Desenvolvimento e Avaliação de uma Estrutura Multitesauro para Recuperação de Informações

Luiz Augusto Sangoi Pizzato e **Vera Lúcia Strube de Lima**
Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação
Av. Ipiranga, 6681 - Prédio 16 - Partenon
90619-900 Porto Alegre - RS - Brasil
{pizzato,vera}@inf.pucrs.br

Abstract

This article concerns the proposition, development and evaluation of a multithesaurus structure to be used in information retrieval applications. This multithesaurus structure was initially based on ISO 2788 standards (since mostly thesauri available today can be easily converted to it) but it was gradually getting its own shape, once its aim is to help information retrieval tasks. In order to evaluate this structure we use information retrieval applications combined with a query expansion method that walks through the multithesaurus structure and assigns different weights to its semantic relations. This article presents the current status of this research, as well as future works to be developed.

Keywords: Information Retrieval, Thesauri, Query Expansion, Semantic Relations.

Resumo

Neste artigo são apresentados a proposta e o desenvolvimento de uma estrutura multitesauro para ser utilizada em aplicações de recuperação de informações. A estrutura multitesauro foi inicialmente baseada no padrão ISO 2788 e foi gradualmente assumindo sua própria forma, uma vez que deve ser utilizada em tarefas de recuperação de informações. De modo a avaliar esta estrutura foi criado um método de expansão de consultas combinado com aplicações de recuperação de informações. O método de expansão de consultas utilizado tem como principal característica utilizar diferentes pesos para as relações semânticas definidas na estrutura multitesauro. Este trabalho apresenta os passos realizados no desenvolvimento e na avaliação da estrutura até a sua forma atual.

Palavras-chave: Recuperação de Informações, Tesauros, Expansão de Consultas, Relações Semânticas.

1 Introdução

Tesauros são importantes na recuperação de informações (RI), segundo Baeza-Yates & Ribeiro-Neto em [1], pois eles podem ser utilizados para obter melhor compreensão de alguns termos de uma consulta em sistemas de RI. Contudo, a utilização de um tesauro depende dos tipos de relações semânticas nele presentes. Portanto, a organização de um tesauro representa um tópico de suma importância para a RI. Neste trabalho mostraremos as etapas de construção de uma estrutura multitesauro que se apresentou útil para a RI, juntamente com um método de expansão de consultas que utiliza as relações definidas nesta mesma estrutura.

Ao desenvolver a estrutura multitesauro nos preocupamos em que esta pudesse ser utilizada em diferentes ambientes e também seus arquivos pudessem ser facilmente transmitidos pela Internet. A preocupação com que a estrutura (e os tesauros definidos com a mesma) seja utilizada na Internet é clara quando a finalidade de sua utilização é a RI pois, atualmente, os *sites* de busca na Internet são os ambientes de RI mais conhecidos e utilizados. O formato XML é facilmente transmitido pela Internet através de servidores HTTP, tornando-se a escolha normal quando se tem a preocupação de uso pela rede.

Optou-se por utilizar a expansão de consulta, para mostrar a utilidade da estrutura na RI, pela grande quantidade de trabalhos que envolvem este tópico e o uso de tesauros (por exemplo, [8], [12], [2], [5], [11] e [7]). O método que apresentaremos para a expansão de consulta foi desenvolvido para utilizar as relações definidas na estrutura e avaliar a importância das mesmas. O método tem como principal característica a atribuição de pesos para cada tipo de relação definida. Desta forma é possível quantificar a importância de uma dada relação, para a RI.

Este artigo está organizado em seis seções, sendo a primeira esta introdução. Na segunda seção são apresentadas as etapas iniciais que levaram à definição atual da estrutura multitesauro. A terceira seção apresenta o método de expansão de consultas desenvolvido. Na quarta seção é demonstrado o funcionamento do método proposto através de um exemplo. A quinta seção apresenta o modo como avaliamos nossa estrutura. Na sexta e última seção são apresentadas uma análise do que foi obtido por nosso trabalho até o presente momento e as futuras etapas que serão desenvolvidas.

2 Desenvolvendo uma estrutura multitesauro

Como descrito na introdução, a estruturação de um tesauro é uma ação importante quando este será utilizado em tarefas de RI. Atualmente estamos desenvolvendo uma estrutura que deverá ser utilizada em um sistema de RI pertencente ao contexto do projeto SEMA[1]. Nossa estrutura é baseada no padrão ISO 2788 definido em [6], dado que, atualmente, a maioria dos tesauros disponíveis pode ser facilmente convertida para este padrão.

Algumas instituições generosamente ofereceram seus tesauros para serem utilizados em nossa pesquisa. Os tesauros obtidos apresentam-se de diferentes formas, mas são igualmente úteis:

- "Vocabulário Controlado Básico do Senado" ou VCBS é um tesauro bastante organizado que contém a grande maioria das características descritas na norma ISO 2788. A lista de palavras contidas no VCBS cobre diferentes áreas do conhecimento, e é utilizada pelos profissionais da Biblioteca do Senado Federal na catalogação do material existente em sua biblioteca. Mais informações sobre este tesauro podem ser encontradas no endereço *http://webthes.senado.gov.br/thes/default.htm*, que funciona como interface *Web* para consultas a este tesauro.

- "Vocabulário Controlado USP" ou VCUSP contém uma grande quantidade de conceitos que são muito úteis aos profissionais de biblioteca ajudando os mesmos no processo de indexação de documentos. Este tesauro cobre muitas áreas diferentes através de relações de equivalência e hierarquia. O VCUSP é um produto elaborado por um grupo de bibliotecários do Sistema Integrado de Bibliotecas da USP (SIBi/USP) e comercializado em CDROM. Mais informações podem ser encontradas no *Web Site* do SIBi/USP em *http://www.usp.br/sibi/*

- "Lista de descritores da PUCRS" ou LDPUCRS é uma lista com 55565 termos contendo somente a relação de equivalência expressa entre os termos. O LDPUCRS não foi construído para ser um tesauro mas para ser uma lista de termos autorizados, a ser utilizado pela biblioteca da Universidade. Por isso muitos termos não contêm relações com outros termos na lista de descritores. Decidiu-se utilizar a lista, pois a relação de equivalência é útil e mesmo os termos não relacionados podem ser utilizados como indicações de termos importantes, principalmente em termos compostos.

[1] Mais informações em *http://www.inf.pucrs.br/grupos/linatural/sema/*

- "Lista de Termos Obtida por Cálculo de Similaridade Sintática" ou LTOCSS é um tesauro construído de forma automática com uso das técnicas descritas por Grefenstette em [4] e adaptadas para o português por Gasperin em [3]. Este tesauro é dependente de corpus e foi construído utilizando um corpus sintaticamente etiquetado do jornal "Folha de São Paulo" do ano de 1994. A principal característica deste tesauro é que seus diversos termos são relacionados com outros termos, de acordo com uma medida de similaridade. Esta medida é representada por um valor real variando de 0 (termo não relacionado) até 1 (termo perfeitamente relacionado).

Depois de analisadas as principais características dos tesauros citados, passamos a definir a estrutura que foi utilizada neste estudo. Decidimos começar pela norma ISO 2788, pois os tesauros obtidos são facilmente transpostos para este padrão. Deste modo foi definida a DTD/XML da Fig. 1, que pode ser melhor entendida na representação de um de seus documento XML na Fig. 2.

```
<!ELEMENT THESAURUS(TERM+)>
<!ELEMENT TERM(SN?,UF*,USE*,BT*,NT*,RT*)>
<!ELEMENT SN(CDDATA)>
<!ELEMENT UF(EMPTY)>
<!ELEMENT USE(EMPTY)>
<!ELEMENT BT(EMPTY)>
<!ELEMENT NT(EMPTY)>
<!ELEMENT RT(EMPTY)>

<!ATTLIST TERM  term CDATA #REQUIRED>
<!ATTLIST BT    term CDATA #REQUIRED>
<!ATTLIST NT    term CDATA #REQUIRED>
<!ATTLIST USE   term CDATA #REQUIRED>
<!ATTLIST UF    term CDATA #REQUIRED>
<!ATTLIST RT    term CDATA #REQUIRED>
```

Fig. 1: Estrutura inicial baseada na norma ISO 2788

```
<THESAURUS>
    <TERM term="A">
        <SN>Scope Note</SN>
        <UF term="A2"/>
        <BT term="B"/>
        <NT term="C"/>
        <RT term="D"/>
    </TERM>
    <TERM term="A2">
        <USE term="A"/>
    </TERM>
    <TERM term="E"/>
</THESAURUS>
```

Fig. 2: Exemplo de um documento XML definido de acordo com a DTD/XML da Fig. 1

A estrutura da Fig. 2 apresenta as definições dos termos "A", "A2" e "E" e suas relações. O termo "A" tem: uma nota explicativa nomeada SN (de *Scope Note*); uma relação de equivalência UF (de *Used For*) com o termo "A2"; uma relação de termo mais genérico BT (de *Broader Term*) com o termo "B"; uma relação de termo mais específico NT (de *Narrower Term*) com o termo "C"; e uma relação de termo associado RT (de *Related Term*) com o termo "D". O termo "A2" tem uma relação de equivalência USE com o termo "A". O termo "E" é definido sem que exista qualquer relação entre ele e o outro termo.

A diferença entre as relações USE e UF é a mesma definida pela norma ISO 2788: um termo "A" preferencial relaciona-se com um termo "A2" não preferencial através de uma relação UF. Enquanto que a relação USE ocorre de modo inverso, um termo não preferencial remete a um termo preferencial através deste tipo de relação. Um termo preferencial, segundo a norma ISO 2788, deve representar um conceito único, enquanto que um termo não preferencial deve estar relacionado a um termo preferencial.

Mesmo que seja possível preencher todos os campos disponíveis nesta estrutura com o conteúdo dos tesauros, considera-se que um dos campos, em particular, não é importante para tarefas de RI. O campo SN é importante quando utilizado no processo manual de catalogação em bibliotecas, ao informar ao profissional

sobre a utilização correta do termo, contudo estas notas não parecem fornecer nova informação semântica que possa ajudar em tarefas automáticas. Por este motivo foi decidido excluir o campo SN da estrutura.

Estudando o tesauro VCUSP, foi constatado que deveria ser oferecida uma outra maneira de representar as relações BT e NT. O VCUSP é distribuído como um banco de dados ordenado, e estruturado como uma árvore de conceitos; então foi decidido adicionar esta mesma característica a nossa estrutura. Para representar as relações BT e NT a estrutura passou a aceitar a inserção das etiquetas de termos entre as etiquetas de início (*start-tag*) e fim (*end-tag*) de termo. Esta característica é mais bem expressa na Fig. 3.

```
<TERM term="A">
    <TERM term="B">
        <UF term="B2"/>
        <TERM term="D"/>
    </TERM>
    <TERM term="C"/>
</TERM>
```

Fig. 3: Uma representação diferente para as relações BT/NT

A representação da Fig. 3 informa que "B" e "C" são NT de "A" (e "A" é BT de "B" e "C"), enquanto "B" tem "B2" como termo equivalente e "D" como NT (e "B" é BT de "D").

Outra característica adicionada à estrutura, que foi verificada durante o processo de análise dos tesauros, foi um valor agregado à relação RT. Este valor é um atributo nas relações RT, importante para representar a medida de similaridade presente em alguns tesauros construídos de forma automática como no caso do LTOCSS. E novamente, o valor agregado é qualquer real entre 0, que representaria uma relação inexistente, e 1, que seria uma relação RT perfeita. Quando o valor não é informado em uma relação RT, será assumido o valor padrão 1 uma vez que, possivelmente, esta relação provém de um tesauro manual. A representação da Fig. 4 exemplifica a utilização da relação RT com e sem um valor agregado.

```
<TERM term = "A">
    <RT term = "B" value = "0.87">
    <RT term = "C">
</TERM>
```

Fig. 4: Exemplo de utilização de um valor agregado à relação RT

O estágio atual de desenvolvimento da estrutura pode ser representado através do DTD/XML mostrado na Fig. 5.

```
<!ELEMENT THESAURUS (TERM+)>
<!ELEMENT TERM (TERM*,BT*,NT*,USE*,UF*,RT*)>
<!ELEMENT BT   (EMPTY)>
<!ELEMENT NT   (EMPTY)>
<!ELEMENT USE  (EMPTY)>
<!ELEMENT UF   (EMPTY)>
<!ELEMENT RT   (EMPTY)>

<!ATTLIST TERM  term CDATA #REQUIRED>
<!ATTLIST BT    term CDATA #REQUIRED>
<!ATTLIST NT    term CDATA #REQUIRED>
<!ATTLIST USE   term CDATA #REQUIRED>
<!ATTLIST UF    term CDATA #REQUIRED>
<!ATTLIST RT    term CDATA #REQUIRED>
<!ATTLIST RT    value CDATA "1">
```

Fig. 5: Atual DTD/XML da estrutura multitesauro

Na próxima seção apresentaremos o modo como avaliamos nossa estrutura, uma vez que esta será utilizada em tarefas de RI.

3 Método proposto

A utilidade da estrutura na área de RI deve ser mensurada, ao nosso ver,através do uso dessa estrutura em algum sistema de RI. O método utilizado para avaliar nossa estrutura não é diferente disto. Foi construído um sistema de expansão de consulta que utiliza diferentes tesauros organizados de acordo com nossa estrutura e foram avaliadas as medidas de precisão[2] e *recall*[3] obtidas em um sistema de RI.

A ferramenta de expansão de consulta foi nomeada QET (um acrônimo para *Query Expansion Tool*), e foi desenvolvida orientada a objetos, em Borland Kylix 2.0 – Open Edition, em uma máquina com sistema operacional Linux.

O QET carrega qualquer tesauro definido de acordo com nossa estrutura, sendo também possível carregar um tesauro separado em diferentes arquivos, o que nos dá duas possibilidades interessantes:

1. poder carregar pequenas porções de um mesmo tesauro que foram transmitidas pela Internet;
2. poder utilizar diferentes tesauros como se fossem um grande e único multitesauro.

A característica de carregar porções de tesauros demonstra que uma estrutura padrão que cobre diferentes aspectos dos tesauros é útil, principalmente, quando a utilização de tesauros provenientes de diferentes fontes acarreta o surgimento de um único multitesauro.

Assim como Mandala et al. em [9], utilizamos diferentes tesauros (e de forma conjunta) em nossos testes. Desta forma todos os tesauros contribuíram para os resultados obtidos. Acreditamos que a utilização conjunta gerou melhores resultados do que obteríamos se os tesauros fossem utilizados de forma separada. Veja a seção 5 onde é feita uma análise dos resultados obtidos.

Diferentes tesauros (ou porções de tesauros) podem ter os mesmos termos e também as mesmas relações entre termos. Quando estas situações ocorrem o sistema somente considera novas relações e termos; portanto, no "grande tesauro" do sistema, os termos não estão duplicados.

Foi desenvolvido um método de expansão de consultas com tesauros utilizando diferentes pesos para diferentes tipos de relações. Este método utiliza a seguinte heurística:

- Consideram-se T_1 como o conjunto dos termos t originais de uma consulta e R_k um valor maior que 0 e menor que 1, referente ao tipo de relação entre dois termos. São encontrados os termos:
  - $t \in T_{n+1}$ relacionados aos termos $t \in T_n$ através de relações R_n.
- Para cada caminho de termos e relações $R_1, R_2, ..., R_{k-1}$ entre os termos $t \in T_1$ e um termo $t \in T_k$, são calculados valores β iguais a:

$$\beta = \prod_{i=1}^{k-1} R_i$$

  .
- Como um termo t pode ser encontrado através de diferentes caminhos, provenientes de um mesmo, ou de diferentes termos da consulta original, existe para cada termo um valor δ que representa a soma de todos os valores β associados a este termo.
- Como o número de conjuntos T_n tende a ser muito grande, limita-se a adicionar valores β que sejam maiores que um determinado valor σ. Desta forma o processo de busca por novos termos em um caminho de relações R_n encerra-se quando o valor de β for menor do que σ. Observa-se que, como R_n está no intervalo $[0, 1)$, o valor β tende a 0.
- Um termo t é adicionado na consulta expandida se o valor δ deste for maior que um limiar λ proposto.
- O valor β representa a importância de um termo $t \in T_n$ dado um termo $t \in T_1$, enquanto que o valor δ representa a importância de um termo $t \in T_n$ dado todos os termos $t \in T_1$. O valor σ regula o mínimo de importância que deve ser considerado para β no cálculo de δ.

Este método é implementado no algoritmo da Fig. 6 e seu funcionamento é melhor visualizado na Fig. 7.

O algoritmo da Fig. 6 implementa o método explicado anteriormente, através de duas funções principais. A função "InsereTermos" recebe como parâmetro um termo t_n e um valor β. O objetivo desta função é inserir, em uma lista de termos L, todos os termos que relacionam-se diretamente, ou indiretamente, com o termo t_n passado como parâmetro. Observa-se que, para que ocorra a inserção de termos relacionados de forma indireta, esta função é chamada de forma recursiva, tendo como parâmetro os termos v diretamente

[2]Número de documentos relevantes encontrados dividido pela quantidade total de documentos encontrados.
[3]Número de documentos relevantes encontrados dividido pela quantidade total de documentos relevantes.

```
Conjunto de Lexemas da linguagem  D = {l_1, l_2, ..., l_k}
Conjunto de Termos do Tesauro     T = {t_1, t_2, ..., t_k  |  t_i ∈ D}
Conjunto de Relações do Tesauro   R = {(u, v)  |  u, v ∈ T}
Consulta                          C = {c_1, c_2, ..., c_k  |  c_i ∈ D}
Lista de Termos                   L = {t_1, t_2, ..., t_k  |  t_i ∈ T}
Lista de Deltas                   D = {δ_t1, δ_t2, ..., δ_tk  |  δ ∈ [0, ∞],  t_i ∈ L}
Consulta Expandida                CE = {t_1, t_2, ..., t_k  |  t_i ∈ T}

Entradas(λ,σ,PesosRelações)

função Expansão(C)
    ∀ c_n ∈ C
        δ_cn ← 1
        InsereTermos(c_n, β = 1)
    ∀ t_n ∈ L
        se δ_tn > λ
            CE = CE ∪ {t_n}
    Retorna CE

função InsereTermos(t_n,β)
    L = L ∪ t_n
    ∀ (t_n, v) ∈ R
        β_v ← β×PesoRelação[TipoRelação[(t_n, v)]]
        se β_v > σ
            δ_v ← δ_v + β_v
            InsereTermos(v,β_v)
```

Fig. 6: Algoritmo para a expansão de consulta

relacionados com t_n. O parâmetro β irá regular o valor δ_v associado a um termo v. O peso dos tipos de relações é um valor igual ou maior que 0 e menor que 1; desta forma o valor β_v para cada termo v, parâmetro na chamada recursiva da função "InsereTermos", será sempre menor a cada novo nível de recursão. Esta função recursiva é interrompida quando o valor de β_v é menor que um valor σ previamente estabelecido.

A função "Expansão", definida no algoritmo, recebe uma consulta C como parâmetro. Depois de utilizada a função "InsereTermos" para todos os termos de C, e assim criada a lista de termos L, são adicionados à consulta expandida os termos de L cujo valor δ for maior que um valor λ pré estabelecido.

4 Um exemplo detalhado

Na Fig. 7 podemos acompanhar o processo de inserção de palavras e cálculo dos pesos das mesmas. Quando pesquisado sobre "Acidente de carro" o sistema separa os *tokens*[4] da consulta e remete aos seus termos relacionados. Considerando os pesos, a fim de exemplo, para as relações USE, UF, NT, BT e RT como sendo respectivamente 1, 1, 0.6, 0.3 e 0.1, um peso 1 aos termos originais da consulta, um valor λ, para inserção do termo na consulta expandida, de 0.7, e um valor mínimo σ de 0.05. As relações dos termos inseridos a partir de "Acidente" estão representadas por setas contínuas e seus pesos β por valores em itálico, enquanto que as relações provenientes de "Carro" são representadas por setas tracejadas.

O processo de inserção de termos na lista de termos, e o cálculo de seus valores δ associados, ocorrem em profundidade no tesauro. Por exemplo, o termo "Acidente de Trânsito" é encontrado pela relação NT do termo inicial da consulta "Acidente". A "Acidente de Trânsito" é associado um valor β, de 0.6, correspondente à relação pela qual este foi encontrado. O processo continua encontrando "Automóvel" pela relação RT de "Acidente de Trânsito", e o valor β, neste caso 0.06, correspondente aos pesos de NT e RT multiplicados. Observa-se que peso de β é equivalente ao produto dos pesos das relações encontradas no caminho entre o termo inicial da consulta e um termo atual.

O processo de busca em profundidade dos termos continua até que o valor β calculado for menor que um valor σ predeterminado. Esta característica pode ser observada no exemplo da Fig. 7, na falta de um valor itálico (proveniente de "Acidente") de β para o termo "Veículo"; isto ocorre, pois o caminho entre este termo e "Acidente" é composto pelas relações NT, RT, BT e o valor β para este caminho é de 0.018, menor que o valor de σ, definido como 0.05 para este exemplo.

São pesquisados todos os termos relacionados direta ou indiretamente com os termos da consulta original.

[4]Itens lexicais mínimos na análise de uma sentença, incluindo palavras, números e sinais de pontuação.

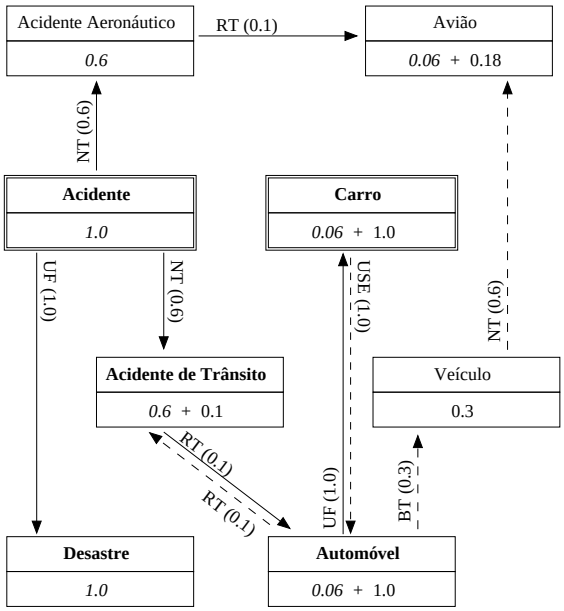

Fig. 7: Exemplo do funcionamento da expansão de consulta

Tabela 1: Definição dos melhores pesos para os tipos de relações

Nro.	T(1)	USE	UF	NT	BT	RT	Abrangência	Precisão
142	1.00	0.90	0.90	0.90	0.90	0.30	0.5319	0.4902
143	1.00	0.90	0.90	0.90	0.90	0.20	0.5319	0.5102
144	1.00	0.90	0.90	0.90	0.90	0.10	0.5319	0.6579
145	1.00	0.90	0.90	0.90	0.80	0.30	0.5319	0.5102
146	1.00	0.90	0.90	0.90	0.80	0.20	0.5319	0.5102
147	1.00	0.90	0.90	0.90	0.80	0.10	0.5319	0.6579
148	1.00	0.90	0.90	0.90	0.70	0.30	0.5319	0.5102
149	1.00	0.90	0.90	0.90	0.70	0.20	0.5319	0.6410
150	1.00	0.90	0.90	0.90	0.70	0.10	0.5319	0.6579
378	1.00	0.90	0.70	0.60	0.60	0.10	0.5319	0.3788
384	1.00	0.90	0.70	0.60	0.40	0.10	0.5319	0.4902
390	1.00	0.90	0.70	0.60	0.20	0.10	0.4894	0.6053

Observa-se que o valor β é armazenado em uma variável δ particular a cada termo. Caso ocorra que um termo seja relacionado com mais de um termo original, os valores de δ será a soma de todos os β encontrados entre o termo e os termos originais. Esta característica ocorre para o exemplo da Fig. 7 nos termos "Carro", "Acidente de Trânsito", "Automóvel" e "Avião".

Ao final de toda a análise os termos com peso acima do valor λ serão inseridos na consulta expandida. Na Fig. 7 temos estes termos representados em negrito.

Uma vantagem deste método sobre outros métodos existentes como, por exemplo, o descrito por Robin & Ramalho em [11], é que o método desenvolvido permite utilizar termos com relação indireta com os termos da consulta.

5 Avaliação

Para avaliar a estrutura assumimos que é possível quantificar a importância de um tipo de relação através dos pesos que forem estabelecidos para ela. Em uma tentativa de quantificar a importância de cada tipo de relação, na expansão de consulta, foi realizado alguns testes utilizando a consulta "Acidente de automóvel". Os pesos para as relações eram modificados automaticamente enquanto eram efetuadas consultas no sistema de RI. Foi gerada uma tabela (uma porção da mesma encontra-se na Tabela 1) com as medidas de precisão e de abrangência para cada combinação de pesos. Da análise dos dados gerados por este processo tivemos as seguintes pistas quanto aos pesos dos tipos de relações:

- Os pesos das relações USE e UF devem ser suficientemente altos de modo que possibilitem a utilização dos termos relacionados como se fossem os termos originais da relação. O peso destas relações deve ser 1 ou qualquer valor próximo de 1. As análises apresentadas por Robin & Ramalho em [11] demonstram

Tabela 2: Resultados obtidos com diferentes combinações do valor λ

λ	Termos	Abrangência	Precisão	Medida-F
0.2000	52	0.5934	0.3871	0.4685
0.2500	50	0.5934	0.3871	0.4685
0.3750	15	0.5714	0.3939	0.4664
0.5000	04	0.5659	0.3946	0.4650
1.0000	01	0.3516	0.3184	0.3342

que relação de sinonímia sempre melhora a resposta dos sistemas de RI, reforçando a nossa definição de pesos altos para estas relações.

- A relação NT é muito importante na expansão da consulta, pois a utilização de um peso alto para esta relação melhora a expansão da consulta. Contudo, em nosso método não é aconselhável dar o valor máximo (1) ao peso de NT, uma vez que a combinação com os pesos de outros tipos de relações pode acarretar uma explosão na quantidade de termos que serão analisados.

- A relação BT não deve ter um peso muito alto. Valor maior para as relações BT demonstra um pequeno aumento na abrangência mas, ao mesmo tempo, uma diminuição significativa da precisão. Este comportamento pode ser observado nos testes 378, 384 e 390 da Tabela 1.

- A relação RT demonstra uma relação semântica diferente da equivalência e hierarquia, que deveria indicar termos importantes para a RI. Contudo, nossos testes sugerem que valores altos para o peso das relações RT diminuem a taxa de precisão. Os testes 142 ao 150 na Tabela 1 demonstram que enquanto o valor para RT aumenta a taxa de precisão é reduzida. Portanto, isto nos leva a conclusão que o peso para esta relação deve ser mantido baixo.

Também, foi feita uma tentativa de estimar um melhor valor de λ. Neste processo foi modificado seu valor para uma consulta e verificado sua precisão, abrangência e Medida-F[5]. Os resultados são apresentados na Tabela 2.

Observa-se na Tabela 2 que um limiar mais baixo aumenta a quantidade de documentos e de documentos relevantes encontrados. Isto ocorre, pois a consulta expandida gerada contém mais termos quando λ tiver um valor baixo. Entretanto a melhor combinação, para esta consulta, entre o número de termos e a quantidade e qualidade dos documentos retornados pode ser verificada no valor do limiar λ equivalente a 0.5, isto é, entre os valores definidos para os pesos das relações BT (0.3) e NT (0.6). Observa-se que na Tabela 2 quando o valor de λ equivale a 1.0 é correspondente a somente utilizar os termos originais da consulta, ou seus termos sinônimos.

A decisão dos valores para os parâmetros da heurística de expansão de consultas deve ser bastante estudada. Os resultados são bastante distintos para as diferentes combinações de pesos. Definir os melhores parâmetros possíveis, se existirem, não fará parte do escopo desta dissertação devido a grande complexidade requerida para isto. Acredita-se que para a realização deste trabalho futuro deve ser utilizada técnicas estatísticas ou redes neuronais.

Foram executados 13 testes utilizando a ferramenta ASPSeek[6] para busca em documentos no mesmo corpus onde LTOCSS foi construído.

O processo de teste realizado foi semi-automático e ocorre da seguinte maneira:

- Primeiro passo: escolha do tópico a ser procurado. Tentamos atuar neste processo da forma mais natural possível, desta forma não tivemos a preocupação de fazer da consulta uma tarefa fácil para o sistema de RI. Basicamente o processo é: (1) queremos uma informação; (2) formulamos a consulta.

- Fizemos uso de um corpus de artigos de jornal[7] onde, através de exaustivas buscas, pode ser encontrado um conjunto de aproximadamente 100% de toda a informação relevante para cada assunto pesquisado. A geração dos conjuntos de documentos relevantes foi realizado por uma bolsista de iniciação científica do projeto Sema. A decisão do assunto a ser marcado no corpus é tomado ao ser confirmada a existência dos termos de uma consulta para este assunto. Nos documentos relevantes[8] encontrados é feita uma marca no nome do arquivo que os contém. No final deste processo podemos encontrar facilmente os arquivos relevantes, pois estes têm um padrão diferente do nome de arquivo.

[5]Do inglês *F-Measure*. Média ponderada da precisão e abrangência.

[6]ASPSeek é uma ferramenta de busca desenvolvida por Swsoft (*http://www.sw-soft.com/*) e licenciada sob os termos da GNU GPL (*http://www.gnu.org/copyleft/gpl.html*). Mais informações sobre a ferramenta podem ser encontradas em *http://www.aspseek.org/*.

[7]Corpus da "Folha de São Paulo" que foi utilizado para gerar o tesauro LTOCSS, gentilmente cedido pelo NILC.

[8]Utilizamos relevância binária, isto é, um documento é relevante ou não.

Consulta		Precisão	Abrangência	Medida-F
Original	Média	0,4499	0,2389	0,3121
	Desvio Padrão	0,3405	0,2508	0,2462
Expandida	Média	0,3778	0,5010	0,4307
	Desvio Padrão	0,2382	0,1728	0,1650

Tabela 3: Resultados parciais

- É formulada a consulta através do QET, que produzirá a consulta expandida.
- Esta consulta expandida é então utilizada em uma ferramenta de RI.

Como todos os documentos relevantes são conhecidos é possível gerar de forma automática, através de pequenos programas, diferentes tipos de estatísticas.

A maioria dos processos "semi-automáticos" é feita manualmente ou utilizando *scripts* e programas fora dos sistemas de RI. Como trabalho futuro deverá ser construída uma ferramenta de consulta junto ao QET, facilitando, assim, a geração de estatísticas, e aumentando a usabilidade da ferramenta.

Os testes realizados demonstraram que, em média, a expansão acarretou em uma degradação na taxa de precisão, mas por sua vez, também acarretou, uma significativa melhora na taxa de abrangência. Observa-se na Tabela 3, que o sistema de expansão de consulta obteve uma melhora na abrangência de 109,71%, ao mesmo tempo que teve degradada sua precisão em 16,02%. Estas medidas representaram um ganho de 38% na medida-F para a consulta expandida em relação a medida-F da consulta original. Veja os resultados obtidos para todas consultas nas suas formas originais e expandidas no Anexo 1.

Observamos na Tabela 3 com as médias das medidas obtidas, que o método aparenta melhorar, de uma forma geral, a RI em corpus estático. Esta conclusão é embasada pela melhora da medida-F, utilizada como um medida comum para avaliar o resultado da RI.

Durante o processo de testes foi possível verificar que todos os tesauros utilizados contribuiram para a geração do resultados finais, uma vez que os termos adicionados a consulta, no processo de expansão, provinham de diferentes tesauros. Contudo, quantificar a importância de cada um destes tesauros na expansão de consultas é uma tarefa ainda a ser desenvolvida.

6 Conclusão

Neste trabalho descrevemos o desenvolvimento de uma estrutura de tesauros para a utilização em sistemas de RI. A estrutura criada demonstrou sua utilidade ou representar diversos tesauros diferentes e possibilitando a utilização conjunta dos mesmos. Para a utilização da estrutura criou-se uma técnica e uma ferramenta de expansão de consulta que possibilitou quantificar a importância de cada relação no processo de RI.

Acreditamos que, no estágio atual, que ao efetuarmos mais testes poderemos afirmar que nossa técnica de expansão de consulta obtem bons resultados. Ao efetuarmos os testes semi-automaticamente o processo de avaliação é razoavelmente rápido para uma única consulta, mas descobrir o conjunto de documentos relevantes leva bastante tempo por ser um processo com grande complexidade.

Como trabalho futuro pretendemos intensificar a análise dos tesauros obtidos, de modo a avaliar a qualidade das relações presentes nos tesauros automáticos como o LTOCSS. Utilizando este tesauro na expansão de consulta e avaliando os resultados obtidos será possível mensurar a qualidade das relações dos termos associados.

Pretende-se, no desenvolvimento da estrutura, distanciar um pouco das relações semânticas definidas na ISO 2788 e testar diferentes relações semânticas, como, por exemplo, as relações definidas na Qualia de Pustejovsky em [10].

A ferramenta QET deve, como já descrito anteriormente, ser acrescida da capacidade de efetuar consulta em bases de dados. Esta característica deve facilitar a geração de dados estatísticos provenientes dos resultados das consultas.

Agradecimentos

O projeto SEMA é patrocinado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). O mestrando Luiz Augusto Sangoi Pizzato é bolsista da Dell Computer Corporation. Gostaríamos de agradecer à Subsecretaria de Biblioteca do Senado Federal, ao departamento técnico do Sistema Integrado de Bibliotecas da USP (SIBi/USP), ao departamento técnico da Biblioteca Central Ir. José Otão da PUCRS e a Caroline Gasperin por fornecer seus tesauros que foram muito importantes para o desenvolvimento deste

trabalho. Um agradecimento especial ao NILC (Núcleo Interinstitucional de Lingüistica Computacional) por ter cedido o corpus utilizado no estudo.

Referências

[1] BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. *Modern Information Retrieval*. ACM Press New York, Addison Wesley, 1999.

[2] CARPINETO, Claudio; et al.. An Information-Theoretic Approach to Automatic Query Expansion. In: *ACM Transactions on Information Systems*. v.19. n.1. 2001. p.1-27.

[3] GASPERIN, Caroline Varaschin. *Extração automática de relações semânticas a partir de relações sintáticas.* Dissertação de Mestrado. Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul. 2001.

[4] GREFENSTETTE, Gregory. *Explorations in automatic thesaurus discovery*. EUA: Kluwer Academic Publishers, 1994. 305 p..

[5] IMAI, H.; COLLIER, N. H.; TSUJII, J.. A Combined Query Expansion Approach for Information Retrieval. In: ASAI, K.; MIYANO, S.; TAKAGI, T.. *Genome Informatics.* Universal Academic Press Inc. 1999, p. 292-293.

[6] International Organization for Standardization. *ISO 2788: Guidelines for the establishment and development of monolingual thesauri*. 2nd ed. Geneva: ISO, 1986.

[7] JING, Yufeng; CROFT, W. Bruce. An Association Thesaurus for Information Retrieval. In: *Intelligent Multimedia Information Retrieval Systems and Management*. RIAO '94. New York, NY, Out. 1994. p. 146-160.

[8] KIMOTO, Haruo; IWADERA, Toshiaki. Construction of a Dynamic Thesaurus and its Use for Associated Information Retrieval. In: VIDICK, Jean-Luc. *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*. SIGIR'90. Brussels, Belgium. Set. 1990. p. 227-240.

[9] MANDALA, Rila; TOKUNAGA, Takenobu; TANAKA, Hozumi. Query expansion using heterogeneous thesauri. *Information Processing and Management.* v.36, n.3. 2000. p. 361-378.

[10] PUSTEJOVSKY, James. *The generative lexicon*. Cambridge: MIT, 1995.

[11] ROBIN, J.; RAMALHO, F. S. *Empirically evaluating WordNet-based query expansion in a web search engine setting*. In: Proceedings of IR'2001, Set. 19-21, Oulu, Finland, 2001.

[12] STRZALKOWSKI, T.; et al.. *Natural Language Information Retrieval: TREC-8 Report*. In: The Eighth Text REtrieval Conference (TREC-8). Gaithersburg, Md. Nov. 1999. p. 275-285.

Anexo 1

As tabelas seguintes demonstram as consultas realizadas na forma expandida e original, com o resultado de suas precisão, abrangências e medidas-F.

Nro	Consulta	Abrangência	Precisão	Medida-F
1	Viagem de Avião	0.1795	0.7000	0.2857
2	Acidente de Automóvel	0.1702	0.5333	0.2581
3	Comércio por Telefone	0.0000	0.0000	0.0000
4	Aposentadoria	0.3939	0.8667	0.5417
5	Animal Doméstico	0.0000	0.0000	0.0000
6	Aluguel de Imóvel	0.3182	0.9333	0.4746
7	Jogo de Futebol	0.4432	0.4588	0.4509
8	Música Brasileira	0.1014	0.5833	0.1728
9	Uso de Computador	0.4043	0.4872	0.4419
10	Doença Grave	0.0112	0.1111	0.0204
11	Frutas Tropicais	0.0000	0.0000	0.0000
12	Viagem Internacional	0.1887	0.6250	0.2899
13	Aumento de Salário	0.8261	0.6333	0.7170

Nro	Consulta Expandida	Abrangência	Precisão	Medida-F
1	VIAGEM DE AVIAO ou EXPEDICAO ou TURISMO ou VIAGEM AO REDOR DO MUNDO ou AEROPLANO ou AVIAO A ENERGIA SOLAR	0.2308	0.0928	0.1324
2	ACIDENTE DE AUTOMOVEIS ou DESASTRE ou ACIDENTE AERONAUTICO ou ACIDENTE DE TRANSITO ou ACIDENTE DO TRABALHO ou ACIDENTE MARITIMO ou ACIDENTE PESSOAL	0.8085	0.4935	0.6129
3	COMERCIO POR TELEFONE ou POLITICA COMERCIAL ou CIRCULACAO DE MERCADORIAS ou ECONOMIA INTERNACIONAL ou COMERCIO INTERNO ou COMERCIO ATACADISTA ou COMERCIO MARITIMO ou APARELHO TELEFONICO	0.3333	0.0106	0.0205
4	APOSENTADORIA ou APOSENTADORIA POR INVALIDEZ ou APOSENTADORIA POR DOENCA ou SEGURO-INVALIDEZ ou APOSENTADORIA POR TEMPO DE SERVICO ou APOSENTADORIA POR VELHICE ou APOSENTADORIA COMPULSORIA ou APOSENTADORIA POR IDADE ou SEGURO-VELHICE ou APOSENTADORIA VOLUNTARIA ou APOSENTADORIA ESPONTANEA ou APOSENTADORIA FACULTATIVA	0.3939	0.8125	0.5306
5	ANIMAL DOMESTICO ou CAPRINO ou COELHO ou EQUINO ou GADO ou OVINO ou SUINO	0.5085	0.2913	0.3704
6	ALUGUEL DE IMOVEL ou LOCACAO ou ALUGUER ou IMOVEL COMERCIAL ou IMOVEL RESIDENCIAL ou IMOVEL RURAL ou IMOVEL URBANO ou IMOVEL (DIREITO CIVIL) ou PROPRIEDADE IMOBILIARIA	0.4091	0.7500	0.5294
7	JOGO DE FUTEBOL ou CONTRATO DE JOGO E APOSTA ou JOGO (DIREITO CIVIL) ou LOTERIA ou LOTERIA ESPORTIVA ou LOTERIA FEDERAL ou LOTO ou FUTEBOL DE CAMPO ou FUTEBOL DE AREIA ou FUTEBOL DE ASFALTO ou FUTVOLEI	0.4432	0.4194	0.4309
8	MUSICA BRASILEIRA ou EVENTO MUSICAL ou FORMA MUSICAL ou HISTORIA DA MUSICA ou MEIO DE EXPRESSAO MUSICAL ou MUSICA TRADICIONAL ou MUSICOS ou TEORIA MUSICAL	0.4928	0.3579	0.4146
9	USO DE COMPUTADOR ou COMPUTADOR ELETRONICO ou COMPUTADOR ANALOGICO ou COMPUTADOR DE GRANDE PORTE ou COMPUTADOR DE QUINTA GERACAO ou COMPUTADOR DIGITAL ou COMPUTADOR GRAFICO ou MICROCOMPUTADOR ou MINICOMPUTADOR ou SUPERCOMPUTADOR ou UNIDADE CENTRAL DE PROCESSAMENTO	0.4468	0.2442	0.3158
10	DOENCA GRAVE ou ENFERMIDADE ou MOLESTIA ou CANCER OCUPACIONAL	0.5393	0.5783	0.5581
11	FRUTAS TROPICAIS ou ABACATE ou FRUTA-DE-CONDE ou GOIABA ou GRAVIOLA ou JABUTICABA ou JACA ou JAMBO ou JENIPAPO ou MAMAO ou MANGA ou MANGOSTAO ou ABACAXI ou ANANAS ou MARACUJA ou NESPERA ou PITANGA ou TAMARA ou TAMARINDO ou UMBU ou ACEROLA ou BANANICULTURA ou CAJA ou CAJU ou CAQUI ou CARAMBOLA ou CUPUACU	0.7222	0.3421	0.4643
12	VIAGEM INTERNACIONAL ou EXPEDICAO ou TURISMO ou VIAGEM AO REDOR DO MUNDO	0.6226	0.2705	0.3771
13	AUMENTO DE SALARIO ou ADICIONAIS ou SALARIO EM UTILIDADES ou SALARIO MINIMO	0.8696	0.3636	0.5128