

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

**UMA ARQUITETURA PARA UTILIZAÇÃO DE ONTOLOGIAS
EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO**

Marlon Candido Guérios

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina como requisito parcial à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Roberto C. S. Pacheco

**FLORIANÓPOLIS
2005**

MARLON CANDIDO GUÉRIOS

**UMA ARQUITETURA PARA UTILIZAÇÃO DE ONTOLOGIAS
EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO**

Esta dissertação foi julgada e aprovada para a obtenção do grau de
Mestre em Engenharia de Produção no
Programa de Pós-Graduação em Engenharia de Produção da
Universidade Federal de Santa Catarina.

Florianópolis, 30 de junho de 2005

Prof. Edson Pacheco Paladini, Dr.
Coordenador do Curso

BANCA EXAMINADORA

Prof. Dr. Roberto C. S. Pacheco
Orientador
Universidade Federal de Santa Catarina

Prof. Vinícius Medina Kern, Dr. Prof.
Universidade Federal de Santa Catarina

Aran Bey Tcholakian Morales, Dr.
Universidade Federal de Santa Catarina

SUMÁRIO

RESUMO.....	V
ABSTRACT.....	VI
LISTA DE FIGURAS.....	VII
LISTA DE TABELAS.....	VIII
LISTA DE SIGLAS E REDUÇÕES.....	IX
1 INTRODUÇÃO.....	10
1.1 CONTEXTUALIZAÇÃO.....	10
1.2 PROBLEMATIZAÇÃO.....	11
1.3 OBJETIVOS.....	11
1.3.1 <i>Objetivo Geral</i>	11
1.3.2 <i>Objetivos Específicos</i>	12
1.4 METODOLOGIA.....	13
1.5 DELIMITAÇÕES.....	14
1.6 JUSTIFICATIVA.....	15
1.6.1 <i>Contextualização na Engenharia de Produção</i>	16
1.7 ESTRUTURA DO TRABALHO.....	16
2 RECUPERAÇÃO DE INFORMAÇÃO.....	18
2.1 INTRODUÇÃO.....	18
2.2 VISÃO GERAL.....	18
2.3 MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO.....	21
2.3.1 <i>Modelo Booleano</i>	22
2.3.2 <i>Modelo Espaço Vetorial (VSM)</i>	24
2.3.3 <i>Modelo Vetor de Contexto (CVM)</i>	25
2.3.4 <i>Modelo Indexação Semântica Latente (LSI)</i>	26
2.3.5 <i>Modelo Difuso</i>	29
2.3.6 <i>Modelo Probabilístico</i>	30
2.4 MEDIDAS DE SIMILARIDADE.....	31
2.5 AVALIAÇÃO DE SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO.....	34
2.6 CONSIDERAÇÕES FINAIS.....	38
3 ONTOLOGIA.....	39
3.1 INTRODUÇÃO.....	39
3.2 WEB SEMÂNTICA.....	39
3.3 ONTOLOGIA.....	42
3.3.1 <i>O que é Ontologia</i>	42
3.3.2 <i>Tesouro, Dicionário e Vocabulário Controlado</i>	46
3.3.3 <i>Representação de Ontologias</i>	47
3.4 ONTOLOGIA EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO.....	52
3.5 CONSIDERAÇÕES FINAIS.....	54
4 ARQUITETURA PARA UTILIZAÇÃO DE ONTOLOGIAS EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO.....	55
4.1 INTRODUÇÃO.....	55
4.2 VISÃO GERAL.....	55
4.3 ARQUITETURA PROPOSTA.....	56
4.3.1 <i>Apresentação da Arquitetura</i>	56
4.3.2 <i>Módulo de Acoplamento de Ontologias</i>	59
4.3.3 <i>Módulo de Expansão do Vetor de Consulta</i>	69
4.3.4 <i>Apresentação do Resultado (Classificação)</i>	71
4.4 CONSIDERAÇÕES FINAIS.....	73
5 APLICAÇÃO DA ARQUITETURA: SITE SCIENTI SAÚDE.....	75

5.1	INTRODUÇÃO	75
5.2	RECUPERAÇÃO DE INFORMAÇÃO EM C&T	75
5.3	APLICAÇÃO DA ARQUITETURA	78
5.3.1	<i>Acoplamento de Ontologias</i>	79
5.3.2	<i>Expansão do Vetor de Consulta</i>	84
5.3.3	<i>Apresentação do Resultado (Classificação)</i>	84
5.3.4	<i>Site de Buscas da Rede ScienTI</i>	86
5.4	ESTUDO COMPARATIVO: SITE SCIENTI SAÚDE	88
5.5	DISCUSSÃO DOS RESULTADOS	93
5.6	CONSIDERAÇÕES FINAIS.....	96
6	CONCLUSÕES E TRABALHOS FUTUROS.....	97
6.1	CONCLUSÕES.....	97
6.2	TRABALHOS FUTUROS.....	98
	REFERÊNCIAS BIBLIOGRÁFICAS.....	100

RESUMO

GUERIOS, Marlon Candido. **Uma Arquitetura para Utilização de Ontologias em Sistemas de Recuperação de Informação**. 2005. 108 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2005.

Sistemas tradicionais de busca estão limitados ao nível léxico. A proposta deste trabalho é a construção de uma arquitetura de sistema de busca que permite o acoplamento de vocabulários para a ampliação do contexto da busca. O fundamento da arquitetura é a inclusão, em um sistema de busca, de um módulo que permite acoplar ontologias. A expansão do contexto da consulta é construída por meio da adição dos termos da ontologia relacionados ao vetor de busca. Os itens recuperados são ordenados segundo a similaridade entre os vetores de busca e do item. Para verificar a viabilidade da proposta, a arquitetura foi empregada em um site de buscas sobre currículos de pesquisadores da Rede ScienTI utilizando o DeCS como ontologia. Os resultados foram comparados aos obtidos por um sistema tradicional de recuperação de informação. A expansão da consulta por meio de ontologias incrementa o contexto semântico do vetor de consulta e, juntamente com o cálculo de similaridade vetorial, promove uma melhor classificação dos documentos.

Palavras-chave: Recuperação de Informação; Ontologias; Similaridade Vetorial.

ABSTRACT

Traditional search systems are limited to lexicon level. This work proposes the development of a system search architecture which allows vocabulary linkage for the search context enlargement. The foundation of the architecture is the inclusion of a module that enables ontology linkage. The expansion of the search context is built by adding ontology terms related to the search vector. The recovered items are arranged according to similarity between the search vectors and the item. To verify the proposal's availability, the architecture was used to build a search website called Rede ScienTI – that contains researchers' resumes – using DeCS ontology. The results of this architecture were compared to results obtained by a traditional information recovering system. The search expansion that ontology offers increases semantic context of the search vector and, amongst the vector similarity calculus (method), promotes a better document assortment (classification).

Keywords: Information Retrieval; Ontologies; Vectorial Similarity.

LISTA DE FIGURAS

Figura 1 - Metodologia.....	13
Figura 2 - Matriz termos x documentos em LSI.....	28
Figura 3 - Espaço vetorial.....	32
Figura 4 - Arquitetura de Camadas da Web Semântica.....	40
Figura 5 - Papel de ontologias e contextualizações na representação de um domínio.....	43
Figura 6 - Recuperação de informação baseada em ontologia.....	53
Figura 7 - Arquitetura conceitual de um sistema de recuperação de informação.....	57
Figura 8 - Arquitetura conceitual de um sistema de recuperação de informação utilizando ontologias.....	57
Figura 9 - Arquitetura proposta no contexto de um sistema de recuperação de informação.....	59
Figura 10 - Módulo de Acoplamento de Ontologias e seus componentes.....	60
Figura 11 - Especificação de uma representação formal para ontologias em XML Schema.....	62
Figura 12 - Fragmento de um arquivo XML representando o termo “Malária”.....	63
Figura 13 - Mapeamento entre a ontologia original e a representação especificada.....	64
Figura 14 - Componente para adaptação das ontologias.....	65
Figura 15 - Componentes de adaptação das ontologias.....	66
Figura 16 - Índice de Ontologias.....	67
Figura 17 - Pool de Ontologias.....	68
Figura 18 - Módulo de expansão do vetor de consulta no contexto da arquitetura proposta.....	70
Figura 19 - Vetor de consulta expandido.....	71
Figura 20 - Módulo de Classificação dos Resultados.....	72
Figura 21 - Exemplos de vetores com peso para cada termo.....	73
Figura 22 - Arquitetura conceitual para projetos de E-Gov.....	77
Figura 23 - Representação do índice de ontologias em XML.....	82
Figura 24 - Tela inicial do site de buscas da Rede ScienTI.....	87
Figura 25 - Tela de resultados do site de buscas da Rede ScienTI.....	88

LISTA DE TABELAS

Tabela 1 - Matriz Documento/Consulta x Termo	27
Tabela 2 - Recall x precisão.....	36
Tabela 3 - Informações necessárias para cada recurso descrito em uma ontologia.....	62
Tabela 4 - Descrição dos elementos da representação de ontologias	80
Tabela 5 - Elementos do índice de ontologias.....	83
Tabela 6 - Vetor de consulta	85
Tabela 7 - Vetor de documento com freqüências absolutas e normalizadas.....	85
Tabela 8 - Número de documentos recuperados para cada termo com e sem a utilização de ontologias para a expansão semântica do contexto da consulta.....	89
Tabela 9 - Vetor de consulta expandido para o termo “aids”.....	90
Tabela 10 - Documentos recuperados para o termo “aids” e seus sinônimos	91
Tabela 11 - Vetor expandido para o termo “obesidade”	92
Tabela 12 - Documentos recuperados para o termo “obesidade” e seus sinônimos	92
Tabela 13 - Vetor expandido para o termo “saúde pública”	93
Tabela 14 - Documentos recuperados para o termo “saúde pública” e seus sinônimos.....	93

LISTA DE SIGLAS E REDUÇÕES

BVS	Biblioteca Virtual em Saúde
C&T	Ciência e Tecnologia
CT&I	Ciência, Tecnologia e Inovação
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CONSCIENTIAS	Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior
CVM	Context Vector Model
DeCS	Descritores em Ciências da Saúde
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
LSI	Latent Semantic Indexing
MESH	Medical Subject Headings
OWL	Ontology Web Language
RBC	Raciocínio Baseado em Casos
RDF	Resource Description Framework
ScienTI	Rede Internacional de Fontes de Informação e Conhecimento para a Gestão de Ciência, Tecnologia e Inovação
SQL	Structured Query Language
TFIDF	Term Frequency X Inverse Document Frequency
TREC	Text REtrieval Conference
URI	Uniform Resource Identifier
VSM	Vector Space Model
WEB	World Wide Web
XML	eXtended Markup Language

1 INTRODUÇÃO

1.1 Contextualização

No meio científico, o acesso irrestrito a toda e qualquer literatura referente aos seus objetos de estudo é o sonho de qualquer pesquisador (SCHATZ, 1997).

Na década de 1940 o problema de armazenamento e recuperação de informações atraiu demasiada atenção, pois a quantidade de informação aumentava gradativamente em detrimento da velocidade de acesso, necessitando de maior demanda de trabalho de pessoas. Essa lentidão ao acesso da informação causava, entre outras coisas, a perda de informação relevante, prejudicando a tomada de decisão.

A funcionalidade básica dos sistemas de recuperação de informação praticamente é a mesma há 30 anos (SCHATZ, 1997). Um conjunto de documentos é indexado e armazenado em um repositório, seja centralizado ou distribuído, o qual pode ser acessado e consultado a partir de um conjunto de palavras-chave. Os documentos recuperados são analisados pelo usuário, sendo os relevantes retidos e os demais descartados. Porém, o problema reside no fato de que os usuários não têm todo o tempo livre para ler e analisar todos os documentos e verificar quais lhe são mais relevantes.

Com o objetivo de resolver tais limitações, muito tem se estudado acerca da utilização de ontologias na construção de sistemas de recuperação de informação. Devido à estrutura que possuem, ontologias podem auxiliar sistemas de recuperação a serem mais inteligentes, pois contam com os conceitos e os relacionamentos dos

mais diversos termos que podem ser utilizados em uma consulta (HAAV; LUBI, 2001; MENA et. al., 2000).

1.2 Problematização

Apesar do constante desenvolvimento e da utilização de ontologias como suporte a sistemas de recuperação de informação, pretende-se verificar como tais sistemas podem ser aprimorados de modo que as consultas realizadas pelo usuário possuam semântica e como o resultado dessas consultas pode ser classificado segundo sua relevância, ou seja, documentos que realmente interessam ao usuário sejam posicionados no início da lista.

1.3 Objetivos

1.3.1 Objetivo Geral

Esta dissertação propõe uma arquitetura que busca minimizar os esforços necessários para o acoplamento de ontologias em sistemas de recuperação de informação (visando ao contexto semântico) e para a classificação (ordenação) dos resultados obtidos, segundo a similaridade com a consulta.

1.3.2 Objetivos Específicos

- Propor uma forma de representação padrão para ontologias na arquitetura.
- Realizar, a partir da utilização de ontologias, a expansão do vetor de consulta baseado na inserção dos termos relacionados.
- Classificar os resultados segundo a sua similaridade, posicionando os itens possivelmente mais relevantes no início.
- Implementar a arquitetura proposta em um site de buscas sobre saúde.
- Realizar um estudo comparativo entre um modelo tradicional e o modelo proposto a fim de se realizar uma discussão sobre as diferenças nos resultados de cada proposta.

1.4 Metodologia

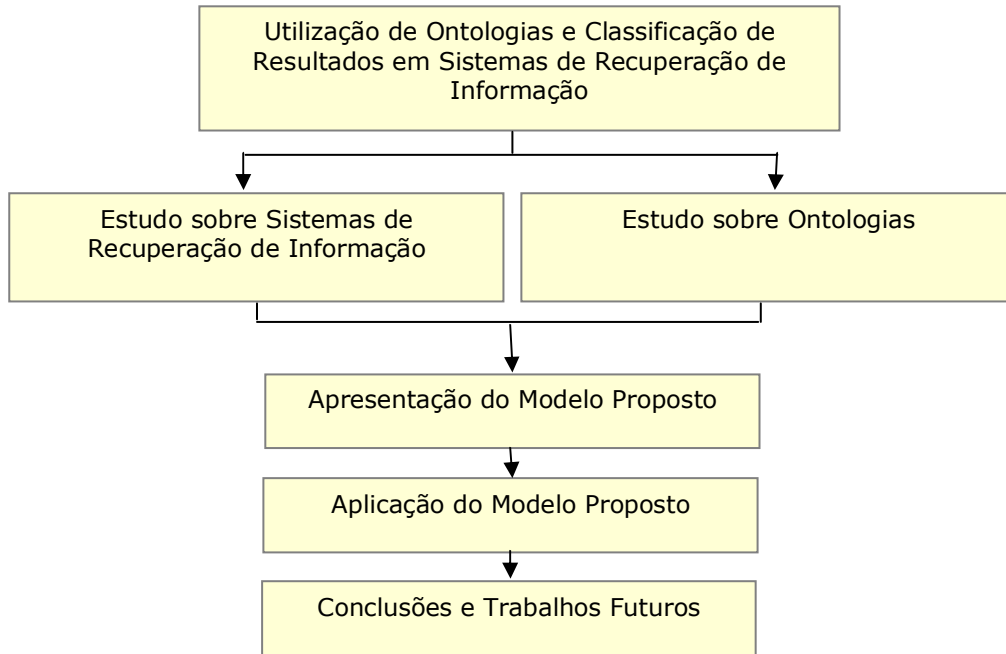


Figura 1 - Metodologia

A Figura 1 apresenta uma visão esquemática da metodologia utilizada para o desenvolvimento deste trabalho, a qual está baseada nas seguintes etapas:

- estudo sobre sistemas de recuperação de informação, relacionando os modelos mais utilizados e as técnicas para classificação e avaliação de resultados;
- estudo sobre ontologia, tipos de ontologias, formas de implementação e representação, assim como um estudo sobre a sua utilização em sistemas de recuperação de informação;

- desenvolvimento de uma arquitetura que visa facilitar a utilização de ontologias em sistemas de recuperação de informação e a classificação dos resultados apresentados ao usuário;
- estudo de caso aplicado ao site de busca da Rede ScienTI – Rede Internacional de Fontes de Informação e Conhecimento para a Gestão de Ciência, Tecnologia e Inovação –, objetivando validar a utilização de ontologias em sistemas de recuperação de informação através da arquitetura proposta;
- ao final do trabalho, são levantados as conclusões e os possíveis trabalhos futuros que podem ser alvo de pesquisa nas áreas de recuperação de informação e ontologia.

1.5 Delimitações

Este trabalho se propõe a demonstrar como as mais diversas ontologias existentes podem ser integradas em um sistema de recuperação de informação. Não faz parte do escopo deste trabalho desenvolver técnicas de recuperação de informação ou demonstrar algoritmos de busca, tampouco pretende construir ontologias ou apresentar metodologias para tal.

A representação de ontologias demonstradas neste trabalho se restringe ao XML, embora existam outras formas que podem ser utilizadas em pesquisas futuras, conforme apresentado no Capítulo 6.

Neste trabalho não serão consideradas questões de performance dos sistemas de recuperação de informação em estudo.

1.6 Justificativa

No processo de recuperação de informação, a necessidade normalmente é expressa como uma consulta composta de poucas palavras (BILLERBECK et al., 2003). Além disso, sistemas de recuperação de informação atuais utilizam-se, em sua maioria, de buscas baseadas em palavras-chave, o que não é satisfatório devido à sua baixa precisão (HAAV; LUBI, 2001; WITTEN et al., 1999; MENA et al., 2000). No entanto, a adição de termos à consulta pode contribuir para uma melhoria significativa na efetividade de um sistema de recuperação de informação (BILLERBECK et al., 2003).

A utilização de uma ferramenta baseada no método proposto, realizando buscas semânticas com o acoplamento de ontologias, que normalmente compreendem um domínio específico, pode contribuir e guiar os usuários em suas buscas de modo que encontrem rapidamente o que necessitam. Além disso, estudos realizados por Paralic e Kostial (2003) sobre a utilização de ontologias em sistemas de recuperação de informação concluíram que os resultados foram satisfatórios em comparação com outros métodos que não utilizam ontologias. Porém, tais estudos ressaltam que se poderia criar um mecanismo baseado na expansão da consulta por meio da agregação de termos correlatos de forma automática, facilitando ainda mais a utilização por parte do usuário (PARALIC; KOSTIAL, 2003; MENA et al., 2000).

Segundo Lancaster (1968), um sistema de recuperação de informação não muda o conhecimento do usuário, apenas informa da existência de documentos inerentes ao assunto pesquisado. No entanto, atualmente se acredita que buscas semânticas dão ao usuário a possibilidade de aumentar o seu conhecimento, uma vez que

novos termos relacionados à sua consulta original são apresentados e podem conduzi-lo a novos caminhos, não pensados antes, em sua busca pela informação desejada.

1.6.1 Contextualização na Engenharia de Produção

Com relação ao problema de pesquisa proposto e ao foco de pesquisa abordado pela Engenharia de Produção, cabe destacar o entendimento comum tanto do American Institute of Industrial Engineering quanto da Associação Brasileira de Engenharia de Produção, os quais consideram a necessidade de implementar melhorias nos sistemas produtivos.

Este trabalho propõe melhorias em sistemas de recuperação de informação, auxiliando o usuário no momento da consulta através da utilização de ontologias e ordenação apropriada da lista de documentos encontrados.

1.7 Estrutura do Trabalho

Este trabalho é composto de seis capítulos, os quais estão descritos a seguir.

No Capítulo 2 é realizado um estudo sobre a Recuperação de Informação, como surgiram os primeiros sistemas, quais são os sistemas atuais e para onde caminham.

No Capítulo 3 apresentam-se o conceito de ontologia bem como suas aplicações, e como isso pode ser aplicado em sistemas de recuperação de informação.

No Capítulo 4 o modelo proposto é apresentado no qual se realiza a integração de ontologias a sistemas de recuperação de informação de maneira dinâmica.

No Capítulo 5 este modelo é confrontado com o modelo tradicional, e seus resultados são analisados.

No Capítulo 6 são apresentadas as conclusões obtidas com a realização deste trabalho e os possíveis trabalhos futuros a serem desenvolvidos na área de recuperação de informação e ontologias.

2 RECUPERAÇÃO DE INFORMAÇÃO

2.1 Introdução

Neste capítulo são abordados a recuperação de informação, as suas definições, o histórico e qual a importância dos sistemas de recuperação de informação em um mundo onde as pessoas estão cada vez mais conectadas entre si através de computadores e onde a informação é disponibilizada em larga escala. Também são apresentadas as principais técnicas utilizadas.

2.2 Visão Geral

Segundo Baesa-Yates e Ribeiro-Neto (1999), a recuperação, a representação, o armazenamento, a organização e o acesso são processos de gestão da manipulação da informação. Pode-se definir “recuperação de informação” como o procedimento de, partindo de uma necessidade de informação, ir em busca de informação em meio a um emaranhado de documentos dos mais variados tipos. Os resultados da busca podem ser apresentados em forma de texto, imagem, som ou vídeo (THE FREE DICTIONARY, 2004).

O termo *recuperação de informação* possui muitas definições. A própria palavra *informação* tem um conceito ambíguo em algumas ocasiões. Shannon e Weaver (1964) em sua teoria da comunicação dizem que tecnicamente, no contexto da recuperação de informação, o significado da palavra *informação* não tem uma definição exata, ou seja, são muitos os casos em que essa palavra pode ser

substituída por *documento*. No entanto, o termo *recuperação de informação* é amplamente aceito na literatura sobre esse tema.

Uma definição simples e bastante aceita é dada por Lancaster (1968):

Um sistema de recuperação de informação não informa (isto é, muda o conhecimento do) (o) usuário a respeito do assunto de sua pesquisa. Apenas informa da existência (ou não) e a localização do documento relacionado à sua requisição¹.

O objetivo dos sistemas de recuperação de informação é encontrar a informação e apontar para ela, que pode ou não ser aquela que irá satisfazer a necessidade corrente do usuário.

No caso de a World Wide Web (Web) tentar encontrar a informação desejada, pode ser uma experiência frustrante e desapontadora. Os recursos na Web são muito heterogêneos e estão disponibilizados e espalhados pela rede. São vários os formatos em que a informação é disponibilizada: texto, hipertexto, imagem, som, vídeo, animação, entre outros. Em muitos casos as coleções de informação são dinâmicas e vão além dos limites físicos por estarem distribuídas em rede. Assim como a informação é heterogênea, as pessoas que utilizam a Web também são. Esses usuários possuem diferentes necessidades de informação bem como conhecimentos diferentes tanto com relação à informação de que necessitam quanto com relação à própria utilização da Web como fonte de pesquisa (WANG et al., 2000).

Conforme apontado por Jenkins et al. (1998), os componentes essenciais para a criação de um sistema de recuperação de informação efetivo são: um mecanismo que recupera e analisa documentos relacionados a outros através de hiperlinks; um

¹ “An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.” (LANCASTER, 1968).

indexador que extrai os termos dos documentos e os disponibiliza de forma a serem facilmente encontrados; um banco de dados com metadados descrevendo os recursos armazenados; um mecanismo de recuperação que percorre os índices em busca dos documentos relevantes à necessidade de informação do usuário; e uma interface para a interação com o usuário, isto é, onde este possa efetuar sua consulta (*query*).

Sistemas de recuperação de informação na Web são essências, pois a forma como os usuários interagem com a interface atualmente disponível se dá em baixo nível, por meio de *hiperlinks*. A utilização de sistemas eficientes de recuperação de informação auxilia o usuário a chegar mais rapidamente ao seu destino, respondendo assim à sua necessidade de informação (JENKINS et al., 1998).

Com o intuito de contribuir para a usabilidade da Web, centenas de mecanismos de busca baseados em palavras-chave têm rapidamente suportado novos processos de busca de informação. Os exemplos mais populares são: Google e Yahoo!, além do Alta Vista, que foi um dos pioneiros nesta área.

Como muitos outros mecanismos de busca, o Alta Vista é composto de três componentes (JENKINS et al., 1998):

- um *webcrawler* que baixa os documentos da Web;
- um *indexador* que extrai os termos-chave desses documentos, que são utilizados para representar o documento recuperado;
- uma *interface de consulta* que recebe os termos, os quais são comparados com o banco de documentos. Assim, aqueles que tiverem maior similaridade com a consulta são apresentados ao usuário.

No caso do Yahoo!, essa abordagem é estendida através da intervenção humana. Uma taxonomia de termos de busca é construída, e os documentos da

Web são classificados segundo essa taxonomia. Isso tende a limitar o escopo dos documentos (JENKINS et al., 1998).

O Google trabalha como o Alta Vista, porém difere deste na determinação da relevância dos documentos. A ordem na qual os documentos são apresentados é especificada pelo seu índice de cotação, isto é, documentos que são mais bem cotados que outros são vistos como sendo mais importantes. Essa cotação é calculada sobre o número de links que apontam para certo documento. Uma consulta normal pode retornar milhares de documentos apresentando aqueles que são mais relevantes primeiramente; isso é muito mais interessante do que representar todos os documentos que são relevantes. Neste caso a qualidade é mais importante que a quantidade (BRIN; PAGE, 1998).

2.3 Modelos de Recuperação de Informação

Basicamente um sistema de recuperação de informação realiza suas buscas de acordo com um termo ou conjunto de termos informados pelo usuário em alguma interface apropriada. Tais termos (que em princípio indicam qual a necessidade de informação do usuário)² são comparados de alguma forma com os documentos previamente armazenados na base de informação com o intuito de se encontrarem documentos que sejam relevantes ao usuário.

Muitas são as técnicas utilizadas na construção de um sistema de recuperação de informação que deve levar em conta um esquema de representação dos

² Muitas vezes um usuário de um sistema de recuperação de informação não sabe exatamente o que deseja, ou caso saiba, o seu vocabulário a respeito do assunto é restrito. O modelo proposto por este trabalho atua justamente neste ponto por ser baseado em ontologias, como pode ser visto em detalhes no Capítulo 4.

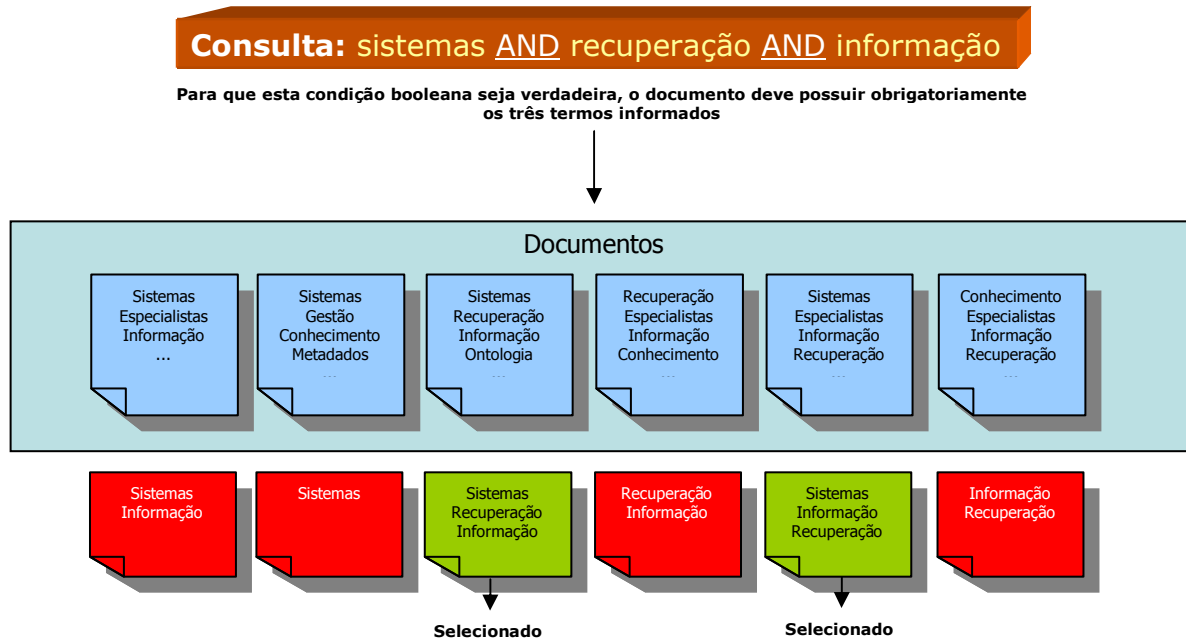
documentos, a formulação da consulta e a construção de uma função de *ranking* (que indica a relevância dos documentos em relação à consulta), são mais conhecidos os modelos de espaço vetorial, de vetor de contexto, de indexação semântica latente, booleano, difuso, probabilístico e de linguagem natural (KORFHAGE, 1997; WONG; YAO, 1995).

2.3.1 Modelo Booleano

O modelo mais comum e mais simples, e largamente utilizado nos sistemas atuais é o booleano. Esse modelo se utiliza de álgebra booleana para realizar a busca sobre a base de informação onde as consultas são especificadas com o auxílio dos operadores lógicos AND, OR e NOT (WITTEN et al., 1999).

Sendo assim, dado um conjunto de termos conectados entre si através dos operadores lógicos, serão recuperados os documentos que satisfazem os termos informados (KORFHAGE, 1997).

Na Figura 1 é apresentado um exemplo usando a consulta: “sistemas” AND “recuperação” AND “informação”, a qual é submetida a um mecanismo de busca que atua sobre uma base indexada de documentos. Neste caso, somente os documentos que possuírem todos os termos informados serão selecionados e exibidos ao usuário.



Esse modelo pode ser utilizado em buscas sobre arquivo-texto ou ainda em bancos de dados relacionais. No caso dos bancos de dados, a consulta é realizada através de uma linguagem própria, a SQL - *Structure Query Language* (FILHO, 2001).

Para se realizar uma busca de pessoas que moram em Santa Catarina mas que nasceram no Paraná sobre um banco de dados onde possua uma tabela com os dados de uma pessoa, pode-se utilizar uma *query*, conforme se demonstra a seguir.

Select nome from tabela where

uf_endereco = 'Santa Catarina' AND uf_nascimento = 'Paraná'

Nesse modelo não é possível estabelecer um *ranking* para a lista de documentos, pois todos os documentos que satisfizerem a condição booleana terão o mesmo peso, ou seja, a mesma relevância diante da consulta (WONG; YAO, 1995).

2.3.2 Modelo Espaço Vetorial (VSM)

Em um esforço em busca de melhores resultados, foi proposto um novo modelo conhecido como “modelo espaço vetorial” (*vector space model*), no qual os vários elementos que compõem o sistema de recuperação de informação são modelados como elementos pertencentes a um espaço vetorial (WONG; RAGHAVAN, 1984; DOMINICH, 2000; CAID; CARLETON, 1994).

Neste modelo, cada vetor se torna uma unidade a ser comparada com outra através da determinação da similaridade entre o vetor do documento e o vetor da consulta.

Entretanto, para a formação dos vetores, tornam-se necessários dois conceitos importantes. O primeiro refere-se à extração de características a partir de uma unidade de informação, e o segundo se refere à redução de dimensionalidade e normalização dos termos.

Neste modelo, podem ser utilizados pesos, a frequência normalizada, para cada termo presente no vetor, indicando a sua relevância dentro do documento. O peso pode ser automaticamente calculado através da frequência intradocumento e sobre toda a coleção de documentos. Tanto a frequência quanto o peso podem ser considerados no cálculo de similaridade. Em uma busca padrão todos os termos da consulta têm pesos iguais, podendo ser oferecida ao usuário a possibilidade de modificar tais pesos, segundo seus critérios e suas necessidades. Uma alteração dos pesos dos termos da consulta pode modificar os documentos que serão recuperados e o seu posicionamento no conjunto (KORFHAGE, 1997).

Segundo Korfhage (1997), quando o modelo vetorial de recuperação é utilizado, medidas de similaridade podem ser associadas com a idéia de distância, seguindo a filosofia de que documentos próximos no espaço vetorial são altamente similares, ou como uma medida angular, baseada na idéia de que documentos “na mesma direção” estão relacionados.

Por adotar uma medida de similaridade aproximada, o modelo espaço vetorial oferece a possibilidade de estabelecer um *ranking* para o resultado obtido da busca, pois, de acordo com a medida utilizada, pode-se definir a relevância do documento em relação ao vetor de consulta apresentando ao usuário uma ordenação dos documentos segundo essa medida (SALTON; MCGILL, 1983; RAGHAVAN; WONG, 1986). Mais detalhes sobre medidas de similaridade serão apresentados na seção 2.4.

2.3.3 Modelo Vetor de Contexto (CVM)

O modelo vetor de contexto é uma extensão do modelo espaço vetorial apresentado no tópico anterior. Os vetores dos sistemas de recuperação de informação baseados no modelo espaço vetorial são compostos apenas da listagem dos termos e de suas freqüências, sendo esses termos os considerados no momento de se medir a similaridade com outros vetores, impondo assim limites ao resultado, pois não consideram os relacionamentos semânticos que porventura existirem entre os termos informados e os outros termos existentes nos documentos. Isso significa que documentos relevantes à consulta em questão poderão ficar de fora do resultado (BILLHARDT et al., 2002; CAID; CARLETON, 1994). Assim sendo,

o modelo espaço vetorial foi expandido para agregar aos termos do vetor (que representa um documento) o contexto em que ele está inserido, ou seja, outros termos que podem indicar alguma relação semântica.

Shütze (1992) considera o significado semântico dos termos e seus contextos como vetores em um espaço vetorial em que as dimensões correspondem aos termos. Tais vetores são denominados “vetores de contexto” por possuírem o contexto onde cada termo está inserido no documento. Esse contexto é obtido através dos termos próximos, dentro do texto, considerando uma janela de termos que indica quantos deles antes e depois serão considerados na definição do contexto. Assim, para cada termo pode ser gerado um vetor contendo os termos próximos no contexto, e o vetor de contexto de um documento é formado por esses pequenos vetores de contexto relacionados a cada termo (CAID; CARLETON, 1994).

2.3.4 Modelo Indexação Semântica Latente (LSI)

A maioria dos métodos considera a ocorrência dos termos, informados na consulta, nos documentos para realizar os cálculos de similaridade que indicarão o grau de relevância de um documento diante dessa consulta. A desvantagem desta abordagem reside no fato de que documentos com alto grau de relevância, porém que não contêm um termo sequer dos que foram informados na consulta, não serão apresentados no resultado da busca. Muitas vezes alguns termos importantes para o sentido da busca que está sendo realizada não são informados por mero desconhecimento ou mesmo esquecimento do usuário no momento de construir a

consulta. Assim, na abordagem em que se considera apenas a ocorrência dos termos para se definir o grau de relevância, muitos documentos relevantes ficarão de fora (DUMAIS et al., 1988; BLAIR; MARON, 1985).

Na tentativa de resolver essa deficiência, o modelo de indexação semântica latente (*latent semantic indexing*) utiliza uma abordagem que leva em consideração a co-ocorrência de termos, isto é, conjuntos de termos que freqüentemente são encontrados nos mesmos documentos. Se tais termos surgem com freqüência nos mesmos documentos relativos a determinada área, isto pode evidenciar que existe aí uma relação semântica latente, ou seja, não é explícita. Com base em técnicas estatísticas, o modelo de indexação semântica latente pode “descobrir” as possíveis correlações existentes (MANNING; SCHÜTZE, 1999).

A utilização das palavras pelo ser humano é caracterizada por um extenso uso de sinônimos. Portanto, uma comparação direta por termos pode ser deficiente. As pessoas normalmente desejam acessar a informação baseada no seu significado, e a comparação direta de palavras não consegue realizar esse trabalho com sucesso (DUMAIS et al., 1988).

Tabela 1 - Matriz Documento/Consulta x Termo

	Termo 1	Termo 2	Termo 3	Termo 4
Consulta	usuário	interface		
Documento 1	usuário	interface	HCI	interação
Documento 2			HCI	interação

Fonte: Adaptada de: MANNING; SCHÜTZE, 1999.

Na Tabela 1 são exibidos um exemplo de consulta, dois documentos e seus respectivos termos. Utilizando o modelo de similaridade por ocorrência simples do termo, o documento que seria recuperado para a consulta dada seria apenas o documento 1, porém, se for levada em conta a co-ocorrência dos termos, deve-se perceber que os termos “*HCI*” e “*interação*” ocorrem em ambos os documentos, logo, estatisticamente, pode-se presumir que os termos “*usuário*” e “*interface*” possuem alguma relação semântica latente com “*HCI*” e “*interação*”. Assim, o documento 2 também deveria ser recuperado. Obviamente esse é um exemplo simples, e em um conjunto maior de documentos os termos citados poderiam não co-ocorrer com tanta frequência ou mesmo não ocorrer. Isso é apenas um exemplo de como o modelo de indexação semântica latente considera não somente a ocorrência dos termos isoladamente mas também a probabilidade de outros termos co-ocorrerem indicando alguma relação semântica.

No modelo de indexação semântica latente, os termos e os documentos/consulta são projetados em uma matriz termos x documentos que indica a ocorrência e a frequência de cada termo em cada documento. Na Figura 2 é apresentada uma matriz com a relação entre os termos e as suas ocorrências nos documentos.

$$A = \begin{pmatrix} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \textit{espaçonave} & 1 & 0 & 1 & 0 & 0 & 0 \\ \textit{astronauta} & 0 & 1 & 0 & 0 & 0 & 0 \\ \textit{lua} & 1 & 1 & 0 & 0 & 0 & 0 \\ \textit{veículo} & 1 & 0 & 0 & 1 & 1 & 0 \\ \textit{pick-up} & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Figura 2 - Matriz termos x documentos em LSI

2.3.5 Modelo Difuso

Este modelo é baseado na Teoria dos Conjuntos Difusos, que redefine a pertinência de um elemento a um conjunto generalizando a função de pertinência do intervalo fechado e discreto $[0,1]_{\in I}$ para o intervalo contínuo $[0,1]_{\in IR}$. Sendo assim, define-se um conjunto difuso por $\{x \in \bar{A} \mid \mu_x(A) \in [0,1]_{\in IR}\}$.

A partir disso, todos os modelos baseados na Teoria dos Conjuntos puderam ser generalizados para a Teoria dos Conjuntos Difusos, que é o caso da busca difusa, uma generalização do modelo booleano.

A lógica difusa fundamenta sistemas de raciocínio aproximado e não exato. No caso da lógica difusa, o raciocínio exato é o caso-limite do raciocínio aproximado. Ao contrário dos modelos booleanos, nos quais o valor verdade só pode assumir *verdadeiro* ou *falso*, no modelo difuso, o valor verdade é um subconjunto difuso de qualquer conjunto (TAKEMURA, 2003).

A utilização do modelo difuso para sistemas de recuperação de informação se justifica devido ao fato de que um determinado documento não contém necessariamente toda a informação relevante para o usuário, ou seja, dois documentos podem ser muito semelhantes, mas não iguais. No entanto, podem ser complementares mutuamente, e embora ambos possam ser de interesse do usuário, ainda assim um pode ser mais relevante do que o outro. Neste caso é interessante que o documento que é mais relevante apareça no topo da lista de documentos encontrados. Isso se deve justamente ao fato de o modelo difuso utilizar raciocínio aproximado e não exato.

Nesse contexto, o modelo difuso possibilita ao usuário especificar um peso para cada termo componente da consulta. Esse peso é utilizado no cálculo da relevância, em que o usuário pode informar ao sistema que determinado termo é mais relevante que outro (WIVES; LOH, 2000).

Dado o exemplo: “(*malária*) 0.5 (*febre amarela*) 0.9”, os documentos que contêm o termo “*febre amarela*” serão considerados mais relevantes que aqueles que contêm “*malária*”. Os documentos que contêm ambos os termos serão ainda mais relevantes.

Uma outra possibilidade do modelo difuso é a especificação de operadores difusos que podem ser qualitativos ou quantitativos (WIVES; LOH, 2000). Os operadores qualitativos incluem descritores numéricos (alto, baixo, grande) e não numéricos (bonito, colorido, etc.). Os operadores quantitativos incluem descritores relacionados a quantidades (pouco, muito, vários). Para serem válidos, esses operadores devem ser exatamente definidos (KORFHAGE, 1997).

2.3.6 Modelo Probabilístico

No modelo probabilístico, o cálculo para se determinar a relevância de um documento é baseado em análises estatísticas e não semântica. É um modelo bastante próximo ao modelo difuso, porém é necessário que algumas regras probabilísticas sejam satisfeitas durante a consulta (RIJSBERGEN, 1999).

Em um modelo probabilístico, o resultado corresponde a documentos que satisfazem a consulta com uma probabilidade mais alta que uma probabilidade mínima especificada (RIJSBERGEN, 1999). É possível utilizar a frequência de

termos no documento para se estimar a probabilidade dos termos de uma *query* em relação a um documento (KORFHAGE, 1997).

2.4 Medidas de Similaridade

A efetividade de um sistema de recuperação de informação está intimamente ligada à sua habilidade em avaliar com precisão a relevância dos documentos em relação a uma consulta realizada pelo usuário. Peça fundamental de um componente que realize tal avaliação, é a medida de similaridade entre os documentos e a consulta (JONES; FURNAS, 1987).

Existem diversas formas de se medir a similaridade entre dois vetores, sendo as principais a medida do produto interno, a medida do co-seno, a medida do pseudoco-seno, a medida de Dice, as medidas de correlação e covariância (JONES; FURNAS, 1987).

Considerando o modelo espaço vetorial, os documentos relevantes para uma determinada consulta são aqueles representados por vetores próximos ao vetor que representa a consulta, contemplando uma representação geométrica (MANNING; SCHÜTZE, 1999).

Na Figura 3 é exibido um plano cartesiano representando uma consulta em relação a alguns documentos indexados no modelo espaço vetorial.

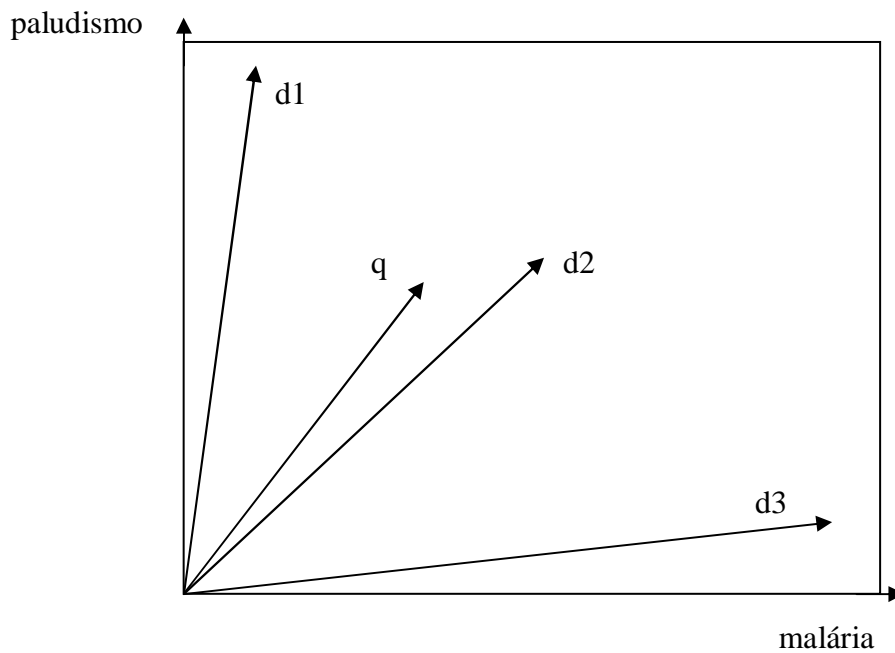


Figura 3 - Espaço vetorial

O espaço vetorial exibido na Figura 3 possui duas dimensões representando as palavras *paludismo* e *malária*. Há três documentos $d1$, $d2$ e $d3$ e um vetor de consulta q . As coordenadas dos vetores representam os pesos em relação aos termos e são calculadas de acordo com a frequência com que esses termos ocorrem nos documentos.

No exemplo, o documento $d1$ possui uma maior frequência do termo *paludismo* do que do termo *malária*, enquanto que o documento $d3$ possui um maior número de ocorrências do termo *malária* do que *paludismo*. Já o documento $d2$ possui um pequeno ângulo em relação ao vetor q , que representa a consulta. Logo, este documento será considerado o mais relevante para a situação apresentada.

A medida da similaridade através do cálculo do co-seno é muito utilizada no modelo de recuperação de informação vetorial (KORFHAGE, 1997) devido ao seu

grau de estabilidade (EGGHE; MICHEL, 2002). Segundo Chiao e Zweigenbaum (2002), a medida do co-seno foi a que apresentou a melhor performance.

A forma utilizada normalmente possui a seguinte definição:

$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n q_i d_i}{\sqrt{\sum_{k=1}^n q_i^2} \sqrt{\sum_{j=1}^n d_i^2}} \quad (1)$$

Neste caso q e d são os vetores de consulta e documento, respectivamente. A correlação é medida entre a ocorrência do termo i (representada por q_i e d_i) na consulta e nos documentos dividida pela medida euclidiana dos vetores, onde k e j são os termos do vetor de cada documento e da consulta. Quando dois vetores que são os termos do vetor de cada documento e da consulta. Quando dois vetores que são comparados possuem os mesmos termos e pesos, ou seja, são idênticos, o ângulo entre eles será zero, e conseqüentemente eles terão a medida máxima de similaridade com a aplicação dessa equação (SALTON et al., 1975). Quanto mais próximo de 1, maior a similaridade entre os vetores.

Como parâmetro para o cálculo do co-seno, utiliza-se normalmente o peso do termo, sendo comumente utilizada a medida denominada TFIDF (*Term Frequency x Inverted Document Frequency*), que considera a freqüência intradocumento e a freqüência de documento, ou seja, o número de documentos da coleção em que cada termo ocorre (WITTEN et al., 1999).

Para TF, pode-se realizar uma normalização das freqüências de cada termo no documento com relação à freqüência do termo com maior número de ocorrências.

TF é definido por:

$$tf_i = \frac{f_i}{\max f_d} \quad (2)$$

onde f_i é a frequência para cada termo i e $\max f_d$ é a maior frequência no documento d (Adaptado de: WITTEN et al., 1999).

O cálculo da frequência inversa do documento (IDF) é definido como:

$$idf_i = \log\left(\frac{N}{n_i}\right) \quad (3)$$

onde N é o número de documentos na coleção e n_i é o número de documentos na coleção em que ocorre o termo i .

Logo, o cálculo de TFIDF fica definido como:

$$tfidf_i = tf_i \times \log\left(\frac{N}{n_i}\right) \quad (4)$$

2.5 Avaliação de Sistemas de Recuperação de Informação

A avaliação de sistemas de recuperação de informação tem um papel muito importante de julgar a eficiência e a efetividade do processo de recuperação. Alguns fatores devem ser levados em consideração no processo de avaliação de um sistema de recuperação de informação, entre eles, o tempo gasto pelo sistema para realizar a busca, o tempo despendido pelo usuário para obter a informação desejada

e a habilidade do sistema para recuperar itens realmente relevantes. Essa é uma tarefa árdua, se não impossível, pois é muito difícil obter todos os parâmetros necessários para se chegar a uma conclusão contundente. E mesmo que fosse possível obter todos os parâmetros ainda seria difícil combiná-los apropriadamente para se chegar a uma simples medida final (RAGHAVAN et al., 1989).

Uma forma de realizar essa avaliação é tomar por referência uma coleção de documentos de teste que contém um conjunto de consultas e a informação sobre a relevância de cada documento para cada consulta (ZOBEL, 1998). Um exemplo desse tipo de coleção é a TREC (*Text Retrieval Conference*), que é uma coleção de teste produzida como resultado de *workshops* na área de recuperação de informação (HARMAN, 1993). Pesquisadores envolvidos receberam gigabytes de dados textuais, um conjunto de consultas e um conjunto de cálculos de relevância que deveriam ser avaliados para se medir a efetividade da recuperação.

Com uma base de referência como a TREC é possível comparar os resultados de um sistema de recuperação de informação com a indicação de relevância oferecida pela TREC a fim de se obterem os valores de *recall* e *precisão* para o sistema avaliado.

A medida denominada *recall* indica a capacidade do sistema de recuperação de informação em tentar recuperar todos os documentos relevantes para a consulta em questão. Quanto maior o percentual de documentos relevantes recuperados pelo sistema, maior será a medida de *recall*.

Calcula-se o *recall* através do percentual de itens relevantes recuperados pelo total de itens relevantes existentes na base.

Tabela 2 - Recall x precisão

Documento	Relevante
D1	√
D2	√
D3	√
D4	X
D5	X
D6	√
D7	X
D8	X
D9	X
D10	X
Recall	75%
Precisão	60%

Para o exemplo demonstrado na Tabela 2, basta dividir o número de documentos relevantes recuperados pelo número de documentos relevantes existentes e multiplicar por 100:

$$r = \frac{3}{4} \times 100 = 75$$

Caso o *recall* calculado seja igual a 100%, todos os documentos relevantes foram recuperados pelo sistema.

A outra medida citada, a precisão, informa a capacidade do sistema em recuperar somente o que é relevante, deixando para trás documentos não relevantes. Portanto, quanto menos documentos irrelevantes forem apresentados no resultado da consulta, maior a precisão.

Na Tabela 2 a precisão foi calculada dividindo-se o número de documentos relevantes pelo número de documentos recuperados, e o resultado multiplicado por 100:

$$p = \frac{3}{5} \times 100 = 60$$

Normalmente existe uma relação inversa entre os valores de *recall* e precisão, quando se obtém um *recall* alto, geralmente o valor da precisão tende a diminuir e o contrário também é verdadeiro, ou seja, com uma maior precisão, normalmente o *recall* decresce. Isso acontece porque, para atingir um *recall* alto, o sistema acaba por recuperar documentos demais, fazendo com que o número de documentos relevantes aumente (*recall* mais alto) assim como o número de documentos irrelevantes (precisão mais baixa). Por outro lado, para se atingir uma precisão mais alta, o número de documentos recuperados normalmente é reduzido, fazendo com que o número de documentos relevantes seja baixo (*recall* mais baixo) assim como o número de documentos irrelevantes (precisão mais alta) (MANNING; SCHÜTZE, 1999; WITTEN et al., 1999). Normalmente os sistemas buscam um equilíbrio entre as duas medidas.

Uma outra medida utilizada é a precisão em um determinado *cutoff*, que é um ponto de corte entre os documentos recuperados, ou seja, a medida de precisão é realizada apenas nos 10 primeiros ou nos 20 primeiros documentos. Isso pode indicar a eficiência do sistema em estabelecer o *ranking* dos documentos, fazendo com que os mais relevantes apareçam no início da lista (MANNING; SCHÜTZE, 1999).

2.6 Considerações Finais

Este capítulo apresentou um levantamento sobre os estudos que vêm sendo realizados na área de recuperação de informação, apresentando a importância de sistemas dessa natureza em um mundo onde a informação cresce rapidamente e se torna cada vez mais difícil encontrar aquilo que é considerado relevante. Apresentou ainda os modelos de recuperação mais utilizados e suas particularidades, as formas para se estabelecer o *ranking* dos documentos e as formas de avaliação dos resultados através das medidas de *recall* e precisão. Para se atingir o objetivo deste trabalho, faz-se necessário ainda um estudo sobre ontologias e sua aplicabilidade a sistemas de recuperação de informação. Este é o assunto do capítulo que segue.

3 ONTOLOGIA

3.1 Introdução

O capítulo anterior apresentou as técnicas mais utilizadas em sistemas de recuperação de informação e em especial as baseadas em ontologias. Neste capítulo são apresentados os conceitos relacionados à criação e utilização de ontologias em sistemas de informação que fazem parte de outro conceito maior e mais abrangente, a Web Semântica. Esta área tem por objetivo principal tornar a Web (e seus documentos) mais compreensível e organizada, não apenas para os leitores humanos mas também para os agentes de software que estejam buscando ou manipulando informações e até mesmo conhecimento na Web.

3.2 Web Semântica

O conceito de ontologias está inserido na arquitetura de camadas da Web Semântica, idealizada por Tim Berners-Lee com o intuito de se criarem novas formas de representação do conhecimento e de possibilitar uma melhor interoperabilidade entre sistemas através da Web. É uma maneira de se organizar a *World Wide Web* levando em consideração a semântica dos muitos documentos que a integram (BENNERS-LEE et al., 2001).

As páginas Web como vêm sido desenvolvidas atualmente não possuem muito significado se não forem lidas por um humano, sendo muito difícil automatizar a interoperabilidade e o fluxo do conhecimento que há nessas páginas. Isso ocorre

porque tais páginas são desenvolvidas utilizando a linguagem HTML, que basicamente é utilizada para a formatação e a diagramação de texto e imagens na página (W3C, 2004).

A Web Semântica vem como uma extensão àquilo que já existe, habilitando computadores e pessoas a trabalharem em cooperação. Como os dados são organizados segundo seu conteúdo semântico, a descoberta, a automação, a integração e o reuso por diversas aplicações se tornam mais fáceis (HENDLER et al. 2002).

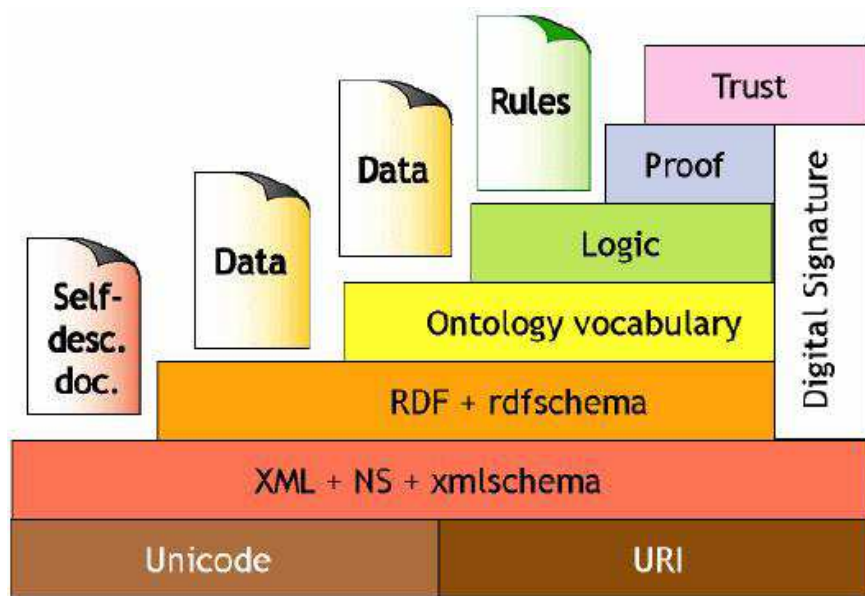


Figura 4 - Arquitetura de Camadas da Web Semântica

FONTE: BENNERS-LEE et al., 2001.

Na Figura 3 são ilustradas as tecnologias que compõem a arquitetura da Web Semântica, descritas brevemente a seguir até a camada de OWL (*Ontology Web Language*).

- **URI:** o URI (*Uniform Resource Identifier*) é o identificador de um recurso qualquer na Web (BENNERS-LEE et al., 2001).

- **XML e XML Schema:** XML é base para a representação dos dados em um formato flexível que possibilite a interoperabilidade entre sistemas. XML Schema pode ser definido como o conjunto de metadados que definem um XML.
- **RDF e RDF Schema:** RDF define como descrever recursos em termos de suas propriedades e valores. É uma forma padronizada de representar ontologias usando formato XML. RDF Schema, por sua vez, define propriedades específicas que podem ser utilizadas para definir esquemas.
- **OWL (*Ontology Web Language*):** é o nível lógico. É necessário que haja caminhos de escrita lógica dentro de documentos que permitam recursos, tais como a checagem de um documento contra um conjunto de regras. Para isso, o W3C está trabalhando na especificação da linguagem.

Neste contexto, ontologias têm um papel fundamental na Web Semântica por oferecerem uma forma de integração com diferentes representações dos recursos Web. A representação implícita de um domínio em uma ontologia pode ser vista como uma estrutura integrada que vai dar à informação uma representação comum e semântica, tornando possível a Web Semântica (DAVIES et al., 2003).

3.3 Ontologia

3.3.1 O que é Ontologia

O termo "ontologia" começou a ser empregado na ciência da computação, dentro da inteligência artificial, no início dos anos 90, em projetos para organização de grandes bases de conhecimento, como CYC (LENAT et al., 1990) e Ontolíngua (GRUBER, 1992). São várias as áreas da inteligência artificial que têm realizado pesquisas em ontologia, incluindo integração inteligente de informação, cooperativa de sistemas de informação, recuperação de informação e gestão do conhecimento (GUARINO, 1998).

Apesar de existirem algumas definições para ontologia, a que melhor caracteriza a essência de uma ontologia, segundo Gruber (1993), é: “Uma ontologia é uma especificação formal e explícita de uma conceituação compartilhada”. E por *conceituação* entenda-se um modelo abstrato de algum fenômeno no mundo que identifique os conceitos relevantes de tal fenômeno.

Segundo Gruber (1993), a especificação deve ser formal, de forma que seja possível a compreensão por um agente não humano (software), e explícita, ou seja, o tipo dos conceitos e as restrições ao seu uso são definidos explicitamente.

Acredita-se que a representação formal do conhecimento tenha começado na Índia do primeiro milênio a. C. com o estudo da gramática de *Shastric Sanscrit*. No entanto, da forma como é vista atualmente, essa disciplina tem uma relação muito próxima aos trabalhos realizados na Grécia Antiga, principalmente por Aristóteles (384-322 a. C.), nos campos da lógica, das ciências naturais e da filosofia metafísica (DAUM; MERTEN, 2002 apud CASTOLDI, 2003).

O produto dessa área, inicialmente criada por Aristóteles com seu abrangente sistema de classificação, de taxonomização e de representação do conhecimento de forma geral, é chamado hoje de *Ontologia*.

Considerando o sentido filosófico, uma ontologia pode ser classificada como um sistema de categorização tentando explicar uma certa visão do mundo, independente da linguagem utilizada para descrevê-la. No entanto, uma ontologia na Inteligência Artificial refere-se a um artefato de engenharia formado por um vocabulário específico que descreve uma certa realidade (GUARINO; WELTY, 1998).

Gruber (1995) sugere o termo *conceituação* para a ontologia filosófica. Assim, uma ontologia pode ser considerada uma especificação de uma conceituação, em que uma conceituação independe da linguagem na qual foi especificada. Isto significa que ontologias podem ser totalmente diferentes em relação ao vocabulário utilizado, mas ainda assim se referirem a uma mesma conceituação. A Figura 5 apresenta a ontologia como uma especificação de uma conceituação.

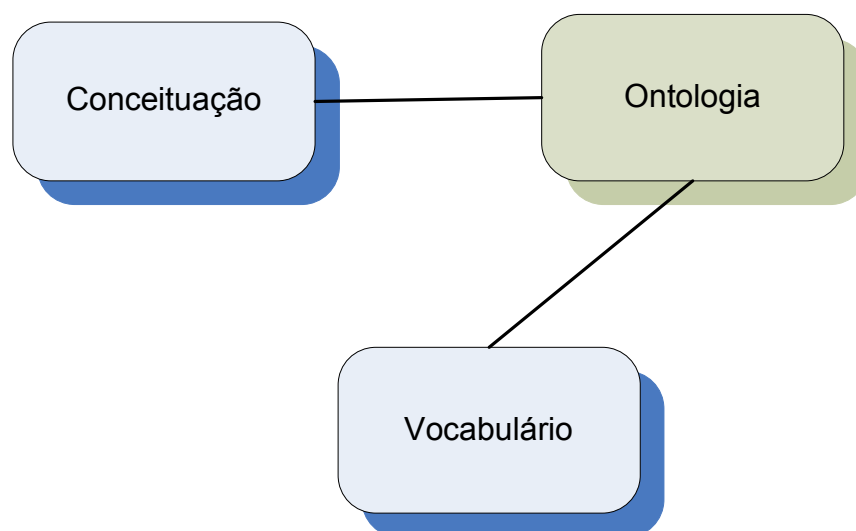


Figura 5 - Papel de ontologias e contextualizações na representação de um domínio

FONTE: CASTOLDI, 2003.

Esquemas de banco de dados relacional podem ser considerados exemplos de ontologias, pois definem uma hierarquia de tipos, especificando classes e seus relacionamentos de subordinação (CASTOLDI, 2003). Ao atuar como um contrato entre parceiros, a ontologia permite que eles se comuniquem com segurança dentro do contexto do domínio (DAUM; MERTEN, 2002 apud CASTOLDI, 2003).

Uma ontologia permite a comunicação entre sistemas computadorizados independentemente de tecnologias, arquiteturas e domínios (ONTOLOGY, 2003), e possui papel crucial no processamento do conhecimento baseado na Web (DECKER, 2000).

A produção de gramáticas e vocabulários comuns à comunidade usuária interessada em um determinado domínio é um dos resultados do trabalho de definição de ontologias (PACHECO; KERN, 2001). Isto promove a uniformização de referências, possibilitando e facilitando o processo de descoberta e geração de conhecimento (PACHECO, 2003).

Um dos resultados do trabalho de definição de ontologias é a produção de gramáticas e vocabulários comuns à comunidade usuária interessada no domínio correspondente (PACHECO; KERN, 2001), uniformizando referências e conseqüentemente possibilitando e facilitando o processo de descoberta e geração de conhecimento (PACHECO, 2003).

Segundo Noy e McGuinness (2000), as principais razões para se utilizarem ontologias são:

- compartilhamento da mesma estrutura de informação entre pessoas e agentes de software;
- permissão do reuso do conhecimento do domínio;

- separação do conhecimento do domínio do conhecimento operacional;
- análise do conhecimento do domínio.

Ontologias podem ser classificadas quanto ao tipo e quanto à profundidade. Os tipos de ontologias existentes, segundo Guarino (1998), são:

- ontologias genéricas: não dependem de um problema ou domínio em particular;
- ontologias de domínio: aplicam-se a um certo domínio do conhecimento, por exemplo, área da saúde;
- ontologias de tarefas: aplicam-se a uma determinada tarefa, por exemplo, análise de requisitos de software;
- ontologias de aplicação: aplicam-se a uma certa aplicação, especializando ontologias de domínio ou de tarefas.

Quanto à profundidade ontológica, Guarino e Welty (1998) classificam as ontologias em vários níveis, são eles:

- nível de vocabulários: é a forma mais simples de ontologia, podendo ser definida através de um XML Schema;
- nível de taxonomia: a definição de relacionamentos entre os termos estabelece seus significados, sendo o mais comum o relacionamento “é um”;
- nível relacional: são relacionamentos não hierárquicos, como, por exemplo, o relacionamento “dirigido por” entre carro e motorista;
- nível axiomático: além de relacionamentos, as ontologias definem restrições, que são conhecidas como axiomas. Um exemplo de axioma em

uma ontologia de casa seria: “uma casa deve ter quatro paredes”, ou seja, se não possuir quatro paredes, não é uma casa.

3.3.2 Tesouro, Dicionário e Vocabulário Controlado

O emprego do termo “ontologia” para denominar uma estrutura de termos e as relações entre eles em um determinado domínio é mais comum na área da ciência da computação e, mais particularmente, na subárea da inteligência artificial. Alguns pesquisadores (JASPER; USCHOLD, 1999; Fensel et al., 2001) consideram os tesouros como ontologias simples, uma vez que uma ontologia complexa, segundo esses autores, exige uma riqueza maior de relações do que as tradicionalmente apresentadas em um tesouro. Pode-se citar como exemplo a seguinte passagem de Fensel et al. (2001): "Grandes ontologias como Wordnet provêm um tesouro para mais de cem mil termos explicados em linguagem natural"³.

Neste sentido, pode-se entender os tesouros como sendo um tipo de ontologia voltada para a organização de termos. Além disso, Stojanovic (2004) também afirma que um vocabulário pode ser considerado uma ontologia.

O termo “tesouro”, que também designa dicionário, vocabulário ou léxico, começou a ser utilizado com mais frequência após a publicação do dicionário analógico de Peter Mark Roget, em Londres, em 1852, intitulado "Thesaurus of English words and phrases". A partir daí, diversas definições e significados para o termo surgiram, como a definição dada pelo programa Unisist (UNESCO, 1973) analisando o termo sob os aspectos estrutural e funcional, sendo o primeiro "um

³ “Larges ontologies such as WordNet provide a thesaurus for over 100,000 terms explained in natural language”.

vocabulário controlado dinâmico de termos relacionados semântica e genericamente cobrindo um domínio específico do conhecimento", e o segundo "um dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem do sistema (linguagem de documentação, linguagem de informação) mais restrita".

No entanto, uma das definições mais aceitas e atuais do momento é dada por Currás (1995):

Tesauro é uma linguagem especializada, normalizada, pós-coordenada, usada com fins documentários, onde os elementos lingüísticos que a compõem – termos, simples ou compostos – encontram-se relacionados entre si sintática e semanticamente.

3.3.3 Representação de Ontologias

As linguagens de marcação XML (*eXtended Markup Language*), XML Schema, RDF (*Resource Description Framework*) e RDF Schema são consideradas facilitadoras na representação e no desenvolvimento de ontologias baseadas na Web (HORROCKS, 2002).

Para Decker (2000), tanto XML como RDF têm seus méritos como fundamento para a realização da Web Semântica. Porém, RDF provê melhores mecanismos para a representação de ontologias, pois sua estrutura objeto-atributo permite uma melhor estruturação semântica dos objetos que são entidades independentes. Um modelo de domínio que define objetos e seus relacionamentos entre si pode ser representado naturalmente em RDF, sem a necessidade de passos adicionais como no caso da utilização de XML (DECKER, 2000).

No entanto, para a arquitetura apresentada neste trabalho, a utilização de XML satisfaz as necessidades, pois a ontologia representada é um vocabulário comum à comunidade usuária (pesquisadores da Rede ScienTI) (PACHECO; KERN, 2001; GUARINO; WELTY, 1998). Entretanto, no quesito interoperabilidade semântica, RDF ou mesmo outras formas de representação como OWL têm vantagens sobre XML (DECKER, 2000).

A seguir é apresentada uma breve introdução sobre XML e RDF.

a. eXtended Markup Language (XML)

XML pode ser definida como uma linguagem de marcação de dados (BERMEJO, 2004), e a sua utilização possibilita representar em um mesmo lugar dados e metadados (DECIO, 2000).

A responsabilidade sobre o desenvolvimento da linguagem XML é do W3C (*World Wide Web Consortium*) (W3C, 2004). Iniciou-se esse desenvolvimento em 1996 com o intuito de construir uma linguagem de marcação que combinasse a flexibilidade e a capacidade do SGML (*Standard Generalized Markup Language*) com a ampla aceitação do HTML (*Hyper Text Markup Language*) (ANDERSON et al., 2001).

Entre as principais características da linguagem XML, pode-se destacar que é um padrão aberto e extensível, independente de tecnologia e plataforma, o que dá bastante flexibilidade às aplicações (BAX, 2001 apud BERMEJO, 2004). Isso tem feito com que XML se popularize em seu segmento como padrão para a interoperabilidade (BERMEJO, 2004).

A linguagem XML também tem sido bastante utilizada e aceita para a padronização de informação no âmbito do governo eletrônico (BERMEJO, 2004), como exemplo pode-se citar o trabalho da Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior - CONSCIENTIAS, que estabeleceu padrões em XML para as unidades de informação da Plataforma Lattes (CONSCIENTIAS, 2004).

Para a definição da estrutura e do conteúdo de um documento XML é utilizado o XML Schema, que expressa vocabulários compartilhados e permite à máquina compreender regras estabelecidas por pessoas (W3C, 2004).

No Capítulo 4 é apresentado exemplos de XML e XML Schema.

b. Resource Description Framework (RDF)

O uso efetivo de metadados pelas aplicações requer convenções em comum sobre a semântica relacionada, a sintaxe e as estruturas bem definidas (MILLER, 1998).

Com a especificação da linguagem RDF por parte do W3C (*World Wide Web Consortium*), é possível também a especificação da modelagem de dados desenvolvida para a descrição de recursos na Web com a utilização de metadados (PÉREZ et al., 2004).

Basicamente, RDF é uma infra-estrutura que possibilita a codificação, o intercâmbio e o reuso de metadados estruturados. É importante lembrar que RDF não define o sentido semântico dos recursos, mas possibilita a construção de elementos de metadados conforme a necessidade (PÉREZ et al., 2004). RDF provê

estruturas comuns para possibilitar a interoperabilidade de dados em XML (POWERS, 2003).

O modelo RDF consiste em três componentes principais (PÉREZ et al., 2004):

a) *recursos*: são quaisquer tipos de dados descritos através de expressões RDF e identificados por meio de URIs (*Uniform Resource Identifiers*);

b) *propriedades (ou predicados)*: definem atributos ou relações usados para descrever um recurso;

c) *objetos*: atribuem um determinado valor a uma propriedade em um determinado recurso.

A partir de RDF é possível descrever recursos, os quais possuem propriedades e são considerados objetos unicamente identificados. As propriedades expressam as relações de valores associados com recursos, que podem ser atômicas como um texto ou um número e também podem ser outros recursos, que por sua vez podem ter suas próprias propriedades. Recursos relacionados por valores de propriedades formam uma sentença (MILLER, 1998).

Como exemplo, consideram-se as duas sentenças abaixo.

"O autor do documento 1 é indivíduo A"

"O indivíduo A é autor do documento 1"

Humanamente chega-se facilmente à conclusão de que as duas sentenças possuem o mesmo significado, porém máquinas não têm essa capacidade e tratam as sentenças como textos completamente distintos.

A primeira sentença possui um recurso simples que é o documento 1, uma propriedade que é o autor e um valor para essa propriedade que é o indivíduo A, ou seja, o sujeito é documento 1, o predicado é autor e o objeto é indivíduo A.

Da mesma forma, a segunda sentença possui um recurso simples que é o indivíduo A, uma propriedade que é o autor e outro recurso que é o documento 1. Porém, essa sentença está invertida com relação à primeira. RDF provê meios de representar ambas as sentenças de uma forma única, o que é lógico, já que elas têm o mesmo significado. Essas sentenças podem ser representadas como um grafo, porque consiste em um fato único que pode ser descrito utilizando a metodologia RDF (POWERS, 2003). Nas sentenças em questão, documento 1 e indivíduo A são nós e autor é uma aresta. A seguinte notação é utilizada para representar o fato:

{“Documento 1”, autor, “Indivíduo A”}

A fim de facilitar a construção de documentos RDF, foi criado um modelo denominado RDF/XML – uma representação formal de dados RDF em formato XML (POWERS, 2003).

Abaixo é apresentado um exemplo de arquivo RDF/XML utilizando as sentenças empregadas no exemplo anterior.

```
<?xml version="1.0"?>
  <rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:ex1="http://ex1.com/elements/1.0">
    <rdf:Description rdf:about="http://ex1.com/documento1">
      <ex1:titulo>Documento 1</ex1:titulo>
      <ex1:autor>Individuo A</ex1:autor>
    </rdf:Description>
  </rdf:RDF>
```

3.4 Ontologia em Sistemas de Recuperação de Informação

Abordagens puramente sintáticas para navegação e buscas baseadas em palavras-chave têm se tornado cada vez mais inadequadas em um ambiente no qual mais e mais pessoas exportam informação (MENA et al., 2000).

A utilização de ontologia em um sistema de recuperação de informação permite a definição de conceitos relacionados entre si a respeito de um domínio específico (MENA et al., 2000; PARALIC; KOSTIAL, 2003). Com isto, torna-se possível calcular a similaridade entre documentos vinculados a tal domínio através da sua estrutura conceitual representada pela ontologia.

Da mesma forma um mecanismo de busca⁴ de um sistema de recuperação de informação pode utilizar essa mesma estrutura conceitual para encontrar conceitos relacionados àqueles especificados pelo usuário. No resultado serão incluídos documentos que possivelmente não conterão nenhum termo informado pelo usuário, porém o mecanismo de busca pode descobrir através da estrutura conceitual definida pela ontologia quais termos têm o mesmo ou quase o mesmo significado (PARALIC; KOSTIAL, 2003).

Para a efetiva utilização de um sistema de recuperação de informação baseado em ontologia é necessário que se realize um pré-processamento dos documentos que serão indexados e disponibilizados para consulta. Esse pré-processamento colhe todos os conceitos na ontologia que podem estar relacionados a cada documento e cria uma nova base de dados mais ampla contendo uma descrição

⁴ Um mecanismo de busca é uma parte de um sistema de recuperação de informação, isto é, um sistema de recuperação de informação pode conter um ou mais mecanismos de busca integrados.

conceitual (conjunto de conceitos obtidos) de cada documento (PARALIC; KOSTIAL, 2003).

No momento da consulta, os termos informados pelo usuário passam pelo mesmo processamento como se a consulta fosse uma pequena fração de um documento. Seus conceitos são igualmente recuperados, e uma nova consulta é criada. Essa nova consulta é então comparada aos documentos processados, e uma lista daqueles que atingirem um determinado grau de similaridade com a consulta que expressa a necessidade do usuário é apresentada.

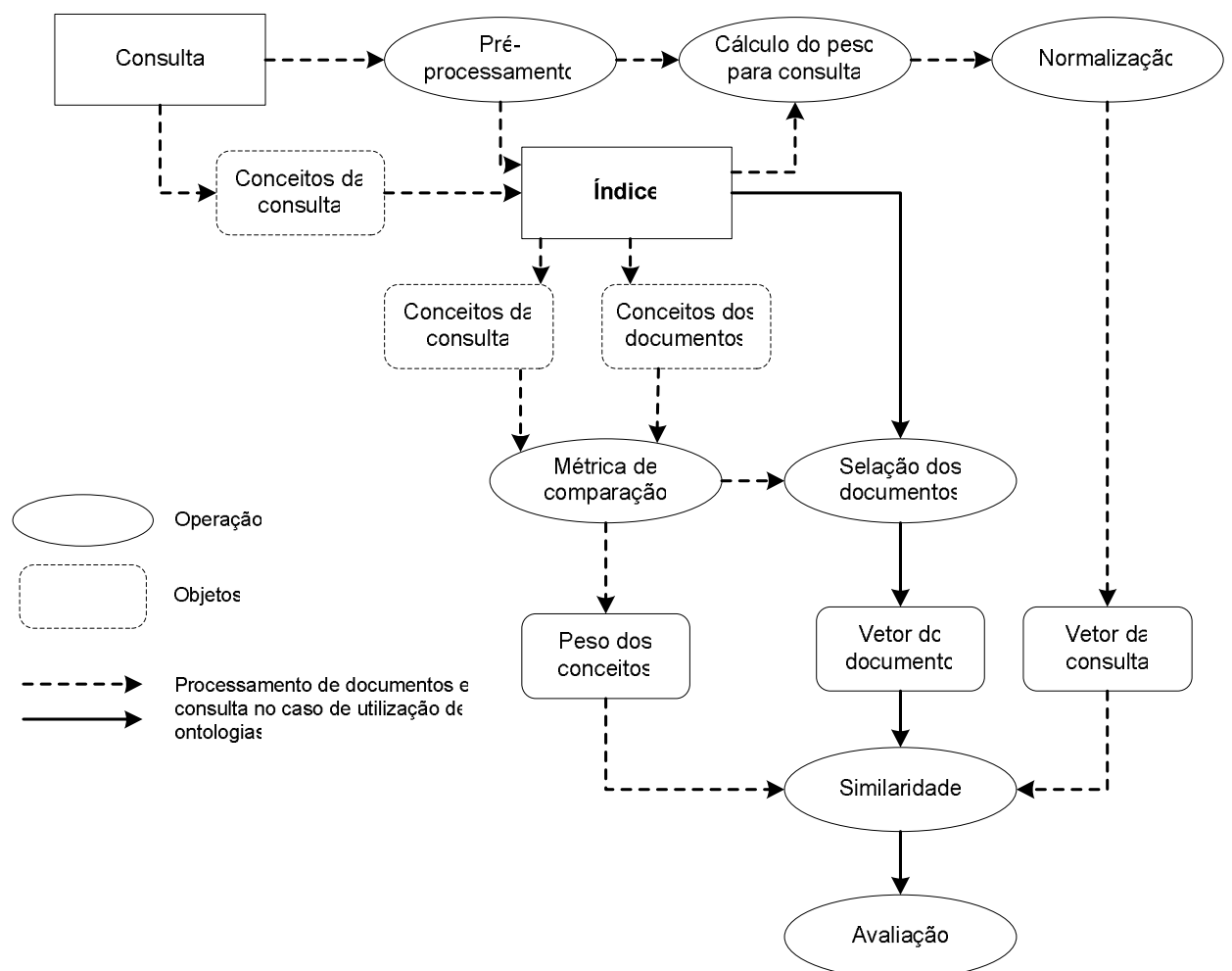


Figura 6 - Recuperação de informação baseada em ontologia

FONTE: Adaptado de: PARALIC; KOSTIAL, 2003.

Em um estudo realizado por Paralic e Kostial (2003), foram realizadas buscas utilizando três métodos: baseado em vetores de contexto, baseado em indexação semântica latente (LSI) e baseado em ontologia. Foram obtidos bons resultados com a utilização dessa abordagem. Porém, nesse estudo os conceitos relacionados foram manualmente atribuídos aos termos de busca.

3.5 Considerações Finais

Este capítulo apresentou uma visão geral sobre Web Semântica e ontologias. Exibiram-se breves definições do que é a Web Semântica e sua arquitetura e o que são ontologias, incluindo seu significado na filosofia, sua importância nas áreas de Inteligência Artificial e Sistemas de Informação e sua utilização em recuperação de informação, que é o foco deste trabalho.

No próximo capítulo será apresentada a arquitetura que se propõe a fim de tornar mais flexível e dinâmica a utilização de ontologias em sistemas de recuperação de informação e a classificação e apresentação dos resultados.

4 ARQUITETURA PARA UTILIZAÇÃO DE ONTOLOGIAS EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO

4.1 Introdução

Diante do que foi exposto nos capítulos anteriores a respeito dos sistemas de recuperação de informação e de ontologias, este capítulo apresenta a arquitetura proposta para utilização de ontologias em sistemas de recuperação de informação, incluindo o tratamento necessário ao resultado para a sua ordenação de forma que os documentos relevantes sejam bem posicionados quando apresentados ao usuário.

4.2 Visão Geral

Uma das grandes dificuldades que os usuários de sistemas de recuperação de informação têm atualmente quando estão diante de uma interface de consulta é saber o que informar para se obter o resultado esperado.

Se tomarmos como exemplo o termo “governo eletrônico”, poderiam também ser utilizados os termos “e-gov”, “e-governance”, “eletronic governance” ou “e-democracy” (PERRI, 2001).

Nesse exemplo são quatro alternativas para o termo inicial, e ainda podem existir outros com menor similaridade semântica, porém com algum nível de relacionamento relevante. O usuário seria obrigado a realizar cinco consultas

diferentes ou informar todos esses termos de uma só vez, isso se lembrasse e/ou conhecesse todas essas alternativas.

Desta forma, é possível que algum documento relevante não seja recuperado, uma vez que possui apenas o termo “e-democracy” e não os demais. O usuário, porém, não informou esse termo por algum motivo, por desconhecimento ou esquecimento.

O objetivo deste trabalho é atuar nesse problema auxiliando o usuário na realização de buscas mais abrangentes, visto que se propõe a desenvolver um modelo de integração dinâmica de ontologias a sistemas de recuperação de informação através da modificação do termo de consulta de maneira automática, por meio da sugestão de termos baseados na ontologia acoplada e selecionada para a busca.

4.3 Arquitetura Proposta

4.3.1 Apresentação da Arquitetura

Um sistema de recuperação de informação padrão é composto de três módulos, sendo eles: a interface com o usuário, o mecanismo de busca e o repositório dos dados e índices onde é realizada a busca. O modelo apresentado neste trabalho propõe o acoplamento de um módulo adicional que deve estar localizado entre a interface com o usuário e o mecanismo de busca.

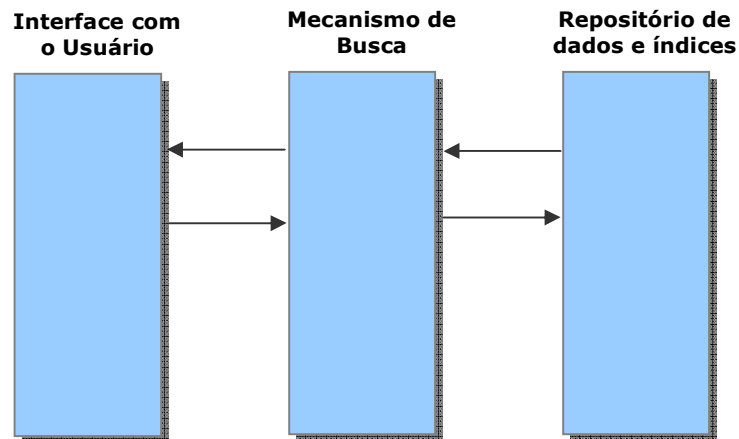


Figura 7 - Arquitetura conceitual de um sistema de recuperação de informação

Na Figura 7 é apresentada a arquitetura simplificada de um sistema de recuperação de informação padrão. Percebe-se que a interface com o usuário comunica-se diretamente com o mecanismo de busca, passando para este os termos informados pelo usuário. Isto significa que o usuário deve saber exatamente qual o conjunto de termos representa melhor aquilo que procura. Caso não informe um determinado termo relevante sobre o assunto, todos os documentos que possuírem tal termo não serão apresentados no resultado da pesquisa.

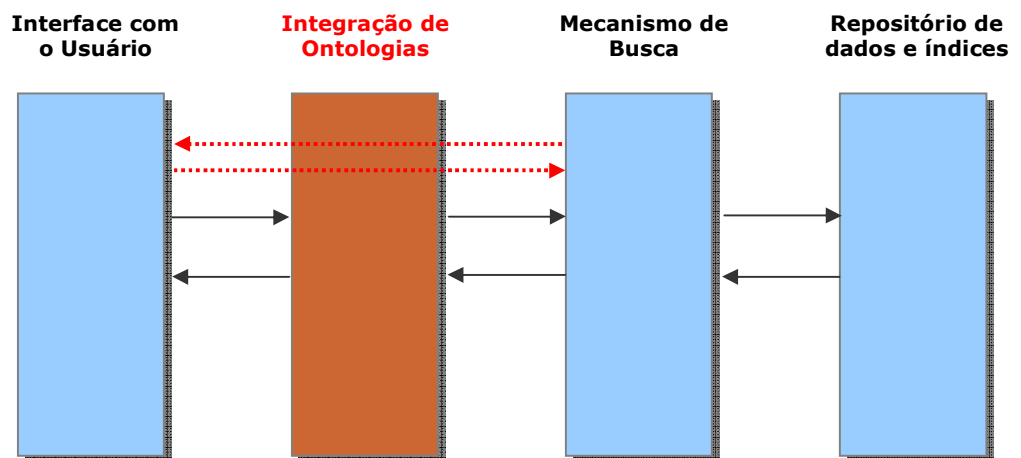


Figura 8 - Arquitetura conceitual de um sistema de recuperação de informação utilizando ontologias

Com o intuito de aprimorar os resultados obtidos por um sistema de recuperação de informação, que a princípio não conhece a necessidade real de informação do usuário, propõe-se que sejam integrados ao sistema os módulos que compõem essa arquitetura (Figura 8). Tais módulos permitem ao sistema inferir os termos informados pelo usuário e dessa maneira conhecer de forma automática mais detalhes a respeito do assunto demandado pelo usuário, sugerindo automaticamente alternativas para a consulta feita. Também é possível realizar a classificação dos resultados para a apresentação ao usuário de tal forma que os documentos mais relevantes sejam exibidos no início da lista. Tais resultados serão discutidos mais adiante quando da aplicação do modelo a um caso real.

A arquitetura proposta pode ser aplicada a qualquer sistema de recuperação de informação, possibilitando a utilização de quaisquer ontologias, ou seja, à medida que surjam novas ontologias ou se faça necessário a expansão do sistema para novos domínios, basta adequá-la à arquitetura para que estejam disponíveis para utilização por parte do sistema de recuperação de informação.

Portanto, a arquitetura proposta está baseada na expansão automática do vetor de consulta através do auxílio de ontologias, que são acopladas, habilitadas ou desabilitadas dinamicamente no sistema de recuperação de informação, que, em conjunto com a correta classificação dos resultados, tem o intuito de atingir um maior índice de “recall” e precisão.

Compõem essa arquitetura três módulos que interagem com o sistema de recuperação de informação, sendo: Módulo de Acoplamento de Ontologias, Módulo de Expansão do Vetor de Consulta e Módulo de Classificação dos Resultados.

Na Figura 9 são apresentados a arquitetura proposta, os módulos que a compõem e os relacionamentos entre os módulos, e o sistema de recuperação de informação.

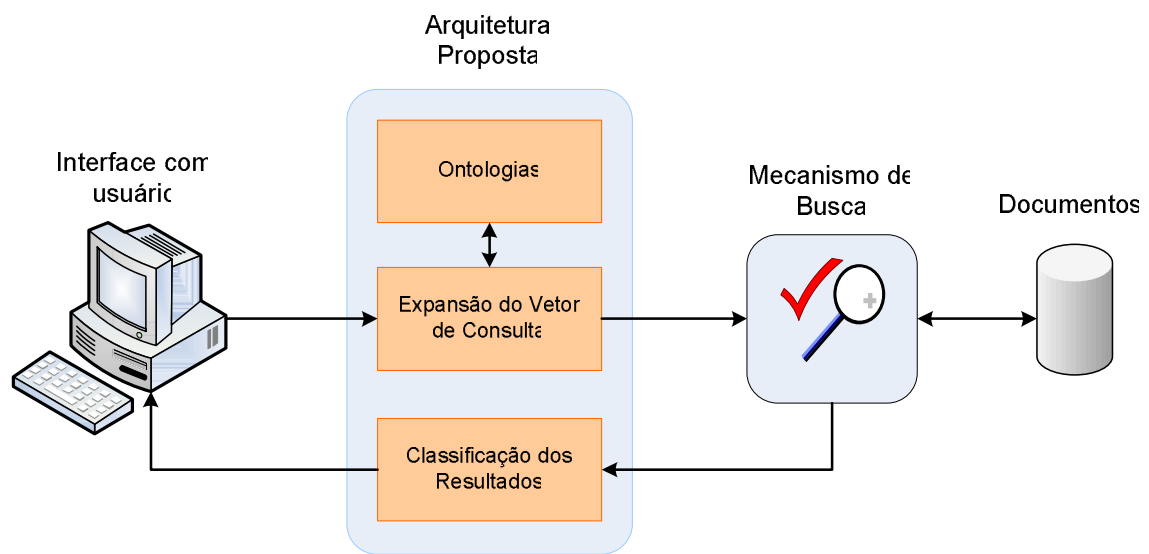


Figura 9 - Arquitetura proposta no contexto de um sistema de recuperação de informação

A seguir serão detalhados os módulos e seus componentes, iniciando-se com o Módulo de Acoplamento de Ontologias.

4.3.2 Módulo de Acoplamento de Ontologias

A utilização de ontologias em sistemas de recuperação de informação pressupõe que essas estejam disponíveis para consulta quando necessário. Para tanto, é previsto na arquitetura um módulo para acoplamento de tais ontologias. Também é importante que esse módulo facilite o acoplamento de novas ontologias sem significativo esforço. Para que isso seja possível, três componentes são essenciais:

a especificação de uma representação formal e padronizada para as ontologias, um componente para a adaptação das ontologias a essa representação formal e um componente para realizar o acoplamento dessas ontologias ao sistema de recuperação de informação.

Na Figura 10 apresentam-se o módulo e seus componentes.

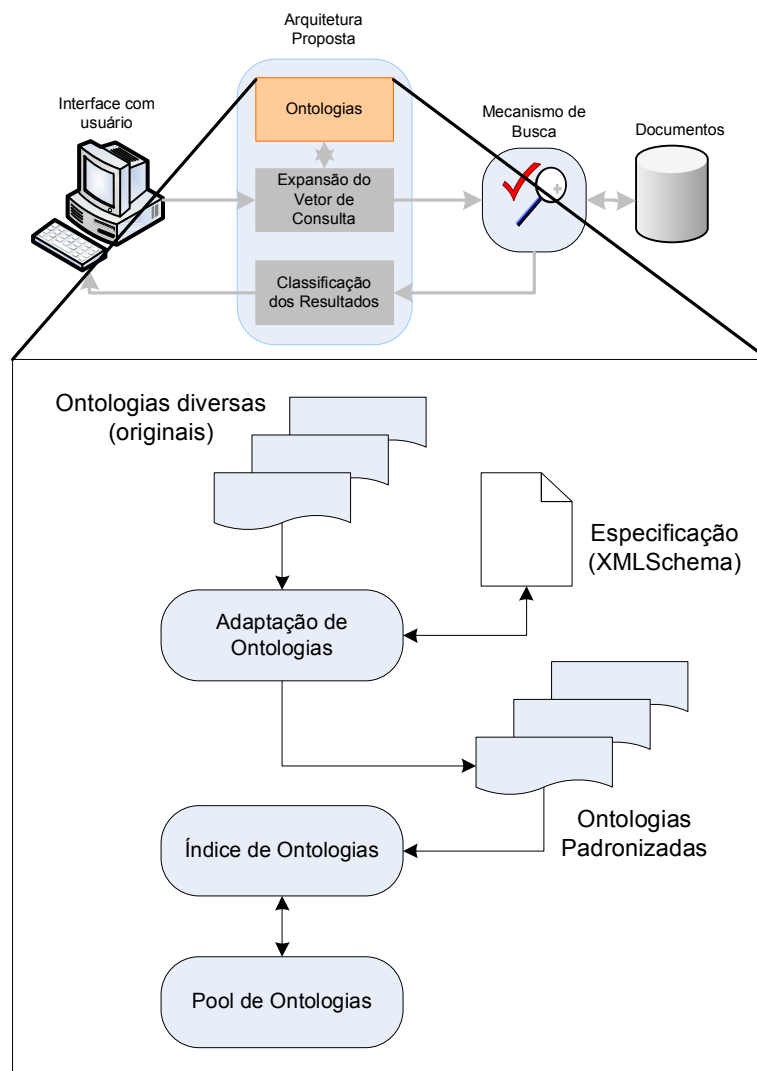


Figura 10 - Módulo de Acoplamento de Ontologias e seus componentes

a. Especificando uma representação formal

A especificação de uma representação padronizada permite que ontologias das mais diversas fontes ou domínios sejam acopladas dinamicamente ao sistema sem qualquer esforço adicional.

Sugere-se que essa especificação seja descrita em uma linguagem de descrição de informação, como XML ou ainda RDF/XML. RDF/XML é a recomendação do W3C para a modelagem de ontologias, além disso, possibilitaria a um sistema qualquer realizar inferências semânticas sobre a ontologia, uma vez que é possível descrever relacionamentos mais complexos utilizando-se RDF do que XML. Inicialmente este trabalho propõe uma representação em XML.

Para cada recurso na ontologia é realizado o levantamento de um conjunto mínimo de informações para suprir os requisitos necessários para a sua utilização em um sistema de recuperação de informação através da arquitetura que está sendo proposta.

- Identificador único
- Variações Idiomáticas do Termo
- Sinônimos do Termo (para cada idioma)
- Localização na hierarquia de áreas
- Nível de profundidade na hierarquia

Na Tabela 3 é apresentado um exemplo do levantamento das informações citadas para o termo “Malária”.

Tabela 3 - Informações necessárias para cada recurso descrito em uma ontologia

Identificador	Termo em Português	Termo em Inglês
000001	MALÁRIA	MALARIA
Sinônimos	INFECCÕES POR PLASMODIUM	PLASMODIUM INFECTIONS
	DOENCA MALÁRICA	
	INFECCÃO MALÁRICA	
Hierarquia	C03.752.250.552	Nível 4

Após realizado o levantamento das informações que serão consideradas, a especificação, no caso da utilização de uma representação em XML, é descrita em XML Schema. Na Figura 11 é apresentada uma proposta de especificação que contempla todas as informações necessárias e os seus relacionamentos.

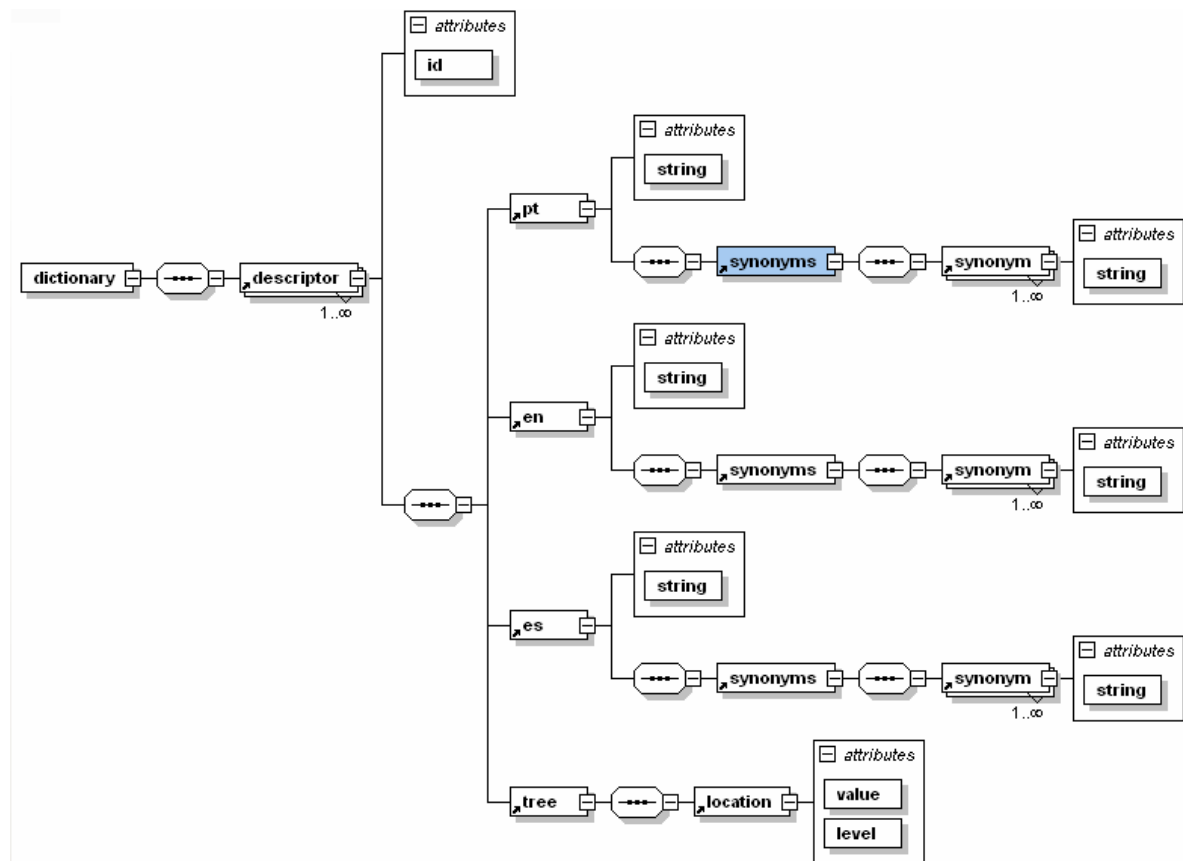


Figura 11 - Especificação de uma representação formal para ontologias em XML Schema

A partir da especificação definida, é possível criar uma representação em XML para as ontologias. Abaixo é apresentado um trecho de um arquivo XML contendo a representação do termo “Malária” com base no XML Schema definido anteriormente.

```

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <dictionary>
- <descriptor id="0000001">
- <pt string="MALARIA">
- <synonyms>
  <synonym string="INFECCOES POR
    PLASMODIUM" />
  <synonym string="DOENÇA MALARICA" />
  <synonym string="INFECCAO MALARICA" />
</synonyms>
</pt>
- <en string="MALARIA">
- <synonyms>
  <synonym string="PLASMODIUM
    INFECTIONS" />
</synonyms>
</en>
- <es string="MALARIA">
- <synonyms>
  <synonym string="INFECCIONES POR
    PLASMODIUM" />
  <synonym string="PALUDISMO" />
</synonyms>
</es>
- <tree>
  <location value="C03.752.250.552" level="4" />
</tree>
</descriptor>

```

Figura 12 - Fragmento de um arquivo XML representando o termo “Malária”

b. Componente para adaptação das ontologias

Tendo definido a representação das ontologias para utilização no sistema, é necessária a construção de um componente que realize a tarefa de adaptar as ontologias, nos seus mais diversos formatos, dos diferentes domínios do conhecimento a essa representação.

Primeiramente é realizado um mapeamento entre as informações existentes na ontologia utilizada e a representação que foi especificada anteriormente, ou seja,

qual informação na ontologia pode ser utilizada como “identificador único” de um termo, qual informação representa o “termo em português”, qual informação representa o “termo em inglês”, qual informação pode ser utilizada como sua localização na “hierarquia”, e assim, sucessivamente, com todas as informações especificadas anteriormente.

A Figura 13 demonstra um exemplo de mapeamento realizado entre uma ontologia representada em arquivo-texto e a mesma ontologia representada segundo a especificação proposta por essa arquitetura.

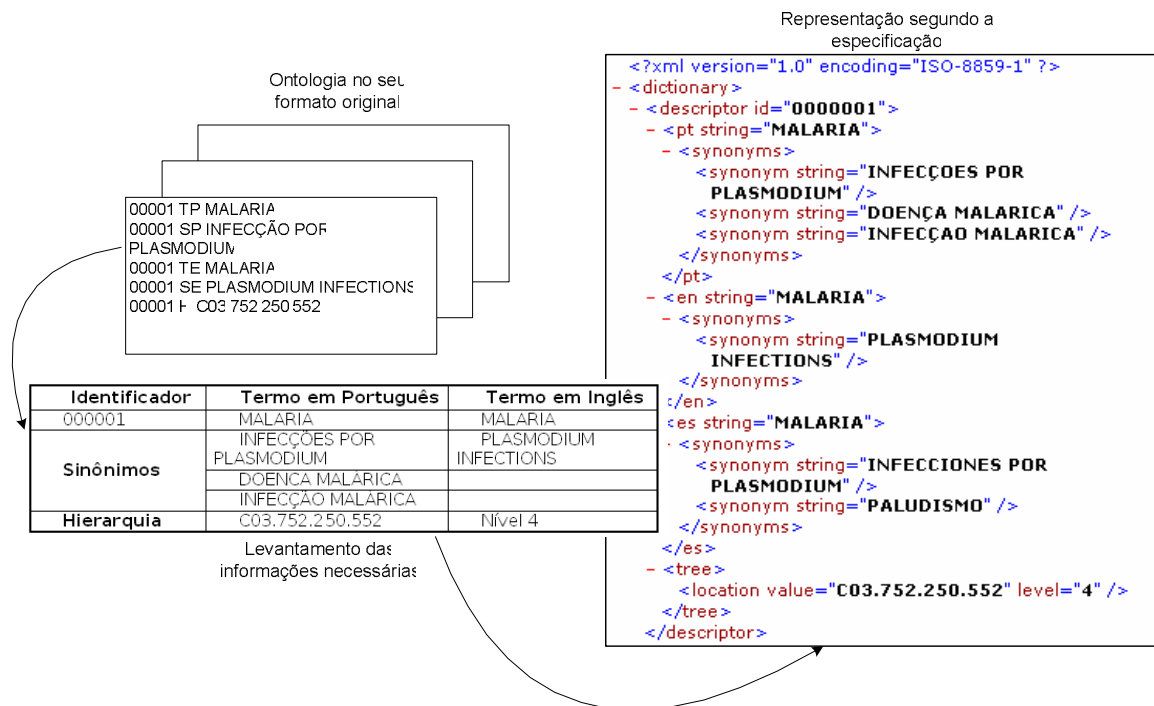


Figura 13 - Mapeamento entre a ontologia original e a representação especificada

Esse mapeamento é importante pois fornece subsídios para a construção de um componente que deve ler a ontologia em sua representação original, coletar os

dados necessários e repassá-los ao componente designado para serializar⁵ a ontologia segundo a especificação padrão proposta anteriormente em um arquivo XML (Figura 14).

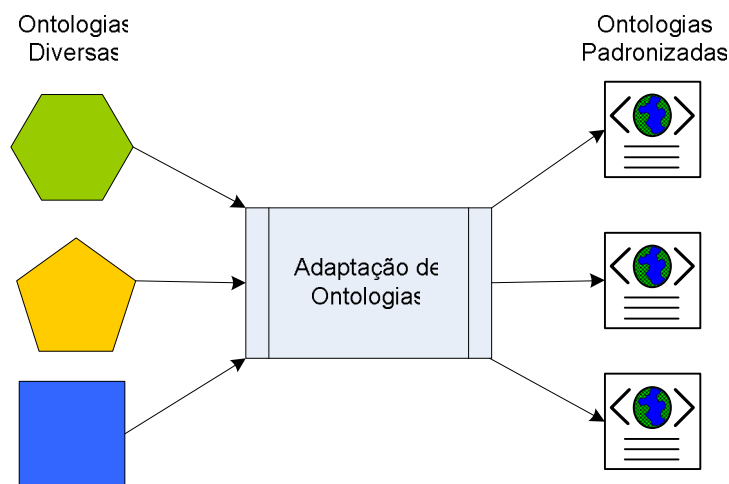


Figura 14 - Componente para adaptação das ontologias

A recomendação para separar essa etapa em dois componentes, um que lê a ontologia original e outro que gera a ontologia em sua nova representação, dá-se pelo fato de que isso facilita o trabalho de adaptação de uma nova ontologia, uma vez que evita que a parte relativa à geração da ontologia na nova representação tenha que ser construída novamente para cada nova ontologia que se deseja acoplar ao sistema. Esse componente de escrita simplesmente recebe uma entrada padrão e gera uma saída padrão. Apenas o componente de leitura deverá ser reescrito, pois toda a lógica de saída estará isolada em um outro componente.

A Figura 15 mostra o relacionamento entre esses componentes.

⁵ Escrever de forma serial, neste contexto, significa escrever um objeto em memória para um meio permanente de forma serial (THE FREE DICTIONARY, 2004).

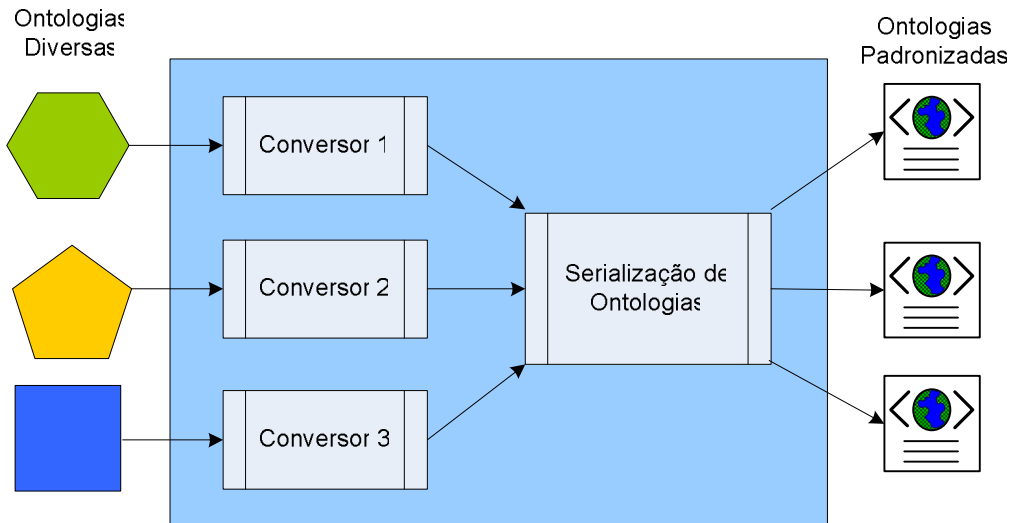


Figura 15 - Componentes de adaptação das ontologias

Após essa etapa, é possível obter diversas ontologias padronizadas para que possam ser acopladas ao sistema.

c. Acoplamento ao sistema

Para possibilitar o acoplamento das ontologias ao sistema, faz-se necessário a construção de um componente que realize tal tarefa. Na Figura 16 é apresentado um diagrama do funcionamento desse componente.

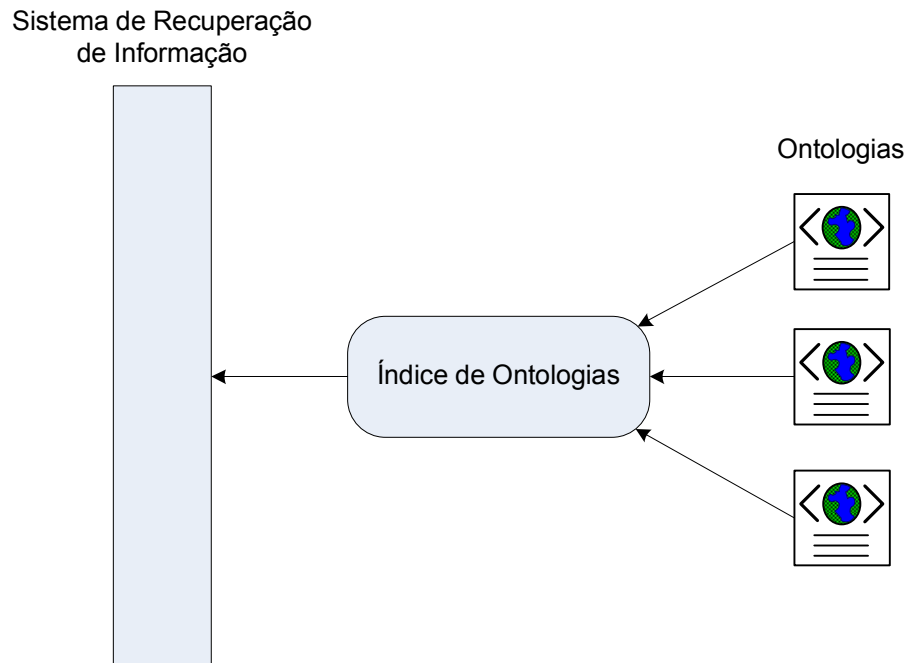


Figura 16 - Índice de Ontologias

São três os requisitos para esse componente: (a) conhecer a localização das ontologias e disponibilizá-las para o sistema de modo a possuir uma representação dessa localização; (b) um componente, que será o meio de manipulação da ontologia por parte do sistema; e (c) assegurar que essas ontologias sejam disponibilizadas de forma a não afetarem o desempenho do sistema, uma vez que muitas delas podem ser extremamente grandes após serializadas para o formato XML.

Para o primeiro requisito é proposta a construção de um índice de ontologias no qual o componente poderá procurar por aquela ontologia solicitada pelo sistema. O componente recebe uma requisição por uma ontologia, verifica nesse índice a localização e disponibiliza a ontologia para a utilização por parte do sistema. Sugere-se que esse índice seja criado em formato XML, contendo uma lista das ontologias disponíveis e suas localizações, tendo o componente a tarefa de ler esse arquivo XML para conhecer a localização das ontologias.

O segundo requisito é a representação da ontologia em uma forma não serializada, ou seja, um componente que possa ser utilizado pelo sistema de recuperação de informação. Esse componente receberá um termo como “malária” e deverá recuperar os sinônimos do termo segundo a ontologia e o idioma desejado.

Para o terceiro requisito é importante criar uma estratégia de carregamento das ontologias para a memória do sistema de modo a evitar perda de desempenho. Portanto, sugere-se a criação de um *pool*⁶ de ontologias que seja carregado em um momento em que não afete a utilização do sistema, como na inicialização dos serviços que disponibilizam o sistema de recuperação de informação em questão.

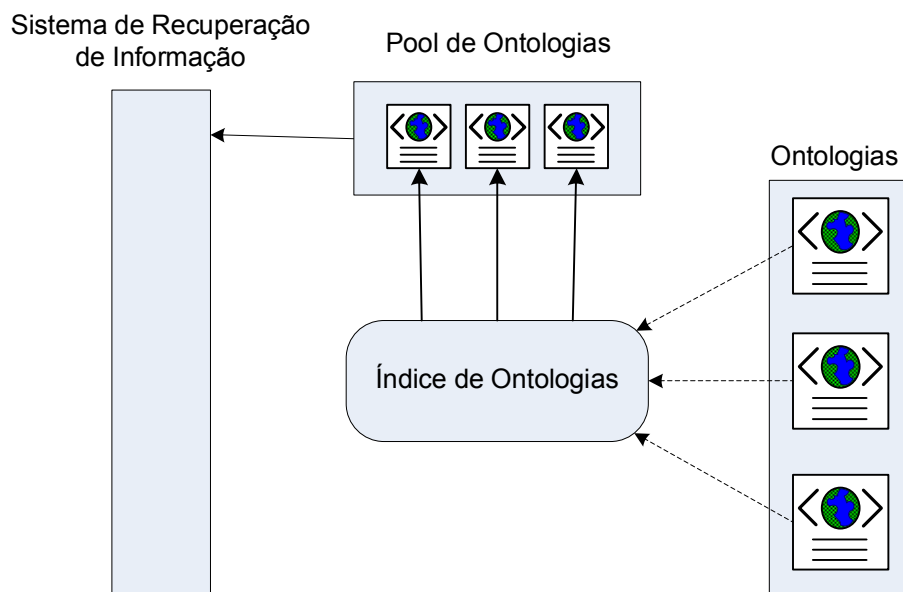


Figura 17 - Pool de Ontologias

Assim que esse componente de acoplamento de ontologias é disponibilizado para o sistema, torna-se mais fácil acoplar uma nova ontologia, bem como o resultado para o usuário tornar-se-á satisfatório, já que o desempenho não é afetado

⁶ Um agrupamento de recursos para a utilização comum por parte dos vários atores (THE FREE DICTIONARY, 2004).

e várias ontologias podem ser utilizadas, permitindo a seleção de uma ou outra, dependendo do domínio com o qual o usuário está interagindo.

4.3.3 Módulo de Expansão do Vetor de Consulta

Uma vez as ontologias devidamente disponibilizadas para o sistema de recuperação de informação, é necessário trabalhar na expansão da consulta informada pelo usuário. Essa é uma etapa relativamente simples, sendo a maior parte do trabalho realizado pelo módulo de acoplamento de ontologias. A Figura 18 apresenta o módulo de expansão do vetor de consulta, baseado em ontologia, no contexto da arquitetura proposta por este trabalho.

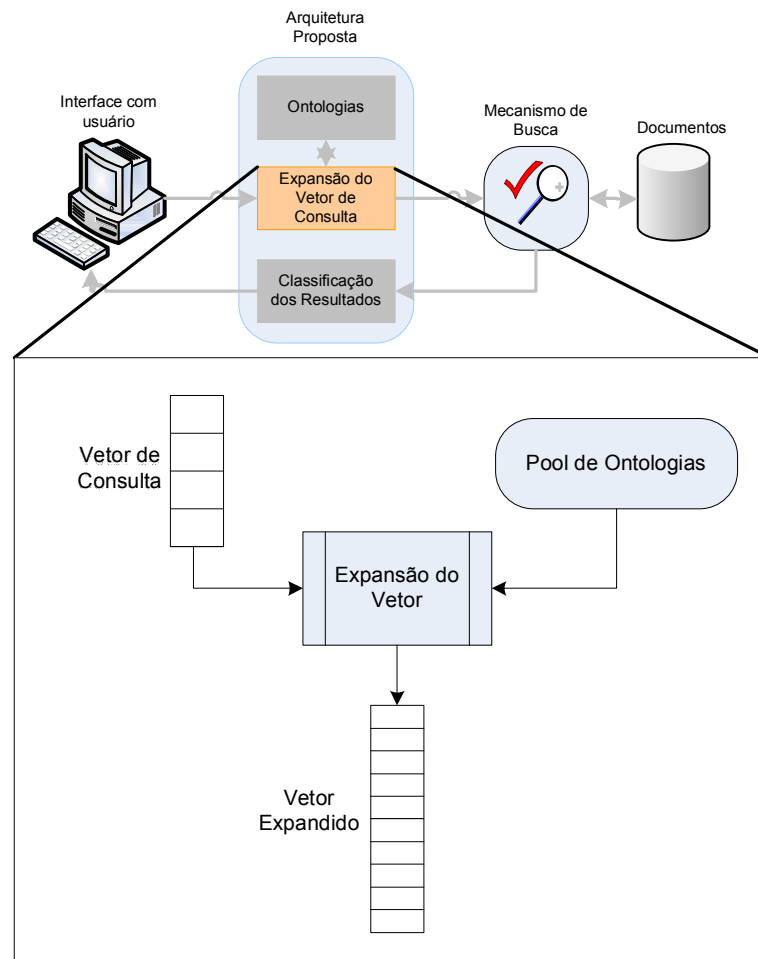


Figura 18 - Módulo de expansão do vetor de consulta no contexto da arquitetura proposta

Basta agora que esse componente de expansão do vetor de consulta faça uma consulta à ontologia segundo os termos informados no vetor de consulta do usuário. A Figura 19 mostra um exemplo do comportamento desse componente.

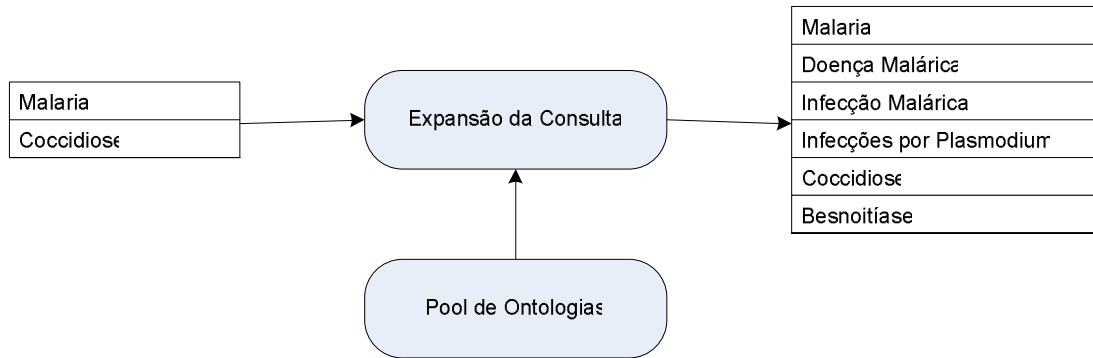


Figura 19 - Vetor de consulta expandido

Recuperados os termos sinônimos, deve-se reconstruir a consulta que será enviada ao mecanismo de busca, segundo a sintaxe especificada para o mecanismo de busca utilizado no sistema.

4.3.4 Apresentação do Resultado (Classificação)

Completando a arquitetura, apresenta-se o módulo que é responsável pela classificação do resultado obtido pelo sistema para uma determinada consulta do usuário.

A Figura 20 apresenta o módulo no contexto da arquitetura proposta.

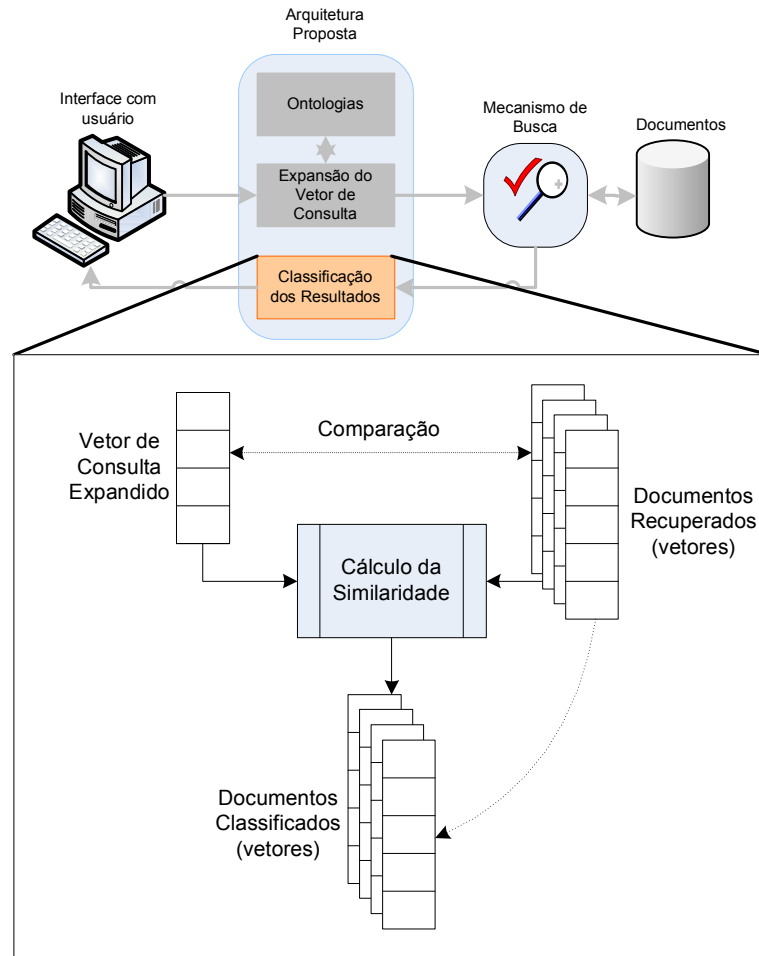


Figura 20 - Módulo de Classificação dos Resultados

Para o desenvolvimento desse módulo é necessária a construção de um componente que realize o cálculo de similaridade através da medida do co-seno, calculando a distância entre cada vetor recuperado e o vetor de consulta. Tais vetores devem ser compostos dos termos e dos respectivos pesos.

Para o peso, utiliza-se a medida TFIDF (equação 4) ou ainda a frequência de cada termo normalizada em relação ao termo de maior frequência no vetor (equação 2) para que se estabeleça a importância do termo em relação ao vetor em uma faixa de 0 a 1, conforme foi apresentado no tópico 2.4.

O vetor de consulta deve iniciar com o peso para cada termo igual a um, porém, para se aprimorar ainda mais o sistema de recuperação de informação, sugere-se a

construção de um componente que permita ao usuário modificar o peso de cada termo na consulta, levando a resultados mais refinados.

Possuindo os vetores dos documentos recuperados e o vetor da consulta normalizado, basta aplicar a equação do co-seno entre o vetor da consulta e o vetor de cada documento para se obter o grau de similaridade existente entre ambos através da equação 1. A Figura 21 apresenta exemplos de vetores e o peso para cada termo.

Malaria	1	Malaria	0,752
Doença Malárica	1	Febre Amarela	0,748
Infecção Malárica	1	Coccidiose	0,621
Infecções por Plasmodium	1	Virologia	0,410
		Avian	0,247
		Besnoitíase	0,012

Coseno = 0,286

Figura 21 - Exemplos de vetores com peso para cada termo

4.4 Considerações Finais

Este capítulo apresentou a arquitetura proposta para a utilização de ontologias e a classificação dos resultados em sistemas de recuperação de informação. A arquitetura é composta de três módulos que se acoplam a um sistema de informação de recuperação, visando melhorar os resultados de uma busca por informação. O primeiro módulo realiza o acoplamento das ontologias ao sistema com o intuito de facilitar novas ontologias a serem utilizadas pelo sistema. O segundo módulo possibilita a expansão do vetor de consulta informado pelo usuário pela utilização das ontologias disponíveis ao sistema, proporcionando uma busca mais abrangente

e considerando termos relacionados que o usuário não conhece ou não lembrou no momento em que formulou a consulta.

Por fim, após a realização da busca, o módulo de classificação dos resultados procura ressaltar os documentos mais relevantes em relação ao vetor de consulta expandido, posicionando-os no topo da lista de resultados para que o usuário rapidamente acesse aqueles que mais lhe interessam.

No próximo capítulo essa arquitetura é aplicada a um site de recuperação de informação da Rede ScienTI e os resultados obtidos são discutidos indicando as principais diferenças entre a utilização e a não utilização da arquitetura proposta.

5 APLICAÇÃO DA ARQUITETURA: SITE SCIENTI SAÚDE

5.1 Introdução

Como a arquitetura proposta foi aplicada ao site de recuperação de informação da Rede ScienTI na área da saúde, com a utilização de uma ontologia da área da saúde, este capítulo apresenta uma breve introdução sobre a recuperação de informação em plataformas de ciência e tecnologia (C&T), além dos passos seguidos para o acoplamento dinâmico da ontologia ao mecanismo de busca utilizado. Não será discutido o mecanismo de indexação e busca utilizado, somente a implementação da arquitetura apresentada.

5.2 Recuperação de Informação em C&T

Com o desenvolvimento tecnológico alcançado nos últimos anos, muito se tem evoluído também em matéria de publicação de pesquisas no Brasil. Cada vez mais têm sido comum a publicação de documentos técnico-científicos na Internet e a interoperabilidade entre fontes de informação heterogêneas e globalmente distribuídas (MARCONDES; SAYÃO, 2001). São diversas as bibliotecas digitais existentes hoje que proporcionam ao pesquisador brasileiro mais visibilidade nacional e internacional, otimizando o fluxo da comunicação científica e reduzindo o ciclo de geração de novos conhecimentos. A Biblioteca Digital Brasileira do IBICT, o Prossiga, o Scielo, o repositório de teses da Universidade de São Paulo, o arquivo e-prints do Impa e o Banco de Teses e Dissertação do Programa de Pós-Graduação

em Engenharia de Produção da Universidade Federal de Santa Catarina são exemplos dessas bibliotecas (MARCONDES; SAYÃO, 2001; ANDRADE, 2003).

Além disso, a área de E-Gov talvez seja a que mais acentua a necessidade de metodologia integrada. Embora já sejam significativos os avanços na organização, publicação e produção de informações em sites de governo, muitos desafios a projetos de informação governo–sociedade permanecem recorrentes em todo novo desenvolvimento (PACHECO, 2003).

Pacheco e Kern (2001) propõem uma arquitetura conceitual para projetos de E-Gov na qual os sistemas de recuperação de informação têm a sua devida importância. Na Figura 22 é apresentada essa arquitetura na qual se nota a sua representação piramidal, que parte de uma base composta de unidades de informação da plataforma e segue por camadas de padronização, sistematização e publicação de informações e serviços, até chegar ao topo, reservado à gestão, produção e publicação de conhecimento no projeto.

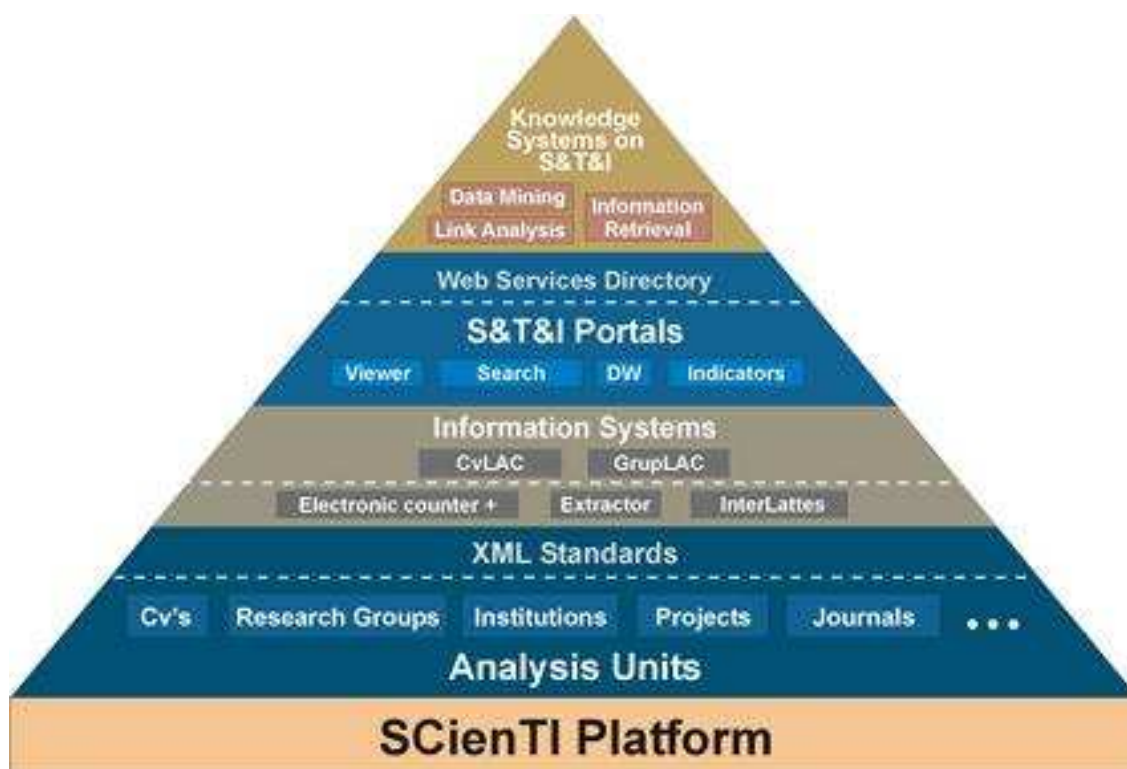


Figura 22 - Arquitetura conceitual para projetos de E-Gov

FONTE: PACHECO, 2003.

No contexto deste trabalho, as ontologias estão estabelecidas no nível da primeira camada, a camada das unidades de informação, por se refletirem ao longo de toda a pirâmide, enquanto que os sistemas de recuperação de informação se estabelecem no topo da pirâmide onde estão os sistemas de conhecimento, possuindo intersecção com a camada de Web Services, onde é publicada ao usuário a informação de corrente das camadas anteriores.

Um exemplo de aplicação dessa metodologia proposta por Pacheco e Kern (2001) é a Plataforma Lattes, um projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) desenvolvido para apoiar as atividades de Ciência, Tecnologia e Inovação (CT&I) do Brasil. A Plataforma Lattes é atualmente formada de bases de dados, sistemas de informação, diretórios de serviços e portais Web,

inovando na maneira de se verem as informações em C&T no Brasil (SABBATINI, 2001).

A Plataforma Lattes tem evoluído promovendo a padronização de cada unidade, a construção e a divulgação de serviços de informação, além do recente intercâmbio internacional para a comunidade científica do Brasil, uma vez que é parte integrante da Rede ScienTI (PACHECO, 2003).

A Rede ScienTI é constituída de organismos de C&T de mais de dez países, grupos de P&D em ciências da informação, organismos internacionais, entre outros, e possui informações sobre currículos de pesquisadores, diretórios de grupos e instituições, agências de fomento, projetos e produção científica que constituem a Plataforma ScienTI, destacando-se a plataforma de currículos de diversos países baseada na Plataforma Lattes do CNPq do Brasil (PELLEGRINI, 2004).

Para a aplicação dessa arquitetura, utilizou-se a base de currículos da área de saúde de pesquisadores do Brasil e da Colômbia.

5.3 Aplicação da Arquitetura

Como detalhado no Capítulo 4, a arquitetura é composta de três módulos: Módulo de Acoplamento de Ontologias, Módulo de Expansão do Vetor de Consulta, Módulo para Classificação do Resultado, os quais foram implementados no site de buscas da Rede ScienTI e são descritos a seguir.

5.3.1 Acoplamento de Ontologias

a. Especificação da Representação

O primeiro passo foi a especificação de uma representação para as ontologias que serão utilizadas no site. A partir do levantamento das informações que são utilizadas no processo de expansão do vetor de consulta, pode-se chegar a uma especificação. As informações necessárias são:

- identificador único (*id*);
- variações idiomáticas do termo (*pt, en, es*);
- sinônimos do termo (para cada idioma) (*synonyms*);
- localização na hierarquia de áreas (*location*);
- nível de profundidade na hierarquia (*level*).

Essas informações são necessárias para cada descritor (*descriptor*). A partir daí, criou-se uma estrutura hierárquica representando a ontologia. Nessa aplicação, a representação foi especificada em XML Schema, conforme pode ser vista na Figura 11 da seção 4.3.2 .

A Tabela 4 apresenta a lista dos elementos utilizados na especificação e uma breve descrição sobre esses elementos.

Tabela 4 - Descrição dos elementos da representação de ontologias

Elemento/Atributo	Descrição
dictionary	Dicionário completo
descriptor	Todas as informações sobre um descritor
Id	Identificador único
Pt	Versão em português do termo
en	Versão em inglês do termo
es	Versão em espanhol do termo
synonyms	Conjunto de sinônimos do termo
synonym	Um sinônimo do termo
string	Valor representando o termo ou sinônimo
tree	Posicionamento na árvore hierárquica
location	Informações sobre o local na hierarquia
value	Atributo com o valor sobre o local na hierarquia
level	Nível de profundidade na hierarquia

As seguintes regras se aplicam à adaptação das ontologias para a especificação apresentada:

- 1 **dictionary** possui 0 ou mais **descriptors**;
- 1 **descriptor** existe em 1 ou mais **idiomas (pt, en, es)**;
- 1 **idioma (pt, en, es)** possui 0 ou 1 conjunto de **synonyms** que contém 1 ou mais **synonyms**;
- 1 **descriptor** possui um conjunto de informações sobre **tree**, que contém 1 ou mais **location**.

Portanto, pode-se dizer que existe uma relação de sinonímia entre dois termos:

{“malária”, **sinônimo**, “infecção por plasmodium”}

b. Adaptação da Ontologia

Assim que a especificação da representação das ontologias foi definida, foi possível criar o componente responsável pela adaptação das ontologias em seu formato original para o novo formato.

Nessa aplicação, a ontologia utilizada como exemplo foi o DeCS (Descritores em Ciências da Saúde) criado pela BVS (Biblioteca Virtual em Saúde) para uso na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos e outros tipos de materiais. O DeCS Foi desenvolvido a partir do MeSH (*Medical Subject Headings*) da US National Library of Medicine com o objetivo de permitir o uso de terminologia comum para pesquisa em três idiomas (BIREME, 2003).

O DeCS é originalmente distribuído em formato texto, portanto, é necessário adaptá-lo à representação definida para utilização na arquitetura que está sendo implementada. Primeiramente foi realizado um mapeamento das informações necessárias no formato em que o DeCS é distribuído para então ser construído o componente que faria a adaptação para a representação especificada.

O passo seguinte refere-se à construção de um componente utilizando a especificação apresentada na seção 4.3.2 . Esse componente varre os arquivos do DeCS e os transcreve para um arquivo XML segundo o XML Schema definido. A Figura 13 apresenta os passos desse mapeamento.

Abaixo é apresentado um trecho do DeCS adaptado à representação especificada para o termo “malária”.

```

<dictionary>
  <descriptor id="0000001">
    <pt string="MALARIA">
      <synonyms>
        <synonym string="INFECÇÕES POR PLASMODIUM" />
        <synonym string="DOENÇA MALARICA" />
        <synonym string="INFECÇÃO MALARICA" />
      </synonyms>
    </pt>
    <en string="MALARIA">
      <synonyms>
        <synonym string="PLASMODIUM INFECTIONS" />
      </synonyms>
    </en>
    <es string="MALARIA">
      <synonyms>
        <synonym string="INFECCIONES POR PLASMODIUM" />
        <synonym string="PALUDISMO" />
      </synonyms>
    </es>
    <tree>
      <location value="C03.752.250.552" level="4" />
    </tree>
  </descriptor>
</dictionary>

```

c. Acoplamento das Ontologias

Conforme apresentado no capítulo anterior, para o acoplamento das ontologias ao sistema é importante a criação de um índice das ontologias, já adaptadas, para que o sistema possa encontrá-las quando necessário.

Para realizar tal função, foi especificado um arquivo XML, conforme ilustra a Figura 23.

```

- <ScientiConfig campoBuscaTermo="Term" caminhoIndice="D:\\xml\\index">
- <CortesTematicos>
  <CorteTematico id="1" nome="Geral" chaveRecurso="busca.corte.label.geral" />
  - <CorteTematico id="2" nome="Saúde" chaveRecurso="busca.corte.label.saude">
    <Ontologia id="1" nome="DeCS" url="http://decs.bvs.br" path="D:\\decs\\OntologiaDecs.xml" />
  </CorteTematico>
</CortesTematicos>
</ScientiConfig>

```

Figura 23 - Representação do índice de ontologias em XML

Entre outras informações utilizadas pelo site da Rede ScienTI, existe o elemento denominado **CorteTematico**, que representa uma ontologia. Esse nome foi escolhido na intenção de se classificarem as ontologias por áreas do conhecimento, que, no caso específico dessa aplicação, é a área de saúde. No entanto, é possível configurar quantos cortes temáticos forem necessários, sendo que cada um poderá conter várias ontologias representadas, nesse caso, pelo elemento **Ontologia**. Para cada elemento **Ontologia** são atribuídos valores para os atributos **id**, **nome**, **url**, **path**. Tais elementos e atributos são detalhados na Tabela 5.

Tabela 5 - Elementos do índice de ontologias

Elemento/atributo	Descrição
CorteTematico	Conjunto de ontologias de uma mesma área
Ontologia	Uma ontologia que pode ser acoplada
id	Uma identificação única
nome	Nome da ontologia para apresentação ao usuário
url	Endereço Web para mais informações sobre a ontologia
path	Caminho absoluto para o arquivo que contém a representação em XML da ontologia

Desses elementos, o mais importante é o atributo **path**, que indica ao sistema onde encontrar a ontologia.

Além do índice de ontologias, existe um componente para gerenciar as ontologias em memória, ou seja, um *pool* de ontologias, no qual cada ontologia é representada por um objeto em memória.

Esse *pool* carrega as ontologias para a memória na primeira vez em que são utilizadas. Considerando que o DeCS possui atualmente 26.851 descritores (BIREME) e que para cada descritor existem informações adicionais, pode-se concluir que não convém ler a ontologia do arquivo a cada acesso. Com a utilização do *pool* de ontologias e os componentes para representação em memória de cada ontologia, segundo os testes realizados nessa aplicação, obtém-se um ganho

aproximado de 95% em desempenho, fator fundamental para qualquer sistema de recuperação de informação.

5.3.2 Expansão do Vetor de Consulta

Para a utilização efetiva das ontologias nessa aplicação, um componente para a expansão do vetor de consulta foi construído seguindo o modelo apresentado na Figura 19, ou seja, um componente que interpreta a consulta informada pelo usuário e para cada termo busca na ontologia utilizada (selecionada pelo usuário e disponibilizada pelo *pool*) os termos sinônimos. Assim, a consulta expandida é submetida ao mecanismo de busca.

Portanto, se o usuário informar o termo “aids”, o componente descrito efetuará uma busca na ontologia por esse termo e encontrará os termos sinônimos, incluindo-os em outros idiomas. A consulta que será submetida ao mecanismo de busca não será simplesmente “aids”, mas, segundo o DeCS, “aids sida síndrome da imunodeficiência adquirida...”. Nas seções 5.4 e 5.5 são apresentados e discutidos os resultados obtidos com a expansão da consulta.

5.3.3 Apresentação do Resultado (Classificação)

Após efetuada a busca aos documentos utilizando a consulta expandida por meio da utilização da ontologia selecionada, é necessário ordenar a lista de documentos que será apresentada ao usuário da melhor forma possível, de maneira que os documentos mais relevantes em relação à consulta sejam mais bem posicionados.

Isso possibilita ao usuário rapidamente ter acesso àqueles documentos que satisfaçam a sua necessidade de informação.

Para essa tarefa existe um componente que ordena a lista de documentos recuperados segundo o seu grau de similaridade em relação à consulta. Esse componente realiza uma varredura por todos os vetores dos documentos recuperados e para cada um calcula a medida do co-seno em relação ao vetor de consulta. Para a realização do cálculo, o peso de cada termo é primeiramente normalizado através da frequência relativa, ou seja, a frequência absoluta dividida pela máxima frequência (equação 2).

Com o valor que representa o grau de similaridade entre cada documento e o vetor de consulta, é realizada a ordenação da lista que contém tais documentos de forma decrescente, sendo 1 o valor máximo possível para essa medida.

Tabela 6 - Vetor de consulta

Termo	Peso
termo2	1
termo3	1

Tabela 7 - Vetor de documento com frequências absolutas e normalizadas

Termo	Frequência	Peso
termo1	10	0,833
termo2	8	0,667
termo3	4	0,333
termo4	2	0,167
termo5	12	1,000
termo6	4	0,333
Valor do co-seno	0,4575	

Na Tabela 6 é apresentado um exemplo de vetor de consulta contendo os termos informados pelo usuário e/ou expandidos através do uso da ontologia e o peso, que, por padrão, é 1. No presente trabalho, a aplicação não permite ao usuário alterar o

peso de cada termo, embora seja um recurso interessante, visto que o usuário pode refinar sua consulta segundo a importância de cada termo.

Já na Tabela 7 é apresentado um exemplo de vetor de um documento que possui diversos termos, incluindo os do vetor de consulta. O vetor possui ainda a frequência absoluta em que ocorrem o termo e a frequência normalizada em relação ao termo mais frequente desse mesmo vetor. A frequência normalizada indica o peso do termo em relação aos demais termos. Ainda na Tabela 7 está presente o valor do co-seno calculado em relação ao vetor de consulta apresentado na Tabela 6. Esse cálculo é efetuado para cada documento recuperado, e seu valor é posteriormente avaliado no momento da ordenação.

5.3.4 Site de Buscas da Rede ScienTI

A Figura 24 apresenta a tela inicial do site de buscas da Rede ScienTI, na qual foi aplicada a arquitetura apresentada neste trabalho cujo desenvolvimento foi apresentado no tópico 5.3. A base de dados do site é composta dos currículos Lattes de pesquisadores da área da saúde do Brasil e da Colômbia.

O site foi construído de forma a ser flexível o suficiente para que o usuário possa utilizá-lo conforme for mais conveniente, ou seja, é apresentada uma lista de áreas do conhecimento ou domínios em que se podem realizar buscas. À medida que uma área é selecionada, as ontologias existentes para essa área são apresentadas em outra lista. Além disso, é possível indicar se o vetor de consulta, informado em caixa de texto própria, deve ou não ser expandido para conter os termos semanticamente

relacionados. Também existe a possibilidade de se indicar de quais países devem ser considerados os currículos a serem pesquisados.



Figura 24 - Tela inicial do site de buscas da Rede SciencTI

Na Figura 25 pode-se ver a tela de apresentação de resultados do site onde são listados os currículos que foram encontrados e que possuem alguma similaridade com a consulta. Ao lado de cada currículo são exibidos dois valores, um indicando o valor do co-seno calculado com base no vetor de cada currículo e no vetor da consulta e outro indicando o escore baseado no cálculo do TFIDF sugerido pelo próprio mecanismo de busca utilizado. Ambos os valores são apresentados apenas

para possibilitar a comparação entre eles, sendo o valor da medida do co-seno já suficiente para indicar o grau de relevância de cada currículo recuperado.

Mais à esquerda na tela é apresentado o vetor de consulta com os termos informados pelo usuário em negrito e os demais termos obtidos a partir da utilização da ontologia.

Consulta expandida

The screenshot shows the search results for 'Saúde' on the Rede SciencTI website. The results are as follows:

Rank	Researcher Name	Similaridade em relação ao vetor de consulta (coseno)	Relevância do documento segundo TFIDF
1	Amanda Elena Maestre Buitrago Currículo Vitae	0,3780	0,08
2	Mary Luz López Pérez Currículo Vitae	0,3780	0,07
3	Marcos Restrepo Isaza Currículo Vitae	0,3780	0,07

Termos utilizados na pesquisa:

- malaria**
infecciones por plasmodium
doença malarica
plasmodium infections
infecciones por plasmodium
paludismo
infeccao malarica

Labels below the screenshot indicate the following elements:

- Documents recovered (Documents recuperados)
- Similarity in relation to the query vector (cosine) (Similaridade em relação ao vetor de consulta (coseno))
- Relevance of the document according to TFIDF (Relevância do documento segundo TFIDF)

Figura 25 - Tela de resultados do site de buscas da Rede SciencTI

5.4 Estudo Comparativo: Site SciencTI Saúde

Para verificar a viabilidade da arquitetura proposta, foram realizadas algumas consultas no site apresentado no tópico acima. A própria interface do site permite

decidir entre a utilização ou não dos termos relacionados semanticamente aos termos de consulta. Além disso, na tela de resultados, o site apresenta tanto a medida obtida do cálculo do co-seno como o valor score atribuído pelo mecanismo de busca utilizado.

Primeiramente serão avaliados os resultados da utilização ou não de ontologia no momento da consulta. Em seguida será avaliado o posicionamento dos documentos recuperados para cada medida.

Para se determinar a diferença no número de documentos recuperados com e sem a utilização de ontologias, foram realizadas 10 consultas cujos resultados são apresentados na Tabela 8.

Tabela 8 - Número de documentos recuperados para cada termo com e sem a utilização de ontologias para a expansão semântica do contexto da consulta

Termo	Com Ontologia	Sem Ontologia	Diferença
aids	108	129	19%
saúde pública	49	68	39%
cegueira	0	1	-
obesidade	22	25	14%
dna	15	20	33%
paludismo	0	67	-
hipotireoidismo	2	5	150%
ortopedia	9	11	22%
gravidez	10	13	30%
infecção	10	16	60%
		Média	46%

A Tabela 8 mostra que para todos os termos utilizados houve um acréscimo no número de documentos recuperados, o que pode indicar que documentos relevantes não são recuperados sem a utilização de ontologias. A menor diferença entre a utilização e a não utilização da abordagem de expansão da consulta através do uso de ontologias foi de 14%, sendo a média de diferença para os casos apresentados de 46%. O fato de recuperar mais documentos não significa que serão relevantes

para o usuário, mas aumenta a possibilidade de mais documentos relevantes serem recuperados. Neste caso a forma como o resultado será ordenado para apresentação ao usuário pode fazer diferença. Os resultados referentes à classificação dos resultados é exibida mais adiante.

Outro aspecto verificado diz respeito ao fato de terem sido utilizados currículos de pesquisadores brasileiros e colombianos. Quando não é realizada a expansão da consulta, se o termo informado estiver em português, a tendência é serem recuperados apenas currículos do Brasil, e, quando é realizada a expansão da consulta, currículos de pesquisadores colombianos também são recuperados. Isto acontece porque pesquisadores colombianos possuem a maioria dos termos em espanhol em seus currículos. E o fato de o DeCS ser trilingüe torna a consulta independente de idioma (português, inglês e espanhol).

Para a avaliação da ordenação do resultado obtido com a utilização da medida do co-seno para o cálculo de similaridade vetorial, foram realizadas três consultas.

A primeira consulta é composta do termo “aids”. Após a expansão da consulta, os termos submetidos ao mecanismo de busca foram (Tabela 9):

Tabela 9 - Vetor de consulta expandido para o termo “aids”

aids
acquired immunodeficiency syndrome
sida
síndrome de imunodeficiência adquirida
síndrome de deficiência imunológica adquirida
immunodeficiency syndrome, acquired
síndrome de imunodeficiência adquirida
síndrome de deficiência imunológica adquirida
immunologic deficiency syndrome, acquired

Entre os documentos recuperados, foram selecionados 10 itens para a análise do resultado, apresentados na Tabela 10 a seguir.

Tabela 10 - Documentos recuperados para o termo “aids” e seus sinônimos

Pesquisador	Co-seno	Ordem Co-seno	Escore⁷	Ordem Escore
Pesquisador 1	0.33333	1	0.03239	5
Pesquisador 2	0.16667	2	0.02450	10
Pesquisador 3	0.14907	3	0.03879	3
Pesquisador 4	0.14907	4	0.02980	6
Pesquisador 5	0.13608	5	0.02888	8
Pesquisador 6	0.12599	6	0.03634	4
Pesquisador 7	0.11785	7	0.04113	1
Pesquisador 8	0.11785	8	0.04113	2
Pesquisador 9	0.11785	9	0.02980	7
Pesquisador 10	0.07454	10	0.02584	9

A partir da Tabela 10 é possível perceber que as medidas do co-seno e escore não classificam os documentos recuperados da mesma forma. Toma-se como exemplo o pesquisador 1, que, pela medida do co-seno, ficou posicionado em primeiro lugar, porém pela medida do TFIDF caiu para quinto lugar. Através de uma análise empírica dos vetores de termos de cada pesquisador listado na Tabela 10, percebe-se que o pesquisador 1 parece ter mais relevância que todos os outros pesquisadores listados. Consta no vetor de termos do pesquisador 1 o termo “aids” com uma frequência igual a 18, sendo a frequência média do termo “aids” ou um de seus sinônimos nos demais vetores igual a 2. Também se nota que os demais termos do vetor do pesquisador 1 também são relacionados a “aids”, ainda que não sejam sinônimos diretos, o que não ocorre com tanta nitidez nos demais vetores.

Outro caso que merece destaque é o pesquisador 2, que, segundo o co-seno, está classificado em segundo lugar, porém pelo escore se posiciona em décimo lugar, ou seja, último na lista apresentada. Mais uma vez, ao analisar empiricamente o vetor desse pesquisador, nota-se uma maior frequência em termos relacionados a “aids”. Pelo fato de o vetor ser grande, com cerca de 180 termos, o escore ficou com

⁷ O escore apresentado é atribuído pelo próprio mecanismo de busca utilizado e é baseado no TFIDF.

um valor baixo, porém, através do cálculo do co-seno, foi possível posicioná-lo apropriadamente.

Através do co-seno mede-se o grau de similaridade entre o vetor de consulta e o vetor do documento, e o escore é calculado através de uma normalização da frequência de cada termo em relação à frequência do termo que conta com mais ocorrências (TFIDF).

A segunda consulta realizada é composta do termo “obesidade”. O vetor expandido é apresentado na Tabela 11.

Tabela 11 - Vetor expandido para o termo “obesidade”

obesidade
obesity
obesidad

Para esta consulta, foram selecionados seis documentos recuperados, os quais são apresentados na Tabela 12.

Tabela 12 - Documentos recuperados para o termo “obesidade” e seus sinônimos

Pesquisador	Co-seno	Ordem Co-seno	Escore	Ordem Escore
Pesquisador 1	0.57735	1	0.80698	1
Pesquisador 2	0.40825	2	0.26726	3
Pesquisador 3	0.28868	3	0.41181	2
Pesquisador 4	0.28868	4	0.26726	4
Pesquisador 5	0.23570	5	0.24006	6
Pesquisador 6	0.17408	6	0.25793	5

Neste caso, é possível observar o reposicionamento de alguns documentos quando ordenados pela medida do co-seno. Um exemplo que merece destaque é o pesquisador 3, que foi apresentado em 22º na listagem de todos os resultados, segundo a medida do co-seno, porém, segundo o escore, ficaria em 4º. Se o vetor do pesquisador 3 for comparado ao vetor do pesquisador 1, nota-se que há muito pouca similaridade, dado que o vetor do pesquisador 1 possui o termo “obesidad”

com uma freqüência igual a 127 e o pesquisador 3 possui o termo “obesidade” com uma freqüência igual a 1 e apenas quatro termos no total.

Uma terceira consulta é apresentada composta do termo “saúde pública”, sendo o seu vetor expandido representado na Tabela 13.

Tabela 13 - Vetor expandido para o termo “saúde pública”

saude publica
public health
salud publica

Tabela 14 - Documentos recuperados para o termo “saúde pública” e seus sinônimos

Pesquisador	Co-seno	Ordem Co-seno	Escore	Ordem Escore
Pesquisador 1	0.23570	1	0.15092	8
Pesquisador 2	0.21822	2	0.31387	1
Pesquisador 3	0.21822	3	0.25097	2
Pesquisador 4	0.20412	4	0.22596	3
Pesquisador 5	0.20412	5	0.19236	4
Pesquisador 6	0.19245	6	0.19236	5
Pesquisador 7	0.18257	7	0.17468	6
Pesquisador 8	0.16013	8	0.16596	7

Na Tabela 14 são apresentados oito documentos recuperados para o termo “saúde pública” e seus sinônimos. Mais uma vez percebe-se o reposicionamento dos documentos quando ordenados pela medida do co-seno. O pesquisador 1, por exemplo, pelo escore seria o último da lista, mas pelo co-seno está posicionado em primeiro. É possível no vetor de termos do pesquisador 1 o termo “saúde pública” ser o mais freqüente e com mais freqüência que os demais vetores.

5.5 Discussão dos Resultados

Após a realização do estudo comparativo entre o modelo tradicional de buscas no site da Rede ScienTI e o modelo baseado na arquitetura proposta, com a adição

de ontologias e o cálculo da similaridade entre vetores pela medida do co-seno, podem-se notar alguns comportamentos nos resultados que merecem destaque.

O primeiro fato a ser observado é o aumento do valor do *recall*, ou seja, mais currículos na coleção são recuperados. Isto acontece porque alguns currículos relevantes à consulta não estavam sendo recuperados sem a utilização do tratamento semântico proporcionado pela utilização de ontologias. Se a consulta informada é composta apenas do termo “malária”, documentos que utilizam apenas outros termos sinônimos de “malária” não são recuperados. Com a expansão do contexto semântico da consulta, tais documentos são recuperados e apresentados ao usuário. Há, portanto, um maior valor para o *recall*.

Com esse aumento no valor do *recall* espera-se um decréscimo no valor da precisão. Isto ocorre porque o mecanismo de busca tende a ampliar a gama de currículos a serem recuperados, porém muitos são pouco relevantes ou mesmo irrelevantes, pois se distanciam bastante da consulta informada, além disso, a ordenação é baseada apenas no índice *score*, que por sua vez é baseado no TFIDF e indica o peso dos termos nos vetores dos currículos, não sendo considerado uma medida de comparação vetorial eficiente.

A utilização da medida do co-seno para o cálculo da similaridade entre os vetores dos currículos encontrados e o vetor de consulta proporciona uma melhor performance (CHIAO; ZWEIGENBAUM, 2002) e uma melhor ordenação da lista de resultados, fazendo com que os currículos mais relevantes sejam apresentados primeiro.

Esse comportamento foi notado ao comparar a utilização do co-seno e a utilização do TFIDF para a ordenação. Observou-se que os currículos considerados relevantes através de análise subjetiva (avaliados pelo usuário) foram posicionados

no início da lista com a utilização do co-seno, diferentemente quando da utilização do escore.

Portanto, se for considerado um valor de *cutoff*, ou seja, se for realizado um corte na lista de resultados considerando-se, por exemplo, os 100 primeiros ou os 150 primeiros, pode-se dizer que há um aumento da precisão, pois os documentos que estiverem após esse corte possuem um baixo grau de similaridade, segundo a medida do co-seno e, portanto, não são muito relevantes ou podem até mesmo ser irrelevantes.

Um outro fato observado na utilização de ontologias no sistema de buscas da Rede ScienTI se refere às sugestões de termos relacionados obtidos da ontologia. A partir deles, o usuário pode aprender novos termos sobre o tema que está pesquisando, dando subsídios até mesmo quando ele utilizar um sistema de recuperação mais simples que não realize a expansão automática do vetor de consulta. Isto indica que de uma certa forma é possível aumentar o conhecimento do usuário ao utilizar um sistema de recuperação de informação que adote a arquitetura proposta neste trabalho. Se forem considerados termos que possuem muitos significados (no DeCS alguns possuem até 10 sinônimos, incluindo variações de idioma), usuários que talvez não conheçam todos os termos estarão adquirindo conhecimento sobre aquele tema.

Pode-se dizer ainda que é válida a utilização de ontologias de múltiplos idiomas em sistemas de busca cujas fontes de dados sejam de vários países e/ou idiomas, como é o caso da Rede ScienTI. Conforme foi apresentado na seção 5.4, a não utilização de ontologias em alguns casos retornou apenas documentos no mesmo idioma da consulta, não considerando as variações dos termos.

5.6 Considerações Finais

Este capítulo apresentou uma descrição da aplicação da arquitetura proposta a um site de buscas da Rede ScienTI cujos resultados foram comparados ao modelo tradicional. Alguns fatos importantes puderam ser destacados a partir da observação do comportamento do site de buscas sobre essa arquitetura. O próximo capítulo apresenta as conclusões de todo o trabalho, propondo alguns caminhos que podem ser alvo de pesquisas no futuro.

6 CONCLUSÕES E TRABALHOS FUTUROS

6.1 Conclusões

Com a apresentação da arquitetura e a sua aplicação a um site de buscas da Rede ScienTI, verificou-se a viabilidade da proposta de acoplamento de ontologias de forma dinâmica. Isto se dá, em parte, como consequência da especificação de uma representação padrão para as ontologias, possibilitando a adaptação de quaisquer ontologias ao sistema.

A partir das ontologias devidamente conectadas ao sistema, é possível realizar a expansão do contexto semântico da consulta, proporcionando um aumento do número de documentos relevantes recuperados, os quais são ordenados através da medida do co-seno calculada entre o vetor de cada documento e o vetor de consulta, movendo os documentos mais relevantes para o topo da lista de documentos recuperados e ocasionando uma maior precisão em um determinado *cutoff*.

A aplicação da arquitetura ao site de buscas da Rede SCienTI demonstrou uma melhora geral nos resultados da busca em comparação com o outro modelo utilizado.

6.2 Trabalhos Futuros

Este trabalho tem como foco o acoplamento de ontologias a sistemas de recuperação de informação. A partir daí, algumas idéias surgem com sugestões para os trabalhos futuros.

Além do acoplamento de ontologias para a recuperação dos conceitos relacionados aos termos de consulta, seria de grande valia oferecer ao usuário a oportunidade de selecionar o sentido do termo, no caso de termos ambíguos, ou seja, com múltiplos significados, de forma que os resultados obtidos pelo mecanismo de busca possam ser mais refinados e relacionados às necessidades de informação do usuário.

Também poderia ser viável o estudo de uma abordagem utilizando o acoplamento dinâmico de ontologias juntamente com a indexação semântica latente, pois o contexto semântico da consulta seria incrementado não apenas com os termos vindos de ontologias mas também com a correlação existente entre eles na matriz de termo X documento gerada no modelo LSI.

A utilização da arquitetura proposta em combinação com uma ferramenta de busca baseada em casos (RBC) como a RNet (BEPPLER, 2002) poderia recuperar muito mais casos relevantes, uma vez que o vetor de consulta é expandido pela utilização de ontologias. Além disso, vislumbra-se a utilização de ferramentas de *link analysis* como as propostas por Bovo (2004), visto que tal recurso poderia prover uma representação gráfica da ontologia acoplada a um sistema de recuperação de informação, estabelecendo uma comunicação mais visual e intuitiva com o usuário e, conseqüentemente, alcançando melhores resultados.

Uma metodologia para a criação automatizada de ontologias pode ser especificada de forma a possibilitar a integração com o modelo proposto, podendo-se iniciar com a criação de vocabulários temáticos (IGARASHI, 2005) e evoluindo na concepção dos relacionamentos entre os conceitos. Tal proposta só é possível tendo-se em mãos um repositório que possua documentos bem categorizados em áreas de conhecimento, é importante que tais documentos sejam de origem confiável, criados por especialistas nas respectivas áreas. Um exemplo de repositório que preenche tais requisitos é a Plataforma Lattes, com seus milhares de currículos de pesquisadores e sua categorização por áreas de conhecimento.

Da mesma forma um algoritmo de busca baseado, nativamente, em ontologias pode ser proposto mesclando-se com o modelo vetorial, incluindo no próprio motor de busca subsídios para o cálculo da similaridade vetorial.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, R. et al. **Professional XML**. Rio de Janeiro: Ciência Moderna, 2001.

ANDRADE Jr., C. R. **A Construção de um Sistema de Informação Baseado em Indicadores de um Banco de Teses e Dissertações para Apoiar a Gestão de Cursos de Pós-Graduação Strictu-Sensu**. 2003. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2003. Disponível em: <<http://teses.eps.ufsc.br>>.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval**. Addison Wesley, 1999.

BENNERS-LEE, Tim et al. **The Semantic Web**. Scientific American. Estados Unidos, maio 2001. Disponível em: . Acesso em: ago. 2004.

BEPPLER, F. D. **Emprego de RBC para Recuperação Inteligente de Informações**. 2002. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2002. Disponível em: <<http://teses.eps.ufsc.br>>.

BERMEJO, P. H. S. **Metodologia para definição de unidades de informação para plataformas de governo eletrônico: uma aplicação à Plataforma Lattes**. 2004. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2004.

BILLERBECK et al. Query expansion using associated queries Conference on Information and Knowledge Management. In: TWELFTH INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 2003, New Orleans, LA, USA. **Proceedings...** New Orleans, LA, USA, 2003. p. 2-9. ISBN 1-58113-723-0.

BILLHARDT, H. et al. A context vector model for information retrieval. **Journal of the American Society for Information Science and Technology**, Hoboken, v. 53, Issue 3, p. 236, 2002.

BIREME - Informação e Conhecimento em Ciências da Saúde. Disponível em: <<http://www.bireme.br>>. Acesso em: jun. 2003.

BLAIR, D. C.; MARON, M. E. An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System. **ACM**, v. 28, p. 289-299, 1985.

BOVO, A. B. **Um método de tradução de fontes de informação em um formato padrão que viabilize a extração de conhecimento por meio de *Link Analysis* e Teoria dos Grafos**. 2004. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2004. Disponível em: <<http://teses.eps.ufsc.br>>.

BRIN, S.; PAGE, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: SEVENTH INTERNATIONAL WORLD WIDE WEB CONFERENCE, 1998, Brisbane, Australia. **Proceedings...** Australia, 1998.

CAID, W. R.; CARLETON, J. L. Context Vector-Based Retrieval. In: 4th IEEE DUAL-USE CONFERENCE, 1994, Utica, NY. **Proceedings...** Utica, NY, May 1994.

CASTOLDI, A. V. **Uma ontologia para enlaces de unidades de informação em plataformas de governo eletrônico**. 2003. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2003. Disponível em: <<http://teses.eps.ufsc.br>>.

CHIAO, Y.; ZWEIGENBAUM, P. Looking for candidate translational equivalents in specialized, comparable corpora. In: 19th INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 2002, Taipei, Taiwan.

CONSCIENTIAS - Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior. **Comunidade CONSCIENTIAS**. Disponível em: <<http://www.lattes.cnpq.br/lmpl>>. Acesso em: ago. 2004.

CURRÁS, Emilia. **Tesauros**: linguagens terminológicas. Brasília: IBICT, 1995. p. 286.

DAVIES, J. et al. **Towards the Semantic Web**. England: John Wiley & Sons Ltd, 2003.

DECKER, S. The Semantic Web: The Role of XML and RDF. In: IEEE INTERNET COMPUTING, set./out. 2000, p. 2-13. Disponível em: <<http://computer.org/internet>>. Acesso em: maio 2005.

DOMINICH, S. Formal Foundation of Information Retrieval. In: WORKSHOP ON MATHEMATICAL AND CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 2000, Grécia. **Proceedings...** Grécia, 2000.

DUMAIS, S. T. et al. Using Latent Semantic Analysis To Improve Access to Textual Information. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 1988, Washington, D.C., United States. **Proceedings...** Washington, D.C., United States, 1988. p. 281-285. ISBN 0-201-14237-6. 1988.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing & Management**, v. 38, issue 6, p.823, 2002.

FENSEL, D. et al. OIL: Ontology Infrastructure to Enable the Semantic Web. **IEEE Intelligent Systems**, 16(2), 2001.

FILHO, J. L. **Estruturação e Modelagem de Bancos de Dados**. 2001. Disponível em: <<http://www.dpi.ufv.br/~jugurta/bd/estmodbd-gisbr2001.pdf>>. Acesso em: maio 2003.

GRUBER, T. R. **Ontologia**: A mechanism to support portable ontologies - Knowledge Systems Laboratory. Stanford University, 1992. Disponível em: <ftp://ksl.stanford.edu/pub/KSL_Reports/KSL-91-66.ps.gz>. Acesso em: mar. 2005.

GRUBER, T. R. A Translation Approach to Portable Ontology Specifications. **Knowledge Acquisition**, 5, p. 199-220, 1993.

_____. Toward principles for the design of ontologies used for knowledge sharing. **International Journal of Human and Computer Studies**, n. 43, p. 71 / 907-928, 1995.

GUARINO, N. Formal ontology and information systems. In: FOIS, 6-8 jun. 1998, Trento, Italia.

GUARINO, N.; WELTY, C. **Conceptual Modeling and ontological analysis**. Padova: LADSEB-CNR, 1998.

HAAV, H. M.; LUBI, T. L. A Survey of Concept-based Information Retrieval Tools on the Web. In: EAST-EUROPEAN CONFERENCE ADBIS, 2001. **Proceedings...** 2001.

HARMAN, D. **The first Text Retrieval Conference (TREC-1)**. Rockville, MD, 1992. Information Processing and Management. 1993.

HENDLER et al. 2002

HORROCKS, I. DAML+OIL: a Description Logic for the Semantic Web. In: IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING, 2002. Disponível em: <<http://ce.sharif.edu/~daneshpajouh/ontology/ieeede2002.pdf>>. Acesso em: maio 2005.

IGARASHI, W. **Construção Automática de Vocabulários Temáticos e Cálculo de Aderência Curricular**: Uma Aplicação aos Fundos Setoriais. 2005. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2005. Disponível em: <<http://teses.eps.ufsc.br>>.

JASPER, R., USCHOLD, M. A. Framework for understanding and classifying ontology applications. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI-99), jul. 31 - aug. 6, 1999, Stockholm, Sweden. **Proceedings...** Stockholm, Sweden, jul. 31 - aug. 6, 1999.

JENKINS, C. et al. Searching the World Wide Web: an evaluation of available tools and methodologies. **Information and Software Technology**, v. 39, p. 985-994, 1998.

JONES, W. P.; FURNAS, G. W. Pictures of Relevance: A Geometric Analysis of Similarity Measures. **Journal of the American Society for Information Science**, 38(6), p. 420-442, 1987.

KORFHAGE, Robert R. **Information Storage and Retrieval**. New York: Wiley Computer Publishing, 1997.

LANCASTER, F. W. **Information Retrieval Systems**: Characteristics, Testing and Evaluation. New York: Wiley, 1968.

LENAT, R. V. et al. CYC: Toward programs with common sense. **ACM**, 1990.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Massachusetts Institute of Technology, 1999.

MARCONDES, C. H.; SAYÃO, L. F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001.

MENA, E. et al. **OBSERVER_ An Approach for Query Processing in Global Information Systems based on Interoperation across Pre_existing Ontologies**. Boston: Kluwer Academic Publishers, 2000. Disponível em: <http://ieeexplore.ieee.org/xpl/abs_free.jsp%3FarNumber%3D554955>. Acesso em: maio 2005.

MILLER, E. **An Introduction to the Resource Description Framework**. D-Lib Magazine, 1998.

NOY, N.; MCGUINNESS, D. L. **Ontology Development 101**: a guide to creating your first ontology. CA: Stanford University, 2000.

ONTOLOGY, Org. **Enabling Virtual Business**. Disponível em: <<http://www.ontology.org/>>. Acesso em: maio 2003.

PACHECO, R. C. S. **Uma metodologia de desenvolvimento de plataformas de governo para geração e divulgação de informações e de conhecimento**. Artigo apresentado em cumprimento a requisito parcial de concurso para professor no INE/UFSC. Florianópolis, 14 jan. 2003. 35 p.

PACHECO, R. C. S.; KERN, V. M. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. **Ciência da Informação**, v. 30, n. 3, p. 56-63, set./dez. 2001. Disponível em: <<http://www.ibict.br/cionline/300301/3030801.pdf>>. Acesso em: dez. 2004.

PARALIC, J.; KOSTIAL, I. **Ontology-based Information Retrieval**. 2003. Disponível em: <<http://neuron-ai.tuke.sk/~paralic/papers/IIS03.pdf>>. Acesso em: nov. 2004.

PELLEGRINI FILHO, A. Health research, health policy and equity in Latin America. **Ciência & Saúde Coletiva**, v. 9, n. 2, p. 339-350, apr./june 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232004000200011&lng=en&nrm=iso>. ISSN 1413-8123.

PÉREZ, A. G. et al. **Ontological Engineering: with examples of Knowledge Management, e-Commerce and the Semantic Web**. Londres: Springer Verlag, 2004.

PERRI. E-governance. Do Digital Aids Make a Difference in Policy Making? In: PRINS, J. E. J. (Ed.). **Designing E-Government**. Kluwer Law International, 2001. p. 7-27.

POWERS, S. **Practical RDF**. Estados Unidos: O'Reilly & Associates, 2003.

RAGHAVAN, V. V. et al. A critical investigation of recall and precision as measures of retrieval system performance. **ACM Transactions on Information Systems**, 7(3), p. 205-229, 1989.

RAGHAVAN, V. V.; WONG, S. K. M. **A critical analysis of vector space model in information retrieval**. J. Am. Soc. Inf. Sci. 37, p. 279-287, 1986.

RIJSBERGEN, C. J. van. **Information retrieval**. 1999. Disponível em: <<http://www.dcs.gla.ac.uk/~iain/keith/>>. Acesso em: jan. 2005.

SABBATINI, M. **Lattes, cómo gestionar la ciência brasileña en la red**. 2001. Disponível em: <<http://www.galeon.com/divulcat/articu/141a.htm>>.

SALTON, G.; MCGILL, M. H. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SALTON, G. et al. A vector space model for automatic indexing. **ACM**, v. 18, issue 11, p. 613-620, 1975. ISSN 0001-0782.

SCHATZ, Bruce R. Information Retrieval in Digital Libraries: Bringing Search to the Net. **Science**, v. 275, p. 327-334, jan. 1997.

SHANNON, C. E.; WEAVER, W. **The Mathematical Theory of Communication**. Urbana: University of Illinois Press, 1964.

SHÜTZE, H. Dimensions of Meaning. In: SUPERCOMPUTING (IEE), 1992, Mineapolis. **Proceedings...** Mineapolis, 1992. p. 787-796.

STOJANOVIC, N. On the query refinement in the ontology-based searching for information. **Science Direct**. Elsevier, 2004. Disponível em: <<http://www.sciencedirect.com>>.

TAKEMURA, R. Y. **Controle Inteligente**: Lógica Difusa. Disponível em: <http://www.din.uem.br/ia/control/fuz_prin.htm>. Acesso em: abril 2003.

THE FREE DICTIONARY. Disponível em: <<http://encyclopedia.thefreedictionary.com/>>. Acesso em: jul. 2004.

UNESCO. **Guidelines for the establishment and development of monolingual thesauri**. Paris, 1973.

W3C - World Wide Web Consortium. Disponível em: <<http://www.w3.org>>. Acesso em: mar. 2004.

WANG, P. et al. Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. **Information Processing and Management**. 36, 2000. p. 229-251. Disponível em: <https://206.191.28.118/docushare/dsweb/Get/Document-1442/Wang_et_al_2001_IR_model.pdf>. Acesso em: maio 2005.

WITTEN, I. H. et al. **Managing Gigabytes** – Compressing and Indexing Documents and Images. Estados Unidos: Academic Press, 1999.

WIVES, L. K.; LOH, S. **Hiperdictionary**: a Knowledge Discovery Tool to Help Information Retrieval. Porto Alegre: UFRGS, 2000.

WONG, S. K. M.; RAGHAVAN, V. V. Vector space model of information retrieval: a reevaluation. In: 7th ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1984. **Proceedings...** 1984. Disponível em: <<http://portal.acm.org/citation.cfm?id=636816&dl=ACM&coll=GUIDE>>. Acesso em: maio 2005.

WONG, S. K. M.; YAO, Y. Y. On modeling information retrieval with probabilistic inference. **ACM Transactions on Information Systems (TOIS)**, v. 13, issue 1, p. 38-68, 1995. ISSN 1046-8188.

ZOBEL, J. How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: 21st ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1998. **Proceedings...** 1998.