

# **Recuperação de Informação e Expansão Automática de Consulta com Thesaurus: uma avaliação**

**Marco Gonzalez**

PUCRS - Faculdade de Informática  
Av.Ipiranga, 6681 – Prédio 16, térreo  
90619-900 Porto Alegre, Brazil  
**gonzalez@inf.pucrs.br**

**Vera Lúcia Strube de Lima**

PUCRS - Faculdade de Informática  
Av.Ipiranga, 6681 – Prédio 16, térreo  
90619-900 Porto Alegre, Brazil  
**vera@inf.pucrs.br**

## **ABSTRACT**

This paper presents a first evaluation of obtained results on automatic query expansion in information retrieval (IR). A thesaurus with semantic structuring and generative operations is used for deriving the lexical set that is related to each query term for automatic expansion. The selection of new terms of the expanded query and the calculation of its weights depend on the intersection of the derived lexical sets and on the depth level for descriptors search with respect to each considered terms. In this evaluation the following issues were considered: (i) a dissertations summaries collection about information systems; (ii) index terms in canonical form; and (iii) a group of test queries where the terms are key-words associated (by a traditional library system) with the dissertations. The results indicate that the query expansion, with the used approach, brings benefits for IR.

**Keywords:** Information Retrieval, Natural Language Processing, Artificial Intelligence, Thesaurus, query expansion, Data Mining

## **RESUMO**

Este artigo apresenta uma primeira avaliação dos resultados obtidos com a expansão automática de consulta em recuperação de informação (RI). Um thesaurus, com estruturação semântica e operações gerativas, é utilizado para gerar o campo lexical de cada termo da consulta e obter a expansão automaticamente. A seleção dos novos termos e o cálculo de seus pesos, na consulta expandida, depende da sobreposição dos campos lexicais e do nível de profundidade que se avança na busca de descritores dos termos considerados. Foram utilizados nesta avaliação: (i) uma coleção de documentos constituída por resumos de dissertações da área de sistemas de informação; (ii) termos de índice em forma canônica; e (iii) um conjunto de consultas de teste oriundas de palavras-chave associadas (por um sistema tradicional de biblioteca) aos documentos. Os resultados indicam que a expansão de consulta, com a abordagem utilizada, traz benefícios para a RI.

**Palavras-chave:** Recuperação de Informação, Processamento de Linguagem Natural, Inteligência Artificial, Thesaurus, Expansão de Consulta, Data Mining

## 1 INTRODUÇÃO

A essência da recuperação de informação (RI) consiste na busca de documentos relevantes a uma dada consulta que expressa a necessidade de informação do usuário. Tais consultas, em alguns sistemas de RI, podem ser expandidas, agregando novos termos aos originalmente inseridos. O objetivo desse procedimento é obter uma classificação, para os documentos recuperados, mais apurada quanto à relevância. Muitas tentativas têm sido feitas nesse sentido [8][1][12][2], incluindo processos automáticos ou manuais, e utilizando técnicas estatísticas de co-ocorrência de termos ou aplicando processamento de linguagem natural. A seleção dos novos termos e o cálculo de seus pesos na consulta expandida ainda é uma discussão em aberto.

Por outro lado, tanto na formulação da consulta, quanto na indexação dos documentos em sistemas de RI, o problema do controle de vocabulário tem tido propostas de solução através do uso de thesauri [4][13][5]. Neste trabalho, para viabilizar a expansão automática de consulta, utilizamos um thesaurus (denominado T-Lex), que possui uma estruturação semântica para implementar relacionamentos lexicais, levando em conta conceitos de modelagem de software orientada a objetos (MSOO) [10] e fundamentos da Teoria do Léxico Gerativo (TLG) de Pustejovsky [7].

A TLG introduz um conjunto de recursos para análise semântica de expressões em linguagem natural, incluindo operações gerativas que possibilitam derivações composicionais das palavras dependentes de contexto. Um léxico semântico, de acordo com a TLG, é caracterizado como um sistema computacional onde a estrutura Qualia<sup>1</sup>, que no T-Lex recebe destaque, é um dos níveis de representação, com quatro campos (ou papéis) de descrição. Assim, na representação de um item lexical  $\alpha$ , o papel Formal distingue  $\alpha$  num amplo domínio, o papel Constitutivo descreve o que faz parte de  $\alpha$ , o papel Agentivo especifica como  $\alpha$  passou a existir e o papel Télico explica qual a função ou o propósito de  $\alpha$ . Para maiores detalhes, ver [7].

No T-Lex, as categorias gramaticais consideradas, substantivo comum concreto (SCC), substantivo abstrato (SAB), substantivo próprio (SPR), verbo (VRB) e adjetivo (ADJ), possuem estruturas Qualia específicas. Os SCCs são descritos como classes de objetos; os SABs como qualidades, atos, sensações e outras abstrações; os SPRs como instâncias de objetos; os VRBs como operações; e os ADJs como estados.

Como exemplo dessas diferenças de representação citamos o papel agentivo, que pode ter descrição dupla, num SCC, e única, num VRB, conforme pode ser visto na figura 1. Diferenças em maior ou menor grau que ocorrem com outros papéis e com outras categorias gramaticais são reconhecidas e tratadas no T-Lex.

<p>Papel agentivo para um <b>SCC</b>, duas informações:  <b>Criação</b>: informa a operação que outros objetos executam para a criação do objeto representado (Exemplo: <u>estruturar</u>, no caso do SCC <u>estrutura</u>), e/ou  <b>Inicialização</b>: informa a operação que o próprio objeto executa ao passar a existir (Exemplo: <u>usar</u>, no caso do SCC <u>usuário</u>).</p> <p>Papel agentivo para um <b>VRB</b>, informação única:  Indica o <b>responsável</b> pela execução da operação correspondente ao VRB (Exemplo: <u>processador</u>, no caso do VRB <u>processar</u>).</p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figura 1. Diferenças de representação do papel agentivo para SCCs e VRBs

Com o objetivo de testar os benefícios do thesaurus que propomos visando a RI, analisamos mecanismos de busca com e sem expansão automática de consulta, e apresentamos os resultados de uma avaliação comparativa dos mesmos. Os novos termos expandidos nas consultas são obtidos pela utilização de critérios semânticos para compor relacionamentos lexicais, conforme estabelecidos no T-Lex.

Na seção 2 é discutida a técnica adotada para expansão automática de consulta, incluindo a seleção e o cálculo dos pesos dos termos expandidos, e as configurações de busca utilizadas; na seção 3 é descrita a avaliação comparativa realizada e são apresentados resultados; na seção 4, são tecidas considerações finais sobre o presente trabalho.

<sup>1</sup> Detalhes sobre a estrutura Qualia e a TLG podem ser obtidos em [7].

## 2 EXPANSÃO AUTOMÁTICA DE CONSULTA

### 2.1 Seleção e Cálculo do Peso dos Termos Expandidos

Os termos obtidos na expansão de uma consulta são selecionados entre os descritores contidos no T-Lex, a partir de operações gerativas. Essas operações tem como objetivo compor um campo lexical<sup>2</sup> (ou campo semântico<sup>3</sup>) de um item lexical. As operações utilizadas são: especialização, co-herança, associação, equivalência, decomposição e agregação [3].

O peso de cada termo, na expansão de uma consulta C, é calculado através de uma matriz de T linhas, onde T é o número de termos originais de C, e T+E colunas, onde E é o número de termos expandidos, conforme pode ser visto na figura 2.

No exemplo da figura 2, a consulta original é constituída por 3 termos: T01, T02 e T03. A partir do termo T01 são expandidos T04 e T09; a partir de T02 são obtidos T05, T06 e T10; e a partir de T03 temos T05, T07, T08 e T11.

		T01	T02	T03	peso total	peso total normalizado
T termos originais	T01	$p_{1,1}=10$	$p_{1,2}=10$	$p_{1,3}=10$	$P_{1,C}=30$	$P_{1,C}=30/30$
	T02	$p_{2,1}=10$	$p_{2,2}=10$	$p_{2,3}=10$	$P_{2,C}=30$	$P_{2,C}=30/30$
	T03	$p_{3,1}=10$	$p_{3,2}=10$	$p_{3,3}=10$	$P_{3,C}=30$	$P_{3,C}=30/30$
E termos expandidos	T04	$p_{4,1}=10/1$	$p_{4,2}=0/1$	$p_{4,3}=0/1$	$P_{4,C}=10$	$P_{4,C}=10/30$
	T05	$p_{5,1}=0/1$	$p_{5,2}=10/1$	$p_{5,3}=10/1$	$P_{5,C}=20$	$P_{5,C}=20/30$
	T06	$p_{6,1}=0/1$	$p_{6,2}=10/1$	$p_{6,3}=0/1$	$P_{6,C}=10$	$P_{6,C}=10/30$
	T07	$p_{7,1}=0/1$	$p_{7,2}=0/1$	$p_{7,3}=10/1$	$P_{7,C}=10$	$P_{7,C}=10/30$
	T08	$p_{8,1}=0/1$	$p_{8,2}=0/1$	$p_{8,3}=10/1$	$P_{8,C}=10$	$P_{8,C}=10/30$
	T09	$p_{9,1}=10/2$	$p_{9,2}=0/2$	$p_{9,3}=0/2$	$P_{9,C}=5$	$P_{9,C}=5/30$
	T10	$p_{10,1}=0/2$	$p_{10,2}=10/2$	$p_{10,3}=0/2$	$P_{10,C}=5$	$P_{10,C}=5/30$
	T11	$p_{11,1}=0/2$	$p_{11,2}=0/2$	$p_{11,3}=10/2$	$P_{11,C}=5$	$P_{11,C}=5/30$

Figura 2. Esquema-exemplo de matriz para cálculo dos pesos dos termos da consulta

O peso total  $P_{j,C}$  de um termo j numa consulta C é calculado pela soma dos elementos de uma linha j da matriz, da seguinte forma:

$$P_{j,C} = p_{j,1} + p_{j,2} + \dots + p_{j,T}$$

onde  $p_{j,i}$  é um elemento da matriz correspondente ao peso parcial de um termo i (original ou expandido) devido à contribuição de um termo original j, na expansão.

Assim, se um item lexical for encontrado, por expansão, a partir de mais de um termo original, seu peso crescerá. O peso parcial de um termo j, obtido pela expansão de um termo original i, será

$$p_{j,i} = 10/n,$$

<sup>2</sup> Campo lexical é o conjunto dos lexemas que estão semanticamente relacionados, paradigmática ou sintagmaticamente, dentro de um sistema lingüístico; é um subconjunto do léxico de uma língua [6][9].

<sup>3</sup> Campo semântico é o conjunto de palavras constituído pela rede de relações tais como homonímia, sinonímia e polissemia [11].

sendo  $n$  o nível de expansão do termo  $j$ , num intervalo<sup>4</sup> de 1 a 10, fazendo com que  $P_{j,i}$  seja 1 (mínimo) quando  $n=10$ .

No exemplo da figura 2, os termos expandidos em nível 1 são T04 a T08 e os de nível 2 são T09 a T11. Termos de nível 1 são aqueles expandidos diretamente dos termos originais; termos de nível 2 são os expandidos a partir dos termos de nível 1; e assim, sucessivamente. São considerados de nível 0 os termos originais da consulta e, excepcionalmente, nestes casos,

$$P_{j,1} = P_{j,2} = \dots = P_{j,T} = 10.$$

O peso total de cada termo  $j$  (de todos os níveis), ao final, é normalizado, ou seja, é dividido pelo maior peso total de forma que seja, no máximo, 1. No caso de não se utilizar expansão, o peso total de cada termo de consulta é 1.

No exemplo da figura 2, observa-se que o termo T05 é o que possui peso maior entre os expandidos, em razão de ter sido obtido em nível 1 e a partir de 2 termos originais.

A seguir é apresentado, na figura 3, um exemplo de expansão para consulta.

**Consulta original:** programação de computadores  
**Consulta expandida:** programação computador dado desenvolver programa linguagem algoritmo codificar programador programar processador dispositivo entrada saída memória ...

Figura 3. Exemplo de expansão de consulta

O resultado da operação de busca é um *ranking* de documentos, classificados decrescentemente, de acordo com pesos calculados como segue:

$$R_{D,C} = f_{1,D} \times P_{1,C} + f_{2,D} \times P_{2,C} + \dots + f_{T,D} \times P_{T,C}$$

onde  $R_{D,C}$  é o peso de um documento  $D$  numa consulta  $C$

$f_{j,D}$  é a frequência do termo  $j$  no documento  $D$

$P_{j,C}$  é o peso do termo  $j$  numa consulta  $C$  com  $T$  termos

## 2.2 Configurações de Busca Utilizadas

Foram avaliadas duas versões (0.0 e 1.0) de um mecanismo de busca batizado informalmente como “Yahinho”. Ambas utilizam índice invertido com termos no formato canônico. Foram testadas quatro configurações diferentes, descritas a seguir. A versão 0.0 corresponde à configuração ZERO, sem expansão de consulta, e a versão 1.0 foi executada nas configurações CT1, CAT1 e CAT3, utilizando o T-Lex, conforme tabela 1.

configuração	Papéis			nível
	constitutivo	agentivo	télico	
CT1	sim	não	sim	1
CAT1	sim	sim	sim	1
CAT3	sim	sim	sim	3

Tabela 1. Configurações da versão 1.0

<sup>4</sup> Níveis maiores de expansão causam um distanciamento semântico muito grande entre o termo original e o expandido. Acredita-se que o valor 10 (do intervalo 1-10) já esteja muito além da necessidade de expansão.

A coluna “nível” da tabela 1 identifica o nível de expansão utilizado. O número de termos expandidos naturalmente cresce, ao ser adotado um nível de expansão maior. Utilizando-se o nível 1, somente são acrescentados à consulta os descritores contidos nas descrições dos termos originais da mesma; utilizando-se o nível 2, são acrescentados também os descritores dos descritores, e assim por diante.

Nas colunas “papéis” da tabela 1, observa-se o uso ou não de um determinado papel para a expansão da consulta. Não utilizar um papel P significa que o conteúdo de P na descrição de um termo  $\alpha$  não é incluído na expansão devida a  $\alpha$ . Por outro lado, ao executar uma operação gerativa  $g(\alpha)$ , se for necessário, os descritores contidos em P nos demais itens lexicais do thesaurus (com exceção de  $\alpha$ ) serão analisados. O conteúdo do papel formal de  $\alpha$  nunca é incluído diretamente como termo expandido a partir de  $\alpha$ , pois há prejuízo no resultado da busca, por torná-la, obviamente, mais genérica.

### 3 AVALIAÇÃO

O objetivo central desta avaliação é verificar o ganho obtido pela expansão automática de consulta, em termos de precisão e resposta, num sistema de RI, utilizando-se um thesaurus com estrutura semântica fundamentada na TLG e em conceitos da MSOO.

#### 3.1 Corpus e Indexação

Foi utilizado um corpus de teste com 7095 palavras, constituído por 34 resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. Em média, os documentos possuem 208 palavras cada um.

Para estes documentos, foi gerado um arquivo invertido de índices com termos em formato canônico. Cada termo de índice possui uma lista com a identificação do documento D, onde o termo ocorre, e a sua frequência  $f_{j,D}$  neste documento, sendo:

$$f_{j,D} = L_{j,D} / N_D$$

onde  $L_{j,D}$  o número de ocorrências do termo j no documento D; e

$N_D$  o número total de palavras (excluídas as *stopwords*<sup>5</sup>) em D.

#### 3.2 Metodologia de Avaliação

Os documentos do corpus utilizado nesta avaliação foram indexados, conforme critérios bibliográficos tradicionais, pela Biblioteca Central da PUCRS, através de expressões-chave. Estes índices (da Biblioteca Central) não foram utilizados mas, do conjunto total destas expressões, foram extraídas algumas para constituir as consultas apresentadas na figura 4. Para a obtenção do gráfico precisão/resposta, foram realizadas as 24 consultas da figura 4, com cada uma das quatro configurações dos mecanismos de busca avaliados.

---

<sup>5</sup> *Stopwords* são palavras como preposições, artigos e conjunções, cujo conteúdo semântico é limitado.

1) hipermídia e multimídia	9) educação à distância	18) sistemas multiagentes
2) informática na educação	10) biblioteca digital	19) análise léxica
3) processamento de linguagem natural	11) java	20) sistemas de informação
4) tradução automática	12) internet	21) processamento paralelo
5) sistema de apoio à decisão	13) contabilidade	22) computação científica
6) interface com o usuário	14) banco de dados	23) arquitetura de computadores
7) ensino colaborativo	15) inteligência artificial	24) programação de computadores
8) orientação a objetos	16) bases de conhecimento	
	17) engenharia de software	

Figura 4. Consultas realizadas

Cada curva de precisão/resposta foi calculada tomando-se os valores médios de precisão ( $n^\circ$  de documentos relevantes recuperados /  $n^\circ$  de documentos recuperados) e de resposta ( $n^\circ$  de documentos relevantes recuperados /  $n^\circ$  de documentos relevantes) para as 24 consultas, de acordo com os procedimentos adotados pela comunidade internacional de RI nas “Text Retrieval Conferences - TREC’s” [14].

Foi, também, utilizado o método *pooling* [14]: um documento D é considerado relevante se estiver associado (pela Biblioteca Central da PUCRS) à expressão-chave utilizada na consulta realizada, e contido no *pool* de documentos recuperados por todas as configurações de busca.

Deve ser observado, ainda, que foram analisadas somente as dez primeiras respostas, em cada teste, já que, com elas, recuperam-se praticamente 1/3 dos 34 documentos do corpus, e este número é mais que o dobro do total de documentos relevantes por consulta que é, em média, quatro. Evitou-se ampliar a quantidade de respostas analisadas já que, quanto maior este número, mais vantagens teriam as configurações que usaram thesaurus, pois, com mais termos devido à expansão de consulta, elas apresentam um conjunto de documentos recuperados maior.

### 3.3 Resultados

O gráfico da figura 5 apresenta as curvas de precisão/resposta médias para as configurações de busca ZERO, CT1, CAT1 e CAT3.

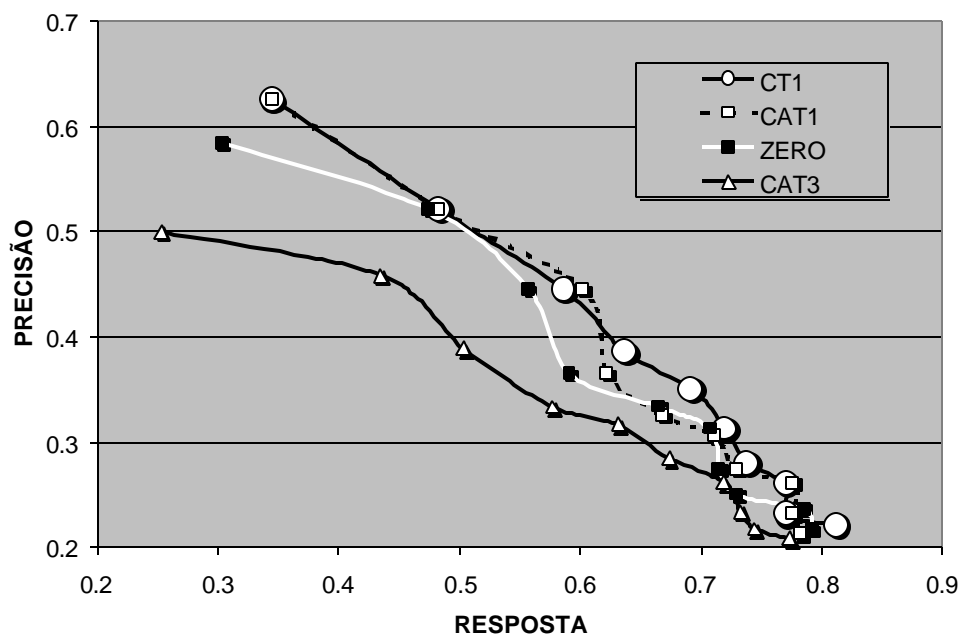


Figura 5. Gráfico Precisão/Resposta

A tabela 2 apresenta os valores máximos e mínimos (para a precisão média e para a resposta média) obtidos com as 24 consultas nas quatro configurações dos mecanismos de busca utilizados.

	Precisão média		Resposta média	
	máxima	mínima	máxima	mínima
<b>valor</b>	63%	21%	81%	25%
<b>Configurações</b>	CT1 e CAT1	CAT3	CT1	CAT3

Tabela 2. Valores máximos e mínimos para precisão e resposta

Conforme o gráfico da figura 5, as configurações CT1 e CAT1 apresentaram alguma vantagem sobre a configuração ZERO.

Pela tabela 2, verifica-se que a configuração CT1 foi a única que contribuiu para ambos os valores máximos de precisão e de resposta. Observa-se facilmente que a configuração CAT3 revelou-se a de pior performance.

Na tabela 3 podem ser observados os valores de precisão e de resposta, obtidos com as configurações CT1 e CAN, nas seguintes condições: (i) o valor de precisão considera o documento do topo da classificação em cada caso; e (ii) o valor de resposta considera os quatro documentos recuperados como mais relevantes em cada configuração. Nessas condições, nota-se que a precisão da configuração CT1 foi 8,6% superior, e a resposta 8,5% superior.

Configuração	precisão	resposta
CT1	63%	64%
CAN	58%	59%

Tabela 3. Valores obtidos na avaliação comparativa entre CT1 e CAN

#### 4 CONSIDERAÇÕES FINAIS

Os resultados obtidos na avaliação indicam que a expansão de consulta pode trazer benefícios à RI. Entretanto, esta expansão não pode ser feita indiscriminadamente, sob um enfoque quantitativo, sob pena de prejudicar os resultados do mecanismo de busca. É o que foi constatado com a configuração CAT3, que expande a consulta até o nível 3. Ela apresenta resultados piores que a configuração ZERO, sem expansão. Também fica abaixo das demais que consideram apenas o nível 1 de expansão.

Como as configurações CAT1 e CT1 tiveram resultados muito próximos, é ainda difícil julgar o efeito do uso do papel agentivo. Por outro lado, foi evitada a inclusão, nas expansões, dos itens lexicais contidos no papel formal, para não generalizar a busca com aumento da resposta e diminuição da precisão.

Todas as configurações que utilizaram expansão de consulta foram testadas com e sem inclusão de verbos nos termos expandidos. Tanto num caso como no outro, entretanto, os verbos foram usados para gerar novos termos na expansão. A inclusão ou não de verbos como termos expandidos não chegou a afetar significativamente a performance das configurações.

Entre os termos expandidos, independentemente da categoria gramatical, alguns podem ser considerados composicionais e outros periféricos. Os composicionais são aqueles que são gerados de mais de um termo original e auxiliam a captura do significado composicional da consulta. Por isso, recebem, como o termo T05 na figura 1, um peso maior. Os periféricos, ao contrário, originam-se de um único termo. Entretanto, ainda que com pequeno peso, os periféricos podem contribuir para a recuperação de documentos que, sem conter em seus textos os itens lexicais originais da consulta, usam com frequência, como é de se esperar, termos comuns naquele contexto. Os resultados desta avaliação

devem-se mais aos termos periféricos que aos composicionais pelas seguintes razões: (i) o T-Lex, na condição de protótipo, possui uma base lexical ainda pequena, sem a descrição completa e necessária em alguns casos; e (ii) as consultas utilizadas, em geral, possuem termos mais ou menos genéricos e com relativamente pequeno grau de ambigüidade.

Esta avaliação, portanto, é considerada preliminar e nossos esforços serão dirigidos, em trabalhos futuros e com o crescimento do thesaurus, (i) para a análise da influência de cada papel da Qualia na expansão; (ii) para o estudo do comportamento dos verbos como termos que viabilizam vínculos entre termos expandidos; e (iii) para examinar a performance de termos composicionais e periféricos na expansão.

## REFERENCIAS BIBLIOGRÁFICAS

1. ARAMPATZIS, A.; TSORIS, T.; KOSTER, C.H.A. Irena: Information Retrieval Engine based on Natural Language Analysis. In: Proceedings of RIAO'97 – Computer-Assisted Information Searching on Internet. Montreal: McGill University, 1997. p.159-175.
2. GAUCH, S.; WANG, J.; RACHAKONDA, S. M. A Corpus Analysis Approach for Automatic Query Expansion and Its Extension to Multiple Databases. ACM Transactions on Information Systems, 1999, v.17, n.3, p.250-269.
3. GONZALEZ, M. O Léxico Gerativo de Pustejovsky sob o Enfoque da Recuperação de Informações. Trabalho Individual I, PPGCC, Faculdade de Informática, PUCRS, maio 2000. 52 p.
4. JOYCE, T.; NEEDHAM, R.M. The Thesaurus Approach to Information Retrieval. American Documentation, 1958, V.9, p.192-197. In: SPARCK JONES, K; WILLET, P. (editores). **Readings in Information Retrieval**. California: Morgan Kaufmann Publishers, Inc., 1997. p. 15-20.
5. LOUKACHEVITCH, N. V.; SALLI, A. D.; DOBROV, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. NLDB'99 – 4<sup>th</sup> Int. Conf. on Applications of Natural Language to Information Systems, 1999. OCG Schriftenreihe, Lecture Notes, v.129, p.203-208.
6. LYONS, J. **Semantics**. Cambridge: Cambridge University Press, 1977. V. I e II.
7. PUSTEJOVSKY, J. **The Generative Lexicon**. Cambridge: The MIT Press, 1995. 298 p.
8. QIU, Y.; FREI, H.P. Concept Based Query Expansion. In: Proceedings of the 16<sup>th</sup> International ACM SIGIR Conference. ACM Press, 1993. p.160-169.
9. REHFELDT, Gládis Knak. **Polissemia e Campo Semântico (estudo aplicativo aos verbos de movimento)**. Porto Alegre: EDURGS / FAPA / FAPCCA, 1980. 172 p.
10. RUMBAUGH, J.; BLAHA, M.; PREMERLANI, W.; EDDY, F.; LORENSEN, W. **Object-Oriented Modeling and Design**. Englewood Cliffs: Prentice-Hall, 1991. 500 p.
11. SCAPINI, Isabel K. Associações Interlexicais: Contribuição para um Dicionário Remissivo. Dissertação de Mestrado. Porto Alegre: Instituto de Letras e Artes, PUCRS, 1997. 169 p.
12. STRZALKOWSKI, T.; WANG, J.; WISE, B. Summarization-based Query Expansion in Information Retrieval. In: Proceedings of 36<sup>th</sup> Annual Meeting of the ACL. Montreal, 1998. V. II, p.1258-1264.
13. SPARCK-JONES, K. **Synonymy and Semantic Classification**. Edinburgh: Edinburgh University Press, 1986. 285 p.
14. VOORHEES, E.M.; HARMAN, D. Overview of the Eighth Text Retrieval Conference (TREC-8). Capturado em dezembro de 2000. Disponível na Internet em <http://trec.nist.gov/pubs.html>.