Revista de
Informática
Teórica e Aplicada

Uma Abordagem para Busca Contextual de Documentos na Internet

Stanley Loh[1], Leandro Krug Wives[2] e Antônio Severo Frainer[3]
{loh, wives}@inf.ufrgs.br frainer@portoweb.com.br

Abstract

This work presents an approach that constructs automatic tools to help users in searching documents in Internet. Contextual use of the words is analyzed in order to eliminate ambiguities and to discover topics interesting to users. Through successive refinements on searches and with user's feedback, the set of documents converge to the most relevant ones. The main contribution of this work is the discussion on related problems and the viability of solutions to information retrieval using a simple semantic search.

Resumo

Este artigo apresenta uma abordagem para construção de ferramentas automatizadas que auxiliam usuários na procura por documentos disponíveis na Internet. Tal abordagem considera principalmente o Contexto dos termos fornecidos como entrada, eliminando ambigüidades e precisando melhor o interesse do usuário. Por refinamentos sucessivos através do *feedback* do usuário, o conjunto dos documentos localizados vai convergindo para os mais desejados. A principal contribuição deste trabalho é a análise dos problemas envolvidos e da viabilidade de soluções ao se utilizar um mecanismo simples de busca semântica baseado em conjuntos de palavras definindo contextos.

Palavras-chave: recuperação de informações, análise de contexto, busca contextual, Internet

[1] Professor Adjunto de ULBRA e UCPEL, doutorando no CPGCC/UFRGS
[2] Mestrando no CPGCC/UFRGS
[3] Professor Adjunto na ULBRA

1. Introdução

Com a difusão da rede INTERNET, aumentaram também o armazenamento e a procura de informações pelos diversos sites desta rede. Agora as pessoas podem obter respostas a suas dúvidas, pesquisando documentos espalhados pela rede, que tratam dos mais diversos assuntos.

Entretanto, tamanho volume de documentos e informações acaba por trazer problemas na hora da pesquisa. [3] cita a frustração dos usuários com o problema da "sobrecarga de informações". Ela ocorre 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.

Já existem ferramentas, tais como AltaVista® e Yahoo®, que auxiliam a busca de informações na Internet. Ferramentas como estas permitem a localização de documentos textuais a partir de palavras ou catálogos de assuntos. Os usuários fornecem as palavras desejadas ou escolhem assuntos de seu interesse, e as ferramentas retornam os documentos correspondentes, bem como os *sites* onde se encontram estes documentos.

Nas principais ferramentas, duas técnicas de recuperação são utilizadas para este processo de busca:

- a indexação de termos ou palavras: que geralmente é full-text, ou seja, todas as palavras dos documentos tornam-se disponíveis no índice, e o usuário recebe como resposta os documentos que contêm as palavras fornecidas como entrada;
- a catalogação de documentos: que ocorre quando alguma pessoa define o assunto do documento; o usuário então precisa escolher um entre os assuntos já pré-definidos para então receber os documentos relativos àquele assunto (catalogados dentro do assunto).

Algumas variações da primeira técnica são comuns. Por exemplo, em algumas ferramentas, há uma linguagem própria para consulta, que utiliza conetivos e símbolos lógicos para eliminar documentos com determinados termos ou para recuperar somente documentos que contenham obrigatoriamente certos termos (na falta de informações mais precisas, bastará ao documento, para satisfazer a consulta, conter um dos termos fornecidos).

Apesar da incontestável utilidade destas ferramentas, alguns problemas podem ocorrer. Primeiro, a maioria dos usuários que utilizam as ferramentas de localização é inexperiente ou leiga, tanto no assunto que procuram quanto na utilização da ferramenta em si. Portanto, têm dificuldades em definir o contexto da informação que necessitam utilizando palavras e conetivos (segundo [14], a utilização de conetivos não é prática, principalmente se a consulta é muito complexa).

Também ocorre de as diversas ferramentas agruparem as informações de diferentes maneiras. Um estudo apresentado em [7] demonstra que aqueles que conhecem o funcionamento interno da ferramenta, e possuem mais experiência com a linguagem de consulta (que é também específica da ferramenta) têm mais facilidade de encontrar

informações úteis.

Além disto, as ferramentas que utilizam a técnica de indexação retornam grandes volumes de documentos sem a certeza de que a informação desejada se encontra em um deles. Isto acontece porque a técnica de indexação é baseada unicamente na presença de termos nos documentos. Assim, podem ser retornados documentos que contêm as palavras fornecidas, mas que se referem a outro contexto, devido à possibilidade de as palavras terem vários significados diferentes. Outro problema é que poderão deixar de ser recuperados documentos relevantes para o assunto escolhido, justamente porque não possuem os termos fornecidos.

Quando a segunda técnica é utilizada, podem ocorrer problemas quando o especialista cataloga documentos de forma errada (por exemplo, interpretando equivocadamente o conteúdo de um documento) ou quando o usuário não consegue encontrar um assunto (dos pré-definidos) que represente precisamente seus interesses de busca (já que as pessoas podem utilizar termos diferentes para associar a mesma idéia ou termos iguais para idéias diferentes – este problema é conhecido como "abismo semântico").

Além disto, uma vez que as ferramentas recuperam um grande número de documentos (na ordem de milhares, em média), haverá a necessidade de refinamentos sucessivos até que o usuário encontre as informações desejadas. Isto ocorre porque, geralmente, o usuário fornece inicialmente apenas alguns poucos termos para pesquisa. À medida que os documentos vão sendo retornados ao usuário, deverão ser feitas novas análises sobre os mesmos (ou pelo próprio usuário ou por ferramentas), para filtrar um conjunto menor de saída.

Uma das alternativas para este último problema é listar os documentos de saída de maneira ordenada segundo algum critério. Por exemplo, os documentos com maior freqüência (presença) dos termos poderão estar no topo da lista.

Este artigo apresenta uma abordagem para um Sistema Inteligente de Busca de Documentos na Internet que utiliza o Contexto para precisar o assunto de interesse do usuário. Através da análise de contexto, as ferramentas de busca poderão precisar melhor o significado dos termos fornecidos pelo usuário para a busca e poderão também filtrar melhor os documentos candidatos à resposta.

2. Uso do Contexto para a Busca

Quando o usuário escolhe palavras ou tópicos para descrever os assuntos pelos quais se interessa, dois problemas principais podem ocorrer:

1. usuário pode ter escolhido termos ou tópicos não adequados para representar suas idéias e interesses;
2. a ferramenta de busca pode entender equivocadamente o significado das palavras fornecidas.

Isto ocorre freqüentemente porque pessoas diferentes podem utilizar os mesmos termos para idéias diferentes ou então utilizar termos diferentes para as mesmas idéias. A

causa de tais imprecisões e ambigüidades está no âmago do processo de comunicação: é impossível transmitir significados; só signos podem ser transmitidos. O significado está na mente das pessoas e não nas marcas gráficas ou nos comprimentos físicos de onda (conforme [13] e Hayakawa em [6]).

Para amenizar tais problemas, deverá ser analisado também o contexto em que os termos são usados. **Contexto** é o resultado da análise do texto pela pessoa (segundo [11]). A comunicação obtém sucesso quando são analisados não só os elementos físicos envolvidos, mas também os aspectos sociais, humanos, emotivos, etc, que acompanham o processo de comunicação. [6] acrescenta: "*o comunicador efetivo é aquele que concerne as idéias e não as palavras*". O Contexto, portanto, é tudo aquilo que envolve um processo de comunicação, seja ou não transmitido explicitamente.

A análise do Contexto será indispensável para o bom entendimento dos termos, caso contrário, a busca poderá retornar documentos não relevantes ou deixar de retornar documentos interessantes para o usuário. [5] sugere que as técnicas para indexação devem gerar índices sensitivos mais intimamente relacionados ao real significado de um texto em particular e não baseados na presença de termos sem identificação do contexto. Outros autores confirmam:

[8] "quanto mais rico for o Contexto de uma mensagem, mais limitada será a perda de informação";

[10] "a ambigüidade das palavras é resolvida, pelos humanos, pelo entendimento do Contexto".

Na abordagem aqui descrita, o Contexto (ou espaço conceitual) será definido por um conjunto de palavras que representam o assunto ou a área do conhecimento (conforme a definição de [4]).

3. Visão Geral da Abordagem

A figura 1 apresenta uma visão geral dos componentes e dos processos da abordagem para busca contextual de documentos na Web. Há três ferramentas. A primeira (interface cooperativa) interage com o usuário recebendo suas consultas (expressas em palavras para busca), apresentando os documentos resultantes e refazendo o processo de interação através de novas seleções de palavras pelo usuário. Esta ferramenta também documenta todo o processo de interação, bem como os documentos da Web escolhidos pelo usuário para visualização.

A segunda (e mais importante) ferramenta (a de Busca Contextual) realiza a comunicação com algum software de indexação já disponível na Internet (no momento, o sistema AltaVista® está sendo utilizado). Antes porém, esta ferramenta deve determinar o espaço de busca, ou seja, o contexto dos termos fornecidos, a fim de evitar mal-interpretações de significados. Para tanto, será utilizada a Base de Contextos. A última ferramenta é a que permite a Definição dos Contextos (na Base de Contextos), ou com a intervenção de um especialista humano ou automaticamente como será descrito mais tarde.

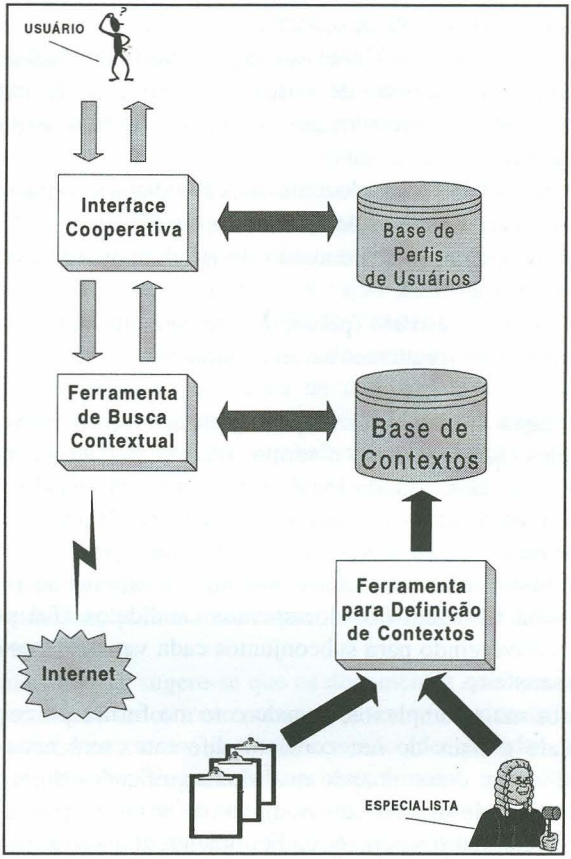

Figura 1- Visão geral dos componentes e dos processos

A seguir, os principais passos e componentes da abordagem serão descritos em detalhes.

3.1 A Ferramenta de Busca Contextual

A Ferramenta de Busca Contextual é responsável por identificar, a partir dos termos fornecidos pelo usuário, o conjunto de termos que serão usados para a busca de documentos na Internet. Este conjunto de termos deve ser o que melhor define o Contexto do assunto desejado e será extraído da Base de Contextos.

Para achar os documentos na rede, esta ferramenta se utiliza dos sistemas de indexação já disponíveis (como AltaVista®, Yahoo®, etc), passando como parâmetros as palavras do contexto (com os símbolos lógicos correspondentes) e recebendo como resposta as URL's dos documentos candidatos.

Na situação mais simples, a Ferramenta de Busca segue os seguintes passos:

1. recebe uma palavra do usuário;
2. procura na Base de Contextos as palavras relacionadas àquela;
3. envia os parâmetros de busca a um sistema de indexação (sendo que os termos serão associados por conjunção, ou seja, basta a presença de um dos termos);
4. recebe as URL's dos documentos candidatos a resposta;
5. busca diretamente os documentos apontados;
6. realiza a análise do conteúdo destes documentos (verifica a freqüência dos primeiros termos);
7. apresenta ao usuário (pela interface cooperativa) os documentos com os seus termos mais freqüentes e a sua pontuação.

A pontuação é determinada pela presença dos termos de busca. A fórmula utilizada é simples: cada vez que o termo aparece no documento, soma-se um ponto. Técnicas mais sofisticadas, usando lógica *fuzzy*, estão em estudo, e resultados parciais de uma ferramenta implementada isoladamente das demais podem ser encontrados em [16]. Uma lista será então apresentada ao usuário, que poderá consultar diretamente os documentos ou então poderá selecionar um novo conjunto de palavras (fornecendo um *feedback)* para uma filtragem dos documentos candidatos. Tal processo pode-se repetir inúmeras vezes, convergindo para subconjuntos cada vez menores de documentos, até que o usuário esteja satisfeito.

Em casos mais complexos, quando o termo fornecido como entrada pelo usuário estiver relacionado a mais de um contexto diferente, será necessário escolher um dos contextos para a busca, determinando qual dos significados do termo é o mais adequado. Para tanto, o contexto do usuário (um conjunto de termos representativos) será consultado na Base de Perfis de Usuários, através da Ferramenta de Interface Cooperativa.

Nesta segunda situação, este conjunto de termos será usado para determinar qual o contexto do interesse do usuário. A técnica utilizada é a da referência cruzada entre os conjuntos com pontuação pela presença dos termos.

No momento, nem todas os passos da ferramenta estão implementados (por exemplo, a parte do *feedback* e dos refinamentos sucessivos por interação com o usuário).

3.2 A Ferramenta para Definição de Contextos

A montagem de contextos manual é delicada, devendo ser tarefa de um especialista no assunto no qual deseja-se identificar o contexto. No fundo, montar um contexto nada mais é do que selecionar as palavras que identificam o assunto.

Por exemplo, se o contexto a ser montado pertence à área de Medicina quem deve selecionar as palavras que definem o contexto é um Médico. Isso porque esta pessoa já está acostumada com o assunto e consegue identificar quais são as palavras que são mais importantes na descrição do assunto. O Médico sabe quais são os termos empregados por seus colegas e todas as outras pessoas que trabalham na mesma área, aumentando assim a

abrangência (capacidade de recuperar todos os documentos relevantes ao assunto) e a precisão (capacidade de recuperar somente documentos relevantes) do contexto.

No caso de uma doença, por exemplo, o que o especialista faz é descrever a doença, indicando seus nomes (sinônimos), sintomas e procedimentos. Assim, cada uma destas palavras será utilizada na busca de documentos que pertençam ao contexto da doença em questão. Cada documento que possuir uma das palavras do contexto (e este documento pode ser um prontuário médico), será retornado.

O maior problema da construção de contextos manual é que ela exige tempo e disposição. Além disto, o especialista pode não ser capaz de identificar todas as palavras relevantes do contexto ou aquela que define o contexto ("cabeça" do conjunto).

Esse tipo de problema pode ser amenizado com a definição automática dos contextos. A princípio, a montagem automática de contextos foi desenvolvida visando substituir o especialista. Apesar disto, sugere-se que esta seja uma etapa complementar à montagem manual, sendo usada para facilitar o trabalho do especialista.

Esta técnica permite a identificação de relações entre palavras e contextos diretamente dos documentos (fontes de informação). Para esta analise é necessário um módulo especial dotado de métodos estatísticos que são capazes de identificar quais são as palavras que estão relacionadas com um contexto qualquer. Estes métodos estatísticos são aplicados sobre os documentos, pois é nestes que poderão ser encontradas quais são as palavras que definem seus assuntos.

Em trabalho anterior [15], sugere-se que os documentos que vão ser analisados na montagem automática pertençam todos a um mesmo contexto (ou assunto). Assim, todas as palavras e relações pertencentes a um único contexto são identificadas. Desta forma, os contextos vão sendo analisados separadamente, um após o outro, tornando a análise mais rápida e eficiente. Para isto, o especialista deve analisar previamente os documentos separando-os por contexto. Há a possibilidade ainda de se utilizar definições já prontas de contextos, como por exemplo dicionários técnicos.

As técnicas empregadas na extração de relações automáticas entre palavras de um mesmo contexto é similar à técnica utilizada na montagem de *Thesaurus* (detalhes em [4]). São métodos estatísticos, que baseiam-se na análise de ocorrência das palavras nos documentos.

Ao todo o processo de montagem automática de contextos possui três etapas distintas: - *a identificação de palavras nos documentos,*

- *a determinação do grau de relação entre as palavras e o documento que as contém,*

- *a análise das relações entre as palavras.*

Segundo *Chen* [3] as palavras que aparecem repetidamente em um único documento e as palavras que aparecem em muitos documentos são boas candidatas. É claro que nem todas as palavras devem ser indexadas. As palavras conhecidas como "*stop-words*", conforme [4], não devem ser adicionadas. As *stop-words* são palavras comuns a todos os textos (por exemplo, artigos e preposições) e portanto não são específicas do assunto tratado pelo documento. As *stop-words* podem variar. Dependendo do domínio a

ser analisado, verbos ou até mesmo expressões podem ser desprezadas.

Após selecionadas as palavras que devem fazer parte do processo, é realizada uma análise de co-ocorrência das palavras nos documentos. É através desta análise que é possível definir o grau de relação de cada palavra com o contexto em questão.

Duas fórmulas são utilizadas para esta análise: a fórmula que analisa o grau de relação entre uma palavra e um documento, e a fórmula que analisa as relações entre palavras (definindo assim os contextos).

A fórmula abaixo define a relação entre uma palavra e o documento em que ela aparece, onde d_{ij} é o valor combinado da palavra *j* no documento *i*:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j}\right)$$

N representa o número total de documentos considerados, tf_{ij} é a freqüência da palavra *j* no documento *i* e df_j é a freqüência inversa de documentos (número de documentos em que a palavra *j* aparece).

A segunda fórmula avalia os resultados gerados pela fórmula anterior, detectando as relações entre as palavras:

$$\text{Valor combinado} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}}, \text{ onde } d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}}\right),$$

sendo que tf_{ijk} representa o número de ocorrências de ambos as palavras *j* e *k* no documento *i* (o menor número de ocorrências entre as palavras deve ser escolhido), df_{jk} representa o número de documentos (em uma coleção de *N*) no qual as palavras *j* e *k* ocorrem ao mesmo tempo.

Desta forma, é possível identificar as relações entre as palavras em diversos contextos, criando uma estrutura (a *Base de Contextos*) capaz de indicar o quanto uma palavra está relacionada com outra em determinado contexto.

Tendo-se esta informação é possível identificar a que contexto determinado documento pertence. É possível também estabelecer um grau de pertinência do documento a determinado contexto, isto é, caso identifique-se mais de um contexto em um único documento, é possível estabelecer *quanto* este documento está relacionado com um e com outro contexto.

É importante lembrar que duas palavras podem estar relacionadas entre si em mais de um contexto, e portanto, em cada contexto existe um grau de relação diferente.

Uma versão inicial da ferramenta de Definição de Contextos foi implementada e testada. Em [15], são relatados experimentos com esta ferramenta, utilizando documentos

com informações sobre os sintomas que podem ser causados pela utilização de drogas. O contexto de cada droga é montado e após utilizado em prontuários médicos a fim de localizar pacientes com sintomas similares. Os resultados foram satisfatórios porque as relações entre as palavras, identificadas automaticamente pela ferramenta, foram consideradas adequadas e realistas por especialistas médicos.

3.3 A Base de Contextos

A Base de Contextos é uma estrutura que armazena os relacionamentos entre as palavras de um mesmo contexto. Assim, é possível percorrer esta estrutura e identificar quais são as palavras que pertencem a um determinado contexto. É possível também saber o quanto uma palavra está relacionada com o contexto ou também o quanto uma palavra está relacionada com outra palavra em determinado contexto.

A implementação atual da base de contextos, como está sendo utilizada pela ferramenta de Busca Contextual, é muito simples. Cada contexto é nomeado e identificado por uma palavra e contém um conjunto de palavras que o representam.

Alternativas para melhorar tal estrutura estão sendo testadas.

Uma implementação possível é estruturar a Base de Contextos como uma rede semântica, onde os nodos são as palavras e os elos representam relações entre palavras de um mesmo contexto. Entretanto, como pode haver o caso de uma palavra estar relacionada com duas outras em contextos diferentes, há a necessidade de se caracterizar o contexto de cada elo.

Em [15], é apresentada uma implementação baseada nesta última alternativa. Assim, duas palavras podem estar relacionadas entre si nos mais diversos contextos. Cada elo possui uma indicação do contexto da relação (inclusive determinando o tipo da relação; por exemplo, causa, efeito, sinônimo, antônimo, etc) e um valor *fuzzy*, o qual caracteriza o grau de associação entre as palavras correspondentes, extraído das fórmulas discutidas na seção anterior. Na figura 2, pode-se ver um exemplo desta situação: o termo "jogador" está associado ao termo "bola" por dois contextos diferentes, e as duas associações possuem graus diferentes de intensidade ou importância.

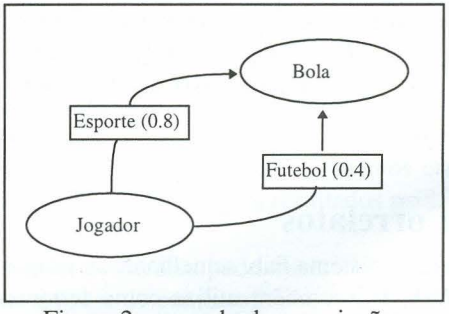

Figura 2: exemplo de associações

Alguns problemas foram detectados, impedindo a continuação do uso deste tipo de alternativa para a ferramenta de busca. O principal deles é que o contexto de cada elo pode ser especificado por uma palavra, mas isto traz de volta o problema da escolha desta palavra, gerando buscas contextuais recursivas. Da mesma forma, ainda não se tem uma maneira adequada para representar o conceito de "contexto" que não o uso de uma única palavra.

3.4 A Interface Cooperativa e a Base de Perfis de Usuários

A Interface Cooperativa registrará todas as interações do usuário com o sistema, mantendo um histórico na chamada Base de Perfis de Usuários. Devem ser documentadas todas as pesquisas realizadas, os documentos recuperados, aqueles que foram lidos e os que foram rejeitados, as palavras utilizadas pelo usuário para a pesquisa, etc, além de algum tipo de identificação do usuário.

Estas informações serão repassadas à ferramenta de busca contextual para que possa inferir objetivos e planos do usuário, a fim de "filtrar" os documentos recuperados, apresentando apenas as informações relevantes ao interesse do usuário.

Uma questão que sempre surge quando da utilização de um modelo do usuário, é referente a seu conteúdo inicial. As alternativas, com algumas variações, se resumem a iniciar com o modelo vazio ou com um conteúdo predeterminado, igual para todos os novos usuários. Iniciar com o modelo vazio reduz o poder de inferência da ferramenta nas interações iniciais. Iniciar com um conteúdo predeterminado pode levar a problemas como a inadequação do modelo ao usuário. Neste trabalho, optou-se pelo modelo inicial vazio, por considerar seu custo menor.

No momento, a implementação da base de perfis apenas contempla o conteúdo das interações. Ou seja, estão sendo armazenados somente a identificação do usuário e um conjunto de palavras que define seus interesses. Portanto, não estão sendo armazenados os tipos de informações (se documentos retornados ou termos fornecidos como entrada), mas somente o seu conteúdo descrito por palavras. Também não estão sendo considerados aspectos de tempo (consultas antigas, consultas mais recentes) e os resultados rejeitados.

As palavras que definem o perfil ou contexto do usuário estão sendo extraídas dos termos fornecidos pelo usuário e dos termos mais freqüentes dos últimos documentos selecionados pelo usuário. Caso haja repetidos processos de *feedback,* estão sendo desconsiderados os aspectos intermediários da interação ferramenta-usuário, na implementação atual.

4. Trabalhos Correlatos

Em [1], é descrito o sistema Fab, semelhante ao proposto aqui para auxiliar na busca de documentos na Web. Fab também utiliza como ferramentas de indexação aquelas já disponíveis na Internet. A técnica de busca também utiliza duas bases de perfis: uma de perfis de usuários (com informações de preferências dos indivíduos) e uma de perfis de

tópicos (criada automaticamente a partir de combinações das bases individuais, e contendo, para cada tópico, uma lista de palavras e seus respectivos graus de importância no tópico).

Fab utiliza ainda um mecanismo de "*feedback*", pelo qual o usuário avalia (segundo uma escola de pontuação) os resultados fornecidos. Esta retroalimentação serve para refinar as bases de perfis (a dos usuários individualmente e a dos tópicos).

O sistema Fab considera-se um sistema do tipo **colaborativo**, além de ser também do tipo **baseado em conteúdo**. Enquanto este tipo concentra esforços na análise de conteúdos, aquele procura combinar o conhecimento e a experiência pessoal de vários indivíduos para realizar as buscas e fazer suas recomendações (sugestões de documentos).

A diferença principal entre o sistema Fab e o proposto aqui é que o primeiro utiliza como entrada a escolha de um tópico (denotando o assunto a ser pesquisado), enquanto que o último recebe como entrada palavras (uma ou mais), cujos contextos serão analisados para se decidir o(s) tópico(s) a ser(em) pesquisado(s).

A vantagem da abordagem proposta neste artigo é que há menos problemas de interpretação das entradas fornecidas pelo usuário, ao se considerar o contexto. Em seus experimentos, o sistema Fab apresentou problemas em alguns resultados devido à ambigüidade dos termos.

Também para tratar o problema de busca contextual, há a técnica de [2], a qual se utiliza de expansões semânticas de palavras. Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto para busca de documentos.

[2] utiliza as definições de um dicionário para achar as palavras que se relacionam, eliminando *stop-words* e modela estas relações através de redes semânticas, criadas manualmente.

Já o Sistema Referral Web [9] usa a combinação de termos próximos para precisar o significado dos mesmos, evitando erros de busca devido à ambigüidade das palavras.

5. Conclusões

Já que não se têm todos os componentes da abordagem trabalhando de maneira integrada e como as ferramentas ainda encontram-se em processo de prototipagem, não foi possível ainda realizar um experimento formal seguindo-se toda a abordagem. Apesar disto, algumas considerações iniciais podem ser feitas com base nos resultados parciais.

A ferramenta de Busca Contextual tem-se mostrado útil ao escolher um conjunto maior de termos para busca e ir refinando o espaço de busca, através da convergência dos documentos que vão sendo localizados. Ao oferecer documentos candidatos e auxílios ao usuário para a tomada de decisão, a ferramenta gera resultados melhores, mesmo que mais demorados.

Além disto, o significado dos termos pode ser melhor precisado com a ajuda das relações entre os termos (tanto na base de contextos, quanto na base de perfis), diminuindo assim os erros por ambigüidade. Experimentos com textos retirados de artigos e reportagens de jornais e sobre prontuários médicos conduzem a estas conclusões. Uma

avaliação mais rigorosa deverá ser feita para determinar o grau de acerto na recuperação dos documentos. Para tanto, podem ser utilizados os critérios de Abrangência (*recall*, que avalia se todos os documentos relevantes foram recuperados) e Precisão (*precision*, que avalia se somente os documentos relevantes foram recuperados e nada mais), definidos por [12].

Cabe salientar que o sucesso desta abordagem depende em muito de como a base de contextos é criada. Uma boa base permitirá melhores interpretações dos interesses do usuário, enquanto que uma base pobre ou mal-definida ocasionará erros no processo de busca (retorno de documentos não desejados ou falta de documentos importantes).

Quando a Base de Contextos é criada por um especialista, a probabilidade de erros pode ser maior (por razões já discutidas anteriormente). Quando a ferramenta de Definição dos Contextos faz esta definição automaticamente, a partir de documentos predeterminados, diminui-se a incerteza, pois são utilizadas técnicas já consagradas na literatura para determinar as palavras representantes de um assunto e podem ser usados volumes maiores de documentos para análise.

Entretanto, tais técnicas somente terão resultados satisfatórios se o conjunto-amostra para extração das relações entre os termos for bem escolhido. De novo, recai-se na dependência de um especialista humano. Também poderá haver problema se o termo escolhido para definir o contexto ("cabeça" do conjunto) não for apropriadamente escolhido. Em parte, tal situação pode ser contornada com o uso de sinônimos.

Uma solução a ser implementada futuramente é a de incrementar e refinar a base de contextos automaticamente a partir de entradas de vários especialistas ou por análise dos documentos selecionados por vários usuários diferentes. Assim, os contextos seriam definidos de maneira a combinar o conhecimento de vários especialistas, tornando a ferramenta também um Sistema Colaborativo.

Problemas também ocorrem quando uma palavra aparece como "cabeça" num contexto e como elemento em outro. Futuramente, serão utilizados conjuntos e operadores *fuzzy* (como definido na lógica *fuzzy* de [17]) para determinar o grau de uma relação *x-z* (caso somente se disponha das relações *x-y* e *y-z*) e para especificar graus diferentes de pertinência dos termos nos contextos.

Já a Base de Perfis de Usuários, por sua vez, pode levar a ferramenta a conclusões equivocadas. Isto pode ocorrer quando o usuário procura informações em um contexto diferente daquele que a ferramenta inferiu ou quando o usuário realmente quer alterar seu assunto de busca. Uma das alternativas possíveis é consultar o usuário toda vez que houver algum conflito a ser resolvido. Desta forma, a ferramenta se tornaria um "assistente de consulta".

Outra limitação da abordagem exposta é que as consultas não levam em conta a sintaxe e a semântica entre os termos, mas apenas o contexto no qual se inserem. Por exemplo, se forem fornecidas como entrada as palavras "ferramentas" + "Internet", podem ser recuperados documentos que tratam de "ferramentas exclusivamente para o uso da Internet" ou de "quaisquer ferramentas disponíveis na Internet". Os refinamentos sucessivos através do *feedback* do usuário podem compensar estas desvantagens.

Referências Bibliográficas

[1] BALABANOVIC, M.; SHOHAM, Y. "Fab: content-based, collaborative recommendation". Communications of the ACM, v.40,n.3, Mar 1997.

[2] CHAKRAVARTHY, Anil S.; HAASE, Kenneth B. *"NetSerf: using semantic knowledge to find Internet information archives"*. **Proceedings.** SIGIR, 1995.

[3] CHEN, Hsinchun. *A textual database/knowledge-base coupling approach to creating computer-supported organizational memory.* MIS Department, University of Arizona, 5 de Julho de 1994. (http:// ai.bpa.arizona.edu/papers/)

[4] CHEN, Hsinchun et alli. *A concept space approach to addressing the vocabulary problem in scientific information retrieval: na experiment on the worm community system.* MIS Department, University of Arizona, 2 de Julho de 1996. (http://ai.bpa.arizona.edu/ papers/)

[5] COWIE, Jim; LEHNERT, Wendy. *"Information extraction"*. **Communications of the ACM**, v.39, n.1, Jan 1996.

[6] DAVIS, K. **Human behavior at work: human relations and organizational behavior**. McGraw-Hill. 4a.ed. 1972.

[7] IIVONEN, M. *Searches and Searches: Differences Between the Most and Least Consistent Searches.* ***SIGIR FORUM*** *95.* p149-157. 1995

[8] JAKOBSON, R. **Lingüística e comunicação**. Ed. Cultrix. s.d.

[9] KAUTZ, H.; SELMAN, B.; SHAH, M. *"Referral web: combining social networks and collaborative filtering"*. **Communications of the ACM,** v.40,n.3, Mar 1997.

[10] KENT, W. **Data and reality**. North-Holland. 1978.

[11] KONITZ, Ruth et alli. "*O signo e sua tipologia*". IN: AZEVEDO, M. C. (coordenador). **Atenção - signos - graus de informação**. Série Cadernos Universitários, n.4. Ed. UFRGS. 1973.

[12] SALTON, G. ***Introduction to Modern Information Retrieval.*** McGraw-Hill. New York. 1983.

[13] STEWART, D. K. **A psicologia da comunicação**. Ed. Forense. 1972.

[14] WILLIE, S; BRUZA, P. *Users' Model of the Information Space: the Case for Two Search Models.* ***SIGIR FORUM*** *95.* P205-211. 1995

[15] WIVES, L. K. *Um modelo de hiperdicionário: estudo de caso em prontuários médicos.* Curso de Bacharelado em Ciências da Computação, UCPEL. Dezembro de 1996. (Trabalho de Conclusão)

[16] WIVES, L. K.; LOH, S. *Recuperação de informações usando a expansão semântica e a lógica difusa.* **Anais.** IV Congresso Internacional en Ingenieria Informatica - ICIE'98. Buenos Aires, Argentina, 16-17 de Abril de 1998.

[17] ZADEH, L. A. *Fuzzy sets.* **Information and Control**, 8, pp.338-353. 1965.

Agradecimentos

[1] Este trabalho é parcialmente apoiado por CNPq/PROTEM e FAPERGS. Os autores aproveitam para homenagear, *in memorian*, o Prof. Dr. José M. V. de Castilho.