

Recuperação de Informações usando a Expansão Semântica e a Lógica Difusa

Leandro Krug Wives (CPGCC/UFRGS) wives@inf.ufrgs.br

Stanley Loh (ULBRA, UCPEL, CPGCC/UFRGS) loh@inf.ufrgs.br

Address:

Curso de Pós-Graduação em Ciência da Computação
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Avenida Bento Gonçalves, 9500
Bloco IV, Prédio 43412 - Campus do Vale
Porto Alegre - RS
BRASIL

Resumo

Este artigo apresenta uma abordagem para recuperação de informações utilizando duas técnicas, a saber: a expansão semântica e a lógica difusa. A expansão semântica permite buscar documentos não somente pelas palavras fornecidas como entrada, mas através de um conjunto maior de termos que define melhor o contexto do assunto requerido. Já a lógica difusa contribui para a definição de quanto cada termo é importante para a consulta, para os contextos e para os documentos da base. Além disto, os operadores da lógica difusa permitirão avaliar a relação entre os dados de entrada, os contextos existentes e os documentos da base. Um software foi implementado com base nesta abordagem e as conclusões sobre os experimentos realizados são discutidas ao final deste artigo.

Abstract

This work presents an approach to information retrieval using two techniques: semantic expansion and fuzzy logic. The first allows to retrieve documents by a set of words bigger than that given at entry. These words are supposed to better define the context of the query. On other side, fuzzy logic and its operators give a better understanding of the degree how words, documents and contexts are related. The software implemented by this way and the conclusions about experiments are discussed at the end of the paper.

* Este trabalho é parcialmente apoiado por FAPERGS, CAPES e PROTEM-CNPq.

* Agradecimentos: os autores gostariam de agradecer, *in memoriam*, ao orientador Prof. Dr. José M. V. de Castilho

1. Introdução

Nos dias atuais, é grande o número de informações que ficam disponíveis para acesso rápido e fácil. O aprimoramento dos meios físicos de armazenamento e da tecnologia computacional contribuiu muito para que as pessoas armazenassem e buscassem mais e mais informações.

Entretanto, muitas destas informações não estão em formatos que possam ser facilmente tratados por meios computacionais (tais como imagens, textos, vídeos, gráficos, desenhos, etc). [WHI96] chega a calcular que 80% das informações que uma empresa utiliza não estão armazenadas em Bancos de Dados na forma de números e caracteres.

Em especial, as informações na forma de textos têm chamado à atenção da comunidade de pesquisa. Existe uma área (que não é recente) que trata especificamente da busca de informações em documentos que contenham textos. Esta área, conhecida como Recuperação de Informações (*Information Retrieval*), pesquisa técnicas para indexar e encontrar documentos (ou partes destes) a partir de determinados padrões pré-estabelecidos ou relevantes para um determinado interesse (segundo [COW96]).

Com a difusão da Internet, ainda mais necessárias se tornaram técnicas de tal tipo. [CHE94] cita a frustração dos usuários com o problema da "sobrecarga de informações". Ela ocorre quando o usuário tem muita informação ao seu alcance, mas não tem condições de tratá-la ou de encontrar o que realmente deseja ou lhe interessa.

Especificamente para a Internet, existem inúmeras ferramentas para recuperar informações ou documentos textuais, entre elas *AltaVista*, *Yahoo*, *Cadê*, etc.

As ferramentas de Recuperação de Informações, geralmente, trabalham com técnicas de indexação capazes de indicar e acessar mais rapidamente documentos de um Banco de Dados textual (conforme [YAT96]).

Existem três tipos principais de indexação (derivados do estudo de [YAT96]):

- indexação tradicional: é aquela onde uma pessoa determina os termos descritivos ou caracterizadores dos documentos, os quais farão parte do índice de busca (com por exemplo no caso de um *Thesaurus*);
- indexação *full-text* (ou indexação do texto todo): onde todos os termos que compõem o documento fazem parte do índice; e
- indexação por *tags* (por partes do texto): onde apenas algumas partes do texto são escolhidas, automaticamente, para gerar as entradas no índice (somente aquelas consideradas mais importantes ou mais caracterizadoras).

Este trabalho trata de técnicas de recuperação de informações e apresenta uma ferramenta automatizada para busca de documentos textuais utilizando duas técnicas principais: a expansão semântica e a lógica difusa (*fuzzy logic*).

Na seção seguinte, os problemas referentes a esta área são discutidos. Depois são apresentados trabalhos correlatos, realizados para solucionar tais problemas. Na seção 4, a solução proposta e implementada é apresentada, descrevendo as técnicas que foram usadas para desenvolver a referida ferramenta. Na seção 5, será detalhada a forma como a ferramenta foi implementada e na seção 6 são apresentados e discutidos alguns estudos de casos realizados com a ferramenta. Por fim, conclusões, contribuições e limitações são avaliadas na seção 7.

2. Explicitação do Problema

Um tipo de problema que geralmente ocorre com as ferramentas de recuperação de informações (principalmente com aquelas voltadas para a Internet) é o retorno de grandes volumes de documentos como resposta a uma consulta. Entre estes, certamente a maioria não é relevante para o interesse do usuário que fez a consulta.

Outro grave problema que ocorre nas técnicas tradicionais de recuperação de informações é que muitas vezes documentos importantes não são recuperados.

Isto se dá porque estas técnicas estão baseadas na presença ou não de palavras nos documentos. Entretanto, pode haver documentos relevantes que não contém as palavras especificadas na consulta (para a busca) e pode haver documentos que contém as palavras de entrada mas que não tratam do assunto desejado.

Este problema é denominado de **indexação imprecisa** e ocorre porque a pessoa ou técnica que descreve e indexa os documentos pode utilizar termos diferentes de quem procura pelos documentos.

A técnica de indexação semântica é usada para melhor compatibilizar o contexto da busca (o interesse do usuário, que é descrito apenas pelos termos de entrada da consulta) e o contexto dos documentos (expressando o conteúdo do documento e caracterizado pelos termos que o compõem).

Já a lógica *fuzzy* (definida em [ZAD73]) serve, entre tantas aplicações na área de recuperação de informações, para amenizar as incertezas advindas do uso de termos lingüísticos e para melhor detalhar a importância dos termos em relação à consulta, a relevância dos documentos para a consulta e o grau em que um termo caracteriza um documento.

3. Trabalhos Correlatos

Algumas técnicas então procuram recuperar documentos baseadas no contexto dos documentos. [CHE96] define contexto ou espaço conceitual como sendo um conjunto de palavras que definem um assunto ou área do conhecimento. [CHE96] discute técnicas baseadas na frequência de termos em documentos para determinar a importância de um termo em um documento e o grau de pertinência de um termo em um contexto (o quanto ele ajuda a definir um contexto).

Estas fórmulas baseadas na frequência relativa (número de vezes em que um termo aparece no documento dividido pelo número total de termos no documento) e na frequência inversa (número de documentos onde o termo aparece) ajudam a definir que termos podem ser usados para recuperar determinado contexto (ou documentos deste contexto).

Se um termo aparece muito em um documento, então o primeiro caracteriza em alto grau o último. Se um termo aparece em muitos documentos, seu grau de discriminação será baixo (pois muitos documentos serão recuperados a partir deste termo), enquanto que, se um termo aparece em poucos documentos, então diz-se que ele caracteriza bem estes documentos. Obviamente, termos que aparecem em todos os documentos não serão analisados (estes são chamados de *stop-words*, e geralmente são as preposições, artigos, pronomes, etc).

Também para tratar o problema de busca contextual, há a técnica de [CHA95], a qual se utiliza de expansões semânticas de palavras. Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto

para busca de documentos. [CHA95] utiliza as definições de um dicionário para achar as palavras que se relacionam, eliminando *stop-words* e modela estas relações através de redes semânticas, criadas manualmente.

Entretanto, os problemas desta técnica são saber que palavras expandir para fazer a busca e se as novas palavras acrescentadas realmente fazem parte do contexto. Segundo os experimentos de [CHA95], algumas das novas palavras não fazem parte do contexto, o que pode fazer com que documentos irrelevantes sejam recuperados.

A intervenção de especialistas humanos pode amenizar em parte tais obstáculos. Os contextos (conjuntos de palavras que caracterizam cada contexto) podem ser definidos por especialistas ou então um especialista seleciona textos de um mesmo contexto e submete a uma ferramenta que, baseada nas fórmulas tratadas acima, extrai os termos que melhor definem o tal contexto.

Problemas podem ocorrer quando houver mais de um contexto possível para uma dada situação, seja porque um documento pertence a mais de um contexto ou porque vários especialistas definiram vários conjuntos diferentes para caracterizar o mesmo contexto.

[WIE96] cita técnicas que utilizam modelagem de contextos alternativos, permitindo que contextos diferentes possam ser explorados em paralelo, e [OLI96] cita uma técnica que combina os conjuntos diferentes em um único resultante, através de operadores *fuzzy* de conjunção e disjunção.

A raiz do problema de contextos diferentes está na imprecisão dos termos (termos com significados diferentes). Este problema pode ser notado tanto no momento da criação do índice, como na hora da recuperação. Isto se dá porque as pessoas utilizam vocabulários diferentes para exprimir suas intenções (conforme [FUR87] comprovou em seus estudos).

A expansão semântica pode ajudar a amenizar tal problema porque permite fazer a busca de documentos com base num conjunto maior de termos. Segundo [IIV95], o usuário deve utilizar o maior número possível de características (termos) para descrever a informação que deseja. Deste modo uma consulta com um conjunto maior de termos tende a recuperar informações mais adequadas ao usuário.

Por sua vez, a lógica *fuzzy* pode ajudar a tratar melhor as incertezas quanto à relevância dos termos dos índices em relação aos documentos e a importância dos termos de entrada para a consulta.

[CRO94] apresenta um *survey* sobre recuperação *fuzzy* de informações. Nos casos apresentados, a lógica *fuzzy* é utilizada para expressar os graus de relevância dos termos no índice em relação aos documentos e para expressar os graus de importância dos termos na consulta.

Os termos de entrada, fornecidos pelo usuário, podem ter relevâncias diferentes para a consulta. Tais diferenças são expressas em valores *fuzzy*, que podem ser determinados por avaliação de termos lingüísticos como “é relevante”, “é muito relevante”, “é pouco relevante”, etc.

Tal abordagem é diferente do modelo probabilístico, o qual avalia se os termos são relevantes ou não (sem graus intermediários) e daí então estima a probabilidade de ocorrência dos termos nos documentos.

[CRO94] cita métodos que usam a frequência relativa dos termos para dar a importância (ou peso) do termo em relação ao documento (independente de outros), expressando tal importância em valores *fuzzy*. Também aborda o uso de consultas complexas com operadores de conjunção, disjunção e negação sobre os valores *fuzzy*.

[CRO94] também cita trabalhos sobre o uso de sinônimos e hierarquias de conceitos (índices tipo *thesaurus*) usando a lógica *fuzzy*, onde termos genéricos são descritos por conjuntos *fuzzy* de termos mais específicos.

Outra técnica citada é a rede semântica para emular o conhecimento do especialista para fazer a expansão semântica da consulta (encontrando termos relacionados semanticamente com os de entrada). São utilizados pesos diferentes (valores *fuzzy*) nas ligações da rede para expressar o quanto um termo se relaciona a outro. O operador de produto é recomendado para juntar (conjunção de) termos. Já o operador de disjunção é usado para a união do conjunto inicial de termos para a consulta com os outros que vão sendo definidos pelo processo de expansão.

Por fim, o grau de satisfação dos documentos em relação à consulta também pode ser expresso em valores *fuzzy*. Pode-se utilizar um limiar (*threshold*) para selecionar documentos na resposta (evitando mostrar todos os documentos recuperados).

Para medir a relação entre o conjunto de entrada, os conjuntos de contextos e os documentos (conjuntos de palavras que os compõem), [CRO94] sugere duas medidas de similaridade ou compatibilidade:

- *set theoretic inclusion*: avalia se um termo está incluso ou não; e
- *Euclidean distance*: representar os conjuntos de termos como vetores no espaço e determinar as distâncias.

Já [OLI96] sugere operadores *fuzzy* complexos para realizar tal medida.

[CRO94] cita ainda as redes neurais *fuzzy* como uma maneira de representar a relação entre termos e documentos. As entradas são os termos da consulta e as saídas são os documentos. Um especialista humano então intervém para treinar a rede.

Para avaliação das técnicas de recuperação de informações são utilizados dois conceitos bastante conhecidos no meio (conforme [SAL84]): precisão (*precision*) e abrangência (*recall*). O primeiro avalia se somente documentos relevantes foram recuperados e o segundo avalia se todos os documentos relevantes foram recuperados.

4. Modelagem da Solução Proposta

A solução proposta neste trabalho, para ajudar a resolver os problemas da busca de documentos textuais, é baseada em duas técnicas principais, como sugestão dos pesquisadores da área e como visto na seção anterior. A saber, são elas:

- a expansão semântica; e
- a lógica *fuzzy*.

A expansão semântica será utilizada para aumentar o conjunto de termos de busca. Para tanto, serão utilizados contextos de busca pré-definidos, para que sejam recuperados documentos pertencentes ao(s) contexto(s) relevante(s) e não apenas os documentos que possuem os termos de entrada. Desta forma, os termos de entrada serão comparados com os contextos existentes, e alguns contextos (os mais significativos) serão selecionados para a busca dos documentos.

Os documentos da base de busca deverão ser representados internamente por conjuntos *fuzzy*, com os termos que os compõem e o grau de pertinência do termo no documento. Este grau é calculado pela frequência relativa do termo no documento, isto é, o número de vezes em que aparece no documento dividido pelo número total de termos no documento.

Nesta proposta, todas as partes do documento contribuem igualmente para a avaliação do documento. Portanto, o peso do termo no documento será calculado sem levar em conta a parte do documento onde aparece, conforme sugestão de [CRO94]. Outra abordagem seria considerar os termos de algumas partes como mais importantes (por exemplo, palavras do

título).

Entretanto, já que alguns termos não têm valor de discriminação, pois aparecem em vários documentos, estes termos, conhecidos como *stop-words*, deverão ser retirados dos documentos para efeitos de cálculo e montagem do conjunto *fuzzy* representativo do documento.

Nesta proposta, os contextos serão representados como conjuntos *fuzzy* de termos ou palavras. Estes conjuntos representarão o conhecimento de especialistas para definir que contextos possuem certos termos ou quais os termos que definem um contexto. Como visto, uma implementação melhor seria utilizar redes semânticas. Para fins de avaliação da proposta como um todo, foi escolhida a alternativa mais fácil de ser implementada.

Associado a cada termo dentro de um contexto, será usado um valor *fuzzy* que representa o grau de pertinência do termo no contexto. Cabe salientar então que poderá haver termos que participam em vários contextos. A forma como os contextos são gerados é discutida em 4.1.

A lógica *fuzzy* ainda será usada para que o usuário expresse a importância dos termos de entrada com relação à consulta. Nesta implementação, o usuário deverá fornecer diretamente um valor *fuzzy* associado a cada termo, mas trabalhos futuros poderão normalizar a entrada através do uso de termos lingüísticos, como visto na seção 3.

Também serão utilizados valores *fuzzy* para determinar o grau de satisfação de um documento resultante da busca em relação à consulta original.

Lembrando que cada contexto é um conjunto *fuzzy* dos termos que o definem e que cada documento também pode ser representado por um conjunto *fuzzy* dos termos que o compõem, pode-se fazer uma comparação entre as abordagens convencionais (que buscam por presença de termos) e a proposta aqui:

(a) Modelo da Solução Convencional (busca por presença de palavras)

$$\{\text{documentos resultantes}\} = \{\text{palavras de entrada}\} \circ [\text{palavras X documentos}]$$

sendo que:

- o símbolo \circ representa uma combinação entre conjuntos *fuzzy* e/ou relações *fuzzy*, utilizada para realizar a inferência (regra de inferência composicional, conforme [NAK93]);
- os símbolos [] representam uma relação *fuzzy* (que pode ser representada por uma matriz);
- e os símbolos { } representam um conjunto *fuzzy*.

(b) Modelo da Solução Proposta neste trabalho

$$\{\text{documentos resultantes}\} = \{\text{palavras de entrada}\} \circ_1 [\text{palavras X contextos}] \circ_2 [\text{contextos X documentos}]$$

O resultado da combinação $\{\text{palavras de entrada}\} \circ_1 [\text{palavras X contextos}]$ é um conjunto *fuzzy* que representa os contextos e seus graus de relevância para a consulta (para os termos de entrada).

Este resultado intermediário $\{\text{contextos}\}$ será combinado com a matriz $[\text{contextos X documentos}]$ através de \circ_2 , resultando no conjunto *fuzzy* final $\{\text{documentos}\}$ que pode ser interpretado como os documentos e seus graus de satisfação em relação à consulta de entrada.

A matriz que relaciona contextos e documentos [contextos **X** documentos] deverá ter sido previamente determinada para acelerar o processo de busca e será resultante da combinação dos conjuntos *fuzzy* de contextos e de documentos. Como os conjuntos de contextos e de documentos são vetores de palavras, pode-se também representá-los por matrizes. Assim, tem-se que:

$$[\text{contextos } \mathbf{X} \text{ documentos}] = [\text{contextos } \mathbf{X} \text{ palavras}] \circ_3 [\text{palavras } \mathbf{X} \text{ documentos}].$$

As relações *fuzzy* resultantes das combinações \circ seguem a sugestão do raciocínio *fuzzy* de [NAK93], onde

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

Os operadores utilizados para as disjunções e conjunções dos conjuntos ou relações *fuzzy* são os seguintes, com as respectivas justificativas da escolha:

- na combinação \circ_1 :

$\bigvee \Rightarrow$ soma limitada = $\min(1, x + y)$, já que os termos de entrada que não estão em um contexto não devem diminuir o grau deste contexto em relação à consulta, pois o contexto pode possuir sinônimos para estes termos;

$\wedge \Rightarrow$ produto algébrico = $(x * y)$, para que ambos os graus (o do termo na entrada e o do termo no contexto) sejam computados, uma vez que ambos são importantes para o resultado final ;

Observação 1: se um termo aparece em um dos fatores e não no outro, ele então aparecerá no resultado mas com grau $\mu = 0$ (zero), pelo operador produto algébrico.

Observação 2: os termos que aparecem no resultado da combinação com grau $\mu = 0$ (zero) não influenciarão na disjunção \bigvee ; se se quiser o contrário (que estes termos diminuam o valor final), pode-se utilizar outro operador, como o de Média ou Média Ponderada, por exemplo (como discutido em [OLI96]).

- na combinação \circ_2 :

$\bigvee \Rightarrow$ máximo = $\sup(x,y)$, porque só interessa o maior contexto no qual o documento está inserido;

$\wedge \Rightarrow$ produto algébrico = $(x * y)$, para que ambos os graus sejam computados (o do contexto em relação à consulta e o do contexto em relação aos documentos), uma vez que ambos são importantes para o resultado final;

- na combinação \circ_3 :

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

$\wedge \Rightarrow$ produto algébrico = $(x * y)$, para que ambos os graus sejam computados (o do termo no contexto e o do termo no documento), uma vez que ambos são importantes para o resultado final.

4.1 Montagem dos Contextos

Como já dito anteriormente, os contextos (representados por conjuntos de palavras) serão definidos por um especialista, o qual escolherá os contextos do Universo de Discurso, os termos que farão parte de cada contexto e seu respectivo grau de pertinência (dentro de cada contexto, pois um termo pode aparecer em mais de um contexto com graus diferentes).

Entretanto, já que esta é uma atividade sujeita a falhas, pode-se utilizar outras duas abordagens: o aprendizado supervisionado e a aprendizado por *clusterização*.

No primeiro caso, um especialista seleciona vários documentos sobre um determinado contexto e submete a uma ferramenta. Esta ferramenta então extrairá o centróide (conforme sugestão de [SAL84]) destes documentos, uma espécie de vetor médio com os termos que mais aparecem nos documentos e um respectivo grau de pertinência, calculado pela média dos graus de pertinência (ou peso) do termo em cada documento. Este centróide então será usado como o conjunto *fuzzy* que define o tal contexto.

A segunda alternativa é utilizar uma ferramenta que agrupa automaticamente, sem intervenção humana, os documentos de um mesmo contexto e então extrai o centróide de cada grupo ou classe (obviamente, alguém deverá selecionar os documentos de entrada, mas não necessitará fazer nenhuma análise sobre eles)

5. Implementação

Foi implementada uma ferramenta para avaliar a proposta de solução para recuperação de informações, usando expansão semântica e lógica *fuzzy*, como especificado na seção 4.

A ferramenta é um protótipo, com algumas limitações de tempo de resposta (que podem ser dirimidas em trabalhos futuros) e pouco tratamento de inconsistências nos dados de entrada. Também não se ateu muito em projetar *interfaces* amigáveis, portanto seu uso pode ser um pouco difícil. A ferramenta foi implementada em Delphi 2.0, para ambiente Windows 95.

6. Experimentos

Foram realizados alguns experimentos de consulta para avaliar as técnicas empregadas e a ferramenta implementada.

Como métricas para avaliação, foram utilizados os conceitos de *precision* e *recall* (conforme sugerido por [SAL84]) com as seguintes fórmulas:

$$\begin{aligned}\text{grau de } precision &= ndr / ndt \\ \text{grau de } recall &= ndr / ndr_u\end{aligned}$$

onde:

ndr é o número de documentos relevantes recuperados (somente os relevantes à consulta dentre todos os que foram recuperados);

ndt é o número total de documentos recuperados; e

ndru é o número de documentos relevantes do universo, os quais deveriam ser recuperados.

A avaliação das métricas foram feitas pelos próprios autores e pelas pessoas que utilizaram a ferramenta, simulando a função de um especialista humano para avaliar o que era relevante no universo de documentos.

Quanto à entrada de dados, os usuários foram instruídos a fornecer os termos para consulta e os respectivos graus de importância.

Para a realização dos experimentos foram utilizadas uma base única de contextos (definida pelos autores sem levar em conta o conteúdo dos documentos) e um conjunto pré-determinado de documentos textuais (escolhidos sem critério algum). Também foi elaborada pelos autores uma lista de *stop-words* usada no início do processo, para “limpar” os documentos textuais (esta lista contém as preposições da língua portuguesa e outros termos julgados comuns).

7. Conclusão

Este trabalho discutiu técnicas de recuperação de documentos e apresentou uma implementação de ferramenta que utiliza a expansão semântica e a lógica *fuzzy* para realizar a busca de documentos textuais.

A expansão semântica permite que um número maior de documentos relevantes seja recuperado e que apenas os realmente relevantes sejam recuperados, já que a análise dos documentos não é feita somente com base na presença dos termos de entrada nos documentos, mas também leva em conta sinônimos e termos semanticamente relacionados. Neste trabalho, a expansão da entrada foi feita utilizando-se uma base de contextos (conjuntos de termos que definem assuntos) previamente definidos.

Já a lógica *fuzzy* permite trabalhar com a incerteza dos resultados (graus diferentes de satisfação dos documentos em relação à consulta) e com graus diferentes de importância para os termos fornecidos como entrada. Ainda, associada à base de contextos, a lógica *fuzzy* permite que os termos tenham graus de pertinência diferentes em relação a cada contexto.

Com base nos experimentos realizados, pode-se concluir que a ferramenta atinge um grau satisfatório de *precision* (precisão) e *recall* (abrangência). Para limiares (*threshold*) próximos de 0,005 (valores que limitam os documentos a serem apresentados como resposta - somente os documentos com grau de satisfação maior que o limiar), a ferramenta consegue alta precisão (próximo de 0,9) e média-alta abrangência (próximo de 0,6).

Concluiu-se também que limiares entre 0,007 e 0,002 são os melhores para avaliar comparativamente os experimentos, já que atingem os melhores desempenhos de *precision* e *recall*. O limiar de 0,007 é o que melhor combina *precision* e *recall*, e 0,002 é o menor limiar para o qual se obtêm resultados significativos.

Para a avaliação dos graus de *precision* e *recall*, os autores atuaram como especialistas, determinando (com base nos conteúdos dos documentos) quais documentos recuperados eram relevantes para a consulta e quais documentos do universo (da base considerada) eram relevantes para a consulta. Obviamente, esta forma de experimentação fica sujeita a interferências dos observadores. Trabalhos futuros devem avaliar a ferramenta e seus resultados de forma mais imparcial.

Outras limitações da ferramenta estão relacionadas com a forma como os documentos são analisados. Uma vez que a lista de *stop-words* não foi criada com critérios científicos, muitos termos com pouco valor de discriminação (aqueles que aparecem em vários documentos e não permitem deduzir o assunto relativo) acabaram sendo considerados nos cálculos.

Da mesma forma, como alguns caracteres especiais (como os de pontuação e as aspas) não foram filtrados, nem erros de acentuação foram distinguidos, algumas análises podem ter perdido em precisão nos resultados. Por exemplo, em documentos sobre prontuários médicos, o termo “paciente:” (com dois pontos no final) acabou sendo considerado na análise léxica, enquanto que “médico” (com acento) e “medico” (sem acento) foram considerados termos diferentes. Entretanto, por análises subjetivas dos documentos, constatou-se que tais problemas não eram freqüentes, portanto não interferindo nos resultados de forma significativa. Para amenizar problemas com acentuação e erros de ortografia, os termos (tanto na entrada quanto nos contextos) podem ser fornecidos nas suas várias várias alternativas.

Outra constatação a que se chegou é de que textos muito pequenos (sem termos repetidos) podem ocasionar desvios nas análises. Portanto, quanto maiores os textos considerados, melhores os resultados.

Problemas também podem ocorrer devido aos procedimentos de determinação dos graus *fuzzy* (tanto para os termos de entrada, quanto para os termos dos contextos). Assim também, os contextos definidos podem causar desvios se não forem bem determinados. O número de contextos deve ser grande suficiente para abranger o maior número de assuntos possíveis. Caso não haja um contexto específico para a consulta fornecida, uma combinação de contextos será utilizada, aumentando assim a incerteza dos resultados.

Da mesma forma, um número pequeno de documentos pode influenciar os cálculos de *precision* e *recall*.

A ferramenta implementada também apresentou limitações quanto ao tempo de resposta. Apesar de as consultas todas terem sido realizadas de maneira rápida (tempo de resposta menor que 1 segundo), a criação da matriz que relaciona contextos e documentos é bastante demorada. Para os casos estudados (5 contextos contra 36 documentos), o tempo de processamento chegou a levar 15 minutos. Trabalhos futuros poderão melhorar tal desempenho.

Como contribuições deste trabalho, vale salientar que as técnicas empregadas (expansão semântica e lógica *fuzzy*) mostraram-se convenientes para o problema de busca, apresentaram um bom grau de satisfação (pelas métricas de *precision* e *recall*) e são uma boa alternativa às técnicas de recuperação de documentos baseadas unicamente na presença de termos e em valores *crisp*.

8. Referências Bibliográficas

- [CHA95] CHAKRAVARTHY, Anil S.; HAASE, Kenneth B. *NetSerf: using semantic knowledge to find Internet information archives*. **Proceedings**. SIGIR, 1995.
- [CHE94] CHEN, Hsinchun. *A textual database/knowledge-base coupling approach to creating computer-supported organizational memory*. MIS Department, University of Arizona, 5 de Julho de 1994. (<http://ai.bpa.arizona.edu/papers/>)

- [CHE96] CHEN, Hsinchun et alli. *A concept space approach to addressing the vocabulary problem in scientific information retrieval: na experiment on the worm community system*. MIS Department, University of Arizona, 2 de Julho de 1996. (<http://ai.bpa.arizona.edu/papers/>)
- [CRO94] CROSS, Valerie. *Fuzzy information retrieval*. **Journal of Intelligent Information Systems**, 5, 1994.
- [COW96] COWIE, Jim; LEHNERT, Wendy. *Information extraction*. **Communications of the ACM**, v.39, n.1, Jan 96.
- [FUR87] FURNAS, G. W. et alli. *The vocabulary problem in human-system communication*. **Communications of the ACM**, v.11, n.30, Nov 1987.
- [IIV95] IIVNEN, Mirja. Searches and Searches: Differences Between the Most and Least Consistent Searches. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 149-157.
- [NAK93] NAKANISHI, H.; TURKSEN, I. B.; SUGENO, M. *A review and comparison of six reasoning methods*. **Fuzzy Sets and Systems**, 57, 1993.
- [OLI96] OLIVEIRA, Henry M. *Seleção de entes complexos usando lógica difusa*. Instituto de Informática da PUC-RS, Porto Alegre, Julho de 1996. **(dissertação de mestrado)**
- [SAL84] SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. New York, McGraw-Hill.
- [WHI96] WHITE, Phillip. *Uma empresa que sabe aonde que chegar*. Entrevista na **Revista Informática Exame Especial**, ano 11, n.6, Set 1996. Editora Abril.
- [WIE96] WIEBE, Janyce; HIRST, Graeme; HORTON, Diane. *Language use in context*. **Communications of the ACM**, v.39, n.1, Jan 96.
- [YAT96] YATES, Ricardo Baeza. *An extended model for full text databases*. **Journal of the Brazilian Computer Society**, v.2, n.3, Abr 1996.
- [ZAD73] ZADEH, Lotfi A. *Outline of a new approach to the analysis of complex systems and decision processes*. **IEEE Transactions on Systems, Man and Cybernetics**, v. SMC-3, n.1, January 1973.