Recuperação Semântica de Documentos Textuais na Internet[1]

Stanley Loh
(Professor ULBRA e UCPEL,
doutorando PPGC/UFRGS)
loh@inf.ufrgs.br

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

Antônio Severo Frainer
(Professor ULBRA, servidor INSS)
asfrainer@cpovo.net

Endereço

Programa de Pós-Graduação em 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 - 91501-970
BRASIL

Resumo

Este artigo apresenta uma abordagem para recuperação semântica de documentos textuais disponíveis na Internet. As ferramentas propostas combinam duas técnicas principais: a expansão semântica e a lógica *fuzzy* (difusa). A expansão semântica analisa o contexto dos termos fornecidos como entrada e determina um conjunto maior de termos para o processo de recuperação, eliminando assim ambigüidades e precisando melhor o interesse do usuário. Já a lógica *fuzzy* (com seus operadores e conjuntos *fuzzy*) ajuda a amenizar as incertezas advindas do uso de termos lingüísticos e também ajudam a detalhar melhor as relações entre termos, contextos e documentos, além de permitir inferências sobre estas relações. As ferramentas implementadas e os resultados dos primeiros experimentos são discutidos ao final do artigo.

Palavras-chave: Internet, Recuperação de Informações

Abstract

This work presents an approach to retrieve textual documents from Internet using two techniques: semantic expansion and fuzzy logic. Semantic expansion analyses the contextual use of the words given as entry and generates a bigger set of words for eliminating ambiguities and imprecision, in order to discover the real interest of users. Fuzzy logic gives a better understanding about how words, documents and contexts are related, besides it allows to do inference from these relations. The software implemented and the conclusions about experiments are discussed at the end of the paper.

Keywords: Internet, Information Retrieval

[1] Este trabalho é parcialmente financiado por FAPERGS, CNPq/PROTEM e CAPES.
Os autores gostariam de agradecer, *in memoriam,* o Prof. Dr. José M. V. de Castilho

1 Introdução

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

Entretanto, tamanho volume de documentos e informações acaba por trazer problemas na hora da pesquisa. Um destes problemas é denominado "sobrecarga de informações" [4], já que o usuário facilmente obtêm muita informação e geralmente não consegue tratá-la, não encontrando o que realmente deseja ou lhe interessa.

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 armazenar e encontrar documentos (ou partes destes), a partir de padrões fornecidos como entrada.

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 [1].

Existem três tipos principais de indexação (derivados do estudo de [1]):

- indexação tradicional: é aquela onde uma pessoa determina os termos descritivos ou caracterizadores dos documentos, os quais farão parte do índice de busca (como 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 para gerar as entradas no índice (somente aquelas consideradas mais importantes ou mais caracterizadoras).

Já as consultas são geralmente feitas através de termos fornecidos pelo usuário ou escolhidos por este dentre alguns apresentados. Estes termos podem significar o assunto ou classe a que pertencem os documentos desejados (na indexação tradicional) ou termos que devem estar presentes nos documentos desejados (nas indexações *full-text* e por *tags*).

Algumas variações de consultas são comuns. Por exemplo, em algumas ferramentas, há uma linguagem baseada na lógica *booleana*, que utiliza conetivos e símbolos lógicos para eliminar documentos com determinados termos ou para recuperar somente documentos que contenham obrigatoriamente certos termos.

Já existem ferramentas, tais como AltaVista® e Yahoo®, que auxiliam a recuperação de documentos na Internet, utilizando as técnicas comentadas anteriormente.

2 Problemas da Área

Apesar da incontestável utilidade das ferramentas de recuperação de informações na Internet, alguns problemas podem ocorrer. Primeiro, a maioria dos usuários que utilizam as ferramentas de localização é inexperiente ou leiga, tanto no assunto que procuram quanto na utilização da ferramenta em si. Portanto, têm dificuldades em definir suas necessidades de informação utilizando palavras e conetivos (a utilização de conetivos não é prática, principalmente se a consulta é muito complexa [20]).

Também ocorre de as diversas ferramentas agruparem as informações de diferentes maneiras. Um estudo apresentado em [10] demonstra que aquelas pessoas que conhecem o funcionamento interno da ferramenta, e possuem mais experiência com a linguagem de consulta (que é também específica da ferramenta) têm mais facilidade de encontrar informações úteis.

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

Quando a técnica de indexação tradicional é utilizada, podem ocorrer problemas quando o especialista cataloga documentos de forma errada (por exemplo, interpretando equivocadamente o conteúdo de um documento) ou quando o usuário não consegue encontrar um assunto (dos pré-definidos) que represente precisamente seus interesses de busca. O mesmo pode acontecer na indexação por *tags*, quando da escolha das partes de onde extrair os termos caracterizadores.

O principal problema comentado acima 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 (podem ser usados termos diferentes para a mesma idéia ou termos iguais para idéias diferentes). Ou seja, as pessoas costumam utilizar vocabulários diferentes para expressar suas intenções [8].

3 Trabalhos Correlatos

Alguns trabalhos procuram solucionar os problemas citados anteriormente.

A técnica de indexação semântica é usada para melhor compatibilizar o contexto da recuperação (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).

Algumas técnicas então procuram recuperar documentos baseadas no contexto dos documentos. O **contexto** ou **espaço conceitual** é definido como sendo um conjunto de palavras que definem um assunto ou área do conhecimento [5]. Há estudos que discutem técnicas baseadas na freqüência de termos em documentos para determinar a importância de um termo em um documento, o grau de pertinência de um termo em um contexto (o quanto ele ajuda a definir um contexto) e o grau de relacionamento entre os termos (para descobrir quais os que melhor definem um contexto) [5].

Estas fórmulas baseadas na freqüência relativa (número de vezes que um termo aparece no documento dividido pelo número total de termos no documento) e na freqüência 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 quase 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 recuperação, há uma técnica que se utiliza de expansões semânticas de palavras [3]. Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto para a recuperação de informações. Para implementar esta técnica, [3] 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 recuperação e se as novas palavras acrescentadas realmente fazem parte do contexto. Segundo os experimentos de [3], algumas das novas palavras não fazem parte do contexto, o que pode fazer com que documentos irrelevantes sejam recuperados. A intervenção de especialistas humanos ou a geração automática dos contextos podem amenizar em parte tais obstáculos, como será discutido mais adiante.

Problemas também 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 de termos para caracterizar o mesmo contexto.

Há técnicas que utilizam modelagem de contextos alternativos [19], permitindo que contextos diferentes possam ser explorados em paralelo, e [16] cita uma técnica que combina conjuntos *fuzzy* diferentes (criados por vários especialistas) em um único resultante, através de operadores *fuzzy* de conjunção e disjunção.

Um *survey* sobre recuperação *fuzzy* de informações é apresentado em [7]. Nos casos apresentados, a lógica *fuzzy* (segundo [22] e [23]) serve 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.

Neste *survey* de [7] também são citados 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 *fuzzy*, a qual serve para representar o conhecimento do especialista no processo de expansão semântica da consulta (processo de encontrar termos relacionados semanticamente com os de entrada). São utilizados pesos (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, e 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 é expresso em valores *fuzzy*. Pode-se utilizar um limiar (*threshold*) para selecionar documentos na resposta.

Para medir a relação entre os conjuntos de termos (o de entrada, os de contextos e os representando documentos), [7] sugere duas medidas de similaridade ou compatibilidade:

- *set theoretic inclusion*: avalia se um termo está incluso ou não no conjunto, somando pontos pelos termos comuns; e
- *Euclidean distance*: representar os conjuntos de termos como vetores no espaço e determinar as distâncias (uma técnica difícil de ser usada com conjuntos bastante diferentes).

[7] cita ainda as redes neurais *fuzzy* como uma maneira de representar a relação entre termos e documentos. As entradas da rede 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 [17]): 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.

Em [18] e [14], existem informações mais detalhadas sobre técnicas e assuntos relativos a Recuperação de Informações.

Para o problema específico de recuperação de informações na Internet, há dois sistemas bastante interessantes. Um, o sistema FAB (descrito em [2]), utiliza como ferramentas de indexação aquelas já disponíveis na Internet. A técnica de recuperação utiliza duas bases de perfis (perfil é um conceito semelhante a contexto ou espaço conceitual): uma de perfis de usuários (com informações de preferências dos indivíduos) e uma de perfis de tópicos (criada automaticamente a partir de combinações das bases individuais, e contendo, para cada tópico, uma lista de palavras e seus respectivos graus de importância no tópico).

O problema principal do sistema Fab é utilizar como entrada a escolha de um tópico (denotando o assunto a ser pesquisado), o que pode levar à imprecisão semântica, como já discutido anteriormente. Conforme relatado no referido artigo, o sistema Fab apresentou problemas em alguns resultados devido à ambigüidade dos termos.

Já o sistema Referral Web (apresentado em [12]) usa uma combinação de termos próximos para precisar o significado dos mesmos, evitando erros de recuperação devido à ambigüidade das palavras fornecidas como entrada.

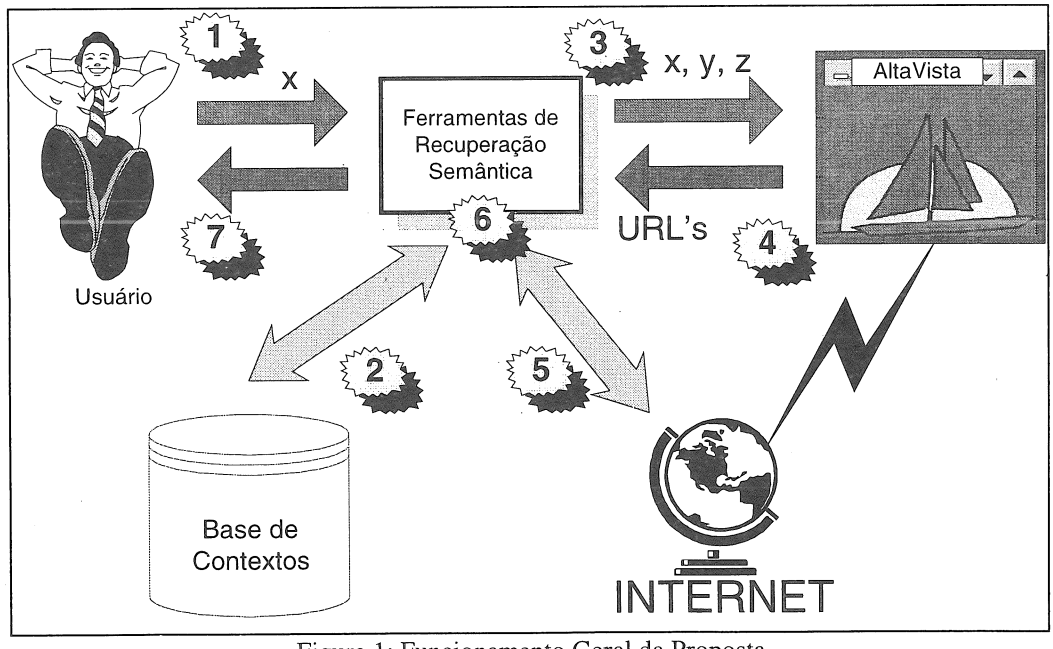

Figura 1: Funcionamento Geral da Proposta

4 A Proposta deste Trabalho

Para amenizar os problemas de imprecisões semânticas, deverá ser analisado também o contexto em que os termos são usados. Conforme [6], as técnicas para indexação devem gerar índices sensitivos mais intimamente relacionados ao real significado de um texto em particular e não baseados na presença de termos sem identificação do contexto. Outros autores

confirmam:

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

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

Na abordagem aqui descrita, o Contexto (ou espaço conceitual) será definido por um conjunto de palavras que representam o assunto ou a área do conhecimento (conforme a definição de [5]).

A recuperação semântica de documentos proposta neste trabalho será feita utilizando-se duas técnicas principais:

- a expansão semântica: para se determinar o contexto dos dados fornecidos como entrada e, conseqüentemente, o real interesse do usuário; e

- a lógica *fuzzy* (ou difusa): para permitir graduações nas relações entre termos, documentos e contextos e para se realizar o raciocínio de inferência.

Para armazenar os contextos, há uma base de contextos (ver figura 1). Já a definição dos contextos poderá ser feita de modo manual ou por uma ferramenta sob aprendizado supervisionado.

A figura 1 apresenta os principais passos e componentes da proposta, os quais serão descritos em detalhes a seguir.

4.1 A Recuperação Semântica

A expansão semântica será utilizada para aumentar o conjunto de termos fornecidos como entrada pelo usuário. Para tanto, serão utilizados contextos pré-definidos (armazenados na Base de Contextos), 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 recuperação dos documentos.

Para achar os documentos na rede, serão utilizados sistemas de indexação já disponíveis na Internet (no momento, está sendo utilizado o Sistema AltaVista®).

Em termos gerais, a Recuperação Semântica proposta aqui segue os seguintes passos (conforme pode ser visto também na figura 1):

1) receber, como entrada, palavras fornecidas pelo usuário;
2) procurar, na Base de Contextos, os contextos que contêm aquelas palavras;
3) enviar os termos de todos os contextos identificados, como parâmetros de busca a um sistema de indexação tipo AltaVista® (sendo que os termos serão associados por conjunção, ou seja, basta a presença de um dos termos);
4) receber as URL's dos documentos candidatos a resposta;
5) analisar o conteúdo de cada documento candidato, recuperado-os pela Internet;
6) comparar (por operadores *fuzzy*) cada documento analisado com os contextos identificados;
7) listar para o usuário os documentos melhor associados aos termos de entrada (por ordem decrescente de grau de relação).

Os contextos são representados como conjuntos *fuzzy* de termos ou palavras. Estes conjuntos representam o conhecimento de especialistas para definir que contextos possuem certos termos ou quais os termos que definem um contexto. 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 mais adiante.

Nesta proposta, os documentos recuperados na Internet deverão ser representados (internamente às ferramentas, depois de analisados os seus conteúdos originais) por conjuntos *fuzzy* compostos pelos termos que compõem o documento e o grau de pertinência de cada termo no documento. Este grau de pertinência é calculado pela freqüência relativa do termo no documento, isto é, o número de vezes em que aparece no documento dividido pelo número total de termos no documento. Entretanto, as *stop-words* deverão ser retiradas dos documentos para efeitos de cálculo e montagem do conjunto *fuzzy* representativo do documento.

A lógica *fuzzy* também permitirá 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 ("muito importante", "pouco relevante", etc).

Valores *fuzzy* também serão usados para determinar o grau de satisfação de um documento resultante da recuperação em relação à consulta original.

Por fim, os operadores da lógica *fuzzy* (conjunção e disjunção) permitirão comparar os vários conjuntos *fuzzy*

utilizados. Em contrapartida às duas medidas de similaridade apresentadas no início deste artigo, nesta proposta, são utilizadas medidas baseadas no raciocínio *fuzzy* discutido em [15] e [16].

Para os operadores de conjunção ($\wedge$), será utilizado o "produto algébrico" dos valores *fuzzy* associados a cada termo. Já para os operadores de disjunção ($\vee$), serão usados os operadores "soma limitada" e "máximo".

Para realizar o processo de recuperação semântica de documentos na Internet, há duas ferramentas implementadas sendo testadas separadamente (ainda não integradas): uma que faz a expansão semântica e o raciocínio *fuzzy* e outra para comunicação com o Sistema AltaVista®.

A primeira ferramenta desempenha os passos 1,2,5,6 e 7. Ou seja, ela recebe os termos de entrada, fornecidos pelo usuário, compara-os com os contextos definidos na Base de Contextos, compara os contextos identificados com os documentos candidatos (sugeridos pela ferramenta AltaVista) e retorna a resposta ao usuário.

A segunda, por sua vez, envia parâmetros de consulta ao Sistema AltaVista® e recebe as páginas HTML como resposta (correspondendo aos passos 3 e 4 definidos anteriormente). Dentre as páginas recebidas, a ferramenta extrai as URL's dos documentos a serem analisados.

A razão da separação destas ferramentas é a demora no processamento de cada uma delas. Entretanto, isto não impede que experimentos sejam feitos e que a abordagem seja testada. Os documentos utilizados pela primeira ferramenta são documentos-texto extraídos de páginas reais da Internet, eliminando-se as marcas da linguagem HTML.

4.2 A Base de Contextos

A Base de Contextos é uma estrutura que armazena as palavras ou termos que caracterizam um mesmo contexto, inclusive relacionando sinônimos para amenizar a imprecisão semântica. Assim, é possível percorrer esta estrutura e identificar quais são as palavras que pertencem a um determinado contexto ou quais os contextos de um certo termo.

A implementação atual da base de contextos, como está sendo utilizada pela ferramenta de recuperação semântica, é muito simples. Cada contexto é um conjunto *fuzzy* de termos (que o definem). Cada contexto é nomeado e identificado por uma palavra no momento de sua criação, mas este identificador fica transparente ao usuário da ferramenta e não interfere no processo de inferência. Associado a cada termo dentro de um contexto, há um valor *fuzzy* que representa o grau de importância do termo neste contexto. A relação entre contextos e termos é do tipo N:M, pois um termo pode aparecer em mais de um contexto.

Alternativas para melhorar tal estrutura estão sendo testadas. Uma delas foi estruturar a Base de Contextos como uma rede semântica, onde os nodos são as palavras e os elos representam relações entre palavras de um mesmo contexto. Entretanto, como pode haver o caso de uma palavra estar relacionada com duas outras em contextos diferentes, há a necessidade de se caracterizar o contexto de cada elo.

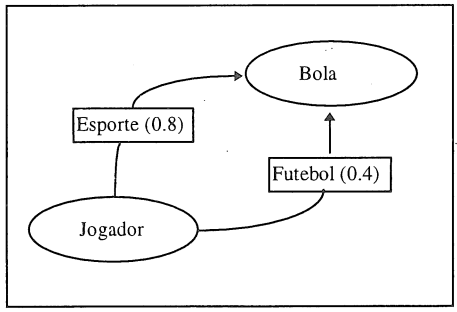

Figura 2: Exemplo de relações entre termos

Em [21], é apresentada uma implementação baseada nesta última alternativa. Assim, duas palavras podem estar relacionadas entre si nos mais diversos contextos. Cada elo possui uma indicação do contexto da relação (inclusive determinando o tipo da relação; por exemplo, causa, efeito, sinônimo, antônimo, etc) e um valor *fuzzy*, o qual caracteriza o grau de associação entre as palavras correspondentes. Na figura 2, pode-se ver um exemplo desta situação: o termo "jogador" está associado ao termo "bola" por dois contextos diferentes, e as duas associações possuem graus diferentes de intensidade ou importância.

Alguns problemas foram detectados, obrigando a maiores estudos desta alternativa. O principal deles é que o contexto de cada elo pode ser especificado por uma palavra, mas isto traz de volta o problema da escolha desta palavra,

gerando buscas contextuais recursivas para tratamento de sinônimos. Da mesma forma, ainda não se tem uma maneira adequada para identificar cada "contexto" que não o uso de uma única palavra.

4.3 A Definição dos Contextos

Os contextos (conjuntos de termos presentes na Base de Contextos) podem ser definidos de duas formas: manualmente (por uma pessoa) ou por aprendizado supervisionado. Este último modo ocorre quando um especialista seleciona textos de um mesmo contexto e submete a uma ferramenta automatizada que extrai os termos que melhor definem o tal contexto.

A montagem de contextos manual é delicada, devendo ser tarefa de um especialista da área, o qual deve conhecer bem os assuntos em questão. Neste caso, o especialista deve selecionar as palavras que identificam cada contexto e associar pesos a elas (grau de importância das palavras no assunto ou probabilidade de que ocorram em documentos do assunto).

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

Um dos problemas da construção de contextos manual é que ela exige tempo e disposição. Além disto, o especialista pode não ser capaz de identificar todas as palavras relevantes do contexto, esquecendo algumas ou então dando ênfases equivocadas aos termos (maior ou menor que a real).

Esse tipo de problema pode ser amenizado com a definição automática dos contextos. A princípio, a montagem automática de contextos foi desenvolvida visando substituir o especialista. Apesar disto, sugere-se que esta seja uma etapa complementar à montagem manual, para facilitar o trabalho do especialista, e que seja supervisionada por um especialista (uma técnica de aprendizado supervisionado).

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

Em trabalho anterior [21], é apresentada uma ferramenta que identifica relações entre termos a partir da análise de um conjunto de documentos pertencentes a um mesmo contexto (ou assunto). Para isto, o especialista deve analisar previamente os documentos separando-os por contexto (como num trabalho de supervisão). Há a possibilidade ainda de se utilizar definições já prontas de contextos, como por exemplo dicionários técnicos.

As técnicas empregadas na extração de relações automáticas entre palavras de um mesmo contexto é similar às técnicas discutidas em [5] e [17]). São métodos estatísticos, que baseiam-se na análise de ocorrência das palavras nos documentos.

Ao todo, o processo de montagem automática de contextos possui três etapas distintas:

- *a identificação de palavras nos documentos,*
- *a determinação do grau de relação entre as palavras e o documento que as contém,*
- *a análise das relações entre as palavras.*

Segundo [4], as palavras que aparecem repetidamente em um único documento e as palavras que aparecem em muitos documentos são boas candidatas. Algumas, entretanto, devem ser desconsideradas (por exemplo, aquelas conhecidas como *"stop-words"*), conforme sugestão de [5]. As *stop-words* podem variar, dependendo do domínio a ser analisado; verbos ou até mesmo expressões podem ser desprezadas.

Após filtradas as palavras que devem fazer parte do processo, é realizada uma análise de co-ocorrência das palavras nos documentos. Esta análise permitirá identificar o grau de relação de cada termo e o contexto em questão.

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

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

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

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

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

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

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

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

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

Um valor limite (um limiar) deverá ser utilizado para cortar fora relações com graus muito baixos.

Uma versão inicial da ferramenta de Definição de Contextos foi implementada e testada. Em [21], são relatados experimentos com esta ferramenta, utilizando documentos com informações sobre os sintomas que podem ser causados pela utilização de drogas. O contexto de cada droga é montado e após utilizado para recuperar prontuários médicos, a fim de descobrir pacientes com sintomas similares. Os resultados foram satisfatórios porque especialistas da área avaliaram os relacionamentos descobertos como sendo válidos e realistas.

Uma alternativa é utilizar a técnica do centróide (discutida em [17]). O centróide de um conjunto de documentos é 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 pode ser usado como o conjunto *fuzzy* que define o tal contexto.

5 Conclusões

Foram realizados experimentos com as ferramentas em separado e, destes, alguns resultados parciais podem ser concluídos.

A ferramenta de recuperação semântica tem-se mostrado útil ao escolher um conjunto maior e mais significativo de termos para a recuperação. Além disto, o significado dos termos pode ser melhor precisado com a ajuda das relações entre os termos (na base de contextos), diminuindo assim os erros por ambigüidade.

Foram feitos experimentos com textos retirados de artigos e reportagens de jornais de páginas da Internet para avaliar a ferramenta de expansão semântica e raciocínio *fuzzy*. A Base de Contextos foi gerada pelos autores, sem conhecimento prévio dos textos e conforme seu próprio entendimento sobre quais termos servem para definir os contextos.

Com base nos experimentos realizados, observou-se que a ferramenta atingiu graus satisfatórios de precisão (próximo de 0,9) e abrangência (próximo de 0,6), na média geral dos experimentos realizados.

Para a avaliação dos graus de *precison* 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.

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

Desvios podem ocorrer se os termos dos contextos não forem bem determinados. Além disso, 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. Algumas consultas nos experimentos retornaram documentos irrelevantes por falta de contextos definidos.

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

Entretanto, tais técnicas somente terão resultados satisfatórios se o conjunto-amostra para extração das relações entre os termos for bem escolhido. De novo, recai-se na dependência de um especialista humano.

Problemas também ocorrem quando uma palavra aparece como "cabeça" num contexto e como elemento em outro, numa situação de transitividade. Futuramente, serão utilizados conjuntos e operadores *fuzzy* para determinar o grau de uma relação *x-z* (caso somente se disponha das relações *x-y* e *y-z*).

Já a lógica *fuzzy* permitiu 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.

Problemas podem ocorrer devido aos procedimentos de determinação dos graus *fuzzy* por pessoas (tanto para os termos de entrada pelo usuário, quanto para os termos dos contextos pelos especialistas).

A ferramenta implementada apresentou limitações quanto ao tempo de resposta. Apesar de as consultas todas terem sido realizadas de maneira rápida (tendo os documentos já analisados e comparados com os contextos previamente), a comparação entre contextos e documentos é bastante demorada. Para alguns casos estudados (5 contextos contra 36 documentos), o tempo de processamento chegou a levar 15 minutos. Trabalhos futuros deverão melhorar tal desempenho.

No que se refere a acesso aos documentos da Internet, o desempenho foi ruim, devido principalmente ao tráfego da rede. Para amenizar situações em que há inúmeros documentos a analisar, poderão ser analisados somente alguns do topo da lista ou então poderão ser analisadas somente algumas partes dos documentos (por exemplo as cem primeiras palavras), conforme sugestão de [9]. A premissa de [9] é a de que no início do documento é que se encontra sua descrição.

Por fim, falta integrar as duas ferramentas (o que não foi feito ainda por razões de desempenho) para observar o comportamento em situações com condições reais.

Como contribuições principais deste trabalho, pode-se citar, em primeiro lugar, a diminuição na imprecisão semântica. Isto porque não é necessário escolher um assunto ou classe, seja na consulta ou seja para classificar documentos. Também porque o problema do uso de sinônimos (na consulta ou nos documentos) é tratado pela definição dos contextos, os quais admitem vários termos com pesos diferentes.

Outra contribuição é a inferência automática (através do raciocínio *fuzzy*), realizada sobre as relações N:M entre consultas, contextos e documentos. Não é necessário obter e representar conhecimento sobre estas relações a partir de um especialista humano. O único trabalho do especialista aparece na definição dos contextos, como discutido na seção 4.3. Consultas, contextos e documentos são comparados automaticamente para realizar a inferência (para descobrir que documentos satisfazem melhor à consulta).

7 Referências Bibliográficas

[1] BAEZA-YATES, Ricardo. An extended model for full text databases. **Journal of the Brazilian Computer Society**, v.2, n.3, Abr 1996.

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

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

[4] 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/)

[5] 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/)

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

[7] CROSS, Valerie. Fuzzy information retrieval. **Journal of Intelligent Information Systems**, 5, 1994.
[8] FURNAS, G. W. et alli. The vocabulary problem in human-system communication. **Communications of the ACM**, v.11, n.30, Nov 1987.
[9] HARDY, Darren R.; SCHWARTZ, Michael F. Essence: a resource discovery system based on semantic file indexing. **Proceedings**. Winter USENIX, San Diego, CA, 25-29 de Janeiro de 1993.
[10] IIVONEN, M. Searches and Searches: Differences Between the Most and Least Consistent Searches. ***SIGIR FORUM*** *95*. p149-157. 1995.
[11] JAKOBSON, R. **Lingüística e comunicação**. Ed. Cultrix. 1970.
[12] KAUTZ, H.; SELMAN, B.; SHAH, M. *"Referral web: combining social networks and collaborative filtering"*. **Communications of the ACM,** v.40,n.3, Mar 1997.
[13] KENT, W. **Data and reality**. North-Holland. 1978.
[14] KORFHAGE, Robert R. **Information storage and retrieval.** John Wiley & Sons. 1997.
[15] NAKANISHI, H.; TURKSEN, I. B.; SUGENO, M. A review and comparison of six reasoning methods. **Fuzzy Sets and Systems**, 57, 1993.
[16] OLIVEIRA, Henry M. **Seleção de entes complexos usando lógica difusa**. Instituto de Informática da PUC-RS, Porto Alegre, Julho de 1996. (dissertação de mestrado)
[17] SALTON, G.; McGILL, M. J. **Introduction to Modern Information Retrieval**. McGraw-Hill. New York. 1983.
[18] SPARCK-JONES, Karen; WILLETT, Peter (eds). **Readings in Information Retrieval**. Morgan Kaufmann Publishers. 1997.
[19] WIEBE, Janyce; HIRST, Graeme; HORTON, Diane. Language use in context. **Communications of the ACM**, v.39, n.1, Janeiro 1996.
[20] WILLIE, S; BRUZA, P. Users' Model of the Information Space: the Case for Two Search Models. ***SIGIR FORUM*** *95*. P205-211. 1995.
[21] WIVES, Leandro K; LOH, Stanley. Hyperdictionary: a knowledge discovery tool to help information retrieval. In: STRING PROCESSING AND INFORMATION RETRIEVAL – A SOUTH AMERICAN SYMPOSIUM, SPIRE, 1998. **Proceedings...** Washington: IEEE Press, 1998.
[22] ZADEH, L. A. Fuzzy sets. **Information and Control**, 8, pp.338-353. 1965.
[23] ZADEH, Lotfi A. Outline of a new approach to the analysis of complex systems and decision processes. **IEEE Transactions on Systems, Man and Cybernetics**, Washington, v.SMC-3, n.1, p.28-44, Jan. 1973