representation: logical, philosophical, and computational foundations, Pacific Grove: Brooks/Cole Publishing Co.

STEWART, Thomas A. (1998). Capital intelectual: a nova vantagem competitiva das empresas. 2ª ed. Rio de Janeiro: Campus.

TAN, Ah-Hwee. (1999). “Text mining: the state of the art and the challenges”. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases – PAKDD’99, p. 65-70, Beijing, April 1999. Disponível por WWW em http://textmining.krdl.org.sg/publications.html.

WANDERLEY, A.V.M. (1999). “Um instrumento de macropolítica de informação: concepção de um sistema de inteligência de negócios para gestão de investimentos de engenharia”. Revista Ciência da Informação, v.28, n.2.

Artigo 2 – Applied Intelligence

"Knowledge discovery in texts for constructing decision support systems"
Co-autores: José Palazzo M. de Oliveira, Maurício Almeida Gameiro
Journal of Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies
Special Issue on Text and Web Mining
Editores convidados: Ah-Hwee Tan e Philip S. Yu
http://textmining.krdl.org.sg/APIN/TWMcfp.html

KNOWLEDGE DISCOVERY IN TEXTS FOR CONSTRUCTING DECISION SUPPORT SYSTEMS

STANLEY LOH[1,2,3], JOSÉ PALAZZO M. DE OLIVEIRA[1]
AND MAURICIO A. GAMEIRO[4]
[1]*Instituto de Informática - PPGC, Federal Univ. of Rio Grande do Sul, Porto Alegre, Brazil*
[2]*Computer Science Department, Catholic University of Pelotas, Pelotas, Brazil*
[3]*Computer Science Department, Lutheran University of Brazil, Canoas, Brazil*
[4]*Olivé Leite Hospital, Pelotas, Brazil*
sloh@zaz.com.br
palazzo@inf.ufrgs.br
magameiro@uol.com.br

Abstract. This paper presents a Text Mining approach for discovering knowledge in texts to later construct decision support systems. Text mining can take advantage of knowledge stored in textual documents, reducing the effort for knowledge acquisition. The approach consists in performing a mining process on concepts present in texts instead of working with words. The assumption is that concepts represent real world events and characteristics better than words, allowing the understanding and the explanation of the reasoning used in decision processes. The proposed approach extracts concepts expressed in natural phrases, and then analyzes their distributions and associations. Concepts distributions and associations are used to characterize classes or situations. After the discovery process, the obtained knowledge can be embedded in automated systems to classify elements or to suggest actions or solutions to problems. In this paper, experiments using the approach in a psychiatric domain are discussed. Concepts extracted from textual medical records represent patients' symptoms, signals and social/behavior characteristics. An automatic system was constructed with the approach: a classifier whose goal is to help physicians in disease diagnoses. Results from this system show that the approach is feasible for constructing decision support systems with satisfactory performance.

Keywords: text mining, knowledge discovery, knowledge acquisition, text analysis, decision support systems

1. INTRODUCTION

Currently, people and organizations have stored a great volume of textual documents, most of them in electronic formats and accessible via Web. Unfortunately, information coded in this documentation is only available for manual analyses, making the discovery of new knowledge a difficult task.

Text Mining or Knowledge Discovery in Texts is an emerging area characterized as a set of techniques and tools that help people to analyze texts and to extract new and useful knowledge implicit in documents [6]. As result, this knowledge can be used to construct decision support systems. Wilcox et al [19] suggest exploring the existing knowledge sources to improve the generation of classifiers, reducing costs with knowledge acquisition processes.

In this paper, we present a Text Mining approach for discovering knowledge in texts. The discovered knowledge can be later embedded in decision support systems. The approach consists in performing a mining process on concepts present in texts instead of operating on isolated words. The assumption is that concepts represent real world events and objects better than words and therefore can be used to improve the reasoning process of a decision support system.

The discovery process extracts concepts expressed in natural language and then analyzes their distributions and associations. Concepts distributions and associations are used to characterize classes or situations. After the discovery process, the knowledge can be embedded in automated systems to classify elements or to suggest actions or solutions to problems. The Text Mining approach, here described, is under the supervised learning paradigm, since it needs human organization of the textual sources for discovering reasoning patterns.

The approach is well suited for applications where there is knowledge coded in textual formats (like free texts) and where it is necessary to construct a decision support system. Wilcox et al [19] state that manual inference processes are expensive (difficult and time-intensive). Then Text Mining can reduce the effort needed in knowledge acquisition processes, taking advantage of knowledge present in textual documents.

In this paper, experiments using the approach in a psychiatric domain are discussed. Concepts extracted from textual medical records represent patients' symptoms, signals and social/behavior characteristics. A decision support system is constructed with the approach: a classifier whose goal is help physicians to diagnose patient's disease. Results from this system show that the approach is feasible to be applied in supporting medical decisions and for training assistant physicians or students.

Section 2 discusses related works and problems, pointing to the proposed solution. Section 3 presents in details the Text Mining approach. Section 4 presents the application of the discovery approach in texts of a psychiatric hospital. Section 5 discusses the construction of a decision support system for the psychiatric domain and presents the experiments carried out. Section 6 analyzes the results of these experiments. Concluding remarks and future work are presented in the section 7.

2. RELATED WORK

Feldman et al [6] [7] use a Text Mining approach over keywords that are assigned to texts (as attributes). The mining techniques use statistical analysis to discover association rules and interesting patterns over keyword distributions and associations. However, keywords should be previously assigned to texts.

By other side, Lin et al [15] use terms automatically extracted from the text to characterize documents and to find associations or co-relations. The most frequent terms are assigned as keywords (attributes). However, texts are not pre-processed (full texts are employed). Thus, the context of terms is not analyzed, generating patterns difficult to understand without a full reading of the text (for example, "operating" is different from "operating system").

Feldman et al [8] suggest extracting terms directly from texts but using syntactic information. The idea is not to use all terms but only the meaningful ones, for example, sequences like "*noun noun*", "*noun preposition noun*" and "*adjective noun*".

However, there is still a problem. When analyzing words, problems arise due to the *vocabulary problem.* As discussed by Chen et al [3] [5] and by Furnas et al [9], the language use may cause semantic mistakes due to synonymy (different words for the same meaning), polysemy (the same word with many meanings), lemmas (words with the same radical) and quasi-synonyms (words related to the same subject, object or event, like "*bomb*" and "*terrorist attack*"). For example, a murder may be described with terms like "*murder*" or "*homicide*". If analyzing only the terms, someone could discover, for example, that "*murder*" is present in 56% and "*homicide*" in 67% of the records of a police department. However, it would be difficult to infer how much this kind of crime (murder/homicide) happens. Experiments of Jensen and Martinez [12] and of Wilcox et al [19] confirmed that the use of synonyms reduces the vocabulary problem and improves text categorization processes.

In addition, although words are used to represent the real world in texts, they are not indicated to be used in a knowledge base. Knowledge-based systems need models more complex to represent knowledge. Furthermore, words difficult the validation of the knowledge embedded in automated systems. It is hard to explain how the system reached a decision.

Consequently, it is necessary to represent knowledge in a higher level than that of words and, if possible, to extract this knowledge directly from the texts. In this way, the work of Subasic and Huettner [18] identifies qualitative attributes (like "*horror*", "*justice*" and "*pain*") in texts about movies. These attributes are then used to categorize the movies (for example, according to that work, in action movies, "*horror*" is more central than "*humor*"). However, the categorization rules have to be defined manually (an expensive and difficult task).

Wilcox et al [19] make automatic knowledge discovery from texts for later generating medical reports classifiers. One proposal is to use natural language processing to convert narrative texts to a set of observations and findings represented by standardized codes. Analyzing the codes in training sets (textual sources), decision rules can be extracted automatically to generate classifiers.

One difference of the proposed approach from Wilcox's one is the application domain:
Wilcox employs texts and classifiers in radiology and the proposed approach uses texts from the psychiatric domain. Unlike other medical areas where there are strong indicators of certain diseases, the diagnose process in Psychiatry is more complex and does not work well with deterministic rules, since symptoms and signals alone do not indicate with accuracy a disease. For example, there is no signal or symptom exclusive of a unique disease and symptoms and signals may be present in more than one disease. Signals and symptoms must be correlated and evaluated in a social and historic context.

The Text Mining approach presented here allows the construction of efficient decision support systems for this complex area with the additional advantage of allowing the explanation of the reasoning used in the process.

3. THE TEXT MINING APPROACH

The proposed approach for Text Mining analyzes high-level characteristics of texts, allowing qualitative and quantitative analyses over the content of a textual collection. Instead of applying mining techniques on terms or keywords extracted from texts, the discovery process works over concepts identified inside texts. Concepts represent real world events and objects, and they help the user to explore, examine and understand the contents (ideas, ideologies, trends, thoughts, opinions and intentions) of talks, texts, documents, books, messages, etc. Chen et al [4], for example, use concepts to identify the content of comments in a brainstorming discussion. In Information Retrieval, concepts are used with success to index and retrieve documents. Lin and Chen [14] comment "*the concept-based retrieval capability has been considered by many researchers and practitioners to be an effective complement to the prevailing keyword search or user browsing*". In this case, its main advantage is to minimize the vocabulary problem.

The proposed approach combines a semi-automatic categorization task with a mining task. Categorization identifies concepts in texts and mining discovers patterns by analyzing and relating concepts distributions in a collection.

3.1 THE CATEGORIZATION TASK

The goal of the categorization task is to identify concepts present in texts. However, documents do not have concepts explicitly stated, but rather they are composed of words [1]. Once concepts are expressed by language constructs (words and grammars) [20], it is possible to identify them in texts by analyzing natural language.

The categorization method identifies concepts analyzing individual phrases of a text. The goal is to verify whether a concept is mentioned in a phrase. Each phrase of the text is compared against rules that define a concept (and against all concepts defined for a domain). Rules combine positive and negative words. To a concept be present in a phrase, all positive words must be present and none negative word may be present. If one of the concept rules is true, then the concept denoted by this rule is present in the phrase and consequently in the text. For example, in a hospital domain, "headache" (a symptom) may be defined as a concept using the following rules (negative words have a '–' before):

(i) headache –deny –denies
(ii) head pain –deny –denies.

The negative term "denies" appears to eliminate false hits like "the patient denies headache".

If the concept is present more than once in a text, the total counting is used as the associative degree between the text and the concept. This degree may be used to indicate how much a concept is referred by a text.

The definition of the concepts (which ones will be used and the rules for identifying them) may be generated in different ways. The approach combines automatic tools and human decision. Automatic tools may help people to have an insight of the language used in the texts (terms and meanings). Humans can augment this vocabulary using technical or Webster-like dictionaries. Software tools can also be useful to analyze textual samples in order to verify if the defined rules work correctly. False hits may help in the definition of negative words. The

final decision about the rules in each concept definition is responsibility of humans. Some ways for defining concepts are discussed in more details in Loh et al [16].

3.2 THE MINING TASK

The present approach uses distribution analyses to discover interesting patterns. One technique used is the key-concept listing, which analyzes concept distributions over the collection. A software tool counts the number of texts where each concept is present, generating a vector (called centroid) of concepts and their frequencies in the collection or in a sub-collection. This technique allows finding which dominant themes exist in a set of texts. Also it is possible to compare one centroid to another (between sub-collections), to find common themes or variations between sub-collections. Another possible usage is to find differences between sub-collections (concepts present in only one sub-collection).

Other technique is the association or correlation. It discovers associations between concepts and expresses these findings as rules in the format X ➔ Y (X may be a set of concepts or a unique one, and Y is a unique concept). The rule means "*if* **X** *is present in a text, then* **Y** *is present with a certain confidence and a certain support*". *Confidence* is the proportion of texts that have X AND Y in relation to the number of texts that have only X, and *support* is the proportion of texts that have X AND Y in relation to all texts in the collection [11] [15]. Confidence works like the conditional probability (*if* X *is present, so there is a certain probability of* Y *being present too*). This allows predicting the presence of a concept according to the presence of another one.

For discovering new knowledge, Feldman and Dagan [7] [8] suggest the inspection of patterns that differ significantly from the full collection, from other related collections or from collections in a different time. Comparisons between sub-collections have been used to discover patterns. Considering that each sub-collection represents a different class, concepts distributions and associations are used as class attributes or descriptors. Therefore, the process works like a supervised learning, where texts that represent a class are mined to discover the class characteristics. That implies in some previous organization in the domain knowledge and in the textual documents for discovering reasoning patterns (decision rules).

4. AN APPLICATION TO THE PSYCHIATRIC DOMAIN

Some experiments have been carried out on a collection of medical records from a psychiatric clinic. In Psychiatry, the diagnosis process is more complex than in other medical specialties. There is not yet a syndrome definition accepted as true by physicians. Symptoms and signals alone can not indicate a disease with accuracy. All characteristics have to be analyzed in a social and historic context.

The goal of the experiments was to discover knowledge about this domain for later constructing a decision support system to the psychiatric diagnosis. The Text Mining approach was used to extract characteristics of patients present in textual records. Analyzing records of patients with a certain disease allows to infer what symptoms, signals and other characteristics may describe that disease.

The advantage of this method is to reduce the knowledge acquisition effort, using automated knowledge discovery tools. The experiment followed four stages intended to:

1) discover knowledge using the text mining approach over a training collection (a supervised learning process);
2) construct an automated system using the discovered knowledge;
3) evaluate the performance of the system in diagnosing patients' disease, using a second collection (test collection); and

4) validate subjectively the results with the help of expert physicians.

In this application, the first medical record of the patient was selected for analysis (one text for each patient). Physicians generated the texts in Portuguese, after interviewing the patient and his/her relatives in the admission moment. Texts contain information about patients' diary activities, social and familiar behavior and past medical history, if readmitted. Symptoms and signals identified by the physician during the interview are also recorded, but the texts do not have explicitly stated the diagnosis.

The experiment employed 65 concepts, corresponding to symptoms, signals and social/behavior characteristics (for example, *inappetence, insomnia, aggressiveness, tobacco use, living alone)* or referencing events, persons or objects (for example, *marriage, husband, wife, children, neighbors, knife, weapon, hanging*).

For selecting the concepts, reports from the International Classification of Diseases, 10th revision (ICD-10) [2] and technical dictionaries from psychiatry were used. In addition, all words and terms used in texts of the training collection were also examined, looking for important references to events, persons or objects. Software tools helped in this last task.

After that processing, the rules for identifying each concept were defined through a similar process. Software tools were used to examine the way in which the words and terms were used in sample texts. Special attention was given to synonyms, which could be identified analyzing samples of texts and with help of a Webster's dictionary. Two professionals helped in this process, which took approximately 30 hours at all, during 2 months. The final decision is responsibility of these professionals.

Following, some examples of the concepts used and their definitions (rules for concept identification) are presented (the symbol '$' indicates a radical and '–' indicates a negative word):

- "*alcoholism*":
  (i) alcohol$ (ii) ethilic (iii) drink (iv) drunk (v) drank; etc.
- "*inappetence*":
  (i) eat not much (ii) feed badly (iii) fed not much; etc.
- "*homicide*":
  (i) kill$ –himself –herself (ii) homicid$; etc.
- "*relatives*":
  (i) mother (ii) father (iii) brother$ (iv) sister$ (v) uncle; etc.

The first collection (for training) was composed of 200 texts, each one corresponding to only one patient. The second collection (for test) had 200 different texts. All texts ranged from 1 to 4 Kbytes in size (minimum of 22 and maximum of 413 words). Each collection corresponds to admissions made during two months, representing a significant amount of knowledge.

The time for categorizing all texts in the training collection took about 1 hour and 20 minutes in a Pentium II 400 MHz with 64 Mbytes of RAM (a comparison of 200 texts against 65 concepts). The mining task took about 15 minutes.

4.1 KNOWLEDGE ABOUT DISEASES

The experiments used four classes of texts, each one corresponding to a disease of the International Classification of Diseases, 10th revision (ICD-10) [2]. These classes are the most frequent in the analyzed hospital. The classes are:

a) ***organic*** mental disturbances (due to brain damage), including codes F00 to F09 of the ICD-10;

b) mental and conduct disturbances due to psychoactive ***substances***, including codes F10 to F19;
c) ***schizophrenia***, schizoid disorders and delirious disturbances, including codes F20 to F29;
d) ***affective*** and mood disturbances, including codes F30 to F39.

For the discovery process, the collection was divided in four sub-collections, each one representing a unique class and containing only medical records corresponding to patients associated with the related disease. The class (disease) of each text was previously determined by a physician in a real diagnosis process but is not explicitly stated in the text.

The mining task was oriented to examine the distributions of concepts in each sub-collection. Interesting patterns include the most frequent concepts of each class and concepts exclusive of only one class. Later, associative rules between concepts were identified in each class in order to find class attributes. A concept was considered representative of a class if it appears at least in two texts in the class sub-collection. Its distribution was used as a relative weight, meaning the strength of a concept to indicate a disease. An associative rule is representative if it appears exclusively in the class sub-collection, with confidence degree greater than 80% and support greater than 40%.

The first collection (for training) was composed of: 27 texts from "*affective*" (13.5%), 103 texts from "*schizophrenia*" (51.5%), 18 texts from "*organic*" (9%) and 52 texts from "*substances*" (26%). The second collection (for test) was very similar (*affective* – 12.4%, *schizophrenia* – 52.7%, *organic* – 8.4% and *substances* – 26.3%).

Part of the discovered knowledge is showed in tables 1 and 2. Table 1 shows the most frequent concepts for the "*affective*" class (percentages indicate the frequency of the concept in the class sub-collection). Table 2 presents examples of exclusive associations for "*schizophrenia*".

insomnia - 85.2%	thought deficit - 70.4%	husband - 48.1%
relatives - 85.2%	attention deficit - 66.7%	homicidal ideas - 48.1%
suicidal - 81.5%	crying - 63.0%	work/job - 40.7%
inappetence – 81.5%	nervousness - 63.0%	children - 40.7%
depression - 74.1%	aggressiveness - 59.3%	death - 40.7%

Table 1: most frequent concepts for affective cases

hearing voices ➔ aggressiveness	hearing voices ➔ nervousness
delusions of persecution ➔ insomnia	delusions of persecution ➔ thought deficit
hearing voices ➔ insomnia	hearing voices ➔ thought deficit

Table 2: exclusive associations for schizophrenia cases

5 EXPERIMENTS WITH A DECISION SUPPORT SYSTEM

The knowledge discovered and presented in the previous section was embedded in a decision support system. The system is a "text classifier", since its goal is to discover the class (disease) associated to a patient, analyzing texts that describe patients.

As suggested by Jensen and Martinez [12], the system was implemented with simple algorithms, so that results are due to the class characteristics (discovered knowledge) more than to the categorization algorithm. Decision trees were left out, because most concepts are not exclusive of one class and since the combination of concepts could form complex rules

with many attributes, making difficult to understand or to explain the rules used in the decision process.

Other alternative was to use a neural net. However, this kind of decision model cannot explain how the decision was obtained.

As a consequence of the former considerations, a classification algorithm based on Rocchio's and Naive Bayes algorithms was selected. A vector is used to represent each class, being composed of concepts and a relative weight per concept (ranging from 0 to 1), meaning the strength of the concept for indicating the association to that class. This weight represents the concept distribution in that class. When analyzing a candidate text, common concepts between the text and the class vector augment the possibility for the text being of that class. The certainty degree augments according to the concept weight in that class (remembering that concepts identified in the texts do not have weights, but are binary decisions).

Several variations of the algorithm have been tested, but maintaining the basic assumption that concepts augment the certainty of a class. One variation was to use pairs of concepts, extracted from the discovered associative rules. The reason is that these associations can help to identify the context of concepts, following an analogy with the conclusions of Apté et al [1] about the usage of words. Two concepts appearing together in a text indicate the class with a greater confidence than if the concepts appear alone. Other variation was the use of a negative value (equal to -1) associated to those concepts that never appear in the class. This helps in the elimination of false candidates. Galavotti et al [10] used with some success negative training instances as negative evidence in text classification (using values equal to –1 instead of zero).

The different methods employed in the experiment are described in the next paragraphs (letters serve to identify the method in the result tables):

- **Wa**: class descriptors are all words present in the corresponding sub-collection of the training set, at least in two texts; weights are associated to words, calculated as the average value of the relative frequency of the word in each document (relative frequency is the count of the word in a document divided by the total number of words in the same document, according to Salton and McGill [17]);
- **Wd**: uses only exclusive words (differences); class descriptors are those words (extracted by analyzing the training texts) that appear only in the class group, at least in two texts;
- **Ca**: uses all concepts and distributions of each class, discovered in the mining task;
- **Ca+n**: uses all concepts (from **Ca**) plus negative concepts with weight = -1;
- **Clf+n**: uses the least frequent concepts of each class plus negative concepts;
- **Cp**: uses all pairs of concepts for each class, extracted from the associative rules discovered in the mining task (confidence equal or greater than 80%);
- **Cp2**: uses pairs of concepts extracted from the associative rules with confidence equal or greater than 50%;
- **Cpd**: uses exclusive pairs, those extracted from associative rules present in only one class (confidence equal or greater than 80%);
- **Ca+p**: uses all concepts (from **Ca**) plus pairs of concepts (from **Cp**);

The method **Cd**, having as class descriptors the concepts present in only one class (the differences), was not used because the mining task did not found exclusive concepts (all concepts belong to more than one class). The first two methods were implemented to compare methods based on words against methods based on concepts.

These methods were chosen to show how concepts and associative rules (the discovered knowledge) might be used in a categorization algorithm. It is important to remember that the experiment goal was not to find the best categorization method, since there are many other alternatives to be implemented. The comparison of the methods helps to demonstrate that "concepts" may be used in decision support systems. The variations were useful to analyze different ways to use the discovered knowledge in a text classifier.

6. RESULTS AND DISCUSSION

The traditional measures precision and recall were used to evaluate the performance. Recall and precision values were calculated using *microaveraging* and *macroaveraging* measures proposed by Lewis [13]. *Microaveraging* considers the whole collection as a unique class and *macroaveraging* first calculates precision and recall inside each class and then extracts the average value for the entire collection. The F-measure was used to join precision and recall and was calculated as 2 * Pr * Rc / (Pr + Rc). An additional value was used to identify the best performance: the average between *microaveraging* F-measure and *macroaveraging* F-measure.

For comparing the methods, they were applied to the 200 texts of the test collection. When more than one class is associated to a text, the class with the greatest degree of relationship was assumed to be the system choice. Table 3 presents the performance of each method.

	Microavg Precision	**Macroavg Precision**	**Microavg Recall**	**Macroavg Recall**	**F-meas. Microavg**	**F-meas. Macroavg**	**Avg**
Wa	0.41	0.54	0.41	0.44	0.41	0.48	0.45
Wd	0.73	0.60	0.72	0.49	0.72	0.54	**0.63**
Ca	0.44	0.50	0.44	0.39	0.44	0.44	0,44
Ca+n	0.51	0.55	0.51	0.42	0.51	0.48	0.50
Clf+n	0.65	0.73	0.61	0.53	0.63	0.61	**0.62**
Cp	0.57	0.51	0.54	0.45	0.55	0.48	0.51
Cp2	0.43	0.47	0.41	0.42	0.42	0.44	0.43
Cpd	0.64	0.54	0.60	0.51	0.62	0.52	**0.57**
Ca+p	0.56	0.51	0.56	0.47	0.56	0.49	0.52

Table 3: results over the test collection

According to table 3, the method **Wd** based on words has a better performance than the best method based on concepts (**Clf+n**). However, the difference was very small, leading to the conclusion that concept-based methods can replace word-based methods in automated systems without losing performance and with the additional advantage of allowing the understanding and the explanation of the knowledge used in the decision process.

Word-based patterns may cause confusion due to the vocabulary problem. For example, a centroid resulting from the mining task had the word "*visual*", but it is difficult to understand whether it was related to "*visual deficit*" or to "*visual illusion*". In the same way, rules about the "*aggressiveness*" symptom may contain the word "*aggressive*" or the word "*aggressiveness*", misleading the mining process. The same problem happens with "*sleep*" (sleeping well or not, walking when sleeping), with "*attention*" and "*thought*" (deficit or normal) and many other cases.

Comparing the performance of the concept-based methods, the method **Clf+n** (using the least frequent concepts plus negative concepts) achieved the best result. Negative evidence helps to improve the class discrimination: the method **Ca+n** performed better than **Ca**.

Pairs of concepts also improve the performance: the method **Cp** performed better than the method **Ca**. However, using only pairs of concepts (**Cp**) do not perform so well as using pairs plus single concepts (**Ca+p**), leading to the conclusion that usage of pairs can improve the performance, but associated with single concepts may achieve better results. This confirms the findings of Apté et al [1] about words: pairs of words can give better results, but using only pairs bring poor results, while single words alone are relatively successful.

Reducing the confidence degree for associative rules does not bring better class descriptors. That can be perceived in the comparison **Cp** x **Cp2**; the former method, using rules with a greater confidence, performed better than the latter.

Differences can improve the performance in word-based methods and in concept-based ones. The method **Wd** performed better than **Wa**, and **Cpd** achieved better results than **Cp, Cp2** and **Ca+p**. The supposition is that the method **Cd** can improve the performance. However, in this application, this method was not implemented because there were not exclusive concepts.

Intending to analyze the performance of the methods in the training collection, two methods were applied over the 200 training texts. Table 4 presents these results.

	Microavg Precision	**Macroavg Precision**	**Microavg Recall**	**Macroavg Recall**	**F-meas. Microavg**	**F-meas. Macroavg**	**Avg**
Wd	0.93	0.95	0.93	0.86	0.93	0.90	**0.92**
Clf+n	0.69	0.70	0.66	0.61	0.67	0.65	0.66

Table 4: results over the training collection

As presented in the table 4, the method **Wd** (words exclusive of one class) achieved the best performance. Results indicate that word-based methods can perform better than concept-based ones when the training and the test collection are the same. However, that is useful only when the collection does not change too much. As discussed early, other problem with the method **Wd** is that words difficult to explain the reasoning used to distinguish one class (disease) from another.

Presenting the results to psychiatrists, they considered the numeric results greater than 60% better than some human performances. It is important to remember that the discovery process was performed quite automatically. Human intervention helped the process only giving training examples, but these records could not contain all the information used by the physician to make the decision; the class associated to training and test texts was the final one. This means that physicians could obtain more information about the patient, after the admission interview, for deciding the disease, while the discovery process and the automated system used only information available in the admission record.

Another subjective evaluation was to present to physicians the knowledge discovered and used by the systems. The physicians considered the class descriptors (concepts and associative rules) similar to the reasoning used in their diary activities. After analyzing patterns word-based and concept-based patterns, physicians considered the latter ones more understandable. Word-based patterns caused confusion with the vocabulary and raised doubts about the meaning of the patterns.

7. CONCLUDING REMARKS

This paper presented a Text Mining approach for discovering knowledge implicit in textual documentation. The experiments demonstrated that software tools could quite automatically discover useful knowledge from this kind of source. The validation of the discovered knowledge was made using it embedded in a decision support system and presenting the results to experts in Psychiatry. Analyzing the performance of the decision support system and remembering that psychiatrists' critics considered good the system results (greater than 60%), it is possible to conclude that the construction of decision support systems is feasible using knowledge discovered by the proposed Text Mining approach.

This does not imply in eliminating the human intervention but has as main advantage the reduction in time and effort for acquiring knowledge and for modeling a decision system. Besides that, people and organizations can make a better use of the great volume of existing texts, taking advantage of the valuable knowledge implicit in this kind of documents.

In this paper, word-based methods were compared against concept-based ones, showing that the latter can achieve results so good as the former, with the additional advantage of allowing the understanding of the discovered knowledge and the explanation of the reasoning process used in decision support systems, as explained in the previous section.

The different concept-based methods implemented in the automated system served to investigate how concepts may be used in a decision support system. Results lead to the conclusion that it is important to use pairs of features, negative evidences and differences as class descriptors for achieving better performance.

The decision support system used in the experiments is a text classifier. Results demonstrate that the Text Mining approach is suitable for helping to construct this kind of system. Future work will evaluate the use of the approach for constructing other kinds of automated systems like control systems and action systems.

A special remark is necessary about the textual collection used in the experiments. The texts in the training collection were selected by time-period (all texts of a time period were used). That could cause noise in the discovery process. The assumption is that it is possible to obtain better results, filtering the collection and retaining only those texts considered good samples for training. Analyzing the volume of texts in each collection, it is important to remember that 200 records correspond to admissions made during two months. In the psychiatric domain, this volume contains significant knowledge. A future work is to evaluate whether greater volumes of texts can improve the results of an automated system.

Other alternative for improving the system performance is to use human intervention for validating and pruning the discovered knowledge before the system construction.

A final conclusion is that the Text Mining approach is better suited for the cases where there is knowledge available in textual documents and when it is necessary to construct decision support systems whose reasoning needs to be explained. A good application is the case where there are great volumes of texts, making difficult the extraction of knowledge using manual tasks. The World Wide Web becomes a special and interesting case, deserving a more accurate investigation.

ACKNOWLEDGMENTS

This research is partially sponsored by: CNPq (Brazilian Council for Scientific and Technological Development) and CAPES. Medical records were provided by Olivé Leite Psychiatric Hospital (Pelotas, RS, Brazil) and have being produced with research support by

FIDEPS (Found of Incentive for the Development of Teaching and Research in Health - Ministry of Health, Brazil).

REFERENCES

1. C. Apté et al., "Automated learning of decision rules for text categorization", ACM Transactions on Information Systems, v.12, n.3, pp.233-251, July 1994.
2. Brazilian Center for Disease Classification, International Classification of Diseases and Health Related Problems in Portuguese, 10th revision, São Paulo: EDUSP, 1989. (in collaboration with the World Health Organization). Online at http://www.datasus.gov.br/cid10/cid10.htm
3. H. Chen, "The vocabulary problem in collaboration", IEEE Computer (special issue on CSCW), v.27, n.5, pp.2-10, May 1994. Online at http://ai.bpa.arizona.edu/papers/ cscw94/cscw94.html
4. H. Chen et al., "Automatic concept classification of text from electronic meetings", Communications of the ACM, v.37, n.10, pp.56-73, October 1994. Online at http://ai.bpa.arizona.edu/papers/ ebs92/ebs92.html
5. H. Chen et al., "A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system", Journal of the American Society for Information Science, v.48, n.1, pp.17-31, January 1997. Online at http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html
6. R. Feldman and I. Dagan, "Knowledge discovery in textual databases (KDT)", in Proc. 1st International Conference on Knowledge Discovery (KDD-95), Montreal, August 1995, pp.112-117.
7. R. Feldman and I. Dagan, "Mining text using keyword distributions", Journal of Intelligent Information Systems, v.10, n.3, pp. 281-300, 1998.
8. R. Feldman et al., "Text mining at the term level", in Proc. 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-98). Lecture Notes in Computer Science Vol. 1510, pp. 65-73, Springer-Verlag, 1998. Online at http://www.wisdom.weizmann.ac.il/~lindell/
9. G.W. Furnas et al., "The vocabulary problem in human-system communication", Communications of the ACM, v.30, n.11, pp. 964-971, November 1987.
10. L. Galavotti et al., "Feature selection and negative evidence in automated text categorization", in Proc. Workshop on Text Mining (KDD-2000), Boston, MA, USA, August 2000. Online at www.cs.cmu.edu/~dunja/wshkdd2000.html
11. M. Garofalakis et al., "Data mining and the web: past, present and future", in Proc. ACM Workshop on Information and Data Management, Kansas City, 1999, pp.43-47.
12. L.S. Jensen and T. Martinez, "Improving text classification by using conceptual and contextual features", in Proc. Workshop on Text Mining (KDD-2000), Boston, MA, USA, August 2000. Online at www.cs.cmu.edu/~dunja/wshkdd2000.html
13. D.D. Lewis, "Evaluating text categorization", in Proc. Speech and Natural Language Workshop, February 1991, pp.312-318. Online at http://www.research.att.com/~lewis
14. C.H. Lin and H. Chen, "An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents", IEEE Transactions on Systems, Man and Cybernetics, v. 26, n.1, pp. 1-14, February 1996. Online at http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

15. S.H. Lin et al., “Extracting classification knowledge of Internet documents with mining term associations: a semantic approach”, in Proc. 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98), Melbourne, August 1998, pp.241-249.
16. S. Loh et al., “Concept-based knowledge discovery in texts extracted from the web”, ACM SIGKDD Explorations, v.2, n.1, pp.29-39, June 2000. Online at http://www.acm.org/sigkdd/explorations
17. G. Salton and M.J. McGill, Introduction to modern information retrieval, New York: McGraw-Hill, 1983.
18. P. Subasic and A. Huettner, “Calculus of fuzzy semantic typing for qualitative analysis of text”, in Proc. Workshop on Text Mining (KDD-2000), Boston, MA, USA, August 2000. Online at www.cs.cmu.edu/~dunja/wshkdd2000.html
19. A. Wilcox et al., “Using knowledge sources to improve classification of medical text reports”, in Proc. Workshop on Text Mining (KDD-2000), Boston, MA, USA, August 2000. Online at www.cs.cmu.edu/~dunja/wshkdd2000.html
20. J.F. Sowa, Knowledge representation: logical, philosophical, and computational foundations, Pacific Grove: Brooks/Cole Publishing Co., 2000.

Artigo 3 – J. Documentation

"Knowledge discovery in textual documentation: qualitative and quantitative analyses".
Co-autores: Jose Palazzo M. de Oliveira, Fábio Leite Gastal
Journal of Documentation, v.57, n.5, September 2001. pp.577-590.
Editor: R. T. Kimber
Publicado pela Aslib, The Association for Information Management (London)
www.aslib.co.uk/jdoc

KNOWLEDGE DISCOVERY IN TEXTUAL DOCUMENTATION: QUALITATIVE AND QUANTITATIVE ANALYSES

STANLEY LOH
sloh@zaz.com.br

Post-Grade in Computer Science, Federal Univ. of Rio Grande do Sul, Porto Alegre, Brazil
Computer Science Department, Catholic University of Pelotas, Pelotas, Brazil
Computer Science Department, Lutheran University of Brazil, Canoas, Brazil

JOSÉ PALAZZO M. de OLIVEIRA
palazzo@inf.ufrgs.br

Institute of Computer Science, Federal Univ. of Rio Grande do Sul, Porto Alegre, Brazil

FÁBIO LEITE GASTAL
flgastal@zaz.com.br

Olivé Leite Hospital, Pelotas, Brazil
Medicine Department, Catholic University of Pelotas, Pelotas, Brazil

This paper presents an approach for performing knowledge discovery in texts through qualitative and quantitative analyses of high-level textual characteristics. Instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in texts. Concepts represent real world events and objects, and they help the user to understand ideas, trends, thoughts, opinions and intentions present in texts. The approach combines a quasi-automatic categorisation task (for qualitative analysis) with a mining process (for quantitative analysis). The goal is to find new and useful knowledge inside a textual collection through the use of mining techniques applied over concepts (representing text content). In this paper, an application of the approach over medical records of a Psychiatric Hospital is presented. The approach helps physicians to extract knowledge about patients and diseases. This knowledge may be used for epidemiological studies, for training professionals and it may be also used to support physicians to diagnose and evaluate diseases.

1. INTRODUCTION

With the growing use of digital resources, people and organisations have stored a great volume of documents. This digital documentation has hidden knowledge, implicit in relations within and among documents (Davies, 1989). In most cases, people and organisations have difficulty to analyse this massive documentation in order to extract new and useful information to improve the knowledge about the domain.

The novel area called *Knowledge Discovery in Texts* (KDT) has emerged to help people to extract knowledge from textual documents, through the application of techniques from *Knowledge Discovery in Databases* (KDD) over texts (Feldman & Dagan, 1995). KDD is the "*nontrivial extraction of implicit, previously unknown, and potentially useful information from given data*" (Frawley et al., 1991) and it has obtained success applying

statistical mining techniques in large databases. However, researches on KDD deals only with structured data (tables, records and fields/attributes). By other side, texts have information coded in natural language sentences (free and unstructured phrases). As a result, information may appear in different styles or formats, making difficult to discover it. The goal of KDT is to extract information from texts and explore the results in order to find new and interesting knowledge. Knowledge is defined as useful information; people receive large amounts of information but only part of this set becomes knowledge since not all information is employed in the discovery process.

This paper presents an approach for knowledge discovery in texts through qualitative and quantitative analyses of high-level textual characteristics exploring the content present in a textual collection. In this approach, instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in texts. Concepts represent real world events and objects, and they help the user to understand ideas, trends, thoughts, opinions and intentions present in texts. The approach combines a quasi-automatic categorisation task (qualitative analysis) with a mining process (quantitative analysis). Categorisation identifies concepts in texts and mining discovers patterns by analysing and relating concept distributions in a collection. The goal is to find new and useful knowledge inside a collection. The discovered knowledge may be used for decision support and evaluation, training of workers and analysis of domain characteristics.

In this paper, an application of KDT concerning medical records of a Psychiatric Hospital is presented. The approach helps physicians to extract knowledge about patients and diseases. This knowledge may be used for epidemiological studies, for training professionals and it may be also used to support physicians to diagnose and evaluate diseases.

Section two analyses related works and explains some problems with the existing approaches. Section three describes the proposed approach in a general view. Section four presents the results of applying the approach in a textual documentation of a Psychiatric Hospital. Section five discusses the quality of the approach and section six presents concluding remarks and future works.

2. RELATED WORK

Lin et al. (1998) use terms automatically extracted from the text to characterise documents and to find associations or co-relations. The most frequent terms are assigned as keywords (attributes). However, when analysing words, problems arise due to the *vocabulary problem*. The language use may cause semantic mistakes due to synonymy (different words for the same meaning), polysemy (the same word with many meanings), lemmas (words with the same radical, like the verb "to marry" and the noun "marriage") and quasi-synonymy (words related to the same subject, object or event, like "bomb" and "terrorist attack") (Chen, 1994), (Chen et al., 1997) and (Furnas et al., 1987). For example, a murder may be described with terms like "murder" or "homicide". If analysing only the terms, the discovery process may be misled by semantic gaps.

Other interesting approach for KDT is to apply KDD techniques after the use of Information Extraction (IE) techniques, which transform information present in texts in attribute values of a structured database (Cowie & Lehnert, 1996). However, IE systems use complex rules, usually based on natural language processing techniques. This process requires a complex computational algorithm and a great human effort of knowledge engineering to

understand how information is coded in natural language (Chinchor et al., 1993) (Gaizauskas & Wilks, 1998).

Feldman & Dagan (1995, 1998) face the KDT problem applying mining techniques over keywords that are assigned to texts (as attributes). These mining techniques use statistical analysis to discover association rules and interesting patterns on keyword distributions and associations. In the cited works, keywords should be previously assigned to texts, either by human or by automatic tasks. When using human assigned keywords, in general only part of all themes present in a text is available for analysis because the human effort concentrates on meaningful aspects. Furthermore, there is the extra work of reading and understanding the text. Automatic tasks correspond to text categorisation and are more suitable for that problem.

However, most text categorisation works rely on a single class method, that is, they find "the class" to which the text belongs. Methods that find more than one class should be employed to get a wider and more significant categorisation. This is a mandatory requirement in medical systems as the symptom complexity leads to a more complex classification process.

Wiener et al. (1995) use neural networks to extract topics from texts (many classes). One problem of this approach is that it is only used for text categorisation (they call it "topic spotting"); no quantitative analysis is done. Another problem is that the extracted knowledge is not used for later analyses or processes, as training, decision support and evaluation.

3. THE APPROACH FOR KDT

The proposed approach for knowledge discovery analyses high-level characteristics of texts, allowing qualitative and quantitative analyses over the content of a textual collection. Instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in the texts. Concepts represent real world events and objects, and they help the user to explore, examine and understand the contents (ideas, ideologies, trends, thoughts, opinions and intentions) of talks, texts, documents, books, messages, etc. Chen et al. (1994), for example, use concepts to identify the content of comments in a brainstorming discussion. In Information Retrieval, concepts are used with success to index and retrieve documents. Lin & Chen (1996) comment "*the concept-based retrieval capability has been considered by many researchers and practitioners to be an effective complement to the prevailing keyword search or user browsing*". In the present approach, the categorisation main advantage is to minimise the vocabulary problem.

The proposed KDT approach combines a quasi-automatic categorisation task with a mining task. Categorisation identifies concepts in texts (qualitative analysis) and mining discovers patterns by analysing and relating concept distributions in a collection (quantitative analysis).

3.1 THE CATEGORISATION PROCESS

The goal of the categorisation is to identify concepts present in texts. However, documents do not have concepts explicitly stated, but instead they are composed of words that represent the concepts. As concepts are expressed by language structures (words and grammars), it is possible to identify concepts in texts analysing phrases (Sowa, 2000). However, we need some straightforward algorithm to accomplish this purpose. The method used in our approach identifies concepts analysing individual phrases of a text. The goal is to verify whether a concept is mentioned in a phrase.

Consequently each phrase of a text is compared against rules that define a concept (and against all concepts defined for a domain). Rules combine positive and negative words. To a concept be present in a phrase, all positive words must be present and none negative word may be present. If one of the concept rules is true, then the concept is present in the phrase and consequently is also present in the text. For example, in a medical domain, "headache" (a symptom) may be defined as a concept using the following rules (negative words have a '-' before):

(i) headache –deny –denies

(ii) head pain –deny –denies

If the concept is present more than once in a text, the total counting is used to define an associative degree between the text and the concept, indicating how much a concept is referred by a text.

The definition of the concepts may be generated in different ways. The proposed approach uses a combination of automatic tools and human decision. Automatic tools help people to have an insight of the language used in the texts (different terms and meanings). Humans, in the other hand, augment this vocabulary using Webster-like dictionaries or technical dictionaries. Software tools can also be useful to analyse textual samples in order to verify if the defined rules work correctly. False hits may help in defining negative words. The final decision about the rules in each concept definition should be responsibility of humans. In other paper more details about some tactics for defining concepts are discussed (Loh et al., 2000).

3.2 THE MINING PROCESS

The approach here described uses distribution analyses to discover interesting patterns. The first technique used is the key-concept listing, which analyses concept distributions over the collection. A software tool counts the number of texts where each concept is present, generating a vector of concepts and their distributions inside the collection (called centroid). Different centroids can be generated for different collections or for parts of a unique collection (sub-collections). This technique allows finding what dominant themes exist in a collection, in a sub-collection or in a single text. In addition, we can compare one centroid to another (between sub-collections), to find common themes or variations between sub-collections. Another possible usage is to find differences between sub-collections (concepts present in only one text). Feldman & Dagan (1998) suggest the exam of distributions that differ significantly from the full collection, from other related collections or from collections in a different time.

The second technique is the association or correlation. It discovers associations between concepts and expresses these findings as rules in the format X ➔ Y (X may be a set of concepts or a unique one, and Y is a unique concept). The rule means, "*if* **X** *is present in a text, then* **Y** *is present with a certain confidence and a certain support*". Following the definitions of Lin et al. (1998), *confidence* is the proportion of texts that have X AND Y in relation to the number of texts that have only X, and *support* is the proportion of texts that have X AND Y in relation to all texts in the collection. Confidence works like the conditional probability (*if* X *is present, so there is a certain probability of* Y *being present too*). This allows predicting the presence of a concept associated to the presence of another concept. Complex rules may be discovered with human intervention. As a result, the precedent part of a rule may be a combination of concepts and/or words, such as WORD_1 AND WORD_2

AND CONCEPT_1 AND CONCEPT_2 ➔ CONCEPT_3. This kind of rule is found using intermediary retrieval tasks, to select sub-collections where some words are present.

The choice for these two techniques is due to their simplicity and extensive use in Data Mining approaches (KDD). One hypothesis is that other different techniques can be used over the concepts, after the categorisation task.

4. EXAMPLE OF AN APPLICATION

Some experiments have being carried out on a collection of medical records from a psychiatric hospital. This domain has special characteristics, as the diagnosis process is more complex than in other medical specialities. Symptoms and signals may be present in different diseases and there are not syndrome definitions, relating symptoms and signals to a specific disease. Other problem is that symptoms and signals may be present in a moment and disappear in other. Besides that, some characteristics may be more predominant than others in a period and a different situation may occur in another time.

The goal of the experiments is to discover knowledge about this domain, so that the results may be used to help physicians in the diagnosis process, to evaluate diagnosis decisions and to qualify students or trainees. From this point of view, the discovery method, described early, has two advantages over others:

1) the analyses are performed on concepts present in the texts instead of on individual words; thus patient symptoms, signals and other characteristics can be analysed (qualitative analysis);
2) the knowledge extracted from the concept distribution analysis (quantitative analysis) may be used to
    a) automatically classify patients in diseases (diagnosis): in the psychiatric domain, inductive decision trees are not well suited, since characteristics may be present in more than one class; consequently methods like ID3 and C4.5 are not indicated in this case; see Ingargiola (1996) for more details on these algorithms;
    b) understand how the process was developed: unlike neural networks, the rules used to identify the class are available to explain why a certain class was associated to a test case.

In this application, the first medical record of patients was used, created in the patient admission. Physicians generated the texts after interviewing the patient and his/her relatives. These records include the patient history and do not have explicitly stated the final diagnosis. Information about the patient concerns the diary activities, social and familiar behaviour and past medical history if readmitted. The records also contain symptoms and signals identified by the physician during the interview.

Two different collections were used. The first one was composed of 200 texts, each one corresponding to only one patient (remembering that it could be a readmission). This collection was used for the training process, to discover knowledge about the domain. The second collection had 200 different texts and was used for the test process, to evaluate the discovered knowledge quality. Some texts in the two collections could correspond to the same patient. Each collection corresponds to admissions made during a two months period.

The texts were classified in one of four major classes, corresponding to diseases of the International Classification of Diseases, 10th revision - ICD-10 (BCDC, 1993). Physicians in a real diagnosis process previously determined the class (disease). The classes were:

e) ***organic*** mental disturbances (due to brain damage), including codes F00 to F09 of the ICD-10;
f) mental and conduct disturbances due to psychoactive ***substances***, including codes F10 to F19;
g) ***schizophrenia***, schizoid disorders and delirious disturbances, including codes F20 to F29;
h) ***affective*** and mood disturbances, including codes F30 to F39.

The first collection (for training) was composed of: 27 texts from "*affective*" (13.5%), 103 texts from "*schizophrenia*" (51.5%), 18 texts from "*organic*" (9%) and 52 texts from "*substances*" (26%). The second collection (for test) was very similar (*affective* – 12.4%, *schizophrenia* – 52.7%, *organic* – 8.4% and *substances* – 26.3%). All texts ranged from 1 to 4 Kbytes in size.

4.1 CONCEPTS USED

The concept definition task selected 65 concepts, corresponding to symptoms, signals and social characteristics (for example, *inappetence, insomnia, aggressiveness, tobacco use, living alone)* or referencing events, persons or objects (for example, *marriage, husband, wife, children, neighbours, knife, weapon, hanging*). The stated goal was to identify references inside the texts, which could be important to characterise the diseases of the patients.

For selecting the concepts, ICD reports and dictionaries from psychiatry were used. Also all words and terms used in texts of the training collection were examined, in order to find important references to events, persons or objects. Software tools were used in this last task.

After that, the rules for each concept were defined through the same process. Additional software tools were used to examine the context of words and terms. Special attention was given to synonyms, which could be identified analysing the documents with help of a Webster's dictionary. These choices are explained by Loh et al (2000). Two professionals helped in the process, taking approximately 30 hours at all, during a 2 months period. The final decision is due to these professionals.

Below some examples of the used concepts and their definitions are showed (each rule is identified by a roman number; the symbol '$' indicates a radical and '–' indicates a negative word):

- "*alcoholism*":
  (i) alcohol$ (ii) ethilic (iii) drink (iv) drunk (v) drank; etc.
- "*inappetence*":
  (i) eat not much (ii) feed badly (iii) fed not much; etc.
- "*homicide*":
  (i) kill$ –himself –herself (ii) homicid$; etc.
- "*relatives*":
  (i) mother (ii) father (iii) brother$ (iv) sister$ (v) uncle; etc.

4.2 ANALYSES

The mining process was oriented to analyse the distribution of the concepts in the whole collection to identify which concepts are the most frequent. Assuming that the collection is a representative sample of all records in the Hospital, we can make predictions about new patients or use this knowledge for epidemiological studies.

In addition, concept distributions for the four classes (representing diseases) were also compared, looking for similar and very different values.

Finally, using additional software tools, the collection was separated by medicine or drug administration. Analysing the concept distributions inside each sub-collection, it was possible to identify which concepts were dominant and thus to infer the symptoms and signals for which the medicines/drugs are indicated.

For the training collection, the time for categorisation (only the identification of concepts in the texts) took about 1 hour and 20 minutes in a Pentium II 400 MHz with 64 Mbytes of RAM (a comparison of 200 texts against 65 concepts). The mining process took about 15 minutes.

4.3 DISCOVERED KNOWLEDGE

Analysis of the whole collection

- Most frequent concepts (above 50%): relatives (84.5%), aggressiveness (77%), inappetence (76%), medicines (74.5%), insomnia (71%), thought deficit (70.5%), nervousness (68.5%), attention deficit (54.5%)

- Interesting observations:

a) "*readmission*" = 33.0% (meaning 1/3 of the internal patients are readmitted);
b) 84.5% of patients have *relatives*;
c) "*aggressiveness*" is the most frequent symptom in the collection.

Analysis by sub-collection (disease)

Comparing the concept distributions among the four classes in the training collection, some patterns arose. Only concepts with very different distributions were considered as interesting patterns, since similar distributions do not help in discriminating different classes. No knowledge from experts was used to select interesting patterns.

Then these patterns were tested in the second collection. The assumption is that a pattern that appears in both collections is more probable to be correct. Below, the patterns verified in both collections are presented: (when not indicated, percentages follow the order: *affective, schizophrenia, organic, substances*)

1- all the concepts appear in more than one class, except "*poison*", which only appears in *schizophrenia* but with a small frequency;
2- "*attention deficit*" is more frequent in the *affective* class;
3- "*suicidal*" is more frequent in *affective* (81.5% against 38.8%, 16.7% and 30.8%)
4- "*depression*" appears more in *affective* (74.1% against 11.7%, 11.1% and 25%)
5- *affective* and *substances* are very similar (similar frequencies for "*insomnia*", "*inappetence*", "*nervousness*", "*aggressiveness*"), except for "*alcoholism*" (high in the latter and low in the former) and for "*depression*", "*suicidal*" and "*crying*" (on the contrary)
6- "*autism*" appears only in *schizophrenia* (37.9%) and in *organic* (16.7%)
7- "*alcoholism*" appears more in *substances* (94.2% against 25.9%, 16.5% and 11.1%)
8- "*normal consciousness*" and "*clouding of consciousness*" had similar distributions in *substances* (17.3%) and in *organic* (11.1%), while in *affective* and in *schizophrenia*,

"*normal consciousness*" is more frequent than "*clouding of consciousness*" (40.7% x 7.4%; 32% x 13.6%)

9- references to "*death*" are lower in *substances* (17.3% against 40.7%, 35% and 33.3%)

10- "*negativism*" has low frequency in *substances* (17.3% against 29.6%, 38.8% and 38.9%)

11- "*insights*" and "*animals*" (*zoopsia*) do not appear in *organic*

12- "*injuries*" are higher in *organic* (38.9% against 18.5%, 13.6% and 21.2%)

13- "*living alone*" does not appear in *organic* and it is low in the others (14.8%, 8.7% and 3.8%)

14- "*marriage*", "*husband*" and "*wife*" do not appear in *organic*

15- "*puerile*" does not appear in *substances*

16- "*mania*" does not appear in organic and is low in *affective* and *substances* (7.4% and 5.8%)

17- "*dromomania*" does not appear in *organic* and *substances* and is low in *affective* (7.4%)

18- "*trembling*" does not appear in *schizophrenia* and *organic*, is high in *substances* (40.4%) and low in *affective* (7.4%)

19- "*tobacco use*" is not cited in *organic*

20- "*delirium*" does not appear in *affective*

21- concept distributions in "*readmission*" sub-collection (records of readmitted patients) are similar to the whole collection

22- other concepts did not showed an interesting pattern

Associative rules (by diagnosis)

Using a confidence threshold of 80% and a support threshold equals to 40%, one set of associative rules was discovered for each class. The sets were compared to find the common rules (true for all diseases) and the exclusive rules (which appear in only one disease). Following, some associative rules per group are presented:

Common Rules:

attention deficit ➔ relatives
attention deficit ➔ thought deficit
inappetence ➔ relatives
insomnia ➔ relatives
thought deficit ➔ relatives
medicines ➔ relatives

Substances:

aggressiveness ➔ inappetence
thought deficit ➔ inappetence
work/job ➔ inappetence

Schizophrenia:

voices ➔ aggressiveness
persecution ➔ insomnia
voices ➔ insomnia
voices ➔ nervousness
persecution ➔ thought deficit
voices ➔ thought deficit

Organic:

injuries ➔ aggressiveness
injuries ➔ nervousness
negativism ➔ aggressiveness

Affective:

inappetence ➔ suicidal
insomnia ➔ suicidal
thought deficit ➔ suicidal

Analysis of drugs/medicines (an example)

Following, concept distributions for the medicine Dienpax are presented (percentages indicate how frequent is the concept in the set of records where this medicine appears): *inappetence* (91.8%), *aggressiveness* (83.7%), *thought deficit* (78.3%), *nervousness* (75.6%), *insomnia* (64.8%), *alcoholism* (62.1%), *voices* (59.4%).

5. RESULTS EVALUATION

It was necessary to evaluate whether the discovered knowledge was true. First the evaluation of the categorisation process was carried out, intending to determine the error rate in identifying concepts inside the texts. If the error was too great, that would mislead the mining process.

Second, the discovered knowledge needed to be validated against the real expertise about the domain. Two approaches were used for this validation: one subjective and other objective. The subjective validation consisted in the presentation of the results to psychiatrists in order to obtain expert feedback. For the objective validation, an automatic system was constructed for determining the disease of test cases representing patients without diagnosis. The discovered knowledge was used in the decision algorithm of this system. The evaluation was to compare the diagnosis indicated by the automatic system against the one predetermined by physicians in the test collection.

5.1 CATEGORISATION PROCESS EVALUATION

A sample of 50 texts extracted from the test collection was examined to evaluate the concept identification. For this evaluation, twelve concepts were selected: those more prone to errors (with complex rules). Recall and precision values were calculated using *microaveraging* and *macroaveraging* measures of Lewis (1991). *Microaveraging* considers the whole collection as a unique class and *macroaveraging* first calculates precision and recall inside each class and then extracts the average value for the entire collection.

The results were:

- *microaveraging* precision = 90%
- *microaveraging* recall = 93%
- *macroaveraging* precision = 89%
- *macroaveraging* recall = 92%

An average error of 10% may be considered a good result. The improvement of these rates is possible by analysing samples of texts with false hits (causing low precision) or the disregarded texts (low recall) and then refining the categorisation rules. The results described above were obtained in a final round, after improving concept definitions. In a previous

process, the results generated low values, respectively 75%, 87%, 71% and 87%. This proves that it is possible to minimise the error, refining the rules for concept identification.

However, a special attention must be given to errors in each concept. For example, the concept "*depression*" had the worst precision (73%) and the concept "*death*" had the worst recall value (73%). These results put in doubt the extracted knowledge concerning these concepts.

5.2 DISCOVERED KNOWLEDGE EVALUATION

The subjective evaluation of the discovered knowledge was done presenting the results to two expert physicians in Psychiatry. The response was that the knowledge is very similar to that used in real processes for diagnosis. This feedback was enough to consider reliable the results of this experiment.

A final objective evaluation was carried out. The discovered knowledge was used in an automatic classification system for identifying the disease of the 200 test texts (diagnosis process). Different classification methods were experimented, for example:

a) using as class descriptors the concept distribution of each class;
b) using the least frequent concepts in each class and their distributions as weights;
c) using negative concepts (those that never appear in a class) to discard a diagnosis;
d) using negative concepts with negative weights;
e) using pairs of concepts (according to exclusive associative rules).

The best method, a combination of (b) and (d), achieved the following results:

- *microaveraging* precision = 65%
- *microaveraging* recall = 73%
- *macroaveraging* precision = 61%
- *macroaveraging* recall = 53%

Presenting these results to the same physicians (an average error of 38%), they considered a very good rate, above some human performances.

6. CONCLUDING REMARKS

This paper presented an approach for performing knowledge discovery in texts through the qualitative and quantitative analyses of high-level textual characteristics. The analyses are performed based on concepts instead of words.

The process is well suited for analysing textual documentation present in organisations. Qualitative analysis discovers concepts referenced in the documentation and quantitative analysis allows the examination of concept distributions and relations.

The goal is to discover new and useful knowledge through this examination. The results may be used for supporting and evaluating decision processes and for training professionals.

The paper presented the application of the approach in textual documents from a psychiatric hospital. Concepts represent symptoms, signals and social characteristics of patients and were used to identify diseases.

Subjective and objective evaluations were carried out to validate the discovered knowledge. Results proved that the knowledge is reliable and thus the approach has obtained success. Results from the objective evaluation (with the automatic diagnosis system) demonstrate that the approach may be used for constructing decision support systems.

Assuming that the collection used in the experiments is representative of all patients, it is also possible to predict the characteristics of new patients and perform epidemiological

studies (for example, to identify geographic causes of diseases). However, the sample could be conditioned to some aspects, for example, seasons and external events. Currently, texts form a unique set and have no time associated. A future work is planned to analyse the concept distributions over different time periods (years, seasons).

In the example application, a few concepts correspond to social characteristics, but other concepts may be generated and analysed. Consequently, other medicine areas may also benefit from this approach.

As physicians considered the discovered patterns similar to the knowledge used in their decision processes, the results of the discovery approach may be used for training students and assistant professionals. Furthermore, the patterns concerning concepts instead of words make the knowledge more clear and understandable.

The psychiatric collection is a special case for applying knowledge discovery. First, because the diagnosis process is very complex and some times, it is normal that physicians disagree about the final decision. Also it is usual to occur later corrections in the diagnosis associated to a patient. In the presented application, the texts correspond to the admission record but the associated diagnosis is the more updated one. That is, more information (not available in the first record) may have been used to make the final decision.

Special attention should be given to some aspects of the approach. First, the presence of a concept inside a text is conditioned to the concept interpretation. For example, when the concept *"alcoholism"* is identified in a text, the correct interpretation is that the concept is cited in the text but the cause may be doubtful. In order to avoid errors, as when the concept is cited because someone in the family uses alcohol, it is necessary to carefully define the rules for each concept.

Regarding the errors in the process, it is possible to minimise the rates, but perhaps never eliminate them at all. However, the error rate can be controlled and the results may be interpreted under a certain degree of reliability.

Other remark is that the language may change along the time (Chen, 1994). The way in which the vocabulary is used may vary according to people and situations. Furthermore, when the own world evolves, concepts and languages also evolve to accommodate the changes. Therefore, the definition of concepts is conditioned by this context and the analyses should be performed under these restrictions.

This work have used two mining techniques (key-concept listing and associative), since they are the most used in knowledge discovery. However, other techniques can be used in the mining process after the extraction of concepts (categorisation task). The next step in this work is to apply the time series technique over textual records describing the evolution of the patient. Analysing the sequence of records, it is possible to find associations between concepts along the time, for example, discovering concepts that appear immediately after some drug administration or some time after a certain symptom registration.

ACKNOWLEDGEMENTS

This research is partially sponsored by: CNPq (Brazilian Council for Scientific and Technological Development) and CAPES. Medical records were provided by Olivé Leite Psychiatric Hospital (Pelotas, RS, Brazil) and have being produced with research support by FIDEPS (Found of Incentive for the Development of Teaching and Research in Health - Ministry of Health, Brazil).

REFERENCES

Brazilian Center for Disease Classification. (1993) *International Classification of Diseases and Health Related Problems in Portuguese*. 10th revision. São Paulo: EDUSP (in collaboration with the World Health Organisation).

Chen, H. (1994). The vocabulary problem in collaboration. *IEEE Computer, special issue on CSCW*, 27(5), p.2-10.

Chen, H. et al. (1994). Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 37(10), p.56-73.

Chen, H. et al. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. *Journal of the American Society for Information Science*, 48(1), p.17-31.

Chinchor, N. et al. (1993). Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). *Computational Linguistics*, 19(3), p.409-449.

Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1), p.80-91.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation*, 45(4), p.273-301.

Feldman, R. & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In: Fayyad, Usama & Uthurusamy, Ramasamy, eds. *Proceedings of the 1st International Conference on Knowledge Discovery* (KDD-95). Cambridge: AAAI/MIT Press, p.112-117.

Feldman, R. & Dagan, I. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems*, 10(3), p.281-300.

Frawley, W. J. et al. (1991). Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G. & Frawley, W.J., eds. *Knowledge discovery in databases*. Menlo Park: AAAI/MIT Press. 1991, p.1-30.

Furnas, G.W. et al. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), p.964-971.

Gaizauskas, R. & Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1), p.70-105.

Ingargiola, G. (1996). Building classification models: ID3 and C4.5. http://www.cis.temple.edu/~ingargiola/cis587/readings/id3-c45.html. (visited August 2000).

Lewis, D. D. (1991). Evaluating text categorization. In: *Proceedings of the Speech and Natural Language Workshop*. San Mateo: Morgan Kaufmann, p.312-318. http://www.research.att.com/~lewis (visited March 2000).

Lin, C.H. and Chen, H. (1996). An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. *IEEE Transactions on Systems, Man and* Cybernetics, 26(1), p.1-14. http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html (visited November 1999).

Lin, S. H. et al. (1998). Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: Croft, W. B. et al, eds. *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR-98). New York: ACM Press, p.241-249.

Loh, S. et al. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations*, 2(1), p.29-39.

Sowa, J.F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove: Brooks/Cole Publishing Co.

Wiener, E.D. et al. (1995). A neural network approach to topic spotting. In: 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), Las Vegas. http://www.stern.nyu.edu/~aweigend/Research/Papers/TextCategorization (visited March 2000).

Artigo 4 – ISKMDM 2000

"Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva"
Co-autores: Leandro Krug Wives, José Palazzo M. de Oliveira
International Symposium on Knowledge Management / Document Management
Proceedings, pp.125-147
PUC-PR, Editora Universitária Champagnat
Eds: Edouard Lethelier, Flávio Bortolozzi, Kival Chaves Weber, Heitor Pereira
Curitiba, Brasil, 26-29 de Novembro de 2000

DESCOBERTA PROATIVA DE CONHECIMENTO EM TEXTOS: APLICAÇÕES EM INTELIGÊNCIA COMPETITIVA

Stanley Loh[1,2,3], Leandro Krug Wives[1], José Palazzo M. de Oliveira[1]

sloh@zaz.com.br
wives@inf.ufrgs.br
palazzo@inf.ufrgs.br

1- Programa de Pós-Graduação em Computação (PPGC)
Universidade Federal do Rio Grande do Sul (UFRGS)
Av. B. Gonçalves, 9500 Bl. IV, Prédio 43412 - Campus do Vale
Porto Alegre - RS - Brasil

2- Universidade Católica de Pelotas (UCPEL)
Escola de Informática
Rua Félix da Cunha, 412
Pelotas – RS - Brasil

3- Universidade Luterana do Brasil (ULBRA)
Departamento de Informática
Rua Miguel Tostes, 101
Canoas – RS - Brasil

Resumo

O imenso volume de documentos textuais armazenados em meios eletrônicos e disponíveis em repositórios públicos ou internamente em uma organização são fontes importantes de conhecimento. As áreas de Inteligência Competitiva e Business Intelligence procuram explorar estas informações para descobrir conhecimento novo e útil. Para auxiliar os processos de "garimpagem" de informações em textos, surgiram técnicas e ferramentas para descoberta de conhecimento em textos. Neste artigo, é apresentada uma estratégia de descoberta de conhecimento em textos para auxiliar os trabalhos de análise em Inteligência Competitiva. A estratégia apresentada está voltada para processos proativos. Nesta classe de sistemas, encontram-se aqueles que iniciam a busca sem hipóteses ou sem um objetivo completamente definido e que contam com a ação efetiva do analista interagindo com o sistema computacional. Neste tipo de processo, o usuário tem problemas que quer resolver mas não sabe exatamente por onde começar ou que tipo de conhecimento lhe pode ser útil. Neste artigo são apresentados exemplos de aplicação utilizando a estratégia proposta e discutidos aspectos que podem influenciar o processo de descoberta.

Palavras-chave:
Descoberta de Conhecimento em Textos, Text Mining, Ferramentas para Inteligência Competitiva

1 INTRODUÇÃO

Com o crescente uso de computadores interligados pela Internet, um grande número de documentos eletrônicos estão sendo armazenados. Em sua grande maioria, estes documentos contêm informações codificadas em forma textual, tais como dicionários, manuais, enciclopédias, guias, mensagens de correio eletrônico e páginas da Web. A necessidade não atendida é como aproveitar o conhecimento humano disponível nestes documentos textuais para inferir conhecimento novo e útil [11] [3]. Em especial, a área de Inteligência Competitiva (IC) trata este tipo de problema.

A Inteligência Competitiva (IC) é uma área que busca suprir as necessidades de informação estratégica de uma empresa [8]. No nível atual de globalização e de informatização, as informações do ambiente interno e externo de uma empresa tornam-se extremamente valiosas. As informações pertinentes ao ambiente interno de uma empresa são todas aquelas relacionadas com os seus produtos, vendas, serviços, estoques, empregados e fornecedores. Já o ambiente externo pode ser compreendido como o mercado em si e suas tendências, novas tecnologias de produção, a opinião e a satisfação dos clientes em relação à empresa e às ações da concorrência.

Quando disponíveis publicamente, essas informações podem ser analisadas por qualquer pessoa de forma lícita. Portanto, coletar e analisar essas informações antes que os concorrentes o façam é extremamente importante para que uma empresa mantenha sua posição no mercado. Torna-se cada vez mais obrigatória a utilização de métodos e ferramentas de inteligência competitiva, correndo-se o risco de se perder posição mercadológica, caso a empresa não as utilize.

A prática da inteligência possibilita aos empresários estarem sempre bem informados e preparados para minimizar riscos, antecipar crises e tornar seus produtos mais competitivos. Por outro lado, por existir

uma quantidade muito grande de informações textuais disponíveis sobre o ambiente externo de uma empresa, aliado ao fato de essa informação não ser facilmente tratável e analisável, o conhecimento necessário à tomada de decisão e ao posicionamento estratégico de uma empresa não é facilmente identificado.

Neste sentido, a área de inteligência busca suprir suas deficiências adotando técnicas provenientes da área de recuperação de informações, extração de informações e descoberta de conhecimento em textos.

A área de Recuperação de Informação – RI – (*Information Retrieval*) tem por objetivo encontrar documentos que contenham informações relevantes às necessidades definidas por um usuário em uma consulta [28]. Entretanto, neste caso, o usuário necessita examinar os documentos resultantes para encontrar informação, o que é uma tarefa demorada. Já a área de Extração de Informação – EI – (*Information Extraction*) estuda metodologias, técnicas e sistemas que possam encontrar dados específicos dentro de textos. Tais sistemas extraem automaticamente valores de atributos, tais como campos de um banco de dados. Infelizmente, em geral, tais sistemas são muito dependentes do domínio, isto é, só apresentam bom desempenho com certas classes de documentos [10]. Além disto, para criar tais sistemas é necessário desenvolver muita engenharia de conhecimento, examinando amostras de textos para identificar como a informação é codificada em frases da língua natural [6].

A partir destas dificuldades, surgiu a área de Descoberta de Conhecimento em Textos (KDT – *Knowledge Discovery in Texts*) [12]. No âmbito comercial, esta área é conhecida como *Text Mining*. Nas propostas iniciais (por exemplo, em [12], [13], [14] e [18]), a estratégia básica de KDT era orientada para a aplicação da técnica de associação ou correlação (comum em *Data Mining*) sobre características extraídas do texto (em geral, termos usados no texto ou palavras-chave associadas ao texto). Pode-se entender a área de KDT como a aplicação de técnicas e ferramentas computacionais com o objetivo de auxiliar na busca de conhecimento novo e útil disponível em coleções textuais.

Neste artigo, é apresentada uma estratégia de KDT para auxiliar trabalhos de Inteligência Competitiva na análise de coleções textuais. A estratégia apresentada está voltada para processos proativos. Na seção 2, é apresentada resumidamente a área de IC e suas estratégias. A seção 3 apresenta as principais técnicas de KDT. Já na seção 4, é apresentada uma estratégia para descoberta proativa de conhecimento. Na seção 5, serão apresentados exemplos de aplicação desta estratégia em problemas relativos a IC e *Business Intelligence.* Os aspectos que podem influenciar o processo de descoberta são discutidos na seção 6.

2 INTELIGÊNCIA COMPETITIVA

Os objetivos da inteligência competitiva são: promover o saber-fazer tecnológico e científico de uma empresa, país ou região; detectar riscos e oportunidades no mercado exterior e interior; monitorar as ações dos concorrentes (modos de pensar, técnicas, cultura, intenções e capacidades) e definir estratégias e ações que devem ser tomadas a fim manter sua estabilidade [8].

Os métodos de inteligência que são utilizados atualmente na área de administração surgiram durante a segunda guerra mundial e foram muito utilizados e aperfeiçoados durante a guerra fria. Com o fim da guerra fria, as pessoas que trabalhavam nos departamentos de defesa passaram a ser contratados pelas grandes empresas para trabalhar em seus departamentos de pesquisa, desenvolvimento e marketing.

Basicamente, a atividade de inteligência consiste na aplicação de métodos de vigília do ambiente externo de uma empresa, buscando monitorar as atividades dos concorrentes, identificar novos produtos e tecnologias que possam ser aproveitados pela empresa. Grande parte das informações necessárias ao processo de inteligência pode ser obtida em fontes públicas (órgãos públicos e na própria Internet) [23]. Com isso, não há o risco de se ser acusado de espionagem inicial (mas, há o risco de um concorrente também coletar informações).

2.1 ETAPAS DO PROCESSO DE IC

As etapas do processo de inteligência podem variar de empresa para empresa. Tudo depende de sua situação atual. Caso a empresa já conheça suas necessidades de informação ela pode partir diretamente para a coleta e monitoração da informação, caso contrário, torna-se necessário identificar essas necessidades e, muito provavelmente, remodelar os sistemas de informação da empresa.

Clerc comenta que o processo de inteligência inicia pela etapa de *coleta* de dados, passando pela *exploração*, *distribuição* e finalizando com a aplicação de mecanismos de *segurança* das informações e conhecimentos descobertos [8]. Já Zanasi aborda a questão da identificação das necessidades de informação, identificando as etapas de *compreensão do problema*, *definição de fontes*, *identificação de dados relevantes* (*pesquisa estratégica*), *análise* e *interpretação* [33].

Combinando-se as propostas, pode-se entender o processo de IC como tendo as seguintes etapas:

a) Identificação da necessidade de informação: nessa etapa deve-se identificar quais são as necessidades de informação de cada pessoa na empresa (principalmente dos tomadores de decisão), quais delas a própria empresa pode *sanar* e quais necessitam de dados externos;
b) Identificação e análise de fontes de informação: uma vez identificadas as necessidades de informação, torna-se necessário identificar onde essas informações podem ser recuperadas (as fontes). Essas fontes podem ser internas ou externas. No caso de fontes externas, deve-se descobrir o formato, o tempo de acesso e o custo das informações, assim como deve ser identificado como elas podem ser agregadas às informações existentes na empresa;
c) Coleta: é a busca, em si, da informação ou dos dados nas fontes identificadas;
d) Filtragem: devido à grande quantidade de dados e informações que podem ser coletadas, muitas podem não ser específicas às necessidades identificadas inicialmente. As informações irrelevantes devem ser descartadas e as relevantes selecionadas;
e) Distribuição: os dados ou informações selecionadas devem ser encaminhadas às pessoas que as necessitam (que expressaram sua necessidade);
f) Exploração: corresponde à transformação dos dados em informação e conhecimento. Para tanto podem e devem ser utilizadas ferramentas computacionais e métodos estatísticos de análise;
g) Segurança: após adquiridos os conhecimentos e informações, estes devem ser, obviamente, postos em prática (utilizados na tomada de decisão) e armazenados em algum local seguro para que não caiam nas mãos dos concorrentes.

Nessa metodologia sugerida, espera-se que a pessoa que recebeu a informação ou os dados saiba tratá-las, ou seja, é ela quem deve realizar a etapa de *exploração*. Eventualmente, essa etapa pode vir antes da distribuição, onde os dados seriam então analisados por um especialista ou departamento de inteligência e seus resultados é que seriam repassados para o tomador de decisão.

Uma consideração importante é a de que as fontes (os ambientes) devem ser constantemente monitoradas, mesmo que as necessidades de informações sejam sanadas. Isso porque toda vantagem competitiva é momentânea: uma mesma tecnologia está igualmente acessível a todos, não sendo possível sustentar vantagem competitiva por muito tempo [31]. Novos processos e produtos surgem a todo o momento e as necessidades de informação podem mudar. Logo, o processo de inteligência deve ser executado constantemente.

Dependendo da empresa, a atividade de inteligência pode requerer um departamento específico. Em outras, ela pode ser realizada por pessoas que, de alguma forma, já trabalhem com a análise de informações (CPD ou departamento de marketing ou pesquisa e desenvolvimento). Há ainda a possibilidade de se contratar uma empresa para realizar uma pesquisa específica de inteligência [33]. Em qualquer um destes casos é extremamente importante que a empresa tenha consciência do papel da pessoa ou departamento encarregado de obter e analisar as informações obtidas. A inteligência envolve, muitas vezes, uma mudança estratégica da empresa (a mudança de mercado, por exemplo). Se a diretoria da empresa não tiver plena consciência do poder dessa atividade, e confiança na pessoa que a faz, a mudança estratégica necessária pode nunca ocorrer.

3 TÉCNICAS PARA KDT

Existem muitas técnicas e ferramentas para suportar KDT. A técnica mais básica é a recuperação de informações (RI), cujo objetivo é encontrar textos que podem conter determinada informação. Métodos para RI são discutidos em [28] e [25]. Uma técnica similar é a recuperação de passagens, que aplica as mesmas técnicas de RI só que sobre partes do texto [4] [16].

Já a técnica de extração de informação (EI) procura valores de atributos dentro dos textos [9]. A técnica de sumarização tem por objetivo extrair resumos de um texto ou de uma coleção, podendo ser uma visão geral ou as partes mais importantes ou mais interessantes [28] [19] [5]. A técnica de listagem de conceitos-chave (*key-concept listing*), por sua vez, analisa uma coleção de textos em busca de características comuns (palavras, palavras-chave, temas), formando o que se convencionou chamar, neste artigo, de *centróide*. A partir desta técnica pode-se fazer o inverso, isto é, descobrir diferenças comparando textos ou coleções; este processo é denominado técnica da diferença.

Já a técnica de agrupamento (*clustering*) é um pouco mais complexa. Ela é utilizada para identificar automaticamente, sem intervenção humana, grupos de textos similares [32]. Sua principal utilidade é permitir encontrar características comuns em subgrupos quando não há nada em comum na coleção toda. Há também a técnica de classificação ou categorização, que procura encontrar temas ou assuntos no conteúdo dos textos (do que os textos estão tratando). A técnica de associação (ou correlação) descobre relações de dependência entre textos ou características dos textos. Existem, também, técnicas para

visualização de resultados, que ajudam o usuário a entender melhor o conhecimento descoberto (ver [29], [2] e [26]).

Apesar de apresentadas em separado, nada impede que as técnicas sejam utilizadas de modo integrado, uma após a outra, de forma que a saída de uma técnica seja a entrada da seguinte. Por exemplo, Moens e Uyttendaele usam a técnica de sumarização associada com EI, para criar resumos de casos jurídicos [20]. Moscarola apresenta uma ferramenta que integra diversas técnicas para KDT [21] [22].

4 ESTRATÉGIA PROATIVA PARA KDT EM IC

Em geral, o processo de descoberta de conhecimento segue algumas etapas, muito parecidas com as etapas do processo de inteligência. Estas etapas são, segundo Goebel e Gruenwald [15]:

a) entendimento do domínio de aplicação e definição do objetivo do processo de descoberta;
b) aquisição ou seleção do conjunto de dados;
c) integração e verificação do conjunto;
d) limpeza dos dados (pré-procesamento e transformação);
e) desenvolvimento de um modelo inicial ou construção de hipóteses iniciais;
f) escolha e aplicação de métodos de mineração (*mining*);
g) visualização e interpretação dos resultados;
h) teste e validação das hipóteses (pode-se refazer parte do processo);
i) uso e manutenção do conhecimento descoberto (tomada de decisão no domínio).

Existem alguns estudos que sugerem a aplicação de técnicas e ferramentas de KDT em problemas de Inteligência Competitiva. Moscarola e outros [21] [22], por exemplo, usam as técnicas para descobrir táticas de empresas concorrentes analisando patentes. Entretanto, tais trabalhos não propõem estratégias de como o usuário deve encaminhar o processo de descoberta. A seleção das técnicas e ferramentas depende da experiência desenvolvida pelo usuário em processos de descoberta com a utilização das ferramentas propostas.

Já Watts [30] apresenta estratégias para análise de tecnologias com auxílio de técnicas de KDT. Medidas bibliométricas (estatísticas sobre textos) são utilizadas para encontrar informações específicas deste problema, ou seja, para identificar o ciclo de vida de certas tecnologias, entidades envolvidas, mudanças no tempo, etc. Neste caso, as estratégias propostas são específicas para um certo tipo de problema e requerem do usuário experiência em atividades de Inteligência Competitiva (o que e como procurar).

O presente artigo diferencia-se dos anteriormente apresentados por descrever uma estratégia proativa para descoberta de conhecimento em textos. De acordo com Choudhury e Sampler [7], existem dois modos de aquisição de informação: o modo reativo e o modo proativo. No primeiro caso, a informação é adquirida para resolver um problema específico do usuário. Nestes casos, o usuário sabe o que quer e tem idéia de como a solução para o problema pode apresentar-se. Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório, para detectar problemas potenciais ou oportunidades. Neste segundo caso, o usuário não tem um objetivo específico ou bem definido mas desenvolve ações específicas utilizando as ferramentas disponíveis.

No modo reativo, o usuário tem uma idéia, mesmo que vaga, do que pode ser a solução ou, pelo menos, de onde pode-se encontrá-la. É possível afirmar, então, que o usuário possui algumas hipóteses iniciais, que ajudarão a direcionar o processo de descoberta. Isto exige entender o interesse ou objetivo do usuário para limitar o espaço de busca (na entrada) ou filtrar os resultados (na saída). É o caso típico da busca de uma informação específica, por exemplo, um valor para um atributo ou um processo (conjunto de passos) para resolver um problema.

Já na abordagem proativa, há um problema ou objetivo, mas o usuário não consegue definir o que precisa para resolver o problema. O objetivo típico do modo proativo poderia ser representado pela expressão: *"diga-me o que há de interessante nesta coleção"*. Neste caso, o usuário não tem de forma definida o que lhe seja de interesse (o que precisa), podendo tal interesse mudar durante o processo. Pode-se dizer que é um processo exploratório, sendo, iterativo (com retroalimentação) e interativo (com ativa participação e intervenção do usuário).

As diferenças em relação ao modelo de Goebel são que:

- o passo (a) pode não ter uma definição precisa do objetivo do processo (há um objetivo ou problema, mas a solução não pode ser prevista);
- o passo (e) não existe na abordagem proativa, ou seja, o usuário não sabe ou não deseja formular hipóteses para a solução de seu problema.

As ferramentas de KDT ajudarão a levantar hipóteses para a solução do problema, as quais serão exploradas, investigadas e testadas durante o processo. Em geral, a falta de hipóteses iniciais se dá porque o usuário não consegue definir exatamente o que está procurando. Ele sabe que tem um problema, mas não tem uma idéia exato do que pode ser a solução. É o caso típico de monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo, talvez sendo necessário avaliar as hipóteses levantadas, para verificar se são verdadeiras ou não.

Um dos problemas do paradigma proativo é definir um plano de uso das técnicas ou de como a coleção textual deverá ser investigada de forma automática pelas ferramentas, a fim de serem descobertas hipóteses. Kuhlthau [17] determinou seis fases em processos de descoberta de informação: iniciação, seleção, exploração, formulação, coleção e apresentação. Cada fase é caracterizada por atitudes diferentes do usuário (por exemplo, em relação a sentimentos, pensamento, ações e tarefas). Uma das descobertas mais interessantes desta pesquisadora é que o usuário inicia procurando algum tipo de conhecimento mais geral, depois ele procura informação relevante em grupos mais restritos e termina procurando informações mais focadas ou específicas. Durante este processo, o usuário reconhece, identifica, investiga, formula, reúne e complementa o conhecimento.

Sugere-se então uma estratégia para descoberta proativa de conhecimento em textos. Não se pode considerar esta estratégia uma metodologia, mas sim um esboço (*framework*), que poderá conduzir os usuários no processo, indicando os passos principais (técnicas ou ferramentas a serem usadas). Os passos são resumidamente descritos a seguir:

1) seleção de textos: o primeiro passo é selecionar uma coleção de textos sobre os quais serão aplicadas as técnicas; as técnicas automáticas mais indicadas são a recuperação de informação (que encontra textos procurando por palavras-chave ou termos presentes nos textos) e a classificação (que separa textos por assunto); outra possibilidade, é o usuário mesmo encontrar ou selecionar os textos, o que demanda mais trabalho manual;
2) análise da coleção toda ou de partes da coleção: neste ponto, o usuário deve decidir se irá aplicar as técnicas de descoberta sobre todos os textos ou sobre partes da coleção; a sugestão é que se comece analisando toda a coleção e depois se examine subcoleções. Em alguns casos, nada de interessante é encontrado na coleção toda, o que leva o usuário, necessariamente, a investigar pequenas subcoleções. A separação em subcoleções pode ser feita de forma automática, com a técnica de agrupamento, ou por algum critério estabelecido pelo usuário;
3) análise de grupos de textos (toda a coleção ou partes): uma boa maneira de começar a análise é extraindo uma lista de termos comuns a todos os textos ou que aparecem em mais de um (técnica de listagem de conceitos-chave ou centróide); a técnica de diferença pode ser usada depois para levantar novas hipóteses; por fim, a técnica de associação, mesmo que demorada, pode ajudar a descobrir algo interessante;
4) comparação de sub-coleções entre si ou em relação à coleção toda: os resultados conseguidos com as técnicas de listagem de centróide, diferença e associação aplicadas a cada grupo particular podem ser comparados entre si ou com os resultados obtidos com a coleção toda;
5) validação de hipóteses: em geral, a técnica de resumos traz bons resultados, pois possibilita ao usuário ler as frases mais significativas e interpretar os resultados;
6) retroalimentação: como o processo é cíclico, os passos ou o processo todo podem ser refeitos.

Apesar de que a estratégia possa ser começada sem que o usuário defina hipóteses iniciais, a intervenção humana é necessária. Por exemplo, o primeiro passo do processo, obrigatoriamente, precisa da intervenção do usuário para selecionar os textos da coleção, seja de forma manual ou fornecendo parâmetros para as ferramentas de recuperação. Também será necessário que o usuário interprete os resultados no contexto da realidade para que as descobertas sejam úteis. Segundo Aamodt and Nygard [1], o conhecimento é imprescindível para que os dados possam ser interpretados e se tornem informação. O conhecimento é subjetivo e depende das pessoas. Por isto, Moscarola and Bolden [22] sugerem o modelo construtivista ao invés do positivista para os processos de descoberta, ou seja, o processo deve ser guiado pelo usuário.

5 EXEMPLOS DE APLICAÇÃO DE KDT EM IC

A seguir são apresentados exemplos de uso da estratégia proativa em problemas de Inteligência Competitiva (IC). Os dois primeiros ajudam a entender como técnicas de KDT podem ser empregadas em processos proativos, isto é, quando não há um objetivo muito definido, mas quando sim se deseja explorar alguma coleção textual em busca de conhecimento novo e útil. Estes exemplos demonstram a viabilidade

de realizar o processo de descoberta sem que haja hipóteses iniciais. O terceiro exemplo é apresentado para demonstrar que a estratégia proativa pode ser utilizada em conjunto com processos reativos, com a finalidade de realizar o reconhecimento do domínio e do estilo de escrita utilizada.

5.1 ESTRATÉGIA AGROALIMENTAR PARA O MERCOSUL

Neste exemplo, o objetivo era descobrir algo de interessante numa coleção de textos sobre potencialidades e oportunidades agroalimentares no MERCOSUL. Os textos foram retirados do Seminário sobre Estratégia Agroalimentar para o Mercosul [27], onde foram discutidas potencialidades, desafios, demandas e oportunidades no setor agrolimentar envolvendo países do Mercosul e associados. A coleção é formada por 6 textos, cada um contendo informações sobre um único país. Cada texto é apresentado por um representante do governo em questão. Participaram representantes dos seguintes países: Argentina, Bolívia, Brasil, Chile, Paraguai e Uruguai.

O processo proativo começou com a aplicação das técnicas de centróide e diferença sobre palavras dos textos, ou seja, procurando encontrar que termos apareciam em mais de um texto e que termos apareciam em somente um texto. Com isto, pode-se levantar hipóteses sobre a existência de temas comuns aos países (por exemplo, preocupações ou estratégias comuns) ou então de temas exclusivos a de um único país (por exemplo, demandas ou estratégias particulares de um país).

Para examinar o significado dos termos, usou-se depois a técnica de resumos, a qual possibilitou confirmar ou refutar as hipóteses levantadas. Termos sinônimos e variações léxicas também foram usados para que o assunto ou tema fosse recuperado de forma completa.

A seguir são discutidos alguns conhecimentos novos resultantes deste processo. Primeiro são apresentados os resultados da técnica de centróide. Cada item apresenta o termo descoberto (e sua variação) e o número de textos nos quais o termo aparece. Os resultados da técnica de diferença são apresentados por país. Ao final de cada item, é apresentada a interpretação dada ao padrão descoberto (termo comum ou exclusivo), a qual foi extraída a partir da leitura dos resumos resultantes da aplicação da técnica de resumos. Nem todos os termos são mostrados, mas somente aqueles que chamaram mais à atenção.

Técnica do Centróide (temas comuns)

a) política(s): 6 textos; todos falam na necessidade de políticas comuns;
b) abertura: 6 textos; a abertura é vista como importação ou exportação e é apontada como uma necessidade ou caminho para o crescimento; alguns países já citam benefícios advindos da abertura do mercado, enquanto outros reclamam da falta ou da pouca abertura; mas há precauções a serem tomadas, especialmente olhando exemplos de fracasso; por isto, os países apontam a necessidade de regulamentação para a abertura do mercado;
c) internacional(is): 6 textos; a maioria dos textos aponta o Mercosul como um caminho, um aprendizado para o comércio internacional; outros ainda falam em uma alternativa quando o mercado mundial não está tão bem;
d) crescimento: 6 textos; 5 países afirmam estar experimentando crescimento econômico;
e) qualidade: 5 textos; em 3 países, há a preocupação com produtos de qualidade;
f) NAFTA: 3 textos; Argentina, Chile e Bolivia têm ou desejam ter relações (importação ou exportação) com o NAFTA (mercado comum da América do Norte).
g) outra observação interessante é que os temas “inflação”, “pobreza/pobres” e “desemprego” aparecem em metade dos textos.

Técnica da Diferença (temas exclusivos a um único país)

a) Argentina:

- frutas: apesar de o termo aparecer somente neste texto, a produção frutífera também é citada em outros 2 países (Chile e Paraguai);
- navegação/rio(s): único texto que fala em utilizar a navegação para escoamento de produção;
- óleo: apesar de o termo aparecer somente neste texto, a produção de oleaginosas também acontece na Bolívia;
- inundações: há uma forte preocupação com a estação chuvosa;
- peixes/pescados: Argentina e Chile dizem ter indústrias pesqueiras;
- alguns tipos de produção foram citados somente neste texto, entre eles: mariscos, bebidas, têxteis e fumo.

b) Bolívia:

- multinacionais: o texto aponta a participação de empresas multinacionais no setor;
- genético: apesar do termo aparecer somente neste texto, pesquisas genéticas também são citadas no Paraguai (uma observação significativa é que não se fala em produtos transgênicos em nenhum dos textos);
- carência: a Bolívia aponta a falta de infra-estrutura como um desafio;
- alguns tipos de produção foram citados somente neste texto, entre eles: castanha, café, borracha e lhamas (produção de gado deste tipo).

c) Brasil:
- cerrado: apontando esta região como a mais promissora nos próximos anos.

d) Chile:
- fertilidade: do solo é um problema pela escassez de água;
- capitais estrangeiros: o texto fala que o capital estrangeiro tem chegado de forma significativa.

e) Paraguai:
- artificiais: é apontado o uso de pastagens artificiais.

f) Uruguai:
- agrotóxicos: fala-se na generalização do uso;
- alguns tipos de produção foram citados somente neste texto, entre eles: malte, arroz, cevada e couro.

Posteriormente, alguns participantes do referido seminário foram chamados a validar os resultados encontrados por este processo de KDT. O consenso foi de que os resultados são significativos. Os avaliadores acharam as hipóteses levantadas viáveis e merecedoras de maiores estudos. Além disto, alguns temas indicados pelo processo de KDT foram discutidos no seminário mas não documentados.

5.2 MARKETING POLÍTICO

Este exemplo mostra como as técnicas de KDT podem auxiliar em análises políticas. Neste caso, um processo de descoberta proativa foi conduzido sobre uma coleção de textos extraídos de um jornal *online*, os quais falavam de um determinado governante público. A coleção foi dividida em duas subcoleções de acordo com o ano de publicação das notícias (1997 e 1999).

As técnicas de centróide e diferença, aplicadas de forma isolada em cada subcoleção, permitiram avaliar a participação de certos temas nas duas subcoleções. Depois, a técnica de resumos foi utilizada para investigar cada hipótese levantada. A seguir, são descritas algumas das principais descobertas.

a) o escândalo de corrupção a que o governante foi associado só aparece na subcoleção de 1997, isto é, não é referenciado em 1999; uma interpretação possível é que o escândalo foi esquecido ou resolvido durante o período;
b) o termo "separação" só aparece em 1999; com uso da técnica de resumos, foi possível descobrir que o governante havia se separado de sua esposa no intervalo de tempo;
c) a presença da palavra "eleição" com freqüências diferentes nas duas subcoleções levantou um interesse particular; uma verificação detalhada com a técnica de resumos (atentando-se para termos relacionados tais como "eleitorado", "reeleição", etc.) permitiu concluir que o tema "eleições" aparecia em 33.7% dos textos de 1999 contra 25% em 1997; a conclusão é de que a proximidade das eleições de 2000 tenha influenciado este aumento;
d) os termos "dívida(s)" e "endividamento" aparecem mais em 1997 do que em 1999; interpretações possíveis podem levar às hipóteses de que o tema perdeu interesse na mídia ou de que o problema da dívida foi minimizado durante este período no governo do político analisado;
e) uma descoberta interessante foi a de que o nome de um outro político aparecia associado ao governante nas duas subcoleções, entretanto, com freqüências diferentes (mais em 1997 do que em 1999); um conhecimento extra-coleção sobre o domínio (de que os dois políticos esfriaram suas relações durante este período) permite validar a descoberta.

Os resultados desta aplicação servem apenas para demonstrar a viabilidade da proposta. Entretanto, resultados de processos semelhantes poderiam ser usados para traçar estratégias de ação ou de publicidade, para melhorar a imagem de políticos ou mesmo para melhorar a atuação dos governantes. Também pode-se ter idéia de como as figuras políticas e seus atos estão sendo vistos pela mídia. Isto pode sugerir mudanças no comportamento pessoal ou na forma de atuação do político em seu cargo.

5.3 BENCHMARKING DE FERRAMENTAS DE KDT

Este terceiro exemplo difere dos anteriores por utilizar a estratégia proativa apenas para reconhecimento do domínio de aplicação e do estilo de escrita da coleção textual. Depois, seguiu-se um processo reativo.

Seguindo o trabalho de Goebel [15], procurou-se fazer um *benchmarking* de ferramentas de KDT divulgadas na Internet. Neste caso, havia um objetivo bem definido, que era encontrar as técnicas utilizadas nas ferramentas, sendo portanto um processo reativo. Entretanto, o usuário não tinha idéia de que técnicas estavam sendo utilizadas pelas ferramentas e como estas técnicas estavam sendo referenciadas nos textos (que termos estavam sendo usados para descrevê-las).

Então, um processo proativo foi conduzido inicialmente para estudar a linguagem utilizada na coleção. Aplicando-se a técnica do centróide, foi extraída uma lista de termos utilizados na coleção. Para minimizar o esforço, as chamadas *stopwords* (palavras muito genéricas e pouco significativas, como preposições, artigos, pronomes, foram retiradas num pré-processamento). Depois, o usuário analisou cada termo para poder identificar as técnicas apresentadas.

Para suportar o posterior processo reativo, a técnica utilizada foi a de classificação. Em geral, esta técnica está associada com o problema de encontrar o assunto principal em um texto. Neste exemplo, entretanto, ela foi utilizada para descobrir a presença ou não de uma característica no texto.

Para a aplicação, foram extraídos textos sobre 9 ferramentas, a partir das páginas Web onde as ferramentas são apresentadas e vendidas. Cada técnica foi definida como uma classe, num total de 13 técnicas definidas. Cada classe continha regras para sua identificação nos textos (em geral, um conjunto de palavras a serem analisadas nas frases). A técnica de centróide foi utilizada somente para analisar o tipo de linguagem que era empregada nos textos. Isto permitiu verificar que termos estavam sendo usados para apresentar as técnicas embutidas nas ferramentas.

Os resultados do processo todo (proativo + reativo) permitiram saber que técnicas eram mais comuns e quais estavam sendo menos utilizadas. Também foi possível identificar que apenas duas ferramentas alegavam ter todas as técnicas definidas.

Tais descobertas poderão ser úteis em estratégias de implementação de futuras ferramentas, cabendo aos desenvolvedores decidir se irão privilegiar as técnicas mais usadas ou as menos empregadas e se irão ou não implementar todas as técnicas disponíveis. Isto demonstra a possibilidade de utilizar técnicas proativas de KDT junto com processos reativos. A estratégia proativa permitiu identificar a linguagem e o estilo utilizados nos textos para que os processos reativos pudessem ser corretamente conduzidos, isto é, para que as hipóteses pudessem ser precisamente definidas (neste caso, as hipóteses eram as técnicas).

5.3 OUTRAS APLICAÇÕES

As possibilidades de aplicação de estratégias proativas de KDT são muitas. A princípio, qualquer coleção de documentos textuais pode ser examinada para se descobrir conhecimento novo e útil, sem que seja necessário ter hipóteses iniciais. Em alguns casos, mesmo com pouco conhecimento do domínio ou da linguagem utilizada na coleção, é possível conduzir um processo de descoberta de conhecimento.

A abordagem proposta neste artigo foi utilizada em outros tipos de aplicações, trazendo bons resultados. Alguns casos são referentes a *Business Intelligence*, que poderia ser visto como a aplicação de estratégias de Inteligência Competitiva em documentação interna de uma organização, visando a descoberta de conhecimento disponível no próprio acervo.

A seguir, são apresentados dois casos em que estratégias um pouco diferentes das anteriores foram usadas:

- Em um sistema de *help desk* de uma empresa, os chamados para atendimento do setor de suporte de informática eram registrados como campos textuais de um banco de dados (*memos*). Nestes registros, havia informações sobre o setor que chamou, o tipo de problema apresentado (queixa) e a solução aplicada, indicando o recurso defeituoso (*hardware* ou *software*). Um processo proativo de descoberta foi desempenhado sobre estes registros textuais (gravados como textos separados para cada chamado). Com as técnicas de centróide e diferença, foi possível saber que setores ou recursos mais apresentavam queixas. Tal descoberta permitiu lançar ações preventivas, seja através de melhores controles sobre equipamentos de *hardware* ou através de treinamento do pessoal usuário de informática. Também foi possível identificar um problema interessante, onde o funcionário técnico que mais atendia a chamados de um certo tipo não era quem deveria realizar tal tarefa (outro funcionário estava designado para isto, por ser mais capacitado e haver sido treinado para tal). Esta

descoberta só foi possível extraindo-se um conjunto de textos referentes a um tipo de chamado (texto em que certo termo aparecia) e aplicando-se a técnica do centróide sobre esta subcoleção.

- Em outro caso, ementas de um tribunal de justiça foram analisadas num processo proativo de descoberta. As ementas contêm resumos dos processos (decisões em andamento ou finais), incluindo juiz, partes, a descrição das características do processo e argumentos usados na decisão. A técnica de *clusterização* foi usada para agrupar automaticamente as ementas (não houve intervenção humana nesta separação). Uma análise posterior de cada grupo (*cluster*) resultante foi realizada com as técnicas de centróide e de diferença. Isto permitiu descobrir que a maioria dos processos de um certo tipo estavam sendo julgados pelo mesmo juiz; um padrão que não deveria ser regra. A análise também permitiu descobrir que as decisões tomadas para certos tipos de processos eram muito parecidas (a chamada jurisprudência). Isto ficou claro pelo grande número de termos comuns utilizados em ementas que foram alocadas em um mesmo grupo.

6 CONCLUSÃO E DISCUSSÕES

Este artigo apresentou uma proposta para aplicação de técnicas de descoberta de conhecimento em textos (KDT) em problemas de Inteligência Competitiva (IC). A proposta segue uma estratégia proativa, isto é, defende o uso de técnicas de KDT sem que seja necessário ter precisamente definido um objetivo. Assim, o analista de informações não precisa definir, previamente ao processo, que informações são importantes. A estratégia proativa se contrapõe ao paradigma reativo com a finalidade de tornar a descoberta um processo exploratório, onde o analista poderá descobrir informações novas e úteis durante o processo.

O diferencial desta proposta é que ela libera o analista de definir precisamente um objetivo ou que informações lhe serão úteis. Isto permite às pessoas ganharem tempo no processo de descoberta, iniciando sem hipóteses. Mas a principal vantagem do processo proativo está em que o analista não é influenciado por suas próprias hipóteses iniciais. Sendo o processo iniciado sem um objetivo preciso, as hipóteses que vão sendo levantadas durante o processo podem levar a descobertas novas para o analista, algo em que ele não teria pensado se utilizasse um processo reativo.

Os benefícios dos processos de KDT serão cada vez mais notados com o aumento de documentos eletrônicos disponíveis internamente às empresas, sejam eles criados a partir de documentos em papel ou diretamente em computadores. Além disto, o crescente uso da Internet tem aumentado o volume de informações publicamente acessíveis em páginas Web. Entretanto, somente com o auxílio de ferramentas computacionais, os analistas de IC e *Business Intelligence* poderão fazer uso adequado das informações disponíveis e encontrar eficientemente conhecimento novo e útil.

Nas aplicações realizadas, pôde-se perceber que há certos cuidados que devem ser tomados para que os resultados do processo de descoberta possam ser confiáveis. [24] sugere que problemas podem ocorrer devido a 5 tipos de imperfeições:

- informação incompleta: quando faltam detalhes de informação (por exemplo, atributos sem valores);
- informação imprecisa: devido à diferença de granularidade (por exemplo, datas com referência ao dia ou somente ao mês);
- informação incerta: quando não pode ser provada;
- informação vaga: devido a imprecisões do vocabulário;
- informação inconsistente: por exemplo, quando há valores contraditórios.

Assim, o primeiro cuidado é com a representatividade da coleção textual. Para interpretar os resultados, assume-se que os textos são completos (contêm todas as informações relevantes) e verdadeiros (fiéis à da realidade). Se os textos não foram escritos ou elaborados corretamente ou de forma compatível com a análise sendo feita, o conhecimento descoberto pode não valer para todos os casos. Este cuidado deve ser tomado, por exemplo, na aplicação relativa a textos sobre o Mercosul.

Com relação ainda à representatividade da coleção, num trabalho de Inteligência Competitiva é importante determinar bem o período de tempo sobre o qual será feita a análise. Por exemplo, no caso do marketing político, os textos foram extraídos de períodos diferentes (1997 e 1999). Então, os resultados do processo de descoberta estão condicionados aos eventos e fenômenos que ocorreram próximos a estes períodos. Decisões ou ações tomadas durante o ano de 1998 certamente influenciaram os eventos descritos na subcoleção de 1999. Desta forma, a análise conjunta das duas subcoleções (sem discriminações) pode levar a hipóteses distorcidas.

Também deve-se ter atenção com o estilo e a linguagem usados nos textos. O uso de sinônimos e variações léxicas podem levantar falsas hipóteses. No caso do Mercosul, houve preocupação com este problema durante a análise de termos exclusivos (resultantes da técnica de diferença).

Outro cuidado a ser tomado é com respeito a termos negativos. No caso do *benchmarking*, a expressão "*unlike clustering*" poderia levar à falsa conclusão de que a técnica de agrupamento (*clustering)* estava presente no texto em questão. Para resolver este problema, foi usado um método de classificação que evita tais distorções.

7 AGRADECIMENTOS

Este trabalho é parcialmente apoiado por CAPES e CNPq.

8 REFERÊNCIAS BIBLIOGRÁFICAS

[1] AAMODT, Agnar & NYGARD, Mads. Different roles and mutual dependencies of data, information and knowledge - an AI perspective on their integration. Data & Knowledge Engineering, v.16, n.3, Setembro de 1995.

[2] BAEZA-YATES, Ricardo e alli. A model and a visual query language for structured text. In: String Processing and Information Retrieval: A South American Symposium - SPIRE'98. Proceedings... 1998.

[3] BOWDEN, Paul R.; HALSTEAD, Peter; ROSE, Tony G. Extracting conceptual knowledge from text using explicit relation markers. In: IX European Knowledge Acquisition Workshop. Proceedings... Lecture Notes in Artificial Intelligence, 1076. Maio de 1996.

[4] CALLAN, James P. Passage-level evidence in document retrieval. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. Proceedings... London: Springer-Verlag. 1994.

[5] CHEN, Z. Let documents talk to each other: a computer model for connection of short documents. Journal of Documentation, v.49. n.1, Março de 1993.

[6] CHINCHOR, Nancy; HIRSCHMAN, Lynette; LEWIS, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). Computational Linguistics, v.19, n.3, Setembro de 1993.

[7] CHOUDHURY, Vivek & SAMPLER, Jeffrey L. Information specificity and environmental scanning: an economic perspective. MIS Quarterly, Março de 1997.

[8] CLERC, Philippe. Inteligencia económica: retos actuales y perspectivas. Paris: UNESCO, 1998. 322-335p. (Informe mundial sobre la informacion)

[9] COWIE, Jim & LEHNERT, Wendy. Information extraction. Communications of the ACM, v.39, n.1, Janeiro de 1996.

[10] CROFT, W. Bruce. Machine learning and information retrieval. In: COLT Conference. Proceedings... Julho de 1995. (invited talk). Online in http://www.ee.umd.edu/medlab/filter/

[11] DAVIES, Roy. The creation of new knowledge by information retrieval and classification. Journal of Documentation, v.45, n.4, Dezembro de 1989.

[12] FELDMAN, Ronen & DAGAN, Ido. "Knowledge Discovery in Textual Databases (KDT)". IN: 1st International Conference on Knowledge Discovery (KDD-95). Proceedings ... pp. 112-117, Montreal, Agosto de 1995. Disponível por WWW em http://www.cs.biu.ac.il:8080/~feldman/

[13] FELDMAN, Ronen & HIRSH, Haym. Exploiting background information in knowledge discovery from text. Journal of Intelligent Information Systems, v.9, n.1, Julho/Agosto de 1997.

[14] FELDMAN, Ronen & DAGAN, Ido. Mining text using keyword distributions. Journal of Intelligent Information Systems, v.10, 1998. pp.281-300

[15] GOEBEL, Michael & GRUENWALD, Le. A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations, v.1, n.1, Junho de 1999.

[16] KASZKIEL, Marcin & ZOBEL, Justin. Passage retrieval revisited. In: XX International ACM SIGIR Conference on Research and Development in Information Retrieval. Proceedings... Philadelphia: ACM Press, 1997.

[17] KUHLTHAU, Carol C. Inside the search process: information seeking from the user's perspective. Journal of the American Society for Information Science, v.42, n.5, Junho de 1991.

[18] LIN, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). Proceedings... 1998.

[19] McKEOWN, Kathleen & RADEV, Dragomir R. Generating summaries of multiple news articles. IN: International ACM-SIGIR Conference on Research and Development in Information Retrieval. Proceedings... Seattle, 1995.

[20] MOENS, Marie-Francine & UYTTENDAELE, Caroline. Automatic text structuring and categorization as a first step in summarizing legal cases. Information Processing & Management, v.33, n.6, Novembro de 1997.

[21] MOSCAROLA, Jean; BAULAC, Yves; BOLDEN, Richard. Technology watch via textual data analysis. Note de Recherche nº 98-14, Université de Savoie. Julho de 1998.

[22] MOSCAROLA, Jean & BOLDEN, Richard. From the data mine to the knowledge mill: applying the principles of lexical analysis to the data mining and knowledge discovery process. Note de Recherche nº 98-15, Université de Savoie. Setembro de 1998.

[23] NASSIF, Mônica Erichssen Borges & CAMPELLO, Bernadete Santos. A organização da informação para negócios no Brasil. Perspectivas em Ciência da Informação, Minas Gerais: Escola de Ciência da Informação. v.2, n.2, 1997. p.149-161.

[24] PARSONS, Simon. Current approaches to handling imperfect information in data and knowledge bases. IEEE Transactions on Knowledge and Data Engineering, v.8, n.3, Junho de 1996.

[25] SALTON, G. & McGILL, M. J. Introduction to modern information retrieval. McGraw-Hill, 1983.

[26] SCHAFFER, Doug et alli. Navigating hierarchically clustered networks through fisheye and full-zoom methods. ACM Transactions on Computer-Human Interaction, v.3, n.2, Junho de 1996.

[27] SOUZA, Inguelore Scheunemann & BELARMINO, Luiz Clovis (eds). Anais do Seminário Estratégia Agroalimentar para o Mercosul. Pelotas: Universitária-UFPEL / EMPRAPA-Clima Temperado / IICA / SARGS, 1999.

[28] SPARCK-JONES, Karen & WILLET, Peter (eds). Readings in Information Retrieval. San Francisco: Morgan Kaufmann, 1997.

[29] VEERASAMY, Aravindan & HEIKES, Russell. Effectiveness of a graphical display of retrieval results. In: XX International ACM SIGIR Conference on Research and Development in Information Retrieval. Proceedings... 1997.

[30] WATTS, Robert J. & PORTER, Alan L. Innovation forecasting. Technological

Forecasting and Social Change, v.56, 1997.

[31] WEBBER, Alam. O que queremos dizer com conhecimento. In: DAVENPORT, T.; PRUSAK, L. Conhecimento Empresarial. 1998. p.1-28.

[32] WILLET, Peter. Recent trends in hierarchic document clustering: a critical review. Information Processing & Management, v.24, n.5, 1988. pp.577-597.

[33] ZANASI, Alessandro. Competitive Intelligence though datamining public sources. Competitive Intelligence Review, Alexandria, Virginia: SCIP. v.9, n.1, 1998.

Artigo 5 – OIA 2000

“Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses”
Co-autores: Leandro Krug Wives, José Palazzo M. de Oliveira
IV Oficina de Inteligência Artificial, pp.143-154
Universidade Católica de Pelotas.
Pelotas-RS, 25 de Agosto de 2000.
Editora EDUCAT. Organizador: Luiz Antônio Moro Palazzo

DESCOBERTA PROATIVA DE CONHECIMENTO EM COLEÇÕES TEXTUAIS: INICIANDO SEM HIPÓTESES

Stanley Loh
(UCPEL, ULBRA, PPGC/II/UFRGS)
loh@inf.ufrgs.br

Leandro Krug Wives
(PPGC/II/UFRGS)
wives@inf.ufrgs.br

José Palazzo M. de Oliveira
(II/UFRGS)
palazzo@inf.ufrgs.br

RESUMO

Este artigo discute o processo de descoberta de conhecimento em textos (KDT) segundo a abordagem proativa, isto é, segundo uma abordagem que inicia sem hipóteses predefinidas e é baseada em uma ação efetiva do pesquisador. A abordagem proativa difere da reativa, porque no primeiro caso o usuário não tem uma necessidade ou problema consciente, enquanto que no segundo o usuário sabe o que está procurando ou do que precisa. Na descoberta proativa, o usuário tem por objetivo encontrar conhecimento novo e útil, mas não sabe o que está à procura, muito menos por onde começar. Neste artigo, exemplos de descoberta proativa são apresentados e discutidos para mostrar como as técnicas de KDT podem ser usadas e que tipos de resultados podem ser obtidos segundo este paradigma. O trabalho apresenta também estratégias para descoberta de conhecimento no modo proativo (sem hipóteses iniciais) e discute a necessidade de intervenção humana no processo e os aspectos que podem influenciar negativamente os resultados (ruídos).

Palavras-chave: *descoberta de conhecimento, text mining, análise de textos*

1 INTRODUÇÃO

Com o crescente uso de computadores e principalmente da Internet, cada vez mais documentos eletrônicos estão sendo armazenados e colocados à disposição das pessoas. Davies [DAV89] afirma que muito conhecimento pode ser inferido a partir destes documentos. Entretanto, encontrar tal conhecimento é uma tarefa árdua.

Existem técnicas e ferramentas para Recuperação de Informação (RI), as quais auxiliam as pessoas a encontrar documentos que contenham informações relevantes [SPA97]. Entretanto, é necessário examinar os documentos resultantes para encontrar a informação desejada, o que não é uma tarefa fácil. Esta dificuldade é causada pelo fato de que documentos são insatisfatórios como respostas, por serem grandes e difusos em geral [WIL94]. Um exemplo prático são os serviços de busca (*search engines)* da Web, que encontram para o usuário um volume enorme de documentos, mas o usuário tem de examiná-los para encontrar o que deseja. Este problema é chamado de "sobrecarga de informações" (*information overload)*.

Alguns trabalhos estão sendo desenvolvidos para tentar resolver ou minimizar tal problema. A área de Extração de Informações (EI – *information extraction*) estuda metodologias, técnicas e sistemas que possam encontrar dados específicos dentro de textos. Tais sistemas extraem automaticamente valores de atributos (como campos de um banco de dados). Infelizmente, em geral, tais sistemas são muito dependentes do domínio, isto é, só funcionam com certos tipos de documentos [CRO95]. Além disto, para criar tais sistemas é necessário muita engenharia de conhecimento, examinando amostras de textos para saber como a informação é codificada em frases da língua natural [CHI93].

Estes trabalhos partem de uma necessidade específica do usuário ou de um objetivo previamente definido em uma aplicação e, portanto, são classificados sob o paradigma

reativo. Ou seja, o usuário deve definir sua necessidade ou problema e fornecer caminhos para a solução (por exemplo, as técnicas e parâmetros a serem utilizados). Em geral, esta é uma premissa falsa que distorce o processo de busca, uma vez que as pessoas não estão aptas a especificar precisamente o que é necessário para resolver seu problema. Pedir ao usuário para formular o que precisa é uma premissa irreal, se é isto justamente o que falta [BEL97]. Belkin e outros [BEL97] chamam a necessidade de informação de um estado anômalo de conhecimento (*ASK - Anomalous State of Knowledge*), portanto (por ser anômalo) difícil de representar.

Contrária a esta abordagem, surge um novo tipo de paradigma (**proativo**) que procura automaticamente informações novas e úteis em uma coleção de documentos, sem que seja necessário que o usuário estabeleça inicialmente uma necessidade. Na área de Banco de Dados, esta abordagem é conhecida como Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases – KDD)* [FAY96]. Em geral, os sistemas de KDD utilizam técnicas estatísticas conhecidas como técnicas de *Data Mining* para encontrar automaticamente padrões nas distribuições de valores em atributos ou campos de um banco de dados

Tais técnicas têm tido sucesso mas trabalham somente sobre dados estruturados. Em se tratando de coleções de textos (dados não-estruturados), o problema é recente e ainda necessita mais estudos. Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts*) é o termo utilizado para designar a aplicação das mesmas técnicas de KDD só que sobre características ou atributos extraídos de textos [FEL95]. Estas características podem ser valores de atributos/campos extraídos dos textos por algum tipo de inferência ou até algo mais simples como as próprias palavras do texto. Lin e outros [LIN98], por exemplo, descobrem associações entre palavras extraídas automaticamente dos textos. Os termos mais freqüentes são utilizados como atributos do texto. Já Feldman e Dagan [FEL98] aplicam as técnicas de KDD sobre palavras-chave associadas a textos. Tal associação já deve ter sido feita antes, seja por pessoas (atividades manuais e intelectuais) ou automaticamente por ferramentas de software. Estas técnicas usam análises estatísticas sobre as distribuições das características para descobrir padrões no formato de regras associativas.

Já Swanson e Smalheiser [SWA97] sugerem algumas estratégias mais complexas para descoberta de conhecimento em textos. Uma delas procura identificar analogias em diferentes textos através da análise de termos comuns. Sua sugestão é omitir termos relacionados à área para reunir documentos de diversas áreas que possam estar relacionados. Tais autores têm relatado sucesso com descobertas de novas e úteis alternativas na área médica. Entretanto, a relação entre os textos não é simples de ser feita, exigindo a intervenção de especialistas humanos no assunto.

Chen [CHE93], por sua vez, sugere a construção automática de resumos combinando partes de distintos textos, usando para isto estruturas internas (redes semânticas) e termos comuns aos textos. Davies [DAV89] sugere examinar as correlações escondidas em textos, isto é, combinações de conceitos através de relações estatísticas. Para tanto, Davies sugere que sejam analisadas as distribuições de termos numa coleção. Assim, por exemplo, foi possível identificar uma hipótese de relação entre um certo tipo de falha num sistema e alguns itens mais freqüentes (possíveis causas das falhas). Davies afirma que o todo é mais que a mera soma das partes, o que permite que conhecimentos novos não explicitamente presentes nos textos possam ser descobertos analisando relações semânticas entre os textos. No caso, as relações são identificadas através da análise dos termos presentes nos textos.

Feigenbaum (*apud* [DAV89]) compara as bibliotecas de hoje com as do futuro. As primeiras são como um armazém de objetos passivos. Já as bibliotecas do futuro serão uma coleção de documentos ativos que ajudarão às pessoas fornecendo conexões desconhecidas, associações e analogias, entendimento de conceitos novos, descoberta de novos métodos e teorias, sem que as pessoas precisem definir claramente quais são suas necessidades de informação. Minsky e Feigenbaum falam que os documentos devem ser capazes de "conversarem entre si" [DAV89] [CHE93].

Este artigo discute o uso de técnicas de KDT sob o paradigma **proativo**, ou seja, é analisado o processo de descoberta sem que seja necessário estabelecer hipóteses iniciais. Para tanto, a seção 2 contém uma breve revisão sobre o assunto, apresentando as principais técnicas de KDT e as diferenças entre os paradigmas **reativo** e **proativo**. Já na seção 3, são apresentados resultados de experimentos sob este paradigma. Na seção 4, são discutidos os aspectos envolvidos em processos deste tipo e que o influenciam (estratégias de descoberta, intervenção humana e ruídos). Por fim, a conclusão discute vantagens e limitações deste tipo de abordagem.

2 REVISÃO SOBRE O ASSUNTO

Nesta seção, é feita uma breve revisão dos assuntos envolvidos neste artigo. Primeiro, são apresentadas as principais técnicas para KDT, as quais são utilizadas nos experimentos apresentados mais adiante. Depois, na segunda subseção, são discutidas as diferenças entre as abordagens reativa e proativa.

2.1 TÉCNICAS PARA KDT

Existem muitas técnicas e ferramentas de software para realizar KDT. Nesta subseção, são apresentadas as principais técnicas.

A técnica mais básica é a recuperação de informações (RI), cujo objetivo é encontrar textos que podem conter determinada informação. Métodos para RI são discutidos em [SPA97] e [SAL83]. Uma técnica similar é a recuperação de passagens, que aplica as mesmas técnicas de RI só que sobre partes do texto [CAL94] [KAS97]. Já a técnica de extração de informação (EI) procura valores de atributos dentro dos textos [COW96]. A técnica de sumarização (*summarization*) tem por objetivo extrair resumos de um texto ou de uma coleção, podendo ser uma visão geral ou as partes mais importantes ou mais interessantes [SPA97] [SAL97] [MCK95] [CHE93].

A técnica de listagem de conceitos-chave (*key-concept listing*), por sua vez, analisa uma coleção de textos em busca de características comuns (palavras, palavras-chave, temas, etc). A ferramenta de Moscarola e outros [MOS98], por exemplo, encontra e apresenta ao usuário uma lista de termos relacionados por proximidade. Já Maarek [MAA92] extrai afinidades léxicas, definidas como relações entre unidades da linguagem, por exemplo sujeito-verbo, substantivo-adjetivo. A partir desta técnica pode-se fazer o inverso, isto é, descobrir diferenças comparando textos ou coleções, o que é chamado de técnica da diferença.

Já a técnica de agrupamento (*clustering*) é um pouco mais complexa. Ela é utilizada para identificar automaticamente, sem intervenção humana, grupos de textos similares [WIL88]. Sua principal utilidade é permitir encontrar características comuns em subgrupos quando não há nada em comum na coleção toda. Já a técnica de classificação ou categorização procura encontrar temas ou assuntos no conteúdo dos textos (do que os textos

estão tratando). A já comentada técnica de associação (ou correlação) descobre relações de dependência entre textos ou características dos textos.

Existem também técnicas para visualização de resultados, que ajudam o usuário a entender melhor o conhecimento descoberto [VEE97] [BAE98] [SCH96b].

Apesar de apresentadas em separado, nada impede que elas sejam usadas de modo integrado, uma após a outra, de forma que a saída de uma seja a entrada da seguinte. Por exemplo, Moens e Uyttendaele [MOE97] usam a técnica de sumarização associada com EI, para criar resumos de casos jurídicos. Moscarola e outros [MOS98] [MOS98b] apresentam uma ferramenta que integra diversas técnicas para KDT.

2.2 DESCOBERTA REATIVA X PROATIVA

A maioria dos pesquisadores concorda que o processo de descoberta é cíclico, tendo como passos principais: [PAR89] [AGR93] [ING96]

e) a formulação de hipóteses;
f) o teste das hipóteses;
g) a observação dos resultados (para refutar ou confirmá-las);
h) a revisão das hipóteses e a sua modificação (reiniciando o processo), até que o usuário se dê por satisfeito.

Entretanto, esta estratégia só pode ser aplicada quando o usuário consegue formular hipóteses iniciais, ou seja, quando ele tem idéia de qual é o seu objetivo ou necessidade e sabe do que precisa.

De acordo com Choudhury e Sampler [CHO97], existem dois modos para aquisição de informação: o modo reativo e o modo proativo. No primeiro caso, a informação é adquirida para resolver um problema específico do usuário (uma necessidade resultante de um estado anômalo de conhecimento). Nestes casos, o usuário sabe o que quer e poderá identificar a solução para o problema quando há encontrar.

Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório, para detectar problemas potenciais ou oportunidades. Neste segundo caso, o usuário não tem um objetivo específico.

Oard e Marchionini [OAR96] classificam as necessidades de informação em estáveis ou dinâmicas e em específicas ou abrangentes (gerais). Taylor (citado em [OAR96]) define 4 tipos de necessidades, os quais formam uma escala crescente para a solução do problema:

- necessidades viscerais: quando existe uma necessidade ou interesse, mas esta não é percebida de forma consciente;
- necessidades conscientes: quando o usuário percebe sua necessidade e sabe do que precisa;
- necessidades formalizadas: quando o usuário expressa sua necessidade de alguma forma;
- necessidades comprometidas: quando a necessidade é representada no sistema.

As necessidades tratadas pela abordagem de descoberta reativa poderiam ser classificadas como estáveis e específicas, segundo a classificação de Oard e Marchioninni, e como conscientes (no mínimo), segundo Taylor. Isto porque o usuário sabe o que quer, mesmo que não consiga formalizar.

Exemplos de objetivos que caracterizam um processo reativo são:

- encontrar atributos comuns nos produtos mais vendidos;
- encontrar motivos que levam à evasão ou a reclamações de clientes;
- achar perfis de grupos de clientes;

- encontrar clientes potenciais para propaganda seletiva;
- encontrar concorrentes no mercado.

No modo reativo, o usuário tem uma idéia, mesmo que vaga, do que pode ser a solução ou, pelo menos, de onde se pode encontrá-la. Pode-se dizer então que o usuário possui algumas hipóteses iniciais, que ajudarão a direcionar o processo de descoberta. Neste caso, é necessário algum tipo de pré-processamento, por exemplo para selecionar atributos (colunas em uma tabela) ou valores de atributos (células). Isto exige entender o interesse ou objetivo do usuário para limitar o espaço de busca (na entrada) ou filtrar os resultados (na saída). É o caso típico de quando se deseja encontrar uma informação específica, por exemplo, um valor para um atributo ou um processo (conjunto de passos) para resolver um problema.

Já as necessidades da abordagem proativa poderiam ser classificadas como dinâmicas e abrangentes, segundo a classificação de Oard e Marchioninni. São dinâmicas porque podem mudar durante o processo, já que o objetivo não está bem claro, e são abrangentes porque o usuário não sabe exatamente o que está procurando. Pela taxonomia de Taylor, as necessidades do modo proativo são viscerais. Isto quer dizer que há uma necessidade ou objetivo, mas o usuário não consegue definir o que precisa para resolver o problema. A necessidade típica do modo proativo poderia ser representada pela expressão: "*diga-me o que há de interessante nesta coleção*". Neste caso, o usuário não tem de forma definida o que lhe seja de interesse (o que precisa), podendo tal interesse mudar durante o processo. Pode-se dizer que é um processo exploratório, sendo, em geral, iterativo (com retroalimentação) e interativo (com ativa participação e intervenção do usuário).

Na abordagem proativa, não há hipóteses iniciais ou elas são muito vagas. O usuário deverá descobrir hipóteses para a solução do seu problema e explorá-las, investigá-las e testá-las durante o processo. Em geral, acontece porque o usuário não sabe exatamente o que está procurando. É o caso típico de quando se quer monitorar alguma situação ou encontrar algo de interessante que possa levar a investigações posteriores. Depois que hipóteses são levantadas, o processo pode seguir como no paradigma reativo, talvez sendo necessário avaliar as hipóteses, para verificar se são verdadeiras ou não.

3 EXPERIMENTOS

Foi implementado um conjunto de ferramentas de software para KDT. Há uma ferramenta diferente para cada técnica descrita na seção anterior. As ferramentas estão integradas de forma que os resultados de uma podem ser usados como entrada de outra. Assim, processos complexos de descoberta podem ser realizados.

Nesta seção, são apresentados experimentos sob a abordagem proativa, usando as ferramentas implementadas. Para os experimentos foram usadas 3 coleções de textos, a saber:

a) coleção política: formada por textos extraídos de um jornal publicado na Web falando sobre um prefeito; os textos foram extraídos usando a ferramenta local de recuperação de informação do *site* e tendo como consulta o nome do prefeito; esta coleção está dividida em dois segmentos (sub-coleções), uma com 180 textos publicados em 1997 e outra com 178 textos publicados em 1999;

b) coleção médica: composta por 1040 prontuários médicos escritos por médicos sobre pacientes de uma clínica psiquiátrica (26 prontuários eram referentes à internação do paciente);
c) coleção sobre guerra: 18 textos versando sobre guerras na história mundial, extraídos de uma enciclopédia (em inglês).

A seguir, são apresentados exemplos de processos de descoberta sobre estas coleções, iniciando com diferentes técnicas de KDT e sem hipóteses ou interesse inicial (abordagem proativa). Estes exemplos ajudam a entender como se pode utilizar a abordagem proativa e que tipo de resultados podem ser alcançados. As técnicas utilizadas durante cada processo aparecem em **negrito**.

- **1º Experimento: coleção política**

Considerando que há dois segmentos nesta coleção, a técnica de **listagem de conceitos-chave** foi utilizada para comparar as palavras que apareciam nas duas sub-coleções (1997 x 1999). Analisando os termos mais freqüentes nas duas coleções, chegou-se a uma descoberta interessante: o nome da esposa do prefeito aparecia mais vezes na sub-coleção referente a 1999 (35 textos) do que na de 1997 (somente 6 textos). Isto suscitou a hipótese de que a esposa do prefeito era citada em situações diferentes. Passou-se então a analisar estes dois subgrupos (35 x 6 textos) em separado, usando a mesma técnica de **listagem**. Notou-se que, no grupo referente a 1997, um termo aparecia em todos os textos. Com um pouco de conhecimento prévio sobre domínio (*background knowledge*), sabia-se que tal termo designava um escândalo no qual o prefeito era acusado de corrupção e no qual sua esposa fora envolvida também. A conclusão final é de que a esposa do prefeito só aparece em 1997, nesta mídia, envolvida neste escândalo. Já no segmento de 1999, não foi possível identificar termos significativos comuns em todos os 35 textos (com a mesma técnica de **listagem**). Então usou-se a técnica de **agrupamento** sobre este pequeno segmento. Analisando a **listagem** de termos comuns a cada *cluster* resultante, notou-se que os termos eram muito genéricos sobre o assunto. Então estes termos foram eliminados dos textos da coleção e a **agrupamento** foi refeita. Como resultado, foram encontrados dois *clusters* com forte coesão, isto é, cujos documentos tinham alto grau de similaridade entre si. Usando a técnica de **sumarização**, verificou-se que o primeiro *cluster* continha textos que tratavam de uma reunião na casa de uma certa pessoa. No segundo *cluster*, notou-se que os termos "*new*" e "*york*" apareciam em todos os textos. Examinando estes textos com a técnica de **sumarização**, foi possível saber que a esposa do prefeito viveu por uns tempos na cidade de Nova York. A conclusão final é que, em 1999, a esposa do prefeito é citada em diferentes situações.

- **2º Experimento: coleção política**

Utilizando-se a técnica da **diferença** para comparar os termos mais freqüentes das duas sub-coleções (1997 x 1999), notou-se que o termo "separação" aparecia somente no segundo segmento. A primeira hipótese levantada era de que o prefeito e sua esposa estavam terminando seu casamento. Para verificar tal hipótese, foram extraídos, com a técnica de **sumarização**, resumos com as frases onde o tal termo aparecia. Os resultados confirmaram a hipótese levantada, o que leva à conclusão de que a separação do casal somente aconteceu após 1997.

- **3º Experimento: coleção médica**

Nesta coleção em particular, havia 26 textos sobre a internação de pacientes diferentes. Realizando um processo de **agrupamento**, descobriu-se um *cluster* forte. Analisando em separado este grupo com a técnica de **listagem**, notou-se a predominância de termos relacionados a "familiares" e "suicídio". A interpretação inicial para estes resultados é de que "a maioria das pessoas com tendências suicidas possuem família". Tal hipótese está sendo avaliada por especialistas médicos da área, os quais acharam a princípio tais descobertas interessantes mas merecedoras de estudos mais profundos.

- **4º Experimento: coleção de guerra**

Utilizando a técnica de **associação** sobre as palavras dos textos desta coleção, encontrou-se que em, 100% dos casos (grau de confiança),

- quando o termo "doença" aparecia, então o termo "verão" também aparecia.
- quando o termo "ditadura" aparecia, então o termo "invasão" também aparecia; e
- quando o termo "ditadura" aparecia, então o termo "assassinato" (ou um de seus correlatos, por exemplo, "assassinar", "matar") também aparecia.

Tais resultados são interessantes mas não permitem grandes conclusões, já que a coleção não era representativa, como será discutido na próxima seção.

4 OBSERVAÇÕES E DISCUSSÃO

Observando os processos de descoberta realizados e seus resultados, pode-se chegar a algumas conclusões e também são levantadas algumas dúvidas. O interesse deste trabalho foi o de analisar três aspectos principais:

d) que tipos de estratégias podem ou devem ser usadas na abordagem proativa (por exemplo, por onde começar e que passos seguir depois);
e) qual a importância da intervenção humana no processo e como o conhecimento prévio sobre o assunto pode influenciar tais tipos de abordagem;
f) que aspectos influenciam tal processo, podendo levar a interpretações erradas.

A seguir, cada um destes aspectos é analisado com base nos experimentos realizados.

4.1 ESTRATÉGIAS PARA DESCOBERTA PROATIVA

Um dos problemas do paradigma proativo é definir um plano de uso das técnicas ou de como a coleção textual deverá ser investigada de forma automática pelas ferramentas, a fim de serem descobertas hipóteses.

Kuhlthau [KUH91] determinou seis fases em processo de descoberta de informação: iniciação, seleção, exploração, formulação, coleção e apresentação. Cada fase é caracterizada por atitudes diferentes do usuário (por exemplo, em relação a sentimentos, pensamento, ações e tarefas). Uma das descobertas mais interessantes desta pesquisadora é que o usuário inicia procurando algum tipo de conhecimento mais geral, depois ele procura informação relevante em grupos mais restritos e termina procurando informações mais focadas ou específicas. Durante este processo, o usuário reconhece, identifica, investiga, formula, reúne e complementa o conhecimento.

Watts e Porter [WAT97] propõem um esboço de metodologia (*framework*) sobre algumas ferramentas de descoberta. Entretanto, a estratégia somente foi testada envolvendo

problemas da área de Inteligência Competitiva. Neste caso, pode-se dizer que a estratégia proposta está ainda no paradigma reativo, pois é apropriada para encontrar nomes de pessoas e companhias e para examinar o vocabulário técnico, necessitando bastante conhecimento específico do domínio.

Seguindo as sugestões destes trabalhos e com base nas observações feitas durante os experimentos, foi possível identificar alguns passos comuns. Sugere-se então uma estratégia para descoberta proativa de conhecimento em textos. Não se pode considerar esta estratégia uma metodologia, mas sim um esboço (*framework*), que poderá conduzir os usuários no processo, indicando os passos principais (técnicas ou ferramentas a serem usadas). Os passos são resumidamente descritos a seguir:

7) seleção de textos: o primeiro passo é selecionar uma coleção de textos sobre os quais serão aplicadas as técnicas; as técnicas automáticas mais indicadas são a recuperação de informação (que encontra textos procurando por palavras-chave ou termos presentes nos textos) e a classificação (que separa textos por assunto); outra possibilidade, é o usuário mesmo encontrar ou selecionar os textos, o que demanda mais trabalho manual;
8) análise da coleção toda ou de partes: neste ponto, o usuário deve decidir se irá aplicar as técnicas de descoberta sobre todos os textos ou sobre partes; a sugestão é que se comece analisando toda a coleção e depois se examine sub-coleções; em alguns casos, nada de interessante é encontrado na coleção toda, o que leva o usuário necessariamente a investigar pequenos grupos; a separação em grupos pode ser feita de forma automática, com a técnica de agrupamento, ou sob algum critério estabelecido pelo usuário, como por exemplo selecionando partes de interesse com as técnicas de recuperação ou classificação;
9) análise de grupos de textos (toda a coleção ou partes): uma boa maneira de começar a análise é extraindo uma lista de termos comuns a todos os textos ou que aparecem em mais de um (técnica de listagem de conceitos-chave); a técnica de diferença pode ser usada depois para levantar novas hipóteses; por fim, a técnica de associação, mesmo que demorada, pode ajudar a descobrir algo interessante;
10) comparação de sub-coleções entre si ou em relação à coleção toda: os resultados conseguidos com as técnicas de listagem de conceitos-chave, diferença e associação aplicadas a cada grupo particular podem ser comparados entre si ou com os resultados obtidos com a coleção toda;
11) validação de hipóteses: em geral, a técnica de resumos traz bons resultados, pois possibilita ao usuário ler as frases mais significativas e interpretar os resultados;
12) retroalimentação: como o processo é cíclico, os passos ou o processo todo podem ser refeitos.

4.2 NECESSIDADE DE INTERVENÇÃO HUMANA E CONHECIMENTOS PRÉVIOS

Uma discussão que surge é se ferramentas de software poderão extrair automaticamente conhecimento a partir de coleções textuais. Os experimentos realizados mostram que é possível automatizar partes do processo de descoberta, minimizando a dependência ao usuário. Entretanto, fica claro que algum tipo de intervenção humana é necessária e útil. Por exemplo, o primeiro passo do processo obrigatoriamente precisa da intervenção do usuário, para selecionar os textos da coleção, seja de forma manual ou

fornecendo parâmetros para as ferramentas de recuperação. Também será necessário que o usuário interprete os resultados no contexto da realidade, para que as descobertas sejam úteis. Segundo Aamodt and Nygard [AAM95], o conhecimento é imprescindível para que os dados possam ser interpretados e se tornem informação. O conhecimento é subjetivo e depende das pessoas. Por isto, Moscarola and Bolden [MOS98b] sugerem o modelo construtivista ao invés do positivista para os processos de descoberta, ou seja, o processo deve ser guiado pelo usuário.

Por outro lado, o conhecimento prévio (*background knowledge*) de que dispõe o usuário ajuda no processo de descoberta, limitando o espaço de pesquisa ou análise, sem que o usuário precise ainda definir hipóteses. Por exemplo, Feldman e Hirsh [FEL97] aplicam as mesmas técnicas descritas em [FEL98] mas permitem que o usuário intervenha no processo, fazendo uso de seus conhecimentos prévios sobre o domínio ou assunto (*background knowledge*). Isto acelera o processo e filtra os resultados de acordo com o interesse do usuário.

Um exemplo de uso do conhecimento prévio nos experimentos, aparece no 2º experimento, quando o usuário interpreta o termo "separação" como algo ligado ao casal citado. Deste caso, conclui-se que não é suficiente encontrar termos comuns ou diferentes. É necessário algum tipo de conhecimento prévio, mesmo que mínimo e limitado à linguagem. Um exemplo de uso de conhecimento sobre o domínio, aparece no 1º experimento, quando o usuário pode verificar que um termo chama mais à atenção que os outros (no caso, o termo que designava o escândalo de corrupção). No 3º experimento, pode-se notar que a falta de conhecimento especializado sobre o domínio pode resultar em descobertas que não podem ser aproveitadas, conseqüentemente também em desperdício de esforços. Pelos experimentos, descobriu-se que uma boa maneira de obter um pouco mais de conhecimento sobre o domínio é examinando os termos mais freqüentes, com a técnica de listagem. Isto permite ao usuário conhecer o estilo dos textos ou o escopo do conteúdo (do que se fala e do que não se fala nos textos). Para maiores discussões teóricas, Choudhury and Sampler [CHO97] discutem tipos de conhecimento prévio em processos de aquisição de conhecimento.

4.3 RUÍDOS NO PROCESSO DE KDT

Durante processos de descoberta, alguns aspectos podem influenciar o resultado final. Estes são chamados de ruídos e diminuem a qualidade das descobertas. O primeiro problema notado nos experimentos é o grau de representatividade da coleção. Por exemplo, analisando os resultados do 4º experimento, não se pode afirmar ou concluir que doenças só acontecem em guerras durante o verão. Isto porque a coleção pode não descrever todas as guerras ou os textos podem não conter todas as informações sobre a guerra que descrevem.

Mesmo que a coleção seja representativa, outros tipos de ruído podem aparecer, como o caso dos sinônimos. Por exemplo, no 4º experimento, após utilizar a técnica de diferença, notou-se que somente um texto continha o termo "fuga", levantando assim a hipótese de que somente uma das guerras teve fuga. Para verificar tal hipótese, foi usada a técnica de sumarização, procurando frases que tivessem termos relativos a fuga (como "fugir", "escapar", etc). Os resultados provaram que a hipótese inicial estava errada.

Outro ruído relativo ao vocabulário pode acontecer quando são usados termos polisêmicos, com mais de um significado. Por exemplo, nos experimentos com a coleção política, o termo "família" aparecia referenciando "parentes" ou como parte da "Secretaria

da Família e do Bem-Estar". Isto pode levar a hipóteses erradas quando usando a técnica de listagem, por exemplo. Problemas com o vocabulário, como sinônimos e termos polisêmicos, são discutidos em [FUR87].

Outro cuidado que se deve ter é com o contexto em que os termos aparecem. Por exemplo, o termo "melhora" aparecia freqüentemente nos prontuários médicos em frases negativas (por exemplo, "o paciente não apresentou melhora"). Outros problemas ainda podem surgir por erros ortográficos.

Além disto tudo há ainda o problema da confiabilidade das fontes de informação. No caso da Web, tal problema é ainda mais preocupante, já que os documentos mudam rapidamente [CHE93]. Alguns trabalhos sugerem estratégias para avaliação da qualidade da informação disponível na Web [SCH96] [OWE97] [SMI97]. Hersh [HER95] discute a falta de qualidade em informações textuais e [PAR96] sugere que problemas podem ocorrer devido a 5 tipos de imperfeições:

- informação incompleta: quando faltam detalhes de informação (por exemplo, atributos sem valores);
- informação imprecisa: devido à diferença de granularidade (por exemplo, datas sem o dia);
- informação incerta: quando não pode ser provada;
- informação vaga: devido a imprecisões do vocabulário;
- informação inconsistente: por exemplo, quando há valores contraditórios.

5 CONCLUSÃO

Este artigo discutiu o processo de descoberta de conhecimento em textos segundo a abordagem proativa, que é aquela que inicia sem que o usuário tenha hipóteses. Os experimentos realizados com as ferramentas de software implementadas permitiram concluir que tal abordagem é viável, ou seja, é possível realizar descoberta e obter resultados interessantes sem ter algum tipo de hipótese ou interesse inicial.

Exemplos de descoberta foram apresentados para mostrar que técnicas podem ser usadas, como elas podem ser usadas e a que tipos de resultados elas podem levar. O trabalho também apresentou estratégias para descoberta de conhecimento no modo proativo (sem hipóteses iniciais). Também foi discutida a necessidade de intervenção humana no processo e como os conhecimentos prévios sobre o domínio ou sobre a linguagem podem ajudar no processo. As contribuições ainda incluem uma análise dos possíveis problemas, chamados de ruídos, que podem interferir no processo, levando a interpretações errôneas.

6 AGRADECIMENTOS

Este trabalho é parcialmente apoiado por CNPq e CAPES.

7 REFERÊNCIAS BIBLIOGRÁFICAS

[AAM95] AAMODT, Agnar; NYGARD, Mads. Different roles and mutual dependencies of data, information and knowledge - an AI perspective on their integration. **Data & Knowledge Engineering**, v.16, n.3, Setembro de 1995.

[AGR93] AGRAWAL, Rakesh; IMIELINSKI, Tomasz. Database mining: a performance

perspective. **IEEE Transactions on Knowledge and Data Engineering**, v.5, n.6, Dezembro de 1993.

[BAE98] BAEZA-YATES, Ricardo e alli. A model and a visual query language for structured text. In: String Processing and Information Retrieval: A South American Symposium - SPIRE'98. **Proceedings...** 1998.

[BEL97] BELKIN, N. J.; ODDY, R. N.; BROOKS, H. M. ASK for information retrieval: part I. background and theory. In: [SPA97]

[CAL94] CALLAN, James P. Passage-level evidence in document retrieval. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** London: Springer-Verlag. 1994.

[CHE93] CHEN, Z. Let documents talk to each other: a computer model for connection of short documents. **Journal of Documentation**, v.49. n.1, Março de 1993.

[CHI93] CHINCHOR, Nancy; HIRSCHMAN, Lynette; LEWIS, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). **Computational Linguistics**, v.19, n.3, Setembro de 1993.

[CHO97] CHOUDHURY, Vivek; SAMPLER, Jeffrey L. Information specificity and environmental scanning: an economic perspective. **MIS Quarterly**, Março de 1997.

[COW96] COWIE, Jim; LEHNERT, Wendy. Information extraction. **Communications of the ACM**, v.39, n.1, Janeiro de 1996.

[CRO95] CROFT, W. Bruce. Machine learning and information retrieval. In: COLT Conference. **Proceedings...** July 1995. (invited talk). Online in http://www.ee.umd.edu/medlab/filter/

[DAV89] DAVIES, Roy. The creation of new knowledge by information retrieval and classification. **Journal of Documentation**, v.45, n.4, Dezembro de 1989.

[FAY96] FAYYAD, Usama M. et alli (ed) **Advances in Knowledge Discovery and Data Mining**. Menlo Park, The MIT Press, 1996.

[FEL95] FELDMAN, Ronen and DAGAN, Ido. Knowledge discovery in textual databases (KDT). In: 1st International Conference on Knowledge Discovery (KDD-95). **Proceedings...** Montreal, Agosto de 1995.

[FEL97] FELDMAN, Ronen and HIRSH, Haym. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, v.9, n.1, Julho/Agosto de 1997.

[FEL98] FELDMAN, Ronen; DAGAN, Ido. Mining text using keyword distributions. **Journal of Intelligent Information Systems**, v.10, n.3, 1998.

[FUR87] FURNAS, G. W. et al. The vocabulary problem in human-system communication. **Communications of the ACM**, v.30, n.11, Novembro de 1987.

[HER95] HERSH, William R. et alli. Towards new measures of information retrieval evaluation. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** 1995.

[ING96] INGWERSEN, Peter. Congnitive perspectives of information retrieval interaction: elements of a cognitive IR theory. **Journal of Documentation**, v.52, n.1, Março de 1996.

[KAS97] KASZKIEL, Marcin; ZOBEL, Justin. Passage retrieval revisited. In: XX International ACM SIGIR Conference on Research and Development in

Information Retrieval. **Proceedings...** Philadelphia: ACM Press, 1997.

[KUH91] KUHLTHAU, Carol C. Inside the search process: information seeking from the user's perspective. **Journal of the American Society for Information Science**, v.42, n.5, Junho de 1991.

[LIN98] LIN, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). **Proceedings...** 1998.

[MAA92] MAAREK, Yoëlle S. Automatically constructing simple help systems from natural language documentation. IN: JACOBS, Paul S. (ed) **Text-based intelligent systems: current research and practice in information extraction and retrieval**. New Jersey: Lawrence Erlbaum, 1992.

[MCK95] McKEOWN, Kathleen; RADEV, Dragomir R. Generating summaries of multiple news articles. IN: International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** Seattle, 1995.

[MOE97] MOENS, Marie-Francine; UYTTENDAELE, Caroline. Automatic text structuring and categorization as a first step in summarizing legal cases. **Information Processing & Management**, v.33, n.6, Novembro de 1997.

[MOS98] MOSCAROLA, Jean; BAULAC, Yves; BOLDEN, Richard. **Technology watch via textual data analysis**. Note de Recherche nº 98-14, Université de Savoie. Julho de 1998.

[MOS98b] MOSCAROLA, Jean; BOLDEN, Richard. **From the data mine to the knowledge mill: applying the principles of lexical analysis to the data mining and knowledge discovery process**. Note de Recherche nº 98-15, Université de Savoie. Setembro de 1998.

[OAR96] OARD, Douglas W.; MARCHIONINI, Gary**. A conceptual framework for text filtering**. Technical Report, University of Maryland. Maio de 1996. Online at http://www.ee.umd.edu/medlab/filter/

[OWE97] OWENS, Janet; RAGAINS, Patrick. **Evaluating Information Sources**. Janeiro de 1997. Online at http://www.library.unr.edu/~ragains/eval.html

[PAR89] PARSAYE, Kamran et alli. **Intelligent databases: object-oriented, deductive hypermedia technologies**. New York: John Wiley & Sons, 1989.

[PAR96] PARSONS, Simon. Current approaches to handling imperfect information in data and knowledge bases. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n.3, Junho de 1996.

[SAL83] SALTON, Gerard; McGILL, M. J. **Introduction to Modern Information Retrieval**. McGraw-Hill, 1983.

[SAL97] SALTON, Gerard et alli. Automatic text structuring and summarization. **Information Processing & Management**, v.33, n.2, Março de 1997.

[SCH96b] SCHAFFER, Doug et alli. Navigating hierarchically clustered networks through fisheye and full-zoom methods. **ACM Transactions on Computer-Human Interaction**, v.3, n.2, Junho de 1996.

[SCH96] SCHOLZ, Ann. **Evaluating World Wide Web Information**. Fevereiro de 1996. Online at http://thorplus.lib.purdue.edu/research/classes /gs175/3gs175/evaluation.html

[SMI97] SMITH, Alastair. **Criteria for evaluation of Internet Information Resources**.

Março de 1997. Online at http://www.vuw.ac.nz/~agsmith/evaln/index.htm

[SPA97] SPARCK-JONES, Karen; WILLET, Peter (eds). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.

[SWA97] SWANSON, D. R.; SMALHEISER, N. R., An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence,** 91 (1997) 183-203.

[VEE97] VEERASAMY, Aravindan; HEIKES, Russell. Effectiveness of a graphical display of retrieval results. In: XX International ACM SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** 1997.

[WAT97] WATTS, Robert J.; PORTER, Alan L. Innovation forecasting. **Technological Forecasting and Social Change**, 56. 1997.

[WIL94] WILKINSON, Ross. Effective retrieval of structured documents. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. **Proceedings...** London: Springer-Verlag. 1994.

[WIL88] WILLET, Peter. Recent trends in hierarchic document clustering: a critical review. **Information Processing & Management**, v.24, n.5, 1988.

Artigo 6 – SIGKDD Explorations

"Concept-based knowledge discovery in texts extracted from the web"
Co-autores: Leandro Krug Wives, José Palazzo Moreira de Oliveira
SIGKDD Explorations, v.2, n.1, July 2000, pp.29-39
Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. ACM Press.
Editor-Chefe: Usama Fayyad
Editor Assistente: Sunita Sarawagi
Editor Convidado: Paul Bradley
Disponível por WWW em http://www.acm.org/sigkdd/explorations

CONCEPT-BASED KNOWLEDGE DISCOVERY IN TEXTS EXTRACTED FROM THE WEB

Stanley Loh, Leandro Krug Wives e José Palazzo Moreira de Oliveira

ABSTRACT

This paper presents an approach for knowledge discovery in texts extracted from the Web. Instead of analyzing words or attribute values, the approach is based on concepts, which are extracted from texts to be used as characteristics in the mining process. Statistical techniques are applied on concepts in order to find interesting patterns in concept distributions or associations. In this way, users can perform discovery in a high level, since concepts describe real world events, objects, thoughts, etc. For identifying concepts in texts, a categorization algorithm is used associated to a previous classification task for concept definitions. Two experiments are presented: one for political analysis and other for competitive intelligence. At the end, the approach is discussed, examining its problems and advantages in the Web context.

Keywords

Knowledge discovery, data mining, information extraction, categorization, text mining.

INTRODUCTION

The Web is a large and growing collection of texts. This amount of text is becoming a valuable resource of information and knowledge. As Garofalakis and partners comment, "*the majority of human information will be available on the Web in ten years*" [21]. To find useful information in this source is not an easy and fast task. People, however, want to extract useful information from these texts quickly and with low cost.

The heterogeneity and the amount of sources may lead to the information overload problem, which happens when we have too much information available that we cannot manage. To minimize the overload and to help people to extract information from texts has emerged the novel area called *Knowledge Discovery in Texts* (KDT) [15], which concerns the application of *Knowledge Discovery in Databases* (KDD) techniques over texts. KDD is the "*nontrivial extraction of implicit, previously unknown, and potentially useful information from given data*" [19]. However, the most researches in KDD work on structured data (like in a database) and cannot be applied over textual data directly.

Feldman and partners [15] [16] [17] face the problem of applying KDD tools over keywords that are assigned to texts as attributes. These mining techniques use statistical analysis to discover association rules and interesting patterns over keyword distributions and associations. To perform the KDT process, keywords should be previously assigned to texts. Authors do not discuss the way in which keywords are assigned to texts, suggesting that this assignment may be done manually by humans or automatically by software tools.

By other side, Lin and partners [31] use terms automatically extracted from texts to categorize documents and to find associations. The most frequent terms are assigned as keywords (attributes). However, when analyzing terms, problems arise due to the *vocabulary problem*, discussed in [5], [7] and [20]. The language use may cause semantic mistakes due to synonyms (different words for the same meaning), polysemy (the same word with many meanings), lemmas (words with the same radical, like the verb "to marry" and the noun "marriage") and quasi-synonyms (words related to the same subject, object or event, like "bomb" and "terrorist attack"). For example, a murder may be described with terms like "murder" or "homicide". If analyzing only the terms (assigned to or extracted from texts), the discovery process may be misled by semantic gaps.

Another interesting approach is to apply KDD techniques after the use of Information Extraction (IE) techniques, which transform information present in texts into a structured database [10]. When textual information is structured into a database, we can do useful analyses only possible with Database Management Systems [33]. For example, using associative techniques, one can discover relations between items examining transactions in a database. In [21], we can see an approach that uses associative techniques over Web pages structures. In this case, URLs are extracted from pages to represent items in a

database. However, Etzioni [14] advises: "*HTML annotations structure the display of Web pages, but provide little insight into their content*". Besides that, texts may even be published without titles, keywords, links or author information, making HTML tags useless.

In some cases, IE has shown to be feasible in exploring textual content. Soderland [43] extracts information about weather forecasting in Web texts. Etzioni [14] cites some successful cases of IE applications using "wrappers" (Web information extractors). Although the promising results in IE, unfortunately the "*majority of today's IE systems rely on hand-coded wrappers to access a fixed set of Web resources*" [21]. This means IE systems are very domain dependent, being useful only for specific applications [11] or working only with a special class of document types. Besides that, to create such systems, a lot of knowledge about the domain is necessary (knowledge engineering), examining text styles and how information is encoded into natural language phrases [8]. Mattox and partners [33] conclude that semantic knowledge about the domain (like ontologies) is essential to IE and that there is a non-trivial effort to generate wrappers, even with tools. When the access to information is infrequently, it is not worth the effort.

This paper presents an approach for performing knowledge discovery in texts extracted from the Web, through the analysis of high level characteristics, minimizing the *vocabulary problem* and the effort necessary to extract useful information. Instead of applying mining techniques on attribute values, terms or keywords labeling texts, the discovery process works over concepts extracted from texts. Concepts represent real world attributes (events, objects, feelings, actions, etc) and, as seen in discourse analysis, they help to understand ideas and ideologies present in texts. The approach combines an automatic categorization task with a mining task. Categorization task identifies concepts present inside texts, without needing too much labor. Mining task discovers patterns by analyzing and relating concept distributions in a collection. A previous classification task is needed to create concept definitions.

For a better communication, we distinguish classification from categorization. Classification is the process of inducing a model or description for each class in terms of its attributes [21]. This model is then used to identify the class of future elements. Categorization refers to the identification of categories, themes, subjects or concepts present in texts. Lewis and Hayes [28] give a different definition for text categorization: "*the classification of units of natural language text into predefined categories*", and Wiener and partners [46] call this problem as "*topic spotting*". Some authors tend to view categorization as part of the classification. Thus, in some papers classification and categorization are used as synonyms. Here, we use these terms as different meanings.

The main goals of the proposed approach are: (a) to do discovery upon concepts instead of words or attribute values, allowing the user to find ideas, ideologies, trends and intentions present in texts; (b) to minimize the effort necessary to identify concepts in texts and to perform knowledge discovery; (c) to find interesting patterns in textual collections using simple statistical techniques; (d) to allow users to perform *ad hoc* discovery (with ill-defined goals) without having to expend time and effort creating formal models.

Applications of the concept-based KDT approach include (but are not limited to): discourse analysis (looking for intentions and structures in textual expressions), sociology (themes and ideas present in a textual study), analogy (same concepts present in different discourses), health research (searching for relations between symptoms present in textual records) and competitive intelligence (strategies used by different companies).

The section 2 presents a general overview of the approach. Subsections discuss each task of the concept-based KDT process. In section 3, results from two experiments are presented. Section 4 discusses the experiments and the final section evaluates the approach.

THE CONCEPT-BASED APPROACH FOR KDT

Although many researchers use the term "concept", it is difficult to see a formal definition of what "*concept*" is. Looking for a definition in dictionaries, we find that "concept" is an "idea, opinion, thought". This confirms the general and intuitive idea that concepts are used to explore and examine the contents of talks, texts, documents, books, messages, etc. Chen and partners [6], for example, use concepts to identify the content of comments in a brainstorming discussion. In Information Retrieval, concepts are used with success to index and retrieve documents. Lin and Chen [30] comment "*the concept-based retrieval capability has been considered by many researchers and practitioners to be an effective complement to the prevailing keyword search or user browsing*". In this case, its main advantage is to minimize the vocabulary problem.

According to [44], concepts belong to the extra-linguistic knowledge about the world. Sowa [44] states: "*the concepts expressed by a language are determined by the environment, activities, and culture of the people who speak the language*". So, the use of concepts depends on who is doing that, for what purpose and in what context. For example, intending to analyze discourses of politicians, one may want to identify concepts like "progress", "problems", "investments", "money", "corruption", etc. By other side, in a psychiatric environment, concepts may be "violence", "drugs", "suicide", "death", etc. Soderland [43] defines concepts inside the weather forecasting domain. Each weather condition ("cloudy", "fair", "precipitation") is a concept and has its own definition. Also there are concepts to time and days.

The way in which concepts are represented in formalisms is suited to particular viewpoints. There are many and different approaches to express mental models. However, we are interested in a simple structure that allows us to represent real world objects, events, thoughts, opinions and ideas, easily and with a certain degree of quality for the discovery process. Following [4], [6] and [42], we use the *vector space model* to represent concepts internally. So each concept is stored as a set or vector of terms. Although it is possible to represent concepts as a network [6] or as an ordered list of terms [9], we have decided to use a non-ordered vector without links, assuming that all terms inside a concept description are related to each other in a same degree. The decision for this structure is to simplify classification and categorization tasks.

Terms in a concept (its descriptors) may include synonyms, quasi-synonyms, lexical variations, plural, verb derivations, semantic related words, etc. Associated to each term in a vector there must be a weight, ranging from 0 to 1 and describing the relative importance of the term to indicate whether a concept is present in a text. According to [4], this approach is better than the binary model since term count information leads to higher accuracy. Terms work like tokens, so it is not necessary that a term have a universal meaning, being possible to use proper nouns and abbreviations, even if they are meaningful only for the domain people (from here, we use "*term*" and "*word*" like synonyms).

Each concept has only one set as a descriptor, but one term may be present in more than one descriptor set. Currently, only single words are allowed. Although we know pairs of terms can give better results [1], our choice for using single words is due to computational limitations. Using simple representations of concepts reduces the time to perform classification and categorization. We expect to achieve good performance by the context analysis. The *fuzzy* function used in the categorization task tends to reward the presence of more than one word. Besides that, according to [1], using only pairs bring poor results, while single words alone are relatively successful.

One assumption is that the concept-based approach tends to minimize the *vocabulary problem* because concepts may be expressed with different words, as in a semantic expansion approach (see benefits of the semantic expansion technique for Information Retrieval in [3], [24] and [45]). So the efficiency of identifying concepts within a text is higher because more terms are covered. Chen [5] argues that people tend to use different terms to describe a similar concept. Furnas and partners [20] discuss the effectiveness of an "*unlimited aliasing*" strategy, which allows unlimited number of aliases for objects, to minimize the *vocabulary problem*. When examining Information Retrieval strategies, Bates [2] found that "*for a successful match, the searcher must somehow generate as much 'variety' in the search as is produced in indexing"*. This kind of *redundancy* allows identifying overlapping words similarly to how people express concepts and ideas (text authors and someone performing knowledge discovery in texts).

Another assumption is that the effort for concept definition and identification can be reduced. Thus knowledge engineers are not necessary and users do not need to expend too much effort and time to define models and rules to extract information, as when using *thesauri,* ontologies and natural language processing. Feldman and Dagan [15] defend the use of simple structures because they allow tasks to be performed with computer aids and with low costs.

In summary, we may compare the KDT approach against the KDD phases suggested in [22]:

a) understanding the application domain and the goals of the data mining process: user must define which concepts are interesting (first part of the classification task);
b) acquiring or selecting a target data set: texts must be gathered, using IR tools or in a manual way;
c) integrating and checking the data set: in our approach, texts must be saved in individual textual files (*.txt), no other validation is done;
d) data cleaning, preprocessing and transformation: concepts must be described (second part of the classification task) and texts need to be analyzed and stored in the internal format (vectors of words with weights representing the relative frequency), after eliminating *stopwords,* following suggestion of [4];
e) model development and hypothesis building: identifying concepts in the collection (the categorization task);

f) choosing suitable data mining algorithms: the application of the statistical techniques (mining task);
g) result interpretation and visualization: humans must interpret the findings;
h) result testing and verification: redoing the process or some stages to validate the discovered knowledge;
i) using and maintaining the discovered knowledge: done by humans.

The main tasks of the approach (categorization, classification and mining) are discussed in the next sections.

HOW TO IDENTIFY CONCEPTS INSIDE TEXTS (CATEGORIZATION)

According to [15], one kind of information extraction is the categorization of a text by meaningful concepts. The goal of the categorization is to identify concepts present in texts. However, documents do not have concepts explicitly, but rather words [1]. Once concepts are expressed by languages (words and grammars) [44], it is possible to identify them in texts by analyzing phrases.

Instead of using complex Natural Language Processing (NLP) to analyze syntax and semantics, our approach is based on a simple technique. We believe concepts may be identified by cues (terms). Using a *fuzzy* reasoning about the cues found in a text, we can calculate the likelihood of a concept being present in that text. This is a kind of Information Extraction, except that it is not necessary to fill fields with values. The extraction only needs to identify the presence of concepts. We consider that this approach is under the statistical NLP paradigm according to the definition in [25], since it uses frequency counting and probability theory. However, syntax analyses are not done.

The categorization algorithm follows Rocchio's one [41], since it uses a prototype-like vector (a centroid) to represent each class/category (in our case, the concepts) and evaluates the membership of an element (the text) in a class using a similarity function that calculates the distance between the element and each centroid. The choice for this algorithm is due to its simplicity, ease to implement and relative efficiency, according to [9]. Ragas and Koster [39] carried out experiments using four different algorithms and found that Rocchio's and Bayes algorithms achieved better results. They suggest a combination of both. The main disadvantage of these algorithms is that the context of words (near words) does not influence the categorization [9]. This may cause problems since the context may change the meaning of a word or the interpretation of a phrase (for example, "*is*" and "*is not*").

The algorithm starts comparing all texts against each concept, assuming that concepts were defined early and texts previously represented in the internal format. The comparison is done through a *fuzzy* reasoning process, following [49] and [37]. Weights of common terms (those present in both text and concept) are multiplied. The overall sum of these products, limited to 1, is the degree of relation between the text and the concept, meaning the relative probability of the concept presence in the text or that the text holds the concept with a specific degree of importance. Terms that are not present in the intersection of the representations are not counted because concept descriptors may be using synonyms.

The fundament behind this process is that each word of a concept contributes with certain strength to the presence of that concept. Strong indicators may receive higher weights in the concept definition (as will be discussed in the next subsection). Indeed, we are working with signs under uncertainty. This is like the relevancy index proposed in [40] whose definition is "*a collection of features that, together, reliably predict a relevant event description*". Similarly, Morris [36] distinguishes between indicator signs and characterizing signs. The first ones point to a specific object or element, while the last ones restrict elements in a set. Alike this thought, McCarthy [34] comments the use of approximate concepts. According to this author, there are sufficient and necessary conditions to certify the presence of a concept. Sufficient conditions (SC) works like the implication SC → CONCEPT, while necessary conditions (NC) are like CONCEPT → NC.

In our approach we consider that terms are characterizing signs and necessary conditions. Terms indicate the presence of a concept with a degree of certainty (TERM → CONCEPT). So the *fuzzy* reasoning must evaluate the likelihood of a concept to be present in a text, analyzing the strength of its indications. The process is like an abductive reasoning. According to [23], in a deduction, if "A → B" and "A is truth" then we can infer "B is truth". In abduction, if "A → B" and "B is truth" then "A is a probable cause for B being truth". That means if words that describe a concept appear in a text, there is a high probability of that concept being present in that text. The decision concerning if a concept is present or not depends then on the threshold used to cut off undesirable degrees. Riloff and Lehnert [40] evaluated three methods for identifying concepts in texts. Two of them consider that a concept is present if and only if there is a keyword or key phrase in the text, but they are prone to false hits due to the *vocabulary problem*. The

third method analyzes the context, using a relevancy degree. They concluded that the choice for one of these methods depends on the collection and language characteristics.

Our approach uses the threshold to decide whether a concept is present or not. As the user may set this threshold, it is possible that only one term indicates the concept presence. However, the more indicators are present, the more likely the concept is present. The decision is then done by the context analysis. Chakrabarti [4] believes that "*using a statistical method for text implies that the learned rules will not be dependent on the presence or absence of specific keywords*". This threshold may be chosen in a training session, before the categorization task.

HOW TO ACQUIRE CONCEPT DEFINITIONS (CLASSIFICATION)

The classification task is responsible by generating concept definitions, that is, the choice of concepts and the description of each concept (terms and their associated degrees of relevance).

Chen and partners [7] suggest to use either an existing controlled vocabulary (like dictionaries, *thesauri* or ontologies) or to automatically generate one. The main problem with *thesauri* is that they are usually very domain dependent and in some cases do not support slight variations because they do not have sufficient vocabulary coverage for all potential applications or specific user groups. Yang and Chute [47] reported problems with a medical *thesaurus*, because physicians used other words in their daily practice. In its turn, ontologies like WordNet [35] fail to include proper nouns. Although Liddy and partners [29] have demonstrated the benefits from using dictionaries, sometimes they do not include important semantic relations. In a previous study (not published yet), we found that definitions present in Webster-like dictionaries use too many general words, as for example in "*soccer = ball game played with feet, disputed by two teams with eleven players each..."* Examining the presence of the *soccer* concept in newspapers, we found that the listed words are not so frequent. By the other side, preexisting vocabularies may not be appropriated to the user's needs (lack of specificity or missing interesting concepts).

The automatic generation of a controlled vocabulary is a learning process [4] and can be done through either a supervised or an unsupervised process. The problem of the supervised process is that a high-quality sample of data must be available [1]. To find such a sample in an environment like the Web is a difficult task and demands a lot of time and effort.

For an unsupervised learning, Etzioni [14] suggests the clustering technique, which does not require labeled inputs. Fisher [18] states that the clustering process "*accepts object descriptions (events, observations, facts) and produces a classification scheme over the observations*". According to the same author, "*a learning of this kind is referred to as learning from observation (as opposed to learning from examples)*". However, classes are identified and created apart from the user's interest and this may not be appropriate to the application goal.

As we need a method that could be efficient (not necessarily the best) but mainly having low cost in terms of time and effort, we have chosen a manual process helped by dictionaries and software tools. We believe that, in the Web environment, users have *ad hoc* needs and do not want to spend time in computations or defining formal models. Besides that, in a manual process, concepts may be pruned to the users' interest.

However, our suggestion is that preexisting vocabularies, such as a *thesaurus* or a technical dictionary, if available, should be used to minimize the effort in this task. Besides that, a general dictionary (like a Webster's) may help the user to find synonyms. Bates [2], for example, proposes the use of a domain-specific dictionary to expand the user's vocabulary. Yang and Chute [47] showed the efficiency of the combination of a technical dictionary, like the CID (International Code/Classification of Diseases) in the medical area, augmented with synonyms specific of the domain.

Also it is important to examine some examples of the language style. In this way, software tools can play an important role, helping users to identify words used in the collection. According to [25], little samples can bring good results in some particular cases. Besides that, software tools minimize the knowledge acquisition bottleneck (according to [7], the cognitive demand required of humans to create controlled vocabularies).

As each word in a concept must have an associated value of importance, the user must define them too. Since it is difficult to assign numeric values, *fuzzy* linguistic variables may be used [49]. However, we use a software tool to help in the weight definition. This tool shows all the words present in a set of texts and the frequency of each one. Thus, user can examine a sample of the collection and verify which words are more common and in which context they occur. Lagus and Kaski [26] state that a good

descriptor must characterize some outstanding property and Salton and McGill [42] suggests that good descriptors are those that are frequent inside a text but infrequent in the whole collection (small inverse frequency). So, we suggest to assign small values to generic words or those present in more than one concept (or even eliminating this kind of word) and to assign higher values (for example, 1) to those that appear only in one concept description.

USING STATISTICAL TECHNIQUES ON CONCEPTS (MINING TASK)

The approach analyzes concept distributions to discover interesting patterns. This is like the IE+KDD paradigm, where IE is performed early to extract text attributes and KDD techniques are then used over these attributes. The difference is that we perform analyses over concepts instead of words or values (concepts work as text attributes). The approach may be considered under the probabilistic and statistical paradigm according to [32], since it is based on the distribution of variables in the collection. Following we discuss the techniques used for the mining task, assuming classification and categorization are finished. These techniques do not consider the degree of relation between a text and a concept (how much a concept is present in a text). We assume that it is only important to know if a concept is present or not inside a text.

The first technique used is the key-concept listing, which analyzes concept distributions over the collection. We have a software tool that extracts a concept-based centroid of a collection. After the categorization task, each text has associated to it a list of concepts with relative degrees. For each concept, the tool counts the number of texts to which the concept is assigned. The degrees are not considered in this step, because it does not matter how much a concept is present in a text but only if it is present (regarding that categorization must have cut off undesirable degrees according to the chosen threshold). This technique allows for finding which dominant themes exist in a collection or in a single text. Also we can compare one centroid to another, to find common themes or changes between sub-collections. Another possible usage is to find differences between sub-collections or concepts present in only one text. We followed Feldman and Dagan's [17] suggestion for examining distributions that differ significantly from the full collection, from other related collections or from collections in a different time.

The second technique is the association or correlation. It discovers associations between concepts and expresses these findings as rules in the format X ➔ Y (X may be a set of concepts or a unique one, and Y is a unique concept). The rule means "*if* **X** *is present in a text, then* **Y** *is present with a certain confidence and a certain support*". Following the definitions of [31] and [21], *confidence* is the proportion of texts that have X AND Y in relation to the number of texts that have only X, and *support* is the proportion of texts that have X AND Y in relation to all texts in the collection. Rules allow predicting the presence of a concept according to the presence of another one. Complex rules may be discovered with human intervention. So the precedent part of a rule may be a combination of concepts and/or words, such as WORD_1 AND WORD_2 AND CONCEPT_1 AND CONCEPT_2 => CONCEPT_3. This kind of rule is found using intermediary retrieval tasks, to select sub-collections where some words are present. In the next section, some examples will be explained.

EXPERIMENTS

We carried out some experiments with textual collections extracted from the Web to validate the concept-based approach presented here. In this paper, we discuss two experiments, one in a political analysis context and another for competitive intelligence (business intelligence) analysis over Text Mining tools.

The way in which the concepts were defined (classification task) was different in each experiment. Under a certain viewpoint, we can say that these two styles are complementary. In the political experiment, concepts were defined through an exhaustive analysis of words used in the collection. That means that we examined every word present in more than one text and classified it into a concept, resulting in a set of 104 concepts. Each word being examined could be classified into an existing concept or generate a new one. *Stopwords* and general terms were eliminated before this examination. By another side, in the competitive experiment, interesting concepts (according to the experiment goal) were first selected and then defined and refined through the examination of words present in the collection, giving a total of 24 concepts.

These experiments show how concepts may be defined. The first alternative generates a bigger set but helps the user to find which concepts are present in the collection. Besides that, this kind of definition intends to cover all relevant words in the collection. The second alternative, by other side,

narrows the set of concepts to only those relevant to the specific goal. This allows the user to choose previously the concepts of his/her interest and therefore tends to generate a smaller set. Classification task took less than 30 minutes in the first experiment and about 10 minutes in the second experiment. Both tasks were supported by a software tool that analyzed the words present in texts using a Pentium II 400 MHz with 64 Mbytes of RAM.

In these experiments, we used as interestingness measures a confidence threshold of 80% and a minimum support equals to 60% or 5 texts. Feldman and Hirsh [16] suggest a minimum support of 5 documents and a confidence threshold of 10% for the association rules.

POLITICAL EXPERIMENT

The goal of this experiment was to extract knowledge about what press is or was telling about the mayor of a big city in Brazil. To represent the press, an online newspaper was used. Texts were written in Portuguese. Using a local search engine and the mayor's name, we gathered 180 texts published in 1997 and 178 texts published in 1999, forming two sub-collections for a later comparison.

Examples of concepts definitions are: "*crimes*" = {crime, crimes, fraud, fraudulent, illegal…} and "*elections*" = {election, elections, term, reelection, voter, elected, electorate,…}.

The most interesting patterns and their interpretation are listed below according to the mining technique used. It is important to say that the mayor is under investigation by the Department of Justice on charges of corruption since the beginning of 2000.

Associative rules (association technique)

a) *drug traffic* ➔ *politicians* (confidence = 93.3%, support = 14 documents)
This means that when a press release dealt with "drug traffic", the name of a politician was cited too. We do not know why the names were cited, whether in a favorable way or not (accusing or accused), unless we read the texts. But this pattern allows us to conclude that the drug problem achieved the political sphere and a high importance degree, when the mayor is involved.

b) *loans* ➔ *politicians* (confidence = 82.1%, support = 23 documents)
This pattern was discovered in the 1997's sub-collection and means that references to "loans" of any kind involved the name of a politician. As the previous finding, it does not allow us to conclude the cause, but we may infer politicians are involved asking for, releasing, criticizing or receiving loans.

c) an interesting combination of 2 patterns,

(1) *loans* ➔ *politicians* (confidence = 82.1%, support = 23 documents)

(2) *education* ➔ *politicians* (confidence = 64.2%, support = 27 documents)

raised the hypothesis of a connection between "education" and "loans". When examining the direct relation between the two concepts, we found

(3) *education* ➔ *loans* (confidence = 4.7%, support = 2 docs)

(4) *loans* ➔ *education* (confidence = 7.1%, support = 2 docs).

These results lead us to conclude that there is not a direct relation between "loans" and "education". Probably, rules (1) and (2) happen in different subsets. However, when analyzing these two concepts together, the following rule was discovered,

(5) *loans* AND *education* ➔ *politicians* (confidence=83,3%, support=5 docs),

allowing for the conclusion that "politicians" are involved when "loans" and "education" are cited together, perhaps influencing decisions. However, not all "loans" with politicians' involvement are related to "education", because only 17,2% of the cases involving "loans" and "politicians" have "education", conform the next rule

(6) *loans* AND *politicians* ➔ education (confidence=17,2%).

Concept distributions (key-concept listing technique)

Analyzing the whole collection (358 texts), we found as the main themes: *politicians* (140 texts, 39.1%), *crimes* (117 texts, 32.6%) and *elections* (105 texts, 29.3%). From that, it is possible to infer that references to "crimes" are common when the mayor is cited, since the theme appears with frequency similar to political themes.

Comparing the distributions of concepts in 1997's to 1999's, some interesting observations arose:

a) comparing the two sub-collections, it can be observed that the 1997's sub-collection had a dominant focus (the presence of politicians associated with the mayor), while in 1999 the themes had a balanced distribution;
b) the weight of the "*elections*" concept rose from 25% (1997) to 33.7% (1999), possibly due to the nearing election in 2000 (the mayor's term finishes in December 2000);
c) the "*debts*" concept reduced its participation from 1997 to 1999, meaning there was a reduction in "debts" or the press changed its interest to other topics;
d) a particular interest exists to observe the distribution of names of people: an interesting pattern is that the presence of the mayor's main associate went down from 1997 to 1999; using our knowledge about the domain, we interpret this event as consequence of the political separation between them, during this time gap; also we can observe over this special sub-collection (when both politicians are cited together) that the main focus has changed from "debts" (40.7% => 12.5%) to "elections" (38.1% => 65%), and that references to "corruption" increased from 7.8% (in 1997) to 22.5% (in 1999), raising hypotheses that the separation was a cause or a consequence for these changes.

In every discovery process, the results from the mining task must be useful for some purpose. Here, we can state that these results may be used to establish political strategies. Examining the distributions of themes, one can evaluate how press is viewing (or manipulating) the events concerning the mayor. For example, "crimes" (32.6%) and "corruption" (10%) are so or more frequent than "education" (26.2%) and "investments" (10.6%). This may help mayor to take care about his declarations and actions or to make a marketing strategy to spread his work. By other side, when looking at the discovered rules, the mayor may decide to stay away from some politicians to avoid having his name associated to "drug traffic" or "loans", for example.

COMPETITIVE INTELLIGENCE EXPERIMENT

The second experiment had as main goal to compare Text Mining (TM) tools, examining the techniques used and the benefits cited by the vendors of these tools. Another goal was to relate techniques and benefits, in order to discover which techniques to use when needing a certain benefit.

Nine tools were selected using the Copernic meta-search engine (www.copernic.com) and the expression "text mining" as query. Texts about the tools were extracted from the linked web pages (initial and subsequent pages, only those telling about the tool). Which specific tools and URLs were used is not important in this discussion.

Concepts were defined as explained early, resulting in 13 concepts about techniques and 11 concepts corresponding to benefits. Techniques do not cover all the existing or possible ones, but were selected according to a previous survey (*summarization, extraction, key-concept listing, clustering, classification, retrieval, filtering, visualization, hypertext navigation, indexing, NL processing, correlation, sampling)*. The 11 benefits (*ease, support, automation, flexibility, analysis, quickness, completeness, consistency, efficiency, accuracy, conciseness)* were defined to cover all words present in the collection which can be semantically related to benefits. The most interesting patterns are presented below according to the mining technique used.

Concept distributions (key-concept listing technique)

a) the most used techniques are "key-concept listing" and "retrieval" (appearing in all cases) and the least used is "clustering";
b) two tools use all the defined techniques;
c) the most cited benefits were "support" and "ease";
d) there is a tool that alleged 10 benefits.

Associative rules (association technique)

When comparing techniques and benefits, 550 rules were found, from which 281 had confidence equals or greater than 80%. Eight rules had 100% of confidence and 90% of support. Examining these, we noted that "classification ⇔ automation" and "classification ⇔ accuracy" (the sign ⇔ means "if and only if"), perhaps indicating some inherent relation between them.

As in the previous experiment, the discovered knowledge needs to be useful for some purpose. One who intends to create a TM tool can analyze the trends above. If wanting to make something new, he/she must use the clustering technique. If wanting to compete with the existing tools, the most common techniques must be implemented. By other side, marketing strategies over benefits may be also

established, using the most cited. And someone who wants to implement a tool to offer a certain benefit should look at the rules comparing benefits versus techniques.

DISCUSSION

The discovered knowledge must be interpreted within the context associated to the concept definition. For example, the concept "*corruption*" may be presented in a situation where the cited mayor was involved in a possible crime or in a situation where the mayor reported a corruption case (for example). Besides that, the findings are relative to the collection. If the texts of the collection are not representative of the real world, we cannot assume that the rules will hold in any situation in a real situation. For example, in the experiments presented in this paper, the results of the KDT process hold only in those collections. Therefore, we cannot state that the same rules apply to the real world, once information present in the texts may be biased by the authors' style, interest, etc.

Given that words in a concept contribute with some weight to the presence of this concept in a text, the decision whether a concept is present or not depends directly on the chosen threshold, but indirectly on the words defined in the concept and their associated weight. We believe that the threshold can cut off undesirable results but may lead to mistakes. One alternative is to use a standard value, proved to be efficient. A future work will evaluate this possibility. Another alternative is to set the threshold using a training sample of the collection, examining errors and hits.

A problem was perceived when a word of a concept appeared in a negative phrase. For example, the expression "*unlike clustering*" erroneously leads to assume that the concept "*clustering*" (a technique in the Text Mining experiment) is present. This is the main problem with Rocchio's and Bayes algorithms, according to [9], once they analyze the whole context (in the entire text) but do not consider the local context (inside a phrase). One simple solution may be the use of negative terms in the concept description, as discussed by [40]. However, texts must be analyzed in the original format or internally represented in other way, perhaps using an ordered list of terms, as proposed by [9]. We know that ambiguity problems can be solved with NLP techniques, however, such algorithms are complex and time consuming, what goes against our initial goals. Our initial solution for this problem is the use of negative values in a concept definition. So, negative words must be selected and included in a concept description with a negative weight. Some experiments have shown that some false hits can be cut off. However, we are still dealing with the whole context and this may still cause false hits and thus more complex methods are necessary (we discuss our directions to solve this problem in the next section).

We also observed problems with ambiguous words. For example, in the TM experiment, the word "*class*" could describe a classification or a clustering technique, what could lead to mistakes. Our suggestion is that the user must examine phrases of a few texts (a sample of the collection) and reduce the weight assigned to this kind of words or eliminate them from the descriptions. We are studying the way in which weights are assigned to terms in a concept description. One alternative is to define them automatically through a training sample, using a supervised method. In an ongoing work, we have implemented different supervised methods for establishing word weights. Initial results suggest that words that appear in all texts within a training set must have a greater degree of importance. Also we are analyzing features such as types of terms (proper nouns, adjectives), sample size and hierarchies of concepts.

In order to compare our approach with other methods to define concepts, we analyzed the Latent Semantic Indexing (LSI), an automatic method that has achieved good results in categorization and information retrieval [12] [13]. The LSI is useful to find relations between terms, where human effort does not bring good results [13]. Thus, the synonymy problem can be solved. However, there are doubts that polysemy can be solved [38]. Deerwester and partners [12] say there is a "partial solution" due to the context analysis, but the consequence may be false hits, like the above example for "*class*". This is because LSI needs a good sample of texts for training, like the most automatic methods. In LSI approach, the sample must be "pure" (each text must be associated to only one category) and "separable" (with a low proportion of terms common to more than one category) [38]. The problem is that good samples (positive and representative cases) are difficult to find, especially in the Web and under the restriction of having only one category per document. Besides that, there is the additional effort (probably by humans) to evaluate the training set and this may also introduce errors.

Even when good samples are available for automatic methods, there is the possibility of negative words being included in a definition, as in the "*unlike*" example, or lexical variations being not considered. This happens because most automatic methods apply statistical techniques and do not use semantic

knowledge about the domain. Dumais [13] has conducted experiments using two different methods for acquiring definitions: one extracts words from natural language sentences describing categories and another uses training sets. Although the latter was best on average, there were 14 categories among 50 (28%) for which the former was better. This brings to discussion the relative efficiency of automatic methods.

One problem specific of the LSI approach is that it cuts off infrequent words. Although Yang [48] has showed that this may bring precision improvement in some cases, there is no prove that this will hold in whatever collection and with whatever training sample. Again we rely on good samples. In Yang's experiments, training sets were created with 20, 25 and 50% of the entire collection. Besides, there is the possibility of important infrequent terms (like lexical variations) being ignored and this may cause a great difference when the collection is composed of short texts.

Due to these problems, we chose to use a manual task supported by automatic tools and by existing vocabularies. Automatic methods can help user to find terms related to categories, lexical variations, local synonyms, frequencies and relations between terms. However, human intervention is important to solve ambiguities and to minimize errors. An important step is to examine samples of false hits (texts assigned to wrong categories), looking for terms that lead to errors, in order to refine the concept definitions. In this step, software tools may bring great benefits. Our suggestion is that the final decision should be responsibility of the user, working as a filter.

To evaluate our categorization method, we carried out formal experiments. The first one followed the idea of Goebel and Gruenwald [22], who have done benchmarking of KDD tools. We chose 5 tools from the original paper and defined 13 concepts, 8 relative to tasks and 5 relative to methods (tasks and methods are features evaluated in the original paper). The concepts were described selecting significant terms used in the paper to define each feature. We then gathered texts extracted from the Web pages referred by the paper. As these texts could have different information from those used by the authors in the original benchmarking, we used experts to decide the correct assignments (tools X features). Using Lewis' measures [27], we got the following results:

- microaveraging precision = 0.59;
- macroaveraging precision = 0.54;
- microaveraging recall = 0.95;
- macroaveraging recall = 0.86;
- fallout = 0.62.

Precision and fallout were not good, although in two concepts/categories we found high precision values (1 and 0.83). So, we carried out another experiment, this time performing classification with more accuracy. Some samples of false hits were analyzed and words that led to them (negative terms and ambiguous words) were eliminated from the definitions. With an adjustment in only one concept, the macroaveraging rates increased from 0.54 to 0.61 in precision and from 0.86 to 0.97 in recall (improvement of 13%). A second round was made, redoing all definitions with more accuracy (examining false hits). The task took about 10 minutes and the final results were:

- microaveraging precision = 0.65;
- macroaveraging precision = 0.69;
- microaveraging recall = 0.89;
- macroaveraging recall = 0.93;
- fallout = 0.28.

There was an increase of 10% in microaveraging precision and of 27% in macroaveraging precision. The fallout rate decreased 45%. Recall rates did not have great variance (-6% and +8%). These results show that categorization quality may be improved by refining more the concept definitions and that human intervention may eliminate precision failures.

To show that is possible to achieve a high precision rate, we performed an experiment with a bigger collection. The collection was composed of 100 texts corresponding to medical records of a psychiatric clinic, each one describing the admission interview of a different patient and written by a physician. We performed the categorization comparing each text (between 1 and 4 Kbytes) against 8 concepts. Concepts were manually defined by two experts of the domain (assistant physicians) along 2 months (not full time). The experts examined samples of texts aided by software tools and used Webster's and technical dictionaries to find synonyms. We estimate that the overall time took about 30 hours. Using as threshold the minor value higher than zero, we found an average error (wrong texts inside a concept) equal to 10.8%. Using Lewis' measures for precision [27], we got a microaveraging precision of 0.915 and a macroaveraging precision of 0.891. This reports an increase in the precision rates in comparison to the

previous experiments, showing that it is possible to achieve high quality by generating better concept descriptions.

Although the average degrees were good, a special attention must be given to measures in each category (concept). For example, in the medical collection evaluation, there was a concept that got a degree of errors equal to 32.5% while other concepts achieved zero error. The ideal situation is to obtain similar degrees in all categories (concepts) to avoid distorted conclusions.

CONCLUDING REMARKS

The paper presented an approach to perform knowledge discovery in textual collections. The process is based on concepts instead of words or attribute values, leading to more real findings, low cost processes and minimizing the *vocabulary problem*. The approach allows users to easily find interesting ideas, ideologies, trends and intentions present in texts, then being useful in sociological studies, discourse analysis, political marketing, competitive intelligence and so on.

An observed advantage is the reduced effort to define and identify concepts in texts, comparing to traditional NLP. The latter is very expensive because it performs complete analysis of a text [40] (using natural language processing) and because it requires large amounts of formally codified knowledge [25] (knowledge models and extraction rules). By other side, our approach uses simple algorithms and structures, helping people to find interesting patterns in a quick way. In the political experiment for example, classification and categorization did not take more than 40 minutes, remembering that 104 concepts were generated and compared against 358 texts in a Pentium II 400 MHz with 64 Mbytes of RAM. The mining task took about 2 minutes. Chinchor and partners [8] comment that the cost (effort) to adapt MUC-3 systems (classification task) to a new domain was of 10 to 11 man/month per system. It is important to say that the categorization algorithm does not work to extract attribute values, like IE systems, but only works to extract concepts when it is possible to infer them analyzing the presence of words in the whole text. In other situations, the approach should be adapted with other methods for categorization and classification.

We believe that the approach is better suited to interactive discoveries, because user does not need to expend too much effort to define concepts, and does not need to wait a long time for the categorization. Besides that, concept definitions may be refined at execution time. Thus, users with *ad hoc* needs (with ill-defined goals) can perform discovery without having to create formal rules or definitions as ontologies, *thesauri* and IE models. With little knowledge about the domain or even with help of software tools and dictionaries (like a Webster's or a technical one) it is possible to define and refine interesting concepts for a specific application or goal.

Although the problems discussed in the previous section, we believe that the discovery process may have quality if the categorization can be controlled. Based on the evaluations discussed early, we conclude that high quality can be achieved if the concepts are well defined (or refined). A good definition may be obtained if there are available experts, time and predefined vocabularies. Refinements are important and may be done by human intervention through analyzing false hits. The quality level depends on how much effort and resources the user wants or has to expend in the classification task. By other side, we can imagine that concept descriptions may evolve to a more accurate model as users become more familiar with the language used in the documents. McCarthy [34] states that approximate concepts may be refined by learning more or by defining more. But it is necessary to say that the improvement will come only to the specific domain and under the application goal as established by the user. We cannot expect that one specific model (for example, the same set of concept descriptions) will always achieve good results, because the Web is very dynamic and chaotic, as posed by [14]. However, even when concepts are ill defined (there is a lack of a precise definition), McCarthy [34] defends that this does not obstruct us to use and reason about concepts. All concepts are approximate, but they are precise for a certain purpose. So, the results from the KDT process are useful for analyzing trends and do not have compromise with the rigor of a scientific method. The discovered knowledge must be interpreted under this viewpoint.

Another remark is that categorization and classification tasks deserve more attention, especially because they may bias the KDT process. Thus, we are studying other algorithms within the same goals (simplicity, efficiency, ease to use, low cost in time and effort). An ongoing work is the implementation of a classification model using pairs of words and negative words and the implementation of a categorization algorithm that analyzes individual phrases (local context). So, concepts are described by a set of simple rules, each of these composed by positive and negative words. If a phrase has all positive words and no negative word, the concept is assumed to be present in the phrase. An overall computation determines

how much the concept is referenced in the whole text. This degree will be used in a further mining task. Initial experiments showed us that some extraction errors (false hits) could be solved. A formal evaluation will be performed to compare the efficiency of the two algorithms.

ACKNOWLEDGMENTS

This work is partially supported by: CNPq (Brazilian Council for Scientific and Technological Development), ULBRA (Lutheran University of Brazil) and UCPEL (Catholic University of Pelotas). We would like to thank our advisor Prof. Dr. José Palazzo Moreira de Oliveira and the anonymous referee for the valuable suggestions.

REFERENCES

[1] Apté, Chidanand; Damerau, Fred; Weiss, Sholom M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[2] Bates, M. J. Subject access in online catalogs: a design model. Journal of the American Society for Information Science, v.37, n.6, November 1986.

[3] Buckley, Chris; Salton, Gerard; Allan, James. The effect of adding relevance information in a relevance feedback environment. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. London: Springer-Verlag. 1994.

[4] Chakrabarti, Soumen. Data mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, v.1, n.2, January 2000.

[5] Chen, Hsinchum. The vocabulary problem in collaboration. IEEE Computer, special issue on CSCW, v.27, n.5, May 1994. Online at http://ai.bpa.arizona.edu/papers/ cscw94/cscw94.html

[6] Chen, Hsinchum e al. Automatic concept classification of text from electronic meetings. Communications of the ACM, v.37, n.10, October 1994. Online at http://ai.bpa.arizona.edu/papers/ ebs92/ebs92.html

[7] Chen, Hsinchum et al. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. Journal of the American Society for Information Science, v.47, n.8, August 1996. Online at http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html

[8] Chinchor, Nancy; Hirschman, Lynette; Lewis, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). Computational Linguistics, v.19, n.3, September 1993.

[9] Cohen, William W. and Singer, Yoram. Context-sensitive learning methods for text categorization. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval SIGIR-96. 1996. Online at http://www.research.att.com/~wcohen/ index.html

[10] Cowie, Jim and Lehnert, Wendy. Information extraction**.** Communications of the ACM, v.39, n.1, January 1996.

[11] Croft, W. Bruce. Machine learning and information retrieval. In: COLT '95 Conference. Lake Tahoe, July 1995. (invited talk) Online at http://www.ee.umd.edu/medlab/filter/

[12] Deerwester, Scott et alli. Indexing by latent semantic analysis. Journal of the Society for Information Science, v.41, n.6, 1990.

[13] Dumais, Susan T. Combining evidence for effective information filtering. In: AAAI Spring Symposium on Machine Learning and Information Retrieval, Tech Report SS-96-07, AAAI Press, March 1996.

[14] Etzioni, Oren. The world-wide web: quagmire or gold mine ? Communications of the ACM, v.39, n.11, November 1996.

[15] Feldman, Ronen and Dagan, Ido. Knowledge discovery in textual databases (KDT). In: 1st International Conference on Knowledge Discovery (KDD-95). Montreal, August 1995.

[16] Feldman, Ronen and Hirsh, Haym. Exploiting background information in knowledge discovery from text. Journal of Intelligent Information Systems, v.9, n.1, July/August de 1997.

[17] Feldman, Ronen and Dagan, Ido. Mining text using keyword distributions. Journal of Intelligent Information Systems, v.10, n.3, 1998.

[18] Fisher, Douglas H. Knowledge acquisition via incremental conceptual clustering. Machine Learning, v. 2, pp.139-172. 1987. Reprinted in Shavlik & Dietterich (eds.), Readings in Machine Learning, section 3.2.1.

[19] Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J. Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G.; Frawley, W. J. (eds.). Knowledge discovery in databases. MIT Press. 1991.

[20] Furnas, G. W. et al. The vocabulary problem in human-system communication. Communications of the ACM, v.30, n.11, November 1987.

[21] Garofalakis, Minos N. et al. Data mining and the web: past, present and future. In: ACM Workshop on Information and Data Management, Kansas City, 1999.

[22] Goebel, Michael and Gruenwald, Le. A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations, v.1, n.1, June 1999.

[23] Gulla, Jon A. et alli. An abductive, linguistic approach to model retrieval. Data & Knowledge Engineering, v.23, n.1, June 1997.

[24] Iivnen, Mirja. Searches and searches: differences between the most and least consistent searches. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval SIGIR'95. Washington: ACM PRESS, 1995.

[25] Knight, Kevin. Mining online text. Communications of the ACM, v.42, n.11, November 1999.

[26] Lagus, K. and Kaski, S. Keyword selection method for characterizing text document maps. In: Ninth International Conference on Artificial Neural Networks – ICANN'99, volume 1, pages 371-376, IEE, London (1999). Online at http://websom.hut.fi/websom/doc/ publications.html

[27] Lewis, David D. Evaluating text categorization. Proceedings of the Speech and Natural Language Workshop, Asilomar, February 1991. Online at http://www.research.att.com/ ~lewis

[28] Lewis, David D. and Hayes, Philip J. Guest editorial. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[29] Liddy, Elizabeth D.; Paik, Woojin; Yu, Edmund S. Text categorization for multiple users based on semantic features from a machine-readable dictionary. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[30] Lin, Chung-hsin and Chen, Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. IEEE Transactions on Systems, Man and Cybernetics, v. 26, n.1, February 1996. Online at http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

[31] Lin, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). 1998.

[32] Mannila, Heikki, Theoretical frameworks for data mining. ACM SIGKDD Explorations, v.1, n.2, January 2000.

[33] Mattox, David; Seligman, Len; Smith, Ken. Rapper: a wrapper generator with linguistic knowledge. In: ACM Workshop on Information and Data Management, Kansas City, 1999.

[34] McCarthy, John. Approximate objects and approximate theories. In: Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000). April 2000. Online at http://www-formal.stanford.edu/jmc

[35] Miller, George A. WordNet: A lexical database for English. Communications of the ACM, v.38, n.11, 1995

[36] Morris, Charles W. Foundations of the theory of signs. Rio de Janeiro, Eldorado Tijuca. 1976. (in Portuguese)

[37] Nakanishi, H.; Turksen, I. B.; Sugeno, M. A review and comparison of six reasoning methods. Fuzzy Sets and Systems, 57, 1993.

[38] Papadimitriou, Christos H. et alli. Latent Semantic Indexing: a probabilistic analysis. In: Seventeenth ACM SIGACT-SIGMOD-SIGART International Conference on Management of Data and Symposium on Principles of Database Systems (PODS). Seattle, June 1998..

[39] Ragas, Hein and Koster, Cornelis H. A. Four text classification algorithms compared on a Dutch corpus. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98). Melbourne, 1998.

[40] Riloff, Ellen and Lehnert, Wendy. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[41] Rocchio, J. J. Document retrieval systems - optimization and evaluation. Ph.D. Thesis, Harvard University, Report ISR-10 to National Science Foundation, Harvard Computation Laboratory. 1966.

[42] Salton, G. and McGill, M. J. Introduction to modern information retrieval. McGraw-Hill, 1983.

[43] Soderland, Stephen. Learning to extract text-based information from the world wide web. In: 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97). 1997.

[44] Sowa, John F. Knowledge representation: logical, philosophical, and computational foundations. Brooks/Cole Publishing Co., Pacific Grove, CA, 2000.

[45] Sparck-Jones, Karen. Assumptions and issues in text-based retrieval. In Jacobs, Paul S. (ed.) Text-based intelligent systems: current research and practice in information extraction and retrieval. New Jersey: Lawrence Erlbaum, 1992.

[46] Wiener, Erik D. et al. A neural network approach to topic spotting. In: 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), Las Vegas, 1995. Online at http://www.stern.nyu.edu/~aweigend/Research/Papers/ TextCategorization

[47] Yang, Yiming and Chute, Christopher G. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[48] Yang, Yiming. Noise reduction in a statistical approach to text categorization. In: ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) Seattle, 1995.

[49] Zadeh, Lotfi A. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man and Cybernetics, v. SMC-3, n.1, January 1973.

Bibliografia

[AAM95] AAMODT, Agnar; NYGARD, Mads. Different roles and mutual dependencies of data, information and knowledge - an AI perspective on their integration. **Data & Knowledge Engineering**, v.16, n.3, p.191-222, Setembro de 1995.

[AGR93] AGRAWAL, Rakesh; IMIELINSKI, Tomasz. Database mining: a performance perspective. **IEEE Transactions on Knowledge and Data Engineering**, v.5, n.6, p.914-925, Dezembro de 1993.

[AMB97] AMBROSIO, Ana P. et al. The linguistic level: contribution for conceptual design, view integration, reuse and documentation. **Data & Knowledge Engineering**, v.21, n.2, p.111-129, Janeiro de 1997.

[APT94] APTÉ, Chidanand; DAMERAU, Fred; WEISS, Sholom M. Automated learning of decision rules for text categorization. **ACM Transactions on Information Systems**, v.12, n.3, p.233-251, Julho de 1994.

[BAT86] BATES, Marcia J. Subject access in online catalogs: a design model. **Journal of the American Society for Information Science**, v.37, n.6, p.357-376, Novembro de 1986.

[BOW96] BOWDEN, Paul R.; HALSTEAD, Peter; ROSE, Tony G. Extracting conceptual knowledge from text using explicit relation markers. In: EUROPEAN KNOWLEDGE ACQUISITION WORKSHOP, 9., 1996, Nottingham. **Proceedings...** Heidelberg: Springer-Verlag, 1996. p.147-162. (Lecture Notes in Artificial Intelligence, n.1076).

[BRA98] BRADLEY, Paul; FAYYAD, Usama M.; REINA, Cory A. Scaling clustering algorithms to large databases. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4., 1998, New York. **Proceedings…** New York: AAAI Press, 1998. p.9-15.

[BUC94] BUCKLEY, Chris; SALTON, Gerard; ALLAN, James. The effect of adding relevance information in a relevance feedback environment. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 7., 1994, Dublin. **Proceedings...** London: ACM/Springer-Verlag. 1994. p.292-300.

[CEN89] CENTRO BRASILEIRO DE CLASSIFICAÇÃO DE DOENÇAS (Centro Colaborador da OMS para a Classificação de Doenças em Português). **Classificação Internacional de Doenças e de Problemas Relacionados a Saúde**. Décima Revisão. São Paulo: EDUSP, 1989. Disponível por WWW em http://www.datasus.gov.br/cid10/cid10.htm

[CHA00] CHAKRABARTI, Soumen. Data mining for hypertext: a tutorial survey. **SIGKDD Explorations**, v.1, n.2, p.1-11, Janeiro de 2000.

[CHE93] CHEN, Z. Let documents talk to each other: a computer model for connection of short documents. **Journal of Documentation**, v.49, n.1, p.44-54, Março de 1993.

[CHE94] CHEN, Hsinchum. The vocabulary problem in collaboration. **IEEE Computer**, v. 27 , n. 5, p.2-10, Maio de 1994. Disponível por WWW em http://ai.bpa.arizona.edu/papers/cscw94/cscw94.html

[CHE94b] CHEN, Hsinchum e al. Automatic concept classification of text from electronic meetings. **Communications of the ACM**, v.37, n.10, p.56-73, Outubro de 1994. Disponível por WWW em http://ai.bpa.arizona.edu/papers/ebs92/ebs92.html

[CHE97] CHEN, Hsinchum et al. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. **Journal of the American Society for Information Science**, v.48, n.1, p.17-31, Janeiro de 1997. Disponível por WWW em http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html

[CHI93] CHINCHOR, Nancy; HIRSCHMAN, Lynette; LEWIS, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). **Computational Linguistics**, v.19, n.3, p.409-449, Setembro de 1993.

[CHO97] CHOUDHURY, Vivek; SAMPLER, Jeffrey L. Information specificity and environmental scanning: an economic perspective. **MIS Quarterly**, v.21, n.1, p.25-50, Março de 1997.

[COH96] COHEN, William W.; SINGER, Yoram. Context-sensitive learning methods for text categorization. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 9., 1996, Zurich. **Proceedings…** Washington: ACM Press, 1996. p.307-315. Disponível por WWW em http://www.research.att.com/~wcohen/ index.html

[COW96] COWIE, Jim; LEHNERT, Wendy. Information extraction**. Communications of the ACM**, v.39, n.1, p.80-91, Janeiro de 1996.

[CRO92] CROFT, W. Bruce; TURTLE, Howard R. Text retrieval and inference. In: JACOBS, Paul S. (ed) **Text-based intelligent systems: current research and practice in information extraction and retrieval**. New Jersey: Lawrence Erlbaum, 1992, p.127-155.

[CRO95] CROFT, W. Bruce. Machine learning and information retrieval (invited talk). In:

COLT CONFERENCE, 1995, Lake Tahoe. **Proceedings…** San Francisco: Morgan Kaufmann, 1995. p.587. Disponível por WWW em http://www.ee.umd.edu/medlab/filter/

[CRO94] CROSS, Valerie. Fuzzy information retrieval. **Journal of Intelligent Information Systems**, v.3, n.1, p.29-56, Fevereiro de 1994.

[DAV89] DAVIES, Roy. The creation of new knowledge by information retrieval and classification. **Journal of Documentation**, v.45, n.4, p.273-301, Dezembro de 1989.

[DEE90] DEERWESTER, Scott et al. Indexing by latent semantic analysis. **Journal of the American Society for Information Science**, v.41, n.6, p.391-407, Setembro de 1990.

[DUM96] DUMAIS, Susan T. Combining evidence for effective information filtering. In: AAAI SPRING SYMPOSIUM ON MACHINE LEARNING IN INFORMATION ACCESS, 1996, Stanford. **Proceedings…** Menlo Park: AAAI Press, 1996. p.26-31.

[ETZ96] ETZIONI, Oren. The world-wide web: quagmire or gold mine ? **Communications of the ACM**, v.39, n.11, p.65-68, Novembro de 1996.

[FEL95] FELDMAN, Ronen; DAGAN, Ido. Knowledge discovery in textual databases (KDT). In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY, 1995, Montreal. **Proceedings…** Cambridge: AAAI/MIT Press, 1995, p.112-117.

[FEL97] FELDMAN, Ronen; HIRSH, Haym. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, v.9, n.1, p.83-97, Julho/Agosto de 1997.

[FEL98] FELDMAN, Ronen; DAGAN, Ido. Mining text using keyword distributions. **Journal of Intelligent Information Systems**, v.10, n.3, p.281-300, Maio/Junho de 1998.

[FEL98b] FELDMAN, Ronen et al. Text mining at the term level. In: EUROPEAN SYMPOSIUM ON PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY, 2., 1998, Nantes. **Proceedings…** Heidelberg: Springer-Verlag, 1998, p.65-73. (Lecture Notes in Computer Science, n. 1510). Disponível por WWW em http://www.wisdom.weizmann.ac.il/~lindell/

[FIS87] FISHER, Douglas H. Knowledge acquisition via incremental conceptual clustering. **Machine Learning**, v.2, n.2, p.139-172, 1987.

[FUR87] FURNAS, G. W. et al. The vocabulary problem in human-system communication.

Communications of the ACM, v.30, n.11, p.964-971, Novembro de 1987.

[GAL00] GALAVOTTI, L. et al. Feature selection and negative evidence in automated text categorization. In: WORKSHOP ON TEXT MINING, 2000, Boston. **Proceedings…** Washington: ACM Press, 2000. (pôster) Disponível por WWW em www.cs.cmu.edu/~dunja/wshkdd2000.html

[GAR99] GAROFALAKIS, Minos N. et al. Data mining and the web: past, present and future. In: ACM WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 1999, Kansas City. **Proceedings…** Washington: ACM Press, 1999. p.43-47.

[GOE99] GOEBEL, Michael; GRUENWALD, Le. A survey of data mining and knowledge discovery software tools. **SIGKDD Explorations**, v.1, n.1, p.20-33, Junho de 1999.

[GRO00] GROBELNIK, Marko; MLADENIC, Dunja; MILIC-FRAYLING, Natasa. Text mining as integration of several related research areas: report on KDD'2000 Workshop on Text Mining. **SIGKDD Explorations**, v.2, n.2, p.99-102, Dezembro de 2000.

[GUL97] GULLA, Jon A. et al. An abductive, linguistic approach to model retrieval. **Data & Knowledge Engineering**, v.23, n.1, p.17-31, Junho de 1997.

[HER95] HERSH, William R. et al. Towards new measures of information retrieval evaluation. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1995, Seattle. **Proceedings...** Washington: ACM Press, 1995. p.164-170.

[HOB79] HOBBS, Jerry R. Coherence and coreference. **Cognitive Science**, v.3, n.1, Janeiro/Março de 1979.

[HWA92] HWANG, Chung H.; SCHUBERT, Lenhart K. Tense trees as the "fine structure" of the discourse. In: MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 30., 1992, Newark. **Proceedings...** ACL, 1992. p.232-240.

[IIV95] IIVNEN, Mirja. Searches and searches: differences between the most and least consistent searches. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1995, Seattle. **Proceedings...** Washington: ACM PRESS, 1995. p.149-157.

[ING96] INGWERSEN, Peter. Congnitive perspectives of information retrieval interaction: elements of a cognitive IR theory**. Journal of Documentation**, v.52, n.1, p.3-50, Março de 1996.

[JEN00] JENSEN, L.S.; MARTINEZ, T. Improving text classification by using conceptual and contextual features. In: WORKSHOP ON TEXT MINING, 2000, Boston. **Proceedings…** Washington: ACM Press, 2000. (pôster). Disponível por WWW em www.cs.cmu.edu/~dunja/wshkdd2000.html

[KAM93] KAMEYAMA, Megumi; PASSONNEAU, R.; POESIO, M. Temporal centering. IN: MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 31, 1993, Columbus. **Proceedings...** ACL, 1993. p.70-77.

[KNI99] KNIGHT, Kevin. Mining online text. **Communications of the ACM**, v.42, n.11, p.58-61, Novembro de 1999.

[KUH91] KUHLTHAU, Carol C. Inside the search process: information seeking from the user's perspective. **Journal of the American Society for Information Science**, v.42, n.5, p.361-371, Junho de 1991.

[LAG99] LAGUS, Krista; KASKI, S. Keyword selection method for characterizing text document maps. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS, 9., 1999, Edinburg. **Proceedings...** London: IEEE, 1999. p. 371-376. Disponível por WWW em http://websom.hut.fi/websom/doc/publications.html

[LAS92] LASCARIDES, Alex, ASHER, N.; OBERLANDER. J. Interferring discourse relations in context. In: MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 30., 1992, Newark. **Proceedings...** ACL, 1992. p.1-8.

[LEW91] LEWIS, David D. Evaluating text categorization. In: SPEECH AND NATURAL LANGUAGE WORKSHOP, 1991, Asilomar. **Proceedings...** San Mateo: Morgan Kaufmann, 1991. p.312-318. Disponível por WWW em http://www.research.att.com/~lewis

[LEW94] LEWIS, David D.; HAYES, Philip J. Guest editorial. **ACM Transactions on Information Systems**, v.12, n.3, p.231, Julho de 1994.. Disponível por WWW em http://www.research.att.com/~lewis/papers

[LEW94b] LEWIS, David D.; GALE, William A. A sequential algorithm for training text classifiers. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 7., 1994, Dublin. **Proceedings...** London: ACM/Springer-Verlag, 1994. p.3-12. Disponível por WWW em http://www.research.att.com/~lewis/papers

[LEW96] LEWIS, David D.; SPARCK-JONES, Karen. Natural language processing for information retrieval. **Communications of the ACM**, v.39, n.1, p.92-101,

Janeiro de 1996.

[LEW98] LEWIS, David D. Naive (bayes) at forty: The independence assumption in information retrieval. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 9., 1998, Chemnitz, Alemanha. **Proceedings...** Heidelberg: Springer-Verlag, 1998. p.4-15. (Lecture Notes in Computer Science, v.1398). Disponível por WWW em http://www.research.att.com/~lewis/papers

[LID94] LIDDY, Elizabeth D.; PAIK, Woojin; YU, Edmund S. Text categorization for multiple users based on semantic features from a machine-readable dictionary. **ACM Transactions on Information Systems**, v.12, n.3, p.278-295, Julho de 1994.

[LIM97] LIMA, Luciano R. S.; LAENDER, Alberto H. F.; RIBEIRO NETO, Berthier A. Um modelo para recuperação de informação especializada aplicado a bases de dados médicas semi-estruturadas. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 1997, Fortaleza. **Anais…** Fortaleza: UFC, 1997. p.241-256.

[LIN96] LIN, Chung-hsin; CHEN, Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. **IEEE Transactions on Systems, Man and Cybernetics**, v. 26, n.1, p.1-14, Fevereiro de 1996. Disponível por WWW em http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

[LIN98] LIN, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. **Proceedings...** Washington: ACM Press, 1998. p.241-249.

[LOH99] LOH, Stanley. **Descoberta de conhecimento em textos**. Exame de Qualificação EQ-29. PPGC/UFRGS, Porto Alegre, Fevereiro de 1999.

[MAA92] MAAREK, Yoëlle S. Automatically constructing simple help systems from natural language documentation. IN: JACOBS, Paul S. (ed) **Text-based intelligent systems: current research and practice in information extraction and retrieval**. New Jersey: Lawrence Erlbaum, 1992. p.243-256.

[MAN00] MANNILA, Heikki, Theoretical frameworks for data mining. **SIGKDD Explorations**, v.1, n.2, p.30-32, Janeiro de 2000.

[MAR88] MARCHIONINI, Gary; SHNEIDERMAN, Ben. Finding facts vs. browsing knowledge in hypertext systems. **Computer**, v.21, n.1, p.70-80, Janeiro de 1988.

[MAT99] MATTOX, David; SELIGMAN, Len; SMITH, Ken. Rapper: a wrapper generator with linguistic knowledge. In: ACM WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 1999, Kansas City. **Proceedings...** Washington: ACM Press, 1999. p.6-11.

[MCC00] McCARTHY, John. Approximate objects and approximate theories. In: INTERNATIONAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING, 7., 2000, Breckenridge. **Proceedings...** San Francisco: Morgan Kaufmann, 2000. p.519-526. Disponível por WWW em http://www-formal.stanford.edu/jmc

[MCK95] McKEOWN, Kathleen; RADEV, Dragomir R. Generating summaries of multiple news articles. IN: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1995, Seattle. **Proceedings...** Washington: ACM Press, 1995. p.74-82.

[MII94] MIIKE, Seiji et al. A full-text retrieval system with a dynamic abstract generation funcion. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 7., 1994, Dublin. **Proceedings...** London: ACM/Springer-Verlag. 1994. p.152-161.

[MIL95] MILLER, George A. WordNet: A lexical database for English. **Communications of the ACM**, v.38, n.11, p.39-41, Novembro de 1995.

[MOR76] MORRIS, Charles W. **Fundamentos da teoria dos signos**. Rio de Janeiro: Eldorado Tijuca, 1976.

[MOR91] MORRIS, Jane; HIRST, Graeme. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. **Computational Linguistics**, v.17, n.1, p.21-48, Março de 1991.

[MOR98] MORGADO, Lina. **O lugar do hipertexto na aprendizagem: alguns princípios para a sua concepção**. Disponível por WWW em: http://www.moderna.com.br (setembro de 1998).

[MOS98] MOSCAROLA, Jean; BAULAC, Yves; BOLDEN, Richard. **Technology watch via textual data analysis**. Université de Savoie, 1998. Note de Recherche nº 98-14.

[NAK93] NAKANISHI, H.; TURKSEN, I. B.; SUGENO, M. A review and comparison of six reasoning methods. **Fuzzy Sets and Systems**, v. 57, n.3, p.257-294, Agosto de 1993.

[NON97] NONAKA, I.; TAKEUCHI, H. **Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação**. Rio de Janeiro:

Campus, 1997.

[OAR96] OARD, Douglas W.; MARCHIONINI, Gary**. A conceptual framework for text filtering**. University of Maryland, 1996. Technical Report EE-TR-96-25. Disponível por WWW em http://www.ee.umd.edu/medlab/filter/ (junho de 1998).

[OWE97] OWENS, Janet; RAGAINS, Patrick. **Evaluating information sources**.. Disponível por WWW emhttp://www.library.unr.edu/~ragains/eval.html (janeiro de 1997).

[PAP98] PAPADIMITRIOU, Christos H. et al. Latent Semantic Indexing: a probabilistic analysis. In: ACM SIGACT-SIGMOD-SIGART INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA AND SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS, 7., 1998, Seattle. **Proceedings...** Washington: ACM Press, 1998. p.159-168.

[PAR89] PARSAYE, Kamran et al. **Intelligent databases: object-oriented, deductive hypermedia technologies**. New York: John Wiley & Sons, 1989.

[PAR96] PARSONS, Simon. Current approaches to handling imperfect information in data and knowledge bases. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n.3, p.353-372, Junho de 1996.

[PED93] PEDRYCZ, Witold. Fuzzy neural networks and neurocomputations. **Fuzzy Sets and Systems**, v.56, n.1, p.1-28, Maio de 1993.

[RAG98] RAGAS, Hein; KOSTER, Cornelis H. A. Four text classification algorithms compared on a Dutch corpus. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1998, Melbourne. **Proceedings...** Washington: ACM Press, 1998. p.369-370.

[RIL94] RILOFF, Ellen; LEHNERT, Wendy. Information extraction as a basis for high-precision text classification. **ACM Transactions on Information Systems**, v.12, n.3, p.296-333, Julho de 1994.

[ROC66] ROCCHIO, J. J. **Document retrieval systems - optimization and evaluation**. Harvard University, 1966. (tese de doutorado)

[SAG95] SAGGION, Horacio; CARVALHO, Ariadne. Análise textual visando a tradução automática. In: CONFERÊNCIA LATINO-AMERICANA DE INFORMÁTICA, 21., 1995, Canela. **Anais...** Porto Alegre: UFRGS, 1995. p.201-212.

[SAL83] SALTON, G.; McGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.

[SAR00] SARMENTO, Cristiane S.; LOH, Stanley (orientador). **Avaliação de métodos para agrupamento de documentos textuais**. Canoas: Universidade Luterana do Brasil, Curso Superior de Tecnologia em Informática. Novembro de 2000. (trabalho de conclusão)

[SCH96] SCHOLZ, Ann. **Evaluating World Wide Web information**. Disponível por WWW em http://thorplus.lib.purdue.edu/research/classes/gs175/3gs175/evaluation.html. (Fevereiro de 1996).

[SMI97] SMITH, Alastair. **Criteria for evaluation of Internet Information Resources**. Disponível por WWW em http://www.vuw.ac.nz/~agsmith/evaln/index.htm. (Março de 1997).

[SOD97] SODERLAND, Stephen. Learning to extract text-based information from the world wide web. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 3., 1997, Newport Beach. **Proceedings...** Menlo Park: AAAI Press, 1997. p.251-254.

[SOW00] SOWA, John F. **Knowledge representation: logical, philosophical, and computational foundations**. Pacific Grove: Brooks/Cole Publishing Co., 2000. Disponível por WWW em http://www.bestweb.net/~sowa/krbook/index.htm

[SPA92] SPARCK-JONES, Karen. Assumptions and issues in text-based retrieval. In JACOBS, Paul S. **Text-based intelligent systems: current research and practice in information extraction and retrieval**. New Jersey: Lawrence Erlbaum, 1992, p.157-177.

[SPA97] SPARCK-JONES, Karen; WILLET, Peter (eds). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.

[SWA97] SWANSON, Don R. Historical note: information retrieval and the future of an illusion. In: [SPA97]. p.555-561.

[SWA97b] SWANSON, Don R.; SMALHEISER, N. R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence**, v.91, n.2, p.183-203, Abril de 1997.

[SUB00] SUBASIC, P.; HUETTNER, A. Calculus of fuzzy semantic typing for qualitative analysis of text. In: WORKSHOP ON TEXT MINING, 2000, Boston. **Proceedings...** Washington: ACM Press, 2000. (pôster). Disponível por WWW em www.cs.cmu.edu/~dunja/wshkdd2000.html

[TAN99] TAN, Ah-Hwee. Text mining: the state of the art and the challenges. In: PACIFIC-ASIA WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999, Beijing. **Proceedings...** Heidelberg: Springer-Verlag, 1999. p.65-70. (Lecture Notes in Computer Science v.1574). Disponível por WWW em http://textmining.krdl.org.sg/publications.html

[WAN99] WANDERLEY, A.V.M. Um instrumento de macropolítica de informação: concepção de um sistema de inteligência de negócios para gestão de investimentos de engenharia. **Ciência da Informação**, v.29, n.2, p.190-199, Maio/Agosto de 1999. Disponível por WWW em http://www.ibict.br/cionline

[WAT97] WATTS, Robert J.; PORTER, Alan L. Innovation forecasting. **Technological Forecasting and Social Change**, v.56, p.25-47, 1997.

[WEB88] WEBBER, Bonnie L. Tense as discourse anaphor. **Computational Linguistics**, v.14, n.2, p.61-73, Junho de 1988.

[WIE94] WIEBE, Janyce M. Tracking point of view in narrative. **Computational Linguistics**, v.20, n.2, p.233-287, Junho de 1994.

[WIE95] WIENER, Erik D. et al. A neural network approach to topic spotting. In: ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, 4., 1995, Las Vegas. **Proceedings...** 1995. p.317-332. Disponível por WWW em http://www.stern.nyu.edu/~aweigend/Research/Papers/ TextCategorization

[WIL00] WILCOX, A. et al. Using knowledge sources to improve classification of medical text reports. In: WORKSHOP ON TEXT MINING, 2000, Boston. **Proceedings...** Washington: ACM Press, 2000. (pôster). Disponível por WWW em www.cs.cmu.edu/~dunja/wshkdd2000.html

[WIL94] WILKINSON, Ross. Effective retrieval of structured documents. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVA, 7., 1994, Dublin. **Proceedings...** London: ACM/Springer-Verlag, 1994. p.311-317.

[WIL88] WILLET, Peter. Recent trends in hierarchic document clustering: a critical review. **Information Processing & Management**, v.24, n.5, p.577-597, 1988.

[WIV99] WIVES, Leandro Krug; OLIVEIRA, José Palazzo M (orientador). **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de "clustering"**. Porto Alegre: PPGC/UFRGS, 1999. (Dissertação de Mestrado).

[YAN94] YANG, Yiming; CHUTE, Christopher G. An example-based mapping method for text categorization and retrieval. **ACM Transactions on Information Systems**, v.12, n.3, p.252-277, Julho de 1994.

[YAN97] YANG, Yiming; PEDERSEN, Jan O. A comparative study on feature selection in text categorization. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 14., 1997, Nashville, USA. **Proceedings...** San Francisco: Morgan Kaufmann, 1997. p.412-420.

[YAN99] YANG, Yiming; LIU, Xin. A re-examination of text categorization methods. In: INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1999, Berkeley. **Proceedings...** Washington: ACM Press, 1999. p.42-49

[ZAD73] ZADEH, Lotfi A. Outline of a new approach to the analysis of complex systems and decision processes. **IEEE Transactions on Systems, Man and Cybernetics,** v. SMC-3, n.1, p.28-44, Janeiro de 1973.

[ZAN98] ZANASI, Alessandro. Competitive Intelligence through datamining public sources. **Competitive Intelligence Review**, v.9, n.1, 1998.