UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

NÚCLEO DE COMPUTAÇÃO ELETRÔNICA

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Expansão semântica de consultas baseada em esquemas terminológicos: uma experimentação no domínio biomédico

Dissertação de Mestrado

André Bechara Elias

Orientador(es): Maria Luiza Machado Campos

Vanessa Braganholo Murta

Rio de Janeiro

2010

André Bechara Elias

EXPANSÃO SEMÂNTICA DE CONSULTAS BASEADA EM ESQUEMAS TERMINOLÓGICOS: UMA EXPERIMENTAÇÃO NO DOMÍNIO BIOMÉDICO

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática DCC / IM-NCE / UFRJ, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Ciência da Computação.

Orientador(es): Maria Luiza Machado Campos

Vanessa Braganholo Murta

Rio de Janeiro

2010

André Bechara Elias

EXPANSÃO SEMÂNTICA DE CONSULTAS BASEADA EM ESQUEMAS TERMINOLÓGICOS: UMA EXPERIMENTAÇÃO NO DOMÍNIO BIOMÉDICO

Rio de Janeiro, 31 de maio de 2010

Profª. _____ (Orientadora)

Maria Luiza Machado Campos, Ph.D.

Profª. _____ (Orientadora)

Vanessa Braganholo Murta, D.Sc.

Prof ª. _____

Maria Luiza de Almeida Campos, Ph.D.

Prof . _____

Pedro Manoel da Silveira, Ph.D.

Prof ª. _____

Maria Cláudia Reis Cavalcanti, D.Sc.

Aos meus familiares.

AGRADECIMENTOS

À Profa. Dra. Vanessa Braganholo Murta, orientadora desta dissertação, por todo empenho, sabedoria, compreensão e, acima de tudo, disciplina. Gostaria de ratificar a sua competência, participação com discussões, correções, revisões, sugestões que fizeram com que concluíssemos este trabalho.

À Profa. Dra. Maria Luiza Machado Campos, co-orientadora desta dissertação, por sua ajuda e interesse no tema, que com sua experiência elevou o nível desta dissertação, nível o qual sem sua ajuda não seria possível chegar.

Ao Prof. Dr. Pedro Manuel da Silveira, Profa. Dra. Maria Luiza Almeida Campos e Profa. Dra. Maria Cláudia Reis Cavalcanti por aceitarem participar da Banca de Defesa desta Tese, proporcionando discussões e sugestões que servirão para crescimento, aprendizado e incentivo à pesquisa.

À secretária Tia Deise Lobo Cavalcante por sua força e dedicação contagiantes na administração do Programa de Pós-Graduação de Informática, sendo uma profissional extremamente competente e dedicada.

Ao Grupo de Pesquisa em Recuperação de Informação e Educação da Universidade de Saúde e Ciência de Oregon pelo fornecimento dos dados necessários para esta pesquisa.

Ao CNPq pelo apoio e financiamento concedidos.

Aos meus familiares que sempre me deram amor e força, valorizando meus potenciais.

À minha esposa que por tantos finais de semana teve que suportar minha falta de atenção.

Ao Jacques Douglas Varaschim, Antônio Maia e Juarez Queiroz meus superiores na Globo.com, os quais me concederam a oportunidade de fazer o mestrado em tempo parcial mantendo o vínculo empregatício.

A todos os meus amigos e amigas que sempre estiveram presentes me aconselhando e incentivando com carinho e dedicação.

A todas as pessoas que, direta ou indiretamente, contribuíram para a execução dessa Tese de Mestrado.

.

RESUMO

ELIAS, André Bechara. **Expansão semântica de consultas baseada em esquemas terminológicos**: uma experimentação no domínio biomédico. 2010. 131 f. Tese (Mestrado em Informática) – PPGI, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, 2010.

Apesar de terem se tornado essenciais para as tarefas do dia á dia, os sistemas de recuperação de informação ainda não utilizam a semântica dos termos para atender as necessidades de informação dos usuários. Este trabalho apresenta um estudo sobre a utilização de ontologias de domínio e outras artefatos de controle terminológico para melhorar a eficiência na recuperação de informação. A proposta deste trabalho é usar o conhecimento de domínio presente nestes artefatos como fonte de termos relacionados para complementar a consulta inserida pelo usuário. O objetivo é desenvolver uma técnica de expansão de consultas que melhore tanto a precisão quanto a cobertura e também evite o problema do descasamento de palavras-chaves.

Para se entender como as diferentes relações léxico-semânticas das ontologias de domínio e outros esquemas terminológicos afetam o desempenho da técnica de expansão de consultas, realizamos dois experimentos usando a TREC Genômica como coleção de testes. No primeiro experimento foi usada apenas a Gene Ontology como base de conhecimento. Para resolver alguns problemas encontrados no primeiro experimento, realizamos um segundo experimento complementando-a com vocabulários e tesauros de ampla utilização na área biomédica.

Os resultados dos experimentos indicam que, para se obter uma melhor precisão, o mecanismo de expansão de consultas deveria examinar características da consulta inserida pelo usuário antes de aplicar a expansão de relações léxico-semântica diferente de sinônimos. Os experimentos mostraram também que os melhores resultados são atingidos quando os pesos dos termos inseridos pelo usuário são muito maiores (em torno de cinco vezes) que os termos sugeridos pelo sistema. Foi identificado também que entre diferentes ontologias e vocabulários existem maneiras distintas de se tratar

certas relações léxico-semânticas com um termo, como por exemplo, na generalização-especialização. Esse fato, por sua vez, dificulta a escolha do termo correto pelo mecanismo de expansão nestes tipos de relações.

ABSTRACT

ELIAS, André Bechara. **Expansão semântica de consultas baseada em esquemas terminológicos**: uma experimentação no domínio biomédico. 2010. 131 f. Tese (Mestrado em Informática) – PPGI, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, 2010.

Despite having become essential to the tasks of everyday life, information retrieval systems do not utilize the semantics of terms to meet the information needs of their users. This work presents a study on using domain ontologies and other terminological artifacts to improve efficiency on information retrieval. The proposal is to use domain knowledge as a source for related words to complement the original user query. The goal is to develop a query expansion technique that enhance both recall and precision and also avoid the word mismatch problem.

In order to understand how the different lexical-semantics relationships in the domain ontologies and vocabularies, affect the performance of the query expansion technique, we ran two experiments using the Genomics TREC as test collection. The first experiment used only Gene Ontology as a knowledge base. To solve some of the problems found in the first experiment, we ran a second experiment using a complementary set of biomedical thesaurus and vocabularies.

The results of the experiments indicate that, in order to achieve better precision, the query expansion mechanism should look for "expansion tips" on the user inputted query before applying a lexical-relation - different from synonyms- on the expansion. The result showed also that the best result occurs when the weight attributed to the user input terms is far bigger (at about five times) than the suggested terms. It was also identified that between different ontologies and vocabularies there are distinct ways of treating some lexical-relationships for a term (like is_a). This, in turn, makes it difficult for the expansion mechanism to choose the right term for this kind of expansion.

Lista de Figuras

Figura 1 - Triângulo Semiótico de Ogden e Richards (1923) 17

Figura 2- Arquitetura de um sistema de RI (BAEZA-YATES; RIBEIRO-NETO, 1999) ... 27

Figura 3- Conjuntos de cobertura e precisão 32

Figura 4 – Curva de Cobertura versus Precisão 33

Figura 5 – Documentos retornados para a consulta C do original (Baeza R., Ribeiro B., 1999) 34

Figura 6 – Curva de cobertura-precisão de 11 pontos do original (Baeza R., Ribeiro B., 1999) 36

Figura 7 – A ontologia de alto nível SOWA 39

Figura 8 - Expansão por sinônimos do original Santon e Lesk (1968) 44

Figura 9 - Expansão por termo genérico do original Salton e Lesk (1968) 45

Figura 10 - Resultado do experimento de Voorhees (1994) retirado do Original 48

Figura 11 - Etapas de um experimento de expansão de consultas 60

Figura 12 - Planejamento de um experimento de expansão de consultas 62

Figura 13 - Recursos necessários para experimentos de expansão de consultas 64

Figura 14 - O ambiente de avaliação 67

Figura 15 - Exemplo de um documento da TREC Genômica 71

Figura 16- Exemplo de uma consulta da base TREC Genômica 72

Figura 17 - Trecho do arquivo qrels TREC Genômica 2004 74

Figura 18 - Exemplo de uma consulta expandida do primeiro experimento 76

Figura 19 - Exemplo de medição feita pelo programa Trec_Eval 83

Figura 20 - O público e suas consultas (Spink 2001) 86

Figura 21 - Consulta acrescida de campo com palavras-chave 88

Figura 22 - Distribuição da quantidade de termos após criação do campo input 89
Figura 23 - Resultado da expansão de sinonímia em cada consulta 95
Figura 24 - Gráfico de cobertura x precisão da expansão por sinonímia 97
Figura 25- Resultado da expansão por termo genérico em cada consulta 100
Figura 26 - Curva de cobertura vs. precisão da expansão por termo genérico 101
Figura 27 - Resultado da expansão por termo específico em cada consulta 104
Figura 28 - Curva de cobertura vs. precisão da expansão por termo específico.....106
Figura 29 - Resultado da expansão por doença associada em cada consulta.....109
Figura 30 - Curva de cobertura vs. precisão da expansão por doença associada.....110
Figura 31 - Resultado da expansão por todo-parte em cada consulta 113
Figura 32 - Curva de cobertura vs. precisão da expansão por relação todo-parte.....114
Figura 33 - Resultado da expansão por relação é parte de na consulta 3 115
Figura 34 - Curva de cobertura vs. precisão da expansão por relação é parte de.....116

Lista de Tabelas

Tabela 1 - Coleções de Testes usadas em Salton e Lesk (1968) 43

Tabela 2 - Categorização das relações do MTM do original Fox (1980) 46

Tabela 3 - Resultados obtidos em Mandala, et. al (1999) 49

Tabela 4 - Resultados obtidos em Lu (2008) 51

Tabela 5 - Resultado do experimento de Stokes (2009) 52

Tabela 6- Resumo dos trabalhos relacionados 53

Tabela 7 - Principais bases de conhecimento da área biomédica 75

Tabela 8 - Estatísticas de expansão do primeiro experimento 77

Tabela 9 - Estatísticas de expansão de consultas do segundo experimento 78

Tabela 10- Quantidade de termos nas consultas 87

Tabela 11 - Resultados obtidos com a relação de sinonímia 94

Tabela 12 - Resultado em cada consulta nas relações de sinonímia 95

Tabela 13 - Resultados das expansões por termo genérico 99

Tabela 14 - Resultado da expansão de cada consulta por termo genérico 99

Tabela 15 - Resultados da expansão por termo específico 103

Tabela 16 - Resultado da expansão de cada consulta por termo específico 103

Tabela 17 - Resultados da expansão por doença associada 107

Tabela 18 - Resultado da expansão por doença associada 108

Tabela 19 - Resultados da expansão por relação todo-parte 111

Tabela 20 - Resultado da expansão por relação todo-parte 112

Tabela 21 - Resultado da expansão por relação é parte de 115

Tabela 22 - Resumo dos resultados de cada relação 117

Lista de Abreviaturas

ADI: American Documentation Institute

API: Application Programming Interface

#**DOCS**: Número de Documentos

CRAN-1: Projeto Cranfield I

MeSH: Medical Subject Headings

MPM: Mediana da Precisão Média

MTM: Meaning to Text Model

P@10: Precisão nos primeiros dez documentos.

SMART: System for the Mechanical Analysis and Retrieval of Text

TF: Term Frequency

IDF: Inverse Document Frequency

TREC: Text Retrieval Conference

RI: Recuperação de Informação

VS: Versus

XML: Extended Markup Language

SUMÁRIO

1. Introdução 17

2. Referencial Teórico 25

2.1 Expansão de Consultas 25

2.2 Avaliação de Desempenho em Recuperação de Informação 29

2.3 Ontologias e outros esquemas conceituais e terminológicos 36

3. Trabalhos em Expansões de Consultas 41

3.1 Experimento de Salton e Lesk de 1968 42

3.2 Experimento de Fox de 1980 45

3.3 Experimento de Voorhees de 1994 47

3.4 Experimento de Mandala et al. de 1999 49

3.5 O experimento de Navigli e Velardi de 2003 50

3.6 O experimento de Lu et. al. de 2009 50

3.7 O experimento de Stokes de 2009 51

3.8 Comparação dos Trabalhos Relacionados 52

4. Metodologia da Pesquisa 55

4.1 Caracterização da pesquisa 55

4.2 Definição do problema e hipótese 57

4.3 Etapas de um experimento de expansão de consultas 58

4.4 Planejamento de um experimento de expansão de consultas 60

5. Ambiente de avaliação experimental 67

5.1 Escolha do mecanismo de recuperação da informação 68

5.2 Escolha da coleção de testes 69

5.3 Escolha das bases de conhecimento 74

5.4 Expansão manual das consultas 75
5.5 Indexação dos documentos da coleção de testes 78
5.6 Aplicação de avaliação 79
5.7 Geração de medições 82
5.8 Geração de curva de cobertura e precisão 83
6. Realização do experimento 85
6.1 Tratamento do número de termos das consultas 85
6.2 Execuções Sucessivas 89
6.3 Classificação de Relevância 90
7. Análise dos resultados 92
7.1 Sinônimos 92
7.2 Expansão por termo genérico 97
7.3 Expansão por termo específico 101
7.4 Doença Associada 106
7.5 Contém 110
7.6 É parte de 114
7.7 Considerações Finais 117
8. Conclusão 118

1. Introdução

Os seres humanos utilizam a língua como uma sistemática para codificar e decodificar informações. Essa sistemática define símbolos que podem se manifestar na forma escrita ou falada. Uma vez compartilhada por um grupo social a língua se torna a base para a sua comunicação.

Dois componentes são essenciais para a formação de uma língua: a gramática que estabelece a forma como se estruturam (organizam) os símbolos; e o significado (semântica) que estabelece o vínculo entre o símbolo e o referente (Figura 1).

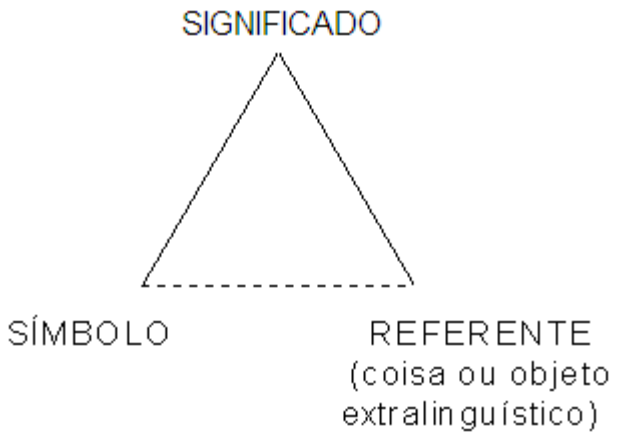

Figura 1 - Triângulo Semiótico de Ogden e Richards (1923)

Mesmo que definidas as regras de construção de sentenças em uma gramática, a interpretação de uma frase só é possível se existe alguma relação semântica entre os termos (símbolos) presentes na frase. Por exemplo, a frase: "Essa criança gosta de brincar de video-game", faz sentido, pois é possível estabelecer as ligações entre os conceitos de criança, gostar e video-game. Por outro lado, a frase "Essa árvore gosta de brincar de video-game", apesar de gramaticalmente semelhante à anterior, não parece fazer sentido, pois não é possível estabelecer relações semânticas entre os conceitos árvore, brincar e video-game.

Dessa forma, ao se comunicarem, os seres humanos não restringem o arranjo dos termos de uma sentença apenas pela gramática da língua, mas também pela relação entre os conceitos que estes termos representam (OGDEN; RICHARDS, 1989).

Para atender a uma necessidade de informação (uma pergunta), precisa-se interpretar esta pergunta através de uma série de relações léxico-semânticas que complementam e ajudam o processo cognitivo necessário para chegar-se à resposta. Antes de elaborar uma resposta para a pergunta "Quais são as profissões dos seus pais?" usam-se as relações existentes entre o conceito *pais* e o conceito *pai* (*pais* inclui *pai*) e também a relação entre o conceito *pais* e o conceito *mãe* (*pais* inclui *mãe*).

Os mecanismos de recuperação de informação disponíveis hoje tiram poucas vantagens da interpretação das relações semânticas existentes entre os termos de uma consulta (BRIN; PAGE, 1998). De maneira geral, os mecanismos são excessivamente dependentes do casamento (*matching*) entre os termos da consulta e os termos do documento.

Nos últimos anos, com a inclusão digital e a consolidação da Internet como um meio de comunicação de massa, a Web tem se tornado uma das principais fontes de busca de informação. Estudos mostraram que dois tipos de estratégias bem distintas são utilizadas nas buscas por informações (MARCHIONINI, GARI; BEN). A estratégia chamada de analítica se caracteriza pela escolha planejada de uma série de termos de consulta e a outra, chamada de estratégia de navegação, se caracteriza pela escolha momentânea de termos conforme estes são disponibilizados.

Um problema comum nesta era da informação é que os usuários geralmente são sobrecarregados com o excesso de informações. A seleção de informações irrelevantes por mecanismos de recuperação de informação é prejudicial tanto do ponto de vista da utilização dos recursos computacionais quanto do ponto de vista do desperdício de tempo e frustração dos usuários. Por outro lado, a falha na seleção de informações relevantes prejudica a eficácia em atender a uma necessidade de informação. Desta forma, um problema chave na recuperação de informações é o desenvolvimento de mecanismos de recuperação que forneçam o mínimo de informações irrelevantes (alta precisão) e ao mesmo tempo garantam que as informações relevantes sejam

recuperadas (alta cobertura)(BAEZA-YATES; RIBEIRO-NETO, 1999). Por este motivo, a eficiência dos mecanismos de recuperação de informação é avaliada segundo as suas medidas de precisão e cobertura.

Para atender as necessidades de informação dos usuários na Web, é comum o uso de máquinas de busca. Tais ferramentas se baseiam nas teorias da área de Recuperação de Informação e são capazes de realizar consultas por palavras-chave em um acervo de documentos. Os resultados são exibidos ao usuário por ordem de relevância, ou seja, os documentos que a máquina de busca julga serem mais importantes são colocados no topo da lista de resultados.

Este julgamento de relevância leva em conta fatores como o número de vezes que um termo aparece no acervo, entre outros, mas não leva em consideração a semântica dos termos presentes nos documentos. Em geral esse julgamento é uma variação do modelo vetorial clássico no qual os pesos dos termos são calculados na forma TF*IDF (*TF-Term Frequency, IDF-Inverse Document Frequency*), ou seja, a quantidade de vezes que um termo aparece em um documento multiplicado pela raridade deste termo na coleção de documentos (MANNING ET AL., 2008). Isso faz com que, em grande parte das vezes, resultados importantes não sejam retornados para o usuário, simplesmente porque aquele documento não contém o termo utilizado na busca, mas sim um sinônimo ou um termo mais genérico.

Esse problema é chamado pela comunidade de recuperação de informação, de problema do descasamento de palavra-chave (XU; CROFT, 1996a). Uma das formas de amenizar esse problema consiste em aplicar técnicas de expansão de consultas. No entanto, vários trabalhos na literatura mostram que a técnica de expansão de consultas pode ter efeito negativo sobre o sistema, diminuindo a precisão do sistema (a fração dos documentos recuperados que são relevantes). Vários parâmetros influenciam nos resultados da técnica: quais características considerar para a expansão das consultas; quantos termos devem ser utilizados na expansão; de onde retirar os termos que serão expandidos.

Na verdade, o uso das relações léxico-semânticas entre os termos de um domínio para apoiar a recuperação de informação é uma técnica bastante antiga (SALTON; LESK, 1968) (FOX, 1980). Essas relações têm sido usadas para a exploração de acervos específicos de domínio. Várias abordagens utilizam taxonomias para a indexação de acervos (CHAKRABARTI ET AL., 1998), de forma a facilitar a navegação pelo acervo. Exemplos podem ser encontrados em bases de dados de grandes empresas, de legislação, em sítios na Internet, entre outros. No entanto, estes usos têm se restringido à navegação. Isso significa que, para encontrar o que deseja, o usuário tem que navegar pelas categorias da taxonomia, utilizando os relacionamentos hierárquicos até encontrar os documentos que deseja. Dependendo da profundidade da hierarquia e do número de conceitos que ela disponibiliza, a tarefa de navegação pode se tornar cansativa e pouco eficiente.

Com a popularização das ontologias, tesauros e outros mecanismos de controle terminológico como um meio de representação de conhecimento de um determinado domínio, aumentaram as possibilidades de sua utilização para a construção de mecanismos de recuperação de informação que explorem esse conhecimento de domínio como meio de atingir uma maior eficiência (LEE ET AL., 2008); (FU ET AL., 2005). Esses mecanismos possibilitariam um direcionamento da recuperação da informação não mais para a co-ocorrência de termos na consulta e no documento, mas sim para um casamento entre uma necessidade de informação e o conteúdo do documento.

Nós acreditamos que as relações presentes em um tesauro ou em uma ontologia (a exemplo de subtipo_de, instância_de, sinônimo_de, supertipo_de) podem ajudar na expansão de consultas, aumentando a cobertura e a precisão do sistema. Neste trabalho, propomos explorar os vários tipos de relações existentes em uma ontologia de domínio ou em esquemas de controle terminológicos do tipo tesauros e estudar seus impactos na técnica de expansão de consulta, quando essas relações são utilizadas para expandir consultas de uma base específica de domínio. Por ontologia de domínio, entende-se um conhecimento explícito e compartilhado pela comunidade de alguma área de conhecimento (GRUBER, 1993), como por exemplo, a área biomédica.

O objetivo deste trabalho é explorar o potencial da aplicação das ontologias de domínio e outros artefatos terminológicos, como apoio à técnica de expansão automática de consultas. Nossa hipótese é que as relações entre os termos contidos

nas ontologias, construídas usando-se uma metodologia formal, permitem uma melhora tanto na precisão quanto na cobertura dos mecanismos de busca.

A comunidade de recuperação de informação não vem dando muita atenção para as possibilidades que as relações léxico-semânticas trazem para a técnica de expansão de consultas. A razão para isso parece estar relacionada a vários experimentos passados (SALTON; LESK, 1968)(FOX, 1980) mostrando que, de maneira geral, a técnica de expansão de consultas baseada nas relações de um domínio, degrada a precisão dos mecanismos de recuperação de informação. Todavia, esta afirmação parece ser prematura e precisa ser melhor analisada.

Para validar a hipótese de que as ontologias de domínio e outros artefatos terminológicos viabilizam a técnica de expansão semântica de consultas, foram realizados dois experimentos usando a coleção de testes TREC Genômica (“ir.ohsu.edu/genomics”) e ontologias de domínio e tesauros da área biomédica. Visando uma melhor análise dos resultados, foram eliminadas variáveis que pudessem distorcer o resultado da técnica, como por exemplo, expansão automática dos termos. Por causa dos termos polissêmicos, nos quais um mesmo termo pode assumir mais de um significado, a correta expansão dos termos envolve análise de contexto. Assim, torná-la automática por meios computacionais envolve um estudo minucioso e será, por conseguinte, tema para trabalhos futuros.

Este trabalho está organizado em 8 capítulos. O primeiro deles corresponde a esta introdução, que exibe a motivação e o objetivo deste trabalho, além de sua organização. O Capítulo 2 apresenta o referencial teórico necessário para compreensão deste trabalho. No Capítulo 3 são apresentados os trabalhos relacionados. No Capítulo 4 é caracterizada a abordagem experimental. O Capítulo 5 apresenta o ambiente de avaliação criado para a realização dos experimentos. No Capítulo 6 é apresentado detalhes da realização do experimento. No Capítulo 7 são apresentados os resultados obtidos e análise dos resultados. Por fim, no capítulo 8 a conclusão contendo uma visão geral deste trabalho, assim como suas contribuições, dificuldades levantadas e suas perspectivas futuras.

2. Referencial Teórico

Para a realização deste trabalho foi necessário combinar o conhecimento de áreas que em geral são apresentadas de formas disjuntas. Desta forma, são apresentadas nas seções que se seguem, as teorias que embasaram o projeto do experimento realizado. Além disso, esse referencial teórico visa garantir uma correta interpretação dos resultados obtidos. Para este fim é apresentada uma breve revisão dos conhecimentos necessários para seu entendimento.

Inicialmente, na seção 2.1, a técnica de expansão de consultas é abordada, discutindo-se os objetivos da técnica. Em seguida, na seção 2.2, são apresentadas as técnicas usualmente empregadas para avaliação de mecanismos de recuperação de informação, visando à compreensão dos resultados obtidos com a aplicação da técnica de expansão de consultas. Finalmente, por serem parte fundamental de nossa abordagem para melhoria da técnica de expansão de consultas, apresentamos uma revisão de ontologias e na seção 2.3.

2.1 Expansão de Consultas

O problema fundamental a ser resolvido pelos sistemas de recuperação de informação (RI) consiste na busca de documentos relevantes para uma dada consulta que expressa a necessidade de informação do usuário. Porém, nem sempre os autores dos documentos usam as mesmas palavras que os usuários para se referir um mesmo

conceito. Esse problema é conhecido na comunidade de recuperação de informação como "Problema do descasamento de palavras" (*word mismatching problem*) (XU; CROFT, 1996b).

Por isso, sem um conhecimento detalhado do conteúdo presente nos documentos do acervo, a maioria dos usuários encontra dificuldades na formulação de consultas que tornem o processo de recuperação de informação mais eficiente. Isso pode ser observado no comportamento dos usuários em mecanismos de busca da Web. Tais usuários encontram as informações que precisam após reformulações sucessivas da consulta original. Este problema tende a diminuir com o aumento dos termos da consulta. Porém, diversos experimentos vêm mostrando que em geral os usuários usam 3 ou menos termos para formular sua consulta (XU; CROFT, 2000)(SPINK ET AL., 2001).

O problema do descasamento de palavras é amenizado através do emprego de duas técnicas: eliminação de variação do radical, chamada de radicalização (*stemming*) e expansão de consultas (XU; CROFT, 2000). Ambas as técnicas são aplicadas como parte da etapa do processamento de consultas (ver Figura 2). A técnica de radicalização consiste em reduzir as palavras ao seu radical eliminando suas variações léxicas. Por exemplo, considere o conjunto de palavras comer, comendo, comem. Após aplicar a técnica de radicalização, todos estes termos seriam reduzidos para o termo come. Dessa forma é possível realizar o casamento dos termos associados ao conceito de comer. Sendo assim, se um usuário busca “porque as crianças não comem”, ele

pode encontrar um documento que tenha o título “fazendo as crianças comerem”. A radicalização dos termos é feita, em geral, aplicando-se alguma variação do algoritmo de Porter (PORTER, 2006). Um exemplo desta variação pode ser encontrado no trabalho de (ORENGO; HUYCK, 2001), onde o algoritmo de Porter foi adaptado para a língua portuguesa.

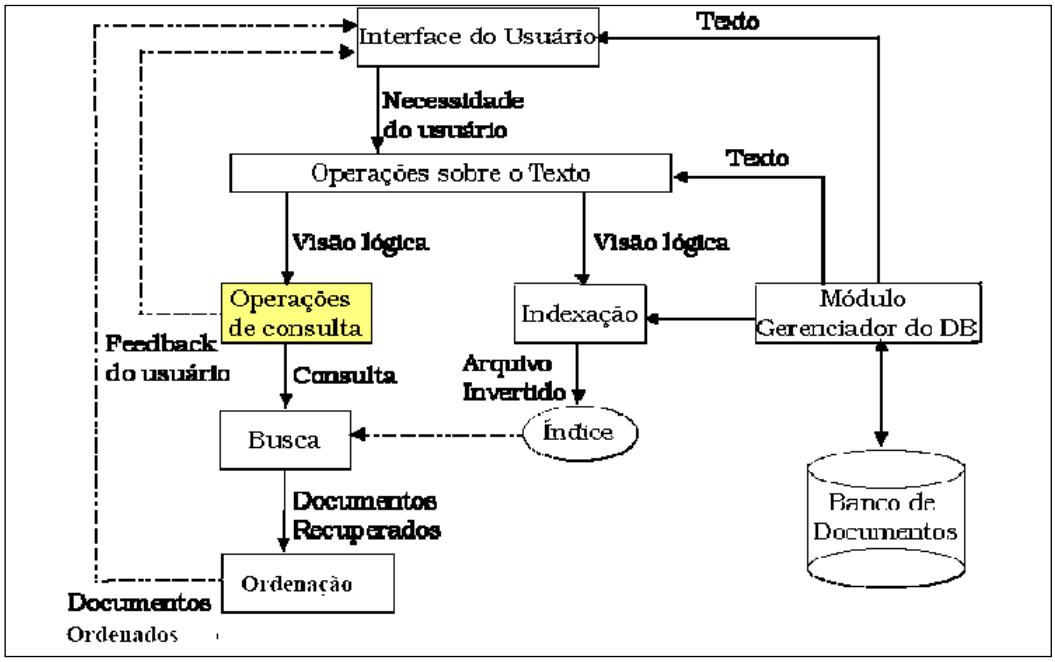

Figura 2- Arquitetura de um sistema de RI (BAEZA-YATES; RIBEIRO-NETO, 1999)

Apesar de melhorar bastante o desempenho da consulta, a técnica de radicalização não lida com problemas de homônimos, sinônimos e termos polissêmicos. Para lidar com esses problemas, utiliza-se a técnica de expansão de consultas. A idéia da técnica de expansão de consultas é buscar palavras relacionadas aos termos da consulta e acrescentar tais palavras à consulta com o objetivo de melhorar a formulação

da consulta original (KHAN; KHOR, 2004). Essa reformulação da consulta é feita em duas etapas: primeiro escolhe-se os termos que serão adicionados à consulta, depois é redefinido o peso dos termos da consulta expandida.

A técnica de expansão de consultas tem sido constantemente revisada, fazendo com que existam diferentes abordagens no seu emprego. A primeira variante refere-se à forma de escolher os termos a serem adicionados. Em geral, utiliza-se uma base de conhecimento léxico-semântico (na forma de taxonomias, tesauros ou ontologias) (VOORHEES, 1994) ou então uma base de conhecimento léxico-estatístico geralmente construída através da medida da co-ocorrência dos termos nos documentos contidos no acervo (PEAT; WILLETT, 1991). A segunda variante é se os termos escolhidos são automaticamente inseridos na consulta expandida (expansão automática) ou se são apresentados ao usuário para que este escolha os termos que serão adicionados (realimentação de relevantes ou expansão semi-automática). A última variante refere-se a se os termos adicionados devem ter pesos iguais ou menores que os termos originais da consulta. Vários experimentos vêm mostrando que em geral quando é atribuído um peso menor para os termos adicionados a eficiência do mecanismo é melhor do que com pesos iguais.

2.2 Avaliação de Desempenho em Recuperação de Informação

Sabemos que existem diversas alternativas no projeto de um sistema de recuperação de informação. Porém como podemos avaliar quais das técnicas estão trazendo benefícios para um determinado tipo de aplicação?

Em geral, o desempenho de sistemas computacionais é medido em relação ao tempo e ao espaço. Quanto menor o tempo de resposta e quanto menor o espaço utilizado, melhor o desempenho do sistema. Por exemplo, esses indicadores são usados nos sistemas gerenciadores de banco de dados para medir o tempo de indexação, o tempo de resposta de consulta e o espaço em memória principal necessário para manter o sistema em execução (CODD, 1990).

Por outro lado, em um sistema projetado para recuperação de informação, outras métricas além de tempo e espaço são importantes. A medida mais importante está relacionada à satisfação do usuário no uso do sistema (MANNING ET AL., 2008). O usuário de um sistema de recuperação de informação utiliza o sistema para atender a uma necessidade de informação. O usuário traduz esta necessidade em palavras-chave que, de forma nem sempre precisa, indicam quais documentos são possivelmente relevantes. Uma vez definido esse conjunto de documentos pelo mecanismo de recuperação de informação, a lista resposta é gerada ordenando-se os elementos deste conjunto por ordem decrescente de relevância. A avaliação da

precisão desta lista é a principal métrica utilizada na comparação dos mecanismos de recuperação de informação (BAEZA-YATES; RIBEIRO-NETO, 1999).

A área de recuperação de informação vem se desenvolvendo como uma disciplina de base empírica. Este processo de teste-avaliação exige um grande esforço para organizar os recursos necessários para sua realização. O primeiro passo para essa validação é definir a tarefa que se quer cobrir com o sistema de recuperação de informação. Isto é necessário para definirmos os tipos de documentos do acervo e que tipos de consultas devem ser definidas. O segundo passo é definir para cada consulta criada no primeiro passo o conjunto de documentos relevantes para cada uma dessas essas consultas. O resultado desse processo é denominado coleção de testes pela comunidade de recuperação de informação.

Boas coleções de testes devem ter um conjunto significativo de documentos e de necessidades de informação. Esse levantamento deve ser feito por especialistas do domínio onde se deseja usar o mecanismo de recuperação, uma vez que as necessidades de informação devem refletir o dia a dia de trabalho desses especialistas. De maneira geral, a comunidade usa um mínimo de 50 necessidades de informação em suas coleções de testes (“trec.nist.gov/data.html”)(“research.nii.ac.jp/ntcir/index-en.html”).

Existem diversas coleções de testes padrões em língua inglesa. A conferência de recuperação de informação textual, *Text Retrieval Conference* (TREC), com suporte do *National Institute of Standards and Technology* (NIST) vem avaliando uma série de

mecanismos de recuperação de informação desde 1992. As diversas edições dessa conferência definiram coleções de testes, as quais eram utilizadas por diversos grupos de pesquisas ao redor do mundo sendo seus resultados posteriormente comparados. Devido ao tamanho e qualidade de suas coleções de testes, as bases TREC são utilizadas em diversas pesquisas em recuperação de informações textuais (VOORHEES, 1994)(MANDALA ET AL., 1999).

Conforme descrito no capitulo 1 as duas medidas mais importantes na avaliação de um mecanismo de recuperação de informação são cobertura e precisão, as quais são definidas como segue (BAEZA-YATES; RIBEIRO-NETO, 1999).

Seja NI uma necessidade de informação e REL o conjunto de documentos relevantes para essa necessidade de informação. Seja RET o conjunto de documentos retornados pelo mecanismo de recuperação sendo avaliado quando testado com NI. Seja ainda RET-REL os documentos relevantes que foram retornados (interseção de RET com REL). A Figura 3 ilustra esses conjuntos.

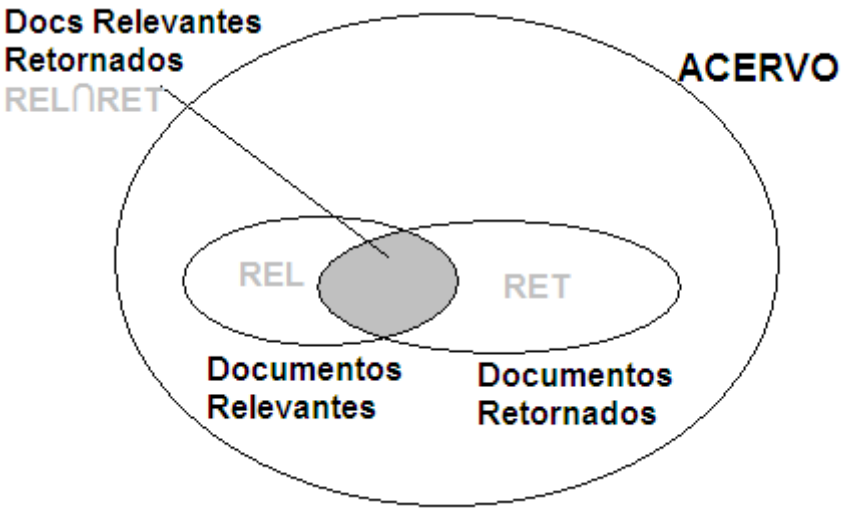

Figura 3- Conjuntos de cobertura e precisão

Considere a função cardinalidade # que retorna o numero de elementos de um conjunto. Por exemplo, #X retorna o número de elementos do conjunto X. Definimos Cobertura como a fração de documentos relevantes (o conjunto REL) que foram recuperados, ou seja:

Cobertura = #REL_RET / #REL

Definimos Precisão como a fração de documentos retornados (o conjunto RET) que foram retornados, ou seja:

Precisão = #REL_RET / #RET

Conforme definido acima, cobertura e precisão são medidas baseadas em conjuntos. Elas são calculadas sem considerar a ordem dos elementos desses

conjuntos. Porém, em um mecanismo de recuperação de informação os documentos retornados são ordenados por ordem de grau de relevância. O usuário desse sistema começa a examinar essa lista do início e vai prosseguindo analisando os documentos posteriores. Nesta situação a cobertura e a precisão variam conforme o usuário vai examinando os documentos retornados. Portanto, uma medida apropriada é traçar a curva de cobertura versus precisão como pode ser vista na Figura 4.

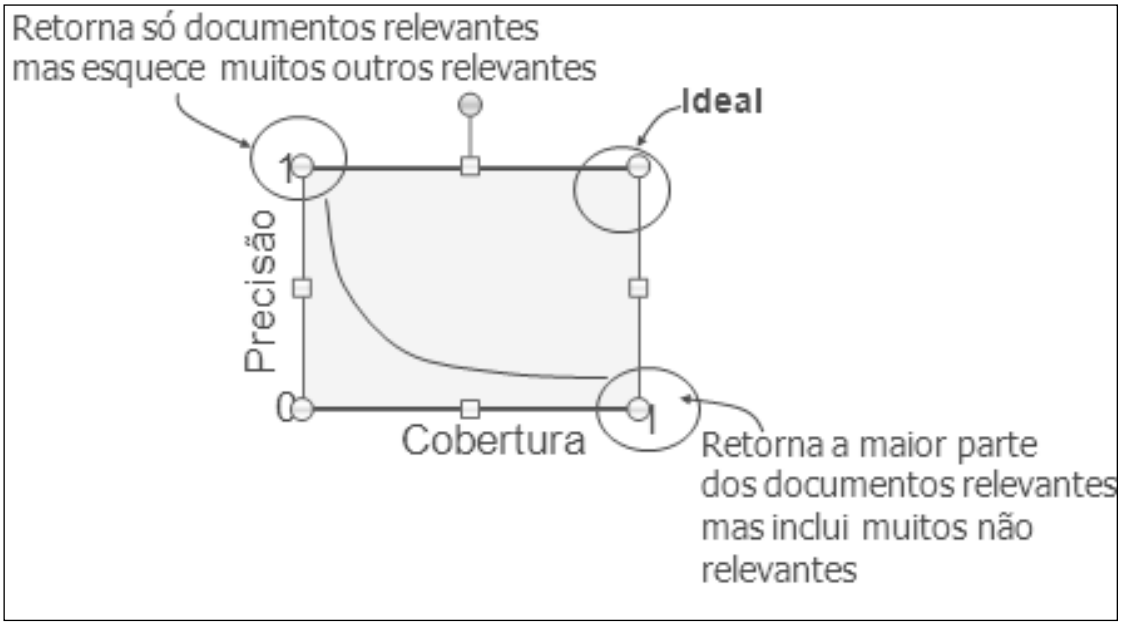

Figura 4 – Curva de Cobertura versus Precisão

Suponha que os documentos relevantes para uma determinada consulta C, sejam representados pelo conjunto

REL(C) = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

Desta forma, existem dez documentos relevantes para a consulta C. Considere que o sistema de recuperação de informação que está sendo avaliado retorne, para essa consulta, o resultado ilustrado na Figura 5. Note que a Figura mostra também a ordem em que os documentos são retornados.

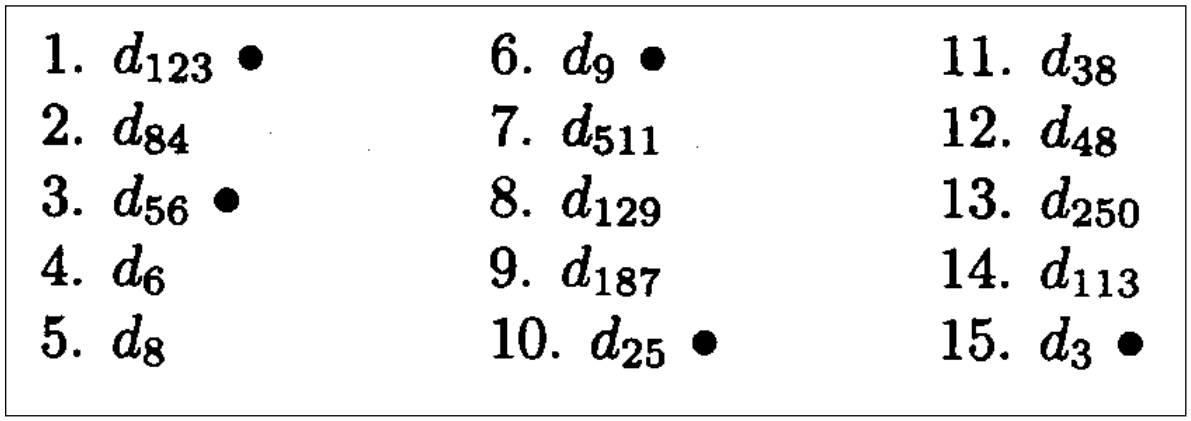

Figura 5 – Documentos retornados para a consulta C do original (Baeza R., Ribeiro B., 1999)

Os documentos que são relevantes estão marcados com um ponto após o número do documento. Sobre este resultado pode-se observar:

I. O primeiro documento retornado (d123) é um documento relevante. Como existem 10 documentos relevantes, 1 documento relevante corresponde a 10% do total de documentos relevantes. Por isso, dizemos que temos uma precisão de 100% em 10% de cobertura.

II. O documento d56 é o próximo documento relevante dos documentos retornados. Como ele está na terceira posição temos uma precisão de 66%

(2 documentos relevantes em 3 retornados) em 20% de cobertura (2 documentos relevantes de um total de 10 documentos relevantes).

III. Se continuarmos este raciocínio podemos traçar a curva de cobertura versus precisão conforme a Figura 6.

IV. A precisão para valores superiores a 50% de cobertura cai a zero porque apenas 5 dos 10 documentos relevantes foram retornados. Dessa forma, não é possível avaliar a precisão em 60% de cobertura.

V. Esta curva de cobertura versus precisão em pontos padronizados de cobertura (0%, 10%, 20%, 30%, 40% 50%, 60%, 70%, 80%, 90%, 100%) é chamada de curva de cobertura-precisão de 11 pontos.

VI. O valor de precisão em 0% de cobertura é calculado através de interpolação linear.

Quando um mecanismo de recuperação de informação é avaliado usando uma coleção de testes é preciso achar a curva de cobertura-precisão de 11 pontos de cada consulta. Só então pode-se calcular a curva de cobertura-precisão do sistema. Esta leva em consideração todas as consultas, e é calculada através da média aritmética de cada ponto de cobertura em cada consulta.

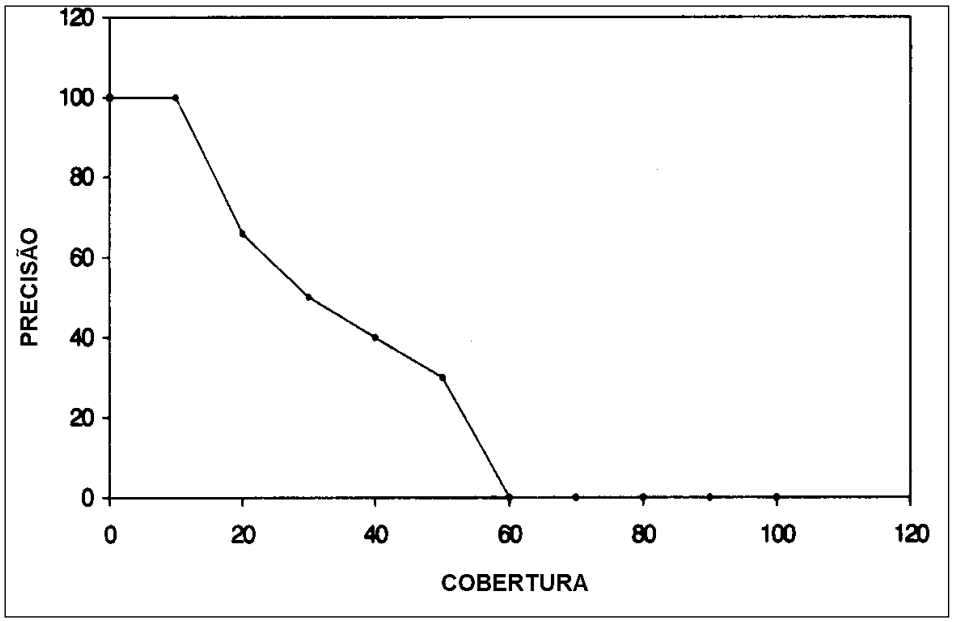

Figura 6 – Curva de cobertura-precisão de 11 pontos do original (Baeza R., Ribeiro B., 1999)

2.3 Ontologias e outros esquemas conceituais e terminológicos

O termo ontologia tem se popularizado bastante na comunidade da computação, embora tenha suas origens na Filosofia. Mas o que seria uma ontologia? Etimologicamente, o termo ontologia vem do grego *ont*, que significa ser, e *logia* significa estudo ou ciência (“Ontologia - Wikipédia, a enciclopédia livre”). Desta forma, podemos dizer que ontologia pode ser definida como o estudo da existência.

Nas ciências da computação e informação o termo ontologia se popularizou a partir da década de 90. As ontologias neste contexto têm um sentido diferente do

utilizado na filosofia. Uma das definições mais populares de ontologia neste contexto foi apresentada por (GUARINO, 1998), que diz:

> *"[...] ontologia se refere a um artefato constituído por um vocabulário usado para descrever uma certa realidade, mais um conjunto de fatos explícitos e aceitos que dizem respeito ao sentido pretendido para as palavras do vocabulário. Este conjunto de fatos tem a forma da teoria da lógica de primeira ordem, onde as palavras do vocabulário aparecem como predicados unários ou binários."*

Agora que sabemos o que é uma ontologia, precisamos definir como uma ontologia é constituída. Existem várias definições sobre o que existe em uma ontologia. Porém, existe uma estrutura comum entre essas composições que pode ser decomposta nas seguintes partes:

I. Um conjunto de conceitos, que pode ser qualquer coisa que faça parte do vocabulário como, por exemplo, meio de transporte, carro, passageiro, pneu, motorista.

II. Um conjunto de relacionamentos que ligam esses conceitos como, por exemplo, carro é uma subclasse de meio de transporte, pneu é uma parte do carro, motorista controla o carro.

III. Um conjunto de funções, que são tipos especiais de relacionamentos como relacionamentos ternários, por exemplo: Um acidente de trânsito envolve um motorista e um meio de transporte

IV. Um conjunto de axiomas, os quais são sentenças sempre verdadeiras. Como por exemplo, “todo carro tem quatro pneus” (MAEDCHE, 2002).

V. Um conjunto de instâncias, que é usado para representar os elementos, por exemplo, o fusca de placa XYZ 4432 (PEREZ; BENJAMINS, 1999).

Uma importante percepção é que o significado de um conceito depende do contexto. Considere o conceito carro. No contexto de trânsito o carro é um veículo que se move, transporta passageiros. Já no contexto de oficina mecânica o carro é um aparelho que tem motor, caixa de marchas e etc. No contexto das concessionárias de automóveis o carro é uma mercadoria que tem preço, cor, acessórios e etc. Isso significa que um dado conceito pode estar representado em diferentes ontologias e associações dependendo do domínio.

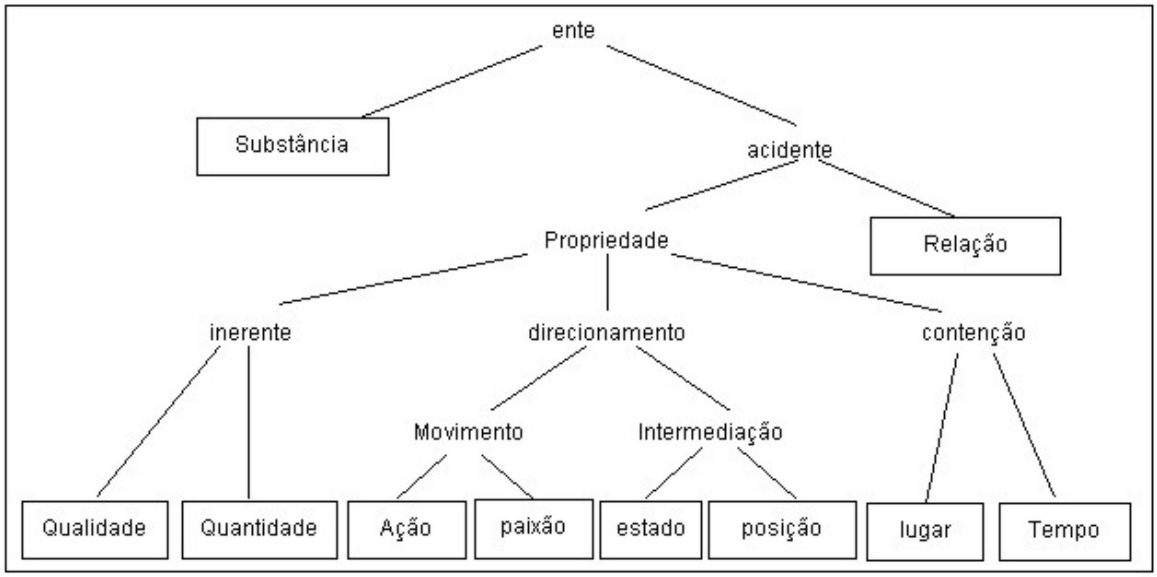

Figura 7 – A ontologia de alto nível SOWA

Existem diversas formas de se classificar uma ontologia. Uma das classificações mais usadas foi proposta por (GUARINO, 1998) que classificou as ontologias de acordo com sua dependência em relação a uma tarefa específica ou a um ponto de vista:

- Ontologias de Alto Nível: descrevem conceitos bem gerais, como por exemplo, espaço, tempo, matéria, objeto, ação, etc., que são independentes de um problema ou domínio em particular. A Figura 7 exemplifica esse tipo de ontologia.

- Ontologias de Domínio e de Tarefas: descrevem um vocabulário relacionado a um domínio genérico (como medicina ou automóveis) ou uma tarefa ou atividade genérica (como diagnosticar ou vender).

- Ontologias de Aplicação descrevem conceitos que dependem de um domínio particular e tarefas, e frequentemente combinam especializações tanto de ontologias de domínios quanto de tarefas. Estes conceitos geralmente correspondem a papeis exercidos pelas entidades do domínio quando este exerce uma determinada tarefa.

Nos últimos anos vem crescendo o interesse por construção de ontologias de domínio. Essas ontologias servem como uma excelente base de conhecimento para a aplicação da técnica de expansão de consultas. É importante ressaltar que diferentes tipos de esquemas conceituais e terminológicos podem ser utilizados como base de conhecimento em experimentos de expansão de consulta. As taxonomias (CAMPOS; HAGAR, 2008), por terem uma estrutura hierárquica, podem ser usadas como fornecedoras de relações de generalização-especialização. Os tesauros clássicos (IMRAN; SHARAN, 2009), com sua organização baseada nas relações de sinonímia, antonímia, generalização-especialização, todo-parte e relacionado_a, também podem ser empregados como base de conhecimento. Existe ainda opção de usar um tesauro estatístico (IMRAN; SHARAN, 2009) nos quais os termos são organizados de acordo com a correlação de ocorrência dos termos nos documentos do acervo.

No próximo capítulo são apresentados trabalhos relacionados à avaliação da técnica de expansão de consulta. Os trabalhos mais antigos usavam bases de conhecimento genéricas, porém os trabalhos mais modernos já utilizam bases de conhecimento específicas de domínio.

3. Trabalhos em Expansões de Consultas

Diversas pesquisas na área de recuperação de informações utilizam uma abordagem empírica, onde são realizados experimentos em um ambiente controlado em laboratório. Antes da realização do experimento, as variantes são definidas e então é realizado o experimento e seus dados de resultados são coletados e analisados. Diferentes combinações destas variantes são testadas buscando evidências na análise dos resultados. Especificamente, na avaliação de estratégias de expansão de consultas, diversas pesquisas tem se apoiado neste tipo de experimento.

Neste capítulo, apresentamos os experimentos que serviram como fonte de inspiração para os experimentos apresentados nesta dissertação. Na primeira seção apresentamos o experimento de Salton e Lesk (1968), um dos primeiros trabalhos publicados em expansão semântica de consultas. Na segunda seção apresentaremos o experimento de Fox (1980). Esse experimento se destacou por ser o primeiro a usar uma abordagem seletiva das relações semânticas. Na terceira seção apresentaremos o experimento de Voorhees (1994), um dos experimentos mais referenciados em expansão semântica de consultas. Posteriormente apresentaremos o experimento de Mandala (1999) que se destaca ao combinar diferentes tipos de tesauros como apoio para a expansão. Na quinta seção é apresentado o trabalho de Navigli e Verlardi (2003) que utilizou a definição dos termos na Wordnet para expansão. Na sexta seção é apresentado o trabalho de Lu (2008) que utilizou a TREC Genômica e o tesauro MeSH.

Por fim, na última seção é apresentado o trabalho de Stokes (2008) que combinou diferentes ontologias de domínio como base de conhecimento.

3.1 Experimento de Salton e Lesk de 1968

Experimentos de recuperação de informação que exploram relações semânticas entre termos têm sido realizados há décadas. Em um dos primeiros experimentos publicados Salton e Lesk (1968) já exploravam essas relações em pequenos acervos específicos de domínio com o auxilio de um tesauro. Estes pesquisadores realizaram vários experimentos utilizando a mesma metodologia de pesquisa que utilizamos nos dias de hoje, onde é utilizada uma coleção de testes e algumas relações léxico-semânticas. Os resultados apresentados foram gerados após dois anos de pesquisa no sistema de recuperação de Informação SMART.

O sistema SMART foi responsável por grandes avanços na área de recuperação de informação. Na época da publicação do trabalho este sistema já apresentava funções bem avançadas para seu tempo, dentre elas:

- Um algoritmo de radicalização
- Um dicionário de sinônimos
- Um tesauro hierárquico
- Um tesauro estatístico

Para realizar esses experimentos foram criadas 3 pequenas coleções de testes: a coleção IRE-3, com um conjunto de 780 *abstracts* de artigos da área de ciência da computação publicados entre 1959-1961 e um conjunto de 34 consultas; a coleção ADI, com 82 pequenos artigos sobre documentação publicados em 1963 do *American Documentation Institute* e 35 consultas; a coleção CRAN-1, com um conjunto de 200 *abstracts* de documentos sobre aerodinâmica do Projeto Cranfield-I e 42 consultas. A Tabela 1 resume as características desses acervos. Pode ser observado na última coluna desta tabela, que as médias do número de termos por consulta destes experimentos foram elevadas, o que está bem distante da média de 2 termos como foi mostrado no trabalho de (SPINK ET AL., 2001).

Tabela 1 - Coleções de Testes usadas em Salton e Lesk (1968)

	Coleção de Testes				
Experimento	**Nome**	**Domínio**	**#Docs**	**#Consultas**	**Média:T/C**
Salton e Lesk (1968) - IRE	IRE-3	Ciência da Computação	780	34	22
Salton e Lesk (1968) - CRAN	CRANFIELD1	Aerodinâmica	200	42	17
Salton e Lesk (1968) – ADI	ADI	Normas de documentação	82	35	14

Um dos experimentos analisou de que forma as relações de sinonímia impactariam a eficiência da recuperação de informação. Os resultados mostraram que estas relações produzem melhoras significativas na recuperação de informação conforme apresentado na Figura 8. Nesta figura foram traçadas as curvas de cobertura

e precisão com e sem a utilização da técnica. É notável que a curva em que foram feitas as expansões está sempre acima da curva que não foi aplicada a técnica.

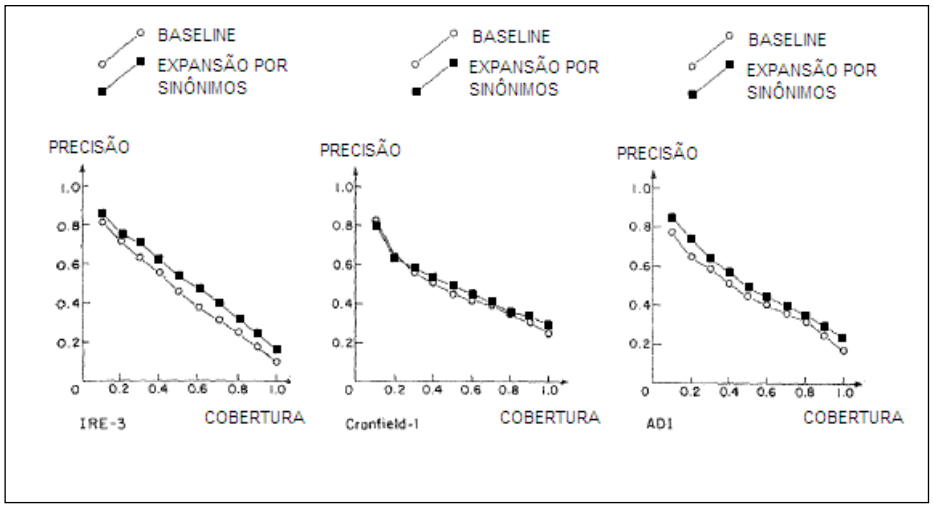

Figura 8 - Expansão por sinônimos do original Santon e Lesk (1968)

Todavia, as expansões por termos genéricos ou irmãos, selecionados a partir de um tesauro hierárquico, apresentaram resultado muito inconsistente para ser considerado útil de forma geral (ver Figura 9).

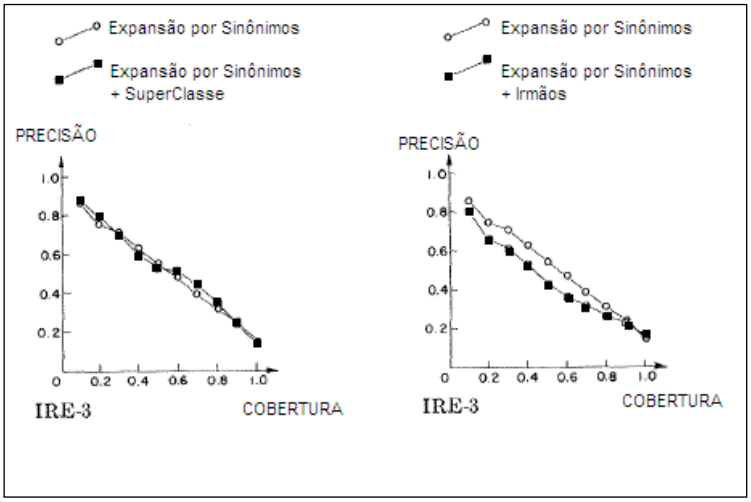

Figura 9 - Expansão por termo genérico do original Salton e Lesk (1968)

3.2 Experimento de Fox de 1980

Um dos primeiros trabalhos a explorar de forma profunda o impacto das relações entre os termos na técnica de expansão de consultas foi o experimento de Fox (1980). Fox adaptou o modelo linguístico “Significado <=> Texto” (*Meaning to Text Model - MTM*) para experimentos em recuperação de informação. Inicialmente foram categorizados os diferentes tipos de relações semânticas do modelo MTM. Essas relações léxicas foram agrupadas de acordo com o impacto esperado na precisão e cobertura de um mecanismo de recuperação de informação (ver Tabela 2).

Fox utilizou a coleção de testes ADI e o sistema de recuperação SMART, ambos descritos na seção anterior. A base de conhecimento para expansão foi gerada de forma *ad hoc* para a realização do experimento. Os resultados obtidos mostraram que a expansão de consultas utilizando todas as categorias de relações léxicas, excetuando-

se antônimos, melhorou a precisão nos primeiros dez documentos (P@10) em torno de 16%. O experimento mostrou também que as expansões baseadas em antônimos acabaram por degradar o desempenho do mecanismo de recuperação.

Este experimento serviu como um exemplo na abordagem de pesquisa para expansões semânticas de consulta. As relações foram testadas em grupos, deixando em aberto uma análise mais profunda de cada relação léxica individualmente. Porém, essa abordagem de análise com foco nos tipos de relações léxico-semânticas foi perpetuada nos experimentos posteriores.

Tabela 2 - Categorização das relações do MTM do original Fox (1980)

Numero Grupo	Descrição	Resultado Experado	Relações Incluídas
1	Fortemente Relacionados	Aumento na cobertura, Aumento na precisão	Predicados, Morfologia, Paradigmático
2	Relacionados	Aumento na cobertura	Causal, Atributo, Situacional, Mesmo Local
3	Generalização	Aumento na cobertura, Piora na precisão	Taxonomia, Todo-parte, Ordenação
4	Oposição	Aumento na Cobertura	Antônimos
5	Sinonimia	Aumento na cobertura, Aumento na precisão	Sinônimos

3.3 Experimento de Voorhees de 1994

O experimento de Voorhees (1994) é um dos trabalhos mais referenciados em expansão semântica de consultas. Seu trabalho foi o primeiro a usar uma base de conhecimento construída de maneira formal nas expansões de consultas. Voorhees utilizou o *WordNet* (“wordnet.princeton.edu”), um grande tesauro livre de domínio criado na Universidade de Princeton. Como coleção de testes foi utilizada a base TREC. Esta base consiste em um acervo livre de domínio totalizando 742.000 documentos, os quais foram extraídos de jornais e revistas. Foram utilizadas para expansão as relações léxicas de sinonímia, hiperonímia, hiponímia e relacionado_a presentes na *WordNet*. Além disso, foram testadas várias combinações de pesos para os termos adicionados pelo sistema.

O experimento foi realizado sobre o sistema de recuperação de informação SMART, que já era baseado no modelo vetorial clássico. De forma a selecionar os termos corretos para a expansão, os conjuntos de sinônimos foram escolhidos manualmente (por um ser humano) levando-se em consideração todo o contexto da consulta. Desta forma, os resultados reportados representam um limite superior do desempenho esperado de um mecanismo de expansão de consultas automático que utiliza esta estratégia de expansão.

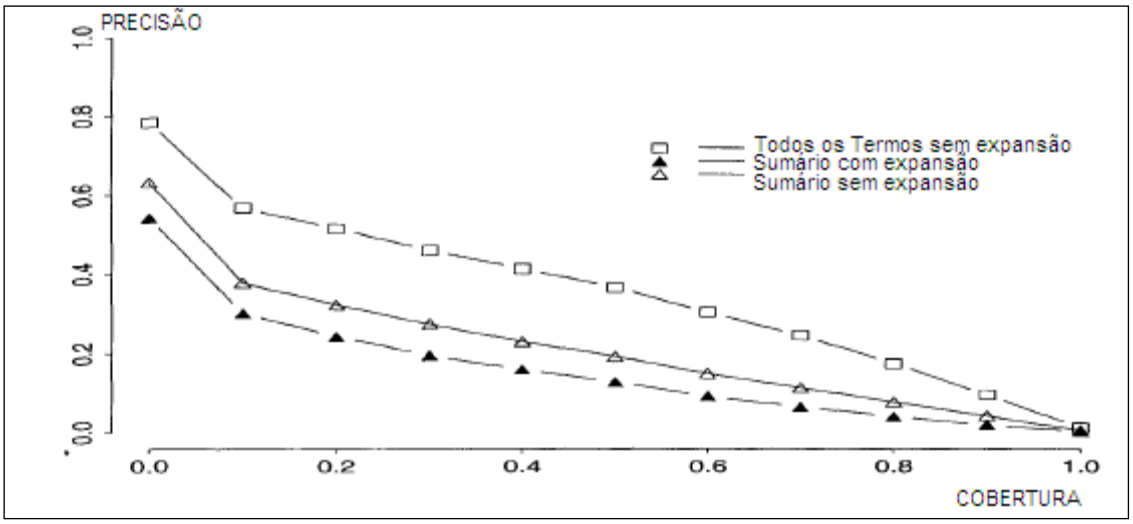

Figura 10 - Resultado do experimento de Voorhees (1994) retirado do Original

Neste experimento, Voorhees conseguiu uma melhora de 35% na precisão média de 11 pontos na expansão do campo sumário, como mostra a Figura 10. Nota-se que o resultado foi pior do que quando foram utilizados todos os termos das consultas da TREC. Porém somente no campo sumário a média de termos de das consultas foi 11,02. Os pesos utilizados na execução da Figura 10 foram 1 para os termos informados pelo usuário e 0,5 para os termos expandidos. Este resultado indicou que o vetor de consulta deve privilegiar os termos inseridos pelo usuário em relação aos termos expandidos automaticamente.

3.4 Experimento de Mandala et al. de 1999

O experimento de Mandala et al. (1999) se destacou ao propor uma técnica de expansão de consultas baseada na combinação de 3 tipos de tesauros: (i) tesauro construído à mão (*WordNet*), (ii) tesauro construído automaticamente do corpus baseado em co-ocorrência de termos nos documentos e (iii) tesauro construído do corpus baseado em co-ocorrência de termos em relações lingüísticas

O experimento foi realizado sobre a coleção de testes TREC-7 com documentos de fontes livres de domínio. Esta coleção contém aproximadamente 530.000 documentos e 50 consultas. Foi utilizada uma técnica de atribuir pesos aos termos expandidos para eliminar efeitos das expansões de termos incorretas. Os resultados mostraram ganhos significativos de aproximadamente 100% na precisão média de 11 pontos como mostra a Tabela 3.

Tabela 3 - Resultados obtidos em Mandala, et. al (1999)

		Expandido com						
Consulta	Base	Apenas WordNet	Apenas Tesauro Linguísti-co	Apenas Tesauro Estatistico	WordNet+ Linguistístico	WordNet+ Estatístico	Linguístico+ Estatístico	Todos
Titulo	0,1175	0,1299 (+10,6%)	0,1505 (+28,1%)	0,1637 (+39,3%)	0,1611 (+37,1%)	0,1698 (+44,5%)	0,1859 (+58,2%)	0,2337 (+98,9%)
Descrição	0,1428	0,1525 (+6,8%)	0,1725 (+19,4%)	0,195 (+33,4%)	0,1832 (+28,3%)	0,1973 (+38,2%)	0,2315 (+62,1%)	0,2689 (+88,3%)
Todos	0,1976	0,2018 (+2,1%)	0,2249 (+13,8%)	0,2395 (+21,2%)	0,2276 (+15,2%)	0,2423 (+22,6%)	0,2565 (+29,8%)	0,2751 (+39,2%)

3.5 O experimento de Navigli e Velardi de 2003

O experimento de (ROBERTO NAVIGLI; PAOLA VELARDI, 2003) foi um pequeno experimento realizado na Universidade de Roma, que utilizou a *Wordnet* 1.6 como base de conhecimento. Foram utilizadas as primeiras 24 das 50 consultas disponíveis na coleção de testes Web TREC 2001. Foi utilizada a API do Google como mecanismo de recuperação.

Os resultados obtidos mostraram uma melhora significativa de 22,76% na precisão dos primeiros 10 documentos (P@10) quando utilizadas as definições dos temos da consulta. Por outro lado, o ganho usando expansão por sinônimos atingiu apenas uma melhora marginal de 3,25%, resultado irrelevante do ponto de vista estatístico.

3.6 O experimento de Lu et. al. de 2009

No experimento (LU ET AL., 2009) foi testada a expansão de consultas baseada em termos do tesauro de temas da área médica MeSH (*Medical Subject Headings*). A coleção de testes utilizada foi a TREC Genômica 2006 e 2007. Essas duas coleções, após ajustes técnicos, totalizaram 160.248 artigos completos publicados pelo periódico Highwire Press e 55 consultas.

Os resultados obtidos (ver Tabela 4) foram inconsistentes mostrando uma pequena melhora de aproximadamente 4% nas consultas da versão 2006 e uma

pequena piora de aproximadamente 4% na versão 2007. Este resultado tem pouca significância do ponto de vista estatístico.

Tabela 4 - Resultados obtidos em Lu (2008)

TREC	Expandido	P@5	P@10	P@20
2006	Sim	0,622	0,616	0,621
	Não	0,613	0,592	0,584
2007	Sim	0,538	0,525	0,503
	Não	0,566	0,548	0,519

3.7 O experimento de Stokes de 2009

O experimento (STOKES ET AL., 2009) comparou o impacto que a escolha de uma determinada ontologia pode produzir na técnica de expansão de consultas. Para realizar este experimento foi utilizada a coleção de testes TREC Genômica 2006 e um mecanismo de recuperação de informação baseado no modelo probabilístico Okapi (ROBERTSON, 1997). Os resultados obtidos mostraram que a melhor opção é combinar as bases de conhecimento, onde foi atingida uma melhora de 20,6% nas expansões por sinônimos (ver Tabela 5). As expansões por termos genéricos e por termos específicos em geral degradaram o desempenho do mecanismo de recuperação de informação.

Tabela 5 - Resultado do experimento de Stokes (2009)

Base de Conhecimento	MPM Exp Sinonimos	Resultado	MPM Exp Genérico	Resultado	MPM Exp Específico	Resultado
Baseline	0,414		0,414		0,414	
E_Gene	0,466	12,56%				
ONIM	0,452	9,18%				
UniProt	0,446	7,73%				
Hugo	0,411	-0,72%				
SNOMEDCT	0,424	2,42%	0,408	-1,45%	0,407	-1,69%
UMLS	0,419	1,21%	0,411	-0,72%	0,395	-4,59%
MTH	0,412	-0,48%	0,399	-3,62%	0,407	-1,69%
MeSH	0,412	-0,48%	0,424	2,42%	0,406	-1,93%
Todas Combinadas	**0,499**	**20,53%**				

3.8 Comparação dos Trabalhos Relacionados

A Tabela 6 mostra de maneira resumida as principais variáveis em cada um dos trabalhos relacionados. Em relação ao número de documentos, existe uma tendência de realização de experimentos com acervos cada vez maiores, o que reflete o aumento de volume de informação disponível com as ferramentas de navegação na Web. Por sua vez, o número médio de consultas por experimento foi de 39,3 e mediana de 37,5, indicando que um número de consultas em torno de 40 é suficiente para a comparação.

Uma divisão bem delineada neste quadro comparativo se refere ao uso de bases de conhecimento específicas de domínio. Com a crescente adoção de ontologias de domínio pela comunidade científica, principalmente na área biomédica, aumentaram-se as expectativas de conseguir melhores resultados na expansão usando estes tipos de ontologias.

Tabela 6- Resumo dos trabalhos relacionados

	Salton	Fox	Voorhees	Mandala	Navigli	Lu	Stokes
Ano	1968	1980	1994	1999	2003	2009	2009
# Docs	780,200,82	82	742K	530K	1,6M	160K	160K
# Consultas	34,42,35	35	50	50	24	55	28
Bases de Conhecimento	Tesauro Harris	Modelo Texto => Significado	Wordnet	WordNet+ Estatístico+ Linguístico	WordNet	MeSH	GO, SnomedCT, Mesh, UMLS, MTH, E_Gene, Uniprot, Onim, Hugo
Tipos de Expansões	Sinônimos, Generali-zação, irmãos	Sinônimos, Antônimos, Todo-Parte, Causa-Efeito, Generalização, Especialização	Sinônimos, Generali-zação, especiali-zação, relacionado a	Sinônimos	Sinônimos, Generali-zação, Definição	Relacionado_a	Sinônimos, Generali-zação, especiali-zação
Variação de pesos para termos expandidos	Não	Não	Sim	Sim	Sim	Não	Não
Processo de Expansão	Manual	Manual	Manual	Automática	Automática	Automática	Automática

As relações usadas para expansão apresentam uma uniformidade no uso de sinônimos e generalização. Excetuando-se o trabalho de Fox (1980), as relações que vinham sendo usadas para expansão, eram as relações tipicamente encontradas em tesauros genéricos, tais como: sinônimos, generalização-especialização e relacionado_a. O uso de ontologias nos possibilita usar relações específicas de domínio. Por exemplo, no domínio da genômica, um gene pode estar associado a uma

doença, logo poderíamos usar a doença associada como expansão de um determinado gene.

Outra variável importante neste tipo de experimento se refere à avaliação de diferentes pesos para os termos expandidos. Geralmente observa-se um melhor comportamento do mecanismo de recuperação de informação quando os termos escolhidos pelo usuário têm um peso maior do que os termos expandidos pelo sistema. Em geral, a expansão por sinônimos apresenta um ponto ótimo com o peso 0.5, ou seja, os termos inseridos pelo usuário possuem um peso duas vezes superior aos sinônimos inseridos pelo sistema.

Por fim, outra característica importante refere-se à forma como os termos são expandidos. Dado que as rotinas de expansão automática de termos possuem uma taxa de erro, tanto pela dificuldade de encontrar os sintagmas nominais quanto na escolha do sentido correto para os termos polissêmicos, a expansão manual, apesar de muito mais trabalhosa, gera resultados que representam um limite superior da técnica de expansão de consultas.

Para realizar esses tipos de experimentos é necessária a aplicação de uma pesquisa de base empírica. No próximo capítulo é caracterizada a metodologia de pesquisa utilizada neste trabalho.

4. Metodologia da Pesquisa

Neste capítulo, é descrita a metodologia de pesquisa que guiou esta investigação, realçando-se seu caráter empírico, ou seja, apoiada em um experimento realizado em laboratório, coleta de dados e observação de suas evidências.

Inicialmente é discutido o caráter científico da investigação assim como sua classificação de acordo com critérios disponíveis na literatura. Em seguida, são apresentadas as principais variantes que se referem ao planejamento de um experimento de expansão de consultas. A seguir, os recursos utilizados na avaliação da hipótese são discutidos, mostrando quais são os módulos necessários para a realização do experimento. Os motivos da utilização de cada módulo são expostos bem como o critério usado na seleção do tipo de cada módulo.

4.1 Caracterização da pesquisa

De acordo com (GIL, 1999), toda pesquisa tem um caráter pragmático, e caracteriza-se como um “processo formal e sistemático de desenvolvimento do método científico. O objetivo fundamental da pesquisa é descobrir respostas para problemas mediante o emprego de procedimentos científicos”. Mas como se classificam os procedimentos científicos empregados neste trabalho?

Em relação à natureza da pesquisa, podemos classificá-la como pesquisa de natureza aplicada, pois objetiva gerar conhecimentos para uma aplicação prática dirigida à solução do problema do descasamento de palavras-chave. Portanto, trata-se de um problema de interesse local da área de recuperação de informação. Por outro lado, sob o aspecto da forma de abordagem do problema, esta pesquisa possui uma característica fortemente quantitativa, pois avalia a veracidade de suas hipóteses através da análise estatística de medições típicas da área de recuperação de informação, tais como cobertura e precisão.

Em relação aos seus objetivos, este trabalho assume um caráter descritivo, pois descreve de uma possível abordagem (expansão de consultas baseada em ontologias de domínio) a um problema bem definido (problema do descasamento de palavras-chaves). Nesta abordagem utilizamos uma nova parametrização para a variável base de conhecimento ao assumir as ontologias de domínio como fonte de informação.

Por fim, em relação aos seus procedimentos técnicos, esta pesquisa assume um formato experimental, ao se definir como objeto de estudos a técnica de expansão de consultas. Foram isoladas as principais variáveis que influenciam este tipo de problema tais como: natureza da base de conhecimento, podendo ser específica de domínio ou genérica; forma de estruturação da base de conhecimento podendo assumir o papel de taxonomias, tesauros ou ontologias; características quanto à natureza do acervo, podendo ser genérico ou específico de domínio; forma de expansão de termos,

podendo ser manual ou automática; pesos para os termos expandidos, podendo ser fixos ou ajustáveis.

Desta forma, esta pesquisa assume um formato de pesquisa aplicada quantitativa descritiva e experimental. Na próxima seção apresentaremos as questões que devem ser respondidas no momento do planejamento do experimento de expansão de consultas, bem como suas principais variáveis.

4.2 Definição do problema e hipótese

Conforme apresentado no capítulo 3, a técnica de expansão de consultas baseada no significado dos termos vem sendo testada há décadas. Os resultados obtidos nestes experimentos não apresentaram uma melhora consistente. Com isso, as comunidades de prática e acadêmica de recuperação de informação acabaram rotulando a técnica de expansão semântica de consultas como uma forma de aumentar a cobertura em detrimento da precisão.

Todavia, os trabalhos passados não tinham à sua disposição as ontologias de domínio, que vêm se popularizando nos últimos anos. Em algumas áreas de conhecimento, como exemplo a área biomédica, esquemas terminológicos e ontologias de domínio são utilizadas como uma forma de vocabulário controlado. As ontologias, em geral, têm uma base científica, com processos de construção e manutenção formal, e são controladas por um comitê de especialistas, tais como as ontologias da OBO.

A princípio, essas ontologias garantiriam que os diferentes grupos de pesquisa se comunicassem usando um mesmo conjunto de significados léxicos-semânticos. Uma vez que os documentos do acervo são escritos usando os termos deste vocabulário controlado, as ontologias de domínio seriam uma excelente base de conhecimento para a expansão de termos das consultas.

Uma forma de validar esta afirmação seria realizar um experimento de expansão de consultas usando estas ontologias como base de conhecimento e comparar os resultados obtidos com o resultado de experimentos anteriores. Dessa forma, pode-se definir que o objetivo deste trabalho é validar ou negar a seguinte hipótese:

"O uso de ontologias de domínio e outros esquemas conceituais e terminológicos como base de conhecimento para a técnica de expansão de consultas produz resultados melhores que os obtidos em experimentos anteriores."

4.3 Etapas de um experimento de expansão de consultas

A realização de um experimento de expansão de consultas é composta de três etapas como mostra a Figura 11. Inicialmente na etapa de projeto de pesquisa é definido o problema a ser atacado bem como a hipótese de solução deste problema. Como o objetivo de aplicação da técnica de expansão de consultas é atingir uma melhora na curva de cobertura e precisão (definição do problema), basicamente o que precisa ser definido é uma nova técnica de expansão que hipoteticamente produza melhores resultados.

Uma vez definida a nova técnica passa-se para a atividade do planejamento do experimento. Nesta atividade as principais variantes são determinadas de forma a comprovar ou refutar a hipótese escolhida. Além disso, são definidos o acervo a ser utilizado, e a base de conhecimento que será usada para sugerir os termos que serão adicionados à consulta do usuário, além de outras variantes que estão detalhadas na seção 4.4.

Uma das atividades mais trabalhosas e fundamentais para a realização deste tipo de experimento é a preparação do ambiente de avaliação. O objetivo desta atividade é criar um ambiente de laboratório que automatize a aplicação da técnica e o armazenamento dos resultados. Nesta atividade são escolhidas as ferramentas computacionais que farão parte deste ambiente laboratorial. Além disso, o acervo é indexado usando o mecanismo de recuperação de informação escolhido.

Uma vez disponibilizado o ambiente de experimentação pode-se iniciar a etapa de realização do experimento. Esta etapa se repete várias vezes de acordo com o número de consultas, o número de relacionamentos léxico-semântico e a quantidade de pesos sendo avaliadas. Ao final desta etapa, todos os resultados estão disponíveis para o início da etapa de análise dos resultados.

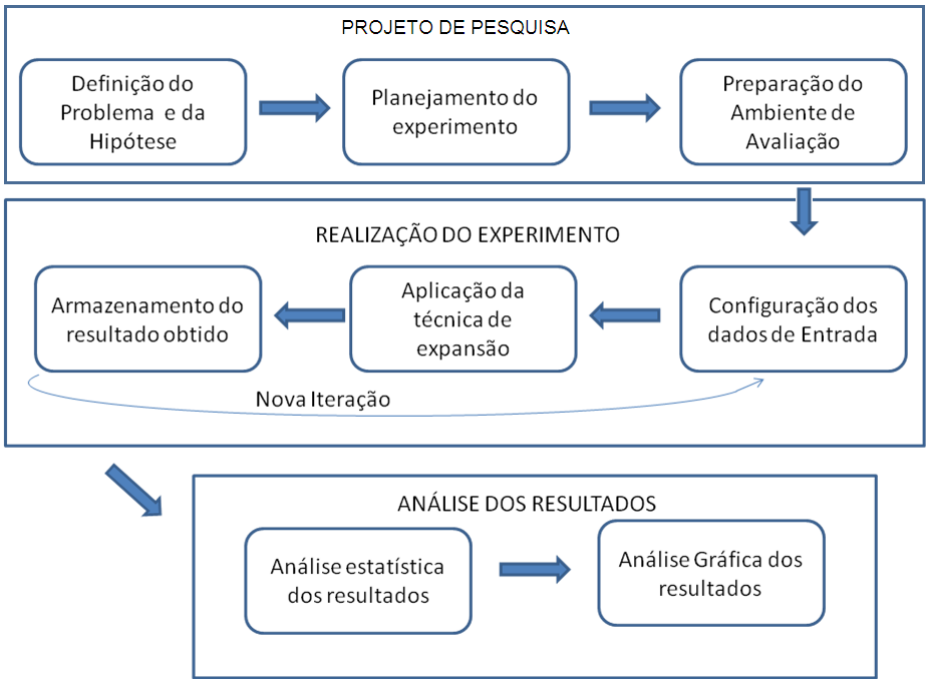

Figura 11 - Etapas de um experimento de expansão de consultas

Enfim, a última etapa de um experimento de expansão de consultas é a avaliação dos resultados obtidos pela aplicação da técnica. Nesta etapa é quantificada a precisão média de 11 pontos para cada relação léxico-semântica, para cada peso. Posteriormente é feita a análise gráfica do resultado gerando-se uma curva de cobertura e precisão.

4.4 Planejamento de um experimento de expansão de consultas

Em um experimento de expansão de consultas, existem diversas variáveis que devem ser definidas para a correta realização da pesquisa. Essas variáveis devem ser escolhidas de forma a comprovar ou refutar as hipóteses levantadas na pesquisa em

questão. Nesta seção são apresentadas as principais variantes e são discutidas também as conseqüências da sua determinação nos experimentos.

A primeira decisão que deve ser tomada se refere à escolha de um acervo específico de domínio ou um acervo genérico como mostra a Figura 12. A escolha desta variante em geral costuma determinar também a escolha da base de conhecimento. Geralmente quando é escolhido um acervo específico de domínio utiliza-se também uma base de conhecimento específica de domínio e vice-versa. Em uma base de conhecimento ideal todos os termos presentes no acervo podem ser encontrados nesta base de conhecimento. Uma grande vantagem de uma base específica de domínio é a redução das expansões incorretas de termos que têm mais de um significado (termos polissêmicos).

Por sua vez, deve ser determinada também a expressividade da base de conhecimento. A escolha de uma taxonomia permite que sejam avaliadas as expansões semânticas baseadas nas relações de generalização-especialização. Por outro lado, a escolha de um tesauro permite não só a expansão por termo genérico e termo específico, como também por relações de sinonímia e antonímia. Por fim a escolha de uma ontologia permite a utilização das relações presentes nos tesauros e outros tipos de relações tais como todo-parte, causa-consequência, regula-regulado e etc.

Outra variante que deve ser definida é a forma como os termos da base de conhecimento são sugeridos. A escolha automática dos termos é a opção menos custosa do ponto de vista do tempo necessário para a realização da expansão. Todavia, como não existem algoritmos de escolha automática que tenham uma taxa de acertos comparável à análise humana, a expansão manual é preferível, pois corresponde ao limite superior da técnica de expansão de consultas sendo avaliada. Finalmente é necessário definir se os pesos serão ajustáveis ou se serão fixos.

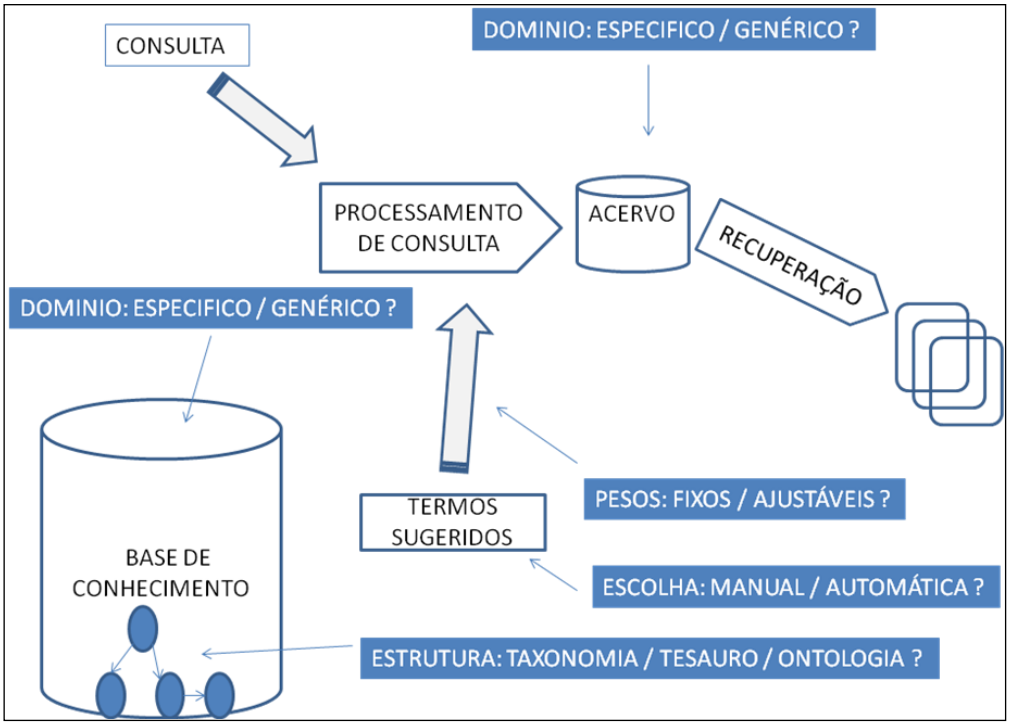

Figura 12 - Planejamento de um experimento de expansão de consultas

A escolha da utilização de pesos fixos tem a vantagem de ter uma implementação mais simples, pois não é necessário dispor de um mecanismo de ajuste de pesos no módulo de expansão de consultas. Por outro lado, a escolha de pesos ajustáveis permite determinar o ponto de configuração ideal para a técnica de expansão de consultas sendo avaliada, determinando assim seu melhor desempenho.

Para realizar experimentos de expansão de consultas é necessário escolher recursos que atendam à escolha das variantes feitas no momento do planejamento do experimento. Conforme demonstrado na Figura 13 os principais recursos utilizados neste tipo de experimento são: uma coleção de testes; um mecanismo de recuperação de informação; um mecanismo para aplicação da técnica de expansão de consultas; uma ferramenta para medir cobertura e precisão; uma ferramenta para geração de gráficos para gerar a curva de cobertura X precisão.

Um dos recursos mais importantes para a realização de um experimento de expansão de consultas é a coleção de testes. Uma coleção de testes é um recurso usado frequentemente em experimentos de recuperação de informação. Uma coleção de testes na verdade é um conjunto de recursos compreendido por um acervo, um conjunto de consultas e o conjunto de documentos do acervo relevantes para cada consulta. A escolha da coleção de testes a ser utilizada deve atender às características definidas no planejamento do experimento. Se no planejamento foi definido que serão avaliadas as expansões específicas de domínio, deve-se escolher uma coleção de testes que possua um acervo de documentos específico de domínio e vice versa.

O mecanismo de recuperação de informação é outro recurso essencial para a realização de experimentos de expansões de consultas. O mecanismo escolhido deve implementar um modelo que seja considerado o estado da arte entre os modelos disponíveis de recuperação de informação. Atualmente existem dois modelos que são considerados estados da arte: o modelo vetorial e o modelo probabilístico (BAEZA-YATES; RIBEIRO-NETO, 1999).

Outro recurso que precisa estar disponível é um mecanismo que implemente a técnica de expansão de consultas que está sendo proposta. Em geral, como os trabalhos de expansão de consultas apresentam uma nova abordagem para a técnica de expansão de consultas, esse mecanismo é construído especificamente para o experimento que foi planejado.

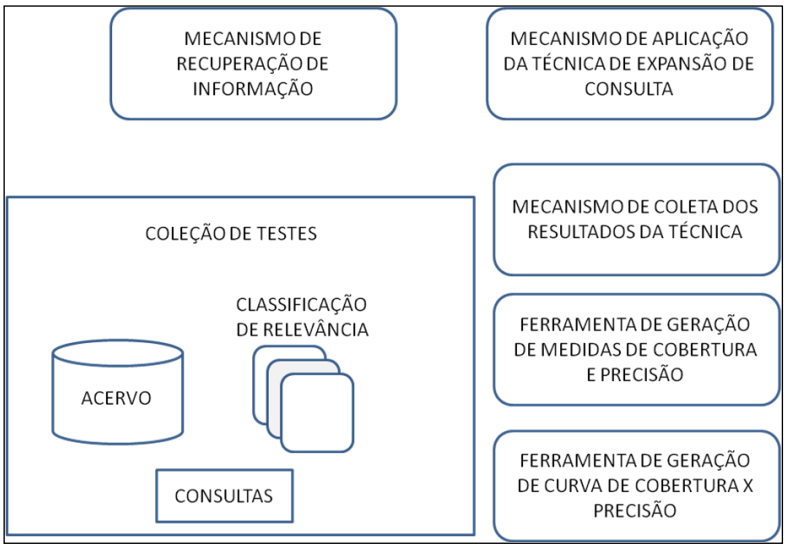

Figura 13 - Recursos necessários para experimentos de expansão de consultas

Um mecanismo de recuperação de informação apresenta os resultados de uma consulta em ordem decrescente de relevância. Esses dados são utilizados para a medição de cobertura e precisão, conforme explicado na seção 2.2. Como em geral são avaliadas pelo menos 40 consultas (mediana do número de consultas por experimento, ver Tabela 6) por execução, o armazenamento destas listas de forma manual se torna inviável neste tipo de experimento. Por isso, deve-se implementar um mecanismo que armazene este resultados em arquivos, os quais serão posteriormente utilizados como entrada pelo programa que gera as medições estatísticas de cobertura e precisão.

Outro recurso de grande importância é a ferramenta de geração das medidas de cobertura e precisão. Conforme demonstrado na seção 2.2, o cálculo de cobertura e precisão leva em consideração duas listas ordenadas, a lista de relevância, fornecida pela coleção de testes, e a lista de resultado. O cálculo para a geração de precisão de uma única consulta com 100 documentos relevantes demanda a computação de 100 comparações. Como são feitas pelo menos 40 consultas por execução, são feitas no mínimo 4000 comparações por iteração. Em um experimento simples de 10 iterações (o experimento realizado neste trabalho totalizou 40 iterações) teríamos 40.000 comparações tornando o problema extremamente custoso.

Para evitar que este trabalho seja feito de forma manual, faz-se necessário a utilização de um programa que automatize o cálculo de cobertura e precisão, o qual recebe como entrada dois arquivos: um contendo a lista de documentos relevantes para cada consulta e outro contendo a lista de documentos retornados usando a técnica de

expansão sendo avaliada para cada consulta. Este programa gera como saída o cálculo de cobertura e precisão média de 11 pontos para cada consulta e totalizado por iteração.

Por fim, o último recurso necessário é a utilização de uma ferramenta de criação de gráficos para a geração da curva de cobertura x precisão. Os dados gerados pelas medições de cobertura e precisão são transportados para esta ferramenta para a geração dos gráficos, onde é possível ter uma visão holística do resultado da expansão.

5. Ambiente de avaliação experimental

Neste capítulo são apresentados os passos que foram necessários para preparar o ambiente de experimentação utilizado na realização dos experimentos de expansão de consultas. A realização deste tipo de experimento requer a coleta de informações de forma minuciosa e controlada. Conforme exposto na seção 4.4, os módulos necessários para este tipo de experimento incluem: um mecanismo de recuperação de informação, uma coleção de testes, um mecanismo que implemente a técnica a ser avaliada, uma aplicação para geração das medições de cobertura e precisão e uma aplicação para geração dos gráficos da curva de cobertura X precisão. A

Figura **14** mostra uma visão geral do ambiente criado para a realização e avaliação dos experimentos.

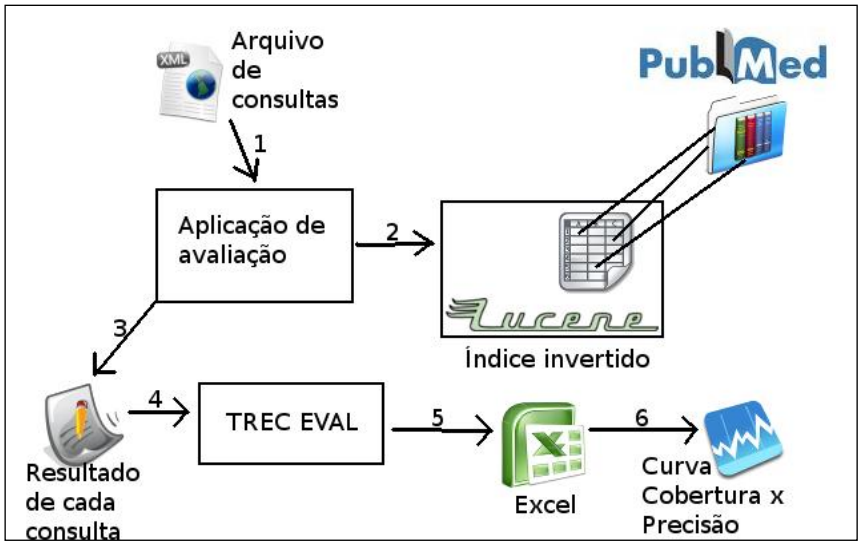

Figura 14 - O ambiente de avaliação

5.1 Escolha do mecanismo de recuperação da informação

A preparação do ambiente de avaliação começa com a escolha de um mecanismo de recuperação de informação. Atualmente, a comunidade de pesquisadores de recuperação de informação define duas categorias de mecanismos que são consideradas como *status quo*: os mecanismos baseados no modelo vetorial e o os mecanismos baseados no modelo probabilístico. Dessa forma, o mecanismo escolhido deve usar um desses dois modelos de recuperação de informação.

Para a realização deste trabalho escolhemos o Lucene (“Welcome to Lucene!”) como mecanismo de recuperação de informação. O Lucene combina as vantagens dos mecanismos baseados em modelo vetorial com as vantagens dos mecanismos baseados no modelo booleano (MANNING ET AL., 2008). O modelo booleano permite o uso de operadores lógicos da álgebra de Boole, tais como: AND, OR e NOT. Além disso, tem a vantagem de ser um módulo de software gratuito e de código aberto e por isso é utilizado como motor de busca em grandes websites da Internet tais como: Monster Jobs (“Monster - Empregos, Networking, Orientação Profissional. Siga a sua Paixão.”), Wikipedia (“Wikipedia”), SourceForge (“SourceForge.net: Find and Develop Open Source Software”) e outros.

5.2 Escolha da coleção de testes

A TREC Genômica foi uma linha de pesquisa que se iniciou em 2003 e foi finalizada em 2007 da Conferência de Recuperação Textual (*Text REtrieval Conference*, www.trec.nist.gov) organizada e fomentada pelo Instituto Americano de Normas e Tecnologias (*U.S. National Institute of Standards and Technology*, http://www.nist.gov/). A Conferência de Recuperação Textual é realizada anualmente e tem um formato de desafio-proposta-avaliação. Inicialmente é definido um problema a ser resolvido no contexto de recuperação de informação e posteriormente é definida uma coleção de testes para esse problema. Essa coleção de testes é submetida aos diferentes grupos de pesquisa em recuperação de informação em todo mundo. Estes grupos elaboram suas propostas de solução e na ocasião do evento é formado um fórum onde os resultados são apresentados e avaliados.

A linha de pesquisa TREC Genômica veio para atender aos recentes avanços da área de pesquisa em biotecnologia. Atualmente, com o desenvolvimento de novas ferramentas para medir expressões gênicas, os pesquisadores têm que lidar com um volume massivo de dados que podem chegar a dezenas de milhares de registros de dados em uma única amostragem biológica. Esses dados são utilizados para descobrir novas associações em potencial entre uma variedade de genes, doenças e outras entidades biológicas. Essa explosão de dados trouxe um aumento significativo no conhecimento científico de biologia, gerando um grande volume de artigos científicos em biologia. Para manterem-se atualizados, os grupos de pesquisa da área biomédica

necessitam de novas abordagens de recuperação de informações em acervos de artigos científicos. Esse novo cenário tem sido a principal motivação de pesquisa de recuperação de informação na área biomédica.

As coleções de testes da pesquisa TREC Genômica podem ser obtidas gratuitamente para fins de pesquisas. O acesso a elas é liberado após assinar o Acordo de Uso dos Dados e enviá-lo ao Departamento de Fornecimento de Dados TREC, do Instituto Americano de Normas e Tecnologias. Como o domínio da biomédica já possui uma variedade de ontologias de grande utilização, tais como: a Gene Ontology, SnomedCT, MeSH e UMLS, a TREC Genômica é uma ótima opção para experimentos de recuperação de informação com apoio de ontologias de domínio.

As coleções de testes mais importantes para os nossos experimentos de recuperação de informação são as coleções *ad hoc 2004* e *ad hoc 2005*. Essas coleções modelam a tarefa de atender à necessidade de informação de um pesquisador da área biomédica que está entrando em uma nova linha de pesquisa. Esse pesquisador busca atender à sua necessidade de informação acessando um sistema de recuperação de informação para encontrar a informação na literatura científica da área biomédica.

Como em geral os pesquisadores da área biomédica iniciam uma linha de pesquisa utilizando o MEDLINE, foi utilizado como coleção de documentos um subconjunto de 10 anos, incluindo os anos de 1994 até 2003, o que representava em

2004 aproximadamente 33% do acervo completo do MEDLINE. Esse subconjunto contém aproximadamente 4,6 milhões (4.591.008) de documentos. Esses documentos foram extraídos e disponibilizados no formato interno do MEDLINE. Sem usar compressão, este acervo de documentos totaliza 9,5GB de dados. A Figura 15 mostra o formato de um destes documentos.

```
PMID- 10793666
DA  - 20000516
DCOM- 20000516
LR  - 20031114
IS  - 0305-1048
VI  - 1
IP  - 1
DP  - 1974 Jan
TI  - The solubility of calf thymus chromatin in sodium chloride.
PG  - 129-39
FAU - Davies, K E
AU  - Davies KE
FAU - Walker, I O
AU  - Walker IO
LA  - eng
PT  - Journal Article
PL  - ENGLAND
TA  - Nucleic Acids Res
JID - 0411011
RN  - 0 (Chromatin)
RN  - 0 (Histones)
RN  - 7647-14-5 (Sodium Chloride)
SB  - IM
MH  - Animals
MH  - Cattle
MH  - Chromatin/chemistry/*isolation & purification
MH  - Circular Dichroism
MH  - Electrostatics
MH  - Histones/isolation & purification
MH  - Magnetic Resonance Spectroscopy
MH  - Sodium Chloride
MH  - Solubility
MH  - Thymus Gland/chemistry
SO  - Nucleic Acids Res 1974 Jan;1(1):129-39.
```

Figura 15 - Exemplo de um documento da TREC Genômica

As consultas (ver Figura 16) para a tarefa *ad hoc* foram geradas através de entrevistas com biólogos e representam as necessidades de informação que estes enfrentam no seu dia a dia de trabalho. Essas necessidades de informação foram agrupadas e seus campos foram organizados no formato a seguir:

- ID – Identificador
- Título (*Title*) – Uma sentença curta que descreve a necessidade de informação
- Necessidade de Informação (*need*) – Uma sentença detalhada da necessidade de informação
- Contexto (*context*) – Informações do entorno para colocar a necessidade de informação em contexto.

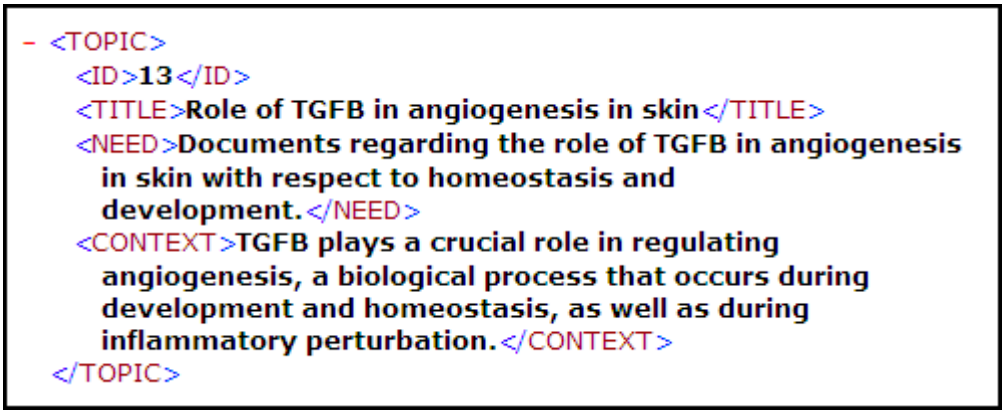

Figura 16- Exemplo de uma consulta da base TREC Genômica

A condição necessária para medir cobertura e precisão é que os elementos do conjunto de documentos relevantes de cada consulta sejam conhecidos. Nas conferências TREC a definição dos documentos relevantes é feita usando o método de *pooling* com vários mecanismos de recuperação de informação diferentes. O método de *pooling* consiste em submeter a consulta para os diferentes tipos de mecanismo de

recuperação de informação, retirar os primeiros K documentos retornados por cada motor de busca (em geral, usa-se K=100) e colocar esses documentos no *pool*. Então, um comitê de especialistas do domínio realiza o julgamento final quanto à relevância ou não relevância de cada documento.

Na coleção TREC Genômica os julgamentos de relevância estão agrupados nos arquivos denominados de “qrels” os quais apresentam o formato:

CONSULTA_ID ITERAÇÃO DOCUMENTO_ID RELEVANCIA

Neste formato, **CONSULTA_ID** é o código da consulta, **ITERAÇÃO** é a quantidade de iterações de feedback (geralmente não usado, definido como 0), **DOCUMENTO_ID** é o código oficial do documento no acervo, **RELEVÂNCIA** é um número que representa o julgamento de relevância pelos especialistas do domínio sendo: 0 para os não relevantes, 1 para os possivelmente relevantes e 2 para os definitivamente relevantes.

A Figura 17 representa um trecho do arquivo qrels da TREC Genômica *Adhoc* 2004. A primeira linha indica que o documento de número identificador 10383147 no Medline é definitivamente relevante para a consulta cujo identificador é 13.

```
13      0       10383147        2
13      0       10411901        1
13      0       10608595        2
13      0       10630313        2
13      0       10631522        2
13      0       10764007        2
13      0       11137399        1
13      0       11137400        1
13      0       11170301        2
13      0       11170303        2
13      0       11424088        2
13      0       11973359        1
13      0       12087057        1
13      0       12391391        2
13      0       12539179        2
13      0       7533322 2
13      0       9049187 2
13      0       8600310 2
13      0       8829826 2
13      0       8842485 2
13      0       8934933 2
13      0       9002217 2
13      0       9364301 2
13      0       9619049 2
```

Figura 17 - Trecho do arquivo qrels TREC Genômica 2004

5.3 Escolha das bases de conhecimento

Conforme definido no Capítulo 2, a técnica de expansão de consultas consiste em adicionar termos aos termos escolhidos pelo usuário para a realização da consulta. Mas qual a origem desses termos e como escolhê-los?

A abordagem escolhida consiste em buscar termos de tesauros, taxonomias e ontologias de domínio. Desta forma, para atender a este objetivo, é necessário achar um esquema terminológico, ou um conjunto de esquemas, que expressem o conhecimento do domínio do acervo. Como utilizamos a coleção de testes TREC Genômica, a qual extraiu seu acervo do MEDLINE, que é o principal acervo para

pesquisadores da área biomédica, utilizamos como base de conhecimento os esquemas terminológicos mais populares desta área (ver tabela 7).

No primeiro experimento foi usada uma ontologia genômica (“The Gene Ontology”) como fonte de conhecimento. A Gene Ontology é uma ontologia criada e mantida pela comunidade de bioinformática. Seu objetivo é unificar, entre as diferentes espécies, a representação dos genes e seus produtos gênicos. Por consequência, essa ontologia funciona como um vocabulário controlado no domínio da genômica.

Tabela 7 - Principais esquemas terminológicos da área biomédica

Esquema Terminológico	Qde de Termos	Disponível em
Gene Ontology (GO)	24.730	http://www.geneontology.org
Medical Subject Headings (MeSH)	97.000	http://www.nlm.nih.gov/pubs/factsheets/mesh.html
SNOMEDCT	930.655	http://www.snomed.org
UMLS	2.000.000+	http://www.nlm.nih.gov/pubs/factsheets/umls.html
MTH	94.450	http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

5.4 Expansão manual das consultas

Como os algoritmos de expansão automática de consultas apresentam um desempenho inferior à expansão manual, tanto no primeiro quanto no segundo experimento, optamos por expandir manualmente as consultas. Isso faz com que os resultados obtidos neste trabalho representem um limite superior para a técnica de expansão de consultas baseada em ontologias de domínio. Dada uma ontologia de

domínio que cubra o domínio do corpus, utilizam-se os conceitos e relações presentes na ontologia para expandir os termos inseridos pelo usuário do sistema em uma consulta, a fim de obter melhores resultados de cobertura e precisão na recuperação de informações.

Cada termo da consulta foi recuperado na Gene Ontology e foram adicionados os termos relacionados, tais como sinônimos, termos genéricos e termos específicos. Este processo foi repetido para todas as cinqüenta consultas da coleção de testes TREC Genômica 2004 resultando em um novo arquivo de consultas de acordo com o formato mostrado na Figura 18.

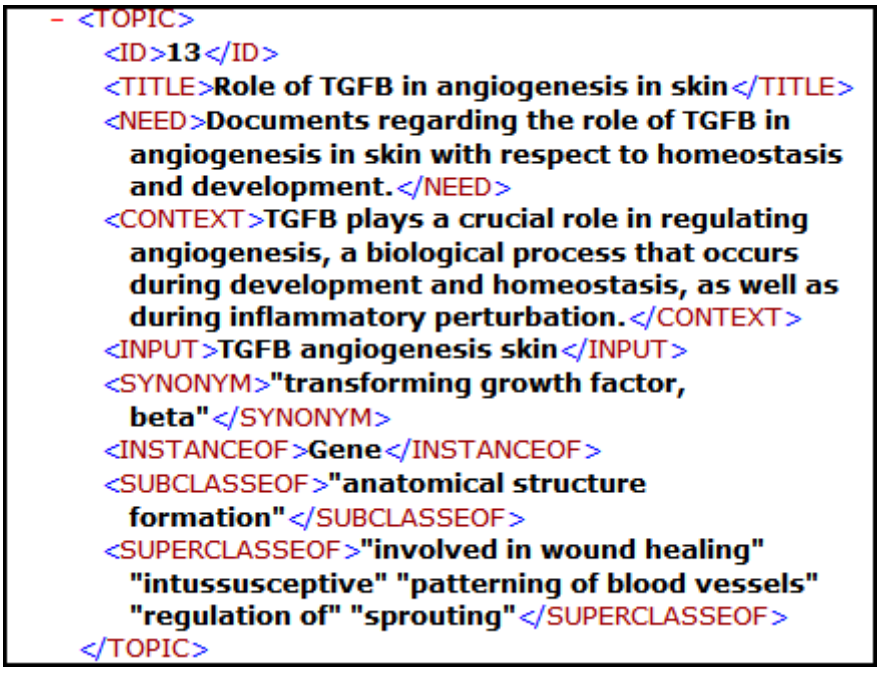

Figura 18 - Exemplo de uma consulta expandida do primeiro experimento

O primeiro resultado empírico foi que a Gene Ontology não é suficiente para expressar as necessidades de informação do dia a dia dos biólogos. Muitas das palavras chaves encontradas nas consultas da TREC Genômica não estão presentes nesta ontologia. Para medir a completude do conhecimento de domínio disponível quando comparada com nossa coleção de testes, foi calculado o número de consultas que puderam ser expandidas em relação ao número total de consultas. Os resultados estão sumarizados na Tabela 8.

Tabela 8 - Estatísticas de expansão usando apenas a GO

Expansão por	Percentual de consultas	Número de consultas
Sinônimo	26 %	13
Instância_de	26 %	13
SubClasse-de	20 %	10
SuperClasse-de	14 %	7
Qualquer Relação	42 %	21
Nenhuma	58 %	29

Devido à baixa completude do conhecimento do domínio quando usada apenas a Gene Ontology como base de conhecimento, foi repetida a fase de expansão em um segundo experimento, usando um conjunto de ontologias, tesauros e taxonomias da área biomédica. Quando um termo existia em mais de um esquema terminológico foi

arbitrada a seguinte regra de precedência: primeiro procura-se na Gene Ontology se encontrado o termo usa-se esta ontologia para expansão, senão procura-se o termo na MeSH e assim sucessivamente na SNOMEDCT, UMLS, MTH. Utilizando-se esse conjunto de esquemas terminológicos, foi possível aumentar significativamente o percentual de consultas expandidas conforme sumarizado na tabela 9.

Tabela 9 - Estatísticas de expansão de consultas usando um conjunto de ontologias

Expansão por	Percentual de Consultas	Número de consultas
Sinônimo	40%	20
Generalização	56%	28
Especialização	42%	21
doença associada	10%	5
parte de	2%	1
Componentes	18%	9
alguma expansão	88%	44
nenhuma expansão	12%	6

5.5 Indexação dos documentos da coleção de testes

Após a escolha do mecanismo de recuperação de informação, o próximo passo foi indexar os documentos da coleção de testes. Cada documento deve ser analisado para que sejam extraídas as principais seções dos documentos tais como título,

autores, *abstract* e etc. Após identificadas as seções de um documento, é gerado o índice invertido com as palavras presentes em cada seção.

A indexação da base TREC Genômica foi feita criando-se um parser, o qual é responsável por identificar as seções dos documentos; um iterador responsável por percorrer os documentos da coleção de testes e por esconder a estrutura do acervo do indexador; e um indexador, responsável por gerar o índice invertido no Lucene. Essas classes, foram geradas na linguagem Java versão 1.5. O trecho de código apresentado no anexo I representa a parte principal do algoritmo de indexação.

5.6 Aplicação de avaliação

Uma vez indexados os documentos da coleção de testes, o próximo passo foi construir uma aplicação que avalieasse a técnica estudada. O papel desta aplicação é facilitar o armazenamento dos resultados das diferentes configurações em um formato que possibilite gerar as estatísticas de cada execução. Esta aplicação deve possibilitar a execução com diferentes parametrizações da técnica proposta. Além disso, a aplicação de avaliação deve permitir a execução com a técnica proposta “desligada”. Essa execução é importante para fins de comparação.

Para realizar nosso experimento, implementamos um módulo de avaliação para a base TREC Genômica. Este módulo foi desenvolvido em Java versão 1.5. O processamento é controlado pela classe Run. Esta classe é responsável pelo controle de fluxo principal deste módulo. Sua tarefa mais importante é usar os termos

relacionados aos termos inseridos pelo usuário (definidos previamente de forma manual no arquivo XML) e usá-los para expandir a busca original. Inicialmente são definidos os parâmetros da execução que são descritos a seguir:

- **Name**: Nome escolhido para esta execução. Este parâmetro é obrigatório.
- **File**: Caminho para o arquivo que contém as consultas. Este parâmetro é obrigatório.
- **Index**: Caminho para o arquivo de índice do Lucene. Este parâmetro é opcional e se estiver ausente é considerado como caminho o subdiretório index.
- **Result**: Caminho onde deve ser armazenado o arquivo com os resultados. Este parâmetro é opcional e se estiver ausente é considerado como caminho o subdiretório result.
- **Operator**: Define o comportamento booleano do motor de busca. Se valor for “and” apenas serão retornados os documentos que contenham todos os termos da busca. Se o valor for “or” serão retornados os documentos que contenham qualquer um dos termos de busca. Este parâmetro é opcional e caso não seja informado, considera-se o valor “or” como padrão.
- **Syno**: Habilita a expansão por sinônimos com o peso informado. Este parâmetro é opcional, caso não informado, não serão feitas as expansões com sinônimos.

- **Subc**: Habilita a expansão por subclasse_de com o peso informado. Este parâmetro é opcional, caso não informado, não serão feitas as expansões com hiperônimos.

- **Supc**: Habilita a expansão por superclasse_de com o peso informado. Este parâmetro é opcional, caso não informado, não serão feitas as expansões com hipônimos.

- **Dise**: Habilita a expansão por doença_associada com o peso informado. Este parâmetro é opcional, caso não informado, não serão feitas as expansões com as doenças associadas.

- **Cont**: Habilita a expansão por contém (inverso do parte_de) com o peso informado. Este parâmetro é opcional e, caso não informado, não serão feitas as expansões com as partes.

- **Part**: Habilita a expansão por parte_de com o peso informado. Este parâmetro é opcional, caso não informado, não serão feitas as expansões com o elemento agrupador.

Uma vez configuradas as variáveis do sistema de execução, inicia-se o processamento de cada consulta. Para extrair cada consulta da coleção TREC Genômica 2004 foi criado um parser XML (classe denominada TopicsXMLParser) baseado na API SAX (classe denominada TopicsSAXHandler). Uma vez identificada a

consulta, essa consulta passa para a fase de expansão, nos quais os relacionamentos e pesos são configurados de acordo com as variáveis de inicialização. Em seguida, a consulta expandida é enviada ao mecanismo de recuperação de informação (Lucene) que retorna o conjunto de documentos resultado em ordem de relevância. Enfim, este resultado é gravado em um arquivo no formato de entrada da aplicação de geração de dados estatísticos (classe denominada TrecEvalResultsFileWriter). Esse processo se repete para cada consulta da coleção de testes. O trecho de código incluído no Anexo II implementa o algoritmo mencionado.

5.7 Geração de medições

Uma vez tendo sido gravado o resultado da busca, o próximo passo é gerar as medições de cobertura e precisão. Para gerar essas medidas usamos o programa trec_eval (“Text REtrieval Conference (TREC) trec_eval”). O trec_eval é o programa oficial de avaliação do NIST para experimentos em recuperação textual. Para usá-lo é necessário um arquivo de relevância (qrels) e um arquivo de resultado da execução. O arquivo de resultado da execução deve respeitar o seguinte formato (separado por espaços):

CONSULTA_ID – ITERAÇÃO – DOCUMENTO_ID – RANK – SIMILARIDADE – EXECUCAO_ID

Onde **CONSULTA_ID** é o número identificador da consulta; **ITERAÇÃO** é a constante 0, que é ignorado mas requerido pelo trec_eval; **DOCUMENTO_ID** é o código que identifica o documento, em nosso experimento é o PubMedID (PMID); **RANK** é um

inteiro entre 0 e 1000 que é ignorado mas requirido pelo trec_eval; **SIMILARIDADE** é um valor de ponto flutuante (*float*) que representa a avaliação de similiaridade calculada pelo mecanismo de recuperação de informação. É assumido que os documentos são retornados em ordem decrescente de similaridade. **EXECUCAO_ID** é uma String que representa o nome dado à execução que está sendo avaliada. A Figura 19 mostra o resultado de uma medição feita pelo programa trec_eval.

```
Total number of documents over all queries
    Retrieved:     1000
    Relevant:        32
    Rel_ret:         13
Interpolated Recall - Precision Averages:
    at 0.00       1.0000
    at 0.10       0.0806
    at 0.20       0.0351
    at 0.30       0.0351
    at 0.40       0.0250
    at 0.50       0.0000
    at 0.60       0.0000
    at 0.70       0.0000
    at 0.80       0.0000
    at 0.90       0.0000
    at 1.00       0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.0945
Precision:
  At    5 docs:   0.6000
  At   10 docs:   0.3000
  At   15 docs:   0.2000
  At   20 docs:   0.1500
  At   30 docs:   0.1000
  At  100 docs:   0.0500
  At  200 docs:   0.0300
```

Figura 19 - Exemplo de medição feita pelo programa Trec_Eval

5.8 Geração de curva de cobertura e precisão

Uma vez gerados os valores de precisão em função da cobertura pode-se gerar o gráfico de cobertura e precisão. Qualquer programa de geração de gráficos pode ser

utilizado. Neste experimento foi utilizado Excel versão 2007. Uma vez criado o ambiente para avaliação, o próximo passo é realizar o experimento. No próximo capítulo são apresentados os passos realizados.

6. Realização do experimento

Uma vez montado o ambiente de avaliação, o próximo passo é realizar o experimento. Neste capítulo é apresentado como foram realizados os dois experimentos apresentados neste trabalho. A necessidade de se fazer um segundo experimento é decorrente de questões levantadas após a realização do primeiro experimento, onde foi constatado que algumas variáveis podiam colocar em dúvida a generalização dos primeiros resultados.

6.1 Tratamento do número de termos das consultas

Para se aproximar da realidade, as consultas presentes em uma coleção de testes devem refletir necessidades de informações do cotidiano de um especialista do domínio. Além disso, os termos que o usuário passa como palavras-chave para o sistema de recuperação de informação devem estar refletidos na coleção de testes.

Considere a Figura 16 como exemplo. Um biólogo deseja saber “Qual o papel do TGFB na angiogênese da pele?”. Porém, como esse biólogo buscaria essa informação em um mecanismo de recuperação de informação? O número de termos das consultas da coleção de testes reflete a forma como os usuários de sistemas de recuperação de informação usam estes mecanismos?

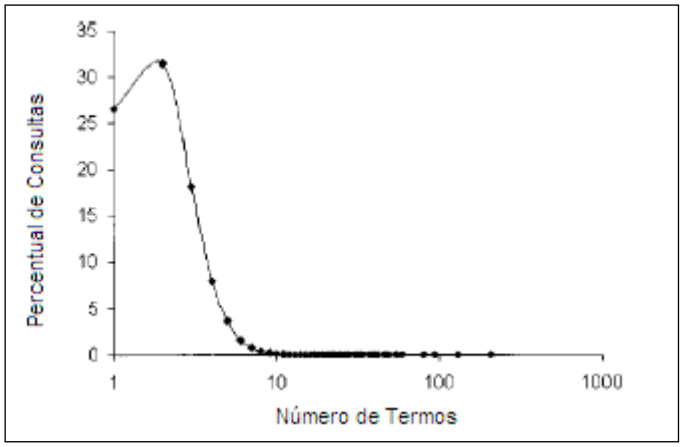

Figura 20 - O público e suas consultas (Spink 2001)

A forma como as pessoas usam os mecanismos de recuperação de informação foi analisada em detalhes por Spink (SPINK ET AL., 2001). Em seu trabalho, foram analisadas aproximadamente um milhão de consultas submetidas ao mecanismo de busca Excite. Os resultados obtidos mostram que aproximadamente 80% das consultas possuem até três termos. A Figura 20 mostra a distribuição do número de termos nas consultas. Como a mediana desta curva de distribuição se aproxima de dois (aproximadamente 60% das consultas têm até dois termos), este valor é considerado o número “mágico” de termos.

Porém, ao analisarmos a distribuição do número de termos da coleção TREC Genômica, verifica-se que esta distribuição não se aproxima do número de termos usados geralmente nas consultas. Após a contagem do número de termos dos campos

título, necessidade e contexto conforme demonstrado na Tabela 10, chega-se à conclusão que apesar do campo título ser o campo que tem a distribuição do número de termos mais próxima da distribuição do trabalho de Spink, ainda assim essa distribuição está com uma curva muito diferente da verificada pelo uso comum de um mecanismo de recuperação de informação.

Tabela 10- Quantidade de termos nas consultas

	Campo TREC				
Qtd Termos	Titulo	Necessidade	Contexto	Spink	Input
1	16,00%	0,00%	0,00%	28,89%	6,00%
2	6,00%	0,00%	0,00%	35,00%	26,00%
3	12,00%	0,00%	0,00%	20,22%	38,00%
4	14,00%	2,00%	0,00%	8,89%	14,00%
5	14,00%	2,00%	0,00%	4,44%	14,00%
6	4,00%	4,00%	0,00%	0,78%	2,00%
7+	34,00%	92,00%	100,00%	1,78%	0,00%
Mediana	**4**	**7+**	**7+**	**2**	**2,5**

Para amenizar esta divergência foi criado um novo campo nas consultas chamado de *input*, como mostra a Figura 21. Neste campo foram adicionados os termos

que um usuário digitaria em um mecanismo de recuperação de informação para atender à sua necessidade de informação.

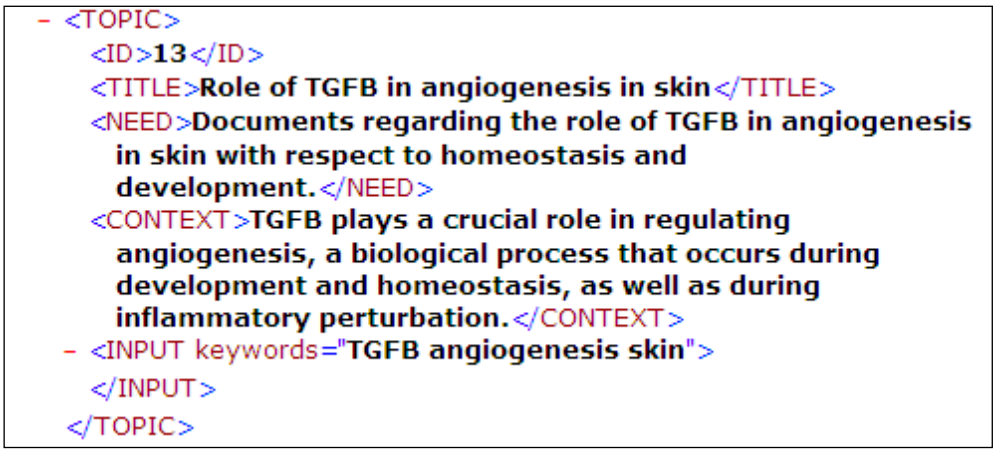

Figura 21 - Consulta acrescida de campo com palavras-chave

O resultado dessa aproximação pode ser confirmado pela curva de distribuição na Figura 22. A curva título que representa a distribuição do número de termos da consulta do elemento título da coleção TREC Genômica 2004. Após o trabalho de criação do campo de palavras-chave, a distribuição do número de termos da consulta (curva input) ficou muito mais próxima a curva publicada no trabalho de Spink representada pela curva Spink.

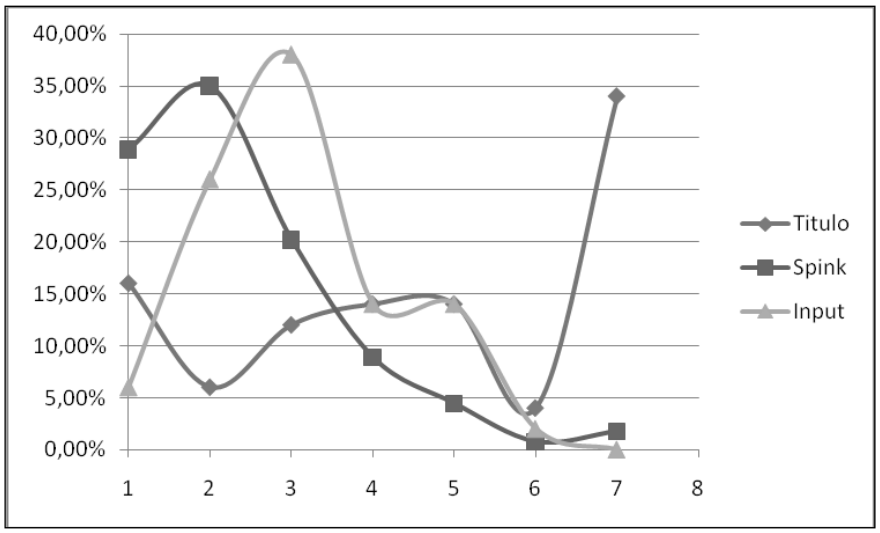

Figura 22 - Distribuição da quantidade de termos após criação do campo input

6.2 Execuções Sucessivas

Passada a etapa mais trabalhosa de expansão manual das consultas, pôde-se iniciar a experimentação. O procedimento de execução do experimento foi projetado com o objetivo de facilitar a análise dos resultados obtidos. De forma a isolar o impacto que cada relação léxico-semântica tem na expansão de consultas foi feita a medição destas relações uma a uma. Além disso, cada relação tinha seu peso alterado entre as execuções variando de um para meio, para um terço e para um quarto. Isso significa

que, no primeiro caso, os termos expandidos têm o mesmo peso que as palavras chaves inseridas pelo usuário e no último caso os termos inseridos pelo usuário têm peso quatro vezes maior que os termos expandidos.

Dessa forma, o procedimento de execução do experimento exige quatro rodadas para cada relação semântica. No primeiro experimento eram 4 relações sendo avaliadas o que totalizou 16 execuções. No segundo experimento eram 6 relações sendo avaliadas o que totalizou 24 execuções.

6.3 Classificação de Relevância

Conforme discutido na seção 5.2, a coleção de testes TREC Genômica classifica os documentos para cada consulta como não relevantes, possivelmente relevantes e definitivamente relevantes. No momento de medição de precisão os trabalhos que utilizam a TREC Genômica consideram como documentos relevantes tanto os documentos definitivamente relevantes quanto os documentos possivelmente relevantes. Apesar do recorrente uso desta metodologia de medição em trabalhos relacionados, optamos no segundo experimento por medir separadamente da forma tradicional e analisar também considerando somente os definitivamente relevantes. Portanto, foram realizadas 2 medições para cada execução.

Como foram realizadas no segundo experimento 24 execuções, gerou-se um total de 48 medições no segundo experimento. As primeiras 24 medições levaram em consideração a mesma classificação de relevância dos trabalhos anteriores. Já a

segunda metade das medições classificou como relevantes apenas os documentos da TREC Genômica que foram considerados como definitivamente relevantes pelos especialistas do domínio. Por isso os ganhos ou prejuízos na técnica de expansão semântica de consultas quando levado em consideração somente os documentos definitivamente relevantes têm um peso maior nas conclusões do que o formato padrão de medição. No próximo capítulo é apresentada uma análise dos resultados obtidos para cada tipo de relação semântica.

7. Análise dos resultados

Neste capítulo, os resultados dos experimentos são analisados, discutindo-se cada relação semântica separadamente. Inicialmente são apresentadas as expansões por sinônimos. Em seguida, são apresentadas as expansões por termo genérico. A próxima expansão analisada é a expansão por termo específico. Depois é analisada a expansão específica de domínio doença associada. Por fim são analisados os dois lados da relação parte_de: a expansão com os componentes e a expansão pelo termo agregador.

7.1 Sinônimos

As expansões baseadas em relações de sinonímia de maneira geral apresentaram bons resultados. No primeiro experimento onde foi usada apenas a Gene Ontology como fonte de conhecimento, 13 consultas possuíam algum termo com relações de sinonímia, como mostra a Tabela 8. Já no segundo experimento, quando foi utilizado um conjunto de bases de conhecimento, 20 consultas possuíam algum termo com relações de sinonímia.

Os resultados obtidos nos dois experimentos mostraram uma melhora significativa em relação à execução sem a expansão de consultas. No primeiro experimento os ganhos foram de aproximadamente 26% na mediana da precisão média de 11 pontos, sendo seu ponto de máximo atingido usando pesos equivalentes entre os

termos expandidos pelo usuário e os termos adicionados automaticamente pelo sistema, quando foi atingida uma melhora de 27,67% como mostra a Tabela 11. Já no segundo experimento, os ganhos obtidos na mediana da precisão média de 11 pontos ficaram em torno de 13%. O ponto de máximo ganho foi obtido quando os termos inseridos pelo usuário tinham um peso 5 vezes maior que o peso dos termos inseridos automaticamente pelo sistema. Nesta configuração o resultado obtido foi uma melhora de 14,25%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou um ganho de aproximadamente 8%, onde o seu ponto de ótimo foi atingido quando os pesos dos termos expandidos eram 0,2 ou 0,33 vezes o peso dos termos inseridos pelo usuário. Em outras palavras, os termos inseridos pelo usuário têm um peso de 3 a 5 vezes maior que os termos inseridos automaticamente pelo sistema.

Fizemos ainda uma terceira medição onde eram considerados relevantes apenas os documentos da coleção de testes TREC Genômica que são considerados como definitivamente relevantes para uma determinada consulta. Desta forma, os documentos classificados no método de *polling* como possivelmente relevantes não são considerados como relevante nesta medição. Esta medição não serve como comparativo com trabalhos anteriores, pois esta medida não é utilizada nos outros trabalhos. Apenas a fizemos para verificar a consistência nos resultados obtidos. Os resultados obtidos quando considerados apenas os definitivamente relevantes foram consistentes com os resultados obtidos da forma tradicional de medição, dado que ambos apresentaram melhoras tanto na mediana da precisão média de 11 pontos

quanto na precisão nos primeiros 10 documentos. O ganho na mediana da precisão média de 11 pontos foi de aproximadamente 2%, sendo o seu ganho máximo de 7,14% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram 0,2. A medição de precisão nos primeiros 10 documentos mostrou uma melhora de aproximadamente 24%, onde o seu ponto de máximo foi de 30,43%, atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 0,33 a 0,2.

Tabela 11 - Resultados obtidos com a relação de sinonímia

	1º. Experimento				2º. Experimento				2º. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,2842	0,2814	0,2784	0,2781	0,2311	0,2377	0,2395	0,2406	0,0615	0,0667	0,0689	0,0705
Baseline	0,2226				0,2106				0,0658			
Ganho	**27,67%**	**26,42%**	**25,07%**	**24,93%**	9,73%	12,87%	13,72%	**14,25%**	-6,53%	1,37%	4,71%	**7,14%**
P@10					0,5000	0,5150	0,5350	0,5350	0,1300	0,1400	0,1500	0,1500
Baseline					0,4850				0,1150			
Ganho					3,09%	6,19%	10,31%	**10,31%**	13,04%	21,74%	30,43%	**30,43%**

Como as médias não são suficientes para dar uma visão minuciosa do que está acontecendo em cada consulta fizemos uma análise individual da precisão nos 10 primeiros documentos de cada consulta. O resultado mostrou que o número de consultas que obtiveram ganhos foi maior do que as que tiveram degradação como resume a Tabela 12. Mais importante do que isso, é o fato de que as consultas onde o resultado obtido foi uma melhora (como pode ser observado nas consultas 8,13 e 37 da Figura 23), o ganho obtido foi muito maior do que as consultas onde o resultado obtido foi uma degradação na precisão como pode ser observado nas consultas 2,3 e 49.

Tabela 12 - Resultado em cada consulta nas relações de sinonímia

Melhorou	5
Piorou	3
Inalterado	12

A consulta 37 obteve o melhor ganho passando de uma P@10 de 20% para uma P@10 de 100% o que representa uma melhora de 400% na P@10. Nesta consulta o pesquisador usou um acrônimo que representa o nome de uma enzima PAM e o sistema expandiu este acrônimo para a sua forma descritiva *"peptidylglycine alpha-amidating monooxygenase"*. Este mesmo cenário de ganho significativo por expansão de acrônimo se repetiu na consulta 13 onde o termo TGFB foi expandido para a sua forma descritiva "*transforming growth factor beta*" resultando em um ganho de 150% na P@10.

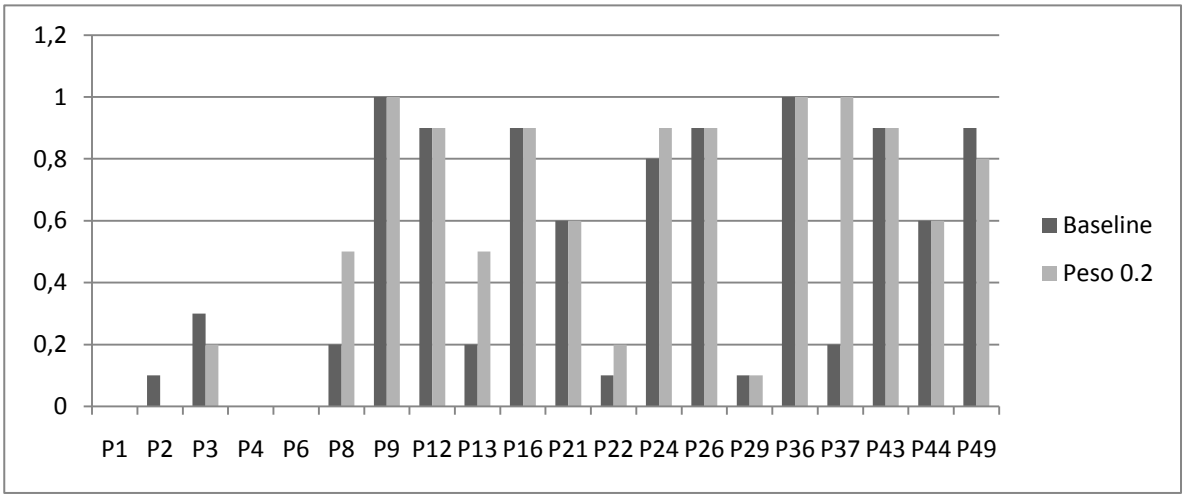

Figura 23 - Resultado da expansão de sinonímia em cada consulta

As consultas 2 e 3 apresentaram uma pequena degradação quando foi expandido o termo “mouse” pelo termo “mice” e vice-versa. A relação que existe entre estes dois termos é uma relação de plural (*mice* é o plural de *mouse*). Existiu também uma pequena degradação na consulta 49 ao expandirmos o termo “*Glyphosate*” por “*Herbicide*”, caso que também não aconteceria em uma ontologia de forte compromisso ontológico (GUIZZARDI, 2005) dado que “*Herbicide*” é um termo mais genérico que o termo “*Glyphosate”*, o que descaracteriza a relação de sinonímia.

Para analisarmos a precisão da expansão usando relações de sinonímia em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja com pesos para os termos expandidos em 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão está sempre acima da curva sem expansão como pode ser visto na Figura 24.

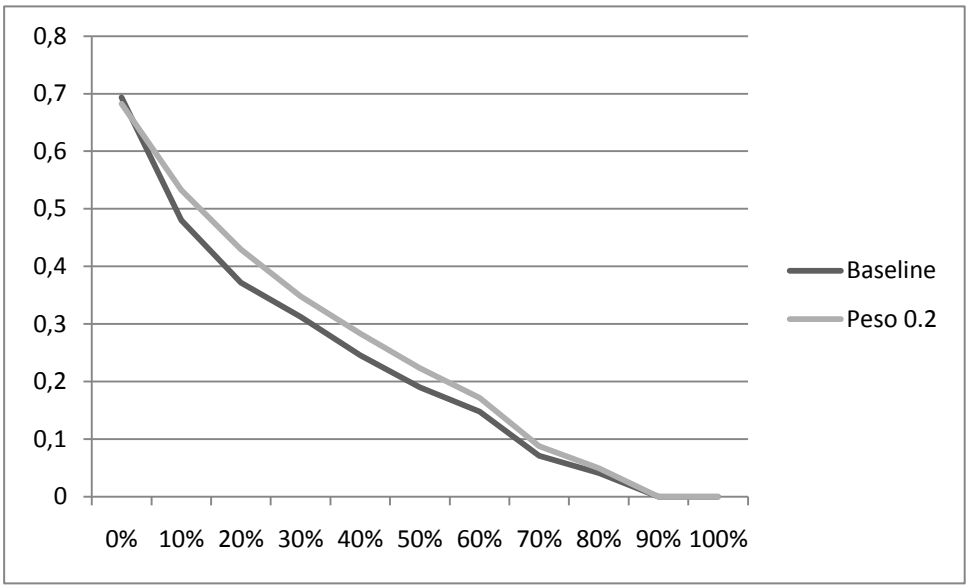

Figura 24 - Gráfico de cobertura x precisão da expansão por sinonímia

7.2 Expansão por termo genérico

As expansões baseadas em relações de subclasse de maneira geral apresentaram resultados insatisfatórios. No primeiro experimento 10 consultas possuíam algum termo com relações de generalização, como mostra a Tabela 8. Já no segundo experimento, 28 consultas possuíam algum termo com relações de generalização como mostra a Tabela 9.

Os resultados obtidos nos dois experimentos mostraram uma degradação significativa. No primeiro experimento as degradações foram de aproximadamente 23% na mediana da precisão média de 11 pontos, sendo seu ponto de melhor desempenho atingido usando pesos 0,2 para os termos adicionados automaticamente pelo sistema,

quando o resultado foi uma degradação de 22,10% como mostra a Tabela 13. Já no segundo experimento, a degradação na mediana da precisão média de 11 pontos ficou em torno de 20%. O ponto de melhor resultado foi obtido quando os termos inseridos pelo sistema tinham peso 0,2. Nesta configuração o resultado obtido foi uma degradação de 13,54%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou uma degradação de aproximadamente 15%, onde o seu ponto de melhor resultado foi atingido quando os pesos dos termos expandidos eram 0,33 quando foi obtida uma degradação de 1,91%.

Os resultados obtidos quando considerados apenas os definitivamente relevantes foram consistentes com os resultados obtidos da forma tradicional de medição, dado que ambos apresentaram degradação tanto na mediana da precisão média de 11 pontos quanto na precisão nos primeiros 10 documentos. A degradação na mediana da precisão média de 11 pontos foi de aproximadamente 10%, sendo o seu melhor resultado uma degradação de 8,58% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram 0,5. A medição de precisão nos primeiros 10 documentos mostrou uma degradação de aproximadamente 18%, onde o seu ponto de melhor resultado foi uma degradação de 16,01% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 0,5 a 0,2.

Tabela 13 - Resultados das expansões por termo genérico

	1o. Experimento				2o. Experimento				2o. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,1654	0,1716	0,1722	0,1734	0,0931	0,1056	0,1079	0,1117	0,0377	0,0394	0,0389	0,0386
Baseline	0,2226				0,1292				0,0431			
Ganho	**-25,70%**	**-22,91%**	**-22,64%**	**-22,10%**	-27,94%	-18,27%	-16,49%	**-13,54%**	-12,53%	-8,58%	-9,74%	**-10,44%**
P@10					0,2750	0,3000	0,3643	0,3143	0,0679	0,0750	0,0750	0,0750
Baseline					0,3714				0,0893			
Ganho					-25,96%	-19,22%	-1,91%	**-15,37%**	-23,96%	-16,01%	-16,01%	**-16,01%**

Ao analisar individualmente a precisão nos 10 primeiros documentos de cada consulta, nota-se que o número de consultas que obtiveram degradação foi muito maior do que as que apresentaram melhorias, como resume a Tabela 13. Mais importante do que isso é o fato de que nas consultas onde o resultado obtido foi uma degradação, a degradação obtida (como pode ser observado nas consultas 3,15 e 20 da Figura 25) é muito maior do que nas consultas onde o resultado obtido foi uma melhora na precisão, como pode ser observado na consulta 4.

Tabela 14 - Resultado da expansão de cada consulta por termo genérico

Melhorou	1
Piorou	3
Inalterado	24

A consulta 4 obteve o melhor ganho passando de uma P@10 de 0% para uma P@10 de 10% o que representa uma pequena melhora na P@10. Nesta consulta o

termo “*gene expression*” foi expandido pelo termo genérico “*macromolecule metabolic* process” relação presente na ontologia de processos biológicos da Gene Ontology.

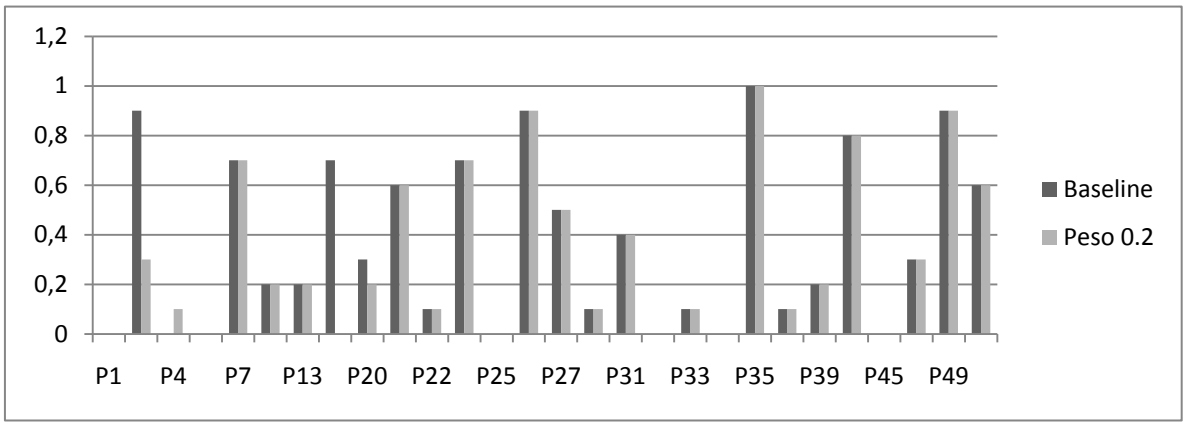

Figura 25- Resultado da expansão por termo genérico em cada consulta

A consulta 15 foi a que obteve a maior degradação passando de uma P@10 de 70% para uma P@10 de 0% após a expansão. Nesta consulta dois termos foram expandidos para os seus respectivos termos genéricos. O termo “*ATPase*” foi expandindo com o termo “*nucleoside-triphosphatase activity*”, e o termo “*apoptosis*” foi expandido pelo termo “*programmed cell death*”.

Para analisarmos a precisão da expansão usando termo genérico em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja com pesos para os termos expandidos em 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão está sempre abaixo da curva sem expansão, como pode ser visto na Figura 26.

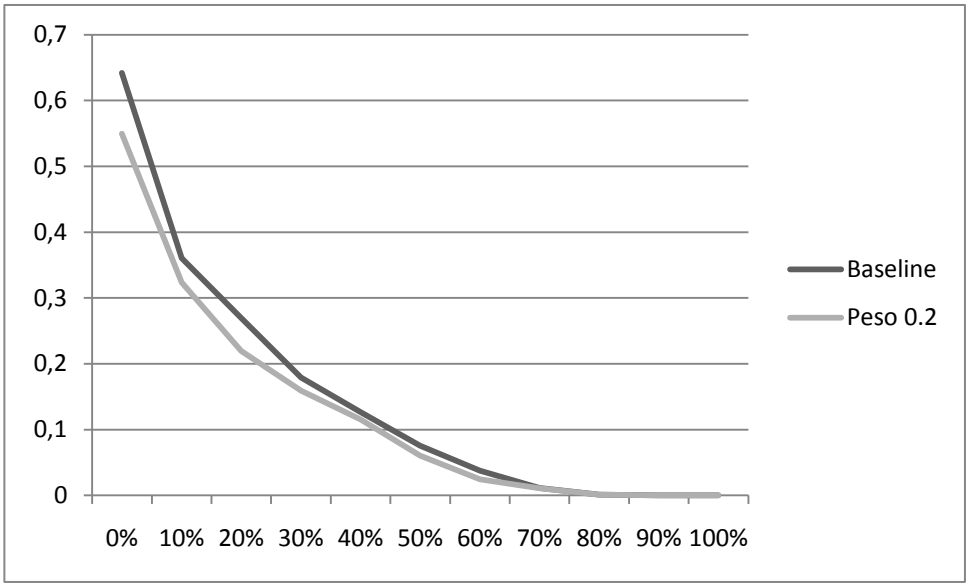

Figura 26 - Curva de cobertura vs. precisão da expansão por termo genérico

7.3 Expansão por termo específico

As expansões baseadas em relações de superclasse de maneira geral apresentaram resultados insatisfatórios. No primeiro experimento, 7 consultas possuíam algum termo com relações de especialização, como mostra a Tabela 8, Já no segundo experimento, 21 consultas possuíam algum termo com relações de especialização como mostra a Tabela 9.

Os resultados obtidos nos dois experimentos mostraram uma degradação significativa. No primeiro experimento as degradações foram de aproximadamente 28% na mediana da precisão média de 11 pontos, sendo seu ponto de melhor desempenho

atingido usando pesos 0,2 para os termos adicionados automaticamente pelo sistema, quando o resultado foi uma degradação de 26,42%, como mostra a

Tabela 15. Já no segundo experimento, a degradação na mediana da precisão média de 11 pontos ficou em torno de 50%. O ponto de melhor resultado foi obtido quando os termos inseridos pelo sistema tinham peso 0,2. Nesta configuração, o resultado obtido foi uma degradação de 39,11%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou uma degradação de aproximadamente 40%, onde o seu ponto de melhor resultado foi atingido quando os pesos dos termos expandidos eram 0,2, quando foi obtida uma degradação de 39,11%.

Os resultados obtidos quando considerados apenas os documentos definitivamente relevantes foram consistentes com os resultados obtidos da forma tradicional de medição, dado que ambos apresentaram degradação tanto na mediana da precisão média de 11 pontos quanto na precisão nos primeiros 10 documentos. A degradação na mediana da precisão média de 11 pontos foi de aproximadamente 45%, sendo o seu melhor resultado uma degradação de 39,81%, atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 0,33 a 0,2. A medição de precisão nos primeiros 10 documentos mostrou uma degradação de aproximadamente 20%, onde o seu ponto de melhor resultado foi uma degradação de 19,96% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 1 a 0,2.

Tabela 15 - Resultados da expansão por termo específico

	1o. Experimento				2o. Experimento				2o. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,1548	0,1607	0,1626	0,1638	0,0421	0,0556	0,0634	0,0735	0,0175	0,0227	0,0254	0,0254
Baseline	0,2226				0,1207				0,0422			
Ganho	**-30,46%**	**-27,81%**	**-26,95%**	**-26,42%**	-65,12%	-53,94%	-47,47%	**-39,11%**	-58,53%	-46,21%	-39,81%	**-39,81%**
P@10					0,1810	0,1905	0,2048	0,2429	0,0762	0,0762	0,0762	0,0762
Baseline					0,3381				0,0952			
Ganho					-46,47%	-43,66%	-39,43%	**-28,16%**	-19,96%	-19,96%	-19,96%	**-19,96%**

Ao analisar individualmente a precisão nos 10 primeiros documentos de cada consulta, nota-se que o número de consultas que obtiveram degradação foi maior do que as que apresentaram melhorias, como resume a Tabela 16. O ponto mais importante é o fato de que as consultas que obtiveram variação significativa no resultado foi muito mais frequente na degradação, as quais podem ser observadas nas consultas 3,15,27,31 e 50 (ver Figura 27). Apenas foi observada uma melhora significativa na consulta 8.

Tabela 16 - Resultado da expansão de cada consulta por termo específico

Melhorou	5
Piorou	8
Inalterado	9

A consulta 8 obteve o melhor ganho passando de uma P@10 de 20% para uma P@10 de 50%, o que representa uma melhora significativa na P@10. Nesta consulta, o termo "*DNA repair*" foi expandido pelos termos específicos "*viral DNA repair, bypass DNA synthesis, non-photoreactive DNA repair, mitochondrial DNA repair, single strand break repair, mismatch repair, postreplication repair, double-strand break repair, DNA dealkylation, DNA replication proofreading, error-prone DNA repair, pyrimidine dimer repair, recombinational repair, non-recombinational repair, base-excision repair, nucleotide-excision repair*". O resultado, nesse caso, pode estar mais relacionado com a repetição dos termos "*repair*" do que com os sintagmas nominais propriamente ditos. Ou seja, este resultado foi um efeito colateral do nome dos filhos e não resultado da técnica.

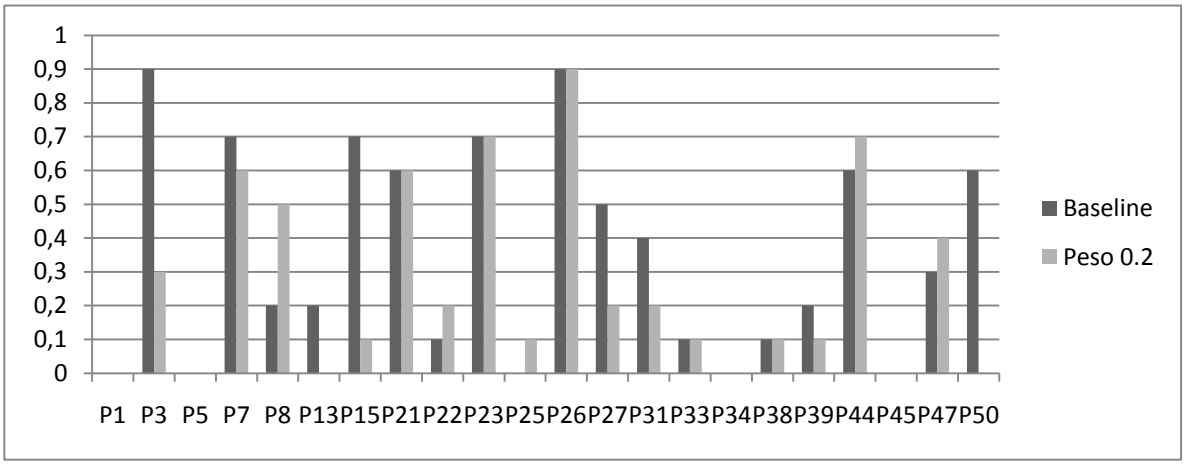

Figura 27 - Resultado da expansão por termo específico em cada consulta

A consulta 50 foi a que obteve a maior degradação, passando de uma P@10 de 60% para uma P@10 de 0% após a expansão. Nesta consulta dois termos foram expandidos para os seus respectivos termos específicos. O termo “*protein expression*” foi expandindo com os termos “*Protein Overexpression Translation*”, já o termo “*E. coli*” foi expandido com os termos “*Shiga-Toxigenic Escherichia coli, Enteropathogenic Escherichia coli, Enterotoxigenic Escherichia coli, e Escherichia coli K12*”.

Para analisarmos a precisão da expansão usando termo específico em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja, com pesos para os termos expandidos em 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão está sempre abaixo da curva sem expansão, como pode ser visto na Figura 28.

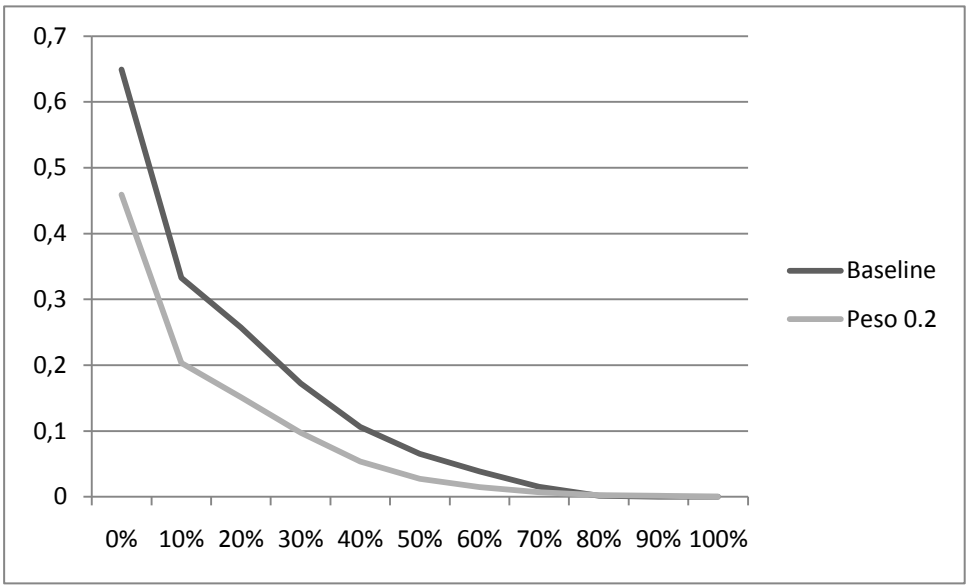

Figura 28 - Curva de cobertura vs. precisão da expansão por termo específico

7.4 Doença Associada

As expansões baseadas em relações de doença associada mostraram resultados insatisfatórios. Este tipo de relação foi explorado apenas no segundo experimento, onde 5 consultas possuíam algum termo com relações de doença associada como mostra a Tabela 9.

Os resultados obtidos foram inconclusivos devido ao baixo número de consultas, mas indicaram um risco elevado no uso deste tipo de relação. No experimento realizado, a degradação na mediana da precisão média de 11 pontos ficou em torno de 40%. O ponto de melhor resultado foi obtido quando os termos inseridos pelo sistema tinham peso 0,2. Nesta configuração, o resultado obtido foi uma degradação de

28,51%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou uma degradação de aproximadamente 50%, onde o seu ponto de melhor resultado foi atingido quando os pesos dos termos expandidos eram 0,33 a 0,2, quando foi obtida uma degradação de 34,62%.

Os resultados obtidos quando considerados apenas os documentos definitivamente relevantes foram inconsistentes com os resultados obtidos da forma tradicional de medição, dado que apresentaram melhora significativa tanto na mediana da precisão média de 11 pontos quanto na precisão nos primeiros 10 documentos. A melhora na mediana da precisão média de 11 pontos foi de aproximadamente 35%, sendo o seu melhor resultado uma melhora de 43,43% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram 0,2. A medição de precisão nos primeiros 10 documentos mostrou uma melhora de aproximadamente 17%, onde o seu ponto de melhor resultado foi uma melhora de 33,33% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 0,33 a 0,2.

Tabela 17 - Resultados da expansão por doença associada

	2o. Experimento				2o. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,0544	0,1606	0,1888	0,1888	0,0335	0,0466	0,0519	0,0535
Baseline	0,2641				0,0373			
Ganho	-79,40%	-39,19%	-28,51%	**-28,51%**	-10,19%	24,93%	39,14%	**43,43%**
P@10	0,1200	0,2600	0,3400	0,3400	0,0600	0,0600	0,0800	0,0800
Baseline	0,5200				0,0600			
Ganho	-76,92%	-50,00%	-34,62%	**-34,62%**	0,00%	0,00%	33,33%	**33,33%**

Ao analisar individualmente a precisão nos 10 primeiros documentos de cada consulta, nota-se que o número de consultas que obtiveram degradação foi ligeiramente maior do que as que apresentaram melhorias, como resume a Tabela 16. Todavia, a única consulta que obteve uma mudança no resultado de forma significativa, foi a consulta 6 onde ocorreu degradação (ver Figura 29). A consulta 1 que apresentou uma melhora, obteve apenas um ganho marginal.

Tabela 18 - Resultado da expansão por doença associada

Melhorou	1
Piorou	2
Inalterado	2

Na consulta 6, onde houve a maior variação de desempenho, o termo "*FancD2*", o qual representa um gene na Gene Ontology, foi expandido com as seguintes doenças relacionadas a este gene: "*Fanconi anemia, Fanconi pancytopenia e Panmyelopathy Fanconi*". O resultado obtido com essa expansão foi desastroso, passando de um bom resultado sem expansão, onde a P@10 era de 90% para uma P@10 de 0% após a expansão.

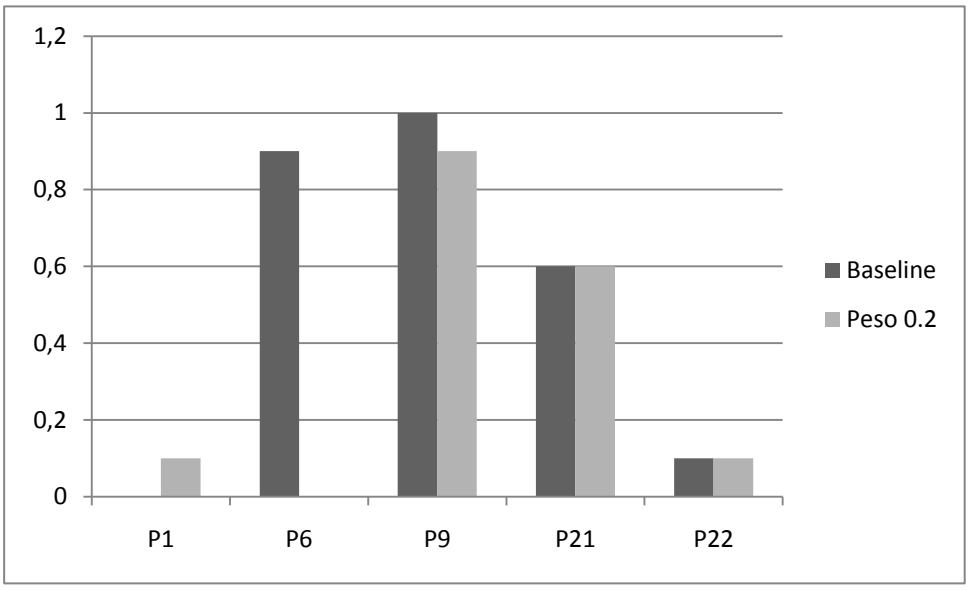

Figura 29 - Resultado da expansão por doença associada em cada consulta

Visando analisar a precisão da expansão usando doença associada em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja com pesos para os termos expandidos em 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão está abaixo da curva sem expansão em baixos níveis de cobertura, porém as curvas se cruzam para níveis altos de cobertura, indicando que a expansão de consultas por doença associada tende a ser uma técnica de aumento de cobertura em detrimento da precisão, como pode ser visto na Figura 29.

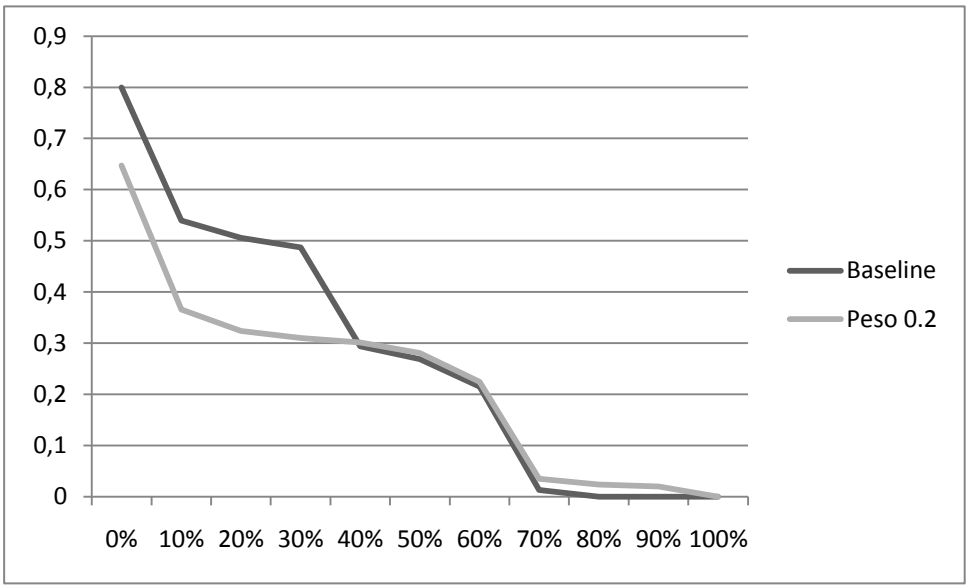

Figura 30 - Curva de cobertura vs. precisão da expansão por doença associada

7.5 Contém

As expansões baseadas em relações de contém, no qual os componentes de um determinado termo são usados na expansão, mostraram resultados insatisfatórios. Este tipo de relação foi explorado apenas no segundo experimento, onde 9 consultas possuíam algum termo com relações do tipo contém como mostra a Tabela 9.

Os resultados obtidos foram ruins no uso deste tipo de relação. No experimento realizado, a degradação na mediana da precisão média de 11 pontos ficou em torno de 13%. O ponto de melhor resultado foi obtido quando os termos inseridos pelo sistema tinham peso 0,2. Nesta configuração, o resultado obtido foi uma degradação de 8,86%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou uma

degradação de aproximadamente 18%, onde o seu ponto de melhor resultado foi atingido quando os pesos dos termos expandidos eram 0,5 a 0,2, quando foi obtida uma degradação de 16,28%.

Os resultados obtidos quando considerados apenas os documentos definitivamente relevantes foram consistentes com os resultados obtidos da forma tradicional de medição, pois ambos apresentaram uma degradação significativa tanto na mediana da precisão média de 11 pontos quanto na precisão nos primeiros 10 documentos. A degradação na mediana da precisão média de 11 pontos foi de aproximadamente 30%, sendo o seu melhor resultado uma degradação de 25,81% atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram 0,2. A medição de precisão nos primeiros 10 documentos mostrou uma degradação de aproximadamente 38%, onde o seu ponto de melhor resultado foi uma degradação de 28,53%, atingido quando os pesos dos termos inseridos automaticamente pelo sistema eram de 0,33 a 0,2.

Tabela 19 - Resultados da expansão por relação todo-parte

	2o. Experimento				2o. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,1302	0,1397	0,1425	0,1461	0,0212	0,024	0,0246	0,0253
Baseline	0,1603				0,0341			
Ganho	-18,78%	-12,85%	-11,10%	**-8,86%**	-37,83%	-29,62%	-27,86%	**-25,81%**
P@10	0,3667	0,4000	0,4000	0,4000	0,0333	0,0556	0,0556	0,0556
Baseline	0,4778				0,0778			
Ganho	-23,25%	-16,28%	-16,28%	**-16,28%**	-57,20%	-28,53%	-28,53%	**-28,53%**

Ao analisar individualmente a precisão nos 10 primeiros documentos de cada consulta, nota-se que o número de consultas que obtiveram degradação foi maior do que as que apresentaram melhorias, como resume a Tabela 20. Apenas a consulta 30 obteve melhora na eficiência do mecanismo. Já em relação à degradação, as consultas 3, 4 e 15 apresentaram piora na eficiência do mecanismo de recuperação. Como pode ser observado na Figura 31, a consulta 3 mostrou uma grande degradação. Por outro lado, a consulta 30 mostrou uma razoável melhora.

Tabela 20 - Resultado da expansão por relação todo-parte

Melhorou	1
Piorou	3
Inalterado	5

Na consulta 30, houve o único ganho de desempenho onde o sintagma nominal "*Nkx gene family members*", o qual representa uma família de genes, foi expandido usando a base de conhecimento UniprotKB com os seguintes membros desta família de genes: "*Nkx-2.1, Nkx-2.2, Nkx-2.3, Nkx-2.4, Nkx-2.5, Nkx-2.6, Nkx-2.7 e Nkx-2.8*". O resultado obtido com essa expansão foi satisfatório, passando de um fraco resultado sem expansão, onde a P@10 era de 30%, para uma P@10 de 50% após expansão.

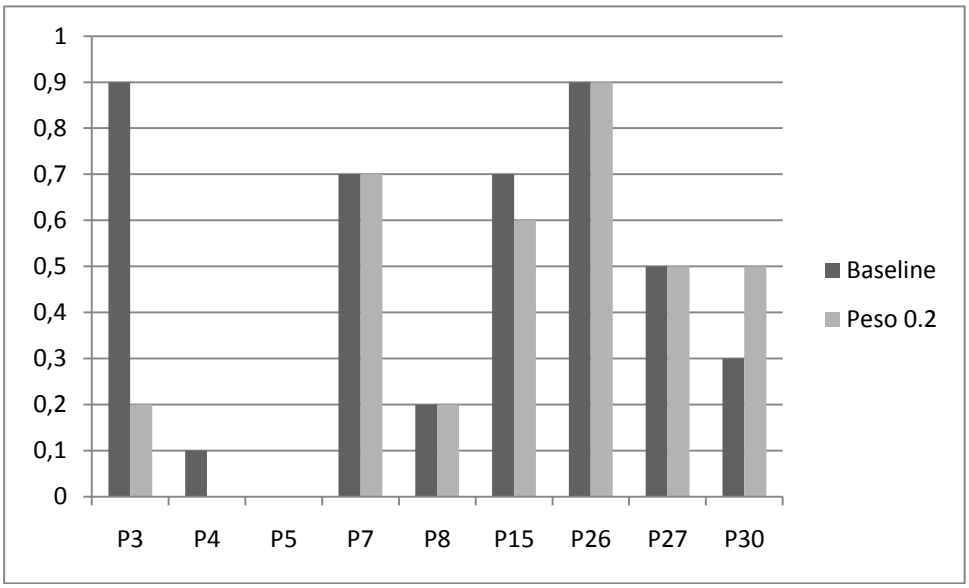

Figura 31 - Resultado da expansão por todo-parte em cada consulta

A consulta 3 foi a que obteve a maior degradação, passando de uma P@10 de 90% para uma P@10 de 20% após a expansão. Nesta consulta o termo "*gene expression*", que representa um processo na Gene Ontology, foi expandindo com as fases deste processo pelos termos "*transcription, protein maturation, RNA processing e translation*".

Para analisarmos a precisão da expansão usando relação todo-parte em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja com pesos para os termos expandidos em 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão está sempre abaixo da curva sem expansão, como pode ser visto na Figura 32.

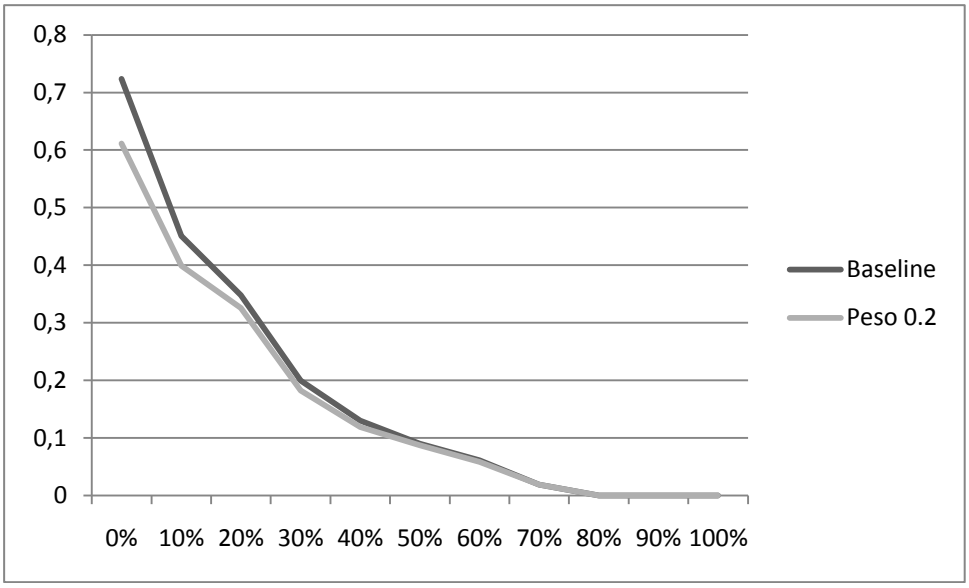

Figura 32 - Curva de cobertura vs. precisão da expansão por relação todo-parte

7.6 É parte de

As expansões baseadas em relações é parte de, no qual um componente é expandido com o termo que o contém, mostrou um resultado insatisfatório. Este tipo de relação foi explorado apenas no segundo experimento. Apenas 1 consulta possuía algum termo com relações do tipo é parte de, como mostra a Tabela 9. Como este resultado foi obtido em apenas uma consulta, seu resultado do ponto de vista estatístico é irrelevante e não pode ser generalizado.

O resultado obtido foi ruim no uso deste tipo de relação. No experimento realizado, a degradação da precisão média de 11 pontos ficou em 30,36%. O ponto de melhor resultado foi obtido quando os termos inseridos pelo sistema tinham peso entre

1 e 0,2. Nestas configurações, os resultados obtidos foram degradações de 30,36%. Além disso, medimos a precisão nos primeiros 10 documentos, a qual mostrou uma degradação de 66,67% com todos os pesos testados.

Tabela 21 - Resultado da expansão por relação é parte de

	2o. Experimento				2o. Experimento só relevantes			
Peso	1	0,5	0,33	0,2	1	0,5	0,33	0,2
MPM11P	0,0944	0,0945	0,0945	0,0945	0,0945	0,0945	0,0945	0,0945
Baseline	0,1357				0,0945			
Ganho	-30,43%	-30,36%	-30,36%	**-30,36%**	0,00%	0,00%	0,00%	**0,00%**
P@10	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000
Baseline	0,9000				0,3000			
Ganho	-66,67%	-66,67%	-66,67%	**-66,67%**	0,00%	0,00%	0,00%	**0,00%**

Os resultados obtidos quando considerados apenas os documentos definitivamente relevantes foram inconsistentes com os resultados obtidos da forma tradicional de medição, pois no primeiro o resultado ficou inalterado enquanto no segundo o resultado foi uma degradação.

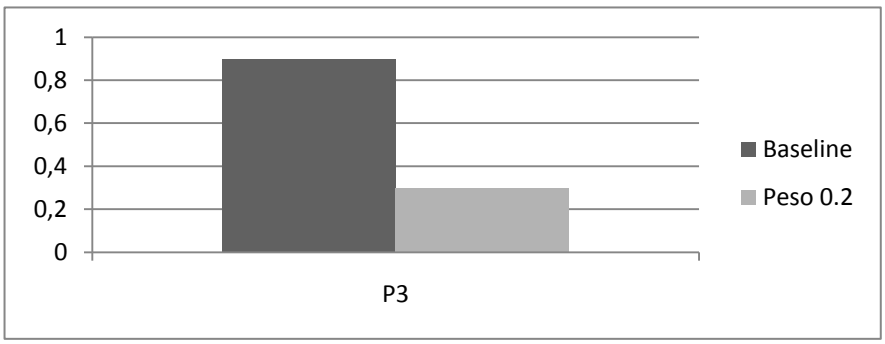

Figura 33 - Resultado da expansão por relação é parte de na consulta 3

A consulta 3 foi a única que apresentou este tipo de relação, onde o termo "*kidney development*" foi expandido pelos termos "*urogenital system development*" com base na ontologia de processos biológicos da Gene Ontology. O resultado obtido com essa expansão foi insatisfatório passando de um bom resultado sem expansão, onde a P@10 era de 90% para uma P@10 de 30% após expansão.

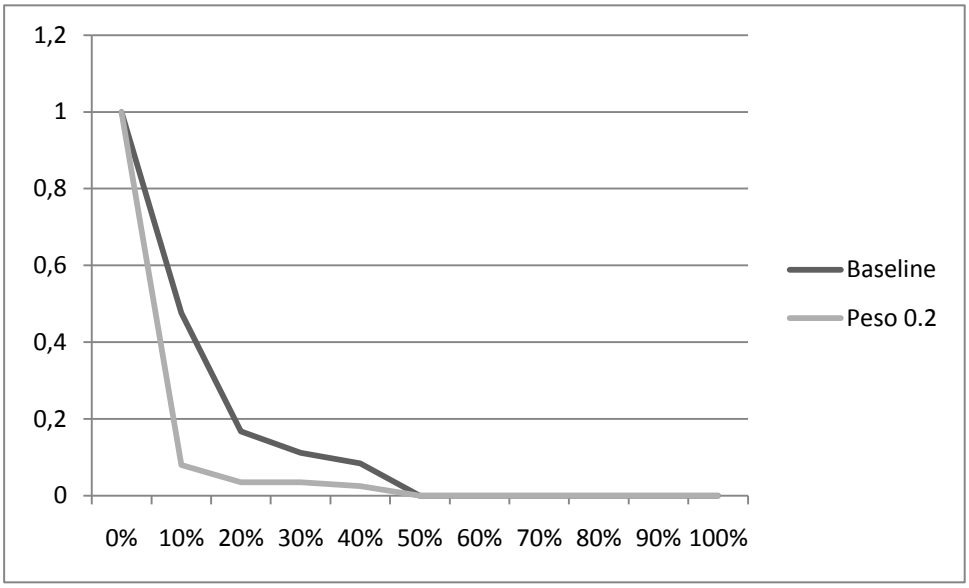

Figura 34 - Curva de cobertura vs. precisão da expansão por relação é parte de

Para analisarmos a precisão da expansão usando relação todo-parte em diferentes pontos da cobertura, traçamos a curva de cobertura precisão para a expansão com os pesos que em geral obtiveram os melhores resultados, ou seja com pesos para os termos expandidos entre 1 e 0,2. O resultado obtido mostrou que a curva resultante pela técnica de expansão com pesos 0,2 está sempre abaixo da curva sem expansão como pode ser visto na Figura 34.

7.7 Considerações Finais

De forma geral podemos observar na Tabela 22, que a única relação que em geral produziu melhoras consistentes no mecanismo de recuperação de informação foi a relação de sinonímia. Este resultado está alinhado com o resultado de trabalhos anteriores em expansão semântica de consultas.

Tabela 22 - Resumo dos resultados de cada relação

	1o. Experimento		2o. Experimento	
Relação	P@10	MPM11P	P@10	MPM11P
Sinonimia	-	**24,93%**	**10,31%**	**14,25%**
Generalização	-	-22,10%	-15,37%	-13,54%
Especialização	-	-26,42%	-28,16%	-39,11%
Doença	-	-	-34,62%	-28,51%
Contém	-	-	-16,28%	-8,86%
É parte de	-	-	-66,67%	-30,36%

Porém, ao fazer uma análise aprofundada de cada relação, foi possível observar que outros tipos de relações produzem melhoras em alguns casos. Isto indica que a técnica de expansão de consultas deveria usar essas relações de forma seletiva. Ou seja, os termos de consulta inseridos pelo usuário deveriam ser analisados para encontrarmos “dicas de expansão” que indiquem qual relação poderia ser utilizada. Dessa forma, o que podemos concluir é que a estratégia expandir sempre com um determinado peso só é eficaz ao utilizar a relação de sinonímia. Para os outros tipos de expansão uma técnica de expansão de consultas seletiva deve ser desenvolvida.

8. Conclusão

A revolução digital nos traz novos desafios em relação à forma como abordamos o problema da sobrecarga de informação. Cada vez mais, são necessárias ferramentas que atendam às nossas necessidades de informação de maneira completa (alta cobertura), e principalmente de forma correta (alta precisão). Vimos que um passo fundamental nesta direção é fazer com que os mecanismos de recuperação de informação “entendam” a semântica presente em nossos idiomas, deixando de ser apenas uma ferramenta de casamento de palavras-chaves. Vimos também que uma das formas possíveis de atingir este resultado é empregar a técnica de expansão de consultas. Todavia, os resultados presentes na literatura mostraram bons resultados para expansões de sinonímia e resultados desanimadores para outros tipos de relações léxico-semânticas.

Com a popularização das ontologias de domínio, as quais vêm se consolidando como um elemento de apoio no controle terminológico, como por exemplo, já podemos observar em sua utilização na área biomédica, surgem novas oportunidades de usar esses artefatos como apoio à técnica de expansão de consultas. Devido a sua expressividade, as ontologias permitem uma representação mais precisa de um domínio do que tesauros e taxonomias. Dessa forma, é possível explorar relações diferentes das relações tradicionais que são empregadas na técnica de expansão de consulta, onde foram feitas várias tentativas de viabilizar a técnica de expansão

semântica de consultas através das relações de sinonímia e generalização-especialização.

Nesta dissertação abordamos o problema da expansão semântica de consultas usando ontologias de domínio e outros mecanismos de controle terminológico como base de conhecimento para os termos inseridos pelo sistema de recuperação de informação. Fizemos dois experimentos usando os principais esquemas terminológicos da área biomédica. A aplicação deste conjunto de conhecimento de domínio na técnica de expansão semântica de consultas mostrou resultados consistentes com outros tipos de base de conhecimento em experimentos anteriores. Vimos que a única relação semântica que apresentou ganho de eficiência consistentemente no mecanismo de recuperação de informação foi a relação de sinonímia.

Semelhantemente aos experimentos anteriores, as relações de generalização e especialização mostram em geral uma degradação. Vários fatores podem explicar tal resultado. Foi identificado que algumas relações são empregadas de forma inconsistente nas ontologias de domínio. Para que as estratégias de expansão de consultas sejam desenvolvidas é necessário que o emprego destas relações seja feita de maneira formal. Por exemplo, na UMLS estava definido que existe uma relação de generalização-especialização (is_a) entre “folha de planta” e “planta”. Essa inconsistência é um exemplo de como as relações de generalização-especialização são usadas equivocadamente, pois a semântica entre estes dois termos é de todo-parte. Quando olhamos o emprego destas relações nas diferentes ontologias do domínio

biomédico o cenário foi ainda pior. Os termos que estavam presentes em mais de uma ontologia ou tesauro têm diferentes termos genéricos nos diferentes artefatos. Isso indica que existe um fraco compromisso ontológico no emprego das relações is_a presentes nas ontologias e vocabulários da área biomédica. Por exemplo, o termo *apoptosis* está presente em várias ontologias. Na SnomedCT seu termo genérico é "*Morphologically altered structure*". Na MeSH seu termo genérico é "*Cell Death*". Já na NCI Thesaurus seu termo genérico é "*Cell Death Process*". Na Gene Ontology seu termo genérico é "*programmed cell death*". Ou seja, apesar do conceito estar no entorno da morte celular, as ontologias divergem em relação ao seu compromisso ontológico, pois existe uma indefinição se o termo "apoptosis" é um processo, é o resultado de um processo ou é um estado de uma matéria (RANGANATHAN, 1933).

Essa falta de compromisso ontológico das ontologias reflete a falta de formalismo na qual a comunidade biomédica emprega a semântica de determinadas relações. Essa deficiência já foi identificada e está sendo trabalhada por pesquisadores da área de ontologias (SMITH ET AL., 2005)(SALES ET AL., 2008). Acreditamos que, uma vez que a comunidade da área biomédica passe a utilizar um discurso com maior precisão semântica, as técnicas de expansão de consultas baseadas em ontologias poderão apresentar resultados melhores. Por isso, acreditamos que as pesquisas relacionadas à área de ontologias abrirão novas oportunidades para pesquisas em expansão de consultas. Desta forma, estes desdobramentos deverão ser acompanhados de perto por pesquisadores.

Este trabalho analisou algumas relações léxico-semânticas que não haviam sido exploradas em trabalhos anteriores. A relação específica de domínio doença associada mostrou uma degradação significativa. As expansões baseadas na relação todo-parte também apresentaram degradações significativas. Porém, uma analise detalhada das expansões diferentes de sinonímia parece indicar para o uso desses tipos de relações em apenas dois tipos de situações: a primeira, quando o resultado da busca apresenta nenhum ou poucos documentos relevantes. Isso pode ser informado como forma de *feedback* do usuário com um botão do tipo, "não achei o que eu queria". Outra situação em que parece ser vantajoso usar estes tipos de relações é quando existe algum termo na consulta que indique o uso de uma determinada relação. Um exemplo desta situação aconteceu na consulta 30, onde o pesquisador queria encontrar informações sobre a membros da família de genes NKx. O termo membros indica uma oportunidade para expandir usando a relação todo-parte. Além disso, a presença de “componentes de” seria um outro indicativo para usar este tipo de relação. Por outro lado, a presença de “tipos de” indica uma expansão por especialização. O mapeamento de termos que sugerem expansões por determinados tipos de relações semânticas é um tema interessante para ser abordado em trabalhos futuros.

Por fim, é importante ressaltar que apesar dos resultados insatisfatórios no emprego da técnica de expansão semântica de consultas baseadas em ontologias de domínio, a pesquisa de recuperação de informação baseada em semântica é uma tendência que vem se fortalecendo nos últimos anos. Parece evidente que as

ontologias terão um papel central neste tipo de pesquisa. Existem diversas outras etapas onde as ontologias poderiam ser empregadas tais como: indexação semântica e organização dos resultados baseada em ontologias. Portanto, acreditamos que a utilização das ontologias em diferentes etapas do processo de recuperação de informação são temas interessantes para trabalhos futuros.

Além disso, nossos resultados indicam que o desenvolvimento de uma técnica de expansão de consultas seletiva poderia reverter os resultados negativos das expansões que não são de sinonímia. Isto pode ser uma ótima oportunidade de trabalhos futuros, pois pode-se desenvolver um mecanismo de interpretação da intenção de busca que procura por “dicas de expansão” na consulta do usuário e utiliza a relação correta para expansão. Por exemplo, se um usuário quer saber as diferentes raças de cachorro ele poderia digitar “tipos de cachorro” em um mecanismo de busca. Após a fase de interpretação o mecanismo identificaria a oportunidade de expandir pela relação is_a, pois a presença dos termos “tipo de” nesta consulta indicam uma grande probabilidade de ganho na expansão por termos específicos.

Analogamente, podemos estender este raciocínio para outros tipos de relação, como por exemplo, vimos que quando um usuário passa a idéia de membros de um conjunto, existe uma oportunidade de expansão pela relação part_of. Por exemplo, se um usuário digita em um sistema de recuperação de informação “PIB dos países

membros do MERCOSUL” poderíamos usar o conhecimento dos elementos deste conjunto e expandir com os termos Brasil, Argentina, Paraguai e Uruguai.

Referências Bibliográficas

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Addison-Wesley Longman Publishing Co., Inc. Recuperado Maio 2, 2010, de http://portal.acm.org/citation.cfm?id=553876&referer=www.clickfind.com.au, 1999.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1-7, p. 107-117. doi: 10.1016/S0169-7552(98)00110-X, 1998.

CAMPOS, M. L.; HAGAR, E. G. Taxonomia e Classificação: princípios de categorização. **DataGramaZero**, v. 9, n. 4, p. 1, 2008, Agosto.

CHAKRABARTI, S.; DOM, B.; AGRAWAL, R.; RAGHAVAN, P. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. **The VLDB Journal**, v. 7, n. 3, p. 163-178. doi: 10.1007/s007780050061, 1998.

CODD, E. F. **The relational model for database management: version 2**. Addison-Wesley Longman Publishing Co., Inc. Recuperado Maio 2, 2010, de http://portal.acm.org/citation.cfm?id=SERIES11430.77708&coll=ACM&dl=ACM&type=book&idx=SERIES11430&part=series&WantType=Proceedings&title=ACM-CBS, 1990.

FOX, E. A. Lexical relations: enhancing effectiveness of information retrieval systems. **SIGIR Forum**, v. 15, n. 3, p. 5-36. doi: 10.1145/1095403.1095404, 1980.

FU, G.; JONES, C. B.; ABDELMOTY, A. I. Ontology-Based Spatial Query Expansion in Information Retrieval. In: **On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE**. p.1466-1482. Recuperado Março 6, 2010, de http://dx.doi.org/10.1007/11575801_33, 2005.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. 6 ed. Atlas, 1999.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowl. Acquis.**, v. 5, n. 2, p. 199-220, 1993.

GUARINO, N. **Formal ontology in information systems**. IOS Press, 1998.

GUIZZARDI, G. Ontological foundations for structural conceptual models. . CTIT, Centre for Telematics and Information Technology. Recuperado Maio 15, 2010, de http://purl.org/utwente/50826, 2005.

IMRAN, H.; SHARAN, A. THESAURUS AND QUERY EXPANSION. **International Journal of Computer Science & Information Technology**, p. 89-97, 2009.

ir.ohsu.edu/genomics. . Recuperado Maio 2, 2010, de http://ir.ohsu.edu/genomics/.

KHAN, M. S.; KHOR, S. Enhanced web document retrieval using automatic query expansion. **J. Am. Soc. Inf. Sci. Technol.**, v. 55, n. 1, p. 29-40, 2004.

LEE, M.; TSAI, K. H.; WANG, T. I. A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. **Computers & Education**, v. 50, n. 4, p. 1240-1257. doi: 10.1016/j.compedu.2006.12.007, 2008.

LU, Z.; KIM, W.; WILBUR, W. Evaluation of query expansion using MeSH in PubMed. **Information Retrieval**, v. 12, n. 1, p. 69-80. doi: 10.1007/s10791-008-9074-8, 2009.

lucene.apache.org. . Recuperado Maio 2, 2010, de http://lucene.apache.org/.

MAEDCHE, A. **Ontology Learning for the Semantic Web (The Kluwer International Series in Engineering and Computer Science, Volume 665)**. 1º ed. Springer, 2002.

MANDALA, R.; TOKUNAGA, T.; TANAKA, H. Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. **Anais...** . p.191-197. Berkeley, California, United States: ACM. doi: 10.1145/312624.312677, 1999.

MANNING, C. D.; RAGHAVAN, P.; SCHTZE, H. **Introduction to Information Retrieval**. Cambridge University Press. Recuperado Maio 2, 2010, de http://portal.acm.org/citation.cfm?id=1394399, 2008.

MARCHIONINI, G.; BEN, S. Finding Facts vs. Browsing Knowledge in Hypertext Systems. **Computer**, v. 1988, n. 1, p. 77-80. doi: 10.1109/2.222119.

MARCHIONINI, G. Exploratory search: from finding to understanding. **Commun. ACM**, v. 49, n. 4, p. 41-46. doi: 10.1145/1121949.1121979, 2006.

OGDEN, C. K.; RICHARDS, I. A. **Meaning Of Meaning**. Mariner Books, 1989.

ORENGO, V.; HUYCK, C. A Stemming Algorithmm for the Portuguese Language. In: String Processing and Information Retrieval, International Symposium on. **Anais...** . v. 0, p.0186. Los Alamitos, CA, USA: IEEE Computer Society. doi: http://doi.ieeecomputersociety.org/10.1109/SPIRE.2001.10024, 2001.

PEAT, H. J.; WILLETT, P. The limitations of term co-occurrence data for query expansion in document retrieval systems. **Journal of the American Society for Information**

Science, v. 42, n. 5, p. 378-383. doi: 10.1002/(SICI)1097-4571(199106)42:5<378::AID-ASI8>3.0.CO;2-8, 1991.

PEREZ, A. G.; BENJAMINS, V. R. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods. **IN**, v. 18, p. 1--1, 1999.

PORTER, M. An algorithm for suffix stripping. **Program: electronic library and information systems**, v. 40, n. 3, p. 211 - 218. doi: 10.1108/00330330610681286, 2006.

pt.wikipedia.org/wiki/Ontologia. . Recuperado Maio 2, 2010, de http://pt.wikipedia.org/wiki/Ontologia.

research.nii.ac.jp/ntcir/index-en.html. . Recuperado Maio 2, 2010, de http://research.nii.ac.jp/ntcir/index-en.html.

ROBERTO NAVIGLI; PAOLA VELARDI. An Analysis of Ontology-based Query Expansion Strategies. In: Proceedings of Workshop on Adaptive Text Extraction and Mining (ATEM) in the 14th European Conference on Machine Learning (ECML). **Anais...** . p.42-49. Cavtat-Dubrovnik, Croatia, 2003.

ROBERTSON, S. Overview of the Okapi projects. **Journal of Documentation**, v. 53, p. 3-7. doi: 10.1108/EUM0000000007186, 1997.

SALES, L. F.; CAMPOS, M. L. D. A.; GOMES, H. E. Ontologias de domínio: um estudo das relações conceituais. **Perspectivas em Ciência da Informação**, v. 13, n. 2. doi: 10.1590/S1413-99362008000200006, 2008.

SALTON, G.; LESK, M. E. Computer Evaluation of Indexing and Text Processing. **J. ACM**, v. 15, n. 1, p. 8-36. doi: 10.1145/321439.321441, 1968.

SMITH, B.; CEUSTERS, W.; KLAGGES, B.; ET AL. Relations in biomedical ontologies. **Genome Biology**, v. 6, n. 5, p. R46. doi: 10.1186/gb-2005-6-5-r46, 2005.

sourceforge.net. . Recuperado Maio 2, 2010, de http://sourceforge.net/.

SPINK, A.; WOLFRAM, D.; JANSEN, M. B. J.; SARACEVIC, T. Searching the web: The public and their queries. **Journal of the American Society for Information Science and Technology**, v. 52, n. 3, p. 226-234. doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R, 2001.

STOKES, N.; LI, Y.; CAVEDON, L.; ZOBEL, J. Exploring criteria for successful query expansion in the genomic domain. **Information Retrieval**, v. 12, n. 1, p. 17-50. doi: 10.1007/s10791-008-9073-9, 2009.

trec.nist.gov/data.html. . Recuperado Maio 2, 2010, de http://trec.nist.gov/data.html.

trec.nist.gov/trec_eval. . Recuperado Maio 3, 2010, de http://trec.nist.gov/trec_eval/.

VOORHEES, E. M. Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. **Anais...** . p.61-69. Dublin, Ireland: Springer-Verlag New York, Inc. Recuperado Novembro 6, 2009, de http://portal.acm.org/citation.cfm?id=188508, 1994.

wordnet.princeton.edu. . Recuperado Maio 2, 2010, de http://wordnet.princeton.edu/.

www.cs.utk.edu/~lsi/corpa.html. . Recuperado Maio 2, 2010, de http://www.cs.utk.edu/~lsi/corpa.html.

www.geneontology.org. . Recuperado Maio 2, 2010, de http://www.geneontology.org/.

www.monster.com.br. . Recuperado Maio 2, 2010, de http://www.monster.com.br/Index.aspx.

www.wikipedia.org. . Recuperado Maio 2, 2010, de http://www.wikipedia.org/.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. **Anais...** . p.4-11. Zurich, Switzerland: ACM. doi: 10.1145/243199.243202, 1996a.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. **Anais...** . p.4-11. Zurich, Switzerland: ACM. doi: 10.1145/243199.243202, 1996b.

XU, J.; CROFT, W. B. Improving the effectiveness of information retrieval with local context analysis. **ACM Trans. Inf. Syst.**, v. 18, n. 1, p. 79-112. doi: 10.1145/333135.333138, 2000.

Anexo I : Código Fonte da Rotina de Indexação

```
static void indexDocs(IndexWriter writer, File file) throws IOException {
    // do not try to index files that cannot be read
    if (file.canRead()) {
        if (file.isDirectory()) {
            String[] files = file.list();
            // an IO error could occur
            if (files != null) {
                for (int i = 0; i < files.length; i++) {
                    indexDocs(writer, new File(file, files[i]));
                }
            }
        } else {
            System.out.println("adding " + file);
            Iterator<GenTrecDocument> iterator = new MedlineAsciiFileIterator(file);
            while (iterator.hasNext()) {
                i++;
                GenTrecDocument genTrecDoc = iterator.next();
                if (i % 1000 == 0) {
                    System.out.println(i + " readed documents...");
                }
                writer.addDocument(GenTrecDocument
                        .toLuceneDocument(genTrecDoc));
            }
        }
    }
}
```

Anexo II : Código Fonte da Rotina de Expansão

```
        TopicsXmlParser topicsParser = new TopicsXmlParser(new File(file));
        List<GenTrecTopic> topics = topicsParser.parse();

        /* Lucene Stuff */
        IndexReader reader = IndexReader.open(index);
        Searcher searcher = new IndexSearcher(reader);
        Analyzer analyzer = new StandardAnalyzer(StopWords.getStopWords());
        String[] documentsFields = new String[] { GenTrecDocument.TITLE,
                GenTrecDocument.ABSTRACT };
        MultiFieldQueryParser qParser = new MultiFieldQueryParser(
                documentsFields, analyzer);

        if ("and".equalsIgnoreCase(operator)) {
            qParser.setDefaultOperator(QueryParser.AND_OPERATOR);
        } else {
            qParser.setDefaultOperator(QueryParser.OR_OPERATOR);
        }

        List<TrecEvalRecord> evalRecords = new ArrayList<TrecEvalRecord>();
        TrecEvalResultsFileWriter resultWriter = new TrecEvalResultsFileWriter(
                new File(result));
for (int i = 0; i < topics.size(); i++) {
    GenTrecTopic topic = topics.get(i);
    System.out.println("Consulta: " + topic.getId());
    String userInputedQuery = topic.getUserInput();
    System.out.println("User Input: " + userInputedQuery);

    String systemExpansion = "";

    if (expandSyn) {
        StringBuffer queryPesificada = new StringBuffer();
        for (Collection<String> domainEntity : topic.getSynonymMap().values()) {
            for (String sinonimo : domainEntity) {
            queryPesificada.append("\"" + sinonimo + "\"^" + synWeight + " ");
            }
        }
        systemExpansion += queryPesificada.toString().trim();
    }

    if (expandSub) {
        StringBuffer queryPesificada = new StringBuffer();
        for (Collection<String> domainEntity : topic.getSubclassMap().values()) {
            for (String subclass : domainEntity) {
            queryPesificada.append("\"" + subclass + "\"^" + subWeight + " ");
            }
        }
```

```
    systemExpansion += queryPesificada.toString().trim();
}

if (expandSup) {
    StringBuffer queryPesificada = new StringBuffer();
    for (Collection<String> domainEntity : topic.getSuperclassMap().values()) {
        for (String supclass : domainEntity) {
        queryPesificada.append("\"" + supclass + "\"^" + supWeight + " ");
        }
    }
    systemExpansion += queryPesificada.toString().trim();
}

if (expandDis) {
    StringBuffer queryPesificada = new StringBuffer();
    for (Collection<String> domainEntity : topic.getDiseaseMap().values()) {
        for (String disease : domainEntity) {
        queryPesificada.append("\"" + disease + "\"^" + disWeight + " ");
        }
    }
    systemExpansion += queryPesificada.toString().trim();
}

if (expandPar) {
    StringBuffer queryPesificada = new StringBuffer();
    for (Collection<String> domainEntity : topic.getPartOfMap().values()) {
        for (String partOf : domainEntity) {
        queryPesificada.append("\"" + partOf + "\"^" + parWeight + " ");
        }
    }
    systemExpansion += queryPesificada.toString().trim();
}

if (expandCon) {
    StringBuffer queryPesificada = new StringBuffer();
    for (Collection<String> domainEntity : topic.getContainsMap().values()) {
        for (String contains : domainEntity) {
        queryPesificada.append("\"" + contains + "\"^" + conWeight + " ");
        }
    }
    systemExpansion += queryPesificada.toString().trim();
}

String resultedQuery = userInputedQuery + " " + systemExpansion;
System.out.println("Resulted query:" + resultedQuery);
Query query = qParser.parse(resultedQuery);
System.out.println("Searching for: " + query.toString());
```