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

Um Estudo Comparativo de Ferramentas de Descoberta de Conhecimento em Texto: a análise da Amazônia

por

ANA CARLA MACEDO DA SILVA

Dissertação submetida à avaliação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Dr. José Palazzo Moreira de Oliveira
Orientador

Porto Alegre, maio de 2002.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Silva, Ana Carla Macedo
Um Estudo Comparativo de Ferramentas de Descoberta de Conhecimento em Texto: a análise da Amazônia/ Ana Carla Macedo da Silva — Porto Alegre: PPGC da UFRGS, 2002.
110 p.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegra, BR-RS, 2002. Orientador: Olivereira, José Palazzo Moreira de.

1. Recuperação de Informações. 2. Descoberta de Conhecimento em Texto. 3. Técnicas de Agrupamento. 4. Avaliação de Sistemas de Recuperação de Informações. I. Oliveira, José Palazzo Moreira de. II. Título.

Agradecimentos

Ao professor Dr. José Palazzo Moreira de Oliveira, que me acolheu como orientanda com toda sua vontade e apesar da distância sempre deu a atenção necessária e os conselhos certos para a realização deste trabalho.

Ao Ms. Leandro Krug Wives, que cedeu o Eurekha e muito contribuiu muito para este trabalho.

À Irvana dos Santos Coutinho, bibliotecária e especialista em Sistema de Informação da Universidade Federal do Pará, que tem co-autoria neste texto.

Aos meus queridos avós (em memória), à minha mãe, ao meu irmão e às minhas tias pelo incentivo e apoio.

Ao meu marido pela compreensão e apoio.

A todos os colegas, que compartilharam as dificuldades e alegrias. Em especial, à Raquel Trindade Borges, amiga de Curso e de local de trabalho.

À agência CAPES, pelo fomento.

À Universidade Federal do Pará e à Universidade Federal do Rio Grande do Sul, pela realização do Mestrado Interinstitucional.

Ao Instituto de Informática da UFRGS, pela utilização de suas dependências, e todo seu pessoal, sempre disposto a cooperar.

Sumário

Lista de Abreviaturas **6**
Lista de Figuras **7**
Lista de Tabelas **10**
Resumo **11**
Abstract **12**
1 Introdução **13**
2 Aspectos relevantes da Recuperação de Informações e Descoberta de Conhecimento em Textos **15**
2.1 Recuperação de Informações (RI) **15**
2.1.1 Paradigma da Recuperação de Informação 15
2.1.2 Sistemas de Recuperação de Informações (SRI) 16
2.2 Recuperação de Informações Textuais **18**
2.2.1 Estruturas de dados de Recuperação de Informação 18
2.2.2 Modelos de Recuperação de Informações Textuais 21
2.2.3 Indexação 25
2.2.4 Ferramentas de auxílio 28
2.2.5 Medidas de avaliação de Recuperação de Informações 29
2.3 Descoberta de Conhecimento em Textos **34**
2.3.1 Definição 35
2.3.2 Etapas do Processo 37
2.3.3 Tipos de Descoberta de Conhecimento em Textos 37
2.3.4 Mineração da Web ("Web mining") 40
3 Estudo de Caso **42**
3.1 Sistemas de Descoberta de Conhecimento em Texto utilizados **43**
3.1.1 Eurekha 5.1 43
3.1.2 Umap 45
3.2 Seleção dos dados e pré-processamento **48**
3.3 Experiência com o Eurekha **49**
3.3.1 Teste com a coleção inteira 51
3.3.2 Teste com as publicações do período de nove meses 53
3.3.3 Teste com as publicações do 1º semestre 53
3.3.4 Teste com as publicações do 1º trimestre 54
3.3.5 Teste com as publicações do mês de janeiro 55
3.3.6 Conhecimento descoberto com o Eurekha 55
3.4 Experiência com Umap **56**
3.4.1 Teste com a coleção inteira 57
3.4.2 Teste com as publicações do período de nove meses 59
3.4.3 Teste com as publicações do 1º semestre 60
3.4.4 Teste com as publicações do 1º trimestre 62
3.4.5 Teste com as publicações do mês de janeiro 63
3.4.6 Conhecimento descoberto com o Umap 63
3.5 Análise dos resultados **64**
3.5.1 Tempo de processamento 64

3.5.2 O grau de similaridade do Eurekha 64
3.5.3 Efetividade dos agrupamentos 67
3.6 Diferenças entre Eurekha e Umap 78
4 Conclusão 82
Bibliografia 85
Anexo Resultados dos Experimentos 88

Lista de Abreviaturas

RI	Recuperação de Informações
SRI	Sistema de Recuperação de Informações
KDT	Knowledge Discovery from Text (Descoberta de Conhecimento em Texto)
KDD	Knowledge Discovery in Database (Descoberta de Conhecimento em Banco de Dados)
SGBD	Sistema de Gerenciamento de Banco de Dados
BD	Banco de Dados
IA	Inteligência Artificial

Lista de Figuras

FIGURA 2.1 - O processo de “matching” 16
FIGURA 2.2 - Fórmula da Lei de Zipf 18
FIGURA 2.3 - Arquivos invertidos 19
FIGURA 2.4 - Exemplo de árvore TRIE 20
FIGURA 2.5 - Strings semi-infinitas 20
FIGURA 2.6 - Exemplo de assinatura de um bloco 21
FIGURA 2.7 - Fórmula para cálculo de similaridade entre termos de documentos e consultas 22
FIGURA 2.8 - Definição do modelo de indexação “fuzzy” 23
FIGURA 2.9 -Fórmula da freqüência absoluta 27
FIGURA 2.10 - Fórmula para cálculo do peso de um termo em um documento 27
FIGURA 2.11 - Fórmula da freqüência relativa 27
FIGURA 2.12 - Exemplo de “Thesaurus” para um único termo (Acústica) 29
FIGURA 2.13 - Fórmula para cálculo da abrangência ou “recall” 31
FIGURA 2.14 - Fórmula para cálculo da abrangência ou “recall” de agrupamentos 31
FIGURA 2.15 - Fórmula para cálculo da média de abrangência ou “macroaverage recall” 32
FIGURA 2.16 - Fórmula para cálculo da média de abrangência ou “microaverage recall” 32
FIGURA 2.17 - Fórmula para cálculo da precisão ou “precision” 32
FIGURA 2.18 - Fórmula para cálculo da Precisão ou “Precision” para agrupamentos 32
FIGURA 2.19 - Fórmula para cálculo da precisão média ou “macroaverage precision” 33
FIGURA 2.20 - Fórmula para cálculo da precisão média global ou “microaverage precision” 33
FIGURA 2.21 - Fórmula para cálculo do “fallout” 33
FIGURA 2.22 - Processo de KDD 35
FIGURA 2.23 - Um modelo para Mineração de Texto 36
FIGURA 2.24 - Taxonomia para Mineração da Web 40
FIGURA 3.1 - Janela inicial do *software* Eurekha 44
FIGURA 3.2 - Janela para visualização dos agrupamentos obtidos no Eurekha 45
FIGURA 3.3 - Área de trabalho do "UMAP" 47
FIGURA 3.4 - Temas mais abordados pelo jornal Folha de São Paulo no ano de 1999 56
FIGURA 3.5 - Palavras relevantes indicadas pelo Umap com usuário “novice” para a coleção inteira 57
FIGURA 3.6 - Palavras relevantes indicadas pelo Umap com usuário “expert” para a coleção inteira 58
FIGURA 3.7 - Palavras relevantes indicadas pelo Umap com usuário “novice” para a coleção do período de nove (9) meses de 1999 59
FIGURA 3.8 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário “novice”, para a coleção do período de nove (9) meses de 1999 59
FIGURA 3.9 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário “novice”, para a coleção do 1o semestre de 1999 60
FIGURA 3.10 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário “expert”, para a coleção do 1o semestre de 1999 61

FIGURA 3.11 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do 1o trimestre de 1999.....62
FIGURA 3.12 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "expert", para a coleção do 1º semestre.....62
FIGURA 3.13 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do mês de janeiro.....63
FIGURA 3.14 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para coleção de 178 documentos.....65
FIGURA 3.15 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 114 documentos.....65
FIGURA 3.16 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 65 documentos.....66
FIGURA 3.17 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 32 documentos.....66
FIGURA 3.18. Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 9 documentos.....66
FIGURA 3.19 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 178 documentos processados pelo algoritmo "best-star".....68
FIGURA 3.20 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 178 documentos processados pelo algoritmo "cliques".....69
FIGURA 3.21 - Comparação entre os valores de "microaveraging" (a) e "macroaveraging" (b) dos algoritmos "best-star' e "cliques" do Eurekha e, do Umap para a coleção de 178 documentos.....70
FIGURA 3.22 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 114 documentos processados pelo algoritmo "best-star".....70
FIGURA 3.23 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 114 documentos processados pelo algoritmo "cliques".....71
FIGURA 3.24 - Comparação entre os valores de "microaveraging" (a) e "macroaveraging" (b) dos algoritmos "best-star' e "cliques" do Eurekha, e do Umap para a coleção de 114 documentos.....71
FIGURA 3.25 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 65 documentos processados pelo algoritmo "best-star".....72
FIGURA 3.26 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 65 documentos processados pelo algoritmo "cliques".....73
FIGURA 3.27 - Comparação entre os valores de "microaveraging" (a) e "macroaveraging" (b) dos algoritmos "best-star' e "cliques" do Eurekha, e do Umap para a coleção de 65 documentos.....73

FIGURA 3.28 - Gráficos para os valores de “microaveraging” (a) e “macroaveraging” (b) dos 32 documentos processados pelo algoritmo “best-star” 74
FIGURA 3.29 - Gráficos para os valores de “microaveraging” (a) e “macroaveraging” (b) dos 32 documentos processados pelo algoritmo “cliques”.....75
FIGURA 3.30 - Comparação entre os valores de “microaveraging” (a) e “macroaveraging” (b) dos algoritmos “best-star’ e “cliques” do Eurekha, e do Umap para a coleção de 32 documentos.....75
FIGURA 3.31 - Quantidade de documentos processados pelo algoritmo “best-star” (a) e “cliques” (b) do Eurekha e respectivos valores de “microaverage precision” 77
FIGURA 3.32 - Quantidade de documentos processados pelo Umap e respectivos valores de “microaverage precision”, tendo a palavra “Amazônia” como ponto focal.....77
FIGURA 3.33 - Quantidade de documentos processados pelo Umap e respectivos valores de “microaverage precision”, eliminando a palavra “Amazônia” do ponto focal 78
FIGURA 3.34 - O “gato” aponta com a calda para uma recomendação ao usuário no Umap.....80

Lista de Tabelas

TABELA 3.1 - Lista de "stopwords" utilizada em todos os experimentos 50
TABELA 3.2 - Grupos originados a partir de ilhas e consultas do Umap para a coleção inteira 58
TABELA 3.3- Grupos originados a partir de ilhas e consultas do Umap para a coleção inteira para o usuário *expert*..... 59
TABELA 3.4 - Grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes aos primeiros 9 meses de 1999..... 60
TABELA 3.5 - Os grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes ao 1o semestre de 1999..... 61
TABELA 3.6 - Grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes ao 1o trimestre de 1999 62
TABELA 3.7 - Grupos formados pelo Umap a partir da coleção do mês de janeiro..... 63
TABELA 3.8 - Número de documentos e o tempo de processamento de cada período pelo Eurekha 64
TABELA 3.9 - Valores de “microaverage recall” e “microaverage precision” do algoritmo “best-star” para os nove documentos 76
TABELA 3.10 - Valores de “macroaverage recall” e “macroaverage precision” do algoritmo “cliques” para a coleção de nove documentos 76
TABELA 3.11 - Comparação entre as ferramentas Eurekha e Umap..... 79
TABELA 3.12 - Comparação entre a recuperação automática e a recuperação humana81

Resumo

Este trabalho faz avaliação de ferramentas que utilizam técnica de Descoberta de Conhecimento em Texto (agrupamento ou "clustering"). As duas ferramentas são: Eurekha e Umap. O Eurekha é baseado na hipótese de agrupamento, que afirma que documentos similares e relevantes ao mesmo assunto tendem a permanecer em um mesmo grupo. O Umap, por sua vez, é baseado na árvore do conhecimento. A mesma coleção de documentos submetida às ferramentas foi lida por um especialista humano, que agrupou textos similares, a fim de que seus resultados fossem comparados aos das ferramentas. Com isso, pretende-se responder a seguinte questão: a recuperação automática é equivalente à recuperação humana? A coleção de teste é composta por matérias do jornal *Folha de São Paulo*, cujo tema central é a Amazônia. Com os resultados, pretende-se verificar a validade das ferramentas, os conhecimentos obtidos sobre a região e o tratamento que o jornal dá em relação à mesma.

Palavras-chave: Recuperação de Informações, Descoberta de Conhecimento em Texto, Técnicas de Agrupamento, Avaliação de Sistemas de Recuperação de Informação

TITLE: "A COMPARATIVE ESTUDY OF KNOWLEDGE DISCOVERY FROM TEXT SOFTWARES: THE ANALYZE OF AMAZON FOREST"

Abstract

This work presents an evaluation of two softwares capable of performing clustering on textual data: Eurekha and Umap. Eurekha, which performs document clustering, is based on Cluster Hypothesis, which states that closely associated documents tend to be relevant to the same requests. Umap does word clustering and is based on dynamic ideography, in which Pierre Lévy states the existence of a new language that would go beyond the distinction between text and image to provide a dynamic representation of thought models. Tree of knowledge is one of these models, and it is used in Umap. The first does document clustering and the last, word clustering.

The outputs of the softwares are compared to the clustering perfomed by a human specialist. Thus, we want to answer the question: is automatic retrieval equivalent to human retrieval? For this purpose, the evaluation will be based on effectiveness, processing time and interface aspects. The collection consists of several publications about the Amazon forest taken from a newspaper called *Folha de São Paulo*. Thus, we see which knowledge is discovered about that too.

Keywords: Information Retrieval, Knowledge Discovery from Text, Clustering, Information Retrieval Systems Evaluation

1 Introdução

O crescimento da Internet e o grande volume de dados disponível aos usuários tornam necessária a criação de métodos mais rápidos e eficientes para acessar estes dados e extrair-lhes informação. Este é o intuito da área de Recuperação de Informação. Assim, um Sistema de Recuperação de Informação (SRI) deve ser capaz de armazenar, recuperar a informação através de uma linguagem de consulta.

Diferentes dos sistemas de gerenciamento de bancos de dados, os sistemas de recuperação de informação textual lidam com informação não estruturada (textos), armazenada em linguagem natural. Desta forma, apresentam problemas decorrentes da ambigüidade da língua como o problema do vocabulário (relativo a erros semânticos no resultado, decorrentes de sinônimos, palavras com mesmo radical, etc.).

Por outro lado, há também o problema de garantir se uma informação é realmente relevante ao usuário. Gerard Kowalski [KOW 97] define cinco tipos de relevância baseadas no julgamento humano:

a) Subjetiva: resulta de um julgamento específico do usuário;
b) Situacional: as necessidades de informação estão relacionadas com as circunstâncias;
c) Cognitiva: depende da percepção humana e do comportamento;
d) Temporal: muda com o tempo;
e) Mensurável: observável em pontos do tempo.

As ferramentas mais utilizadas para recuperação de documentos textuais são os motores de busca da Internet como “Yahoo!”, “Altavista”, “Infoseek”, “Google” e outros, mas a “interface” de consulta é limitada pela pesquisa sintática por palavras-chave. Além disso, o resultado apresenta os problemas do vocabulário e da sobrecarga de informação (centenas de documentos são retornados em resposta a uma consulta).

Na área de Descoberta de Conhecimento em Bancos de Dados ou Mineração de Dados, estão sendo desenvolvidas técnicas para extrair informação não trivial ou armazenada implicitamente em bancos de dados estruturados. A descoberta de conhecimento é utilizada por empresas para identificar perfis de usuários, padrões e tendências em grandes bancos de dados. Por este motivo, tornou-se interessante formas de aplicar estas técnicas a textos, que são muito mais ricos em informação. A área de Descoberta de Conhecimento em Textos (KDT) ou Mineração de Textos aplica técnicas de mineração de dados em textos e técnicas específicas. É considerada a evolução de Recuperação de Informações, uma vez que minimiza o esforço do usuário na busca por informações relevantes.

Descoberta de Conhecimento em Textos apresenta técnicas para executar tarefas que eram realizadas apenas por seres humanos como:

a) Sumarização: abstração das partes mais importantes do conteúdo de um documento ou conjunto de documentos;
b) Agrupamento: identifica similaridades entre documentos, alocando-os em grupos de acordo com o grau de similaridade;
c) Classificação: identifica a classe ou categoria (assunto) a que pertence determinado documento. Esta classe deve ser previamente definida.

d) Filtragem: espécie de classificação de informações, que pode ser por recomendação ou filtragem colaborativa.

Estas técnicas são extremamente úteis para o ser humano quando o volume de textos a ser analisado é muito grande. Portanto, é muito importante avaliar os seus resultados.

O objetivo deste trabalho é fazer um estudo comparativo de ferramentas, que utilizam técnicas de Descoberta de Conhecimento em Textos: Umap e Eurekha. O Eurekha utiliza técnica de agrupamento ou "clustering" e o Umap executa o processo de lematização, filtra palavras-chave e, em seguida, gera grupos de assuntos.

Este estudo comparativo se baseia na avaliação de *efetividade* das ferramentas através da comparação de seus resultados com os resultados de um especialista humano e avaliação da "interface" ou canal de comunicação oferecido pelas mesmas, que também influencia bastante os resultados. Com isto, pretende-se saber se a recuperação automática é equivalente à recuperação humana.

Para tanto, os capítulos de dois a cinco apresentam o levantamento bibliográfico e os aspectos mais importantes sobre Recuperação de Informações (RI), Recuperação de Informações Textuais e Descoberta de Conhecimento em Textos (KDT) respectivamente. No capítulo seis, são descritas as principais características das ferramentas utilizadas. E, finalmente, no capítulo sete, é feito um estudo de caso, tomando como coleção de teste matérias do jornal *Folha de São Paulo*, publicadas em 1999, que permitem fazer uma análise sobre os problemas e características da Amazônia.

2 Aspectos relevantes da Recuperação de Informações e Descoberta de Conhecimento em Textos

Vannevar Bush, em seu artigo "As we may think" de 1945 [BUS 45], tornou popular a idéia da recuperação de informação. Porém, o termo "Information Retrieval" (Recuperação de Informação) só apareceu em 1952, em um artigo escrito pelo empresário Calvin Moores, que trabalhava na área [WIV 2000]. Em 1962, o termo se tornou realmente popular entre pesquisadores através dos escritos de Fairthorn [SPA 97].

Nesta etapa, são apresentadas as bases teóricas de Recuperação de Informações como as contribuições de H.P. Luhn, Gerard Salton e outros. Cyril Cleverdon (na época, membro do "Cranfield College of Aeronautics"), por exemplo, desenvolveu, na década de 60, as medidas de avaliação "recall" e "precision", que são utilizadas até hoje. Além disso, são estudados os aspectos principais da Descoberta de Conhecimento em Textos, "Knowledge Discovery from Text", termo utilizado pela primeira vez por Ronen Feldman e Haym Hirsh em 1997. O próximo tópico apresenta definições relativas à área de Recuperação de Informações.

2.1 Recuperação de Informações (RI)

Segundo Sparck Jones e Willet [SPA 97], "Recuperação de informação é freqüentemente considerada sinônimo de recuperação de documento e, hoje, de recuperação de texto, implicando que a tarefa do sistema de recuperação de informação é recuperar documentos ou textos com *conteúdo relevante* à necessidade de informação do usuário". Um documento pode conter texto, página da WWW, imagem, som ou vídeo.

A maior dificuldade da área de RI é definir o conteúdo relevante para o usuário, posto que *relevância* é uma medida abstrata do quanto um documento satisfaz a necessidade de informação do usuário.

A área de RI possui um modelo abstrato de representação, que mostra quais são seus principais componentes e processo. Este assunto é abordado na seção a seguir.

2.1.1 Paradigma da Recuperação de Informação

Os componentes básicos do modelo de processo de RI são o usuário, o sistema de recuperação de informação e o documento.

O usuário é o elemento que está em busca de conhecimento para realizar uma tarefa. O sistema de recuperação de informações é a interface entre o usuário e a coleção de documentos a ser consultada. Sua função é receber a consulta e retornar informação relevante. O documento é a unidade básica do sistema a ser consultada. A interação destes componentes é feita através de três processos: abstração de informações, descrição das necessidades do usuário e "casamento" ou "matching", que são abordados a seguir.

Abstração de informações

Um SRI precisa de uma forma de representar um documento para poder manipulá-lo corretamente. Esta representação é um modelo (abstração), que deve ser

especificado com cautela, sob pena de o usuário não conseguir recuperar a informação desejada.

A indexação é a técnica mais utilizada pelos SRI para representar informações de um documento. Um índice funciona como um marcador ("tag") através do qual o conteúdo da informação do documento pode ser identificado. É neste contexto que pode surgir o *ruído semântico*, quando a correspondência entre o conteúdo da informação do documento e seu conjunto de índices não é exata [MAR 59]. Há dificuldade em especificar precisamente o conteúdo do assunto de um documento por meio de uma ou mais palavras (índices).

Assim, dado um termo qualquer, há muitos assuntos possíveis que podem ser denotados por ele. Há também muitos termos que podem denotar um assunto. Da mesma forma, a correspondência entre a formulação de consulta feita pelo usuário e sua necessidade pode não ser exata.

Descrição da necessidade do usuário

Para obter a informação desejada de um sistema de RI automatizado, o usuário precisa descrever sua necessidade através de uma linguagem de consulta. Em decorrência disto, podem surgir problemas como: o usuário pode não saber descrever suas necessidades em termos das especificações formais da consulta ou mesmo não saber quais são elas (por desconhecimento do conteúdo completo dos textos); o próprio formalismo pode não ser adequado à necessidade do usuário (neste caso, o problema é de modelagem do sistema); a descrição fornecida pelo usuário pode não combinar com nenhuma das descrições realizadas pelas pessoas que criaram as representações dos documentos. Este problema é conhecido como problema do vocabulário.

O processo de "casamento" ou "matching"

"Casamento" ou "matching" é o processo de identificação de informações relevantes, que se dá comparando a expressão de consulta do usuário e as representações de cada documento (índice). A função de similaridade é responsável por fazer a comparação entre ambos, como mostra a figura 2.1.

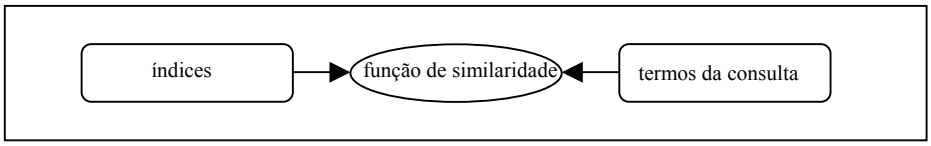


FIGURA 2.1 - O processo de "matching"

2.1.2 Sistemas de Recuperação de Informações (SRI)

Inicialmente, a tarefa de recuperação de informação era realizada apenas por bibliotecários, que eram obrigados a executar pesquisas bibliográficas, usando ferramentas manuais como catálogos e esquemas de classificação universal. Entre os anos 50 e 60, surgiram teorias e modelos computacionais para a área de sistemas de recuperação de informações, que serviram de base para implementação de sistemas como o SMART (cujo modelo é abordado na seção 2.2.2). Mas, foram os motores de busca, utilizados para localizar documentos (página Web) na Internet, que popularizaram estes sistemas.

Um Sistema de Recuperação de Informação (SRI) automático é qualquer sistema, envolvendo computadores, que realiza recuperação de informação sobre documentos em um formato padrão através de uma consulta. Uma consulta é a

expressão formal da necessidade do usuário. Assim, seus objetivos são minimizar o tempo de espera do usuário pela localização da informação e usar noção de relevância para apresentação do resultado. A avaliação de seu desempenho se baseia em métricas denominadas “precision” e “recall”, abordadas mais adiante.

Um SRI se diferencia de um Sistema de Gerenciamento de Banco de Dados (SGBD), porque um SGBD trata dados estruturados, armazenados em tabelas e cujas consultas retornam resultados exatos. Já um SRI trata dados não estruturados, cujas consultas freqüentemente retornam resultados imprecisos. Na maioria dos casos, um SRI recupera somente uma aproximação, encontrando diversas respostas possíveis e elaborando um “ranking”, em que os documentos são listados de acordo com sua estimativa de relevância. O sucesso disso é subjetivo [WIV 2000].

Existem dez requisitos importantes, que devem ser disponibilizados em um SRI. Eles são em ordem decrescente de importância [CRO 95]:

a) Soluções integradas: Oferecer ferramentas para manipular e digitalizar textos, recursos multimídia, gerenciar bancos de dados estruturados e “workflow”.

b) RI Distribuída: Um ambiente distribuído que englobe a Internet.

c) Indexação e recuperação eficientes e flexíveis: Permitir indexar documentos em diferentes formatos e recuperá-los com maior velocidade.

d) “Magic”: Um dos grandes problemas em SRI são incompatibilidades de vocabulários, isto é, geralmente a informação é descrita pelo usuário com palavras diferentes dos índices de documentos relevantes. Este problema requer que as técnicas de expansão de vocabulário, como indexação semântica latente e “thesaurus”, se tornem mais confiáveis.

e) Interfaces e “Browsing”: Melhoramento da interface dos SRI, para apresentarem suporte a um conjunto de funções como formulação de consulta, visualização de informações recuperadas, “feedback” e “browsing” conceitualmente mais simples.

f) Roteamento e filtragem: São sinônimos e referem-se ao problema padrão de RI: recuperar a informação que o usuário necessita. Hoje, existem máquinas de pesquisa genéricas e específicas em relação ao assunto, trazendo o problema do roteamento de consulta ou “query routing problem” [SUG 2000], que se refere à pesquisa de motores de busca apropriados para uma determinada consulta.

g) Recuperação Eficiente: Um sistema, que funciona bem para muitas pesquisas, mas não deixa que o usuário recupere erros ou compreenda por que eles ocorrem, não é eficiente. Esses erros ocasionais têm muito pouco impacto sobre a média de “recall”/“precision” usadas nos testes padrão de RI, mas têm considerável impacto sobre usuários finais.

h) Recuperação Multimídia: Há necessidade de se descobrir técnicas mais eficientes para indexar e recuperar vídeo, som e imagem sem descrições.

i) Extração de Informação: Técnicas de extração de informações são projetadas para identificar entidades, atributos e relacionamentos de bancos de dados textuais. Elas servem, por exemplo, para extrair informações financeiras ou comerciais de textos da Internet. XML (*Extensible Markup Language*) permite que as informações sejam trocadas de forma mais fácil.

j) "Relevance Feedback": É o processo em que usuários identificam documentos relevantes em uma lista inicial de documentos recuperados. O sistema cria uma nova consulta baseada nestes exemplos. Os problemas centrais são selecionar palavras e frases dos documentos relevantes e calcular os pesos destes no contexto de uma nova consulta. Hoje, os usuários especificam apenas um único documento, que nem sempre apresenta os aspectos de documentos realmente relevantes.

2.2 Recuperação de Informações Textuais

Recuperação de Informação envolve duas diferentes, mas relacionadas, atividades: indexação e pesquisa [SPA 97]. A indexação é a forma como os documentos e consultas são representados para serem recuperados posteriormente. A pesquisa é a forma como um arquivo é examinado e seus itens tomados quando relacionados por uma consulta.

Um sistema de recuperação de informações textuais, portanto, é desenvolvido para indexar e recuperar documentos do tipo texto, isto é, cujas informações estão em linguagem natural e cuja unidade básica é a *palavra*. Para tanto, ele deve permitir indicar a relevância de um termo em relação ao documento para tornar possível a indexação.

A definição de relevância dos termos envolve o estudo de como eles estão distribuídos dentro do texto, ou seja, qual a freqüência de ocorrência de um termo no texto e como isso pode ser explorado. Muitas coleções de documentos possuem características estatísticas similares.

Em 1949, surgiu a lei de Zipf: a freqüência de uma dada palavra multiplicada pela sua ordem de classificação ("rank") é aproximadamente igual à freqüência de outra palavra multiplicada por sua ordem de classificação [SAL 83], como mostra a figura 2.2.

$$\text{Freqüência} * \text{rank} \approx \text{constante}$$

FIGURA 2.2 - Fórmula da Lei de Zipf

Esta lei é explicada pelo Princípio do Mínimo Esforço, o qual afirma que é mais fácil para um escritor ou locutor de uma língua repetir certas palavras para expressar suas idéias do que criar novas. O princípio também mostra que palavras menos freqüentes (com menor "rank") têm menos importância no texto.

A seguir, são abordadas as estruturas de dados utilizadas nos Sistemas de Recuperação de Informações.

2.2.1 Estruturas de dados de Recuperação de Informação

Uma estrutura de dados mostra a organização da informação, geralmente, na memória do computador, que serve para melhorar a eficiência do algoritmo. As mais simples são as *pilhas, vetores, listas e árvores*. A área de recuperação de informações, por sua vez, também possui estruturas de dados que são: "Stemming", "n-gram", arquivos invertidos, árvore TRIE, árvore PAT, método da assinatura e hipertexto.

"Stemming"

O algoritmo de normalização morfológica ou "Stemming" identifica o radical das palavras para eliminar suas variações morfológicas. Uma forma de identificar radicais de palavras consiste na definição de uma lista de prefixos e sufixos.

A desvantagem é que o padrão encontrado nem sempre é o prefixo ou sufixo. Por exemplo, o sufixo *ual* deve ser retirado de fat*ual*, mas não de ig*ual* [WIV 2000]. Outra solução é o uso de dicionários, contendo os radicais das palavras, porém nem sempre eles são completos.

Estruturas "N-Gram"

As estruturas "n-gram" se baseiam no cálculo da medida de similaridade do "string" para eliminar variações morfológicas. Elas fragmentam uma palavra em uma seqüência de "n-grams", isto é, "strings" de n caracteres adjacentes, e assim estimam a similaridade entre um par de palavras pela similaridade entre os conjuntos de "n-grams" correspondentes.

As palavras-chave são quebradas em segmentos de tamanho fixo, por exemplo, três caracteres são "trigams", "sea colony" fica "sea + col + olo + lon + ony".

Arquivos invertidos

A estrutura de arquivos invertidos é uma lista ordenada de palavras. Cada palavra possui um ponteiro ("link") para os documentos, em que ela aparece, como na figura 2.3.

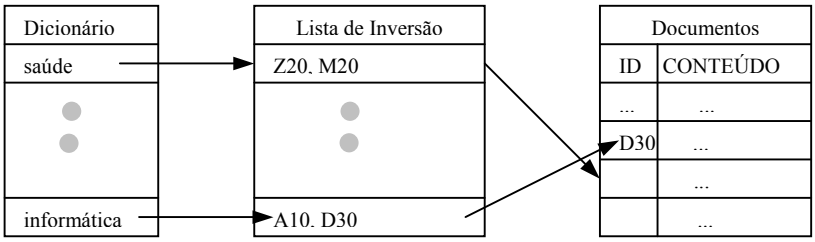


FIGURA 2.3 - Arquivos invertidos

A estrutura é implementada com três arquivos geralmente: dicionário ou lista de palavras indexadas, lista de inversão e documentos. O dicionário contém a entrada do índice e a lista de inversão, o endereço dos documentos em que o índice aparece. Sua desvantagem é consumir bastante espaço, mas é muito eficiente em termos de acesso, por isso, é bastante utilizada.

Árvore TRIE

Uma árvore TRIE[1] é uma estrutura de dados utilizada para fazer busca rápida em um texto extenso. Seus nós são vetores com campos correspondentes aos valores que compõem a chave. O conteúdo de cada componente de um nó pode ser um número ou os caracteres que indicam o conteúdo do nó.

Cada nó, em um nível, representa as chaves com mesma seqüência inicial de valores e com o número equivalente ao do nível. Além disso, cada nó interno pode especificar tantos caminhos quanto o número de componentes na chave. Os nós externos apresentam as chaves completas e não apontam para mais nada.

[1] Derivada do termo "retrieval" (recuperação).

Dependendo do tamanho das chaves, os galhos podem-se tornar muito grandes. Uma grande desvantagem da TRIE é a formação de caminhos de uma só direção para chaves com grande número de bits em comum, como as chaves B e C da figura 2.4.

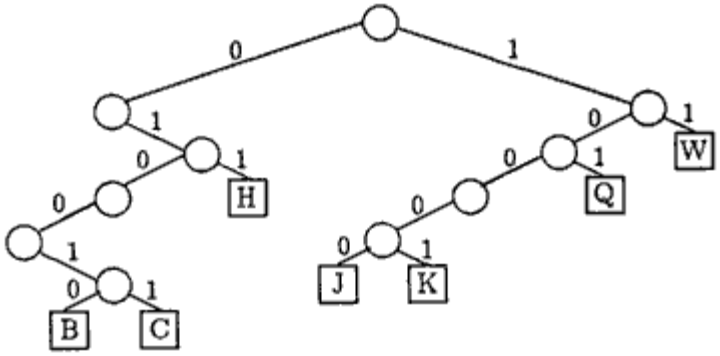


FIGURA 2.4 - Exemplo de árvore TRIE [2]

Árvore PAT

Nas estruturas manipuladas pelos algoritmos PATRICIA ("Pratical Algorithm To Retrieve Information Coded In Alphanumeric", quer dizer, Algoritmo Prático para Recuperação de Informação Codificada em Alfanumérico), o documento é tratado como uma cadeia de caracteres, em que cada posição pode ser um ponto de busca. Cada posição da "string" define uma "substring" que começa em um ponto e vai até o final do documento, incluindo todo o texto intermediário.

As "substrings" são denominadas "semi-infinite strings" ("strings" semi-infinitas) ou simplesmente "sistrings". A figura 2.5 mostra o exemplo obtido a partir do primeiro parágrafo desta subseção:

Sistring1	Nas estruturas manipuladas pelos algoritmos PATRICIA...
Sistring2	as estruturas manipuladas pelos algoritmos PATRICIA...
Sistring3	s estruturas manipuladas pelos algoritmos PATRICIA...
Sistring4	Estruturas manipuladas pelos algoritmos PATRICIA...
	...

FIGURA 2.5 - Strings semi-infinitas

As árvores PAT, entretanto, aparecem em formato digital (0s e 1s).

Método da assinatura

O método da assinatura fornece um teste rápido, que indica os arquivos (provavelmente) mais relevantes à consulta do usuário. O resultado é exibido ou serve de entrada para outro método de filtragem.

Os documentos são divididos em blocos para evitar assinaturas muito grandes e colisões (assinaturas iguais). Quanto maior a assinatura, menor a probabilidade de ocorrerem colisões.

[2] Exemplo extraído de http://www.dcc.pucmg.br/computacao/disciplinas/atp2/semestre1_2001/paginas/051141/estruturas_de_armazenamento.htm

Dentro de um bloco, cada palavra é mapeada para um código de tamanho fixo de bits (a assinatura), estabelecido por uma função "hash", que determina quais posições do código devem receber um valor igual a 1. Os códigos das palavras são combinados em uma função OR para gerar a assinatura do bloco, como no exemplo da figura 2.6.

Computer	0001	0110	0000	0110
Science	1001	0000	1110	0000
Graduate	1000	0101	0100	0010
Students	0000	0111	1000	0100
Study	0000	0110	0110	0100
Assinatura do bloco	1001	0111	1110	0110

FIGURA 2.6 - Exemplo de assinatura de um bloco

Os blocos de cada arquivo são armazenados de modo contíguo no arquivo de assinaturas. Cada nova assinatura é armazenada no final do arquivo. A vantagem é que o arquivo de assinaturas, nesta forma comprimida, ocupa pouco espaço.

Hipertextos

A estrutura de hipertexto é composta por nós e "links" (ligações). Seus arquivos de dados necessitam de um interpretador, já que são escritos em uma linguagem de marcadores (HTML – "HyperText Markup Language"). Para serem carregados em um "browse", como o "Netscape Navigator" ou "Internet Explorer" (ferramentas de visualização), precisam que se informem o protocolo (http://), o servidor (www.ufpa.br/) e a localização (secom/).

2.2.2 Modelos de Recuperação de Informações Textuais

Desde a década de 70, percebeu-se que era vantajoso modelar vários aspectos de recuperação de informação, porque [SPA 97]:

a) qualquer modelo se basearia em um conjunto de hipóteses importantes diferentes da implementação;

b) não só se tornaria possível declarar qual estratégia de recuperação seria mais apropriada, como explicar por que;

c) não haveria mais dificuldade de relacionar um modelo geral a técnicas de implementação específicas, principalmente quando estratégias e dispositivos individuais aparecessem combinados.

Estas vantagens aparecem porque um modelo é apenas uma abstração de um mecanismo, ou seja, está distante dos detalhes de uma implementação. Eles incluem não só o mecanismo de recuperação usado para comparar uma consulta com um conjunto de documentos, mas também os meios pelos quais uma necessidade de informação do usuário pode ser formulada como uma consulta a ser executada por aquele mecanismo.

Nas próximas seções, são descritos os seguintes modelos de recuperação de informações: Booleano, Espaço-vetorial, Probabilítico, Difuso, Busca Direta, Aglomerado e Contextual.

Booleano

O termo *booleano* advém do nome do matemático britânico George Boole, que realizou trabalhos de lógica e álgebra matemática no século XIX.

O modelo booleano é a base para a maioria dos sistemas de banco de dados e dos sistemas de recuperação de informação. Ele se fundamenta na teoria dos conjuntos, representando conceitos e características de um documento em um conjunto finito. Estas representações são binárias, isto é, o documento possui ou não uma determinada propriedade.

As consultas são definidas por expressões booleanas, cujos operandos são os conceitos ou características do conjunto finito e operadores tradicionais da lógica booleana (AND, OR, NOT), que conectam os termos (índices) procurados na consulta. Assim, um SRI expressa cada consulta como uma combinação de índices, que são conectados por operadores lógicos.

O problema do modelo booleano é que o usuário leigo tem dificuldade de expressar sua necessidade através da linguagem de consulta. Além disso, há pouco controle sobre o número de documentos retornados. Sua resposta é uma simples partição do conjunto de documentos em dois subconjuntos: os registros que atendem a consulta e os que não atendem.

As ferramentas que utilizam este modelo são: Excite (www.excite.com), AltaVista (www.altavista.com), Lycos (www.lycos.com.br), Yahoo (www.yahoo.com).

Espaço-Vetorial

O modelo espaço-vetorial ("vector-space model") foi desenvolvido por Gerard Salton [SAL 83a] e implementado por ele mesmo no sistema SMART, enquanto trabalhava na universidade de *Cornell*.

No sistema SMART, cada registro ou documento é representado por um vetor de termos. Um documento (D_i) é identificado por uma coleção de termos T_{i1}, T_{i2}, ... T_{ij}, onde T_{ij} é o elemento de uma matriz ou um vetor que assume o valor do peso (importância) do termo (índice) j assinalado ao documento i. Um termo é positivo se ocorre no documento, e é negativo, caso contrário.

Uma consulta é representada também como um vetor C_{j1}, C_{j2}, ... C_{jt}, onde C_{jt} representa o peso do termo t assinalado a uma consulta j. Assim, ao invés de fazer uma comparação entre todos os termos dos documentos e das consultas, a recuperação do termo é feita a partir da magnitude de uma medida de similaridade entre um vetor de documento e um vetor de consulta. Esta medida de similaridade é dada pelo cosseno do ângulo entre documentos ou entre documentos e consultas, como na figura 2.7:

$$\text{Similaridade} = \text{Cosseno}\,(D_i, C_j) = \frac{\sum_{k=1}^{t} (D_{ik} . C_{jk})}{\sqrt{\sum_{k=1}^{t} (C_{ik})^2 . \sum_{k=1}^{t} (C_{jk})^2}}$$

FIGURA 2.7 - Fórmula para cálculo de similaridade entre termos de documentos e consultas

O documento é recuperado dependendo de um limiar de similaridade ou um número específico de itens a serem recuperados. Se o limiar for 0,5, por exemplo, serão recuperados os itens com ângulo, cujo cosseno é maior ou igual a 0,5. Os documentos recuperados podem ser apresentados em ordem decrescente de relevância.

Apesar de, na época, tentar eliminar limitações do modelo booleano, o modelo espaço-vetorial tinha suas próprias limitações [SPA 97]: necessidade de vários termos de consulta, enquanto o modelo booleano precisava de dois ou três termos na expressão

lógica; dificuldade de especificar relacionamentos entre sinônimos e frases; os termos são sempre considerados ortogonais (isto é, pouco relacionados).

Probabilístico

Apesar do sucesso do modelo anterior, em meados da década de 70, surgiu o modelo probabilístico, cuja característica era considerar que a principal função de um SRI era classificar documentos em ordem decrescente de relevância para o usuário ("Probability Ranking Principle").

Segundo Sparck Jones e Willet [SPA 97], a técnica de *indexação probabilística* foi o primeiro desenvolvimento sólido deste modelo, cujo objetivo era eliminar o problema do ruído semântico, já mencionado, e tentar realmente satisfazer as necessidades do usuário.

Esta técnica permitia que um computador, dada uma requisição de informação, fizesse uma inferência estatística e derivasse um número (chamado *número de relevância*) para cada documento, que é uma medida da probabilidade do documento satisfazer aquela requisição. O resultado da pesquisa é uma lista ordenada daqueles documentos que satisfazem a requisição, ordenados de acordo com sua provável relevância.

Difuso ("Fuzzy")

A idéia básica da teoria do conjunto "fuzzy", utilizada em RI, é que, ao invés de incluir um elemento em um dado conjunto ou excluí-lo, uma função-membro expressa o grau de pertinência de um elemento a um conjunto [SAL 83]. Assim, esta teoria adequa-se à imprecisão característica da avaliação de relevância dos sistemas de recuperação de informação.

Um conjunto "fuzzy" é uma generalização do conjunto Crisp, no entanto, cobre apenas um intervalo [0,1]. O grau, assumindo valores de 0 a 1, indica o quanto um objeto pertence ao conjunto "fuzzy". Se o grau for 0, o objeto não pertence; se for 1, pertence totalmente.

Um conjunto "fuzzy" A, que é subconjunto do universo de discurso, U = $\{u_1, u_2, \ldots u_n\}$, é definido pela função membro μ_A: U $\rightarrow$[0,1] em que $\mu_A(u_i)$ é o grau do membro u_i pertencer a A, apresentando a seguinte notação: A= $\{\mu_A(u)/u \mid w \in W\}$ [CRO 94].

O modelo de recuperação "fuzzy" representa os documentos com uma relação de indexação "fuzzy" binária de acordo com a figura 2.8:

$$F_I = \{\mu_{FI}(d,w)/(d,w) \mid d \in D \text{ e } w \in W\},$$

onde:

d: é o documento;
w: é o índice;
μ_{FI}: é a função membro μ_{FI}: D x W $\rightarrow$ [0, 1], que especifica para o par (d, w), o grau (de relevância ou de importância) do termo w para o documento d.

FIGURA 2.8 - Definição do modelo de indexação "fuzzy"

Há vários esquemas para determinar o grau de relevância dos termos, que não são abordados aqui.

Em um SRI booleano, o resultado da consulta é um conjunto de documentos que são iguais à consulta, já no difuso, eles são parcialmente iguais. A representação de uma

consulta simples tem a seguinte notação: Fq = $\{\mu_{Fq}(w)/w \mid w \in W\}$ [CRO 94], em que a função μ_{Fq} especifica a relevância de um documento para uma consulta.

Uma das vantagens dos SRI difusos é que os documentos podem ser ordenados na ordem dos graus (ou pesos), isto é, em ordem decrescente de relevância. Ele também permite que o usuário especifique um limiar $\alpha \in [0,1]$, para delimitar o número de documentos a serem apresentados no resultado da consulta.

Busca Direta

O modelo de busca direta ou de busca de padrões ("pattern search") utiliza métodos de busca de "strings" (cadeias de caracteres) no documento, para localizar aqueles que são relevantes.

Neste modelo, geralmente não se opera com índices, uma vez que as buscas são realizadas diretamente nos textos dos documentos. Portanto, seu uso é aconselhável para pequenas quantidades de documentos. Por outro lado, quando há o emprego de índices, grandes esforços são aplicados às etapas de normalização e padronização, porém as buscas se tornam mais rápidas.

Aglomerado ("Clustering")

Uma forma de facilitar a manipulação de grandes volumes de dados característicos dos SRI é a classificação. Quando se classificam ou agrupam dados, mais rapidamente eles são recuperados. O modelo de aglomerado ou "clustering" utiliza técnicas de agrupamento de documentos para tentar solucionar os problemas de recuperação de informação.

O principal motivo para utilizar métodos de "clustering" em RI está na *Hipótese de Agrupamento* ("Cluster Hypothesis"), definida por Rijsbergen [RIJ 79]: documentos relacionados tendem a ser relevantes às mesmas consultas. Isto é, documentos relevantes a uma consulta tendem a ser mais parecidos entre si, do que os não relevantes e portanto devem ser agrupados, a fim de acelerar o processamento da mesma.

Há vários métodos para gerar agrupamentos. Para escolher um deles, é preciso levar em consideração que [SAL 83]:

a) a classificação (ou agrupamento) deve ser estável, de forma que a adição de novos itens ou a alteração de itens antigos não altere a classificação;

b) a classificação deve ser bem definida, isto é, conjuntos de dados devem pertencer a uma única classe.

Existem duas abordagens distintas, nas quais se baseiam os algoritmos para gerar agrupamentos [RIJ 79]:

a) baseada em medidas de similaridades entre os objetos a serem agrupados;

b) baseada nas descrições dos objetos.

A segunda é a mais utilizada por sua eficiência, uma vez que estes algoritmos não perdem tempo de processamento com cálculos de similaridade. Eles se caracterizam por não procurar estrutura nos dados, mas impor uma, restringindo o número e o tamanho dos grupos.

Neste caso, o mais importante conceito é o de "cluster representative" (grupo representativo) ou vetor de classificação ou, simplesmente, centróide, em que um item (uma entrada de documento ou consulta) é comparado com o centróide de um grupo existente para ser agrupado. O centróide pode ser qualquer documento do grupo.

Para recuperar uma informação, classificada de forma hierárquica, por exemplo, pode-se usar uma simples estratégia de pesquisa. Define-se uma regra de decisão e outra de parada e começa-se a busca pela raiz da árvore. A busca é expandida para o nó que apresenta o valor máximo da função "matching". Um grupo é recuperado, quando o valor máximo da função "matching" obtido é menor do que o anterior, neste caso, apenas um grupo pode ser alcançado.

A desvantagem do modelo de aglomerados "é identificar os grupos de documentos mais coesos e mantê-los assim durante a utilização do sistema. Todo documento inserido ou modificado deve ser analisado novamente a fim de ser colocado no grupo correto" [WIV 2000].

Contextual

Os modelos anteriores apresentam uma característica comum: o "matching" ("casamento") entre documentos e consultas é restritivo, porque só é realizado se os termos da consulta aparecem no documento. O que leva ao "problema do vocabulário" (sinônimos).

O modelo contextual, por sua vez, considera que todo documento possui um assunto específico ou um contexto. Da mesma forma, a consulta do usuário também possui um contexto, que define sua necessidade de informação. Assim, o processo de "matching" é feito em nível destes dois contextos.

Determinar o conceito de um documento ou uma consulta não é uma atividade fácil. Atualmente, um contexto é determinado por um conjunto de palavras, sendo que um peso indica a relevância de uma palavra para um contexto.

Os contextos definidos são utilizados na indexação dos documentos. Da mesma forma, as palavras existentes na consulta do usuário ativam um determinado conceito.

Na realidade, o problema do vocabulário só será eliminado se o conjunto de palavras que define o contexto for bem definido. De outro modo, se os índices não forem bem definidos em termos de contextos corretos, acarretará falhas no sistema.

2.2.3 Indexação

Para se construir um SRI, a primeira operação a ser realizada é a Catalogação, quando são adicionados ao sistema (em dispositivos eletrônicos ou não) as características dos documentos (como resumo, título, autor) ou o próprio texto. Em seguida, vem a etapa de indexação.

Indexação é o processo de análise para extrair um conjunto de palavras do texto capaz de identificar o seu conteúdo, para ser recuperado por uma consulta baseada nestas mesmas palavras. Em um SRI, elas são denominadas *índices*, *descritores* ou *palavras-chave*.

Índices podem ser compostos a partir de vocabulários controlados ou não. Um vocabulário não controlado é menos indicado, porque pode levar a erros e ambigüidades característicos das linguagens naturais.

Os termos de um índice podem ser termos individuais ou de contexto. Quando são termos individuais, a unidade básica para compor o índice é a palavra; quando de contexto, a unidade básica pode ser uma palavra composta ou uma sentença. Em uma consulta, os termos individuais são combinados ou *coordenados* para recuperar documentos; este processo é denominado *pós-coordenação*. Uma página da Web pode ser indexada pelo termo: *educação*. Quando os termos combinados são de contexto, o processo é denominado *pré-coordenação*, como exemplo, o termo *informática educativa*.

A *exaustividade* mede o grau com que todos os conceitos e noções incluídos em um documento são descritos pelos índices. Quanto mais exaustiva for a indexação, maior pode ser a proporção de itens relevantes recuperados, já que todos os aspectos importantes do conteúdo foram considerados [SAL 83]. Já a *especificidade* se refere ao nível genérico dos índices usados para caracterizar o conteúdo de um documento. Se o vocabulário for muito específico, uma grande proporção de termos não relevantes podem ser rejeitados em uma consulta [SAL 83].

A seguir, são mostrados os passos da indexação automática.

Indexação Automática

Inicialmente, a indexação era feita manualmente, mas depois se tornou automática, principalmente com a inclusão de métodos estatísticos no processo. Luhn foi o precursor de muitas das idéias originais sobre análise de texto automática [RIJ 79]. Ele propôs que a freqüência de ocorrência de uma palavra em um artigo fornece uma medida útil da importância da palavra e que a posição relativa dentro de uma sentença tendo dados valores de importância (*pesos*) fornece uma medida útil para determinar a importância das sentenças [LUH, 58][3] apud [RIJ, 79]. A vantagem de se utilizar métodos estatísticos associados à análise de texto automática é que os índices passam a ser descritos pelas palavras do próprio documento [SPA 97].

A indexação automática se constitui das seguintes etapas [WIV 2000]:

a) identificação de palavras;

b) identificação de termos compostos;

c) remoção de "Stopwords";

d) normalização morfológica/Lematização ("Stemming");

e) cálculo de relevância;

f) seleção de termos.

Na etapa de i*dentificação de palavras*, as palavras de um documento são identificadas, eliminando-se caracteres de controle de arquivo ou de formatação. Um dicionário pode ser utilizado para validar as palavras e corrigir erros de ortografia. O resultado desta etapa é um documento normalizado.

Termos compostos são aqueles que possuem um significado diferente quando utilizados em conjunto, por exemplo, *sistema hidráulico.* Assim, na etapa de *identificação de termos compostos,* pode-se utilizar uma de duas formas existentes para identificá-los. Na primeira, o sistema apresenta os termos que sempre ocorrem juntos e o usuário checa os mesmos, aceita ou não. Na segunda, é utilizado um dicionário de termos compostos, que faz a checagem. Neste caso, para evitar problemas, os termos compostos e simples devem estar armazenados no dicionário.

"Stopwords" ou palavras negativas são aquelas que ocorrem com grande freqüência no texto, porém não servem para identificar o conteúdo do mesmo, por isso, a necessidade da etapa de *remoção de "stopwords"*. São as preposições, as conjunções, os pronomes. Uma lista de "stopwords" é chamada de "stoplist" ou dicionário negativo.

Na *normalização morfológica/Lematização ("Stemming"),* sufixos e prefixos são retirados das palavras para evitar variações morfológicas.

---

[3] LUHN, H.P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, [s.l.], v. 2, p. 159-165, 1958.

A *relevância ou peso de um termo* pode ser calculado com base na sua freqüência no texto. As fórmulas de cálculo mais comuns são:

a) *Freqüência absoluta de um termo* ($F_{abs}t$) (figura 2.9): mede quantas vezes um termo aparece em um documento. É a mais simples, porém apresenta as seguintes desvantagens: não faz distinção entre termos que aparecem em poucos e em muitos documentos; o que é importante, já que os que aparecem em muitos são péssimos índices (identificadores); a uma palavra pouco freqüente, em um documento pequeno, pode ser dada a mesma importância de uma palavra muito freqüente em um documento grande.

$$F_{abs}t = n_D,$$

onde:

n_D: é o número de vezes que o termo aparece no documento.

FIGURA 2.9 -Fórmula da freqüência absoluta

b) *Freqüência inversa de documentos* (figura 2.10): visa resolver o primeiro problema da freqüência absoluta, calculando o peso dos termos nos documentos:

$$Peso_{td} = \frac{F_{td}}{DocFreq_t}$$

onde:

$DocFreq_t$: número de documentos em que o termo t aparece,
F_{td}: ocorrência do termo t no documento d,
$Peso_{td}$: peso do termo t no documento d.

FIGURA 2.10 - Fórmula para cálculo do peso de um termo em um documento

c) *Freqüência relativa* (F_{rel}) (figura 2.11): visa diminuir o último problema da freqüência absoluta, já que normaliza os pesos a partir do tamanho do documento.

$$F_{rel} = \frac{F_{abs}}{N}$$

onde:

N: é o número de palavras no documento,
F_{abs}: freqüência absoluta do termo no documento.

FIGURA 2.11 - Fórmula da freqüência relativa

Após o processo de identificação dos índices e determinação de pesos, é necessário selecionar aqueles índices mais importantes para que o arquivo de índices não ocupe muito espaço nem se torne tão grande a ponto de comprometer o desempenho do sistema. É a etapa de *seleção de termos*.

Esta seleção pode ser baseada no peso dos termos ou em sua posição sintática. Quando baseada no peso do termo, a aplicação ou o usuário determina um limiar para a eliminação do termo. Após esta filtragem, pode ser feita a *truncagem*, técnica para estabelecer um número máximo de índices, selecionando aqueles com maior grau.

Indexação Sintática

A indexação automática e o processamento de linguagem natural são freqüentemente usados para extrair palavras significativas ou frases dos textos automaticamente. Com o processamento de linguagem natural, podem ser feitas análises sintática e semântica, que permitem fazer a indexação sintática e semântica de um documento.

A indexação sintática faz a análise sintática dos documentos para descobrir as palavras mais importantes de uma oração. As posições dos termos sintáticos (sujeito, predicado, por exemplo) devem ser pré-definidas. Assim, só os termos importantes são inseridos no índice. É necessária uma *Base de Conhecimento* para que todas as combinações sintáticas possíveis sejam testadas.

Indexação Semântica

O problema das técnicas comuns de indexação automática é que os usuários formulam suas pesquisas com base em conceitos pessoais, enquanto os sistemas são indexados por palavras-chave que não são eficientes para expressar conceitos. Além disso, muitas palavras têm múltiplos significados (polissemia), então os termos de uma consulta podem ser igualados a índices de documentos que não têm o mesmo significado.

Várias técnicas têm sido usadas para análise semântica de textos, como aprendizado de máquina, análises estatísticas (indexação semântica latente, etc.), computação baseada em redes neurais e outros. Os resultados do processo de análise semântica de um texto podem ser representados na forma de redes semânticas, regras de decisão e outros.

"A indexação semântica baseia-se no princípio de que o documento já possui estruturas que indicam a semântica dos termos" [WIV 97]. Palavras-chave, títulos, por exemplo, são marcados por "tags" no caso de ser um documento HTML ou XML. Assim, o processo de indexação deve identificar estas marcações e indexar os termos mais importantes entre estas marcações. Se a marcação for feita de forma errada, ocorre o problema da *indexação incerta*.

2.2.4 Ferramentas de auxílio

Durante a etapa de formulação da consulta, pode surgir o problema da busca incerta ("search incertainly"), quando o usuário não sabe qual a melhor palavra para descrever o assunto que ele quer localizar em um sistema de recuperação de informação. A técnica mais comum para solucionar este problema é a *expansão semântica*, que adiciona à consulta termos correlacionados aos termos informados pelo usuário.

As ferramentas de auxílio, retro-alimentação por relevância ("relevance feedback"), "Thesaurus", hiperdicionário e dicionário, são abordadas a seguir.

Retro-alimentação por Relevância ("Relevance Feedback")

"Relevance feedback" é um processo em que o usuário indica quais documentos, daqueles já retornados em resposta a uma consulta, são mais relevantes. O sistema geralmente tenta encontrar termos comuns àqueles já informados pelo usuário, e adiciona-os à consulta anterior. Assim, novos documentos são retornados a partir da consulta reformulada. Este processo, introduzido na metade da década de 60, também é conhecido como "query by example" (consulta por exemplo).

Existem duas formas de "relevance feedback". A primeira é denominada retro-alimentação positiva, em que o sistema adiciona à consulta termos que aparecem nos

documentos selecionados. A segunda é retro-alimentação negativa, em que os termos que não aparecem no documento são excluídos da consulta ou seu peso (importância) é diminuído.

"Thesaurus"

"Thesaurus" é uma estrutura hierárquica de palavras, similar a um dicionário, mas, ao invés de informar o significado de uma palavra, informa o relacionamento semântico entre palavras como mostra a figura 2.12 para o vocábulo *Acústica*. Estes relacionamentos podem ser de equivalência (sinônimos), hierárquico (por exemplo, boi *é um tipo de* mamífero; dedo *é parte da* mão; Halley *é um exemplo de* cometa) e associativo (simétrico: ouro *está relacionado com* dinheiro e dinheiro *está relacionado com* com ouro; assimétrico: controle populacional *está relacionado com* planejamento familiar, mas não o contrário).

Acústica	
SN	Ciência do som – inclui o estudo da transmissão do som através de vários meios e ambientes fechados
U	Som
	Transmissão de som
	Ondas de som
BT	Ciência
	...

SN: nota; U = usado como; BT = no sentido mais amplo

FIGURA 2.12 - Exemplo de "Thesaurus" para um único termo (Acústica)

A maior dificuldade é manter o "Thesaurus" atualizado, principalmente quando ele contém uma grande quantidade de palavras armazenadas, porque a operação exige uma análise complexa das informações existentes.

Hiperdicionário

É uma estrutura que armazena relacionamentos entre palavras de um mesmo contexto. É bastante empregado juntamente com o modelo de RI contextual. Sua estrutura é composta por nós, que representam palavras. Estes nós estão ligados por elos, que representam o relacionamento entre palavras. Eles também possuem informações como o tipo e o grau do relacionamento.

Dicionários

Dicionários são geralmente utilizados com técnicas de linguagem natural, as quais necessitam de informações sobre morfologia das palavras, sintaxe e sinônimos. Durante a formulação de consulta, o usuário pode dispor deles para selecionar termos adequados ao sistema e verificar sinônimos de palavras simples ou compostas.

2.2.5 Medidas de avaliação de Recuperação de Informações

Os seguintes componentes são necessários em um teste de sistema [SAL 83]:

a) uma descrição do sistema e seus componentes ou um modelo do sistema a ser examinado;

b) um conjunto de hipóteses a ser testado, ou um protótipo particular, em relação ao qual o modelo deve ser medido;

c) um conjunto de critérios refletindo os objetivos de desempenho do sistema e medidas permitindo uma quantificação dos critérios de desempenho.

d) métodos para obter e avaliar dados.

A utilização comercial dos SRI tornou a avaliação um tópico de interesse para pesquisadores. Até 1993, os testes eram feitos em ambientes acadêmicos com poucos dados [KOW 97]. O processo padrão de avaliação mudou com a criação da TREC ("Annual Text Retrieval Evaluation Conference"), que fornece um padrão de banco de dados de teste com "gigabytes" de informação, consultas e resultados esperados.

Avaliar um SRI é muito difícil, porque quando uma consulta retorna documentos irrelevantes, não se pode determinar que isto se deva à existência ou não de documentos sobre o tópico de pesquisa do usuário ("coverage collection"), à classificação de documentos relevantes ou à má formulação da consulta pelo usuário [SPA 97]. Além disso, também é difícil saber quais são os critérios utilizados pelo usuário para julgar um documento relevante.

Sistemas de Recuperação de Informações são avaliados através dos critérios de *efetividade* e *eficiência*. Efetividade é a habilidade de um sistema de informação oferecer os serviços que o usuário necessita. Eficiência é uma medida do custo ou do tempo necessário para realizar um dado conjunto de tarefas.

A efetividade de um sistema é avaliada a partir das medidas como "recall", "precision", "fallout". Por outro lado, a avaliação de eficiência envolve a questão de custo-benefício. A análise de custo-benefício requer uma comparação sistemática entre custos de operações individuais e os seus benefícios.

Ao final da avaliação de um SRI, os resultados devem ser apresentados da seguinte forma [TAG 81]:

a) Objetivo do teste: Mostrar por que fazer o teste e o que se quer descobrir a partir do mesmo.

b) Visão geral do teste: Apresentar referência de trabalhos anteriores relatando especificamente o teste.

c) Metodologia: Descrever o ambiente e procedimentos detalhadamente de forma que possa ser repetido por outro investigador. Além disso, mostrar problemas decorrentes da metodologia no teste atual.

d) Apresentação de resultados: Deve ser de forma verbal, tabular, gráfica, claramente identificada e rotulada.

e) Conclusões: A conclusão é um resumo do que foi feito, a explicação da principal contribuição do trabalho e suas implicações para pesquisas futuras.

As medidas de avaliação "Recall", "Precision", "Macroaveraging" e "Microaveraging" são descritas a seguir.

Abrangência ("Recall")

"Recall" (abrangência ou revocação) mede a quantidade de itens relevantes, dentre os existentes na base de dados, que foram recuperados [WIV 2000] como mostra a figura 2.13.

$$Recall = \frac{número\_de\_documentos\_relevantes\_recuperados}{número\_de\_possíveis\_documentos\_relevantes\_na\_base\_de\_dados}$$

FIGURA 2.13 - Fórmula para cálculo da abrangência ou "recall"

Se o número de documentos relevantes recuperados é igual a 5 e o número de documentos relevantes existentes na base de dados é 6, então o valor de "recall" é igual a 0,83 (5/6). Se o número de documentos recuperados é 9 e o número de documentos relevantes existentes é igual a 7, então o valor de "recall" é 1,2 (9/7). Logo, conclui-se que quanto maior o "recall", maior a probabilidade de serem recuperados documentos irrelevantes, já que o denominador é o número de documentos potencialmente relevantes estimados armazenados. Por isso, no caso de coleções de milhões de documentos, é quase impossível medir "recall".

Uma boa medida de "recall" pode ser obtida com uma indexação exaustiva, porque maior é a proporção de itens relevantes, recuperados, já que os índices foram bem definidos. Para melhorar o "recall", a linguagem de indexação deve fornecer reconhecimento de sinônimo e de relações entre termos. Um valor de "recall" alto recupera itens além dos que podem ser interessantes para o usuário.

Neste trabalho, são importantes as medidas de avaliação para agrupamentos de documentos realizados por *softwares*. Assim, quanto à efetividade do agrupamento, existem dois tipos de medidas de avaliação de SRI: as externas e as internas [STE 2000]. As medidas internas são aquelas que não precisam de uma coleção classificada ou agrupada para comparação, por exemplo, a medida "overall similarity". As medidas externas, ao contrário, necessitam de um parâmetro. É o caso de "microaveraging" e "macroaveraging" [LEW 91] e da entropia [FUJ 2001]. "Microaveraging" é uma medida para avaliar o agrupamento feito em relação à coleção inteira, enquanto "macroaveraging" calcula, primeiro, o "recall" (abrangência) e "precision" (precisão) de cada grupo para depois obter a média. Estas são as mais importantes. Já a entropia estima a distribuição de documentos relevantes, então, quanto menor o seu valor, melhor o agrupamento proposto.

Para avaliar os resultados de agrupamentos, a fórmula da figura 2.13 é alterada como mostra a figura 2.14.

$$Recall = \frac{número\_de\_documentos\_agrupados}{número\_de\_documentos\_que\_deveriam\_ter\_sido\_agrupados}$$

FIGURA 2.14 - Fórmula para cálculo da abrangência ou "recall" de agrupamentos

A fórmula da figura 2.14 calcula a abrangência (“Recall”) para um único grupo. Para saber a média de abrangência ou “macroaverage recall” do agrupamento realizado, deve ser utilizada a fórmula 2.15.

$$Macroaverage\ Recall = \frac{\sum_{n}^{1} \mathrm{Re}call}{n}$$, onde n é o número de grupos

FIGURA 2.15 - Fórmula para cálculo da média de abrangência ou “macroaverage recall”

Para obter uma média global de abrangência do agrupamento realizado, utiliza-se a medida de avaliação denominada “microaverage recall” da figura 2.16.

$$Microaverage\ Recall = \frac{número\_total\_de\_documentos\_agrupados}{número\_total\_de\_documentos\_que\_deveriam\_ter\_sido\_agrupados}$$

FIGURA 2.16 - Fórmula para cálculo da média de abrangência ou “microaverage recall”

Precisão (“Precision”)

“Precision” (precisão) é a habilidade de um SRI recuperar apenas itens relevantes. Esta medida indica o esforço que seria desperdiçado pelo usuário ao fazer uma busca, como mostra a figura 2.17.

$$Precision = \frac{número\_de\_documentos\_relevantes\_recuperados}{número\_total\_de\_documentos\_recuperados}$$

FIGURA 2.17 - Fórmula para cálculo da precisão ou “precision”

Ou seja, se em um conjunto de 100 documentos recuperados aparecem 70 documentos relevantes. O usuário tem de desperdiçar esforço com outros 30% irrelevantes. Assim, quanto maior a precisão, menor é o esforço do usuário na busca por documentos relevantes, porque é menor a taxa de erro.

Uma boa medida de “precision” requer maior especificidade na criação da linguagem de indexação. Com um vocabulário mais específico, maior a proporção de itens relevantes recuperados, portanto, maior a precisão. Por outro lado, uma alta precisão pode deixar de fora documentos possivelmente relevantes ao usuário.

Para avaliar a precisão dos agrupamentos realizados por *softwares*, a fórmula da da figura 2.17 é alterada como mostra a figura 2.18.

$$Precision = \frac{número\ de\ documentos\ corretamente\ agrupados}{número\ de\ documentos\ agrupados}$$

FIGURA 2.18 - Fórmula para cálculo da Precisão ou “Precision” para agrupamentos

A fórmula da figura 2.18 calcula a precisão de um único grupo. Para saber a precisão média ou "macroaverage precision" do agrupamento realizado, deve ser utilizada a fórmula 2.19.

$$Macroaverage\ Precision = \frac{\sum_{n}^{1} \Pr ecision}{n}$$, onde n é o número de grupos

FIGURA 2.19 - Fórmula para cálculo da precisão média ou "macroaverage precision"

Para obter uma média global da correção do agrupamento realizado, utiliza-se a medida de avaliação denominada "microaverage recall" da figura 2.20.

$$Microaverage\ Precision = \frac{número\ total\ de\ documentos\ corretamente\ agrupados}{número\ total\ de\ documentos\ agrupados}$$

FIGURA 2.20 - Fórmula para cálculo da precisão média global ou "microaverage precision"

"Fallout"

Esta medida permite que se verifique se a quantidade de documentos relevantes continua a mesma, quando o número total de documentos é alterado. A fórmula é mostrada na figura 2.21. Um sistema efetivo em RI apresenta o máximo "recall" e mínimo "precision".

$$Fallout = \frac{número\_de\_documentos\_não\_relevantes\_recuperados}{número\_total\_de\_documentos\_não\_relevantes\_existentes}$$

FIGURA 2.21 - Fórmula para cálculo do "fallout"

"Effort"

"Effort" é a medida do esforço do usuário durante a interação com o sistema. Isto envolve a preparação da consulta, o processo de análise dos resultados e a reformulação da consulta.

Medidas Subjetivas

Um problema de determinar "recall" e "precision" é a interpretação de relevância. Relevância, sob um ponto de vista subjetivo, leva em consideração o estado de conhecimento do usuário durante a pesquisa e os documentos que foram recuperados por ele anteriormente. Sob um ponto de vista objetivo, preocupa-se apenas com o que é requerido diretamente na formulação de consulta do usuário.

As medidas "precision" e "recall" medem a quantidade de informação útil obtida por um SRI sob o ponto de vista objetivo. Também não é feita nenhuma avaliação quanto à existência de documentos relevantes ou contraditórios.

Existem quatro tipos de relevância subjetiva [LOH 99]:

a) Tópica: (relativa "aboutness") avalia a correspondência entre tópicos julgada por uma pessoa;

b) Pertinente: relativa à necessidade de informação como percebida pelo usuário;

c) Situacional: relativa à utilidade para uma tarefa;

d) Motivacional: relativa à intencionalidade do usuário.

O critério mais utilizado para avaliar a relevância dos documentos é o julgamento de especialistas humanos.

2.3 Descoberta de Conhecimento em Textos

Em 1989, foi criado o termo "Knowledge Discovery in Database" (KDD) ou Descoberta de Conhecimento em Banco de Dados. O objetivo desta área de pesquisa, também conhecida como Mineração de Dados ("Data Mining"), é a extração não trivial de informação implícita, previamente desconhecida, e potencialmente útil de um dado [FRA 91][4] apud [FEL et. al. 98]. Técnicas de aprendizado de máquina e análise estatística são aplicadas para a descoberta automática de padrões em BD dentro de ambientes para exploração e navegação.

O processo de KDD, mostrado na figura 2.22, inicia com a seleção de um conjunto, um subconjunto ou uma amostra de dados, que dependem do domínio da aplicação e dos objetivos do processo. Estes dados podem ou não passar por um pré-processamento para eliminar ruídos e dados incompletos. Na etapa de transformação, são encontradas características úteis para representar dados.

A etapa mais importante é a de mineração de dados, cujos objetivos podem ser:

a) descobrir a dependência entre dados;
b) descrever conceitos para construir regras;
c) detectar elementos que são exceções às regras;
d) identificar similaridades entre os elementos para classificação e
e) descrever fórmulas.

O resultado da mineração deve passar por um processo de interpretação de especialistas humanos para validar ou não os padrões gerados.

[4] FRAWLEY, W.J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C.J. Knowledge Discovery in Databases: an Overview. In: PIATETSKY-SHAPIRO, G.; FRAWLEY, W.J. (Ed.). **Knowledge Discovery in Databases**, Massachusets,: MIT Press, 1991.

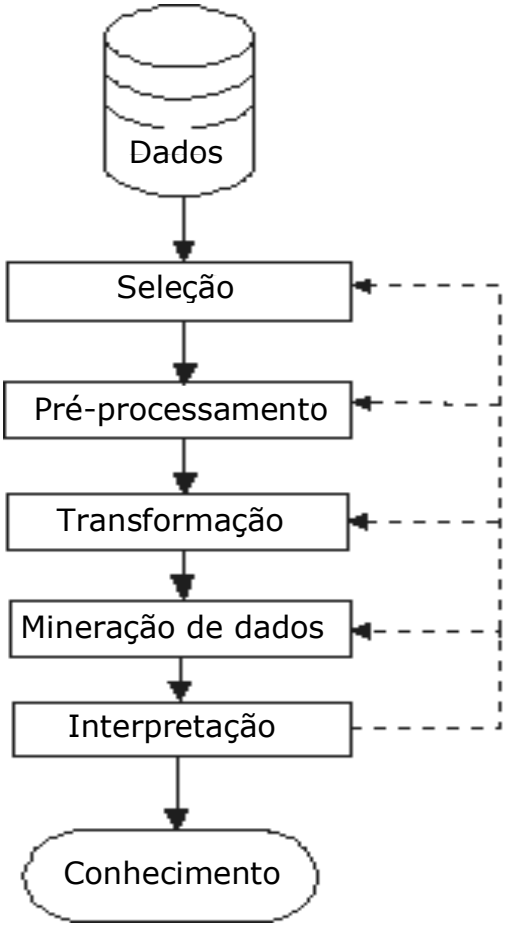


FIGURA 2.22 - Processo de KDD[5]

Os resultados financeiros obtidos com este processo tornaram a mineração de dados uma tecnologia central para as empresas globalizadas. Entretanto, percebeu-se a necessidade de estender o uso destas técnicas a dados não estruturados, visto que mais de 80% das informações de uma empresa estão em bancos de dados textuais [TAN 99]; o que se explica principalmente pelo advento da Internet. Assim, surgiu a área de Descoberta de Conhecimento em Textos ("Knowledge Discovery from Text" ou KDT).

A seguir, são apresentadas a definição, as etapas do processo, os tipos e alguns dos sistemas existentes (e, finalmente, uma discussão sobre medidas de avaliação dos resultados) de Descoberta de Conhecimento em Textos.

2.3.1 Definição

O termo Descoberta de Conhecimento em Texto ("Knowledge Discovery from Text" ou KDT) foi utilizado pela primeira vez por Ronen Feldman [FEL 97]. No entanto, há alguns anos vem se tentando resolver problemas da área e termos afins surgiram como *Busca de Informação* ou "Information Seeking" (descrição de processos automatizados de busca de informação pelo usuário) e *Recuperação de Conhecimento* ou "Knowledge Retrieval" (processo que requer poder de inferência).

[5] Extraído de [BRU 98].

Descoberta de Conhecimento em Textos, também conhecida como mineração de texto (“text mining”), pode ser entendida como a aplicação de técnicas de KDD sobre dados não estruturados (textos). O problema da área é encontrar a informação de interesse do usuário, minimizando o tempo de pesquisa gasto com informação irrelevante [FEL 99]. Por isso, diz-se que o KDT é conseqüência da evolução da área de Recuperação de Informações. KDT, entretanto, não visa apenas extrair a informação de um texto ou coleção de textos, mas também detectar fenômenos, padrões e tendências.

KDT é útil na área judicial, de publicidade, de inteligência competitiva[6] e outras. Serve para desvendar a informação “escondida” em patentes, contratos, títulos de seguros e outros repositórios de textos. Uma empresa que vende produtos na Internet e disponibiliza um e-mail para consumidores através de KDT pode extrair informações preciosas das mensagens, que de outro modo, provavelmente, ficariam sem serem lidas.

Tan [TAN 99] apresenta um modelo para o processo de mineração de texto ilustrado na figura 2.23, consistindo de duas fases:

a) Refinamento do texto: converte documentos em formato de texto em uma forma intermediária escolhida.

b) Refinamento de conhecimento: deduz padrões ou conhecimento da forma intermediária.

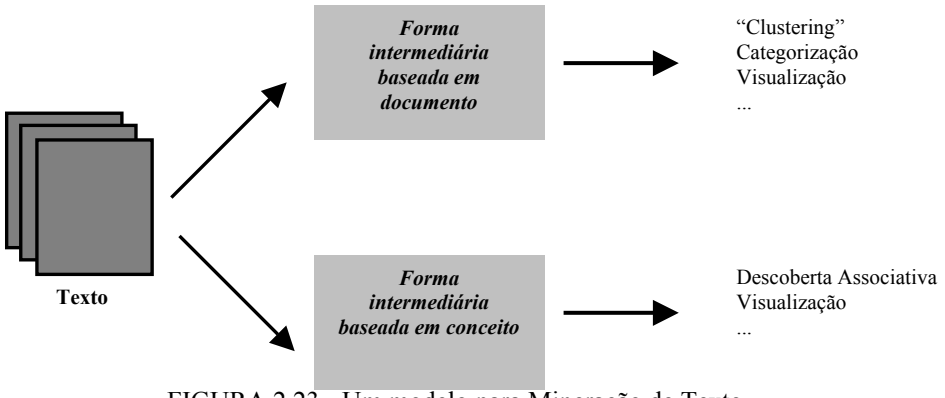


FIGURA 2.23 - Um modelo para Mineração de Texto

A forma intermediária pode ser baseada em documento (cada entidade representa um documento) ou em conceito (cada entidade representa um objeto ou conceito em um domínio específico). A primeira é independente do domínio, a segunda não.

Dado um conjunto de notícias de jornal, a fase de refinamento de texto converte cada documento em uma forma intermediária baseada em documento. A fase de refinamento de conhecimento pode ser aplicada com o propósito de agrupar os artigos por assunto (“clustering”) para visualização e navegação. Por outro lado, caso se desejasse extrair informação dos artigos relacionada a uma determinada empresa, da forma intermediária de documento se passaria à forma intermediária de conceito para construir um banco de dados ao qual seria aplicada Descoberta Associativa, por exemplo, para visualização.

[6] Aplicação de métodos de vigilância ao ambiente externo de uma empresa, buscando monitorar concorrentes, tecnologias e produtos [WIV 2000].

2.3.2 Etapas do Processo

As etapas do processo de KDT são similares às etapas do processo de KDD, porém são genéricas e podem deixar de ser seguidas pelo projetista. Basicamente, são as seguintes [WIV 2000]:

a) Definição de objetivos: definição do que pode ou deve ser descoberto.

b) Seleção de um subconjunto de dados: não se deve utilizar uma grande quantidade de dados, porque pode influenciar no resultado de forma negativa ou tornar o processo mais demorado.

c) Pré-processamento ou limpeza de dados: prepara os dados, eliminando ruídos, por exemplo, caracteres indesejados, correção ortográfica, análise semântica e normalização de vocabulário.

d) Redução ou projeção dos dados: escolha de termos ou partes importantes de texto para análise, a fim de que o processamento seja mais eficiente.

e) Escolha da técnica, método ou tarefa de mineração.

f) Mineração: aplicação do método escolhido.

g) Interpretação dos resultados.

h) Consolidação do conhecimento descoberto e aplicação prática do mesmo.

Qualquer processo de descoberta de conhecimento é dito cíclico, porque após a etapa de interpretação, pode-se voltar às etapas anteriores caso os resultados não sejam satisfatórios. É que, na etapa de definição de objetivos, são formuladas hipóteses que podem ou não ser confirmadas. Esta estratégia só pode ser aplicada quando o usuário é capaz de formular hipóteses iniciais, ou seja, quando ele sabe o que quer ou pode extrair dos dados. Neste caso, o modo de aquisição de informação é reativo.

No modo reativo, a informação é adquirida para resolver um problema específico do usuário, que sabe o que quer e pode identificar a solução do problema quando a encontra [LOH 2000].

O outro modo de aquisição de informação é o modo proativo. Neste, o usuário não tem um objetivo específico; o processo deve encontrar problemas potenciais ou oportunidades, já que o usuário não formula hipóteses iniciais.

2.3.3 Tipos de Descoberta de Conhecimento em Textos

Os tipos de KDT, analisados nesta seção, utilizam uma das duas formas de aprendizado: supervisionado ou não supervisionado. O aprendizado supervisionado apresenta exemplos e resultados esperados ao algoritmo de aprendizado. Já no aprendizado não supervisionado, não existem modelos ou exemplos a serem aprendidos. O sucesso do primeiro depende da boa formulação dos exemplos a serem aprendidos, porém "apresenta resultados melhores e mais refinados" [WIV 2000].

Os métodos ou tipos de KDT abordados nesta seção são os seguintes: extração de informações, sumarização, "clustering", classificação, categorização e filtragem.

Descoberta por Extração de Informações

As técnicas de Extração de Informações (EI) podem ser enquadradas tanto na área de RI, quanto na área de KDT, uma vez que sua aplicação pode retornar informações que não são facilmente identificadas pelo usuário.

Muitos componentes de EI são usados para indexação, por isso, ambos processos são confundidos. Entretanto, a indexação busca identificar palavras capazes de caracterizar o documento e colocá-las em um índice; enquanto a EI transforma a informação extraída para o formato do banco de dados alvo.

O objetivo da EI é transformar dados semi-estruturados ou não estruturados em dados estruturados (geralmente registros). Ela extrai tipos específicos de informações (marcadas por "tags" sintáticas ou semânticas) contidas nos textos, havendo a possibilidade de construir regras de extração genéricas ou específicas do domínio (como datas). Depois, estas informações podem ser usadas como simples dados de entradas para processos como o de mineração de texto.

O outro método de Extração de Informações é por Passagens se caracteriza por ser menos dependente do domínio do que a extração tradicional. Assim, o usuário pode utilizar regras gerais ou ele mesmo definir suas regras de busca. Por exemplo, para encontrar o objetivo de um artigo, pode-se procurar no texto palavras como *objetivo*, *finalidade* (sinônimos) ou passagens como "pretende-se", "quer-se", "será discutido". Os resultados devem ser lidos e interpretados pelo usuário.

Descoberta por Sumarização

Sumarização é a abstração das partes mais importantes do conteúdo de um documento ou conjunto de documentos, ou seja, um resumo ou sumário gerado a partir das palavras ou frases mais importantes.

Muitos dos programas de sumarização disponíveis se fundamentam em uma simples extração de fragmentos importantes do texto para produzir resumos. Eles podem ser classificados nas seguintes categorias gerais: abordagens dependentes do domínio e abordagens independentes do domínio. A primeira usa conhecimento de um domínio específico e a estrutura do texto (da área médica, financeira, ...) para produzir resumos de alta qualidade. A segunda emprega várias técnicas lingüísticas e estatísticas para identificar as sentenças-chave do documento. O problema com extratores de textos é que eles produzem geralmente resumos sem consistência e incoerentes.

Existem os chamados "text abstractors" que visam transpor estas limitações através da análise léxica do texto original, encontrando novos e menores conceitos para descrevê-lo. Ao invés de extrair sentenças, eles as transformam automaticamente, produzindo um resultado mais conciso. Um exemplo é o *auto-resumo* do MS-Word ®, embora existam produtos com algoritmos mais avançados.

As técnicas de sumarização podem ser empregadas também após os processos de "clustering" na análise do centróide (documento utilizado como parâmetro de comparação de um grupo). A partir da análise do centróide, pode-se obter uma visão geral do assunto tratado na coleção de documentos.

Descoberta por Agrupamento ("Clustering")

Agrupamento é um método que identifica similaridades entre documentos, alocando-os em grupos de acordo com o grau de similaridade, obtendo grupos de assuntos. Portanto, o agrupamento pode ser usado antes da classificação (método estudado a seguir).

O agrupamento pode ser hierárquico ou isolado. No agrupamento hierárquico ou partição hierárquica ("hierarchic partition"), os grupos formados possuem um relacionamento entre si, gerando uma árvore, em que as folhas representam os grupos mais específicos e os nós intermediários, os grupos mais abrangentes. Já no agrupamento isolado ou agrupamento por partição ou de partição total ("flat partition"), os grupos não apresentam um relacionamento entre si.

Os algoritmos de agrupamento podem ser classificados quanto à complexidade em relação ao tempo de processamento como *constantes*, *lineares* e *exponenciais de ordem quadrática*.

Os algoritmos de tempo constante tentam limitar o tempo máximo de processamento ou, pelo menos, descobrir o tempo gasto por cada elemento e estimar um tempo ou número máximo de comparações necessárias. Não existem algoritmos deste tipo que possam ser usados em processamento de tempo real.

Os algoritmos de tempo linear aumentam linearmente o tempo de processamento de acordo com o número de elementos, já que nem todos precisam ser comparados mutuamente.

Os algoritmos exponenciais de ordem quadrática gastam tempo de forma exponencial, porque sempre que um elemento é adicionado, é comparado com todos os outros.

As etapas básicas de agrupamento são:

a) Identificação e seleção de características: identifica palavras nos documentos e seleciona as de maior grau de discriminação. O resultado são listas de palavras relevantes que identificam o documento.

b) Cálculo de similaridade: identifica o grau de similaridade entre os documentos.

c) Agrupamento: identifica correlações entre os elementos da matriz gerada a partir do cálculo de similaridades.

Descoberta por Classificação e Categorização

Classificação é uma técnica para identificar a classe ou categoria (assunto) a que pertence determinado documento. Para isto, as classes devem ser pré-definidas.

Classificação e Categorização são consideradas sinônimos. Entretanto, alguns autores consideram Classificação um processo que aloca o documento em uma classe, enquanto a Categorização identifica os assuntos de que tratam os documentos, criando as classes.

Os sistemas de classificação utilizam geralmente as seguintes técnicas (WIVES, 2000):

a) regras de inferência: baseiam-se em um conjunto de características para identificar a classe de um documento. São específicas para cada domínio.

b) modelos conexionistas (redes neurais artificiais): induzem um conjunto de regras a partir de arquivos de treinamento. São capazes de detectar mudanças nos dados. Seus resultados não são facilmente compreendidos.

c) método de similaridade de vetores ou centróides: as classes são representadas por vetores de palavras (centróides). O documento é comparado com cada vetor para ser agrupado.

d) árvores de decisão: utilizam técnicas de aprendizado de máquina para induzir regras.

e) classificadores de Bayes: têm como base a teoria da probabilidade. Informam a probabilidade de um documento pertencer a uma classe.

Descoberta por Filtragem de Informação

Filtragem de informação é uma espécie de classificação de informações. Podem ser identificados dois tipos de sistemas de filtragem: sistemas de recomendação e sistemas de filtragem colaborativa.

Os sistemas de recomendação ("recommender"/ "recommendation systems") são capazes de analisar uma série de alternativas e escolher aquelas que são úteis ou combinam com o perfil do usuário ou de vários usuários que já tenham usado o sistema.

Existem dois tipos de sistemas de recomendação: por conteúdo e colaborativo. O sistema de recomendação por conteúdo recomenda itens similares aos que o usuário escolheu no passado. O segundo identifica usuários cujas preferências são similares ao do usuário atual e os recomenda.

Já os sistemas de filtragem colaborativa são mais simples porque analisam as recomendações definidas por outros usuários do sistema, filtrando as adequadas e encaminhando-as aos usuários interessados.

2.3.4 Mineração da Web ("Web mining")

Mineração da Web pode ser definida como a descoberta e análise de informação útil da "World Wide Web". Refere-se tanto à pesquisa e recuperação automáticas de informação e recursos disponíveis em milhões de "sites" e bancos de dados on-line ("Web content mining") quanto à descoberta e análise de padrões de usuários de um ou mais servidores Web ou serviços on-line ("Web usage mining"). Esta taxonomia é mostrada na figura 2.24.

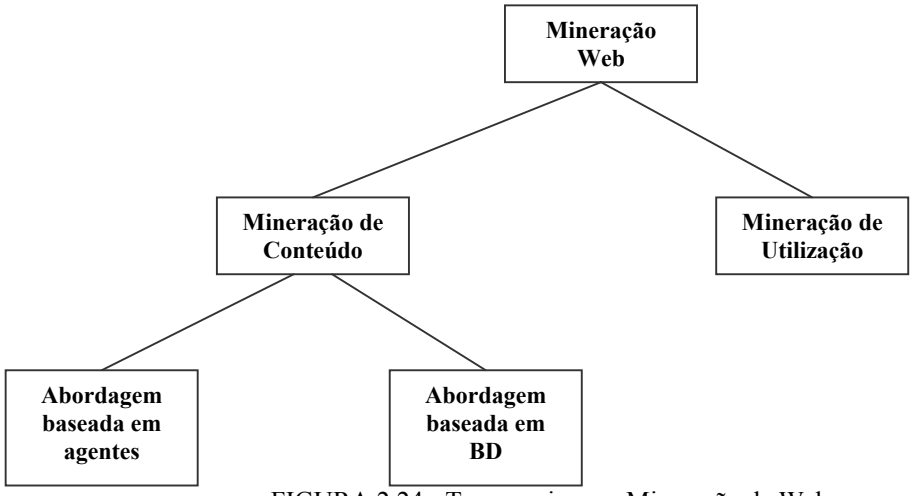


FIGURA 2.24 - Taxonomia para Mineração da Web

Desta forma, a Mineração da Web pode ser classificada como Mineração de Conteúdo ("Web content mining") e Mineração de Utilização ("Web usage mining") [COO 97]:

a) Mineração de Conteúdo: Há duas abordagens: uma baseada em agentes inteligentes e outra em banco de dados. A abordagem baseada em agentes inteligentes envolve o desenvolvimento de sistemas sofisticados de IA, que podem agir de forma autônoma ou quase autônoma sobre o comportamento do usuário, para descobrir e organizar a informação da Web. Sistemas de mineração Web baseados em agentes se apresentam nas seguintes categorias: agentes inteligentes de pesquisa; agentes para categorização e filtragem de informações e agentes de personalização (que aprendem as preferências do

usuário e descobrem fontes de informação Web baseados nestas preferências).

A abordagem baseada em bancos de dados se preocupa com técnicas para integrar e organizar dados heterogêneos e semi-estruturados da Web, transformando-os em bancos de dados relacionais, por exemplo, e usando mecanismos de consulta padrão, além de técnicas de mineração de dados para analisar as informações. As categorias desta abordagem são bancos de dados multiníveis e sistemas de consulta Web. A primeira categoria mantém, em um nível mais baixo, um banco de dados semi-estruturados armazenado em vários repositórios Web como documentos hipertexto. No nível mais alto, aparecem meta-dados extraídos dos níveis mais baixos e organizados em coleções estruturados (bancos de dados relacionais ou orientados a objetos). Já a segunda categoria, envolve a tentativa de usar no ambiente Web linguagens de consulta padrão como SQL e mesmo processamento de linguagem natural para lidar com as consultas formuladas pelos usuários.

b) Mineração de Utilização: Há duas ferramentas: de descoberta de padrão e de análise de padrão. As ferramentas de descoberta de padrão usam sofisticadas técnicas de IA, mineração de dados, Psicologia e teoria da informação para minerar conhecimento a partir dos dados coletados. Já as ferramentas de análise de padrão analisam os resultados das ferramentas de descoberta de padrão, oferecendo aos analistas técnicas apropriadas para compreender, visualizar e interpretar estes padrões.

Mineração da Web levanta várias questões importantes como a necessidade de integrar várias fontes de dados tais como “logs” de acesso de um servidor, registros de usuários, solucionar problemas de identificação do usuário e de suas sessões e transações, topologias dos “sites” e modelo de comportamento dos usuários. Trata-se de um trabalho de pesquisa de várias áreas como RI, BD, IA e outras.

3 Estudo de Caso

O objetivo deste estudo de caso é avaliar a efetividade, ou seja, a habilidade das ferramentas Eurekha e Umap de tomarem decisões corretas ao agrupar documentos. O Eurekha utiliza a técnica de agrupamento ou "clustering", para agrupar documentos similares. Foram escolhidos os algoritmos "best-star" e "cliques", uma vez que o algoritmo "best-star" é um aperfeiçoamento do algoritmo "stars" e o "cliques" é mais rigoroso na alocação de documentos. O Umap executa o processo de lematização, filtragem de palavras-chave e através de sua "interface", o mapa dinâmico, é capaz de gerar grupos de palavras através dos quais se pode obter grupos de documentos.

Ao contrário de Evans [EVA 98], a avaliação não é feita a partir de julgamento de relevância, ou seja, grupos gerados pelas ferramentas não foram analisados por usuários. Isto poderia influenciar os julgamentos. A bibliotecária e especialista em Sistemas de Informação da Universidade Federal do Pará Irvana dos Santos Coutinho fez a leitura da coleção de textos e agrupou os mais similares. Estes grupos foram tomados como parâmetros de comparação com os resultados das ferramentas, a fim de verificar se a recuperação automática pode ser comparada à recuperação humana. Devido à existência de um parâmetro de comparação, foram escolhidas as *medidas externas* para a avaliação das duas ferramentas: "microaveraging" (calcula "recall" e "precision" para a coleção inteira) e "macroaveraging" (calcula "recall" e "precision" para cada grupo, a fim de obter a média).

A coleção de teste é composta por 178 textos do Jornal *Folha de São Paulo* publicados no ano de 1999, que apresentam os termos "Amazônia" e/ou "Amazônica". Com o processamento, espera-se também obter conhecimento sobre os temas que marcaram o ano de 1999 e saber qual a visão de um jornal, com sede fora do território amazônico a respeito dele.

O processo de descoberta de conhecimento foi proativo, uma vez que não foram formuladas hipóteses iniciais. Já o processamento destes documentos é feito da seguinte maneira: a coleção inteira de 1999, os documentos correspondentes aos nove primeiros meses, aos seis primeiros meses, aos três primeiros meses e ao mês de janeiro (semelhante a [JIA 2001]). Com isto, quer-se:

a) avaliar o tempo de processamento para diferentes quantidades de documentos;
b) verificar se os valores de "microaveraging precision" para cada uma das ferramentas foram influenciados pela quantidade de documentos;
c) descobrir os temas que marcaram cada período do ano de 1999.

Por outro lado, os métodos de avaliação de SRI, durante muito tempo, basearam-se apenas nas medidas de efetividade, quando deveriam também atentar para a usabilidade da "interface" com o usuário e se o mesmo consegue alcançar seus objetivos [HAN 98]. Usabilidade está relacionada à:

a) facilidade de aprendizado (intuitiva e "natural");
b) flexibilidade de interação (multiplicidade de formas);
c) e robustez de interação (acompanhamento e recuperação).

Os requisitos citados abaixo foram definidos por usuários para avaliação de "interfaces" de SRI [HAN 98] e serão utilizados neste trabalho:

a) Suporte à exploração e à pesquisa: O usuário não deve ser forçado a fazer uma nova consulta, quando quiser explorar os dados que ele já possui.

b) Nível de usuário ("novice" vs. "expert"): O SRI deve oferecer meios para identificar o perfil do usuário e saber quanto conhecimento ele possui a coleção.

c) Suporte à aprendizagem: O SRI deve oferecer meios de aprender sobre como manipular o mesmo de forma correta.

d) Apoio à decisão: A ferramenta deve fornecer meios de o usuário decidir se o resultado obtido é ou não satisfatório.

e) Nível de controle: Está relacionado com o grau de interatividade do SRI durante a pesquisa.

f) Meio de comunicação: O SRI deve fazer recomendações aos usuários e permitir estabelecer contatos com outros pesquisadores para colaboração e comunicação.

O próximo tópico discute as principais características das ferramentas utilizadas: Eurekha e Umap.

3.1 Sistemas de Descoberta de Conhecimento em Texto utilizados

Os sistemas de Descoberta de Conhecimento em Textos (KDT) utilizados neste estudo de caso foram: Eurekha e Umap. O Eurekha é um *software* acadêmico, que apresenta quatro tipos de algoritmos para agrupar documentos similares: "stars", "full-star", "best-star" e "cliques". O Umap é um *software* comercial, que, a partir das palavras-chave encontradas em uma coleção de documentos, cria um mapa através do qual se pode extrair os grupos de documentos existentes. A seguir, os aspectos mais importantes destas ferramentas são descritos.

3.1.1 Eurekha 5.1

Eurekha, desenvolvido por Wives [WIV 99], foi implementado em linguagem C++ (com algumas características de orientação a objetos) no ambiente de programação do *Borland CBuilder 3.0*. Seu objetivo é realizar o processo de KDT a partir do agrupamento de documentos. Além disso, gera um arquivo em formato ASCII, com as palavras mais relevantes do grupo. Os documentos de entrada podem estar em formato ASCII ou HTML.

O Eurekha é baseado na *Hipótese de Agrupamento*: documentos similares e relevantes ao mesmo assunto tendem a permanecer em um mesmo grupo ("cluster"). As etapas são as seguintes:

a) identificar palavras;
b) remover palavras negativas ("stopwords");
c) calcular a importância das palavras;
d) selecionar as palavras mais relevantes;

e) calcular a similaridade entre os objetos;
f) escolher o agrupamento: “best-star”, “stars”, “cliques”, “full-star”.

A figura 3.1 mostra a janela inicial do Eurekha. Cada coleção de documentos testada é considerada um projeto. Entretanto, apenas a versão mais recente do Eurekha permite nomear e gravar o projeto testado no disco rígido. Após isto, é possível verificar as palavras mais relevantes de cada documento, dando um duplo clique no mesmo.

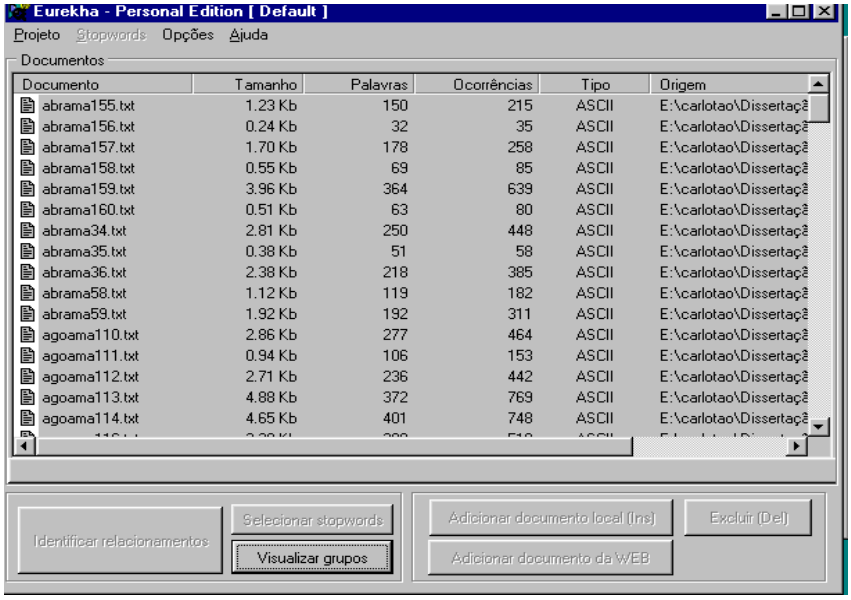


FIGURA 3.1 - Janela inicial do *software* Eurekha

O Eurekha permite definir a lista de “stopwords” ou palavras negativas (palavras que não expressam significado relevante no texto). É possível agrupá-las, identificando as categorias como por exemplo: advérbios, pronomes, conjunções, interjeições e outros. Porém, o programa só as leva em conta na hora do processamento, quando são selecionadas. Pode-se selecionar também a opção “Considerar números como stopwords” para acelerar o processamento.

O botão “Identificar relacionamentos” calcula o número de comparações a serem feitas entre os documentos submetidos, estima o tempo a ser gasto e informa o tempo de processamento no final. Ele também gera a matriz de similaridades, que contém o grau de similaridade entre todos os documentos. Os grupos gerados podem ser visualizados, clicando no botão “Visualizar grupos”.

Os grupos podem ser visualizados a partir de quatro algoritmos: “stars”, “best-star”, “cliques” e “full-star”. A figura 3.2 mostra o agrupamento gerado e um gráfico com a porcentagem de documentos para cada grupo. Dependendo da coleção e do grau de similaridade (coeficiente de sensibilidade, na figura 3.2, que vai de 0 a 100%), os algoritmos podem ou não gerar agrupamentos diferentes. Eles são:

a) “Stars”: faz a seleção de um documento e identifica todos os documentos conectados (similares) a ele, como uma estrela. O problema do algoritmo “stars” é que cada documento é atribuído ao primeiro grupo, cujo grau de similaridade com o documento central é maior que o mínimo exigido, não havendo garantia de que o novo documento seja alocado ao grupo de maior afinidade. Desta forma, a ordem dos elementos na matriz de similaridade afeta o resultado.

b) "Best-star": Foi desenvolvido para solucionar o problema do algoritmo "Stars", pois este algoritmo atribui o documento ao grupo de maior similaridade e não simplesmente ao primeiro testado.

c) "Cliques": caracteriza-se pelo fato de um documento só ser alocado a um grupo, se o grau de similaridade entre este e os documentos já alocados, for um valor maior ou igual ao indicado pelo usuário. Assim, é melhor e mais demorado.

d) "Full-star": atribui um documento a todos os grupos cujo grau de similaridade ultrapasse o valor mínimo estabelecido, ou seja, permite que um documento pertença a mais de um grupo.

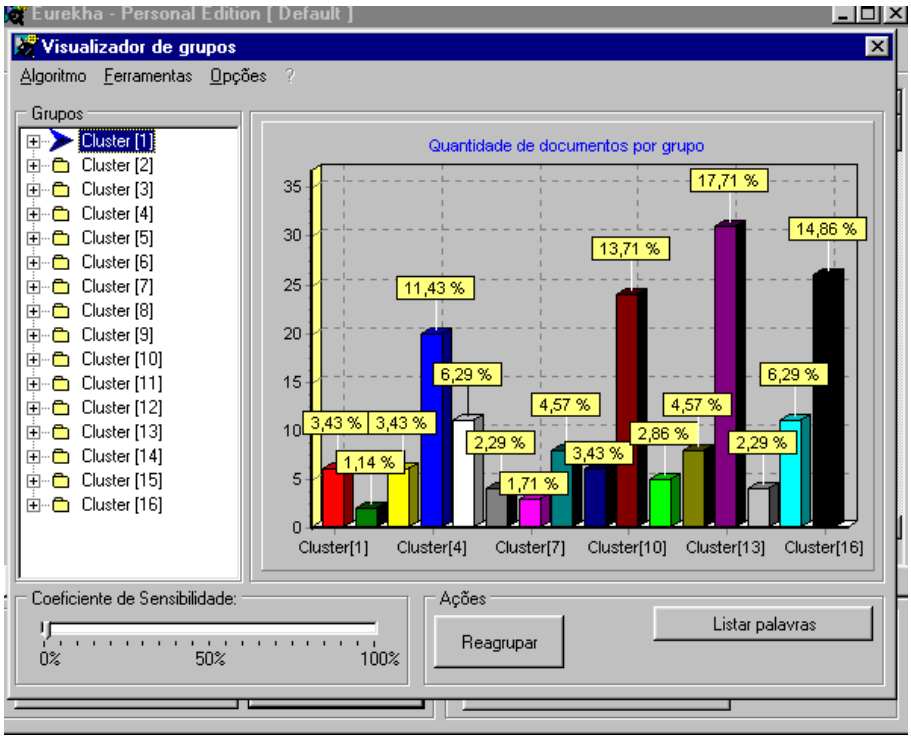


FIGURA 3.2 - Janela para visualização dos agrupamentos obtidos no Eurekha

O botão "Listar palavras" apresenta o número de documentos em que a palavra ocorre e sua relevância, a partir do cálculo da freqüência relativa. Estas informações podem ser exportadas para o arquivo *palavras.txt*. Já o botão "Reagrupar" reagrupa os documentos cada vez que um novo algoritmo é selecionado no *menu* "Algoritmo".

3.1.2 Umap

O UMAP foi desenvolvido pela "Trivium" (http://www.trivium.com.fr). Ele se baseia na tecnologia de *Árvores de Conhecimento*, cujos fundamentos tecnológicos e filosóficos foram dados por Pierre Levy e Michel Authier. A *Árvore do Conhecimento* é uma forma técnica para a *Ideografia Dinâmica*, que visa fornecer uma representação dinâmica (linguagem) para modelos mentais. É capaz de alterar radicalmente o papel do criador que trabalha sobre "interfaces", transformando o espectador em um ator criativo. O componente que representa esta tecnologia, no UMAP, é o *mapa dinâmico*, que é construído após o processo de mineração dos documentos, composto por lematização e filtragem de palavras-chave.

Umap é apresentado nas seguintes versões:

a) "Umap Universal": mapeia documentos texto em mais de sessenta formatos como *Word*, *Excel*, *Adobe Acrobat* e muitos outros.

b) “Umap for Word”: auxilia na tarefa de ler longos documentos do *Microsoft Word*, sendo disponível apenas para o *MS WinWord 97*.

c) "Umap for Outlook": auxilia a descobrir e explorar a informação armazenada no *Outlook*, *software* para leitura de *e-mails*.

d) "Umap Web": submete à Internet uma questão do usuário e depois fornece uma visão compreensível da informação disponível obtida.

A área de trabalho do UMAP Web é apresentada na figura 3.3. A janela esquerda contém a lista de palavras-chave ("List of Keywords"), cuja função é estabelecer coerência temática ("thematic coherence"), ou seja, identificar as palavras comuns a todos os documentos ou a grupos de documentos. Cada vez que documentos ou páginas são eliminados, a lista é redefinida.

A janela da direita é denominada *documentos* ("The Documents"), que contém os documentos a serem explorados ou o domínio sobre o qual se quer aprender. Os valores da coluna “Rating” mostram quantas palavras, daquelas que estão selecionadas no mapa, existem em cada documento. Um duplo clique no nome do documento exibe o seu conteúdo. Um clique, no sinal de “+”, à esquerda de cada documento, mostra as palavras mais relevantes do mesmo. As colunas “Selection” e “Prox.” indicam em que documentos e qual a freqüência no documento (através da cor dos quadrinhos) das palavras selecionadas no mapa.

A janela central apresenta o *arquipélago* ou mapa ("map"), cujo objetivo é fornecer uma visão global do conteúdo dos documentos. O mapa é uma ferramenta de compreensão visual, um modelo para todas as palavras-chave. A forma do mapa depende das ligações existentes entre as palavras (quando uma é eliminada, ele toma novo formato). A palavra que representa cada pedaço do mapa pode ser visualizada, pressionando o botão esquerdo do "mouse" sobre o mesmo. A palavra aparece em legenda ("landmark"), como na figura 3.3.

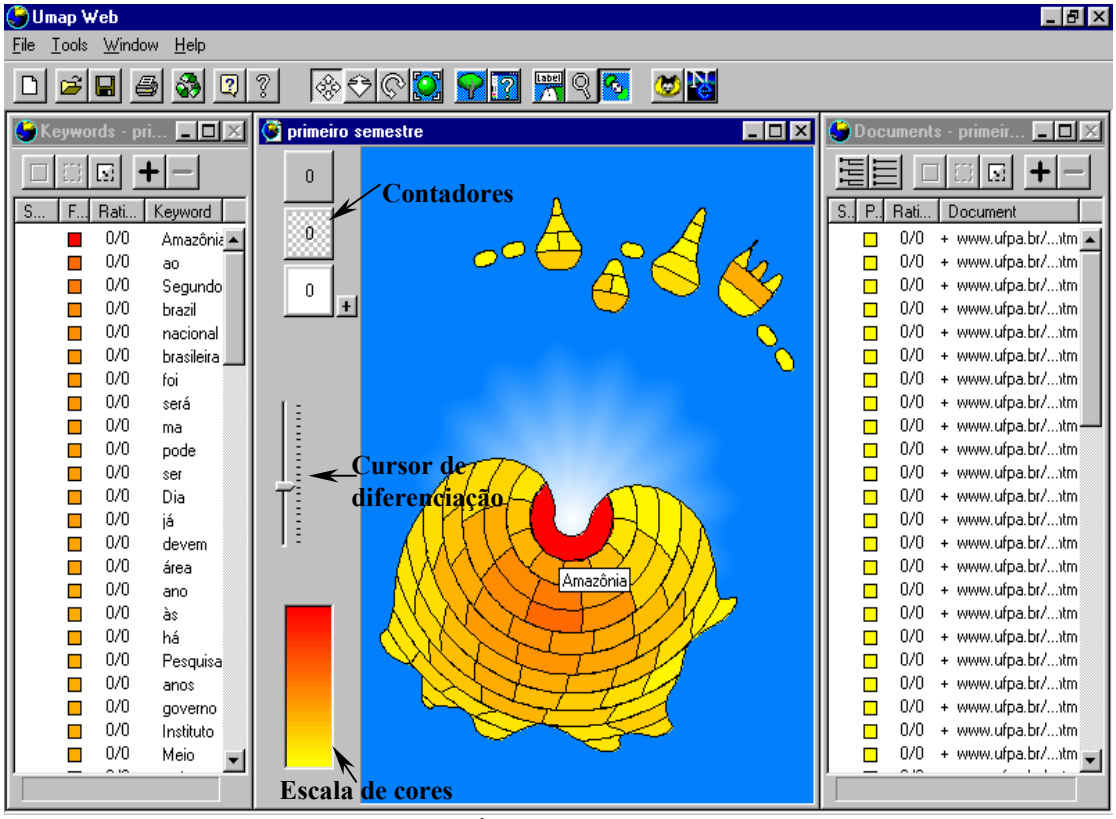


FIGURA 3.3 - Área de trabalho do "UMAP"

Através de um menu suspenso, pode-se mudar características do mapa como a cor e as preferências. O analisador de preferências ("Analyzer Preferences") permite indicar quantas palavras devem ser mostradas na janela "Keywords": 100, 200, 300, 400 ou 500. Além disso, permite que o usuário informe o seu nível de conhecimento sobre a coleção: "novice" (a lista de "keywords" mostra muitos tópicos comuns) e "expert" (a lista de "keywords" mostra tópicos que ocorrem freqüentemente em alguns textos específicos e menos freqüentemente em outros).

O ponto inicial de exploração do mapa é o *ponto focal* ("focal point"). Quanto mais próxima uma palavra está do ponto focal mais geral e distribuído é o assunto a que ela se refere, quanto mais distante mais ela representa um tópico específico. Por isso, na figura 3.3, a palavra *Amazônia* surge no ponto focal, já que todos os documentos abordam o tema *Amazônia*.

O mapa, dependendo das ligações entre as palavras e da posição do cursor de diferenciação, é composto de continente ("continents"), ilhas ("islands") e "subilhas" ("islets"). Em cada ilha, as palavras que estão mais próximas do centro (a base) identificam o tópico básico dos documentos. Um continente representa tópicos secundários de um tópico geral; uma ilha representa o desenvolvimento específico de

um tópico importante e “subilha” representa uma palavra usada de forma dominante em um texto específico (pode ser apenas um ruído).

O padrão é que esteja sempre visível o cursor de diferenciação ("differentiation slider") e oculto o cursor de hierarquia ("hierarchy slider"). O Umap calcula um valor ótimo para a razão entre visibilidade e taxa de erro para cada valor do cursor. O cursor de diferenciação, próximo de 0, significa que a taxa de erro é nula, formando um mapa difícil de interpretar por possuir poucas zonas diferenciadas, ou seja, poucas ilhas e, portanto, menos grupos. Mas, há maior confiança neste mapa por preservar as reais ligações entre palavras. Um cursor de diferenciação, próximo de 100, permite visualizar no mapa apenas as ligações mais fortes entre as palavras. Apresenta taxa de erro maior, porém mais zonas diferenciadas e maior facilidade de interpretação.

O cursor de hierarquia é utilizado para verificar qual dos tópicos é o mais importante. Pode ser utilizado para encontrar o assunto mais representativo da coleção de documentos explorada. Próximo de 0, os grupos são empacotados em um único nível. Próximo de 100, o assunto mais importante é evidenciado.

Por outro lado, a cor de uma palavra-chave indica o número de textos em que ela aparece. A cor do topo da *Escala de Cores* (vermelha é o padrão) é a cor das palavras com maior ocorrência nos textos. A última cor da escala identifica as palavras com menor freqüência. Clicando sobre a escala, são selecionadas as palavras correspondentes àquela cor. Já os contadores informam o número de palavras-chave “marcadas”, “selecionadas” e “marcadas e selecionadas”.

Um personagem importante surge a cada alteração nos cursores: o gato. Ele descreve através de conselhos enumerados a situação do mapa e diz o que deve ser feito. Assim, o Umap auxilia o usuário a alcançar um cursor de diferenciação ótimo para a análise da coleção de documentos.

3.2 Seleção dos dados e pré-processamento

Os textos foram obtidos através do CD “Folha edição 2000”. Dos 609 artigos, que continham os termos *Amazônia* e/ou *Amazônica,* 178 foram selecionados. Informações irrelevantes foram eliminadas antes do processamento como autor, data, palavras-chave e outras. Porém, há possibilidade de existirem outros ruídos como erros ortográficos.

Os arquivos foram gravados em formato HTML, para os testes com Umap, e em formato TXT (ASCII), para os testes com o Eurekha. Os arquivos receberam a denominação através do ***mês*** da publicação mais um número de ***seqüência***. Ex: jan179.txt (Este documento foi publicado no mês de janeiro, seu número seqüencial é 179 e seu formato é ASCII).

O processador Pentium II-MMX de 333 Mhz e 65Mbytes de RAM foi utilizado nos testes.

3.3 Experiência com o Eurekha

O primeiro teste foi realizado com o Eurekha. Os seguintes dados foram anotados:

a) período;

b) número de documentos;

c) tempo de processamento;

d) tipo do algoritmo "best-star" ou "cliques";

e) grau de similaridade;

f) número de grupos gerados por cada algoritmo;

g) os documentos de cada grupo;

h) as palavras que identificam o tópico tratado por aquele grupo em ordem decrescente de relevância;

i) o número de documentos do grupo em que a palavra ocorre e o grau de importância ou relevância da mesma para o grupo;

j) o grupo mais representativo (ou seja, aquele que tem maior número de documentos, expresso em porcentagem), para os testes em que não existem documentos isolados (desagrupados).

Conforme visto anteriormente, o Eurekha permite visualizar os grupos através de quatro algoritmos: "best-star", "cliques", "full-star" e "stars". Neste experimento, só foram analisados os grupos dos algoritmos "best-star" e "cliques". Os graus de similaridade foram configurados em 0%, 2,5%, 5%, 10% e 15%, pois além de 15% o agrupamento se tornou mais difícil.

O Eurekha permite que se defina uma lista de "stopwords" ou palavras negativas. As categorias são apresentadas na tabela 3.1.

TABELA 3.1 - Lista de "stopwords" utilizada em todos os experimentos[7]

CATEGORIA	PALAVRAS NEGATIVAS
Pronomes Pessoais	eu, tu, ele (s), ela (s), nós, vós, me, mim, comigo, te, ti, contigo, nos, vos, conosco, convosco, lhe (s), se, si, consigo, você (s), Sr, senhor
Pronomes Possessivos	meu (s), teu (s), minha (s), tua (s), nosso (s), nossa (s), vosso (s), vossa (s), seu (s), sua (s)
Pronomes Demonstrativos	este (s), esta (s), esse (s), essa (s), isto, isso, aquele (s), aquela (s), aquilo, mesmo (s), mesma (s), próprio (s), própria (s), semelhante (s), tal, tais
Pronomes Relativos	onde, como, quando, que, cujo (s), cuja (s)
Pronomes Indefinidos	algo, alguém, ninguém, muito, cada, outrem, quem, algum, alguma (s), alguns, todo (s), toda (s), nenhum (a), muito (s), muita (s), pouco (s), pouca (s), certo (s), certa (s), diverso (s), diversa (s), vário (s), vária (s), quais, qual, outro (s), outra (s), quanto (s), quanta (s), quaisquer, qualquer, tudo, outrem, cada
Pronomes Interrogativos	quem, cadê
Advérbios de Afirmação	sim, deveras, certamente, realmente
Advérbios de Negação	não, absolutamente, tampouco
Advérbios de Dúvida	talvez, quiçá, decerto, porventura, acaso, provavelmente, possivelmente
Advérbios de Intensidade	muito, pouco, bastante, mais, menos, demais
Advérbios de Lugar	aqui, ali, aí, acolá, lá, atrás, perto, longe, abaixo, acima, adiante, dentro, fora, além
Advérbios de Modo	bem, mal, depressa, devagar, calmamente, preconceituosamente
Advérbios de Tempo	hoje, amanhã, nunca, jamais, breve, logo, antes, depois, agora, já, sempre, cedo, tarde, outrora, diariamente, anualmente, antigamente, novamente, entrementes, imediatamente, raramente
Numerais Cardinais	um, dois, três, quatro, cinco, seis, sete, oito, ... milhão (ões), bilhão (ões), trilhão (ões)
Numerais Ordinais	primeiro, segundo, terceiro, quarto...
Meses	jan, fev, mar, abr, ..., dez, janeiro, fevereiro, ..., dezembro
Artigos	a (s), o (s), um, uns, uma, umas
Dias da semana	seg, ter, qua, qui, sex, sab, dom, segunda-feira, terça-feira, quarta, quinta-feira, sexta-feira, sábado, domingo
Preposições	a, ante, até, após, com, contra, de, desde, por, para, perante, sem, sobre, sob, trás, ao (s), à (s), da (s), do (s), no (s), na (s), pelo (s), pela (s), dum (ns), duma (s), num (ns), numa (s), daquilo, daquele (s), daquela (s), afora, entre, nisso, aquilo, naquele (s), naquela (s), pra (o), em, defronte, através, invés, àquele (s), àquela (s), àquilo
Conjunções	Que, mas, porém, ou, e, portanto, todavia, entretanto, nem, contudo, ora, pois, logo, assim, porque, como, apesar, ainda, conforme, segundo, consoante, tão, tamanho, quando, mal, enquanto, desde, salvo, conseguinte, contanto, (no) entanto, contanto, já
Geral	Anexo (s), anexa (s), obrigado (a), junto (s), quite (s), meio

Locuções prepositivas, adverbiais e outras não foram inseridas, uma vez que a versão do Eurekha permite inseri-las (por exemplo: “no entanto”, “uma vez que”) como

[7] A lista de palavras negativas da tabela 3.1 foi definida com o auxílio da seguinte bibliografia: SACCONI, Luiz Antônio. **Gramática Essencial da Língua Portuguesa**: teoria e prática, 4 ed. revisada. São Paulo: Atual, 1989.

palavras negativas, mas não as leva em consideração durante o processamento. A nova versão do Eurekha oferece solução para este problema.

Os próximos tópicos apresentam a descrição dos experimentos com cada coleção de documentos.

3.3.1 Teste com a coleção inteira

Para os graus de similaridade 0 e 2,5%, o algoritmo "best-star" alocou os 178 documentos em 52 grupos. Já que nenhum documento ficou isolado, pode-se afirmar que há dois grupos representativos do ano de 1999, os grupos 6 e 17 com percentagem de 4% cada um; o primeiro aborda *queimadas* e *pesquisa* e o segundo, *narcotráfico*. Porém estas palavras não caracterizaram apenas estes grupos. Também apresentaram a palavra *queimadas* os grupos 2, 15, 33 e 51 e a palavra *narcotráfico*, os grupos 9, 14, 47, 45, 46 e 48. Isto quer dizer que estas palavras não apresentaram o mesmo grau de relevância ou importância nos documentos, por isso, foram alocadas em grupos diferentes.

No grupo 3, as palavras *satélite* e *satélites* demonstraram a ausência do processo de lematização, discutido anteriormente. Assim, o Eurekha não apresenta as desvantagens decorrentes deste processo. O grupo 26 apresentou outro exemplo com as palavras *Ticuna* e *Ticunas* (tribo indígena).

Os grupos 28 e 31 apresentaram como tema principal *acidente* de *barco*, porém seus documentos foram alocados separadamente, no grupo 28, as palavras *barco*, *pessoas*, *afundou*, *desaparecidas* e, no grupo 31, *barco*, *naufrágio*, *corpos*, *acidente.* Trata-se do problema do vocabulário.

O Eurekha apresenta palavras soltas para descrever o conteúdo dos grupos. Isto dificulta o entendimento do usuário quando da presença de siglas desconhecidas como LBA (sigla em inglês do Experimento de Grande Escala da Biosfera-Atmosfera na Amazônia), IMAZON (Instituto do Homem e Meio Ambiente da Amazônia), IPAM (Instituto de Pesquisa Ambiental da Amazônia), SENAD (Secretaria Nacional Antidrogas) e outras. Por outro lado, as palavras *rio* e *madeira* do grupo 28 podem levar o usuário a pensar que se trata de dois substantivos comuns, não relacionados, já que *madeira* é um produto de extração da Amazônia. No texto, a palavra *madeira* é um nome próprio, denominação do rio, local onde ocorreu o acidente de barco. Neves [NEV 2000] classifica as palavras "Concílio de Granges" e "Avenida Quinze" como *substantivos próprios compostos*, que devem ser considerados como um "conjunto unitário". Este é o caso das palavras *rio* e *madeira*, portanto não devendo ser consideradas separadamente como o faz o Eurekha.

Um caso diferente ocorreu no grupo 2 com as palavras "intoxicação" e "mercúrio", em que não se trata de substantivos próprios compostos e sim substantivos comuns, portanto não havendo, gramaticalmente, a obrigatoriedade de considerar como um "conjunto unitário". Pode-se utilizar de conhecimento prévio do domínio ("background knowledge") para identificar problemas como *intoxicação* por *mercúrio*. A intoxicação por mercúrio é causada pelo consumo de água de rios, que atravessam áreas de garimpo, onde o mercúrio é utilizado para separar o ouro da terra. É um problema característico da região.

Também há a possibilidade de se fazerem afirmações incorretas, porque o Eurekha não propõe solução para o problema da negação. Um termo pode aparecer na matéria jornalística e estar sendo negado. Uma solução foi proposta em [LOH 2000a] através da Descoberta de Conhecimento em Texto baseada em Conceito. Apesar disso, pode-se afirmar que a presença da palavra *malária* demonstra que a doença ainda não

foi erradicada na região, novamente, valendo-se de conhecimento prévio do domínio ("background knowledge"). A palavra ocorreu nos grupos 27 e 37 associada às palavras *índios*, *casos, Roraima* e *Acre*.

Por outro lado, às vezes, é difícil identificar o assunto de que trata um determinado grupo só a partir das palavras mais relevantes. Mas o Eurekha fornece uma maneira rápida de ler os documentos que compõem qualquer grupo. Na figura 3.2, basta clicar no sinal de "+" para que o grupo seja expandido. Selecionando um documento, seu conteúdo pode ser visualizado no lugar do gráfico de colunas.

Inicialmente, o usuário é levado a pensar em determinar verbos como palavras negativas. O Eurekha permite transformar uma palavra em palavra negativa na lista de palavras do grupo através do botão "Transformar em sw". Alguns verbos foram considerados "stopwords" como *ser*, *estar*, *ir*, *afirmar* e outros; mas alguns foram essenciais para o entendimento do grupo: descobrir (grupo 2), afundar (grupo 28), asfaltar (grupo 7).

Pelo menos, quatro grupos apresentam um assunto bem delimitado. O grupo 7 abordou o problema da ausência de infra-estrutura na região, a necessidade de *asfaltar* a rodovia *Santarém-Cuiabá*, importante principalmente para os municípios do sul do Pará. A presença da sigla *FHC* (iniciais de Fernando Henrique Cardoso, presidente da república) e o substantivo próprio *Almir* (Almir Gabriel, governador do Pará) demonstraram que o problema foi (ainda é) de âmbito estadual e federal. O grupo 22 aborda o problema da venda da Eletronorte, que atende parte da região. Os grupos 49 e 50 abordam o manejo sustentável (o extrativismo de forma inofensiva à floresta) e o selo de certificação concedido a empresas que exploram a floresta de forma sustentável.

A palavra *narcotráfico* ocorreu em vários grupos: 9, 14, 16, 17, 44, 46, 47 e 48. Isto demonstra que o tema apresentou temas secundários. Com exceção dos grupos 47 e 48, que abordaram a *CPI* (Comissão Parlamentar de Inquérito) do narcotráfico realizada em Brasília, os outros estavam relacionados à Colômbia; ressalte-se o fato de que a Amazônia é uma floresta densa que avança sobre vários países da América Latina. No ano de 1999, houve o agravamento da crise colombiana, que envolve grupos *guerrilheiros* como as *FARC* (Forças Armadas Revolucionárias da Colômbia), cuja base de sustentação financeira é o *narcotráfico*. Este agravamento transcendeu as fronteiras do país. O risco de *intervenção militar* norte-americana se tornou iminente, já que os norte-americanos são um dos grandes consumidores do produto. *Mccafrey* é o sobrenome do general Barry Mccafrey, conhecido como czar norte-americano do combate às drogas, que também se destacou nos grupos.

Em 10%, o grupo 4 apresentou uma peculiaridade: a presença do numeral cardinal *mil* escrito por extenso. Em todos os testes, os numerais cardinais foram considerados "palavras negativas", mas, no grupo 4, a palavra *mil* faz referência à "Marcha dos 100 mil", promovida pelo MST (Movimento dos Sem-Terra).

Em 15%, apenas 10 documentos foram agrupados. Três grupos trataram do *desmatamento* e os outros dois, sobre *queimadas* e acidente de *barco.*

O algoritmo "cliques" apresentou um comportamento totalmente diferente do algoritmo "best-star" para o grau de similaridade 0%. Enquanto o "best-star" gerou vários grupos, o "cliques" gerou apenas um grupo. Apesar disso, ele apresentou as palavras mais relevantes para a coleção. Neste caso, a palavra mais mencionada foi *Amazônia*, ocorrendo em 148 documentos com relevância de 0,04. Da mesma forma, o problema regional mais abordado na coleção é o *desmatamento* com 24 ocorrências e grau de relevância em 0,003.

Em 2,5%, o grupo 3 apresentou as palavras *dalai* e *lama*, que formam o substantivo próprio composto "Dalai Lama". O Eurekha não tem suporte para manter

estas palavras como um “conjunto unitário”. Em 5%, o grupo 47 apresentou as palavras *barco, pessoas, causa, superlotação*, levando à suposição de que a causa do acidente de barco noticiado foi a superlotação. O grupo 48 apresentou uma nova palavra: a *biopirataria* (juntamente com as palavras *holandeses, plantas, floresta, promotor, ...*), que não ocorreu no algoritmo “best-star”. A biopirataria é um problema da região, que ocorre quando estrangeiros (no caso, *holandeses*) roubam plantas, animais ou até mesmo conhecimentos da cultura local para industrializar, patentear e vender no mercado internacional.

Em 10%, o grupo 3 mostrou a necessidade de cautela na solução do problema do vocabulário. *Queimadas* e *incêndios*, embora pareçam sinônimos, podem ter significados bastante diferentes. A palavra *acidentais* é um adjetivo que substitui a locução adjetiva *por acidente* e qualifica o substantivo comum incêndio. O texto se refere a *incêndios acidentais*, que ocorrem devido a condições climáticas (a seca). Já as *queimadas* são decorrentes da atividade agropecuária, comum na região.

Em 15%, o algoritmo “cliques” retornou o mesmo resultado do algoritmo “best-star”.

3.3.2 Teste com as publicações do período de nove meses

A segunda etapa do experimento submeteu ao programa uma coleção de 114 documentos, correspondente a nove meses do ano de 1999. Os mesmos problemas mostrados no tópico anterior se repetiram.

Os graus de similaridade 0 e 2,5% do algoritmo “best-star” apresentaram agrupamentos comuns. Dois grupos apresentaram maior representatividade em 5,26%, já que não houve documentos isolados. O grupo 25 tem relação com o tráfico de armas e a guerrilha na Colômbia (FARC), e o grupo 27, cujas palavras mais relevantes foram *saúde* e *índios*.

O grupo 23, em 0, 2,5 e 5%, apresentaram uma combinação curiosa de palavras: *ambiental*, *ambiente*, *sustentável*, *gestão*, *futuro*, *modernidade*, *modernização*, *contradição*, *interesses*, *projeto*, *progresso*. Esta combinação pode remeter à necessidade de uma gestão ambiental de forma sustentável, que esbarra em interesses e apresenta contradições.

Em 10%, o número de grupos caiu para 15 e, em 15%, caiu para 5 com a mesma configuração do experimento anterior. Neste último, novamente, ocorreram como assuntos principais: *desmatamento*, *queimadas* e acidente de *barco*.

O algoritmo “cliques”, em 0%, também apresentou a palavra *Amazônia* como a de maior ocorrência em 84,21% dos documentos e, em seguida, *desmatamento* (16,6% dos documentos). Já, em 2,5%, o grupo 8 teve maior representatividade (9,65%), tratando da crise colombiana. Em 5%, foram compostos 34 grupos, em 10%, 15 grupos. Em 15%, os 5 grupos apresentaram a mesma configuração do experimento anterior, apesar do número de documentos testado ser diferente.

3.3.3 Teste com as publicações do 1º semestre

Nesta etapa, são relatados os resultados do processamento da coleção de textos relativa ao primeiro semestre. Foram submetidos 65 documentos. Para os graus de similaridade 0, 2,5 e 5%, o algoritmo "best-star" apresentou 18 grupos e nenhum documento ficou isolado. O grupo 3 foi o mais representativo com 10,77% dos documentos, cujo tema foi *estudo* sobre o *desmatamento*. Os grupos 8 e 10 apresentaram o mesmo tema acidente de *barco*, descrito através de palavras diferentes, levando mais uma vez a considerar a presença do problema do vocabulário. As palavras

do grupo 16 estavam relacionadas com o *tráfico* de *armas*. O grupo 17 apontou a palavra *interdição* juntamente com *FUNAI* (Fundação Nacional de Assistência ao Índio), fazendo supor que, em 1999, a FUNAI sofreu ameaça de interdição.

Em 10%, o número de grupos diminuiu para 16 e 17 documentos ficaram isolados. No grupo 1, o tema *queimadas* prevaleceu com eliminação das palavras *intoxicação* e *mercúrio*. Novamente, registrou-se a presença de dois grupos com o mesmo contexto: 7 e 9. Já no grau de similaridade 15%, ocorreram apenas 7 grupos e 48 documentos ficaram isolados. A palavra *desmatamento* se destacou nos grupos 1, 4 e 7. Dois grupos abordaram acidente de *barco* e os outros dois, *queimadas* e *incêndios*.

O algoritmo "cliques", em 0%, alocou todos os documentos em um único grupo. As palavras com maior grau de importância ou relevância foram *Amazônia*, que ocorreu em 89,23% dos documentos com 0,013 de relevância, e o *desmatamento*, em 23,07% dos documentos com 0,004 de relevância. Em 2,5%, foram formados 5 grupos. O grupo mais representativo foi o grupo 3 com 33,85% do documentos, cujo tema foi *desmatamento*.

Em 5%, foram gerados 9 grupos. O grupo 3, abordando o problema do *desmatamento*, foi o mais representativo, com 38,46% do total de documentos, já que não houve documentos isolados. O grupo 6 apontou o tema *narcotráfico*. Do grupo 2, destacaram-se os nomes de bancos estrangeiros: *BID* (Banco Internacional de Desenvolvimento) e *Interamericano*. Os países em desenvolvimento precisam de empréstimos destas instituições para desenvolver regiões como a Amazônia. Em 10% de similaridade, foram criados 16 grupos. Desta vez, 34 documentos ficaram isolados.

Em 15%, os documentos foram alocados em apenas 8 grupos, 48 documentos ficaram isolados. Os resultados deste grau de similaridade diferiram dos outros, porque as palavras relevantes convergiram para um único tema: grupo 1 (devastação da Floresta Amazônica), 2 (autorização para o desmatamento), 3 (incêndios acidentais), 4 (acidente de barco), 5 (desmatamento), 6 (naufrágio), 7 (proibição do desmatamento) e 8 (queimadas).

3.3.4 Teste com as publicações do 1º trimestre

O algoritmo “best-star” alocou 32 documentos em 11 grupos para os graus de similaridade 0, 2,5 e 5%. O grupo 4 foi o de maior representatividade (18,75%) com os assuntos: *intoxicação* por *mercúrio*, *desmatamento* e *dengue*.

Em 10%, foram formados 9 grupos e 7 documentos ficaram isolados. Em 15%, apenas 12 documentos foram alocados em 5 grupos. Os grupos 1 e 3 abordaram acidente de barco, os grupos 2 e 5, *desmatamento*, e o grupo 4, *queimadas*.

O algoritmo “cliques”, em 0%, gerou apenas um grupo. As palavras *Amazônia* (84,37% dos documentos) e *desmatamento* (28,12% dos documentos) apresentaram as maiores ocorrências e graus de relevância. Em 2,5%, os 32 documentos foram alocados em 2 grupos. O de maior representatividade (75%) teve como palavra mais relevante *desmatamento*. O de menor representatividade (25%) também estava relacionado ao *desmatamento*.

Em 5%, 6 grupos foram gerados. O grupo 3 apresentou maior número de documentos (31,25%) com as palavras *dengue, armas, Colômbia, LBA*. Em 10%, 9 grupos foram formados e 9 documentos ficaram isolados. Em 15%, 5 grupos foram formados e 21 documentos ficaram isolados.

3.3.5 Teste com as publicações do mês de janeiro

Para o mês de ***janeiro***, o algoritmo "best-star", configurado em graus de similaridade iguais a 0 e 2,5%, alocou os 9 documentos em 3 grupos. O grupo 1 apresentou, em ordem decrescente de relevância, as palavras: *ambiental, terras, indígenas, futuro, gestão, índios*. Sua representatividade foi de 22,22%. O grupo 2 teve como palavras relevantes: *Amazônia, armas, dengue, Brasil, Colômbia, infra-estrutura, FARC*. Recebeu 56,55% do total de documentos. O grupo 3 apresentou as seguintes palavras relevantes: *LBA*, *Amazônia, avião, ER-2* (avião utilizado no LBA); com 22,22% de representatividade.

Quando ocorreu um aumento da similaridade para 5%, os grupos gerados foram quase os mesmos, porém o documento 178 ficou isolado. Alterando o grau de similaridade para 10%, apenas permaneceu o grupo 3, os outros foram eliminados e seus documentos ficaram isolados.

O algoritmo "cliques", com graus de similaridade 0 e 2,5%, não gerou agrupamento. As palavras mais relevantes foram *Amazônia*, *armas*, *Brasil*, *dengue.* Em 5%, foram gerados 2 grupos e, em 10%, apenas um grupo: o grupo 3, com a mesma alocação do algoritmo "best-star". Isto quer dizer que os dois documentos do grupo 3 foram os únicos a apresentarem 15% de similaridade.

3.3.6 Conhecimento descoberto com o Eurekha

Os resultados do Eurekha permitiram conhecer o que aconteceu no ano de 1999 na Amazônia e entender um pouco a postura do jornal *Folha de São Paulo* em relação à região.

A presença das palavras *Estados Unidos* e *EUA*, tanto em grupos que abordavam *pesquisa* quanto a crise colombiana e o *narcotráfico*, demonstram a grande influência que este país tem na América Latina. Na época, a grande preocupação dos Estados Unidos era com as fronteiras (palavra constantemente mencionada): impedir que o *nacrotráfico* se espalhasse através de uma *intervenção militar* na região. Na realidade, a verdadeira intenção dos Estados Unidos era manter uma base militar na Amazônia e assim controlar suas riquezas. Por outro lado, isto poderia servir de pretexto para novos golpes militares nos países da América Latina. Esta intervenção não aconteceu.

Quanto aos acidentes de *barco*, ficou clara a sua causa: a *superlotação*. A palavra aparece logo no primeiro teste. A região possui locais difíceis de serem fiscalizados e a navegação é um importante meio de transporte.

O jornal também registrou a presença de biopiratas na região no ano de 1999. A *biopirataria* é um crime típico na Amazônia, em que não só as matérias-primas são roubadas por estrangeiros, mas o conhecimento dos povos da floresta.

Notícias a respeito das *queimadas* e do *desmatamento* estavam sempre relacionados com a divulgação de *pesquisa* de institutos como *INPE* ou publicações de revistas científicas internacionais como a *Nature* ou com a atuação do *IBAMA*. Assim, publicações desta natureza foram sempre apoiadas em estudos ou fatos ocorridos, propiciando grande credibilidade às matérias. No caso do desmatamento, o ministério do meio ambiente, assim como o seu titular, José Sarney Filho, estavam quase sempre presentes como palavras relevantes. Por outro lado, registrou-se a existência de empresas certificadoras, que concedem o selo de certificação a empresas que exploram a floresta de forma sustentável. Além disso, a atuação de ONGs como IMAZON, IPAM, “Greenpeace”, órgãos de pesquisa como EMBRAPA, INPE e de proteção ao índio (FUNAI) na região.

Pode-se notar que o jornal não abordou apenas os problemas da região, mas também a cultura, a geografia, a fauna e as características locais. A palavra Amazônia apareceu relacionada inclusive à *exposição* no *Museu Britânico*.

No ano de 1999, o *clima* da Amazônia foi tema de matérias jornalística, quando abordava, por exemplo, o experimento *LBA*, cuja finalidade era saber o que aconteceria em caso de *seca* na região. Além disso, seminários a respeito da *biodiversidade* também foram registrados.

Já as palavras *índio* e *índios* (ianomâmis e ticunas) sempre ocorreram relacionadas às palavras: *malária*, *cegueira*, *tracoma*, *suicídios*, *terras*; o que demonstrou que os índios enfrentaram problemas sérios de saúde na região em 1999.

O gráfico da figura 3.4 apresenta uma comparação da ocorrência dos temas mais abordados para cada mês do ano de 1999, que foi processado separadamente. A palavra de maior freqüência relativa no jornal *Folha de São Paulo*, logo após *Amazônia*, foi o ***desmatamento***, segundo os resultados do algoritmo "cliques" do Eurekha.

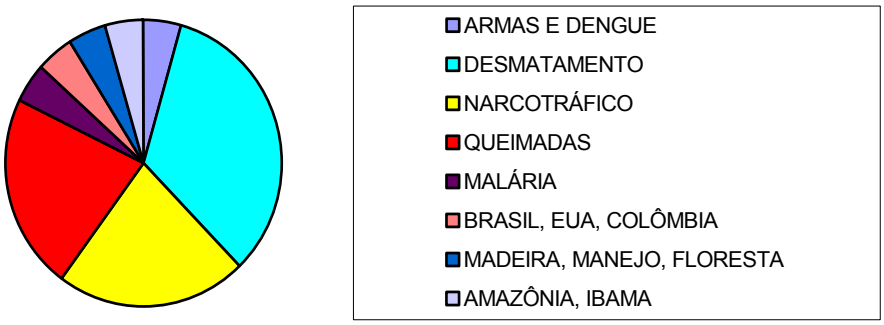


FIGURA 3.4 - Temas mais abordados pelo jornal Folha de São Paulo no ano de 1999

3.4 Experiência com Umap

Os documentos em formato ASCII foram convertidos para o formato HTML através do Microsoft Word do Microsoft Office 97 ®. Os documentos tiveram de ser publicados em um servidor *Web* com acesso através de senha, já que os textos pertencem ao jornal *Folha de São Paulo* e a versão disponível para o processamento é o Umap Web.

O Umap permite salvar cada teste como um arquivo do tipo "Trivium Toolkit Document", com extensão .IMP. Cada conjunto de documentos testado no Eurekha foi submetido também ao Umap.

O objetivo do Umap é mostrar um ponto inicial para explorar um conjunto de documentos. Ele gera ilhas através das quais se podem obter grupos e fazer associações entre palavras, verificando se estas associações ocorrem nos documentos.

Inicialmente, os documentos foram submetidos às seguintes configurações:

a) o número máximo de palavras para construção do mapa foi 100;

b) o conhecimento do usuário sobre a coleção foi considerado baixo ("novice"), a fim de que se pudesse obter uma visão geral dos temas relativos à coleção;

c) Com relação aos idiomas foram selecionadas as seguintes opções: eliminar numerais, filtrar *stopwords*, suporte a substantivos compostos, identificar o idioma, considerar somente substantivos.

Na segunda etapa, o usuário foi selecionado como “expert”. As outras configurações não sofreram modificação. Os próximos tópicos descrevem os experimentos.

3.4.1 Teste com a coleção inteira

Após o Umap construir o primeiro mapa dinâmico, surgiram na janela “Keywords” palavras negativas, que tiveram de ser eliminadas manualmente. Com a eliminação, um novo processamento automático foi disparado. As 100 palavras mais freqüentes e, portanto, mais relevantes são mostradas na figura 3.5.

Amazônia, região, Brasil, brasileiro, Estado, área, Meio, governo, casos, Estados, floresta, pesquisa, problema, Ambiente, mundo, presidente, controle, madeira, empresas, Rio, Recursos, Organização, produção, internacional, Amazonas, ambiental, terra, ministro, Ministério, público, EUA, política, trabalho, água, Colômbia, droga, Filho, fronteira, índios, Polícia, queimadas, Desenvolvimento, narcotráfico, acre, desmatamento, indígenas, Unidos, relação, dinheiro, Ibama, militar, Encontro, questão, avião, sustentável, cocaína, colombiano, comunidade, Roraima, Saúde, exploração, barco, extração, guerrilheiros, crise, cultura, FHC, seca, aldeia, deputado, fogo, mostra, Rio de Janeiro, CPI, clima, crianças, ONGS, seminário, certificação, Imazon, piloto, qualquer, exército, malária, médico, Barry Mccaffrey, Fumaça, Gestão, sigilo, Curica, educação, FNS, pescadores, Suriname, exposição, soja, atendimento, ticunas, Guaraná, suicídios

FIGURA 3.5 - Palavras relevantes indicadas pelo Umap com usuário “novice” para a coleção inteira

Houve uma divergência entre as palavras mais relevantes apresentadas pelo Umap e o Eurekha com o algoritmo “cliques” em 0%. Para o Umap as palavras: *pesquisa*, *madeira*, *Colômbia, índios, queimadas* e *narcotráfico* vieram antes do *desmatamento*. Entretanto, a palavra de maior relevância permaneceu *Amazônia.* Os cálculos de relevância do Umap são diferentes do Eurekha.

A presença de palavras como “Rio de Janeiro” e “Barry Mccafrey” pode ser justificada pelo fato da opção *suporte a “substantivos compostos”* estar selecionada. Este não é o caso das palavras “meio” e “ambiente”, que embora tenham sentido juntas são apenas substantivos comuns, que, morfologicamente, não são consideradas um “conjunto unitário”. A palavra *qualquer*, embora sendo palavra negativa, foi mantida porque não é possível controlar totalmente as 100 palavras que aparecem na janela “Keywords”.

Em 0%, todas as palavras apareceram ligadas a um continente. Portanto, não foi possível analisar o mapa. Em 20%, dezesseis (16) “subilhas” (formadas por uma só palavra) ocorreram: *suicídios, ticunas, soja, Barry Mccafrey, malária, direito, FHC, avião, índios, produção, barco, certificação, exército, fumaça, exposição, Suriname, guaraná*. E apenas uma ilha: *CPI* e *Curica* (base), *deputado*, *sigilo*. Em 40%, surgiram duas (2) ilhas:

a) *CPI* (base), *deputado*, *sigilo*;

b) *desmatamento* (base), *IBAMA* (ligado a fogo).

Em 60%, dezenove (19) “subilhas” apareceram, mas a palavra *Amazônia* prevaleceu em relação às outras, isto é, a maior parte das palavras-chave estavam ligadas à palavra *Amazônia*. Foi necessário aumentar ainda mais o cursor de diferenciação. Em 80%, ocorreram poucas “subilhas”.

O ambiente dinâmico do Umap permite formular consultas a partir da seleção de palavras do mapa mesmo que elas não estejam em uma mesma ilha ou continente. Na tabela 3.2, as linhas de 1 a 4 mostram as ilhas formadas pelo Umap; enquanto as outras resultam de consultas formuladas, selecionando palavras-chave diretamente no mapa, e

criando associações entre estas. Por exemplo, na linha 5, os documentos apresentados são aqueles que ao mencionar a palavra *madeira* também mencionam a palavra *certificação.* O selo de certificação é dado à empresa que explora a floresta de forma sustentável. Os grupos aparecem na tabela 3.2.

TABELA 3.2 - Grupos originados a partir de ilhas e consultas do Umap para a coleção inteira

Grupos	Palavras mais relevantes	Documentos
1.	Brasil (base), governo, narcotráfico, Colômbia	Jul128, nov44, ago117, set26, ago55, out48, mar62, ago119, nov84, ago30, ago113, ago49, nov78, nov9
2.	Desmatamento (base), IBAMA	Jun145, fev170, fev171, dez109, jul127, mar37, fev177, abr157, mar163
3.	CPI (base), deputado, sigilo	Nov89, nov91, nov41, nov85
4.	IBAMA, fogo	Jul127, set27, out17
5	Madeira, certificação	Out14, out15, out92, out100, out17
6.	Clima, seca	Abr59, ago53, set104, ago29, dez68, jun141, jan183, set22, jul129, nov83, fev169, mai33, fev166, abr36, dez76, out16, dez5, ago49, nov77, out18
7	Queimadas, fumaça	Set105, ago53, set106, set107

Um cursor de hierarquização em 50% apresentou a palavra *desmatamento* no mais alto nível. Neste aspecto, Umap confirmou os resultados do Eurekha.

Alterando o nível de conhecimento do assunto para *expert*, as palavras mais relevantes tendem a indicar contextos mais específicos e menos usuais. A palavra *Amazônia*, por exemplo, que ocorria no ponto focal, foi eliminada. Além disso, surgiram ilhas formadas com as palavras *privatização* e *Eletronorte*, que pertencem a poucos documentos, assim como as palavras *biopirataria*, *Fernandinho Beira-mar, grilagem* e *infanticídio* destacadas.

Organização, aldeia, CPI, piloto, malária, Barry Mccaffrey, Fumaça, sigilo, Curica, FNS, pescadores, quebra, Suriname, democracia, exposição, Greenpeace, Modernidade, Sâmia Haddock, sementes, soja, atendimento, capoeira, cartel, condenado, holandeses, interdita, juiz, Maués, mediador, Museu, pacientes, privatização, professora, Sustentabilidade, teatro, ticunas, Tuma, urbana, várzea, África, biopirataria, Campinas, Darly, dengue, Eletronorte, etnia, ex-ditador, gene, Guaraná, lagos, lenda, mandante, modernização, Móveis, pasta-base, pastor, Sesc, site, suicídios, Antologia, aulas, Aviadora, baterias, BC, **Beira-Mar**, Bertazzo, Bosque, boto, Bouterse, Braztoa, carvão, Carvoeiros ,cegueira, celulares, cidadãos-dançantes, dispersão, Eletrobrás, **Fernandinho**, Fiocruz, Fnac, grilagem, Guaporé, Incra, Infanticídio, interdição, Ipiranga, Ipram, isenção, Palhares, Parintins, patenteado, Petrobras, piolhos, pólen, rivalidade, saneamento, Suframa, tracoma, triturada, vivax

FIGURA 3.6 - Palavras relevantes indicadas pelo Umap com usuário “expert” para a coleção inteira

Na figura 3.6, nem todos os substantivos compostos foram identificados pelo Umap: *Fernandinho Beira-mar* (traficante) e *Plasmodium Vivax* (espécie de mosquito transmissor da malária).

A tabela 3.3 mostra quais grupos puderam ser gerados a partir da configuração do mapa, cujo ponto focal foi a palavra *organização*:

TABELA 3.3- Grupos originados a partir de ilhas e consultas do Umap para a coleção inteira para o usuário *expert*

Grupos	Palavras mais relevantes	Documentos
1	Malária, vivax, FNS	Fev166, jul133, jul122, jul126, jul125
2	Holandeses, biopitaria, patenteado	Mai150, nov84
3	CPI, sigilo, Palhares, quebra, Fernandinho	Nov89, nov91, nov41, nov85, abr58
4	Piloto, Suriname, ex-diretor, Bourtes	Nov84, nov9, nov42, nov43, ago54
5	Ticunas, aldeia, pastor	Dez5, nov46, dez75
6	Privatização, Eletronorte, Eletrobrás	Mar60, dez65
7	Museu, exposição	Dez73, dez67

Aumentando o cursor de hierarquização para 100%, as palavras *CPI*, *Barry Mccafrey*, *piloto* são as mais representativas para toda a coleção. Estas palavras estão relacionadas com o *narcotráfico*.

3.4.2 Teste com as publicações do período de nove meses

Cento e catorze documentos foram submetidos ao Umap. As palavras mais relevantes aparecem na figura 3.7.

Amazônia, região, brasileiro, Brasil, nacional, área, governo, ano, Meio, anos, Instituto, pesquisa, Ambiente, presidente, mundo, problema, Projeto, casos, controle, Estados, Recursos, floresta, internacional, ministro, política, ambiental, Naturais, queimadas, Rio, acordo, desmatamento, Ministério, produção, público, pesquisadores, planeta, programa, Desenvolvimento, EUA, Filho, madeira, norte, sistema, terra, militar, Unidos, acre, indígenas, índios, mato, Colômbia, Grosso, Serviço, árvores, defesa, revista, avião, madeireiras, Saúde cidade, crise, Encontro, focos, tema, ambientais, Rondônia, Roraima, biodiversidade, Entidade, incêndios, necessidade, verbas, Agricultura, comunidade, FHC, modelo, semanas, autorizações, Sarney, guerrilha, instituição, Sivam, tecnologia, aeroporto, cumprir, Fumaça, Gestão, regime, terceiro, Barry Mccaffrey, FNS, malária, **mapa**, **mapas**, médico, rota, contradição, atendimento, capoeira, soja

FIGURA 3.7 - Palavras relevantes indicadas pelo Umap com usuário "novice" para a coleção do período de nove (9) meses de 1999

A presença das palavras *mapa* e *mapas,* em negrito, no quadro, demonstrou que o processo de lematização não foi feito de forma coerente para com a língua portuguesa. Foram alteradas configurações e eliminadas algumas palavras. A figura 3.8 mostra novas palavras relevantes.

Amazônia, região, brasileiro, Brasil, área, governo, Meio, anos, Instituto, pesquisa, Ambiente, presidente, mundo, problemas, Projeto, caso, controle, Estados, Recursos, floresta, internacional, ministro, política, ambiental, Naturais, queimadas, Rio, acordo, desmatamento, Ministério, produção, pública, pesquisadores, planeta, programa, Desenvolvimento, EUA, Filho, madeira, norte, sistema, terra, militar, suspeita, Unidos, acre, indígenas, índios, mato, Colômbia, Grosso, Serviço, árvores, defesa, revista, avião, madeireiras, Saúde, Agricultura, cidade, crise, Encontro, focos, tema, ambientais, Rondônia, Roraima, biodiversidade, Entidade, incêndios, necessidade, verbas, comunidade, FHC, modelo, semanas, autorizações, Sarney, guerrilha, instituição, mapa, Sivam, tecnologia, aeroporto, Cobrat, cumprir, Fumaça, Gestão, nossas, regime, terceiro, Barry Mccaffrey, FNS, malária, médico, rota, contradições, atendimento, capoeira, soja

FIGURA 3.8 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do período de nove (9) meses de 1999

O cursor de diferenciação, em 20 e 40%, gerou apenas "subilhas". Em 85%, foi apresentada uma configuração, com muitas "subilhas" e apenas 4 ilhas:

a) "subilhas": *malária, soja, FHC, focos, mapa, avião, fumaça, Barry Mccafrey*

b) ilhas: (1) *índios*, *médico*, (2) *desmatamento*, *autorizações*, (3) *Brasil*, *governo*, *Colômbia*, *militar*, (4) *queimadas* e *incêndios*

Os grupos da tabela 3.4 foram formados por ilhas e também pela associação de palavras-chave através de consultas.

TABELA 3.4 - Grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes aos primeiros 9 meses de 1999

Grupos	Palavras mais relevantes	Documentos
1	Brasil, governo, Colômbia, militar, Barry Mccafrey, avião	Ago49, ago119, ago30, ago116, ago117, ago55, ago114, ago51, ago31, ago113
2	Queimadas, focos, fumaça	Jan183, set106, set107, set105
3	Incêndios, queimadas	Jun32, ago28, abr36
4	Queimadas, focos, incêndios	Ago53, set27, fev38, abr157, fev177
5	Desmatamento, autorizações	Fev171, fev170,fev175, fev169
6	Desmatamento, soja	0
7	Índios, médico, malária	Jul125, jul126
8	Índios, médico	Jul124, Jul125, jul126

Não houve documentos para representar a associação soja e desmatamento (linha 6 da tabela), embora o cultivo da soja seja considerado nocivo à região Amazônica. As linhas 7 e 8 demonstraram que nem sempre houve associação entre *médico, índios* e *malária*, podendo existir outros problemas de saúde que estejam relacionados a índios. Em um cursor de hierarquização de 62%, os grupos 1 e 8 apareceram no mesmo nível.

Elevando o conhecimento do usuário para *expert*, obteve-se uma configuração com uma ilha central cuja base era a palavra *índios* associada às palavras *interdição*, *médico*, *cegueira*. Estas associações confirmaram que, em 1999, os índios da região amazônica sofreram com problemas diversos de saúde. Nenhum valor do cursor de diferenciação gerou grupos para o usuário "expert".

3.4.3 Teste com as publicações do 1º semestre

As palavras mais relevantes dos 65 documentos do 1º semestre de 1999 aparecem na figura 3.9.

Amazônia, Brasil, brasileiro, estado, área, pesquisa, governo, Instituto, Meio, Ambiente, Projeto, Inpe, país, caso, Rio, desmatamento, floresta, Espaciais, ministro, Recursos, ambiental, madeira, queimadas, pesquisadores, produção, controle, Filho, Pará, revista, satélite, setor, acordo, Agricultura, Desenvolvimento, Ibama, índios, madeireiras, Ministério, sistema, terra, dinheiro, indígenas, norte, Polícia, cultura, decisão, Encontro, futuro, Grosso, incêndios, modelo, países, preservação, programa, Roraima, sustentável, econômico, Entidade, experiências, extração, investimentos, política, reservas, Rondônia, americana, direito, Fundo, prevenção, semanas, solo, aéreo, aldeia, autorizações, barco, desaparecidas, Farc, fazenda, FHC, focos, Gestão, Justiça, LBA, necessidade, propriedades, rota, Sarney, Sivam, verbas, capoeira, crime, funcionários, Greenpeace, pasta, Procuradoria, República, tecnologia, Colômbia, interdita, teatro, remessa

FIGURA 3.9 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do 1o semestre de 1999

Em 20%, palavras de um único documento dominaram o mapa, por exemplo, a respeito da privatização da Eletronorte. Estas palavras foram eliminadas. Um cursor de diferenciação em 85% gerou a seguinte configuração:

a) "Subilhas": *teatro, barco, aéreo, LBA, focos*;

b) Ilhas: (1) *governo*, *verbas*, *prevenção*, (2) *Colômbia*, *FARC*, (3) *desmatamento*, *autorizações*, *ministro*, *Sarney*, *produção*, (4) *FHC*, "*Greenpeace*", *entidade*, (5) *índios*, *política*.

A tabela 3.5 mostra os grupos formados. A linha 4 demonstra que não se escreveram notícias sobre o que governo fez para prevenir o *desmatamento*. A linha 7 da tabela não apresentou documentos. Isto quer dizer que não se discutiu sobre verbas para preservação da Amazônia. A linha 8 demonstra que se discutiu em 1999 sobre investimentos em pesquisa para a região. Um cursor de hierarquização em 100% apontou o *desmatamento* como assunto mais relevante da coleção.

TABELA 3.5 - Os grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes ao 1o semestre de 1999

Grupos	Palavras mais relevantes	Documentos
1	Greenpeace, FHC	Jun147, jun138
2	Índios, política	Mar165, jun142
3	Desmatamento, autorizações, ministro, Sarney	Fev175, fev169, fev177, fev171, fev170, mar161
4	Governo, prevenção, desmatamento	0
5	Pesquisa, queimadas, prevenção	Abr59, fev177
6	Preservação, Amazônia	Jun146, jun135, mar163
7.	Preservação, governo, Amazônia, verbas	0
8.	Investimentos, pesquisa	Abr59, jun142, mar163
9	Amazônia, LBA	Jan181, jan182, jun144, fev38

Elevando o conhecimento do usuário para *expert*, as palavras mais relevantes aparecem na figura 3.10.

Recursos, Desenvolvimento, índios, terra, indígenas, autorizações, fazenda, FHC, focos, Gestão, LBA, rota, Sarney, Sivam, verbas, capoeira, funcionários, Greenpeace, Procuradoria, República, tecnologia, aeronave, afastados, buracos, Colômbia, desflorestada, desflorestamento, etnia, governador, hidrelétricas, instituição, interdita, intervenção, isolados, mort, regime, teatro, BNDES, cegueira, check-up, cidadãos-dançantes, contaminação, contradições, Darly, dengue, Eletrobrás, Eletronorte, embarcação, ermitão, espetáculo, fazendeiro, feira, Fiocruz, Fnac, FNS, força-tarefa, fugitivo, Gonçalves, Goya, Guaporé, guerrilheiros, hepatite, hotel, indigenistas, indústria, infectados, interdição, Ipiranga, isenção, isto, Itacoatiara, Itaituba, juiz, lojas, maconha, malária, mandant, Melo, Mucajaí, nação, oncocercose, palhoça, Parintins, parque, Plástica, Pochmann, privatização, produtividade, remessa, semi-aberto, Sesc, soja, Sorcel, Suframa, Tapajós, tracoma, triturada, Tucuruí, Turismo, vacinas

FIGURA 3.10 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "expert", para a coleção do 1o semestre de 1999

Em qualquer cursor de diferenciação, para o usuário *expert*, as ilhas formadas estavam relacionadas com um único documento.

3.4.4 Teste com as publicações do 1º trimestre

Nesta etapa, o número de documentos testados foi diminuído para 32, correspondentes ao primeiro trimestre de 1999. As palavras mais relevantes aparecem na figura 3.11.

Amazônia, região, nacional, segundo, Área, Instituto, Meio, também, amazônica, anos, governo, Inpe, ontem, passado, Ambiente, brasileira, ministro, nas, Pesquisas, projeto, Rio, caso, desmatamento, outro, Filho, mil, país, recursos, suspendeu, floresta, Manaus, medida, representantes, buscas, econômico, essa, Ibama, índios, ministério, porque, Roraima, trabalho, agricultura, ambiental, aos, chefe, crescimento, estudos, isso, lado, modelo, Rondônia, satélites, Seus, terra, Ana Maria, autorizações, barco, combate, desaparecidas, estimativa, mudança, números, Observação, Polícia, propriedades, próximo, público, queimadas, Sarney, Avião, desenvolvimento, Farc, fazenda, focos, gestão, incêndio, indígenas, LBA, pasta, setor, sustentável, ambientais, Colômbia, etnia, futuro, interdição, investimento, isolados, Justiça, necessidade, Nobre, países, política, pontos, prevenção, rota, sobrevivente, Verba, verde

FIGURA 3.11 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do 1o trimestre de 1999

O cursor de diferenciação teve de ser aumentado de 40% para 85% a fim de que gerasse um maior número de ilhas e menor número de "subilhas". As "subilhas" foram *interdição*, *barco*, *índios* e *focos*. As ilhas são mostradas na tabela 3.6.

TABELA 3.6 - Grupos originados a partir de ilhas e consultas do Umap para os documentos correspondentes ao 1o trimestre de 1999

Grupos	Palavras mais relevantes	Documentos
1	Desmatamento, autorizações	Fev170, fev175, fev171, fev169
2	Desmatamento, ministério, agricultura, propriedades	Mar162, mar161, mar37
3	Colômbia, FARC	Não formam grupo
4	Rio, desaparecidas, Manaus	Fev168, fev167, fev173, fev164, fev172
5	Avião, LBA, caso	Jan182, jan181, jan64
6	Terras, indígenas	Jan40, mar165

A palavra *governo* ocorreu em uma ilha associada às palavras *verbas, setor* e *prevenção*, mas não formaram um grupo. Aumentando o curso de hierarquização para 90%, as palavras que apareceram no nível mais alto foram: *Amazônia*, *rota*, *governo*, *desaparecidas*, *polícia*.

Elevando-se o conhecimento do usuário para *expert* e eliminando palavras negativas, as palavras da figura 3.12 apareceram como mais relevantes:

Amazônia, Meio, ministro, Rio, caso, desmatamento, Filho, recursos, floresta, índios, ministério, Roraima, agricultura, ambiental, modelo, Rondônia, terra, autorizações, barco, propriedades, queimadas, Sarney, Avião, desenvolvimento, Farc, fazenda, focos, gestão, incêndio, indígenas, LBA, pasta, setor, sustentável, afastados, aldeia, ambientais, Colômbia, desflorestada, desflorestamento, doença, etnia, funcionários, futuro, governadores, interdição, investimento, isolados, Justiça, necessidade, política, prevenção, rota, sobrevivente, Verba, verde, agenda, Almir, arrecadação, cegueira, chance, Chaves, científico, contradição, demarcadas, dengue, detectados, discussão, economia, elétrico, Eletrobrás, Eletronorte, embarcação, ermitão, Experimento, FHC, FNS, força-tarefa, fugitivo, Guaporé, guerrilheiros, hidrelétrica, imposto, indigenistas, infectados, infra-estrutura, km², maconha, malária, Marcelo, Mucajaí, nação, oncocercose, privatização, soberania, Sorcel, Suframa, tracoma, Tucuruí, venda

FIGURA 3.12 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "expert", para a coleção do 1º semestre

Em um cursor de diferenciação de 50%, o único grupo obtido foi com os documentos jan181, jan182, que estão relacionados com o projeto LBA. Aumentando o cursor de hierarquia, a palavra *desmatamento* tomou o nível mais alto.

3.4.5 Teste com as publicações do mês de janeiro

A figura 3.13 mostra os resultados para o processamento dos nove (9) documentos do mês de janeiro. As palavras mais relevantes estão relacionadas com dois grupos de documentos da tabela. O primeiro está relacionado com o LBA e o segundo com gestão ambiental. A palavra *queimadas* aparece no mais alto nível de hierarquia. Isto se repete para o usuário "expert".

Avião, LBA, check-up, decola, Amazônia, brazil, nacional, Estado, região, brasileira, internacionais, questão, ambiental, calor, Cientistas, estudado, Farc, Inpe, Interesse, Previsão, reunião, satélites, SP, terra, território, velho, Almir, Ambiente, áreas, armada, científico, Colômbia, confirmada, contradição, demarcadas, dengue, detectados, difícil, dívidas, Doente, efeito, encanta, estariam, EUA, exemplo, Experimento, febre, feita, FHC, floresta, focos, fotossíntese, fronteiras, futuro, gestão, governador, grupo, guerrilheiros, Higuchi, ianomâmis, imagens, indígenas, índios, infra-estrutura, inglês, jogo, junho, líquida, Meio, metabolismo, nação, Nasa, nature, nega, novembro, ONU, Paul Crutzen, período, PF, Polícia, política, população, progresso, provocada, queimadas, Química, reconhecidas, renegociação, reservas, seca, sistema, Sivam, soberania, sociedade, Suspeita, sustentável, Tapias, tarefa, tráfico, verdadeiras

FIGURA 3.13 - Palavras relevantes indicadas pelo Umap, para um novo teste com usuário "novice", para a coleção do mês de janeiro

A tabela 3.7 apresenta os grupos obtidos:

TABELA 3.7 - Grupos formados pelo Umap a partir da coleção do mês de janeiro

Grupos	Palavras mais relevantes	Documentos
1	LBA, experimento	jan181, jan182
2	Gestão, futuro, ambiental	jan40, jan179

No "expert", o número de ilhas é igual ao número de documentos. Isto demonstrou que para o Umap os assuntos da coleção de documentos foram bastante específicos e diversos.

3.4.6 Conhecimento descoberto com o Umap

No Umap, novas palavras foram descobertas *grilagem* e *Fernandinho Beira-mar*. A ocorrência da palavra *grilagem* demonstrou que o termo foi pelo menos uma vez mencionado no noticiário do jornal *Folha de São Paulo. Grilagem* é a apropriação de terras com documentos falsificados, problema que há tempos remotos atinge a região.

Fernandinho Beira-mar (traficante) foi um personagem da CPI do narcotráfico. Daí, o fato das palavras *CPI, deputado* e *narcotráfico* aparecerem associadas. Por outro lado, a expressão *Fernandinho Beira-mar*, embora sendo um substantivo próprio composto, não foi considerado como um "conjunto unitário" como as palavras "Rio de Janeiro", "Barry Mccafrey" e outras.

O *desmatamento* foi o assunto mais abordado pelo jornal, em 1999, de acordo com o cursor de hierarquia do Umap, coincidindo com o algoritmo "cliques" do Eurekha com grau de similaridade igual a 0%.

Outras descobertas foram realizadas através de consultas. Por exemplo, ao se fazer a associação de palavras *governo*, *prevenção* e *desmatamento*, não houve documentos representativos. Pode-se dizer que não houve documentos representativos

da associação *preservação, governo, Amazônia* e *verbas*. Isto quer dizer que nada foi noticiado a respeito de quanto o governo gastou pela preservação da floresta.

O Umap destacou os problemas de saúde dos índios da região. A palavra *índios* apareceu associada às palavras *médico* e *malária.* Também se descobriu que a *malária* não foi o único problema de saúde que afligiu os índios em 1999, mas também a cegueira e outros.

3.5 Análise dos resultados

A presente análise de resultados é feita com base nos seguintes aspectos:

a) tempo;
b) grau de similaridade, número de grupos e número de documentos não agrupados (isolados);
c) valores de medidas de avaliação ("macroaveraging" e "microaveraging").

3.5.1 Tempo de processamento

O *tempo de processamento* é um aspecto de considerável importância na avaliação de SRIs, que utilizam técnicas de Descoberta de Conhecimento em Textos, uma vez que o seu objetivo principal é minimizar o tempo de resposta [FEL 99]. Apesar disso, o tempo de processamento do Umap não foi registrado, uma vez que apresentou valores muito baixos em relação ao do Eurekha. Por outro lado, uma nova versão deste *software* reduz estes tempos. A tabela 3.8 apresenta para cada período o tempo de processamento do Eurekha e o número de documentos processados.

TABELA 3.8 - Número de documentos e o tempo de processamento de cada período pelo Eurekha

Período	Tempo de processamento	Número de documentos processados
Ano de 1999	1 dia 8h 42min	178
9 meses	20h 30min	114
1º semestre	5h 19min 41s	65
1º trimestre	1h 49min 10s	32
Janeiro	5min 54 s	9

Os dados mostram que o tempo é diretamente proporcional ao número de documentos, ou seja, quanto maior o número de documentos maior é o tempo de processamento. É certo que o número de palavras de cada documento também influencia. Enquanto o pior tempo foi de um dia oito horas e quarenta e dois minutos, a especialista levou duas semanas para analisar a coleção inteira.

3.5.2 O grau de similaridade do Eurekha

As figuras 3.14 a 3.18 apresentam os gráficos com as relações entre o *grau de similaridade* e o *número de grupos* e o *grau de similaridade* e o *número de documentos isolados* para os algoritmos "best-star" e "cliques". Nos gráficos 3.14a a 3.18a, o eixo horizontal x é representado pelo grau de similaridade e o eixo vertical y pelo número de grupos. Os gráficos 3.14b a 3.18b têm como eixo horizontal x o grau de similaridade e como eixo vertical y, o número de documentos isolados.

As curvas dos gráficos 3.14a a 3.18a para o algoritmo "best-star", em termos gerais, apresentaram comportamentos semelhantes, apesar da diferença na quantidade

de documentos. No grau de similaridade 0%, o número de grupos alcançado é sempre maior. À medida que o grau de similaridade aumenta o número de grupos se mantém constante, depois, decresce. Os gráficos 3.14b a 3.18b mostram que o número de documentos isolados aumenta com o aumento do grau de similaridade (maior restrição para o agrupamento). Entretanto, nota-se que, inicialmente, o número de documentos isolados é nulo, uma vez que o grau de similaridade 0% significa uma restrição também nula para o agrupamento.

O algoritmo "cliques" inicia sempre com o número de grupos igual a 1 (figuras 3.14a a 3.18a), já o número de documentos isolados inicia em 0. Em seguida, com o aumento do grau de similaridade, o número de grupos aumenta atingindo o valor máximo nos graus 5% ou 10%. Como se vê nos gráficos das figuras 3.14b a 3.18b, o número de documentos isolados aumenta com o aumento do grau de similaridade, como acontece com o algoritmo "best-star".

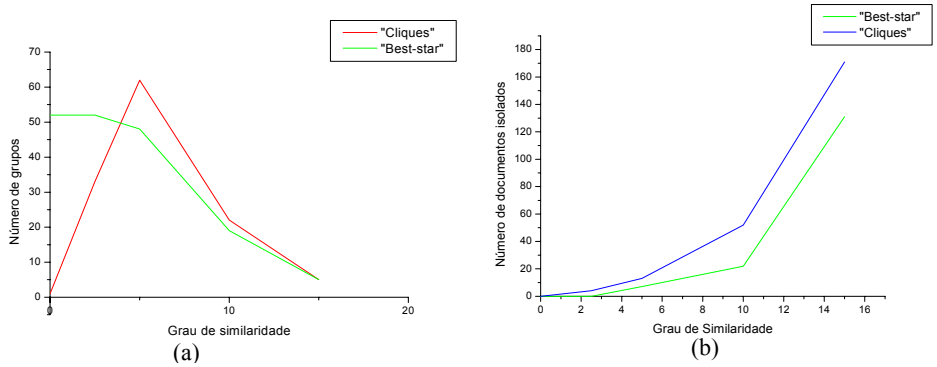


FIGURA 3.14 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para coleção de 178 documentos

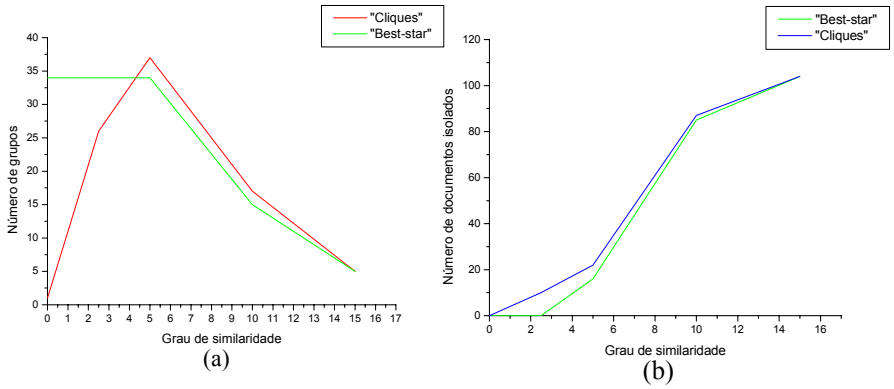


FIGURA 3.15 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 114 documentos

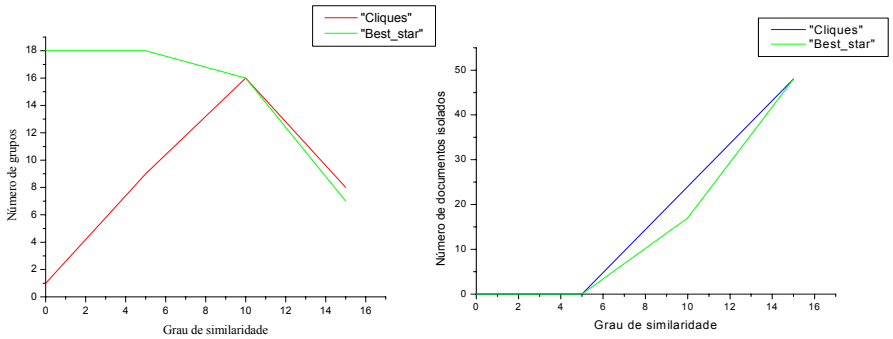


FIGURA 3.16 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 65 documentos

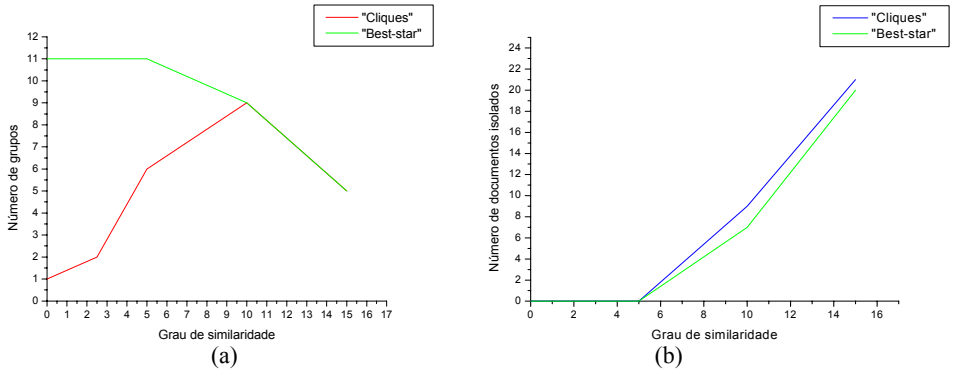


FIGURA 3.17 - Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 32 documentos

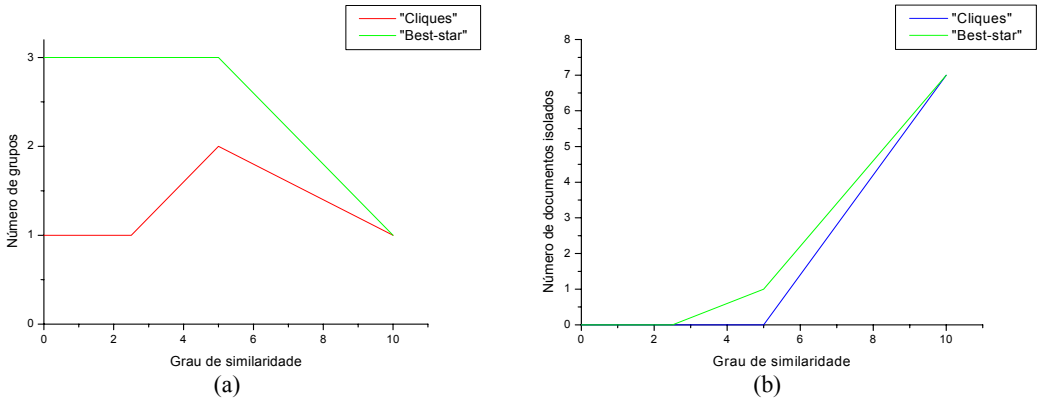


FIGURA 3.18. Relação entre o grau de similaridade e o número de grupos (a) e o grau de similaridade e o número de documentos isolados (b) para a coleção de 9 documentos

Para os graus de similaridade mais baixos (0 e 2,5%), pode-se notar a diferença entre os dois algoritmos. O algoritmo "best-star" tende a selecionar documentos para serem o centro da estrela a ser formada com outros documentos (cada estrela é um grupo), daí o grande número de grupos resultantes. O algoritmo "cliques", entretanto, não seleciona documentos centrais para formar os grupos. Ele aloca os documentos, quando todos são similares entre si. Por isso, para um grau de similaridade nulo (ou baixo), todos os documentos são alocados em um único grupo. Isto se repetiu independente da quantidade de documentos testada como mostram os gráficos das figuras 3.14a a 3.18a.

Para obter uma visão mais detalhada sobre a coleção de documentos, o algoritmo "best-star" é a melhor opção com graus de similaridade 0 e 2,5%, por exemplo (o valor depende do tamanho da coleção). Isto não é vantajoso quando a coleção é muito extensa com muitos assuntos diferentes, pois gera um grande número de grupos, recaindo, assim, no problema da sobrecarga de informação. De outro modo, quando se quer os assuntos mais relevantes, deve-se aumentar os graus de similaridade, em quaisquer dos dois algoritmos.

3.5.3 Efetividade dos agrupamentos

Para avaliar a *efetividade dos agrupamentos*, foram feitos os seguintes cálculos :

a) "Microaverage recall" (MiR) foi obtido dividindo-se o número total de documentos agrupados pelo número total de documentos que deveriam ter sido agrupados;

b) "Microaverage precision" (MiP) foi obtido dividindo-se o número total de documentos atribuídos corretamente pelo número total de documentos efetivamente atribuídos (independente do fato de terem sido atribuídos corretamente ou não) [WIV 99].

c) "Macroaverage recall" (MaR) (abrangência média) foi calculado dividindo a somatória do valor de "Recall" para cada grupo e o número de grupos.

d) "Macroaverage precision" (MaP) (precisão média) foi calculado dividindo a somatória do valor de "Precision" para cada grupo e o número de grupos.

Quando da ocorrência de anomalias aritméticas para quaisquer das fórmulas, o grupo foi eliminado, como sugerido por Lewis [LEW 91]. Por outro lado, quando o número de grupos para um grau de similaridade foi maior que o número de grupos da especialista, foi escolhido o grupo de maior número de documentos equivalentes.

Os resultados apresentados a seguir foram submetidos à apreciação da especialista. Para ela, os valores de "microaveraging" e "macroaveraging" válidos obedecem as seguintes regras:

a) $20\% \leq MiR \leq 100\%$

b) $MiP \geq 25\%$

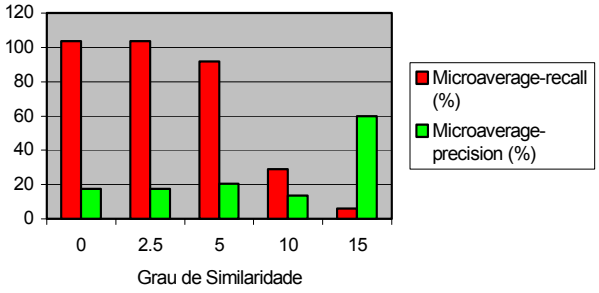
c) $20\% \leq MaR \leq 100\%$

d) $MaP \geq 50\%$

Irvana dos Santos Coutinho justificou estes valores pelo fato de sua metodologia de agrupamento e as metodologias do Eurekha e do Umap serem bastante diferentes.

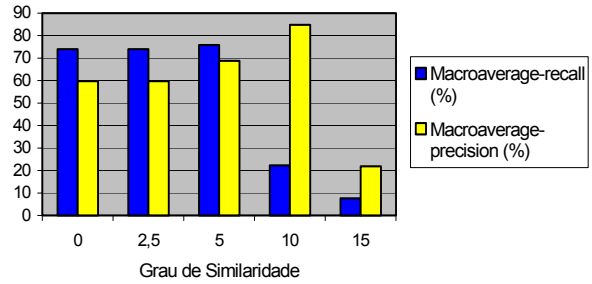
Basicamente, as ferramentas apresentam os resultados a partir de estatísticas sobre as palavras mais freqüentes, enquanto ela realiza a leitura e análise do texto para determinar o tema central e a partir disto selecionar as palavras-chave do texto.

A figura 3.19 apresenta gráficos e tabelas com os valores de "microaveraging" e "macroaveraging" resultantes do processamento da coleção de 178 documentos pelo algoritmo "best-star" do Eurekha.



Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)
0	103,4	17,41
2,5	103,4	17,41
5	91,86	20,46
10	29,06	13,46
15	5,81	60

(a)

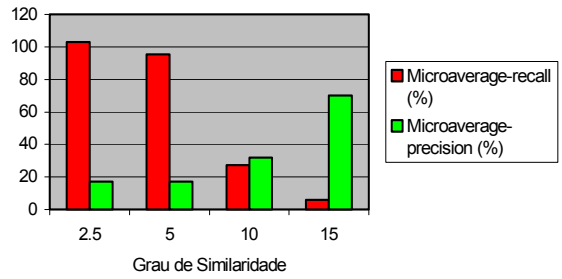


0	73,97	59,77
2,5	73,97	59,77
5	75,81	68,82
10	22,27	84,81
15	7,63	21,87

(b)

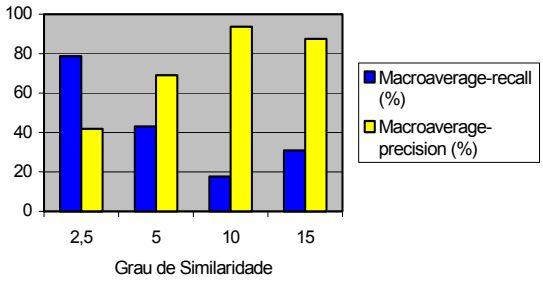
FIGURA 3.19 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 178 documentos processados pelo algoritmo "best-star"

Os valores de "microaveraging" são apresentados na figura 3.19a. O algoritmo "best-star" não apresentou resultados compatíveis com as regras definidas pela especialista para "microaveraging". Os graus de similaridade mais baixos (0% e 2,5%) permitiram alocar todos os documentos em um número de grupos maior (52) do que o número de grupos criados pela especialista (catorze grupos (14) e seis (6) documentos isolados). Já em 5%, foi apresentado o mais alto valor de MiP, porque o número de grupos diminuiu (22) e o número de documentos isolados deixou de ser nulo. (O resutlado, neste caso, aproximou-se um pouco do resultado da especialista). Apesar disso, MiR e MiP não obedeceram os valores definidos. É provável que a grande quantidade de documentos tenha dificultado o agrupamento. Os valores de "macroaveraging" aparecem no gráfico 3.19b. A maior taxa de MaP (84,81%) foi obtida para o menor valor de MaR (22,27%).



Grau de		
0	-	-
2,5	102,9	17,24
5	95,34	17,07
10	27,32	31,91
15	5,81	60

(a)



Grau de similaridade	Macroaverage-recall (%)	Macroaverage-precision (%)
0	-	-
2,5	78,79	41,9
5	43,11	69,04
10	17,70	93,75
15	30,90	87,5

(b)

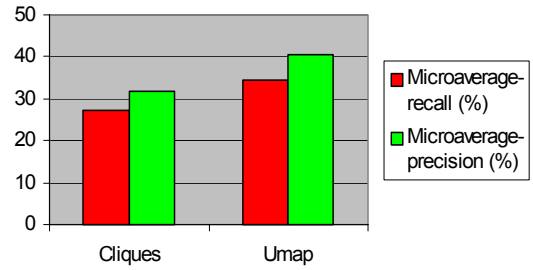
FIGURA 3.20 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 178 documentos processados pelo algoritmo "cliques"

Como mostra a figura 3.14a, o algoritmo "cliques" alocou todos os documentos em um único grupo para o grau de similaridade 0%. Quando isto ocorreu, não foram calculados os valores de "microaveraging" e "macroaveraging" (figura 3.20). Em 2,5% de similaridade, a restrição para o agrupamento também foi pequena, por isso foram agrupados mais documentos do que deveriam ter sido, daí o fato do valor de MiR ultrapassar 100%. Em 10%, o número de grupos se aproximou mais do número de grupos da especialista (14), por isso ambos valores obedeceram as regras definidas acima. O maior MiP foi alcançado por 15%; o maior o grau de similaridade. Os valores de "macroaveraging" da figura 3.20b apresentam o valor de MaP inválido em 2,5% e MaR em 10%, conforme definições da especialista. O maior valor de MaP foi obtido para o menor valor de MaR.

No Umap, os valores do cursor de diferenciação tiveram de ser alterados até que fosse alcançada uma configuração mais fácil de entender o mapa. Foram apresentados agrupamentos apenas para um cursor de diferenciação em 85%. Isto quer dizer que a taxa de erro é grande, as ligações entre as palavras, que formam "ilhas", são menos fortes. Daí, a necessidade de analisar estas ligações e considerar aquelas que têm sentido, utilizando-se de conhecimento prévio ("background knowledge").

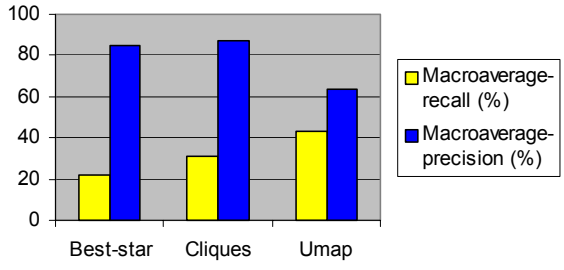
Os gráficos da figura 3.21 apresentam os valores de "microaveraging" e "macroaveraging" para o Umap. Embora o Umap tenha conseguido um valor de MiP maior do que o do algoritmo "cliques", não chegou à metade. Isto quer dizer que poucos contextos secundários (grupos) surgiram mesmo para um alto cursor de diferenciação. Os valores de MaR e MaP para o Umap foram, respectivamente, 42,92% e 63,44%.

No gráfico da figura 3.21a, o melhor valor de MiP foi do Umap. Já o melhor valor de MaP foi do algoritmo "cliques" do Eurekha na figura 3.21b.



	Microaverage-recall (%)	Microaverage-precision (%)
"Cliques"	27,32	31,91
Umap	34,3	40,67

(a)

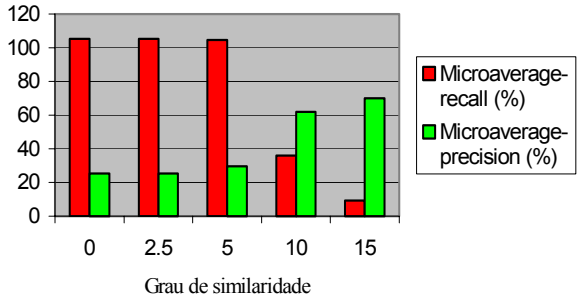


Best-star	22,27	84,81
Cliques	30,9	87,5
Umap	42,92	63,44

(b)

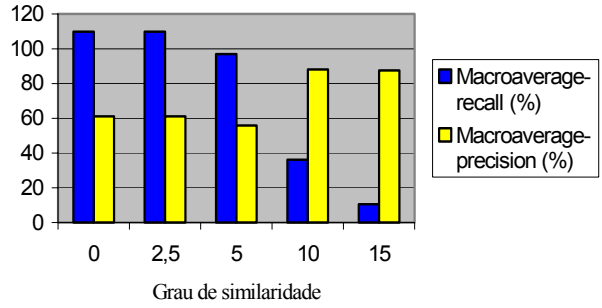
FIGURA 3.21 - Comparação entre os valores de "microaveraging" (a) e "macroaveraging" (b) dos algoritmos "best-star' e "cliques" do Eurekha e, do Umap para a coleção de 178 documentos

Para a coleção de 114 documentos, os resultados do algoritmo "best-star" aparecem na figura 3.22.



Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)
0	105,5	25,43
2,5	105,5	25,43
5	104,6	29,59
10	36,11	62,06
15	9,25	70

(a)

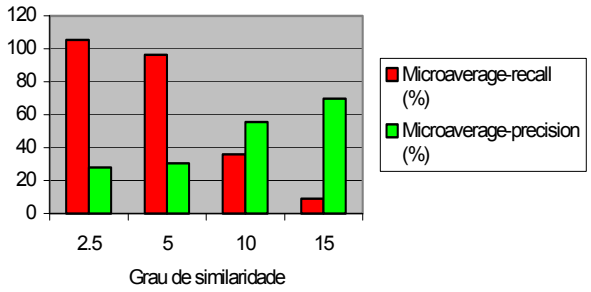


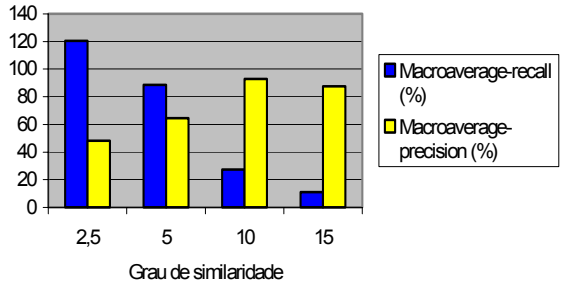
Grau de similaridade	Macroaverage-recall (%)	Macroaverage-precision (%)
0	109,8	61,07
2,5	109,8	61,07
5	96,88	55,95
10	36,13	88,09
15	10,56	87,50

(b)

FIGURA 3.22 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 114 documentos processados pelo algoritmo "best-star"

Desta vez, o algoritmo "best-star" apresentou valores de "microaveraging" válidos (pelas regras acima indicadas) apenas para o grau de similaridade 10%, quando o número de grupos mais se aproxima do número de grupos da especialista e o número de documentos isolados não é nulo. O maior valor de MiP foi obtido para o maior grau de similaridade, em que a restrição de agrupamento é maior. Os valores de MaR e MaP foram validados para os graus de similaridade 5 e 10%. Em 10%, obteve-se a melhor MaP.





Grau de		
0	-	-
2,5	105,5	27,88
5	96,29	30,43
10	36,11	55,55
15	9,25	70,0

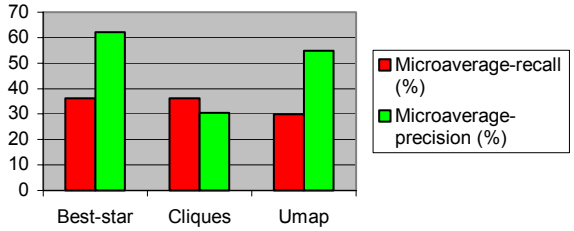
(a)

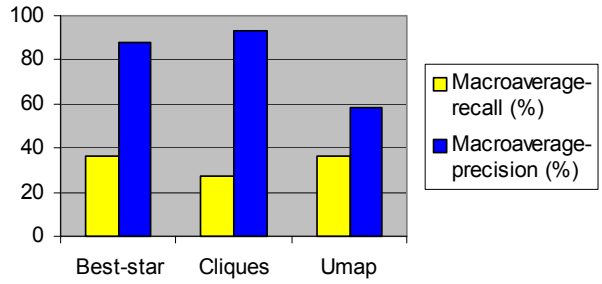
Grau de similaridade	Macroaverage -recall (%)	Macroaverage-precision (%)
0	-	-
2,5	120,4	48,1
5	88,48	64,52
10	27,36	92,85
15	11,02	87,5

(b)

FIGURA 3.23 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 114 documentos processados pelo algoritmo "cliques"

De acordo com as regras definidas pela especialista, o algoritmo "cliques" apresentou o maior valor de MiP (55,55%) no grau de similaridade 10%, como mostra a figura 3.23a, uma vez que foi onde o número de grupos do algoritmo (17) se aproximou mais do número de grupos da especialista (14). O maior valor de MaP foi obtido para o menor valor de MaR com grau de similaridade 10% (figura 3.23b).





	Microaveraging-recall (%)	Microaveraging-precision (%)
Best-star	36,11	62,06
Cliques	36,11	55,55
Umap	29,8	54,83

(a)

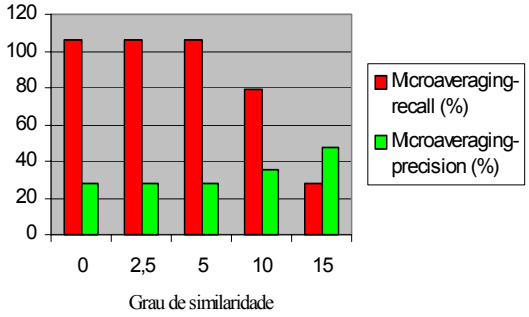
	Macroaverage-recall (%)	Macroaverage-precision (%)
Best-star	36,13	88,09
Cliques	27,36	92,85
Umap	35,99	58,33

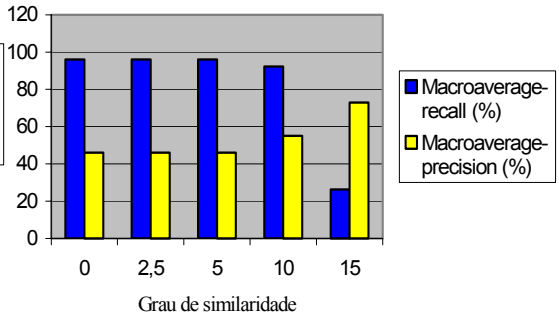
(b)

FIGURA 3.24 - Comparação entre os valores de "microaveraging" (a) e "macroaveraging" (b) dos algoritmos "best-star' e "cliques" do Eurekha, e do Umap para a coleção de 114 documentos

De acordo com a figura 3.24a, a maior precisão global média (MiP) foi apresentada pelo algoritmo "best-star", enquanto o maior valor de MaP foi do algoritmo "cliques" (3.24b). Ou seja, em termos gerais o "best-star" teve maior efetividade ("microaveraging"), mas avaliando o agrupamento realizado para cada grupo, o "cliques" apresentou melhores resultados ("macroaveraging").

Tomando a coleção de sessenta e cinco (65) documentos, referentes ao 1º semestre de 1999, os valores de "microaveraging" e "macroaveraging" para o algoritmo "best-star" são apresentados na figura 3.25.





0	106,5	27,69
2,5	106,5	27,69
5	106,5	27,69
10	78,68	35,41
15	27,86	47,05

(a)

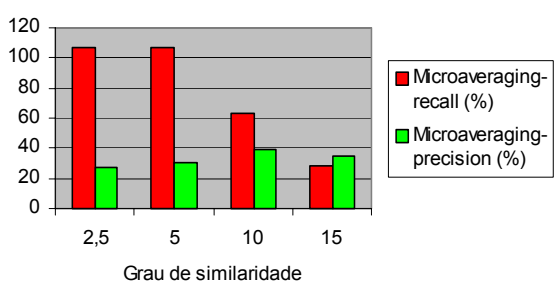
Grau de		
0	96	46
2,5	96	46
5	96	46
10	92,26	55,09
15	26,25	72,91

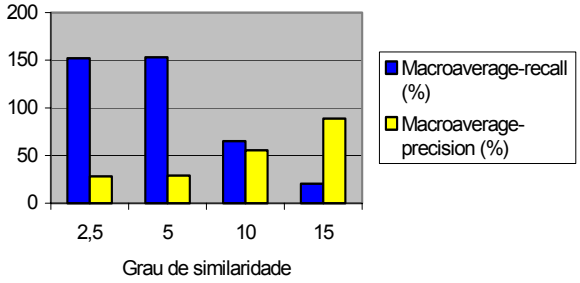
(b)

FIGURA 3.25 - Gráficos para os valores de "microaveraging" (a) e "macroaveraging" (b) dos 65 documentos processados pelo algoritmo "best-star"

O algoritmo "best-star" apresenta valores de "microaverage recall" no intervalo 20% ≤ MiR ≤100% apenas para os graus de similaridade 10 e 15%, quando o agrupamento é dificultado (figura 3.25a). O maior valor de MiP foi encontrado em 15% para um baixo MiR. Também o maior valor de MaP foi encontrado para o menor valor de MaR (figura 3.25a).

A figura 3.26 apresenta os resultados do algoritmo “cliques”. Para os valores 0% e 2,5% de similaridade foram apresentados valores de MiR que excederam os 100%. O maior valor de MiP foi obtido em 10% para o maior MiR. O maior valor de MaP (88,88%) foi obtido quando se agrupou apenas 20,51% dos documentos que deveriam ter sido alocados em cada grupo.





Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)
0	-	-
2,5	106,5	27,69
5	106,5	30,76
10	63,07	39,02
15	27,86	35,29

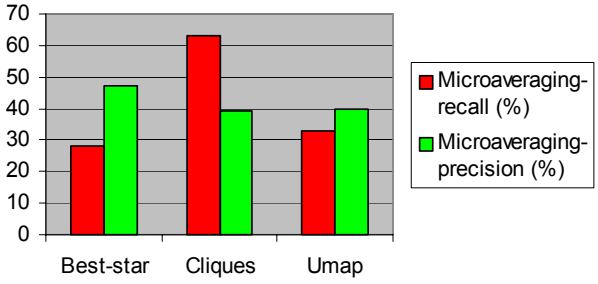
(a)

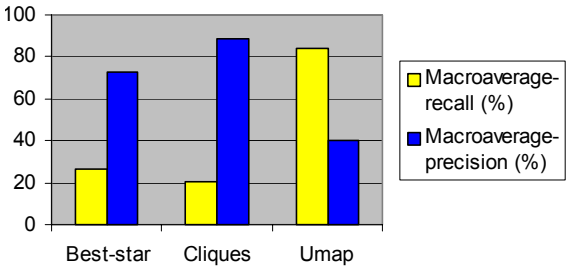
0	-	-
2,5	152	28,2
5	153	29
10	65,11	55,55
15	20,51	88,88

(b)

FIGURA 3.26 - Gráficos para os valores de “microaveraging” (a) e “macroaveraging” (b) dos 65 documentos processados pelo algoritmo “cliques”

A figura 3.27a demonstra que o algoritmo “best-star” apresentou o maior valor de MiP. Já o maior valor para MaP foi apresentado pelo “cliques”.





	Microaverage-recall (%)	Microaverage-precision (%)
Best-star	27,86	47,05
Cliques	63,07	39,02
Umap	32,78	40

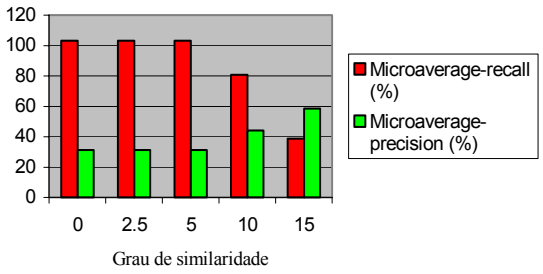
(a)

	Macroaverage-recall (%)	Macroaverage-precision (%)
Best-star	26,25	72,91
Cliques	20,51	88,88
Umap	84,08	40,27

(b)

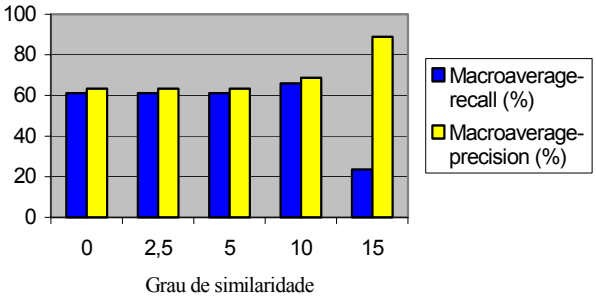
FIGURA 3.27 - Comparação entre os valores de “microaveraging” (a) e “macroaveraging” (b) dos algoritmos “best-star’ e “cliques” do Eurekha, e do Umap para a coleção de 65 documentos

Os valores das medidas de avaliação calculados para o algoritmo “best-star” pelo processamento de trinta e dois (32) documentos do 1º trimestre de 1999 aparecem na figura 3.28.



Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)
0	103,2	31,25
2,5	103,2	31,25
5	103,2	31,25
10	80,64	44
15	38,7	58,33

(a)



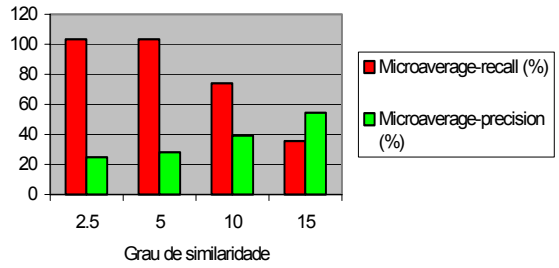
Grau de similaridade	Macroaverage-recall (%)	Macroaverage-precision (%)
0	61,06	63,33
2,5	61,06	63,33
5	61,06	63,33
10	66	68,66
15	23,54	88,88

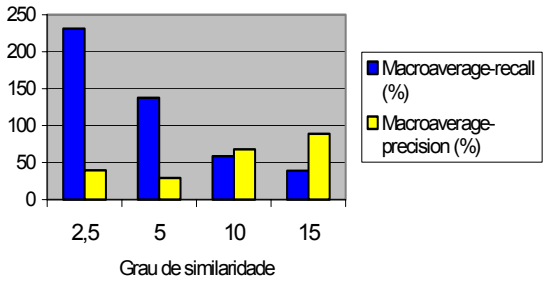
(b)

FIGURA 3.28 - Gráficos para os valores de “microaveraging” (a) e “macroaveraging” (b) dos 32 documentos processados pelo algoritmo “best-star”

Na figura 3.28a, o algoritmo “best-star” apresentou MiR muito alto (MiR>100), tendo em vista que o número de grupos para graus de similaridade baixos tendem a ser grandes. Desta vez, o valor de MiP ultrapassou a metade, porém ainda com um MiR baixo. Todos os valores de “macroaveraging” respeitaram as regras definidas pela especialista. O maior valor de MaP foi alcançado pelo menor valor de MaR (figura 3.28b) em15% de similaridade.

O algoritmo “cliques” tem seus valores mostrados na figura 3.29 Apenas os valores de MiR para os graus de similaridade 10% e 15% respeitaram as regras citadas anteriormente. O maior valor de MiP foi obtido para o menor valor de MiR. Também os valores de MaR obedecem as regras nos graus de similaridade 10% e 15%. Em 15%, obteve-se a maior MaP.





Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)
0	-	-
2,5	103,2	25
5	103,2	28,12
10	74,19	39,13
15	35,48	54,54

(a)

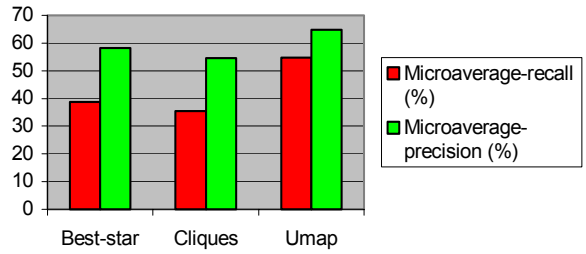
Grau de similaridade	Macroaverage-recall (%)	Macroaverage-precision (%)
0	-	-
2,5	231,2	39,58
5	137,5	29,23
10	58,56	68
15	38,69	88,88

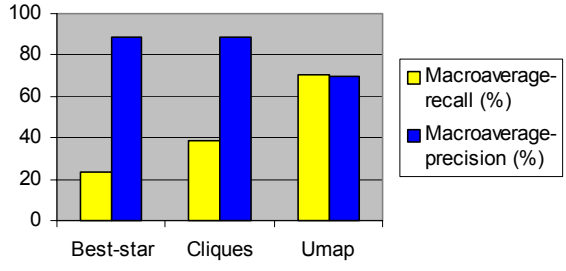
(b)

FIGURA 3.29 - Gráficos para os valores de “microaveraging” (a) e “macroaveraging” (b) dos 32 documentos processados pelo algoritmo “cliques”

Para as duas últimas coleções, foram apresentados valores que obedeceram as regras para dois graus de similaridade: 10% e 15%. Isto ocorreu porque para coleções menores, o agrupamento da especialista foi menos genérico: seis (6) grupos foram criados para a coleção de trinta e dois (32) documentos e onze (11) para a de sessenta e cinco (65); enquanto para a coleção de 178 e 114 documentos foram criados catorze (14) grupos apenas.

A figura 3.30 aponta que os maiores valores de “microaveraging” foram alcançados pelo Umap, que agrupou 68% dos documentos com 64,7% de acerto, quando comparado aos resultados da especialista.





	Microaverage-recall (%)	Microaverage-precision (%)
Best-star	38,7	58,33
Cliques	35,48	54,54
Umap	68	64,70

(a)

	Macroaverage-recall (%)	Macroaverage-precision (%)
Best-star	23,54	88,88
Cliques	38,69	88,88
Umap	70,48	69,99

(b)

FIGURA 3.30 - Comparação entre os valores de “microaveraging” (a) e “macroaveraging” (b) dos algoritmos “best-star’ e “cliques” do Eurekha, e do Umap para a coleção de 32 documentos

A tabela 3.9 apresenta os valores de "microaveraging" e "macroaveraging" para os documentos referentes às publicações de janeiro.

TABELA 3.9 - Valores de "microaverage recall" e "microaverage precision" do algoritmo "best-star" para os nove documentos

Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)	Macroaverage-recall (%)	Macroaverage-precision (%)
0	150	22,22	250	40
2,5	150	22,22	250	40
5	133	25	200	50
10	33,33	0	100	0
15	33,33	0	100	0

O algoritmo "best-star" não alcançou valores, que respeitassem as regras da especialista para "microaveraging" e "macroaveraging".

A tabela 3.10 apresenta os resultados do algoritmo "cliques". Também não foram alcançados valores satisfatórios, segundo as regras definidas acima.

TABELA 3.10 - Valores de "macroaverage recall" e "macroaverage precision" do algoritmo "cliques" para a coleção de nove documentos

Grau de similaridade	Microaverage-recall (%)	Microaverage-precision (%)	Macroaverage-recall (%)	Macroaverage-precision (%)
0	150	0	0	0
2,5	150	0	0	0
5	133	22,22	250	40
10	33,33	0	100	0
15	33,33	0	100	0

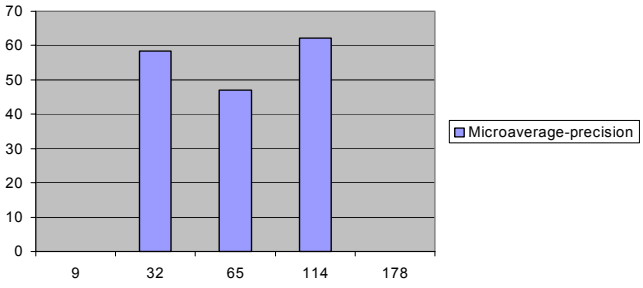
No Umap, para a coleção de nove (9) documentos, foram alocados quatro (4) documentos em dois (2) grupos. Apenas um dos documentos de cada grupo estavam corretos, ou seja, caso se tratasse de classificação o valor de "macroaverage recall" seria igual a 100% e o "macroaverage precision" igual a 50%. Como se trata de agrupamento, ambos não têm semelhança alguma com o que foi definido pela especialista. Assim, os valores MiP e MaP são nulos.

Assim, os piores resultados do Eurekha e do Umap foram encontrados no processamento da coleção de nove (9) documentos. A especialista encontrou três grupos cada um com dois documentos, deixando três (3) documentos isolados. Nenhum destes grupos, entretanto, é igual ao grupo dos documentos jan181 e jan182, que foram criados pelo Umap e foram apontados pelo Eurekha como o de maior grau de similaridade. Isto pode ser explicado pela diferença de metodologia da especialista e das ferramentas. As ferramentas geram os resultados a partir da estatística das palavras-chave repetidas, enquanto a especialista escolhe as palavras-chave através do assunto principal.

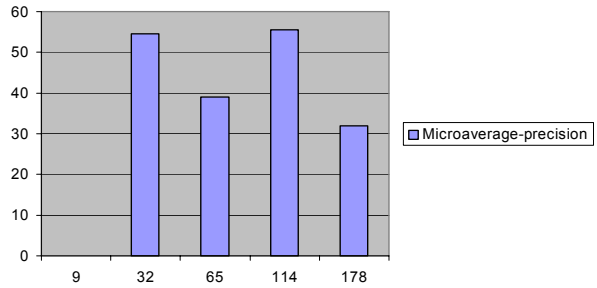
O documento jan181, intitulado "Cientistas iniciam 'check-up' da Amazônia", se atém na definição do LBA[8] e menciona a chegada do ER-2, avião utilizado pela NASA para espionagem, que seria utilizado pelo projeto. Já o documento jan182 intitulado, "Avião [ER-2] só decola com supervisão", aborda especificamente o problema do avião. Realmente, há termos comuns entre os documentos, porém o assunto principal de ambos é diferente; o primeiro aborda o LBA e o segundo, o problema do ER-2, utilizado no LBA. Mas, segundo a especialista, o agrupamento proposto pelos programas é possível, uma vez que o termo LBA pode ser utilizado em uma consulta e o documento jan182 seria uma resposta correta.

[8] LBA: sigla em inglês do Experimento de Grande Escala da Biosfera-Atmosfera na Amazônia

Os gráficos da figura 3.31 mostram para cada coleção os maiores valores de MiP obtidos pelos algoritmos do Eurekha "best-star" (3.31a) e "cliques" (3.31b). Para graus de similaridade 0% e 2,5%, os valores baixos de MiP podem ser explicados pelo grande número de grupos ocorridos. Já para graus de similaridade altos, o número de documentos isolados foi bastante alto como mostram os gráficos das figuras 3.14b a 3.18b. Por outro lado, pode-se afirmar que o agrupamento realizado pelo Eurekha foi menos genérico do que o da especialista, que agrupou, por exemplo, os 178 documentos em apenas catorze (14) grupos. No caso do algoritmo "best-star", percebe-se que para a maior coleção o valor de MiP não obedeceu as regras definidas pela especialista. É possível que o tamanho da mesma tenha dificultado a seleção dos documentos centrais para cada grupo.



(a)



(b)

FIGURA 3.31 - Quantidade de documentos processados pelo algoritmo "best-star" (a) e "cliques" (b) do Eurekha e respectivos valores de "microaverage precision"

No Umap, figura 3.32, ocorreram muitas ligações fortes de palavras-chave com a palavra "Amazônia", que é o *ponto focal*[9] do mapa, dificultando o surgimento de contextos secundários (grupos).

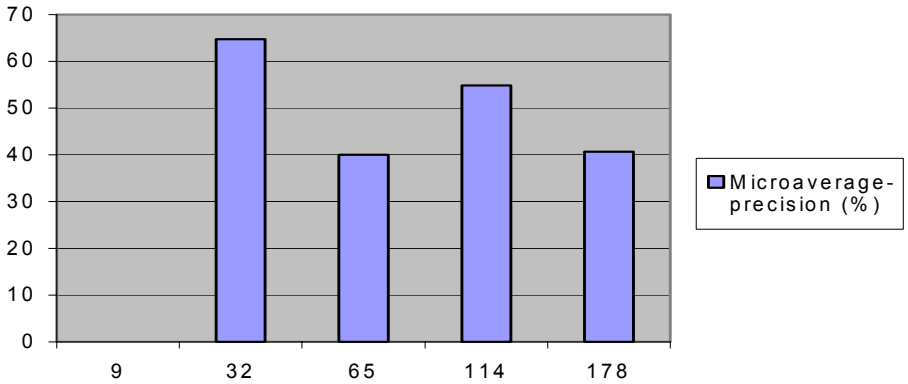


FIGURA 3.32 - Quantidade de documentos processados pelo Umap e respectivos valores de "microaverage precision", tendo a palavra "Amazônia" como ponto focal

[9] Ponto inicial de exploração do mapa, ou seja, a palavra mais freqüente.

Eliminando, a palavra “Amazônia” (figura 3.33), na coleção de 178 documentos, MiR resultou em 22,6% e MiP aumentou para 51,28%, superando os 40,67%. Já para a coleção de 114 documentos, MiP foi de 60,86% e MiR, 21,29%. O valor de MiP não ultrapassou o valor de MiP do algoritmo “best-star”, porém melhorou. Para a coleção de 65 documentos, o valor de MiP ultrapassou o do “best-star”, 52,63%, com um MiR igual a 31,14%. Para a coleção com trinta e dois documentos (32), MiP apresentou o valor de 55,55% com MiR igual a 29,03%. A menor quantidade de documentos permitiu o surgimento de muitas palavras concentradas em um único documento. Assim, os melhores resultados de MiP foram encontrados para o Umap, enquanto os melhores valores de “macroaveraging” foram os do algoritmo “cliques”.

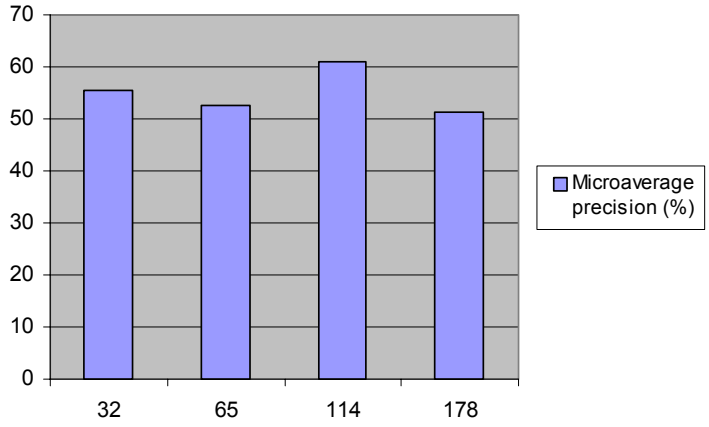


FIGURA 3.33 - Quantidade de documentos processados pelo Umap e respectivos valores de “microaverage precision”, eliminando a palavra “Amazônia” do ponto focal

O próximo tópico apresenta as diferenças entre o Eurekha e o Umap.

3.6 Diferenças entre Eurekha e Umap

A primeira diferença entre o Eurekha e o Umap é que o Eurekha é um *software* acadêmico, enquanto o Umap é um *software* comercial, disponível em quatro versões previamente citadas. Apesar disso, ambos fazem um processamento robusto (não ocorreram “bugs” durante os testes).

O Umap realiza agrupamento de palavras-chave de uma coleção através do qual se podem obter grupos de documentos, utilizando a tecnologia da *Árvore de Conhecimento*. O Eurekha realiza agrupamento de documentos com base na *Hipótese de Agrupamento* (já mencionada anteriormente).

O Umap, como mostra a tabela 3.11, apresentou tempo de processamento menor do que o Eurekha, embora a nova versão do Eurekha tenha melhorado o seu desempenho.

O Umap permite definir o *nível do usuário* como “novice” ou “expert”. Se o nível é “novice”, as palavras-chave exibidas são as mais freqüentes da coleção, caso contrário, são mais específicas. Assim os resultados apresentados levam em consideração o grau de conhecimento do usuário a respeito da coleção. O Eurekha, não.

TABELA 3.11 - Comparação entre as ferramentas Eurekha e Umap

ORDEM	ITENS	UMAP	EUREKHA
1	Utiliza técnica de agrupamento		✔
2	Permite definir nível de usuário	✔	
3	Suporte à exploração e pesquisa	✔	
4	Suporte à aprendizagem	✔	
5	Apoio à decisão	✔	
6	Nível de controle	✔	
7	Meio de comunicação		
8	Permite definir lista de “stopwords”		✔
9	Considera substantivos próprios compostos	✔	
10	Processo de lematização	✔	
11	Apresenta tratamento para o problema do vocabulário		
12	Relevância da saída total (maior valor de “microverage precision”)	✔	
13	Relevância da saída média (maior valor de “macroverage precision”)		✔
14	Menor tempo de processamento	✔	

No Umap, as palavras negativas são eliminadas manualmente. Quando eliminadas, são registradas em um “dicionário”. O Eurekha, por sua vez, permite criar categorias de palavras negativas e as elimina durante o processamento.

O Eurekha ainda não possui um sistema de ajuda embutido no programa. Já o Umap apresenta um tutorial simples para que o usuário possa ter uma noção de como manipular o sistema. São dicas simples fornecidas pelo “gato”.

O Umap oferece *apoio à decisão* como mostra a figura 3.34. A cada mudança do cursor de diferenciação ou hierarquia, o “gato” explica a situação atual do mapa e sugere o que deve ser feito. Ou seja, ele recomenda alterações em ambos os cursores. O Eurekha não apresenta apoio à decisão porque o *nível de controle* do usuário sobre o resultado da mineração é baixo. Este controle só pode ser exercido pelo usuário antes do processamento, por exemplo, definir o grau de similaridade; e após o processamento, escolher entre quatro saídas geradas por cada um dos algoritmos já citados (“best-star, “stars”, “full-star”, “cliques”).

.

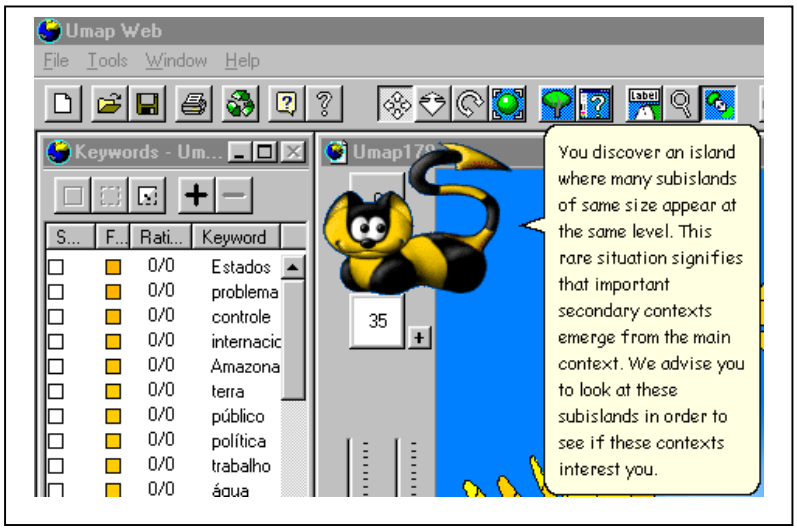


FIGURA 3.34 - O "gato" aponta com a calda para uma recomendação ao usuário no Umap

O Umap e o Eurekha não apresentam solução para o problema do vocabulário. Apesar disto, este problema é mais evidente no Eurekha. O Umap permite por exemplo a formulação de pesquisa sobre as palavras do texto, havendo menor possibilidade de ocorrerem erros semânticos. Assim pode-se dizer que o Umap fornece *suporte* não só *à pesquisa* mas também *à exploração* do resultado da mesma. Os usuários podem formular consultas através das palavras-chave do mapa. Segundo Hansen [HAN 98], os usuários reclamam por serem obrigados pelos sistemas a fazerem novas pesquisas, quando precisam apenas conhecer melhor os resultados já obtidos. O Eurekha processa os documentos e exibe os resultados, que não podem mais ser alterados. Assim, a solução do problema do vocabulário, no Eurekha, e a impossibilidade de identificar *termos compostos*, não só substantivos próprios, mas termos característicos da região como *intoxicação por mercúrio*, *desenvolvimento sustentado*, poderiam ser solucionados se o Eurekha permitisse que o usuário refinasse os resultados ("relevance feedback").

Quanto ao *meio de comunicação*, o Umap e o Eurekha não fornecem meios de receber e nem oferecer colaboração a outros usuários. O Eurekha não está disponível para analisar documentos na Internet, mas possui forma alternativa, ou seja, documentos HTML podem ser submetidos a ele. Por outro lado, o Umap utiliza-se do processo de lematização e, portanto, apresenta suas desvantagens (muito raramente); o Eurekha, não.

Quanto aos valores das medidas de *efetividade*, os melhores resultados de "macroaverage precision" foram obtidos pelo algoritmo "cliques" do Eurekha, uma vez que este é mais rigoroso no agrupamento, já que verifica a similaridade entre todos os documentos de um grupo para fazer a alocação. Os melhores resultados para "microaverage precision" foram obtidos pelo Umap, uma vez que o Umap faz um agrupamento mais genérico analisando as palavras-chave da coleção inteira, não se preocupando com sua organização. Neste caso, o Eurekha permite, para graus de similaridade mais baixos, obter maior conhecimento sobre os temas da coleção, porque ele mostra as palavras-chave de cada grupo. (Para coleções maiores, é preferível trabalhar com altos graus de similaridade.)

Sabe-se, entretanto, que os resultados de ambas medidas de avaliação foram encontrados para baixos valores de "microaveraging recall" e "macroaveraging recall". Isto se deve às diferenças entre a recuperação humana e a recuperação automática como mostra a tabela 3.12.

TABELA 3.12 - Comparação entre a recuperação automática e a recuperação humana

Recuperação humana	Recuperação Automática
Subjetiva	Objetiva
Capacidade para lidar com o contexto	Dificuldade para lidar com o contexto
Incapacidade para lidar com a sobrecarga de informação	Capacidade para lidar com a sobrecarga de informação
Maior tempo consumido	Menor tempo consumido

A recuperação humana é bastante subjetiva. Lewis [LEW 91] afirmou que muitos estudos comprovaram que os próprios profissionais aptos a classificar ou agrupar documentos apresentavam divergências nos resultados finais. Estas divergências aparecem porque cada ser humano apresenta uma carga de conhecimento ("background knowledge") diferente, fazendo a análise do texto a partir de pontos de vista diferentes. Já as ferramentas fazem a análise literal. Neste caso, é interessante que a ferramenta permita o refinamento da pesquisa ("relevance feedback") para melhorar os resultados.

Por outro lado, afirmou-se que a especialista fez um agrupamento genérico dos 178 textos, gerando apenas catorze (14) grupos. A capacidade humana de lidar com a sobrecarga de informação é bastante limitada, por isso, o tempo gasto pelas ferramentas é menor do que para o ser humano.

Assim, não se pode negar a validade de tais ferramentas no auxílio à recuperação de informações, principalmente porque o conhecimento descoberto a partir da análise dos resultados de ambas ferramentas não pode ser descartado. As palavras-chave encontradas em ambas as ferramentas permitiram analisar associações e criar novas, que remeteram aos problemas reais da região. Por exemplo, foi descoberta a causa dos acidentes de barco da região ocorridos em 1999: a superlotação. Soube-se que biopiratas "atacaram" a região naquele ano. Ficaram claros os problemas de saúde dos índios. A malária ainda não foi erradicada. *Clima* e *biodiversidade* foram temas discutidos em seminários. A grilagem de terras também foi tema do noticiário. O desmatamento, por outro lado, foi o tema mais abordado no ano de 1999 e na maior parte dos períodos testados.

4 Conclusão

Este trabalho apresentou uma visão geral das áreas de Recuperação de Informações (RI) e Descoberta de Conhecimento em Textos (KDT). Descreveu a aplicação das etapas do processo de KDT com duas ferramentas: Eurekha e Umap. E, finalmente, avaliou os resultados obtidos, confrontando a recuperação automática e a recuperação humana. Desta forma, foram encontradas conclusões importantes.

O Eurekha descreve os tópicos de cada grupo a partir de palavras-chave selecionadas através de cálculos da freqüência relativa e da relevância no grupo. Como estas palavras aparecem desprendidas do contexto, dificultam o entendimento do usuário quanto à identificação dos substantivos próprios compostos (substantivos próprios compostos devem ser considerados um "conjunto unitário" [NEV 2000]). Por exemplo, as palavras "rio" e "madeira", que se referem a "Rio Madeira", podem ser encaradas como substantivos comuns; madeira pode ser lida como um produto de extração da floresta e não o nome de um rio. Para solucionar o problema, a ferramenta poderia permitir que após o processamento o usuário identificasse estas palavras para formar um "conjunto unitário" e pedir um novo processamento. Isto serviria não apenas para substantivos próprios compostos, mas também para termos cuja junção apresentam significado importante como *desenvolvimento sustentado*, *incêndios acidentais*, *intoxicação* por *mercúrio* e outros.

O Umap, por sua vez, consegue identificar alguns substantivos próprios compostos como "Barry Mccafrey", "Rio de Janeiro", etc., mas não consegue reconhecer outros "Plasmodium Vivax" (espécie de mosquito transmissor da malária), "Fernandinho Beira-mar" (traficante), "José Sarney Filho" (ministro do meio ambiente). Esta deficiência não influencia significativamente os resultados do Umap, já que são mais gerais, apresentando agrupamentos (associações) de palavras e não, propriamente, de documentos.

O problema do vocabulário ocorre quando da presença de grupos diferenciados para o tema "acidentes de barco" na ferramenta Eurekha. Uma proposta de solução para este problema é apontado por [LOH 2000a], quando sugere que se trabalhe com conceitos e não com palavras. Mais uma vez, este problema poderia ser solucionado pelo refinamento do resultado pelo usuário (retro-alimentação por relevância), sugerindo a junção dos grupos semelhantes, por exemplo. O Umap não trabalha com conceitos, porém, como apresenta palavras-chave para a coleção inteira (ao invés, de para cada grupo gerado como o Eurekha) e uma forma de visualizar as associações entre as palavras-chave, consegue minimizar o problema. Daí, conclui-se que tanto a "interface" ou canal de comunicação quanto a forma de visualização de resultados são muito importantes em um sistema de recuperação de informações, principalmente em decorrência do contexto ou do próprio domínio.

Em comparação ao Eurekha, o Umap possui em sua "interface" importantes requisitos de usabilidade como nível de usuário ("expert" e "novice"), nível de controle, suporte à aprendizagem, apoio à decisão. Apesar disso, o Umap não é, a princípio, uma ferramenta que permita o aprendizado "natural" e intuitivo. O ser humano não está acostumado a interpretar texto através de modelos gráficos; ele precisa aprender a interpretar o mapa para obter as informações necessárias. Existe, entretanto, alguma flexibilidade de interação (multiplicidade de formas). As palavras-chave podem ser selecionadas tanto na janela de "Keywords", quanto no mapa, para se verificar a quais documentos ela pertence, por exemplo. Quanto à robustez de interação, o Umap apresenta a característica do acompanhamento com o "gato", que sugere as atitudes a

partir de um estado atual. Mas não permite recuperar (retornar) ações anteriores, apenas atualizar o mapa.

O Eurekha, por sua vez, oferece facilidade de aprendizado porque apresenta uma "interface" com características padronizadas do ambiente "Windows". Como apenas as funções são diferenciadas, pois os objetos da "interface" são os mesmos (botões, barras de rolagem, menus e outros), exigem que o usuário esteja familiarizado com o ambiente "Windows". Não possui robustez de interação, já que não há acompanhamento nem formas de recuperação. Quando o processamento inicia, o usuário não pode alterar nenhuma das configurações como o valor do coeficiente de similaridade. A flexibilidade de interação também não é uma característica que possa lhe ser atribuída. Assim, o Umap possui uma "interface" um pouco mais usável que o Eurekha.

Quanto à efetividade, os melhores resultados de "macroaverage precision" foram obtidos pelo algoritmo "cliques" do Eurekha, uma vez que este é mais rigoroso no agrupamento, já que verifica a similaridade entre todos os documentos de um grupo para fazer a alocação. Os melhores resultados de "microaverage precision" foram obtidos pelo Umap, uma vez que o Umap faz um agrupamento mais genérico analisando as palavras-chave da coleção inteira, não se preocupando com a organização da coleção. Neste caso, pode-se dizer que o Eurekha permite, para graus de similaridade mais baixos, obter maior conhecimento sobre os temas da coleção, porque ele mostra as palavras-chave de cada grupo. Para coleções maiores, é preferível trabalhar com altos graus de similaridade.

Sabe-se, entretanto, que os resultados de ambas medidas de avaliação foram encontrados para valores muitas vezes abaixo de 50% "microaveraging recall" e "macroaveraging recall". Isto se deve às diferenças entre a recuperação humana e a recuperação automática. Enquanto, a recuperação humana tem caráter subjetivo, isto é, a carga de conhecimento do indivíduo influencia o resultado; a recuperação automática é objetiva, não tendo compromisso com o contexto. Por outro lado, é óbvio que o ser humano se tornou incapaz de lidar com a sobrecarga de informações atual. O tempo de processamento das ferramentas é bem menor do que o tempo que um ser humano consome para ler e entender todos os textos; a especialista levou duas semanas para ler os 178 textos, enquanto o maior tempo de processamento das ferramentas foi de trinta e duas horas e quarenta e dois minutos, menos do que dois dias. Este tempo foi reduzido na nova versão do Eurekha. O Umap consome apenas alguns segundos. Assim, os resultados obtidos neste trabalho servem para validar as ferramentas avaliadas como ferramentas de auxílio à recuperação de informações.

Não obstante, o conhecimento obtido através das ferramentas não pode ser descartado. Foi possível obter um leque de problemas que marcaram a região Amazônica no ano de 1999, e que se tornaram matéria jornalística para *Folha de São Paulo*: desmatamento (tema mais abordado pelo jornal), queimadas, incêndios acidentais, clima, narcotráfico, CPI do narcotráfico, irregularidades, intoxicação por mercúrio, grilagem, problemas de saúde dos índios, problema da terra, privatização da Eletronorte, crise colombiana, malária, influência norte-americana, biopirataria, necessidade de asfalto na rodovia Santarém-Cuiabá e outros. Além disso, a postura do jornal se tornou evidente sob alguns aspectos. O fato das matérias jornalísticas sobre queimadas e desmatamento serem pautadas em trabalhos de pesquisa de institutos brasileiros e internacionais conferem credibilidade à notícia. E as matérias a respeito da geografia, cultura, biodiversidade da Amazônia mostram que o jornal não se ateve apenas aos problemas da região, mas também às suas características.

O trabalho que pode ser desenvolvido futuramente é a comparação do conhecimento descoberto com a análise de discurso dos textos feita por um especialista. As contribuições deste trabalho são: o estudo dos temas mais importantes das áreas de RI e KDT; proposta para solução dos problemas do Eurekha e validação das técnicas de agrupamento.

Bibliografia

[BRU 98] BRUSSO, Marcos José. **O Paralelismo na Mineração de Regras de Associação**. 1998. Trabalho Individual (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[BUS 45] BUSH, Vannevar. As we may think. **Atlantic Monthly**, [S.l.], v.176, n. 1, p. 101-108, 1945.

[COO 97] COOLEY, Robert; MOBASHER, Bamshad; SRIVASTAVA, Jaideep. **Web Mining**: Information and Pattern Discovery on the World Wide Web. Minneapolis: Department of Computer Science University of Minnesota Minneapolis, 1997. Disponível em:<http://www-users.cs.umn.edu/~mobasher/webminer/survey/ survey.html> Acesso em: 2000.

[CRO 95] CROFT, W.Bruce. What do people want from Information Retrieval? **D-Lib Magazine**, 1995. Disponível em: <http://www.dlib.org/dlib/ november95/11croft.html> Acesso em: 2000.

[CRO 94] CROSS, Valerie. Fuzzy Informaton Retrieval. **Journal of Intelligent Information Systems**, [S.l.], v.3, n. 1, Feb. 94.

[EVA 98] EVANS, David A.; HUETTNER, Alison; TONG, Xiang; JANSEN, Peter; BENNETT, Jeffrey. Effectiveness of Clustering in Ad-Hoc Retrieval. In: TEXT RETRIEVAL CONFERENCE, TREC, 7., 1998. **Proceedings...** Gaithersburg: CLARITECH Corporation, 1998. p. 143.

[FEL 99] FELDMAN, Ronen. Tutorial 4. Mining Unstructured Data. **ACM SIGKDD**, New York, v.1, p. 182-236, Feb. 1999.

[FEL 98] FELDMAN, Ronen et al. **Text mining at the Term Level**. 1998. Disponível em: <http://citeseer.nj.nec.com/feldman98text.html> Acesso em: 2001.

[FEL 97] FELDMAN, Ronen; HIRSH, Haym. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, Boston, v.9, n.1, p.83-97, July/Aug. 1997.

[FUJ 2001] FUJII, Atsushi; ISHIKAWA, Tetsuya. Evaluating Multi-lingual Information Retrieval e Clustering at ULIS. In: WORKSHOP MEETING ON EVALUATION OF CHINESE & JAPANESE TEXT RETRIEVAL AND TEXT SUMMARIZATION, NTCIR, 2., 2001. **Proceedings...** [S.l.: s.n.], 2001.

[HAN 98] HANSEN, Preben. **Evaluation of IR User Interface:** Implications for User Interface Design. 1998. Disponível em: <http://www.hb.se/bhs/ith/2-98/ph.htm>. Acesso em: 2001.

[JIA 2001] JIANG, Haifeng; LOU, Wenwu; WANG, Wei. Three-tier Clustering: an Online Citation Clustering System. In: WEB-AGE INFORMATION MANAGEMENT, WAIM, 2., 2001. **Proceedings...** [S.l.: s.n.], 2001.

[KOW 97] KOWALSKI, Geraldi. **Information Retrieval Systems:** Theory and Implementation. Massachusetts: Kluwer Academic Publishers, 1997.

Implementation. Massachusetts: Kluwer Academic Publishers, 1997.

[LEW 91] LEWIS, David D. Evaluating Text Categorization. In: SPEECH AND NATURAL LANGUAGE WORKSHOP, 1991. **Proceedings...** San Mateo, CA: [s.n.], 1991. p.312-318.

[LOH 99] LOH, Stanley. **Descoberta de Conhecimento em Textos**. 1999. 140p. Exame de Qualificação (Doutorado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[LOH 2000] LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL, OIA, 4., 2000. **Anais...** Pelotas: EDUCAT, 2000.

[LOH 2000a] LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. Concept-based knowledge discovery in texts extracted from the WEB. **ACM SIGKDD Explorations**, NewYork, v.2, n.1, p. 29-39, July 2000.

[MAR 59] MARON, M.E.; KUHNS, J.L. On relevance, probabilistic indexing and information retrieval. In: SPARCK JONES, Karen; WILLET, Peter (Ed.). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.

[NEV 2000] NEVES, Maria Helena de Moura. **Gramática de Usos do Português**. São Paulo: Fundação Editora Unesp, 2000. p.108-109.

[RIJ 79] RIJSBERGEN, C. J. van. **Information Retrieval**. 1979. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html> Acesso em: 2000.

[SAL 83] SALTON, Gerard; McGILL, M.L. **Introduction to Modern Information Retrieval.** New York: McGraw-Hill Book Company, 1983.

[SAL 83a] SALTON, Gerard; McGILL, M.J. **The SMART and SIRE Experimental Retrieval Systems**. In: SPARCK JONES, Karen; WILLET, Peter (Ed.). Readings in Information Retrieval. San Francisco: Morgan Kaufmann, 1997.

[SPA 97] SPARCK JONES, Karen; WILLET, Peter (Ed.). **Readings in Information Retrieval.** San Francisco: Morgan Kaufmann, 1997.

[STE 2000] STEINBACH, Michael; KARYPIS, George; VIPIN, Kumar. A Comparison of Document Clustering Techniques. In: TEXT MINING WORKSHOP, KDD, 2000. **Proceedings...** Disponível em: <www.researchindex.com/>. Acesso em: 2001.

[SUG 2000] SUGIURA, Atsushi; ETZIONI, Oren. Query Routing for Web Search Engines: Architecture and Experiments. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 9., 2000. **Proceedings...** Disponível em: <http://www.www9.org/w9cdrom/139/139.html> Acesso em: 2001.

[TAG 81] TAGUE-SUTCLIFFE, Jean. The pragmatics of information retrieval experimentation, revisited. In: SPARCK JONES, Karen; WILLET, Peter

(Ed.). **Readings in Information Retrieval.** San Francisco: Morgan Kaufmann, 1997.

[TAN 99] TAN, Ah-Hwee. Text Mining: the state of the art and the challenges. In: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, PAKDD, 1999. **Proceedings...** [S.l.: s.n.], 1999.

[WIV 2000] WIVES, Leandro Krug. **Tecnologias de descoberta de conhecimento em textos aplicadas a inteligência competitiva**. 2000. 100p. Exame de Qualificação (Doutorado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[WIV 97] WIVES, Leandro Krug. **Um Estudo sobre Técnicas de Recuperação de Informações com Ênfase em Informações Textuais.** 1997. Trabalho Individual I (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[WIV 99] WIVES, Leandro Krug; OLIVEIRA, José Palazzo M. de. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de "clustering".** 1999. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Anexo Resultados dos Experimentos

Algoritmo: "Best-star"
Período: 1999
Graus de similaridade: 0 e 2.5%

Grupo	Arquivos	Palavras relevantes	%
1	abr155, ago112	AMAZÔNIA (2; 0,015), HOTEL (1; 0,012), GOYA (1; 0,012), MÓVEIS (1; 0,011), MADEIRA (1; 0,009), MELIÁ (1; 0,008), FESTIVAL (1; 0,008), CIDADE (2; 0,007), BRASILEIRA (2; 0,007), REGIÃO (2; 0,007), MOSTRA (1; 0,007), ZANINE (1; 0,007), INFLUÊNCIAS (2; 0,005)	1,14
2	abr156, fev174, jun139	AMAZÔNIA (3; 0,10), QUEIMADAS (1; 0,04), OPÇÕES (1; 0,04), DESCOBRE (1; 0,04), EMBRAPA (2; 0,03), INTOXICAÇÃO (1; 0,03), MERCÚRIO (1; 0,03), ASSENTAMENTO (1; 0,02), PROJETO (1; 0,02), TRITURAR (1; 0,02), QUEIMÁ-LA (1; 0,02), CAPOEIRA (1; 0,02), SINTOMAS (1; 0,01), SCIENTIST (1; 0,01)	1,71
3	abr157, abr59	FLORESTA (2; 0,024), DEVASTAÇÃO (2; 0,011), PESQUISA (2; 0,010), INCÊNDIOS (2; 0,008), IBAMA (2; 0,008), NATURE (2; 0,008), EXTRAÇÃO (2; 0,008), ÁRVORES (2; 0,008), SATÉLITES (1; 0,006), SATÉLITE (1; 0,006)	1,14
4	abr158, abr160, abr35, jan178	AMAZÔNIA (4; 0,030), BRASIL (4; 0,030), DESMATAMENTO (2; 0,018), ESTUDO (2; 0,016), DEVASTAÇÃO (2; 0,011), REVISTA (2; 0,011), IMAGENS (2; 0,011), EUA (2; 0,011), PUBLICADO (2; 0,011), SATÉLITE (2; 0,011), NATURE (2; 0,011)	2,29
5	abr159, jun138	FNAC (1; 0,014), AMAZÔNIA (2; 0,007), SÃO (1; 0,006), CAPA (1; 0,006), REVISTAS (1; 0,006), PLÁSTICA (1; 0,004), SEMANAIS (1; 0,004), ÁTICA (1; 0,004), PAULO (1; 0,004), PAULISTA (1; 0,004), BRASIL (2; 0,004)	1,14
6	abr36, jun32, nov80, nov83, out16, set104, set20	AMAZÔNIA (7; 0,016), FLORESTA (6; 0,016), PESQUISA (7; 0,011), FOGO (5; 0,009), IPAM (4; 0,007), REGIÃO (5; 0,006), INSTITUTO (7; 0,006), MADEIRA (6; 0,005), ÁRVORES (5; 0,005), AMBIENTAL (7; 0,004), PARÁ (5; 0,004), SECA (4; 0,004), INCÊNDIOS (3; 0,004), VÁRZEA (1; 0,004), QUEIMADAS (4; 0,004)	4
7	ago111, jan39	FHC (2; 0,025), INFRA-ESTRUTURA (1; 0,019), ESTADO (2; 0,015), AMAZÔNIA (2; 0,015), REUNIÃO (1; 0,014), GOVERNADOR (1; 0,014), ALMIR (1; 0,014), SANTARÉM-CUIABÁ (1; 0,011), PRESIDENTE (1; 0,011), ASFALTAR (1; 0,011)	1,14
8	ago113, ago114, ago117, dez76	BRASIL (4; 0,012), COLÔMBIA (3; 0,008), MILITAR (3; 0,006), INTERVENÇÃO (3; 0,005), POLÍTICA (2; 0,005), EUA (2; 0,005), CARVOEIROS (1; 0,005), GOVERNO (3; 0,005), WROBEL (1; 0,004), AMAZÔNIA (4; 0,004), PAÍS (4; 0,004), PRODUÇÃO (2; 0,004), FOLHA (2; 0,003), NORTE-AMERICANA (2; 0,003)	2,29
9	ago116, nov78, nov82, out96	BRASIL (4; 0,014), PAÍS (3; 0,011), COLÔMBIA (2; 0,007), ESTADOS (3; 0,007), POLÍTICA (3; 0,007), ARGENTINA (1; 0,006), EUA (3; 0,006), UNIDOS (2; 0,006), NARCOTRÁFICO (2; 0,006), AMAZÔNIA (4; 0,006), MUNDO (4; 0,006), EMBAIXADOR (1; 0,005)	2,29
10	ago118, jul123, mar163, nov77	DESENVOLVIMENTO (2; 0,007), RECURSOS (3; 0,007), MUNDO (2; 0,007), LIMITES (2; 0,005), AMBIENTE (4; 0,005), AMBIENTAL (4; 0,005), MEIO (3; 0,004), MAIOR (3; 0,003), CRESCIMENTO (3; 0,003), SUSTENTÁVEL (3; 0,003), SUSTENTABILIDADE (2; 0,003), AMBIENTAIS (3; 0,003)	2,29
11	ago134, jun136	AMAZÔNIA (2; 0,043), INTERVENÇÃO (2; 0,014), NATURAIS (1; 0,013), RECURSOS (1; 0,013), BRASIL (2; 0,012), MILITAR (2; 0,010), COMANDANTE (2; 0,010), LESSA (2; 0,010), BRASILEIRO (1; 0,009)	1,14

Grupo	Arquivos	Palavras relevantes	%
12	ago28, dez109, dez120, mar63, nov115	AMAZÔNIA (4; 0,028), PAULO (2; 0,011), SÃO (2; 0,011), ÁREA (3; 0,011), REGIÃO (2; 0,011), FAZENDA (1; 0,010), PROGRAMA (2; 0,009), RONDÔNIA (3; 0,009), ACRE (3; 0,009), PARÁ (3; 0,009), PRODUÇÃO (2; 0,009), AUTORIZAÇÃO (2; 0,008), IBAMA (2; 0,008), GROSSO (2; 0,008), MATO (2; 0,008), RORAIMA (2; 0,008), AMAZONAS (2; 0,008), AMAPÁ (2; 0,008), TOCANTINS (2; 0,008), GOVERNO (1; 0,007), QUEIMADAS (1; 0,007), DESMATAMENTO (1; 0,006)	2,86
13	ago29, jun137	CLIMA (1; 0,039), AMAZÔNIA (2; 0,031), PAÍSES (1; 0,019), LBA (1; 0,019), INGLÊS (1; 0,019), INFLUÊNCIA (1; 0,019), EXPERIMENTO (1; 0,019), ESCALA (1; 0,019), REGIÃO (1; 0,019), PROJETO (1; 0,019), PLANETA (1; 0,019), BIOSFERA (1; 0,019), ATMOSFERA (1; 0,019), AMAZÔNICA (1; 0,019), PESQUISADORES (1; 0,019)	1,14
14	ago30, ago49, set26	MCCAFFREY (2; 0,014), BRASIL (3; 0,013), LEI (2; 0,013), COLÔMBIA (3; 0,011),,MINISTRO (2; 0,011), GOVERNO (3; 0,010), DEFESA (1; 0,010), NARCOTRÁFICO (3; 0,009), NORTE (2; 0,009), EUA (2; 0,009), BRASILEIRO (2; 0,008), PRESIDENTE (2; 0,008), GENERAL (2; 0,008), COMBATE (3; 0,008), FRONTEIRAS (1; 0,008), AMAZÔNIA (2; 0,007)	1,71
15	ago50, ago52, mai153, mai33, out98, set102, ago53	PRESIDENTE (5; 0,013), FHC (3; 0,009), BRASIL (6; 0,007), PAÍS (6; 0,006), GOVERNO (3; 0,005), ACRE (4; 0,004), BATERIAS (1; 0,004), REGIÃO (5; 0,004), MEIO (4; 0,004), FUMAÇA (1; 0,003), AMBIENTE (3; 0,003), NACIONAL (3; 0,003), POVO (3; 0,003), QUEIMADAS (1; 0,002), SANEAMENTO (1; 0,002)	4
16	ago51, mar61	GOVERNO (2; 0,036), PROGRAMA (2; 0,022), DROGAS (2; 0,021), EUA (1; 0,017), BRASILEIRO (1; 0,017), BRASIL (1; 0,017), MILITAR (2; 0,015), PREVENÇÃO (1; 0,014), NACIONAL (1; 0,014), FORÇA-TAREFA (1; 0,014), COOPERAÇÃO (1; 0,011), COMBATE (1; 0,011), PAÍSES (1; 0,011), AJUDA (1; 0,011), ESTRATÉGIA (1; 0,011), MCCAFFREY (1; 0,011)	1,14
17	dez1, dez66, jan180, jul121, jun143, mai151, nov8	AMAZÔNIA (7; 0,017), IBAMA (3; 0,010), MADEIREIRAS (2; 0,009), MEIO (4; 0,008), PRESIDENTE (2; 0,008), DENGUE (1; 0,007), AMBIENTE (3; 0,007), RIO (2; 0,007), SP (3; 0,006), INSTITUTO (3; 0,006), SENAD[10] (1; 0,005), NARCOTRÁFICO (1; 0,005), ESPAÇO (1; 0,005), CHELOTTI (1; 0,005), AÉREO (1; 0,005), SECRETARIA (2; 0,005)	4
18	dez2, jul130, nov81	BOSQUE (1; 0,012), PARQUE (2; 0,010), AMAZONAS (2; 0,009), ANIMAIS (2; 0,008), ANO (3; 0,008), TERRAS (1; 0,007), ÁREA (1; 0,006), ZOOLÓGICO (1; 0,006), PEIXES (1; 0,006), TRILHAS (1; 0,006), ÁGUA (3; 0,006), JACARÉS (2; 0,005)	1,71
19	dez3, dez68	CHEIA (1; 0,018), ILHAS (2; 0,016), TERRA (1; 0,014), NEGRO (2; 0,014), RIO (2; 0,014), METROS (1; 0,014), LAVADAS (1; 0,014), ÉPOCA (1; 0,014), ANOS (1; 0,014), LOCAL (2; 0,011), ANAVILHANAS (2; 0,011)	1,14
20	dez4, dez70, mai152	MAUÉS (2; 0,012), GUARANÁ (1; 0,012), LENDA (2; 0,010), RIOS (2; 0,008), PESCADORES (1; 0,008), ÍNDIOS (3; 0,008), CRIANÇA (2; 0,008), BOTO (1; 0,007), TEATRO (1; 0,007), SESC (1; 0,007), IPIRANGA (1; 0,007), BERTAZZO (1; 0,007), FLORESTA (2; 0,006), TUPI (1; 0,006)	1,71

[10] Secretaria Nacional Anti-drogas

Grupo	Arquivos	Palavras relevantes	%
21	dez5, mar165	ALDEIA (2; 0,008), ÍNDIOS (2; 0,008), ÍNDIO (2; 0,007), TICUNAS (1; 0,007), RIO (2; 0,006), CRIANÇAS (1; 0,005), FUNAI (1; 0,004), INTERDIÇÃO (1; 0,004), FLORESTA (2; 0,004), SANTOS (2; 0,004), ÁREA (1; 0,003), TICUNA (1; 0,003)	1,14
22	dez65, mar60	GOVERNO (2 ;0,026), ELETRONORTE (2; 0,024), VENDA (2; 0,022), PRIVATIZAÇÃO (2; 0,02), FEDERAIS (1; 0,018), SETOR (2; 0,014), CHESF (2; 0,014), ANO (2; 0,014), ENERGÉTICAS (1; 0,012), ENERGIA (2; 0,012), FURNAS (1; 0,012)	1,14
23	dez67, dez73	MUSEU (2; 0,038), BRITÂNICO (2; 0,02), AMAZÔNIA (2; 0,018), EXPOSIÇÃO (2; 0,017), BRASIL (2; 0,014), ARTE (2; 0,011), MOSTRA (2; 0,011), PEÇAS (2; 0,011), ASSOCIAÇÃO (2; 0,007), ARTES (2; 0,007)	1,14
24	dez69, jun141	SISTEMAS (1; 0,018), PRODUTOR (1; 0,014), EMBRAPA (1; 0,014), AMAZÔNIA (2; 0,013), EXEMPLO (2; 0,013), PEIXES (1; 0,013), MEIO (1; 0,013), PESQUISADOR (1; 0,009), PASTO (1; 0,009), MANDIOCA (1;0,009), SILVOPASTORIS (1; 0,009)	1,14
25	dez71, mai150, mai154, nov7	JUSTIÇA (3; 0,007), DARLY (1; 0,007), ANOS (2; 0,007), GOVERNO (3; 0,007), EMPRESA (2; 0,006), PROMOTOR (2; 0,005), PARÁ (2; 0,005), JUIZ (1; 0,005), REGIME (1; 0,005), TERRAS (1; 0,005), INCRA (1; 0,005), BRASIL (4; 0,005), AMAZÔNIA (4; 0,005), PLANTAS (1; 0,005), CASO (4; 0,004), DIREITO (2; 0,004), MORTE (1; 0,004)	2,29
26	dez75, jul129, nov46	COMUNIDADE (2; 0,015), ÍNDIOS (3; 0,011), FUNAI (2; 0,01), SOLIMÕES (3; 0,01), SUICÍDIOS (1; 0,008), ANOS (3; 0,007), HOMENS (2; 0,007), TICUNAS(1; 0,007), TICUNA (2; 0,007), FEDERAL (2; 0,007), ÍNDIO(2; 0,007), IPRAM (1; 0,006), LENY(1; 0,006), AULAS (1; 0,006)	1,71
27	fev166, jul133, mar164, out99	MALÁRIA (3; 0,028); CASOS (4; 0,02), DOENÇA (3; 0,013), RORAIMA (3; 0,01), SAÚDE (4; 0,0080), VIVAX (1; 0,008), ÍNDIOS (2; 0,007), FUNDAÇÃO (4; 0,006), CEGUEIRA (1; 0,006), BRASIL (3; 0,005), MANAUS (2; 0,005), PESSOAS (2; 0,005), TRACOMA (1; 0,005)	2,29
28	fev167, fev168	BARCO (2; 0,062), AMAZÔNIA (2; 0,057), PESSOAS (2; 0,052), AFUNDOU (2; 0,047), RIO (2; 0,033), DESAPARECIDAS (2; 0,028), SELVA (2; 0,028), MADEIRA (2; 0,028), HUMAITÁ (2; 0,023), MANICORÉ (2; 0,023)	1,14
29	fev169, fev176, jul127, jun144, out101	AMAZÔNIA (5; 0,028), INSTITUTO (4; 0,014), INPE (3; 0,013), ÁREAS (4; 0,012), ANO (4; 0,012), KM2 (2; 0,011), DESMATADAS (2; 0,011), PESQUISAS (3; 0,01), ÁREA (3; 0,008), AVIÃO (2; 0,008), ESPACIAIS (3; 0,008), NACIONAL (3; 0,008)	2,86
30	fev170, fev171	DESMATAMENTO (2; 0,044), AUTORIZAÇÕES (2; 0,037), FUNCIONÁRIOS (2; 0,024), IBAMA (2; 0,024), AFASTADOS (2; 0,024), ARIPUANÃ (2; 0,018), AMAZÔNIA (2; 0,018), FAZENDA (2; 0,018), AMBIENTE (2; 0,015), MEIO (2; 0,015), FILHO (2; 0,013), SARNEY (2; 0,013), ACUSADOS, (2; 0,012), INSTITUTO (2; 0,01)	1,14
31	fev173, fev172, mai149, nov87, nov88	BARCO (4; 0,024), NAUFRÁGIO (4; 0,018), RIO (4; 0,017), PESSOAS (4; 0,015), CORPOS (2; 0,013), ÁGUAS (2; 0,013), AMAZÔNIA (5; 0,011), MADEIRA (4; 0,01), PASSAGEIROS (3; 0,009), ACIDENTE (3; 0,008), EMBARCAÇÃO (2; 0,007)	2,86
32	fev177, jun145	AMAZÔNIA (2; 0,018), MINISTRO (2; 0,018), DESMATAMENTO (2; 0,017), MEIO (2; 0,017), AMBIENTE (2; 0,015), SARNEY (2; 0,014), CAUSAS (1; 0,012), FILHO (2; 0,012), WWW (1; 0,011), ANO (2; 0,009), SITE (1; 0,009), PRESIDENTE (2; 0,008)	1,14

Grupo	Arquivos	Palavras relevantes	%
33	fev38, jan183, set105, jan181	QUEIMADAS (3; 0,018), AMAZÔNIA (4; 0,014), ANO (3; 0,01), INSTITUTO (4; 0,01), INPE (3; 0,009), FOCOS (2; 0,007), NACIONAL (4; 0,007), ÁREA (3; 0,006), BRASIL (4; 0,006), PESQUISAS (4; 0,005)	2,29
34	jan179, jul131, out93	AMBIENTAL (2; 0,009), INFANTICÍDIO (1; 0,008), AMBIENTE (2; 0,005), SUSTENTÁVEL (2; 0,005), GESTÃO (1; 0,005), FUTURO (1; 0,005), MODERNIDADE (1; 0,004), FILHOS(1; 0,004), CRIANÇA (1; 0,004), SÉCULO (2; 0,004), MODERNIZAÇÃO (1; 0,004), CONTRADIÇÃO (1; 0,004)	1,71
35	jan40, jun135	NACIONAL (2; 0,011), TERRAS (1; 0,01), INDÍGENAS (1; 0,01), PAÍSES (2; 0,009), NORTE (2; 0,006), REGIÃO (2; 0,006), ÁREAS (1; 0,005), ÍNDIOS (1; 0,005), BRASILEIRA (1; 0,005), TENDÊNCIA (1; 0,005), PROFESSOR (1; 0,005), POLÍTICA (1; 0,005)	1,14
36	jan64, jul128, nov10, nov47, ago31	COLÔMBIA (5; 0,023), BRASIL (4; 0,019), POLÍCIA (3; 0,015), ARMAS (1; 0,014), FARC (3; 0,012), GUERRILHEIROS (3; 0,009), GUERRILHA (4; 0,009), PERU (4; 0,008), FEDERAL (2; 0,008), GOVERNO (4; 0,007), FRONTEIRA (2; 0,007)	2,86
37	jul122, jul126	CASOS (2; 0,058), ANO (2; 0,038), SEMESTRE (2; 0,026), PRIMEIRO (2; 0,026), MALÁRIA (2; 0,026), IANOMÂMIS (1; 0,022), REGISTRADOS (2; 0,019), RORAIMA (1; 0,018), ACRE (1; 0,018), ÍNDICE (1; 0,015), DISSEMINAÇÃO (1; 0,015), CONTROLE (1; 0,015)	1,14
38	jul124, jul125, jul57, jun142, out97	SAÚDE (4; 0,016), RIO (5; 0,01), PAÍS (3; 0,008), ORGANIZAÇÃO (2; 0,008), TRABALHO (2; 0,007), FNS (2; 0,007), ÍNDIOS (2; 0,007), MÉDICO (2; 0,007), FUNDAÇÃO (4; 0,007), PESQUISA (2; 0,006), FUNAI (2; 0,006), INDÍGENAS (3; 0,006)	2,86
39	jul132, mai148, jan182	SIVAM (3; 0,022), AMAZÔNIA (3; 0,016), PETROBRAS (1; 0,016), REGIÃO (2; 0,014), ACORDO (2; 0,014), AVIÃO (2; 0,011), VIGILÂNCIA (3; 0,01), LBA (1; 0,009), SISTEMA (3; 0,009), ER-2 (1; 0,007)	1,71
40	jun140, nov12, out95, set25	PESQUISA (4; 0,009), ÁRVORES (3; 0,009), AMAZÔNIA (3; 0,008), FLORESTA (3; 0,007), CAPOEIRA (1; 0,006), PESQUISADORES (3; 0,005), ANOS (3; 0,005), SEMENTES (1; 0,005), IPAM (3; 0,005)	2,29
41	jun146, out94, set108, abr34	GOVERNO (3; 0,023), AMAZÔNIA (4; 0,022), PROGRAMA (3; 0,01), ANO (3; 0,009), BRASIL (4; 0,008), INVESTIMENTOS (2; 0,008), PAÍSES (2; 0,008), ORÇAMENTO (2; 0,008), ESTUDO (1; 0,007), GREENPEACE (1; 0,006)	2,29
42	jun147, out48	FHC (2; 0,022), PAÍSES (2; 0,018), GREENPEACE (1; 0,017), BRASIL (2; 0,013), PAZ (1; 0,012), COLÔMBIA (1; 0,012), ENTIDADE (1; 0,011), TEMA (2; 0,011), GOVERNO (2; 0,011), REGIÃO (2; 0,01), PASTRANA (1; 0,01), PROCESSO (1; 0,01), COLOMBIANO (1; 0,01), AMAZÔNIA (1; 0,008); ONGS (1; 0,008)	1,14
43	mar161, mar162, mar37, fev175	DESMATAMENTO (4; 0,058), AMAZÔNIA (4; 0,038), MINISTÉRIO (3; 0,031), REGIÃO (4; 0,028), PROIBIÇÃO (3; 0,025), FAMILIAR (3; 0,023), MEDIDA (3; 0,023), MEIO (4; 0,023), AMBIENTE (4; 0,023), AGRICULTURA (3; 0,022), PROPRIEDADES (3; 0,02), SARNEY (3; 0,02), FILHO (3; 0,02), MINISTRO (3; 0,019)	2,29
44	mar62, nov11, nov44, nov84	COCAÍNA (4; 0,03), REGIÃO (3; 0,013), COLÔMBIA (4; 0,012), TRÁFICO (4; 0,012), PERU (3; 0,012), FRONTEIRA (2; 0,011), AMAZÔNICA (3; 0,11), RIO (2; 0,11), NARCOTRÁFICO (3; 0,011), SURINAME (1; 0,009)	2,29

Grupo	Arquivos	Palavras relevantes	%
45	nov42, nov43, abr58	AVIÃO (3; 0,03), DROGA (3; 0,021), COCAÍNA (3; 0,019), PISTA (2; 0,015), SURINAME (2; 0,015), TRÁFICO (2; 0,011), OLIVEIRA (1; 0,011), BANCOS (2; 0,01), COLÔMBIA (2; 0,01), PILOTO (2; 0,009), CLANDESTINA (2; 0,009)	1,71
46	nov45, nov79	NARCOTRÁFICO (2; 0,011), SOLIMÕES (2; 0,011), TRAFICANTE (1; 0,011), SERRARIA (1; 0,011), MADEIRA (2; 0,01), EXTRAÇÃO (2; 0,009), GOVERNADOR (1; 0,009), ALTO (2; 0,009), ÁREA (2; 0,009), TABATINGA (2; 0,009), BISPO (1; 0,008), MANAUS (2; 0,008), COCAÍNA (2; 0,008)	1,14
47	nov85, nov86, nov89, nov9, nov41	CPI[11] (4; 0,016), SÂMIA (4; 0,009), CURICA (4; 0,009), CAMPINAS (2; 0,007), DEPUTADO (4; 0,006), TRAFICANTE (4; 0,006), DINHEIRO (4; 0,005), SÃO (5; 0,005), FEDERAL (5; 0,005), PRISÃO (4; 0,005), PAULO (4; 0,005), ORGANIZAÇÃO (2; 0,004), NARCOTRÁFICO (4; 0,004), RIO (2; 0,004), DEPOIMENTO (5; 0,004), AMAZÔNIA (4; 0,004)	2,86
48	nov90, nov91	CPI (2; 0,038), NARCOTRÁFICO (2; 0,022), MILITARES (2; 0,015), ANTITRÁFICO (2; 0,015), OPERAÇÃO (1; 0,012), EXÉRCITO (1; 0,012), AÇÃO (1; 0,012), MÃOS (1; 0,012), ARMADAS (2; 0,011), FORÇAS (2; 0,011), BC (1; 0,01), PRESIDENTE (2; 0,01), PROPOSTA (2; 0,009), COMISSÃO (2;0,009), COMBATE (2; 0,009)	1,14
49	out100, out15, out17, out92, ago110	MADEIRA (4; 0,015), MANEJO (4; 0,012), BRASIL (4; 0,01), CERTIFICAÇÃO (4; 0,009), AMAZÔNIA (5;0,008), FLORESTAL (4; 0,008), FSC (4; 0,006) , FLORESTAS (3; 0,006), PRODUÇÃO (3; 0,005), MERCADO (3; 0,005), SUSTENTÁVEL (3; 0,005)	2,86
50	out13, out14	MADEIRA (2; 0,019), IMAZON (1; 0,011), PARAGOMINAS (1; 0,011), AMAZÔNIA (2; 0,011), ÁRVORES (1; 0,011), EQUIPE (1; 0,011), BELÉM (1; 0,007), UHL (1; 0,007), MANEJO (2; 0,007), SUSTENTÁVEL (2; 0,007), REGIÃO (2; 0,007)	1,14
51	set106, set107, set27, dez6	QUEIMADAS (2; 0,021), IBAMA (4; 0,018), PRISÃO (2; 0,013), MEIO (4; 0,011), AMBIENTE (4; 0,010), FOCOS (3; 0,008), PORTARIA (2; 0,008), FEDERAL (4; 0,007),, PAULO (3; 0,007), SÃO (3; 0,006), SARNEY (3; 0,006), FILHO (3; 0,006), MINISTRO (3; 0,006)	2,29
52	set19, set21, set22, set24, out18	AMAZÔNIA (5; 0,014), SEMINÁRIO (5; 0,012), SOJA (3; 0,011), BIODIVERSIDADE (5; 0,009), DESENVOLVIMENTO (5; 0,007), CONSERVAÇÃO (5; 0,007), ÁREAS (3; 0,007), MAPA (3; 0,006), EIXOS (4; 0,006), GOVERNO (4; 0,006), ÁREA (3; 0,006), FEDERAL (4; 0,005), AMBIENTE (4; 0,005), MAPAS (3; 0,005)	2,86

Grau de similaridade: 5%

Grupo	Arquivos	Palavras relevantes	%
1	abr156, jun139	AMAZÔNIA (2; 0,11), QUEIMADAS (1; 0,06), OPÇÕES (1; 0,06), DESCOBRE (1; 0,06), EMBRAPA (2; 0,05), ASSENTAMENTO (1; 0,04), PROJETO (1; 0,03), TRITURAR (1; 0,03), QUEIMÁ-LA (1; 0,03), CAPOEIRA (1; 0,03), SUSTENTÁVEL (1; 0,02), DESENVOLVIMENTO (1; 0,02),, BID (1; 0,02)	1,28
2	abr157, abr59	FLORESTA (2; 0,024), DEVASTAÇÃO (2; 0,011), PESQUISA (2; 0,010), INCÊNDIOS (2; 0,008), IBAMA (2; 0,008), NATURE (2; 0,008), EXTRAÇÃO (2; 0,008), ÁRVORES (2; 0,008), SATÉLITES (1; 0,006), SATÉLITE (1; 0,006)	1,28

[11] Comissão Parlamentar de Inquérito

Grupo	Arquivos	Palavras relevantes	%
3	abr158, abr160, abr35	BRASIL (3; 0,034), DESMATAMENTO (2; 0,024), AMAZÔNIA (3; 0,021), ESTUDO (2; 0,021), NATURE (2; 0,015), REVISTA (2; 0,015), IMAGENS (2; 0,015), EUA (2; 0,015)	1,92
4	abr36, jun32, nov80, nov83, out16, set104, set20	AMAZÔNIA (7; 0,016), FLORESTA (6; 0,016), PESQUISA (7; 0,011), FOGO (5; 0,009), IPAM (4; 0,007), REGIÃO (5; 0,006), INSTITUTO (7; 0,006), MADEIRA (6; 0,005), ÁRVORES (5; 0,005), AMBIENTAL (7; 0,004), PARÁ (5; 0,004), SECA (4; 0,004), INCÊNDIOS (3; 0,004), VÁRZEA (1; 0,004), QUEIMADAS (4; 0,004)	4,49
5	ago111, jan39	FHC (2; 0,025), INFRA-ESTRUTURA (1; 0,019), ESTADO (2; 0,015), AMAZÔNIA (2; 0,015), REUNIÃO (1; 0,014), GOVERNADOR (1; 0,014), ALMIR (1; 0,014), SANTARÉM-CUIABÁ (1; 0,011), PRESIDENTE (1; 0,011), ASFALTAR (1; 0,011)	1,28
6	ago113, ago114, ago117, abr159	BRASIL (4; 0,010), COLÔMBIA (3; 0,008), MILITAR (3; 0,006), POLÍTICA (3; 0,006), INTERVENÇÃO (3; 0,005), EUA (2; 0,005), GOVERNO (3; 0,005), WROBEL (1; 0,004), FOLHA (3; 0,004), PAÍS (4; 0,004), AMAZÔNIA (4; 0,004)	2,56
7	ago116, nov78, nov82, out96	BRASIL (4; 0,014), PAÍS (3; 0,011), COLÔMBIA (2; 0,007), ESTADOS (3; 0,007), POLÍTICA (3; 0,007), ARGENTINA (1; 0,006), EUA (3; 0,006), UNIDOS (2; 0,006), NARCOTRÁFICO (2; 0,006), AMAZÔNIA (4; 0,006), MUNDO (4; 0,006)	2,56
8	ago118, mar163, nov77	DESENVOLVIMENTO (2; 0,010), MUNDO (2; 0,009), RECURSOS (2; 0,008), LIMITES (2; 0,007), AMBIENTAL (3; 0,006), CRESCIMENTO (3; 0,004), SUSTENTÁVEL (3; 0,004), AMBIENTE (3; 0,004), SUSTENTABILIDADE (2; 0,004), AMBIENTAIS (3; 0,004)	1,92
9	ago134, jun136	AMAZÔNIA (2; 0,043), INTERVENÇÃO (2; 0,014), NATURAIS (1; 0,013), RECURSOS (1; 0,013), BRASIL (2; 0,012), MILITAR (2; 0,010), COMANDANTE (2; 0,010), LESSA (2; 0,010)	1,28
10	ago28, dez109, nov115	SÃO (2; 0,019), PAULO (2; 0,019), ÁREA (3; 0,019), FAZENDA (1; 0,018), AMAZÔNIA (3; 0,016), CIDADE (2; 0,013), AUTORIZAÇÃO (2; 0,013), PARÁ (2; 0,013), ACRE (2; 0,013), RONDÔNIA (2; 0,013), MATO (2; 0,013), IBAMA (2; 0,013), GROSSO (2; 0,013), QUEIMADAS (1; 0,011), REGIÃO (1; 0,011), RORAIMA (1; 0,010), TOCANTINS (1; 0,010), MARANHÃO (1; 0,010), FONTE (1; 0,010), DESMATAMENTO (1; 0,010)	1,92
11	ago30, ago49, set26	MCCAFFREY (2; 0,014), BRASIL (3; 0,013), LEI (2; 0,013), COLÔMBIA (3; 0,011), MINISTRO (2; 0,011), GOVERNO (3; 0,010), DEFESA (1; 0,010), NARCOTRÁFICO (3; 0,009), NORTE (2; 0,009), EUA (2; 0,009), BRASILEIRO (2; 0,008), PRESIDENTE (2; 0,008), GENERAL (2; 0,008), COMBATE (3; 0,008), FRONTEIRAS (1; 0,008), AMAZÔNIA (2; 0,007)	1,92
12	ago50, ago52, mai153, mai33, out98, set102, ago53	PRESIDENTE (5; 0,013), FHC (3; 0,009), BRASIL (6; 0,007), PAÍS (6; 0,006), GOVERNO (3; 0,005), ACRE (4; 0,004), BATERIAS (1; 0,004), REGIÃO (5; 0,004), MEIO (4; 0,004), FUMAÇA (1; 0,003), AMBIENTE (3; 0,003)	4,49
13	ago51, mar61	GOVERNO (2; 0,036), PROGRAMA (2; 0,022), DROGAS (2; 0,021), EUA (1; 0,017), BRASILEIRO (1; 0,017), BRASIL (1; 0,017), MILITAR (2; 0,015), PREVENÇÃO (1; 0,014), NACIONAL (1; 0,014), FORÇA-TAREFA (1; 0,014), COOPERAÇÃO (1; 0,011), COMBATE (1; 0,011)	1,28
14	dez1, dez66	MADEIREIRAS (2; 0,032), IBAMA (2; 0,032), AMAZÔNIA (2; 0,02), INSTITUTO (2; 0,017), AMBIENTE (2; 0,017), MEIO (2; 0,017), LEVANTAMENTO (2; 0,017), ILEGAL (2; 0,014), PARÁ (2; 0,014), EMPRESAS (2; 0,014), FISCALIZAÇÃO (2; 0,011),EXTRAÇÃO (2; 0,011), MADEIRA (1; 0,01), IMAZON (2; 0,008), RENOVÁVEIS (2; 0,008)	1,28

Grupo	Arquivos	Palavras relevantes	%
15	dez2, jul130	BOSQUE (1; 0,018), PARQUE (2; 0,015), ANIMAIS (2; 0,013), ZOOLÓGICO (1; 0,009), ÁREA (1; 0,009), TRILHAS (1; 0,009), PEIXES (1; 0,009),JACARÉS (2; 0,008), ONÇAS (1; 0,007), VISITANTES (1; 0,007), CIGS (1; 0,007), METROS (1; 0,006), LAGO (1; 0,006), PESQUISAS (1; 0,006)	1,28
16	dez3, dez68	CHEIA (1; 0,018), ILHAS (2; 0,016), TERRA (1; 0,014), NEGRO (2; 0,014), RIO (2; 0,014), LAVADAS (1; 0,014), ÉPOCA (1; 0,014), ANOS (1; 0,014), LOCAL (2; 0,011), ANAVILHANAS (2; 0,011)	1,28
17	dez4, dez70	MAUÉS (2; 0,019), GUARANÁ (1; 0,018), LENDA (2; 0,015), RIOS (2; 0,013), PESCADORES (1; 0,013), CRIANÇA (2; 0,012), BOTO (1; 0,011), FLORESTA (2; 0,01), ANSELMO (1; 0,007), AMAZONAS (2; 0,006), MORADORES (2; 0,006), TRIBO (1; 0,006)	1,28
18	dez5, mar165	ALDEIA (2; 0,008), ÍNDIOS (2; 0,008), ÍNDIO (2; 0,007), TICUNAS (1; 0,007), RIO (2; 0,006), CRIANÇAS (1; 0,005), FUNAI (1; 0,004), INTERDIÇÃO (1; 0,004), FLORESTA (2; 0,004), SANTOS (2; 0,004)	1,28
19	dez65, mar60	GOVERNO (2; 0,026), ELETRONORTE (2; 0,024), VENDA (2; 0,022), PRIVATIZAÇÃO (2; 0,02), FEDERAIS (1; 0,018), SETOR (2; 0,014), CHESF (2; 0,014), ANO (2; 0,014), ENERGÉTICAS (1; 0,012), ENERGIA (2; 0,012)	1,28
20	dez67, mar60	MUSEU (2; 0,038), BRITÂNICO (2; 0,02), AMAZÔNIA (2; 0,018), EXPOSIÇÃO (2; 0,017), ANOS (2; 0,015), BRASIL (2; 0,014), ARTE (2; 0,011), MOSTRA (2; 0,011), PEÇAS (2; 0,011)	1,28
21	dez71, mai150, mai154, nov7	JUSTIÇA (3; 0,007), DARLY (1; 0,007), ANOS (2; 0,007), GOVERNO (3; 0,007), EMPRESA (2; 0,006), PROMOTOR (2; 0,005), PARÁ (2; 0,005), JUIZ (1; 0,005), REGIME (1; 0,005), TERRAS (1; 0,005), INCRA (1; 0,005), BRASIL (4; 0,005), AMAZÔNIA (4; 0,005)	2,56
22	dez75, jul129, nov46	COMUNIDADE (2; 0,015), ÍNDIOS (3; 0,011), FUNAI (2; 0,01), SOLIMÕES (3; 0,01), SUICÍDIOS (1; 0,008), ANOS (3; 0,007), HOMENS (2; 0,007), TICUNAS (1; 0,007), TICUNA (2; 0,007), FEDERAL (2; 0,007), ÍNDIO (2; 0,007), IPRAM (1; 0,006)	1,92
23	fev166, jul133, mar164, out99	MALÁRIA (3; 0,028), CASOS (4; 0,02), ANO (2; 0,015), DOENÇA (3; 0,013), RORAIMA (3; 0,01), SAÚDE (4; 0,008), VIVAX (1; 0,008), ÍNDIOS (2; 0,007), FUNDAÇÃO (4; 0,006), CEGUEIRA (1; 0,006)	2,56
24	fev167, fev168	BARCO (2; 0,062), AMAZÔNIA (2; 0,057), PESSOAS (2; 0,052), AFUNDOU (2; 0,047), RIO (2; 0,033), DESAPARECIDAS (2; 0,028), SELVA (2; 0,028), MADEIRA (2; 0,028)	1,28
25	fev169, jul127, jun144, out101, dez76	AMAZÔNIA (5 0,018), INPE (2; 0,008), AVIÃO (2; 0,008), EXÉRCITO (2; 0,007), ÁREAS (3; 0,006), ANO (3; 0,006), MADEIRA (3; 0,006), DESMATAMENTO (3; 0,006), REGIÃO (2; 0,005)	3,21
26	fev170, fev171	DESMATAMENTO (2; 0,0448), AUTORIZAÇÕES (2; 0,037), FUNCIONÁRIOS (2; 0,0248), IBAMA (2; 0,0248), AFASTADOS (2; 0,0248), ARIPUANÃ (2; 0,0185), AMAZÔNIA (2; 0,0185), FAZENDA (2; 0,0185), AMBIENTE (2; 0,0154), MEIO (2; 0,0154), FILHO (2; 0,0138), SARNEY (2; 0,0138)	1,28
27	fev173, fev172, mai149, nov87, nov88	BARCO (4; 0,024), NAUFRÁGIO (4; 0,018), RIO (4; 0,017), PESSOAS (4; 0,015), CORPOS (2; 0,013), ÁGUAS (2; 0,013), AMAZÔNIA (5; 0,011), MADEIRA (4; 0,01), PASSAGEIROS (3; 0,009)	3,21

Grupo	Arquivos	Palavras relevantes	%
28	fev177, jun145	AMAZÔNIA (2; 0,018), MINISTRO (2; 0,018), DESMATAMENTO (2; 0,017), MEIO (2; 0,017), AMBIENTE (2; 0,015), SARNEY (2; 0,014), CAUSAS (1; 0,012), FILHO (2; 0,012)	1,28
29	fev38, jan183, set105, fev176, jan181	AMAZÔNIA (5; 0,022), INSTITUTO (5; 0,019), QUEIMADAS (3; 0,015), ANO (4; 0,014), INPE (4; 0,012), NACIONAL (5; 0,011), DESMATADAS (1; 0,01), PESQUISAS (5; 0,009)	3,21
30	jan179, jul131	AMBIENTAL (2; 0,014), AMBIENTE (2; 0,008), SUSTENTÁVEL (2; 0,007), GESTÃO (1; 0,007), FUTURO (1; 0,007), MODERNIDADE (1; 0,007), GRANDES (2; 0,006), MODERNIZAÇÃO (1; 0,006), CONTRADIÇÃO (1; 0,006), INTERESSES (2; 0,005), PROJETO (2; 0,005), PROGRESSO (2; 0,005)	1,28
31	jan40, jun135	NACIONAL (2; 0,011), TERRAS (1; 0,01), INDÍGENAS (1; 0,01), PAÍSES (2; 0,009), NORTE (2; 0,006), REGIÃO (2; 0,006), ÁREAS (1; 0,005), ÍNDIOS(1; 0,005)	1,28
32	jan64, jul128, nov10, nov47, ago31	COLÔMBIA (5; 0,023), BRASIL (4; 0,019), POLÍCIA (3; 0,015), ARMAS (1; 0,014), FARC (3; 0,012), GUERRILHEIROS (3; 0,009), GUERRILHA (4; 0,009), PERU (4; 0,008), FEDERAL (2; 0,008), GOVERNO (4; 0,007), FRONTEIRA (2; 0,007)	3,21
33	jul122, jul126	CASOS (2; 0,058), ANO (2; 0,038), SEMESTRE (2; 0,026), PRIMEIRO (2; 0,026), MALÁRIA (2; 0,026), IANOMÂMIS (1; 0,022), REGISTRADOS (2; 0,019), RORAIMA (1; 0,018), ACRE (1; 0,018), ÍNDICE (1; 0,015), DISSEMINAÇÃO (1; 0,015), CONTROLE (1; 0,015), MORTES (2; 0,013)	1,28
34	jul124, jul125, jun142, out97	SAÚDE (4; 0,021), ORGANIZAÇÃO (2; 0,01), RIO (4; 0,01), FNS (2; 0,009), ÍNDIOS (2; 0,009), MÉDICO (2; 0,009), FUNAI (2; 0,008), INDÍGENAS (3; 0,008), FUNDAÇÃO (3; 0,007), ATENDIMENTO (1; 0,007), FIOCRUZ (1; 0,007)	2,56
35	jul132, mai148, jan182	SIVAM (3; 0,022), AMAZÔNIA (3; 0,016), PETROBRAS (1; 0,016), REGIÃO (2; 0,014), ACORDO (2; 0,014), AVIÃO (2; 0,011), VIGILÂNCIA (3; 0,01), LBA(1; 0,009), SISTEMA (3; 0,009), ER-2 (1; 0,007)	1,92
36	jun140, nov12, out95, set25	PESQUISA (4; 0,009), ÁRVORES (3; 0,009), AMAZÔNIA (3; 0,008), FLORESTA (3; 0,007), CAPOEIRA (1; 0,006), PESQUISADORES (3; 0,005), ANOS (3; 0,005), SEMENTES (1; 0,005), IPAM (3; 0,005), RAÍZES (2; 0,004), PRODUTIVIDADE (2; 0,004), MADEIRA (2; 0,004)	2,56
37	jun146, out94, set108, abr34	GOVERNO (3; 0,023), AMAZÔNIA (4; 0,022), PROGRAMA (3; 0,01), ANO (3; 0,009), BRASIL (4; 0,008), INVESTIMENTOS (2; 0,008), PAÍSES (2; 0,008), ORÇAMENTO (2; 0,008), ESTUDO (1; 0,007), GREENPEACE(1; 0,006)	2,56
38	jun147, out48	FHC (2; 0,022), PAÍSES (2; 0,018), GREENPEACE (1; 0,017), BRASIL (2; 0,013), PAZ (1; 0,012), COLÔMBIA (1; 0,012), ENTIDADE (1; 0,011), TEMA (2; 0,011), GOVERNO (2; 0,011), REGIÃO (2; 0,01), PASTRANA (1; 0,01), PROCESSO (1; 0,01), COLOMBIANO (1; 0,01), AMAZÔNIA (1; 0,008), ONGS(1; 0,008)	1,28
39	mar161, mar162, mar37, fev175	DESMATAMENTO (4; 0,058), AMAZÔNIA (4; 0,038), MINISTÉRIO (3; 0,031), REGIÃO (4; 0,028), PROIBIÇÃO (3; 0,025), FAMILIAR (3; 0,023), MEDIDA (3; 0,023), MEIO (4; 0,023), AMBIENTE (4; 0,023), AGRICULTURA (3; 0,022), PROPRIEDADES (3; 0,02), SARNEY (3; 0,02), FILHO (3; 0,02), MINISTRO (3; 0,019), FAMÍLIAS (2; 0,017)	2,56
40	mar62, nov11, nov44, nov84	COCAÍNA (4; 0,013), REGIÃO (3; 0,012), COLÔMBIA (4; 0,012), TRÁFICO (4; 0,012), PERU (3; 0,012), FRONTEIRA (2; 0,011), AMAZÔNICA (3; 0,011), RIO (2; 0,011), NARCOTRÁFICO (3; 0,011), SOLIMÕES (2; 0,01), SURINAME (1; 0,009)	2,56

Grupos	Arquivos	Palavras relevantes	%
41	nov42, nov43, abr58	AVIÃO (3; 0,03), DROGA (3; 0,021), COCAÍNA (3; 0,019), PISTA (2; 0,015), SURINAME (2; 0,015), TRÁFICO (2; 0,011), OLIVEIRA (1; 0,011), BANCOS (2; 0,01), COLÔMBIA (2; 0,01), PILOTO (2; 0,009), CLANDESTINA (2; 0,009)	1,92
42	nov45, nov79	NARCOTRÁFICO (2; 0,011), SOLIMÕES (2; 0,011), TRAFICANTE (1; 0,011), SERRARIA (1; 0,011), MADEIRA (2; 0,01), EXTRAÇÃO (2; 0,009), GOVERNADOR (1; 0,009), ÁREA (2; 0,009), TABATINGA (2; 0,009), BISPO (1; 0,008)	1,28
43	nov85, nov86, nov89, nov9, nov41	CPI (4; 0,016), SÂMIA (4; 0,009), CURICA (4; 0,009), CAMPINAS (2; 0,007), DEPUTADO (4; 0,006), TRAFICANTE (4; 0,006), DINHEIRO(4; 0,005), SÃO (5; 0,005), FEDERAL (5; 0,005), PRISÃO (4; 0,005), PAULO (4; 0,005), ORGANIZAÇÃO (2; 0,004), NARCOTRÁFICO (4; 0,004)	3,21
44	nov90, nov91	CPI (2; 0,038), NARCOTRÁFICO (2; 0,022), MILITARES (2; 0,015), ANTITRÁFICO (2; 0,015), OPERAÇÃO (1; 0,012), EXÉRCITO (1; 0,012), AÇÃO (1; 0,012), MÃOS (1; 0,012), ARMADAS (2; 0,011), FORÇAS (2; 0,011), BC (1; 0,01), PRESIDENTE (2; 0,01)	1,28
45	out100, out15, out17, out92	MADEIRA (4; 0,019), MANEJO (4; 0,016), CERTIFICAÇÃO (4; 0,012), BRASIL (3; 0,011), FLORESTAL (4; 0,01), AMAZÔNIA (4; 0,009), FSC (4; 0,008), FLORESTAS (3; 0,008), PRODUÇÃO (3; 0,006), MERCADO (3; 0,006), SUSTENTÁVEL (3; 0,006)	2,56
46	out13, out14, jan180, jun141, nov8	AMAZÔNIA (5; 0,013), DENGUE (1; 0,011), REGIÃO (5; 0,008), MADEIRA (2; 0,007), SISTEMAS (1; 0,007), SP (2; 0,007), LIMA (1; 0,007), ANOS (4; 0,006), ÁRVORES (2; 0,006), EQUIPE (2; 0,006), M3 (1; 0,006), PRODUTOR (1; 0,005), EMBRAPA (1; 0,005)	3,21
47	set106, set107, set27, dez6	QUEIMADAS (2; 0,021), IBAMA (4; 0,018), PRISÃO (2; 0,013), MEIO (4; 0,011), AMBIENTE (4; 0,01), FOCOS (3; 0,008), PORTARIA (2; 0,008), DIAS (3; 0,008), FEDERAL (4; 0,007), SARNEY (3; 0,006), FILHO (3; 0,006), MINISTRO (3; 0,006)	2,56
48	set19, set21, set22, set24, out18	AMAZÔNIA (5; 0,014), SEMINÁRIO (5; 0,012), SOJA (3; 0,011), BIODIVERSIDADE (5; 0,009), DESENVOLVIMENTO (5; 0,007), CONSERVAÇÃO (5; 0,007), ÁREAS (3; 0,007), MAPA (3; 0,006), EIXOS (4; 0,006), GOVERNO (4; 0,006)	3,21
0		Os outros não foram agrupados	

Grau de similaridade: 10%

Grupos	Arquivos	Palavras relevantes	%
1	abr158, abr35	ESTUDO (2; 0,032), BRASIL (2; 0,032), DESMATAMENTO (1; 0,027), PUBLICADO (2; 0,022), NATURE (2; 0,022), SATÉLITE (2; 0,022), IMAGENS (2; 0,022), EUA (2; 0,022), REVISTA (2; 0,022), AMAZÔNIA (2; 0,022)	4
2	abr36, jun32, set20, abr157, abr59, fev169	FLORESTA (6; 0,022), PESQUISA (5; 0,013), AMAZÔNIA (5; 0,012), INCÊNDIOS (4; 0,007), MADEIRA (5; 0,007), EXTRAÇÃO (4; 0,006), ÁRVORES (5; 0,006), INSTITUTO (5; 0,006), ESTUDO (4; 0,006)	12
3	ago30, ago49	MCCAFFREY (2; 0,021), LEI (2; 0,02), BRASIL (2; 0,017), EUA (2; 0,013), GENERAL (2; 0,012), COLÔMBIA (2; 0,011), NARCOTRÁFICO (2; 0,011), PAÍS (2; 0,009), GOVERNO (2; 0,009), COMBATE (2; 0,009), ABATE (2; 0,008), PAÍSES (2; 0,007), DROGAS (2; 0,007), ESTADOS (1; 0,007), REGULAMENTAÇÃO (2; 0,006)	4

Grupos	Arquivos	Palavras relevantes	%
4	ago50, ago52	PRESIDENTE (2; 0,029), FHC (2; 0,02). GOVERNO (2; 0,012), MARCHA (2; 0,01). POVO (2; 0,009), BRASIL (2; 0,009), VISITA (2; 0,006), GOVERNADOR (2; 0,006), OPOSIÇÃO (2; 0,006), DEMOCRACIA (2; 0,006)	4
5	dez1, dez66	MADEIREIRAS (2; 0,032), IBAMA (2; 0,032), AMAZÔNIA (2; 0,02), INSTITUTO (2; 0,017), AMBIENTE (2; 0,017), MEIO (2; 0,017), LEVANTAMENTO (2; 0,017), ILEGAL (2; 0,014), PARÁ (2; 0,014), EMPRESAS (2; 0,014), FISCALIZAÇÃO (2; 0,011), EXTRAÇÃO (2; 0,011)	4
6	dez67, dez73	MUSEU (2; 0,038), BRITÂNICO (2; 0,02), AMAZÔNIA (2; 0,018), EXPOSIÇÃO (2; 0,017), ANOS (2; 0,015), BRASIL (2; 0,014), ARTE (2; 0,011), MOSTRA (2; 0,011), PEÇAS (2; 0,011)	4
7	fev167, fev168	BARCO (2; 0,062), AMAZÔNIA (2; 0,057), PESSOAS (2; 0,052), AFUNDOU (2; 0,047), RIO (2; 0,033), DESAPARECIDAS (2; 0,028), SELVA (2; 0,028), MADEIRA (2; 0,028)	4
8	fev170, fev171	DESMATAMENTO (2; 0,044), AUTORIZAÇÕES (2; 0,037), FUNCIONÁRIOS (2; 0,024), IBAMA (2; 0,024), AFASTADOS (2; 0,024), ARIPUANÃ (2; 0,018), AMAZÔNIA (2; 0,018), FAZENDA (2; 0,018), AMBIENTE (2; 0,015), MEIO (2; 0,015), FILHO (2; 0,013), SARNEY (2; 0,013)	4
9	fev172, fev173, nov88	BARCO (3; 0,031), NAUFRÁGIO (3; 0,027), RIO (3; 0,025), PESSOAS (3; 0,023), CORPOS (2; 0,022), ÁGUAS (2; 0,022), MADEIRA (3; 0,013), AMAZÔNIA (3; 0,013), DESAPARECIDAS (2; 0,012)	6
10	fev38, jan183	INPE (2; 0,016), ANO (2; 0,016), FOCOS (2; 0,015), QUEIMADAS (2; 0,014), INSTITUTO (2; 0,013), AMAZÔNIA (2; 0,012), ÁREA (2; 0,011), IMAGENS (2; 0,008), PONTOS (1; 0,007), REGISTROU (1; 0,007), CALOR (1; 0,007), PESQUISAS (2; 0,007)	4
11	jan64, jul128, nov47	COLÔMBIA (3; 0,032), ARMAS (1; 0,023), BRASIL (3; 0,019), FARC (2; 0,016), GUERRILHEIROS (2; 0,014), GUERRILHA (3; 0,013), AFIRMOU (3; 0,012), EXÉRCITO (1; 0,012), PAÍS (3; 0,011), PERU (3; 0,011), GOVERNO (3; 0,01), SENDERO (1; 0,01), AÇÃO(2; 0,009)	6
12	jul122, jul126	CASOS (2; 0,058), ANO (2; 0,038), SEMESTRE (2; 0,026), PRIMEIRO (2; 0,026), MALÁRIA (2; 0,026), IANOMÂMIS (1; 0,022), REGISTRADOS (2; 0,019), RORAIMA (1; 0,018), ACRE (1; 0,018), ÍNDICE (1; 0,015), DISSEMINAÇÃO (1; 0,015), CONTROLE (1; 0,015), MORTES (2; 0,013)	4
13	jul124, jul125	SAÚDE (2; 0,030), FNS (2; 0,019), ÍNDIOS (2; 0,019), MÉDICO (2; 0,018), FUNAI (2; 0,016), ATENDIMENTO (1; 0,014), INDÍGENAS (2; 0,013), INDÍGENA (2; 0,012), PAÍS (2; 0,011), GOVERNO (2; 0,011), ÍNDIO (2; 0,01), FUNDAÇÃO (2; 0,01)	4
14	jun147, out48	FHC (2; 0,022), PAÍSES (2; 0,018), GREENPEACE (1; 0,017), BRASIL (2; 0,013), PAZ (1; 0,012), COLÔMBIA (1; 0,012), ENTIDADE (1; 0,011), TEMA (2; 0,011), GOVERNO (2; 0,011), REGIÃO (2; 0,01), PASTRANA (1; 0,01), PROCESSO (1; 0,01), COLOMBIANO (1; 0,01), AMAZÔNIA (1; 0,008), ONGS (1; 0,008)	4
15	mar161, mar162, mar37, fev175	DESMATAMENTO (4; 0,058), AMAZÔNIA (4; 0,038), MINISTÉRIO (3; 0,031), REGIÃO (4; 0,028), PROIBIÇÃO (3; 0,025), FAMILIAR (3; 0,023), MEDIDA (3; 0,023), MEIO (4; 0,023), AMBIENTE (4; 0,023), AGRICULTURA (3; 0,022), PROPRIEDADES (3; 0,02), SARNEY (3; 0,02), FILHO (3; 0,02), MINISTRO (3; 0,019), FAMÍLIAS (2; 0,017), REVOGOU (3; 0,016)	8

Grupos	Arquivos	Palavras relevantes	%
16	nov95, nov89, nov9	CPI (2; 0,015), CAMPINAS (2; 0,012), CURICA (3; 0,01), ORGANIZAÇÃO (2; 0,008), RIO (2; 0,007), SÂMIA (3; 0,007), NARCOTRÁFICO (3; 0,007), FEDERAL (3; 0,006), DEPUTADO (2; 0,006), DINHEIRO (3; 0,006), QUEBRA (2; 0,006), PF (2; 0,006), SIGILO (3; 0,005), QUADRILHA (3; 0,005), TRAFICANTE (3; 0,005), COCAÍNA (2; 0,005)	6
17	out100, out17	MADEIRA (2; 0,023), MANEJO (2; 0,017), CERTIFICAÇÃO (2; 0,015), PRODUÇÃO (2; 0,009), AMAZÔNIA (2; 0,009), MUNDIAL (2; 0,008), MERCADO (2; 0,007), SETOR (2; 0,007), FSC (2; 0,007), GOVERNO (2; 0,007), BRASIL (1; 0,007), BANCO (2; 0,006), MADEIREIRAS (2; 0,006)	4
18	set106, set107, set27, dez6, set105	QUEIMADAS (3; 0,026), IBAMA (4; 0,014), MEIO (5; 0,011), PRISÃO (2; 0,01), AMBIENTE (5; 0,01), FILHO (4; 0,007), MINISTRO (4; 0,007), SARNEY (4; 0,007), FOCOS (3; 0,006), PORTARIA (2; 0,006)	10
19	set21, set19	SOJA (1; 0,017), SEMINÁRIO (2; 0,016), MAPA (1; 0,012), ÁREAS (1; 0,012), GOVERNO (2; 0,01), AMAZÔNIA (2; 0,01), EIXOS (2; 0,01), CONSERVAÇÃO (2; 0,008), RISCO (2; 0,008), REGIÃO (2; 0,008), FEDERAL (2; 0,008), BIODIVERSIDADE (2; 0,008)	4
0		Os outros não foram agrupados	

Grau de similaridade: 15%

Grupos	Arquivos	Palavras relevantes	%
1	abr158, abr35	ESTUDO (2; 0,032), BRASIL (2; 0,032), DESMATAMENTO (1; 0,027), PUBLICADO (2; 0,022), NATURE (2; 0,022), SATÉLITE (2; 0,022), IMAGENS (2; 0,022), EUA (2; 0,022), REVISTA (2; 0,022)	20
2	fev167, fev168	BARCO (2; 0,062), AMAZÔNIA (2; 0,057), PESSOAS (2; 0,052), AFUNDOU (2; 0,047), RIO (2; 0,033), DESAPARECIDAS (2; 0,028), SELVA (2; 0,028), MADEIRA (2; 0,028)	20
3	fev170, fev171	DESMATAMENTO (2; 0,044), AUTORIZAÇÕES (2; 0,037), FUNCIONÁRIOS (2; 0,024), IBAMA (2; 0,024), AFASTADOS (2; 0,024), ARIPUANÃ (2; 0,018), AMAZÔNIA (2; 0,018), FAZENDA (2; 0,018), AMBIENTE (2; 0,015), MEIO (2; 0,015), FILHO (2; 0,013), SARNEY (2; 0,013)	20
4	mar161, mar162	DESMATAMENTO (2; 0,059), AMAZÔNIA (2; 0,059), PROIBIÇÃO (2; 0,045), MINISTÉRIO (2; 0,045), FAMÍLIAS (1; 0,032), REVOGOU (2; 0,029), MEIO (2; 0,029), INSTRUÇÃO (2; 0,029), FAMILIAR (2; 0,029), REGIÃO (2; 0,029), AMBIENTE (2; 0,029), PROPRIEDADES (2; 0,029), AGRICULTURA (2; 0,029)	20
5	set106, set107	QUEIMADAS (2; 0,043), PRISÃO (2; 0,026), PORTARIA (2; 0,017), IBAMA (2; 0,015), MEIO (2; 0,013), MATO (2; 0,013), GROSSO (2; 0,013), DIAS (2; 0,013), SÃO (2; 0,011), PAULO (2; 0,011), MINISTRO (2; 0,011), FILHO (2; 0,011), SARNEY (2; 0,011), AMBIENTE (2; 0,011)	20
0		os outros não foram agrupados	

Algoritmo: “Cliques”
Período: 1999
Grau de similaridade: 0%

Grupo	Arquivos	Palavras relevantes	%
1	TODOS	AMAZÔNIA (148; 0,04), REGIÃO (94; 0,005), BRASIL (81; 0,049), GOVERNO (61; 0,003), DESMATAMENTO (24; 0,003), MADEIRA (39; 0,002), PAÍS (57; 0,002), NACIONAL (63; 0,002), FLORESTA (46; 0,002), AMBIENTE (44; 0,002), QUEIMADAS (18; 0,002), PRESIDENTE (38; 0,002), AMAZÔNICA (51; 0,002), COLÔMBIA (26; 0,001), IBAMA (24; 0,001), ... NARCOTRÁFICO (24; 0,001)	100%

Grau de similaridade: 2.5%

Grupo	Arquivos	Palavras relevantes	%
1	abr155, abr147, abr34	AMAZÔNIA (3; 0,015), FLORESTA (2; 0,008), GOYA (1; 0,008), HOTEL (1; 0,008), ESTADO (2; 0,006), ENCONTRO (1; 0,006), COBRAT (1; 0,006), BRAZTOA (1; 0,006), DEVASTAÇÃO (1; 0,006)	1,72
2	abr156, abr35	AMAZÔNIA (2; 0,056), ASSENTAMENTO (1; 0,043), RURAL (1; 0,021), RIO (1; 0,021), OCIDENTAL (1; 0,021), NOVO (1; 0,021), MODELO (1; 0,021), INVESTIR (1; 0,021), INTERAMERICANO (1; 0,021), EMBRAPA (1; 0,021), DESENVOLVIMENTO (1; 0,021)	1,15
3	abr158, abr160	BRASIL (2; 0,037), DESMATAMENTO (2; 0,037), LAMA (1; 0,019), DESIGUALDADE (1; 0,019), DALAI (1; 0,019), AMAZÔNIA (2; 0,018), ESTUDO (1; 0,018)	1,15
4	abr159, abr36, abr59, ago110, ago112, ago113, ago114, ago116	FLORESTA (3; 0,007), AMAZÔNIA (7; 0,005), EUA (4; 0,005), COLÔMBIA (3; 0,005), MADEIRA (3; 0,005), BRASIL (6; 0,005), POLÍTICA (4; 0,004), PESQUISA (2; 0,004), UNIDOS (4; 0,003), ESTADOS (3; 0,003), EXTRAÇÃO (2; 0,003)	4,6
5	abr58, ago111	OLIVEIRA (1; 0,016), CPI (1; 0,012), AVIÃO (1; 0,012), SANTARÉM-CUIABÁ (1; 0,011), PRESIDENTE (1; 0,011), MEDIDA (1; 0,011), GRUPO (1; 0,011), FHC (1; 0,011), ASFALTAR (1; 0,011)	1,15
6	ago117, ago118, ago134	AMAZÔNIA (3; 0,013), RECURSOS (2; 0,01), BRASIL (2; 0,01), MILITAR (2; 0,009), NATURAIS (1; 0,009), MUNDO (2; 0,008), FOLHA (2; 0,007), INTERVENÇÃO (2; 0,007), LIMITES (1; 0,007)	1,72
7	ago28, ago29	CLIMA (1; 0,039), REGIÃO (2; 0,037), PAÍSES (2; 0,024), AMAZÔNICA (2; 0,024), AMAZÔNIA (2; 0,024), PROJETO (1; 0,019), PLANETA (1; 0,019), PESQUISADORES (1; 0,019), LBA (1; 0,019), INGLÊS (1; 0,019), INFLUÊNCIA (1; 0,019), GRANDE (1; 0,019), EXPERIMENTO (1; 0,019)	1,15
8	ago30, ago31, ago49, ago50, ago51, ago52, ago53, dez1, dez2, dez3	BRASIL (6; 0,012), PRESIDENTE (5; 0,008), GOVERNO (6; 0,008), BRASILEIRO (4; 0,007), EUA (4; 0,006), MCCAFFREY (3; 0,006), AMAZÔNICA (0; 0,005), REGIÃO (7; 0,005), FHC (3; 0,004), COMBATE (3; 0,004), LE (2; 0,004)	5,75

Grupo	Arquivos	Palavras relevantes	%
9	dez109, dez120	AMAZÔNIA (2; 0,063), TOCANTINS (1; 0,016), SÃO (1; 0,016), RORAIMA (1; 0,016), RONDÔNIA (1; 0,016), PAULO (1; 0,016), PARÁ (1; 0,016), MATO (1; 0,016), MARANHÃO (1; 0,016), IBAMA (1; 0,016)	1,15
10	dez4, dez5, dez6, dez66, dez68, dez69, dez70, dez71	MEIO (6; 0,007), IBAMA (2; 0,006), SÃO (5; 0,006), MADEIRA (3; 0,005), REGIÃO (6; 0,005), RIO (4; 0,005), AMAZÔNIA (7; 0,005), RIOS (4; 0,005), PEIXES (4; 0,005), PARÁ (3; 0,005), FLORESTA (5; 0,005), MAUÉS (2; 0,004), GUARANÁ (1; 0,004)	4,6
11	dez65, dez67, dez73, dez75, fev166	MUSEU (2; 0,015), AMAZÔNIA (5; 0,011), CASOS (2; 0,008), ANOS (3; 0,008), BRITÂNICO (2; 0,008), GOVERNO (1; 0,007), FEDERAIS (1; 0,007), VENDA (1; 0,007), EXPOSIÇÃO (2; 0,007)	2,87
12	dez76. fev169, fev171, fev173, fev177, fev38	DESMATAMENTO (4; 0,012), AMAZÔNIA (6; 0,011), FILHO (5; 0,008), MINISTRO (3; 0,008), SARNEY (3; 0,008), ÁREA (4; 0,007), AUTORIZAÇÕES (2; 0,007), INSTITUTO (4; 0,006), INPE (4; 0,006)	3,45
13	fev167, fev168, fev172	BARCO (3; 0,05), AMAZÔNIA (3; 0,047), PESSOAS (3; 0,044), RIO (3; 0,031), AFUNDOU (2; 0,031), DESAPARECIDAS (3; 0,028), MADEIRA (3; 0,028), CORPOS (2; 0,021), SELVA (2; 0,019)	1,72
14	fev170, fev175, fev176	DESMATAMENTO (2; 0,037), AMAZÔNIA (3; 0,031), INSTITUTO (3; 0,028), AUTORIZAÇÕES (2; 0,019), AUMENTO (3; 0,019), INPE (3; 0,019), DESMATADAS (1; 0,018), MINISTRO (2; 0,017), ESPACIAIS (2; 0,016), PESQUISAS (2; 0,016)	1,72
15	fev174, jul122	AMAZÔNIA (2; 0,051), MERCÚRIO (1; 0,045), INTOXICAÇÃO (1; 0,045), CASOS (1; 0,043), ANO (1; 0,03), OBSERVADOS (1; 0,022), NEW (1; 0,022), NEUROLÓGICOS (1; 0,022), MUSCULAR (1; 0,022), SINTOMAS (1; 0,022), SCIENTIST (1; 0,022)	1,15
16	jan178, jan182, jan183, jan64, jul125, jul126	AMAZÔNIA (4; 0,013), ARMAS (1; 0,011), BRASIL (4; 0,009), NACIONAL (5; 0,008), PAÍS (4; 0,008), COLÔMBIA (1; 0,007), ANO (3; 0,007), IANOMÂMIS (1; 0,007), MÉDICO (2; 0,006), AVIÃO (3; 0,006), CASOS (2; 0,005), CONTROLE (2; 0,005), FUNAI (1; 0,005), SAÚDE (1; 0,005)	3,45
17	jan179, jan181, jan40, jul123, jul124, jul127, jul129, jul131, jul132, jul133, jul57, jul135, jul136	AMAZÔNIA (11; 0,011), REGIÃO (7; 0,006), ACORDO (3; 0,004), ANOS (10; 0,003), AMBIENTE (5; 0,003), AMBIENTAL (6; 0,003), SIVAM (1; 0,003), PETROBRAS (1; 0,003), MEIO (7;0,003), ÁREA (8;0,003), MALÁRIA (1;0,003), AFIRMOU (4;0,003), COMUNIDADE (4;0,003)	7,47
18	jan180, jan39	DENGUE (1; 0,027), INFRA-ESTRUTURA (1; 0,019), AMAZÔNIA (2; 0,018), REUNIÃO (1; 0,014), GOVERNADOR (1; 0,014), FHC (1; 0,014), ALMIR (1; 0,014), SP (1; 0,013), DOENTE (1; 0,013), RENEGOCIAÇÃO (1; 0,009), GOVERNO (1; 0,009), GOVERNADORES (1; 0,009)	1,15

Grupo	Arquivos	Palavras relevantes	%
19	jul121, jul128, jul130, jun138, jun140, jun141, jun142	AMAZÔNIA (7; 0,01), PESQUISA (3; 0,006), EMBRAPA (2; 0,005), SISTEMAS (1; 0,005), BOSQUE (1; 0,005), CAPOEIRA (2; 0,005), MATO (2; 0,004), GROSSO (2; 0,004), SENDERO (1; 0,004)	4,02
20	jun137, jum143	AMAZÔNIA (2; 0,027), PRESIDENTE (1; 0,024), ARIRANHA (1; 0,017), REMESSA (1; 0,016), RIO (1; 0,016), FHC (1; 0,016), ESTRANGEIRA (1; 0,016), DEFENDE (1; 0,016)	1,15
21	jun139, mar162	AMAZÔNIA (2; 0,098), OPÇÕES (1; 0,066), DESCOBRE (1; 0,066), QUEIMADAS (1; 0,066), PROJETO (1; 0,033), TRITURAR (1; 0,033), EMBRAPA (1; 0,033), QUEIMÁ-LA (1; 0,033), CAPOEIRA (1; 0,033), PROIBIÇÃO (1; 0,032), MINISTÉRIO (1; 0,032), FAMÍLIAS (1; 0,032), DESMATAMENTO (1; 0,032), REVOGOU (1; 0,016)	1,15
22	jun144, jun145, jun146, jun147, jun32	AMAZÔNIA (5; 0,031), GREENPEACE (2; 0,012), FHC (1; 0,008), ENTIDADE (2; 0,008), INPE (2; 0,007), AVIÃO (1; 0,007), FLORESTA (3; 0,007), GOVERNO (3; 0,006), BRASIL (4; 0,006)	1,72
23	mai148, mai149, mai150	SOJA (1; 0,012), DARLY (1; 0,01), ANOS (1; 0,008), REGIME (1; 0,007), JUIZ (1; 0,007), OPÇÕES (1; 0,006), ERJ-145 (1; 0,006), EMBRAER (1; 0,006), AÉREA (1; 0,006), AIRWAYS (1; 0,006), AMAZÔNIA (3; 0,006), REGIÃO (2; 0,006), MORTE (1; 0,006)	1,72
24	mai151, mai154	SENAD (1; 0,02), NARCOTRÁFICO (1; 0,02), ESPAÇO (1; 0,02), CHELOTTI (1; 0,02), AÉREO (1; 0,02), EX-DIRETOR (1; 0,013), TRÁFICO (1; 0,013), SECRETARIA (1; 0,013), AMAZÔNIA (2; 0,009)	1,15
25	mai152, mai153, mai33, mar163, mar164, mar165, mar62, nov10, nov12, nov41, nov42, nov45	POLÍCIA (7; 0,006), ÍNDIOS (5; 0,005), RIO (5; 0,005), REGIÃO (8; 0,004), AMAZÔNIA (8; 0,003), NACIONAL (7; 0,003), FEDERAL (6; 0,003), COCAÍNA (4; 0,003), AMAZÔNICA (6; 0,002), TRÁFICO (3; 0,002)	6,9
26	mar161, mar37, mar60, mar63, nov47, nov7, nov79	DESMATAMENTO (; 0,014), REGIÃO (; 0,014), MINISTÉRIO (; 0,01), AMAZÔNIA (; 0,01), FILHO (; 0,01), GOVERNO (; 0,01), MADEIRA (; 0,01), FAMILIAR (; 0,008), SARNEY (; 0,008), AGRICULTURA (; 0,008), TRANSPORTE (; 0,007), MINISTRO (; 0,007), MEIO (; 0,007), PROPRIEDADES (; 0,007)	4,02
27	mar61, nov11, nov43, nov44	COCAÍNA (3; 0,016), AVIÃO (1; 0,014), COLÔMBIA (3; 0,014), AMAZÔNICA (4; 0,013), TRÁFICO (4; 0,013), FRONTEIRA (2; 0,013), GOVERNO (2; 0,012), DROGA (3; 0,012), REGIÃO (3; 0,011), NARCOTRÁFICO (2; 0,01), PERU (2; 0,008)	2,3
28	nov115, nov46, nov78, nov80, nov81, nov82, nov83, nov84, nov85	SÃO (6; 0,006), FAZENDA (2; 0,006), PAÍS (4; 0,006), AMAZÔNIA (7; 0,005), VÁRZEA (2; 0,005), NARCOTRÁFICO (4; 0,004), SURINAME (1; 0,004), ILHA (2; 0,004), PAULO (3; 0,004), BRASIL (4; 0,004), CPI (2; 0,003)	5,17

Grupo	Arquivos	Palavras relevantes	%
29	nov77, nov8, nov86, nov89, nov9	CPI (2; 0,009), RIO (3; 0,008), LIMA (1; 0,007), SÂMIA (3; 0,007), ANOS (4; 0,005), AMAZÔNIA (5; 0,005), AFIRMOU (2; 0,005), PLANTAS (1; 0,005), TRAFICANTE (3; 0,005), DINHEIRO (3; 0,005), ACRE (3; 0,004), ORGANIZAÇÃ (2; 0,004)	2,87
30	nov87, nov88, nov91, out100, out101, out13, out14, out15, out16, out17	MADEIRA (8; 0,011), AMAZÔNIA (10; 0,009), BARCO (2; 0,007), SÃO (8; 0,006), MANEJO (5; 0,006), PESSOAS (5; 0,005), ANOS (6; 0,004), CERTIFICAÇÃO (3; 0,004), BRASIL (3; 0,004), FLORESTAL (5; 0,004), FLORESTA (4; 0,004)	5,75
31	nov90, out48, out94, out98, out99, set105, set106, set107	QUEIMADAS (3; 0,016), GOVERNO (6; 0,01), AMAZÔNIA (7; 0,008), PAÍSES (3; 0,008), MEIO (5; 0,007), AMBIENTE (5; 0,007), MINISTRO (7; 0,007), PRESIDENTE (6; 0,006), PRISÃO (2; 0,006), BRASIL (5; 0,006), SÃO (6; 0,005), MALÁRIA (1; 0,005), CASOS (1; 0,005)	4,6
32	out18, out92, out93, out95, out96, out97, set102, set104	AMAZÔNIA (7; 0,008), FLORESTA (4; 0,007), BRASIL (4; 0,006), ORGANIZAÇÃO (1; 0,004), TRABALHO (3; 0,004), REGIÃO (4; 0,004), ARGENTINA (1; 0,003), MADEIRA (4; 0,003)	4,6
33	set108, set19, set20, set21, set22, set24, set26, set27	AMAZÔNIA (8; 0,014), GOVERNO (6; 0,01), ÁREA (6; 0,007), SOJA (2; 0,006), SEMINÁRIO (4; 0,006), ORÇAMENTO (2; 0,005), REGIÃO (6; 0,005), BIODIVERSIDADE (4; 0,005), DESENVOLVIMENTO (5; 0,005), DIAS (4; 0,005), ESTUDO (3; 0,004)	4,6
0	set25		

Grau de similaridade: 5%

Grupo	Arquivos	Palavras relevantes	%
1	ab155, ago122	AMAZÔNIA (2; 0,015), HOTEL (1; 0,012), GOYA (1; 0,012), MÓVEIS (1; 0,011), MADEIRA (1; 0,009), MELIÁ (1; 0,008), FESTIVAL (1; 0,008)	1,21
2	abr156, jun139	AMAZÔNIA (2; 0,11), QUEIMADAS (1; 0,066), OPÇÕES (1; 0,066), DESCOBRE (1; 0,066), EMBRAPA (2; 0,055), ASSENTAMENTO (1; 0,043), PROJETO (1; 0,033), TRITURAR (1; 0,033), QUEIMÁ-LA (1; 0,033), CAPOEIRA (1; 0,033), SUSTENTÁVEL (1; 0,021), DESENVOLVIMENTO (1; 0,021), BID (1; 0,021)	1,21
3	abr157, abr35, abr36, abr59	FLORESTA (4; 0,025), DEVASTAÇÃO (4; 0,013), AMAZÔNIA (3; 0,013), PESQUISADORES (4; 0,012), ESTUDO (3; 0,012), NATURE (4; 0,011), BRASIL (3; 0,01), SATÉLITE (3; 0,01), MADEIREIRAS (4; 0,01), REVISTA (4; 0,01), AÇÃO (4; 0,01)	2,42
4	abr158, abr160	BRASIL (2; 0,037), DESMATAMENTO (2; 0,037), LAMA (1; 0,019), DESIGUALDADE (1; 0,019), DALAI (1; 0,019), AMAZÔNIA (2; 0,018), ESTUDO (1; 0,018)	1,21
5	abr159, ago113, ago116, ago117	COLÔMBIA (4; 0,01), BRASIL (5; 0,008), EUA (3; 0,006), POLÍTICA (4; 0,006), UNIDOS (4; 0,006), ESTADOS (3; 0,005), GOVERNO (4; 0,005), MILITAR (3; 0,005), PAÍS (5; 0,005), INTERVENÇÃO (3; 0,004), AMAZÔNIA (5; 0,003)	2,42

Grupo	Arquivos	Palavras relevantes	%
6	abr34, ago30, ago53, dez6	LEI (1; 0,008), IBAMA (2; 0,007), ESTADO (3; 0,007), BRASIL (3; 0,007), FUMAÇA (1; 0,006), ANO (3; 0,006), PRESIDENTE (3; 0,006), MADEIRA (1; 0,005), GREENPEACE (1; 0,005), PARÁ (2; 0,005)	2,42
7	abr58, jan64	ARMAS (1; 0,035), COLÔMBIA (1; 0,023), OLIVEIRA (1; 0,016), AVIÃO (2; 0,015), FARC (1; 0,014), TRÁFICO (2; 0,014), POLÍCIA (2; 0,013), FEDERAL (2; 0,013), CPI (1; 0,012), BRASIL (1; 0,011)	1,21
8	ago110, jan181, jun135, out100, out18, set24, set27	AMAZÔNIA (7; 0,011), MADEIRA (4; 0,006), BRASIL (5; 0,005), NACIONAL (6; 0,004), ANOS (6; 0,003), AMBIENTE (3; 0,003), FLORESTA (3; 0,003), BIODIVERSIDADE (2; 0,003)	4,24
9	ago111, jan39	FHC (2; 0,025), INFRA-ESTRUTURA (1; 0,019), ESTADO (2; 0,015), AMAZÔNIA (2; 0,015), REUNIÃO (1; 0,014), GOVERNADOR (1; 0,014), ALMIR (1; 0,014), SANTARÉM-CUIABÁ (1; 0,011), PRESIDENTE (1; 0,011), ASFALTAR (1; 0,011)	1,21
10	ago118, ago50	PRESIDENTE (; 0,011), LIMITES (; 0,01), FHC (; 0,01), DESENVOLVIMENTO (; 0,008), MUNDO (; 0,008), SUSTENTABILIDADE (; 0,006), GOVERNO (; 0,005), DESMATAMENTO (; 0,005), BRASIL (; 0,005)	1,21
11	ago134, jun136, out48, out94	AMAZÔNIA (3; 0,029), PAÍSES (4; 0,018), GOVERNO (3; 0,013), BRASIL (4; 0,012), INTERVENÇÃO (2; 0,007), RECURSOS (1; 0,006), NATURAIS (1; 0,006), TROPICAIS (2; 0,006), PROGRAMA (1; 0,006), COLÔMBIA (1; 0,006), PAZ (1; 0,006)	2,42
12	ago28, ago51, dez1	REGIÃO (3; 0,019), PROGRAMA (2; 0,017), BRASILEIRO (2; 0,015), IBAMA (2; 0,015), EUA (2; 0,014), ESTADOS (2; 0,012), MADEIREIRAS (1; 0,012), QUEIMADAS (1; 0,011), BRASIL (1; 0,011), GOVERNO (1; 0,011), AMAZÔNIA (2; 0,011)	1,82
13	ago29, jun137	CLIMA (1; 0,039), AMAZÔNIA (2; 0,031), PAÍSES (1; 0,019), LBA (1; 0,019), INGLÊS (1; 0,019), INFLUÊNCIA (1; 0,019), GRANDE (1; 0,019), EXPERIMENTO (1; 0,019), ESCALA (1; 0,019), REGIÃO (1; 0,019)	1,21
14	ago31, ago49, ago47	COLÔMBIA (3; 0,024), BRASIL (3; 0,021), FRONTEIRA (2; 0,012), EXÉRCITO (1; 0,012), MCCAFFREY (2; 0,012), GUERRILHEIROS (3; 0,011), FARC (2; 0,01), GUERRILHA (2; 0,01), PAÍS (2; 0,008)	1,82
15	ago52, jan40, jul125	INDÍGENAS (3; 0,017), PRESIDENTE (1; 0,011), TERRAS (3; 0,011), ÍNDIOS (2; 0,01), SAÚDE (1; 0,01), FUNAI (1; 0,01), PAÍS (2; 0,008), MÉDICO (1; 0,008)	1,82
16	dez109, fev170	DESMATAMENTO (2; 0,039), IBAMA (2; 0,034), AMAZÔNIA (2; 0,025), ÁREA (2; 0,02), AUTORIZAÇÃO (2; 0,02), FUNCIONÁRIOS (1; 0,018), AUTORIZAÇÕES (1; 0,018), AFASTADOS (1; 0,018)	1,21
17	dez2, dez68	CHEIA (1; 0,018), ANOS (2; 0,016), LAVADAS (1; 0,014), ILHAS (1; 0,014), ÉPOCA (1; 0,014), PARQUE (1; 0,009), ZOOLÓGICO (1; 0,009), ÁREA (1; 0,009), FLORESTA (2; 0,009)	1,21
18	dez3, fev173	RIO (2; 0,02), TERRA (1; 0,014), POLÍCIA (1; 0,01), EMBARCAÇÃO (1; 0,01), MANAUS (2; 0,01), BARCO (2; 0,01), ÍNDIOS (1; 0,009), FARINHA (1; 0,009), MILITAR (1; 0,007)	1,21
19	dez4, dez70, dez75	MAUÉS (2; 0,012), GUARANÁ (1; 0,012), LENDA (2; 0,01), RIOS (2; 0,008), PESCADORES (1; 0,008), ÍNDIOS (3; 0,008), CRIANÇA (2; 0,008), SUICÍDIOS (1; 0,008), FUNAI (1; 0,008), BOTO (1; 0,007), FLORESTA (3; 0,007)	1,82

Grupo	Arquivos	Palavras relevantes	%
20	dez5, mai33, mar165	ÍNDIO (3; 0,01), ÍNDIOS (3; 0,009), SÃO (3; 0,007), ALDEIA (3; 0,006), FLORESTA (3; 0,004), TICUNAS (1; 0,004), RIO (2; 0,004), AMAZÔNIA (3; 0,004), CRIANÇAS (2; 0,004), INDÍGENA (2; 0,003), CULTURA (3; 0,003), INTERDIÇÃO (1; 0,003), FUNAI (1; 0,003)	1,82
21	dez65, mar60	GOVERNO (2; 0,026), ELETRONORTE (2; 0,024), VENDA (2; 0,022), PRIVATIZAÇÃO (2; 0,02), FEDERAIS (1; 0,018), SETOR (2; 0,014), CHESF (2; 0,014), ANO (2; 0,014), ENERGÉTICAS (1; 0,012), ENERGIA (2; 0,012), FURNAS (1; 0,012)	1,21
22	dez66, dez71, fev166	CASOS (2; 0,0135), ANO (1; 0,0125), PARÁ (2; 0,0114), LEVANTAMENTO (3; 0,0097), MUNICÍPIOS (3; 0,0092), ESTADO (3; 0,0091), IBAMA (1; 0,0091), MADEIREIRAS (1; 0,0091), AMAZÔNIA (3; 0,0091), MALÁRIA (1; 0,0088)	1,82
23	dez67, dez73	MUSEU (2; 0,038), BRITÂNICO (2; 0,02), AMAZÔNIA (2; 0,018), EXPOSIÇÃO (2; 0,017), ANOS (2; 0,015), BRASIL (2; 0,014), ARTE (2; 0,011), MOSTRA (2; 0,011), PEÇAS (2; 0,011)	1,21
24	dez69, jun141	SISTEMAS (1; 0,018), PRODUTOR (1; 0,014), EMBRAPA (1; 0,014), AMAZÔNIA (2; 0,013), EXEMPLO (2; 0,013), PEIXES (1; 0,013), MEIO (1; 0,013), SILVA (1; 0,009), PLANTOU (1; 0,009), PESQUISADOR (1; 0,009), PASTO (1; 0,009)	1,21
25	dez76, fez169, nov45, nov78	NARCOTRÁFICO (2; 0,008), ÁREA (3; 0,007), PAÍS (3; 0,006), FLORESTA (3; 0,006), MADEIRA (3; 0,006), CARVOEIROS (1; 0,005), BRASIL (2; 0,005), ESTADO (3; 0,004), GOVERNADOR (1; 0,004), AMAZÔNIA (3; 0,004)	2,42
26	fev167, fev168, fev172	BARCO (3; 0,05), AMAZÔNIA (3; 0,047), PESSOAS (3; 0,044), RIO (3; 0,031), AFUNDOU (2; 0,031), DESAPARECIDAS (3; 0,028), MADEIRA (3; 0,028), CORPOS (2; 0,021), SELVA (2; 0,019), ÁGUAS (1; 0,018), NAUFRÁGIO (1; 0,018)	1,82
27	fev171, fev175, fev177	DESMATAMENTO (3; 0,044), MINISTRO (3; 0,03), SARNEY (3; 0,021), FILHO (3; 0,02), AUTORIZAÇÕES (2; 0,019), REGIÃO (3; 0,019), AMAZÔNIA (3; 0,017), MEDIDA (2; 0,016), MEIO (3; 0,015), AMBIENTE (3; 0,014), INPE (3; 0,013)	1,82
28	fev176, fev38	AMAZÔNIA (2; 0,036), INSTITUTO (2; 0,035), DESMATADAS (1; 0,027), ÁREAS (2; 0,02), ANO (2; 0,02), INPE (2; 0,02), PESQUISAS (2; 0,018), ANOS (2; 0,018), QUEDA (2; 0,016), CRESCEU (2; 0,016), AUMENTO (2; 0,016), LEVANTAMENTO (2; 0,016)	1,21
29	jan179, jul123, jul131	AMBIENTAL (3; 0,01), AMBIENTE (3; 0,008), VIDA (2; 0,006), SUSTENTÁVEL (2; 0,005), FUTURO (1; 0,005), GESTÃO (1; 0,005), GRANDES (3; 0,004), MODERNIDADE (1; 0,004), RECURSOS (3; 0,004), MEIO (3; 0,004), MODERNIZAÇÃO (1; 0,004), CONTRADIÇÃO (1; 0,004)	1,82
30	jan180, jan183, jul127	DENGUE (1; 0,018), ANO (3; 0,017), AMAZÔNIA (3; 0,015), REGIÃO (3; 0,012), EXÉRCITO (1; 0,009), SP (1; 0,009), DOENTE (1; 0,009), ÁREA (2; 0,009), FOCOS (1; 0,008), INSTITUTO (3; 0,008), DESMATAMENTO (1; 0,007)	1,82
31	jan182, jul132, out13	SIVAM (2; 0,02), AMAZÔNIA (3; 0,017), PETROBRAS (1; 0,016), REGIÃO (2; 0,013), ACORDO (1; 0,01), AVIÃO (1; 0,009), LBA (1; 0,009), IMAZON (1; 0,007), PARAGOMINAS (1; 0,007), ANOS (3; 0,007), SISTEMA (2; 0,007), ER-2 (1; 0,007), PESQUISAS (1; 0,005)	1,82

Grupo	Arquivos	Palavras relevantes	%
32	jul121, mai149	MATO (2; 0,02), GROSSO (2; 0,02), SOJA (1; 0,019), CAMINHÕES (2; 0,013), COMBUSTÍVEIS (1; 0,012), INTERIOR (1; 0,012), NORTE (2; 0,012), REGIÃO (2; 0,009), PROBLEMAS (1; 0,008), COMBUSTÍVEL (1; 0,008), LITROS (1; 0,008), ABASTECIMENTO (1; 0,008)	1,21
33	jul122, jul126, out99	CASOS (3; 0,053), ANO (3; 0,033), MALÁRIA (3; 0,032), SEMESTRE (3; 0,02), PRIMEIRO (3; 0,019), REGISTRADOS (3; 0,016), IANOMÂMIS (1; 0,015), ACRE (2; 0,015), CONTROLE (2; 0,011), DADOS (3; 0,011), AUMENTO (2; 0,011), NÚMERO (2; 0,011), DISSEMINAÇÃO (1; 0,01)	1,82
34	jul124, jul129, jul133	MALÁRIA (1; 0,014), SAÚDE (3; 0,012), VIVAX (1; 0,01), COMUNIDADE (1; 0,01), MANAUS (3; 0,009), ATENDIMENTO (1; 0,009), ÍNDIOS (2; 0,008), FNS (1; 0,007), INDÍGENA (2; 0,007), AMAZÔNIA (2; 0,006)	1,82
35	jul128, mar62, nov10, nov11	POLÍCIA (3; 0,016), PERU (4; 0,014), TRÁFICO (3; 0,011), RIO (1; 0,01), COLÔMBIA (4; 0,009), BRASIL (3; 0,009), FEDERAL (3; 0,008), COCAÍNA (3; 0,008), REGIÃO (3; 0,008), SENDERO (1; 0,007)	2,42
36	jul57, jun142, jun32	PESQUISA (3; 0,015), FLORESTA (2; 0,01), FIOCRUZ (1; 0,009), INCÊNDIOS (1; 0,008), ACORDO (2; 0,007), AMAZÔNIA (2; 0,007), BIOTECNOLOGIA (1; 0,006), EXTRACTA (1; 0,006), GLAXO (1; 0,006), PAÍS (1; 0,006), ACIDENTAIS (1; 0,006)	1,82
37	jun138, set21	SOJA (1; 0,017), FNAC (1; 0,014), AMAZÔNIA (2; 0,01), SÃO (2; 0,008), EIXOS (1; 0,008), DESENVOLVIMENTO (2; 0,007), BRASIL (2; 0,007), PAULO (2; 0,006), PPA (1; 0,006), SEMINÁRIO (1; 0,006)	1,21
38	jun140, jun145, jun146	AMAZÔNIA (3; 0,026), GREENPEACE (1; 0,009), CAPOEIRA (1; 0,008), MEIO (1; 0,007), AMBIENTE (1; 0,007), CAMPANHA (2; 0,006), SITE (1; 0,006), PESQUISA (1; 0,006), MADEIREIRAS (1; 0,005)	1,82
39	jun143, jun147	FHC (2; 0,036), PRESIDENTE (2; 0,03), AMAZÔNIA (2; 0,024), PAÍS (2; 0,021), RIO (2; 0,019), GREENPEACE (1; 0,017), DEFENDE (1; 0,016), REMESSA (1; 0,016), ESTRANGEIRA (1; 0,016), ENTIDADE (1; 0,011)	1,21
40	mai150, mar164	ANOS (2; 0,018), DARLY (1; 0,015), ÍNDIOS (1; 0,012), CEGUEIRA (1; 0,012), JUIZ (1; 0,011), REGIME (1; 0,011), DOENÇA (1; 0,01), TRACOMA (1; 0,01), FUNDAÇÃO (2; 0,009), MORTE (1; 0,009), GABRIEL (1; 0,007), ONCOCERCOSE (1; 0,007)	1,21
41	mai151, nov86	SENAD (2; 0,023), AÉREO (2; 0,023), ESPAÇO (1; 0,02), CHELOTTI (1; 0,02), NARCOTRÁFICO (1; 0,02), CPI (2; 0,017), SECRETARIA (2; 0,016), TRÁFICO (2; 0,016), USA (1; 0,013), EX-DIRETOR (1; 0,013), SÂMIA (1; 0,012), DEPOIMENTO (2; 0,011), AMAZÔNIA (2; 0,011)	1,21
42	mai153, mar163	RECURSOS (1; 0,011), BRASIL (2; 0,011), DESENVOLVIMENTO (2; 0,01), PAÍS (2; 0,009), AMBIENTAL (2; 0,007), SIVAM (1; 0,007), CONTRATO (1; 0,007), NACIONAL (2; 0,006), MEIO (1; 0,005), PROGRAMA (1; 0,005), ACORDO (1; 0,005), NUCLEAR (1; 0,005), TECNOLOGIA (1; 0,005)	1,21
43	mai154, nov91	CPI (2; 0,02), BC (1; 0,01), NARCOTRÁFICO (1; 0,009), BANCO (2; 0,008), BANCÁRIO (1; 0,008), REPÚBLICA (1; 0,006), GOVERNO (1; 0,006), EMPRESAS (2; 0,006), FEDERAL (2; 0,006), INFORMAÇÕES (1; 0,006), FUNDO (1; 0,005), VALE (1; 0,005), BNDES (1; 0,005)	1,21

Grupo	Arquivos	Palavras relevantes	%
44	mar161, mar162, mar37	DESMATAMENTO (3; 0,055), AMAZÔNIA (3; 0,043), MINISTÉRIO (3; 0,042), PROIBIÇÃO (3; 0,034), FAMILIAR (3; 0,031), AGRICULTURA (3; 0,029), PROPRIEDADES (3; 0,027), AMBIENTE (3; 0,023), REGIÃO (3; 0,023), MEIO (3; 0,023), FAMÍLIAS (2; 0,023), INSTRUÇÃO (3; 0,021), REVOGOU (3; 0,021), DIAS (3; 0,021), SUSPENDEU (3; 0,021), FILHO (2; 0,02)	1,82
45	mar61, nov90, set26	GOVERNO (2; 0,016), PRESIDENTE (3; 0,016), MINISTRO (2; 0,014), NARCOTRÁFICO (3; 0,013), CARDOSO (3; 0,012), CPI (1; 0,012), ANTITRÁFICO (2; 0,011), AMAZÔNIA (2; 0,01), DEFESA (1; 0,01), PREVENÇÃO (1; 0,009), FORÇA-TAREFA (1; 0,009), NACIONAL (1; 0,009), HENRIQUE (3; 0,009), FERNANDO (3; 0,009)	1,82
46	mar63, set106, set107	QUEIMADAS (2; 0,028), PRISÃO (2; 0,017), GOVERNO (2; 0,015), FEDERAL (3; 0,012), MINISTRO (3; 0,012), PORTARIA (2; 0,011), GOVERNADORES (1; 0,01), SUFRAMA (1; 0,01), REGIÃO (3; 0,01), IBAMA (2; 0,01)	1,82
47	nov115, nov88	FAZENDA (1; 0,027), BARCO (1; 0,026), SÃO (2; 0,017), PAULO (1; 0,013), PESSOAS (1; 0,013), CAUSA (1; 0,013), SUPERLOTAÇÃO (1; 0,013), ACIDENTE (1; 0,013)	1,21
48	nov12, nov7	PLANTAS (1; 0,01), FLORESTA (1; 0,01), PROMOTOR (1; 0,008), VALSTAR (1; 0,008), HOLANDESES (1; 0,008), EMPRESA (1; 0,008), ESPÉCIES (2; 0,008), BRASILEIRO (1; 0,006), BIOPIRATARIA (1; 0,006), RECURSOS (2; 0,006)	1,21
49	nov41, nov84, nov85, nov89, nov9	CPI (4; 0,015), SURINAME (2; 0,008), CAMPINAS (2; 0,007), CURICA (3; 0,006), NARCOTRÁFICO (5; 0,006), FEDERAL (5; 0,005), BOUTERSE (1; 0,005), DEPUTADO (4; 0,005), PRISÃO (4; 0,005), PAÍS (3; 0,005), RIO (3; 0,005), PF (3; 0,005)	3,03
50	nov42, nov43	AVIÃO (2; 0,032), DROGA (2; 0,024), PISTA (2; 0,023), SURINAME (2; 0,022), COCAÍNA (2; 0,021), BANCOS (2; 0,015), COLÔMBIA (2; 0,015), PILOTO (2; 0,014), CLANDESTINA (2; 0,014)	1,21
51	nov44, nov79	SOLIMÕES (2; 0,023), FRONTEIRA (1; 0,021), ALTO (2; 0,02), REGIÃO (2; 0,018), NARCOTRÁFICO (2; 0,018), COCAÍNA (2; 0,017), SERRARIA (2; 0,016), TRAFICANTE (2; 0,016), AMAZÔNICA (1; 0,016), BISPO (2; 0,014)	1,21
52	nov77, out16	FLORESTA (1; 0,008), SECA (1; 0,007), BROWN (1; 0,007), FOGO (1; 0,006), AMAZÔNIA (2; 0,006), MUNDO (1; 0,006), ÁGUA (2; 0,005)	1,21
53	nov8, out101	LIMA (1; 0,018), RIO (2; 0,016), PLANTAS (1; 0,013), AFIRMOU (1; 0,013), EMPRESÁRIO (1; 0,009), PORTUGUÊS (1; 0,009), CLARO (1; 0,009), PASSEANDO (1; 0,009), ACRE (1; 0,009), HOLANDESES (1; 0,009), PESCA (1; 0,007)	1,21
54	nov80, nov81	VÁRZEA (2; 0,023), ILHA (2; 0,019), TERRAS (2; 0,013), PROJETO (2; 0,012), LAGOS (1; 0,011), AMAZONAS (1; 0,011), PESCADORES (2; 0,01), SANTARÉM (2; 0,008)	1,21
55	nov82, out15	BRASIL (2; 0,023), PAÍS (2; 0,018), MUNDO (2; 0,013), AMAZÔNIA (2; 0,013), IMAFLORA (1; 0,011), FLORESTAS (1; 0,011), FLORESTAL (1; 0,011), EMBAIXADOR (1; 0,01), SELO (1; 0,007), FSC (1; 0,007), MANEJO (1; 0,007), MADEIRA (1; 0,007), EMPRESAS (1; 0,007), CERTIFICAÇÃO (1; 0,007), CERTIFICADORES (1; 0,007), CERTIFICADO (1; 0,007)	1,21

Grupo	Arquivos	Palavras relevantes	%
56	nov83, out17, out92	MANEJO (2; 0,01), MADEIRA (3; 0,009), FOGO (2; 0,008), SUSTENTÁVEL (2; 0,007), MERCADO (2; 0,006), FLORESTA (2; 0,006), EXPLORAÇÃO (2; 0,006), CERTIFICAÇÃO (2; 0,006), ÁRVORES (2; 0,005), AMAZÔNIA (3; 0,005)	1,82
57	out14, set20, set25	AMAZÔNIA (3; 0,015), ÁRVORES (2; 0,013), PESQUISA (2; 0,012), EQUIPE (2; 0,008), MADEIRA (2; 0,008), IMAZON (1; 0,008), SEMENTES (1; 0,007), FLORESTA (3; 0,007), IPAM (1; 0,007)	1,82
58	out95, set104	FLORESTA (2; 0,017), REGIÃO (1; 0,013), AMAZÔNIA (1; 0,013), MOUTINHO (2; 0,012), RAÍZES (1; 0,007), POÇOS (1; 0,007), TERRA (2; 0,007), IPAM (2; 0,007), GÁS (2; 0,007), CARBÔNICO (2; 0,006), RESSECAMENTO (1; 0,005)	1,21
59	out96, set102	BRASIL (2; 0,017), ARGENTINA (1; 0,013), POLÍTICA (2; 0,009), RELAÇÕES (2; 0,008), ANTOLOGIA (1; 0,007), RIVALIDADE (1; 0,006), LIVRO (1; 0,006), TEXTOS (1; 0,006)	1,21
60	out97, out98	ORGANIZAÇÃO (1; 0,019), BATERIAS (1; 0,015), MEIO (2; 0,013), RIO (1; 0,013), TRABALHO (1; 0,013), FHC (1; 0,012), SANEAMENTO (1; 0,01), PRESIDENTE (1; 0,01), AMBIENTE (1; 0,01), COMUNIDADES (1; 0,009), SAÚDE (2; 0,009), BRASIL (2; 0,009), AMAZÔNIA (2; 0,009)	1,21
61	set105, set108	GOVERNO (2; 0,027), QUEIMADAS (1; 0,023), ANO (2; 0,014), ORÇAMENTO (1; 0,014), INVESTIMENTOS (1; 0,014), ESTUDO (1; 0,014), NACIONAL (2; 0,012), PROGRAMA (2; 0,01), AMAZÔNIA (2; 0,01)	1,21
62	set19, set22	AMAZÔNIA (2; 0,016), ÁREA (2; 0,014), SEMINÁRIO (2; 0,014), DIAS (2; 0,012), ÁREAS (1; 0,012), MAPA (1; 0,012), CONSERVAÇÃO (2; 0,01), GOVERNO (2; 0,01), TECNOLOGIA (1; 0,008), SOJA (1; 0,008), PRODUÇÃO (1; 0,008), LIMITAÇÕES (1; 0,008)	1,21
0			

Grau de similaridade: 10%

Grupo	Arquivos	Palavras relevantes	%
1	abr157, abr36	FLORESTA (2; 0,022), AMAZÔNIA (2; 0,012), EXTRAÇÃO (2; 0,011), PESQUISA (2; 0,011), DEVASTAÇÃO (2; 0,011), PESQUISADORES (2; 0,009), MADEIRA (1; 0,008), SATÉLITE (2; 0,008), CAMPO (2; 0,008), BRASIL (2; 0,008), DESMATAMENTO (2; 0,007)	4,26
2	abr158, abr35	ESTUDO (2; 0,032), BRASIL (2; 0,032), DESMATAMENTO (1; 0,027), PUBLICADO (2; 0,022), NATURE (2; 0,022), SATÉLITE (2; 0,022), IMAGENS (2; 0,022), EUA (2; 0,022), REVISTA (2; 0,022), AMAZÔNIA (2; 0,022)	4,26
3	abr59, jun32	FLORESTA (2; 0,028), INCÊNDIOS (2; 0,017), PESQUISA (2; 0,014), ACIDENTAIS (2; 0,012), ÁRVORES (2; 0,011), AMAZÔNIA (1; 0,009), INCÊNDIO (2; 0,008), MATO (2; 0,008), AMAZÔNICA (2; 0,008)	4,26
4	ago116, ago49	COLÔMBIA (2; 0,019), MCCAFFREY (2; 0,018), ESTADOS (2; 0,017), UNIDOS (2; 0,016), GENERAL (2; 0,01), EUA (2; 0,008), PAÍS (2; 0,008), BRASIL (2; 0,008), GOVERNO (2; 0,007)	4,26
5	ago52, ago52	PRESIDENTE(2; 0,029), FHC(2; 0,02), GOVERNO(2; 0,012), MARCHA(2; 0,01), POVO(2; 0,009), MIL(2; 0,009), BRASIL(2; 0,009)	4,26

Grupo	Arquivos	Palavras relevantes	%
6	dez1, dez66	MADEIREIRAS (2; 0,032), IBAMA (2; 0,032), AMAZÔNIA (2; 0,02), INSTITUTO (2; 0,017), AMBIENTE (2; 0,017), MEIO (2; 0,017), LEVANTAMENTO (2; 0,017), ILEGAL (2; 0,014), PARÁ (2; 0,014), EMPRESAS (2; 0,014), FISCALIZAÇÃO (2; 0,011), EXTRAÇÃO (2; 0,011)	4,26
7	dez6, set106, set107	QUEIMADAS (2; 0,028), IBAMA (3; 0,019), PRISÃO (2; 0,017), MEIO (3; 0,011), PORTARIA (2; 0,011), DIAS (3; 0,011), AMBIENTE (3; 0,01), PAULO (3; 0,01), FEDERAL (3; 0,008), SÃO (3; 0,008), MATO (2; 0,008), GROSSO (2; 0,008), GREENPEACE (1; 0,007), MADEIRA (1; 0,007)	6,38
8	dez67, dez73	MUSEU (2; 0,038), BRITÂNICO (2; 0,02), AMAZÔNIA (2; 0,018), EXPOSIÇÃO (2; 0,017), ANOS (2; 0,015), BRASIL (2; 0,014), ARTE (2; 0,011), MOSTRA (2; 0,011), PEÇAS (2; 0,011)	4,26
9	fev167, fev168, fev172	BARCO (3; 0,05), AMAZÔNIA (3; 0,047), PESSOAS (3; 0,044), RIO (3; 0,031), AFUNDOU (2; 0,031), DESAPARECIDAS (3; 0,028), MADEIRA (3; 0,028), CORPOS (2; 0,021), SELVA (2; 0,019)	6,38
10	fev170, fev171	DESMATAMENTO (2; 0,044), AUTORIZAÇÕES (2; 0,037), FORAM (2; 0,026), FUNCIONÁRIOS (2; 0,024), IBAMA (2; 0,024), AFASTADOS (2; 0,024), ARIPUANÃ (2; 0,018), AMAZÔNIA (2; 0,018), FAZENDA (2; 0,018), AMBIENTE (2; 0,015), MEIO (2; 0,015), FILHO (2; 0,013), SARNEY (2; 0,013)	4,26
11	fev173, nov88	BARCO (2; 0,034), RIO (2; 0,024), PESSOAS (2; 0,021), ACIDENTE (2; 0,015), DESAPARECIDOS (2; 0,014), PASSAGEIROS (2; 0,014), NAUFRÁGIO (2; 0,014), SUPERLOTAÇÃO (1; 0,013), CAUSA (1; 0,013), AFUNDOU (2; 0,011), EMBARCAÇÃO (1; 0,01), POLÍCIA (1; 0,01), NAVIO (2; 0,009), EXCESSO (2; 0,009)	4,26
12	fev175, fev176	INSTITUTO (2; 0,038), AMAZÔNIA (2; 0,038), DESMATAMENTO (1; 0,033), DESMATADAS (1; 0,027), ESPACIAIS (2; 0,024), NACIONAL (2; 0,024), AUMENTO (2; 0,024), PESQUISAS (2; 0,024), INPE (2; 0,024), ESTIMA (2; 0,024), REGIÃO (1; 0,022), MINISTRO (1; 0,022)	4,26
13	fev38, jan183	INPE (2; 0,016), ANO (2; 0,016), FOCOS (2; 0,015), QUEIMADAS (2; 0,014), INSTITUTO (2; 0,013), AMAZÔNIA (2; 0,012), ÁREA (2; 0,011)	4,26
14	jan64, jul128	ARMAS (1; 0,035), COLÔMBIA (2; 0,027), BRASIL (2; 0,023), SENDERO (1; 0,015), FARC (1; 0,014), PERU (2; 0,014), AFIRMOU (2; 0,013), GOVERNO (2; 0,013), VIEGAS (1; 0,011), RIVERO (1; 0,011), FILHO (1; 0,011)	4,26
15	jul122, jul126	CASOS (2; 0,058), ANO (2; 0,038), SEMESTRE (2; 0,026), PRIMEIRO (2; 0,026), MALÁRIA (2; 0,026), IANOMÂMIS (1; 0,022), REGISTRADOS (2; 0,019), RORAIMA (1; 0,018), ACRE (1; 0,018), ÍNDICE (1; 0,015), DISSEMINAÇÃO (1; 0,015), CONTROLE (1; 0,015)	4,26
16	jul124, jul125	SAÚDE (2; 0,03), FNS (2; 0,019), ÍNDIOS (2; 0,019), MÉDICO (2; 0,018), FUNAI (2; 0,016), ATENDIMENTO (1; 0,014), INDÍGENAS (2; 0,013), INDÍGENA (2; 0,012), PAÍS (2; 0,011), GOVERNO (2; 0,011), ÍNDIO (2; 0,01), FUNDAÇÃO (2; 0,01), FUNCIONÁRIOS (2; 0,008)	4,26
17	jul127, set27	IBAMA (2; 0,016), REGIÃO (2; 0,015), ANO (2; 0,015), AMAZÔNIA (2; 0,015), EXÉRCITO (1; 0,014), FOGO (2; 0,013), DESMATAMENTO (1; 0,011), CONVÊNIO (1; 0,011), MEIO (2; 0,011), AMBIENTE (2; 0,011)	4,26
18	jun147, out48	FHC (2; 0,022), PAÍSES (2; 0,018), GREENPEACE (1; 0,017), BRASIL (2; 0,013), PAZ (1; 0,012), COLÔMBIA (1; 0,012), ENTIDADE (1; 0,011), TEMA (2; 0,011), GOVERNO (2; 0,011), REGIÃO (2; 0,01), PASTRANA (1; 0,01), PROCESSO (1; 0,01), COLOMBIANO (1; 0,01), AMAZÔNIA (1; 0,008)	4,26

Grupo	Arquivos	Palavras relevantes	%
19	mar161, mar162, mar37	DESMATAMENTO (3; 0,055), AMAZÔNIA (3; 0,043), MINISTÉRIO (3; 0,042), PROIBIÇÃO (3; 0,034), FAMILIAR (3; 0,031), AGRICULTURA (3; 0,029), PROPRIEDADES (3; 0,027), AMBIENTE (3; 0,023), REGIÃO (3; 0,023), MEIO (3; 0,023), FAMÍLIAS (2; 0,023), INSTRUÇÃO (3; 0,021), REVOGOU (3; 0,021)	6,38
20	nov85, nov89	CPI (2; 0,023), CAMPINAS (2; 0,018), CURICA (2; 0,013), DEPUTADO (2; 0,01), RIO (1; 0,009), QUEBRA (2; 0,009), NARCOTRÁFICO (2; 0,009), SÂMIA (2; 0,008), SIGILO (2; 0,008), EMPRESAS (2; 0,008), TRAFICANTE (2; 0,008)	4,26
21	out100, out17	MADEIRA (2; 0,023), MANEJO (2; 0,017), CERTIFICAÇÃO (2; 0,015), PRODUÇÃO (2; 0,009), AMAZÔNIA (2; 0,009), MUNDIAL (2; 0,008), MERCADO (2; 0,007), SETOR (2; 0,007), FSC (2; 0,007), GOVERNO (2; 0,007)	4,26
22	set19, set21	SOJA (1; 0,017), SEMINÁRIO (2; 0,016), MAPA (1; 0,012), ÁREAS (1; 0,012), GOVERNO (2; 0,01), AMAZÔNIA (2; 0,01), EIXOS (2; 0,01), CONSERVAÇÃO (2; 0,008), RISCO (2; 0,008), REGIÃO (2; 0,008), FEDERAL (2; 0,008), BIODIVERSIDADE (2; 0,008)	4,26
0		Os outros não foram agrupados	

Grau de similaridade: 15%

Grupo	Arquivos	Palavras relevantes	%
1	abr158, abr35	ESTUDO (2; 0,032), BRASIL (2; 0,032), DESMATAMENTO (1; 0,027), PUBLICADO (2; 0,022), NATURE (2; 0,022), SATÉLITE (2; 0,022), IMAGENS (2; 0,022), EUA (2; 0,022), REVISTA (2; 0,022), AMAZÔNIA (2; 0,022)	20
2	fev167, fev168	BARCO (2; 0,062), AMAZÔNIA (2; 0,057), PESSOAS (2; 0,052), AFUNDOU (2; 0,047), RIO (2; 0,033), DESAPARECIDAS (2; 0,028), SELVA (2; 0,028), MADEIRA (2; 0,028)	20
3	fev170, fev171	DESMATAMENTO (2; 0,044), AUTORIZAÇÕES (2; 0,037), FUNCIONÁRIOS (2; 0,024), IBAMA (2; 0,024), AFASTADOS (2; 0,024), ARIPUANÃ (2; 0,018), AMAZÔNIA (2; 0,018), FAZENDA (2; 0,018), AMBIENTE (2; 0,015), MEIO (2; 0,015), FILHO (2; 0,013), SARNEY (2; 0,013)	20
4	mar161, mar162	DESMATAMENTO (2; 0,059), AMAZÔNIA (2; 0,059), PROIBIÇÃO (2; 0,045), MINISTÉRIO (2; 0,045), FAMÍLIAS (1; 0,032), REVOGOU (2; 0,029), MEIO (2; 0,029), INSTRUÇÃO (2; 0,029), FAMILIAR (2; 0,029), REGIÃO (2; 0,029), AMBIENTE (2; 0,029), PROPRIEDADES (2; 0,029)	20
5	set106, set107	QUEIMADAS(2; 0,043), PRISÃO(2; 0,026), PORTARIA(2; 0,017), IBAMA(2; 0,015), MEIO(2; 0,013), MATO(2; 0,013), GROSSO(2; 0,013), SÃO(2; 0,011), PAULO(2; 0,011), MINISTRO(2; 0,011), FILHO(2; 0,011), SARNEY(2; 0,011), AMBIENTE(2; 0,011)	20
0		Os outros não foram agrupados	