PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA E EXTENSÃO

ÁREA DE CIÊNCIAS TECNOLÓGICAS

Curso de Mestrado em Nanociências

HENRIQUE TAMIOSSO MACHADO

ARQUITETURA DE UM SISTEMA DE CONSULTAS E VISUALIZAÇÃO GRÁFICA DA REPRESENTAÇÃO DO CONHECIMENTO CONTIDO NO PUBMED

Santa Maria, RS

2009

HENRIQUE TAMIOSSO MACHADO

ARQUITETURA DE UM SISTEMA DE CONSULTAS E VISUALIZAÇÃO GRÁFICA DA REPRESENTAÇÃO DO CONHECIMENTO CONTIDO NO PUBMED

Dissertação apresentada ao Curso de Mestrado em Nanociências do Centro Universitário Franciscano de Santa Maria como requisito parcial para obtenção do título de Mestre em Nanociências.

Orientador: Prof° Dr° **GIOVANI RUBERT LIBRELOTTO**

Santa Maria, RS

2009

CENTRO UNIVERSITÁRIO FRANCISCANO

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA E EXTENSÃO
CURSO DE MESTRADO EM NANOCIÊNCIAS

A COMISSÃO EXAMINADORA, ABAIXO-ASSINADA, APROVA A DISSERTAÇÃO:

ARQUITETURA DE UM SISTEMA DE CONSULTAS E VISUALIZAÇÃO GRÁFICA DA REPRESENTAÇÃO DO CONHECIMENTO CONTIDO NO PUBMED

Elaborada por:
HENRIQUE TAMIOSSO MACHADO

COMISSÃO EXAMINADORA

_____

Profº. Dr. Giovani Rubert Librelotto
Presidente

_____

Profª. Dra. Iara Augustin (Computação - UFSM)

_____

Profº. Dr. José Carlos Merino Mombach

Santa Maria-RS, 12 de Março de 2009.

Dedico este trabalho aos meus pais,
João Carlos e Maria Luiza
a quem tudo devo.

AGRADECIMENTOS

Dedico meus agradecimentos a todos que de alguma forma colaboraram com a realização deste trabalho:

– a Deus, sou grato por todas as oportunidades que me foram dadas;

– aos colegas de graduação e pós-graduação da UNIFRA;

– aos demais Professores do mestrado em Nanociências, que colaboraram na minha formação acadêmica;

– aos meus familiares, pelo apoio em as minhas decisões;

– à Profa. Dra. Juliana Vizzotto, pela co-orientação neste trabalho;

–ao Prof°. Dr°. Giovani Rubert Librelotto, pela orientação, tolerância, paciência, amizade e pelos ensinamentos.

- aos colegas e amigos Mirkos Martins, Renato Azevedo e Alexandre Roos pela constante ajuda e participação neste trabalho.

RESUMO

Na bioinformática existe uma grande quantidade de informações biológicas e genéticas que servem de suporte para pesquisas, e a cada dia essa quantidade de informações cresce ainda mais. Diversas pesquisas realizadas por inúmeros pesquisadores de diferentes áreas como biologia molecular, bioquímica estrutural, enzimologia, fisiologia, patologia, entre outras, vem gerando vários resultados e informações que devem ser armazenadas para serem utilizadas de diversas formas. Mas aí que surge o problema, como armazenar, manipular, visualizar todas essas informações? A bioinformática usa o poder computacional para catalogar, organizar, estruturar e manipular essas informações de uma forma que facilite a utilização dessas informações que são de extrema importância para a Biologia.

O PubMed é um serviço da Biblioteca Nacional de Medicina dos Estados Unidos (*U.S. National Library of Medicine*) que fornece acesso para mais de 18 milhões de citações para artigos científicos de jornais da área de ciências da saúde. De modo a propor uma nova abordagem para a busca e representação do conhecimento encontrado no resultado obtido, este trabalho apresenta uma arquitetura de um sistema para consultas utilizando prioridades e visualização das informações através de redes semânticas para representar o conhecimento contido no PubMed, para isso, foi utilizado a norma ISO 13250 Topic Maps para a criação de redes semânticas envolvendo os conceitos encontrados no sistema de informação do PubMed, realizado o desenvolvimento de uma arquitetura composta por um banco de dados com informações oriundas do PubMed disponibilizadas pela Biblioteca Nacional de Medicina dos Estados Unidos (NLM), uma interface para pesquisa diferente da disponibilizada pelo sistema *Entrez*, onde seja possível definir prioridades para as consultas. E também fazer a representação do conhecimento contido no PubMed através de redes semânticas e técnicas de *data mining*.

Palavras-chave: PubMed, *Data Mining*, Redes Semânticas.

ABSTRACT

In the bioinformatics has a great amount of biological and genetic information that serve of support for research, and each day this amount of information grows more. Diverse research made for many researchers in different areas, as molecular biology, structural biochemist, enzymology, physiology, pathology, among others, comes generating some results and information that must be stored to be used in the diverse forms. But there the problem appears, how to store, to manipulate, to visualize all these information? The bioinformatics uses the computational power to catalogue, to organize, to structuralize and to manipulate these information facilitating the use of these information that are of extreme importance for Biology.

The PubMed is a service of the National Library of Medicine of United States that supplies access to more than 18 million citations for periodical scientific articles in the area of health science. In order to consider a new boarding for the search and representation of the knowledge found in the gotten result, this work presents the use of norm ISO 13250 Topic Maps for the creation of semantic nets involving the concepts found in the system of information of the PubMed, the construction of a composed architecture for a data base with deriving information of the PubMed by the National Library of Medicine of the United States (NLM), an interface for different research of the available one for the Entrez system, where it is possible to define priorities for the consultations. Also is possible to make the representation of the knowledge contained in the PubMed through semantic nets and techniques of date mining.

Keywords: *PubMed, Data Mining, Semantic Networks.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O rápido crescimento da Bioinformática está ocorrendo devido às necessidades de manipulação de grandes quantidades de informações biológicas e genéticas. Sem o auxílio do computador seria muito difícil processar e manipular essa grande quantidade de informações, que a cada dia aumenta bastante. Com a descoberta do código genético e do fluxo da informação biológica, dos ácidos nucléicos para as proteínas, tais polímeros passaram a ser os principais objetivos de estudo da Biologia Molecular. Depois dessas descobertas, diversas pesquisas realizadas por inúmeros pesquisadores de diferentes áreas como biologia molecular, bioquímica estrutural, enzimologia, fisiologia e patologia, vêm gerando vários resultados e informações que devem ser armazenadas para serem utilizadas de diversas formas. Então, surge o problema: como armazenar, manipular, visualizar todas essas informações? A bioinformática usa o poder computacional para catalogar, organizar, estruturar e manipular essas informações de uma forma que facilite a utilização dessas informações que são de extrema importância para a Biologia.

Existem diversas ferramentas que foram criadas para armazenar, manipular e visualizar essas informações, uma delas é o PubMed. O PubMed (MEDICINE, 2007) é uma ferramenta para pesquisa da literatura científica das ciências biológicas, que é disponibilizada pelo NCBI (*National Center for Biotechnology Information* – Centro Nacional de Informações de Biotecnologia), da Biblioteca Nacional de Medicina dos Estados Unidos. É um dos recursos mais utilizados por profissionais da área na Internet, pois é possível acessar mais de 4000 revistas científicas de diversas áreas *(*GIBAS; JAMBECK, 2001), contendo, no total, mais de 18 milhões de citações.

Um dos problemas encontrados pelos usuários do PubMed é a dificuldade de analisar as informações disponibilizadas como resposta às consultas, pois as mesmas são mostradas de forma textual e não graficamente. A forma como as consultas são realizadas pelos usuários através do sistema disponível pela NCBI é outro fator que dificulta de se ter um resultado mais apropriado para a pesquisa realizada. Neste contexto, o presente trabalho apresenta uma arquitetura de um sistema para consultas utilizando prioridades e visualização das informações através de redes semânticas para representar o conhecimento contido no PubMed.

Após ser firmado um protocolo entre o Centro Universitário Franciscano e a Biblioteca Nacional de Medicina dos Estados Unidos (NLM), iniciou-se o desenvolvimento de uma arquitetura para trabalhar com o dados armazenados no PubMed. Dessa forma, desenvolveu-se o chamado PubMed Local.

O PubMed Local é uma arquitetura composta por um banco de dados, contendo as informações disponibilizadas pela NLM, sobre a qual foram aplicadas técnicas sobre os dados adquiridos, com o objetivo de gerar uma rede semântica, para representar o conhecimento encontrado nesse domínio.

Além disso, essa dissertação utiliza técnicas de mineração de dados para achar padrões nas publicações, o que auxilia os pesquisadores a interpretar e compreender os dados e informações disponíveis no PubMed.

1.1. OBJETIVOS

1.1.1. Objetivo Geral

O objetivo geral deste trabalho é a representação do conhecimento contido no PubMed através de linguagens ontológicas, de modo que todos os termos citados nos artigos científicos nele catalogados possam definir a criação de uma rede semântica, o que tornaria as pesquisas neste ambiente mais eficientes e intuitivas.

1.1.2. Objetivos Específicos

- Construção de uma ontologia para a representação do conhecimento encontrado no PubMed;
- Estruturação e criação de um banco de dados com as informações contidas no PubMed;
- Utilização das normas ISO 13250 Topic Maps e W3C RDF e OWL para a representação da ontologia;
- Desenvolvimento de uma interface para consultas;
- Mineração dos dados do PubMed de forma a encontrar padrões nas publicações.

1.2. ESTRUTURA DA DISSERTAÇÃO

Para realizar o objetivo proposto, o capítulo 2 descreve o PubMed. Após, o capítulo 3 descreve o processo de criação do banco de dados do PubMed Local. O capítulo 4 apresenta como é feita a representação do conhecimento contido no PubMed. Em seguida, o capítulo 5 apresenta a interface, assim como explica as suas funcionalidades e os métodos utilizados em sua criação. O capítulo 6 aborda como foram feitas as descobertas de conhecimento no PubMed, utilizando técnicas de *data mining*. O capítulo seguinte, mostra os resultados reais, os quais foram submetidos a profissionais da área. Por fim, no capítulo 8 discuti alguns trabalhos relacionados e o capítulo 9 apresenta as conclusões.

2. PUBMED

O acesso rápido à informação científica atualizada é de fundamental importância para o profissional da área da saúde. Essa informação científica está disponível através de artigos publicados em revistas científicas. Hoje em dia, quase todas as revistas estão publicando seus conteúdos em um formato eletrônico, além da cópia impressa.

No final da década de 50, a Biblioteca Nacional de Medicina criou o MEDLARS (*Medical Literature Analysus and Retrieval System),* um sistema feito para automatizar a composição da biblioteca médica. Em 1971, a MEDLINE (MEDICINE, 2006) – que significa MEDLARS On-Line – começou a oferecer acesso on-line às referências da base de dados MEDLARS. Porém, apenas em 1997, o MEDLINE foi disponibilizado para todos, de forma gratuita (TIDIA, 2006).

Um excelente recurso para pesquisa de literatura científica sobre ciências biológicas é o serviço gratuito patrocinado pelo NCBI (*National Center for Biotechnology Information* – Centro Nacional de Informações de Biotecnologia) da Biblioteca Nacional de Medicina dos Estados Unidos, conhecido como PubMed (MEDICINE, 2007). Esse serviço permite acesso ao banco de dados MEDLINE, que é um banco de dados de citações referentes à literatura científica sobre ciências biológicas. As consultas no PubMed são realizadas através do sistema Entrez (ENTREZ, 2008).

2.1. ORGANIZAÇÃO DO PUBMED

O PubMed é um dos recursos disponíveis mais utilizados por biólogos na Web. Mais de 4.000 revistas científicas estão indexadas no PubMed, incluindo a maioria das mais respeitadas revistas científicas sobre biologia celular e molecular, bioquímica, genética e campos relacionados, assim como muitas publicações clínicas de interesse para profissionais médicos (GIBAS; JAMBECK, 2001). Nele estão incluídos mais de 18 milhões de citações de MEDLINE e outros jornais relacionados com artigos científicos, como mostra a Figura 2.1.

United States National Library of Medicine
National Institutes of Health

Yearly Citation Totals from 2008 MEDLINE/PubMed Baseline: 16,880,015 Citations Found

1865	1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880	1881	1882
3	11	36	95	58	69	62	38	47	39	29	58	46	80	122	75	55	57
1883	**1884**	**1885**	**1886**	**1887**	**1888**	**1889**	**1890**	**1891**	**1892**	**1893**	**1894**	**1895**	**1896**	**1897**	**1898**	**1899**	**1900**
72	101	182	170	212	152	160	205	171	157	251	231	232	277	243	254	241	196
1901	**1902**	**1903**	**1904**	**1905**	**1906**	**1907**	**1908**	**1909**	**1910**	**1911**	**1912**	**1913**	**1914**	**1915**	**1916**	**1917**	**1918**
242	283	304	297	273	335	335	336	325	286	337	345	392	348	540	574	532	470
1919	**1920**	**1921**	**1922**	**1923**	**1924**	**1925**	**1926**	**1927**	**1928**	**1929**	**1930**	**1931**	**1932**	**1933**	**1934**	**1935**	**1936**
490	547	520	544	606	725	782	854	908	970	853	899	1,019	1,073	1,161	1,138	1,264	1,218
1937	**1938**	**1939**	**1940**	**1941**	**1942**	**1943**	**1944**	**1945**	**1946**	**1947**	**1948**	**1949**	**1950**	**1951**	**1952**	**1953**	**1954**
1,213	1,306	1,238	1,143	1,080	1,022	993	951	939	1,020	1,016	1,291	16,150	81,236	101,369	106,572	107,452	104,066
1955	**1956**	**1957**	**1958**	**1959**	**1960**	**1961**	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**	**1968**	**1969**	**1970**	**1971**	**1972**
106,896	105,291	109,981	107,728	108,165	109,408	117,658	124,839	140,481	159,310	173,913	175,015	186,910	203,984	210,986	214,097	218,807	222,600
1973	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**
226,545	230,223	244,361	249,305	255,909	265,638	274,357	272,508	274,493	285,528	299,286	308,194	318,451	330,918	348,624	366,212	382,100	389,045
1991	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
389,864	392,905	399,352	408,416	418,242	423,994	434,615	449,205	462,020	490,626	519,896	542,504	568,939	601,557	627,549	641,604	434,237	55

Last updated: 17 December 2007
First published: 17 December 2007
Metadata| **Permanence level:** Permanent: Dynamic Content

Copyright, Privacy, Accessibility
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Health & Human Services

Figura 2.1. Número de citações ao final do ano de 2007. Fonte: http://www.ncbi.nlm.nih.gov/pubmed

As consultas no PubMed são realizadas através do sistema Entrez, que é um sistema de busca e recuperação que integra as informações das bases de dados da NCBI, onde possui bases da MEDLINE. Através do PubMed, é possível consultar seqüências de nucleotídeos, seqüências de proteínas, estruturas macromoleculares e genomas completos, taxonomia e outros, como mostra a Figura 2.2, onde cada círculo representa uma das bases de dados do MEDLINE, e a cor do círculo representa o número de dados inseridos nessa base de dados. (TIDIA, 2006).

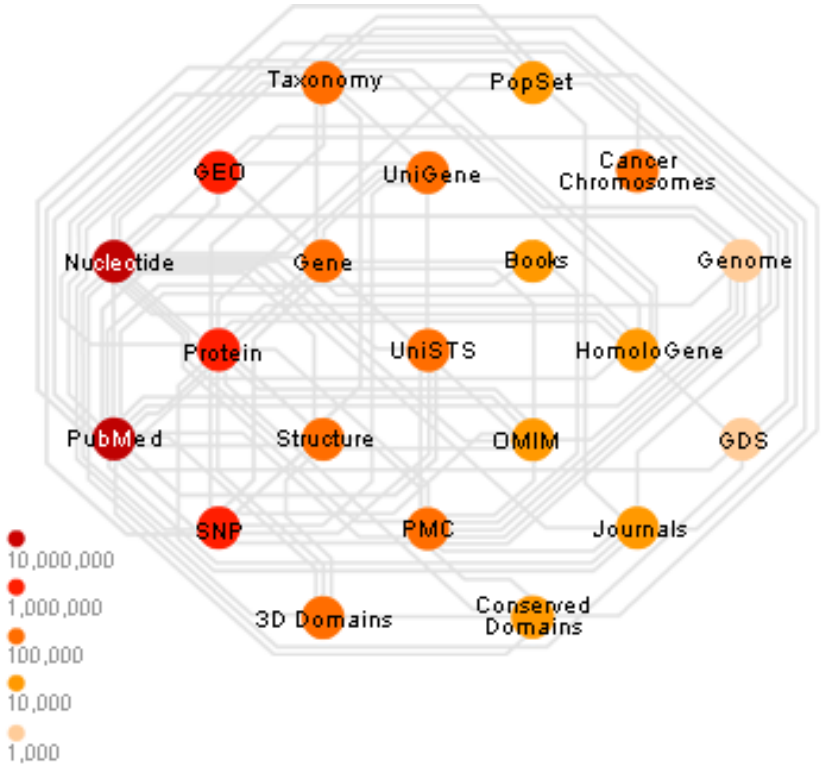


Figura 2.2. Integração do Entrez entre as diversas bases de dados do Medline onde a legenda indica o número de dados em cada base. Fonte: http://www.ncbi.nlm.nih.gov/pubmed

Através do PubMed, atualmente, tem-se acesso a várias bases de dados incluindo a OLDMEDLINE (com registros anteriores a 1966), *Serials Database* (sobre as revistas indexadas), NLM *Catalog* e a MEDLINE (de 1966 em diante).

Abaixo, listam-se algumas diferenças entre MEDLINE e PubMed (TIDIA, 2006):

- MEDLINE é o maior componente do PubMed, a base de citações biomédicas e *abstracts* da *National Library of Medicine* dos Estados Unidos (NLM).
- A base MEDLINE é pesquisável na web, pelo PubMed, de forma gratuita.
- A MEDLINE cobre mais de 4800 revistas publicadas nos Estados Unidos e em mais de 70 outros países de 1966 até o presente.
- MEDLINE inclui referências a artigos indexados com termos do vocabulário controlado MeSH (*Medical Subject Headings*) (NCBI, 2007).
- As citações da MEDLINE são de revistas selecionadas para a inclusão na base de dados através de um comitê.

Para se ter uma idéia da evolução do número de consultas à base PubMed, em 1997, o total era de cerca de 2 milhões de consultas por mês; atualmente é de mais de 3 milhões por dia (CANESE, 2006).

A base de dados PubMed, é composta por 3 camadas de *software* como mostrado na Figura 2.3. A primeira camada é um Sistema de Gerenciamento de Banco de Dados (DBMS) que gerencia uma coleção de dados. No topo está o navegador Web que transmite os pedidos para o banco de dados e retorna as respostas como páginas Web. No meio está uma camada de *software* que fica entre a DBMS e o navegador Web para transformar requisições em consultas ao banco de dados, e para transformar o resultado da consulta e linguagem de marcação de hipertexto (HTML).

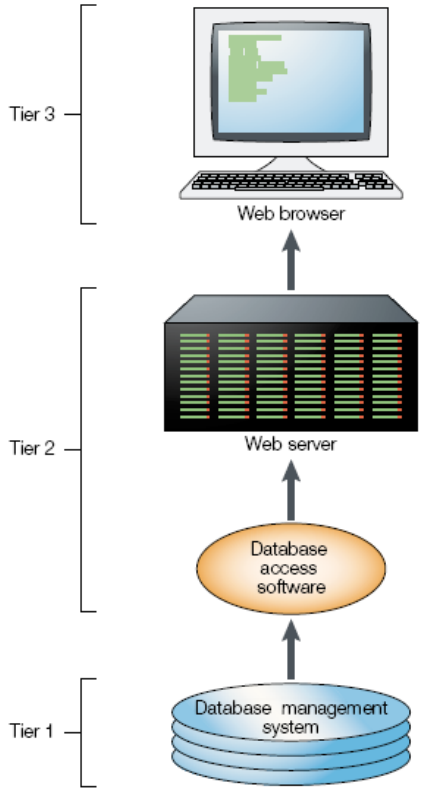


Figura 2.3. Arquitetura do PubMed

Os arquivos do PubMed são destinados ao tratamento automatizado e, portanto, disponíveis no formato XML. Cada conjunto de 30.000 citações do PubMed é armazenado como um arquivo XML definido por um DTD (*Document Type Definition*) (PEPPER; GRAHAM, 2001). Foi definido um esquema XML para os arquivos do PubMed. A visualização desta estrutura é apresentada no Anexo A e um exemplo de documento está no Anexo B.

2.2. CONSULTAS NO PUBMED

Conforme citado anteriormente, as buscas no PubMed são realizadas através do sistema Entrez, onde o usuário pode realizar uma pesquisa simples, apenas digitando

palavras-chave, ou pode utilizar uma pesquisa avançada, usando filtros para limitar a pesquisa, definindo em qual jornal foi publicado, ano da publicação, autores, idioma, entre outros limites.

Pode-se ainda realizar consultas utilizando os termos MeSH ou *Medical Subject Headings* no Pubmed. MeSH é um vocabulário controlado usado na indexação dos artigos do MEDLINE. Também, pode ser visto como uma biblioteca de termos padronizados que podem ajudar a localizar manuscritos que utilizem termos alternativos para se referirem ao mesmo conceito (GIBAS; JAMBECK, 2001). O vocabulário MeSH provê uma forma consistente de recuperar informações que pode aparecer de diferentes formas terminológicas para descrever o mesmo conceito. Um artigo indexado para o MEDLINE tem os termos MeSH assinalados como descritores para cada um dos artigos indexados; vale ressaltar que nem todos os artigos armazenados no PubMed estão indexados.

Entretanto, ao efetuar-se uma busca por artigos usando somente descritores MeSH, será encontrado somente aqueles artigos que foram indexados por termos MeSH, e perderá aqueles artigos que estão no PubMed, mas que não estão indexados pela MEDLINE.

Pode-se também restringir a busca a termos como assunto principal (*Major Topic Headings*). Assim, os itens encontrados serão primariamente nos termos principais e não naqueles que estão relacionados ao assunto como termos secundários ou que mencionam o assunto de passagem.

Quando é efetuada uma consulta no Entrez, os resultados são visualizados de várias formas, como: sumário, documento XML, *abstract*, *abstract plus*, entre outras. Após essa consulta, existe a possibilidade de salvar o resultado dessa consulta em um arquivo texto, enviar para um e-mail ou imprimir, por exemplo.

O PubMed sempre mostrará o resultado de forma padronizada, ou seja, quase sempre as citações conterão os nomes dos autores, o título do trabalho em inglês, o título da revista juntamente com outras informações sobre a publicação, idioma original do documento, um identificador número do PubMed, e a situação do documento na base de dados. Na Figura 2.4 é apresentada a página de resultados do PubMed.

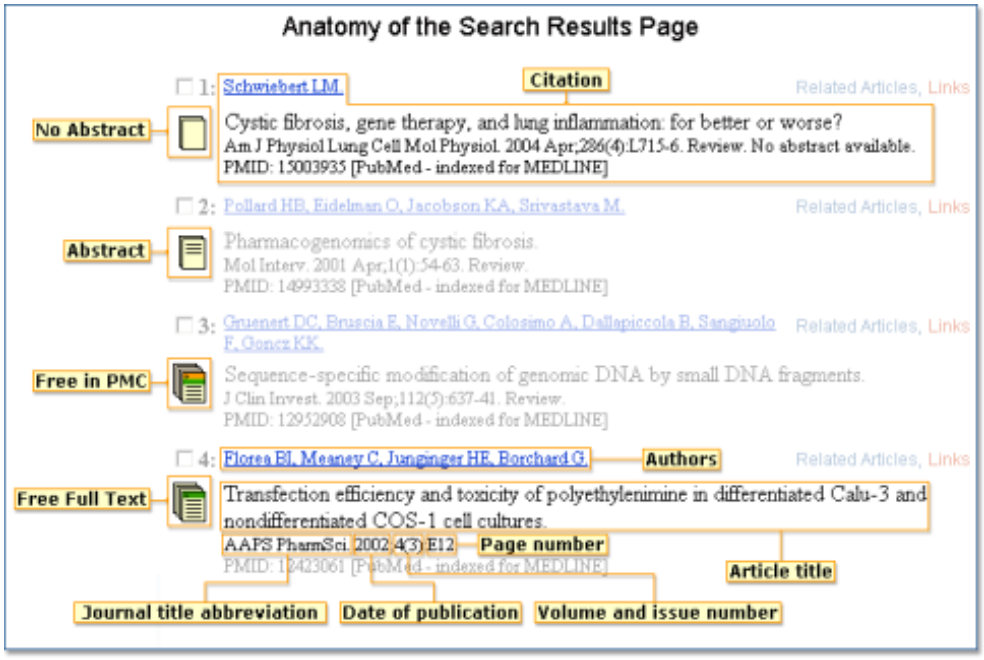


Figura 2.4. Anatomia da pagina de resultados do PubMed. Fonte: http://www.ncbi.nlm.nih.gov/pubmed/

2.3. PROBLEMAS DETECTADOS NO PUBMED

No PubMed tem-se uma busca estática, através da qual o usuário entra com as palavras que deseja procurar, e tem vinte resultados mostrados por página (por padrão, podendo ser alterado pelo usuário). O resultado e a ordem dos artigos mostrados é a mesma, sempre que for utilizado o mesmo termo para a busca. Isso torna a busca de artigos bastante complexa em áreas abrangentes, pois o artigo desejado pode estar situado no final da busca, fazendo com que o usuário perca muito tempo passando por artigos que não são de seu real interesse, antes que encontre o resultado efetivamente desejado.

A utilização do sistema de filtros pode trazer o resultado desejado mais facilmente, pois a busca é restringida em alguns pontos; mas, ainda assim, ao executar uma busca em áreas bastante abrangentes, o artigo desejado pode não ser de fácil acesso.

As buscas por termos MeSH apresentam problemas, pelo fato de nem todos os artigos no PubMed estarem indexados com esse vocabulário. Então, ao efetuar uma busca pelos qualificadores MeSH, somente serão retornados, como resultados, os artigos que já foram indexados. Aqueles que não foram indexados serão ignorados nesse tipo de busca; então, é possível que o artigo desejado não tenha sido indexado, e fique fora do resultado, fazendo com que o usuário perca todo o tempo de realização da busca.

Para ilustrar melhor o problema descrito no parágrafo anterior, toma-se o seguinte exemplo: supondo-se que o usuário execute uma busca utilizando "*anaphylaxis*" como termo, e no banco de dados haja cadastrados os artigos com os seguintes títulos:

- "*Anaphylactoid reactions and bradykinin generation in patients treated with LDL-apheresis and an ACE inhibitor*."
- "*Phase I trial of paclitaxel in children with refractory solid tumors: a Pediatric Oncology Group Study*."
- "*Latex anaphylaxis masquerading as fentanyl anaphylaxis: retraction of a case report*."

Uma vez que somente o primeiro e o segundo estão indexados por termos MeSH, a busca retorna ambos artigos como resultado da consulta. O usuário, não satisfeito com o resultado obtido, opta por executar novamente a busca, só que dessa vez da forma padrão (através da consulta por palavras-chaves), e obtém como resultado os três artigos citados. Ao analisar os três artigos, o usuário chega à conclusão que o artigo que é mais relevante para suas necessidades é o artigo "*Latex anaphylaxis masquerading as fentanyl anaphylaxis: retraction of a case report*", justamente o artigo que não está indexado pelos termos MeSH.

Outro problema encontrado no PubMed é a grande quantidade de informações disponíveis ao usuário, o que dificulta os usuários a interpretar e compreender esses dados e informações. Para exemplificar o problema, considere que um usuário pesquise por um determinado assunto, e que o resultado desta busca traga inúmeras citações, mas não mostre uma relação entre os autores, para identificar qual autor se co-relaciona com o outro.

O usuário do PubMed também tem a dificuldade de analisar as informações, pois elas são mostradas de forma textual e não graficamente.

Em termos de pesquisa na base de dados, quando é efetuada uma pesquisa no PubMed, o número de referências retornadas é muito grande. Por exemplo, a busca de artigos com a palavra *bioinformatics* retorna 16189 referências. Caso a busca seja efetuada com *bioinformatics* DNA, o número de referências passa a ser 4357, o que ainda é considerado um número excessivo de referências. Muito provavelmente, nenhum pesquisador verificaria uma a uma dessas 4357 referências. Caso a resposta do sistema separasse os artigos em sub-temas ou em alguma outra classificação, a busca pelo(s) artigo(s) desejado(s) seria mais eficiente.

Outro exemplo pode ser dado quando forem buscados termos que são sinônimos. Quando se buscam artigos com o termo *homo sapiens*, retorna-se 1438 artigos; entretanto, quando se busca pelo termo *human*, o número de referências retornado é 9883146. Conclui-se

que o motor de busca do PubMed não faz relações entre termos que são sinônimos, pois nesse caso, em particular, as duas procuras deveriam retornar o mesmo conjunto de referências.

Outro item que o PubMed não aborda é a representação do conhecimento, podendo ser utilizadas técnicas de *data mining* para descobrir informações sobre co-autorias mais frequentes, palavras-chaves mais utilizadas em conjunto.

Algumas das questões colocadas anteriormente podem ser respondidas por estatísticas simples. Em (CORLAN, 2007) um *web-service* fornece estatísticas para determinado entradas de palavras-chave. Procurando relações mais complexas, como por exemplo a co-autoria de redes, simples estatísticas têm uso limitado. Mineração de dados aparece como uma promissora abordagem para descobrir padrões de dados.

Propõe-se a aplicação de técnicas de *data mining*, ou seja, regras de associação (AGRAWAL; SRIKANT, 1994) e redes de dependência (HECKERMAN *et al*, 2000), para os dados fornecidos pelo PubMed, incididos sobre o período de 2003 a 2005.

2.4. SUMÁRIO DO CAPÍTULO

Neste capítulo foi abordado o PubMed, como ele está organizado, como são realizadas as consultas através do sistema Entrez e alguns problemas encontrados no PubMed. O objetivo deste capítulo foi de definir a estrutura e o funcionamento do PubMed, bem como realizar os levantamentos sobre os problemas encontrados neste ambiente, que é a base para este trabalho. No próximo capítulo, apresenta-se a organização o banco de dados do PubMed e como foi realizada a criação do modelo Entidade Relacionamento do PubMed Local.

3. O PUBMED LOCAL

Os dados armazenados no sistema de informação do PubMed tem uma estrutura rígida e bem definida. A mesma pode ser formalizada por uma gramática livre de contexto (MENEZES, 2000), da forma apresentada abaixo, onde são apresentados apenas os seus conceitos principais:

MedlineCitation	==> PMID, DateCreated, DateCompleted, Article, MedlineJournalInfo, ChemicalList, CitationSubset, MeshHeadinList
Article	==> Journal, ArticleTitle, Pagination, Abstract, Affiliation, AuthorList, Language, PublicationTypeList
Journal	==> ISSN, JournalIssue, Title
JournalIssue	==> Volume, Issue, PubDate
PubDate	==> Year, Month, Day, Hour?, Minute?, Second?
MedlineJournalInfo	==> Country, MedlineTA, NlmUniqueID
ChemicalList	==> Chemical+
Chemical	==> RegistryNumber, NameOfSubstance
MeshHeadingList	==> MeshHeading+
MeshHeading	==> DescriptorName, QualifierName?
AuthorList	==> Author+
Author	==>LastName, ForeName, Initials
PublicationTypeList	==> PublicationType

Esta gramática livre de contexto apenas demonstra a estrutura básica da informação encontrada no PubMed. Na prática, essas informações estão em banco de dados estruturados de uma maneira similar.

O PubMed disponibiliza acesso aos seus dados representados em arquivos XML *(eXtensible Markup Language*). A partir destes arquivos, utilizou-se a ferramenta *Exult* (Novixys, 2007) para a conversão dos dados para o modelo relacional. Entretanto, devido ao processo automatizado, diversos problemas foram detectados no banco de dados gerado, o que acarretou na construção manual de código SQL (*Structured Query Language*) para a eliminação de redundância de dados e otimização das tabelas geradas. A partir desse banco de dados criado, chamado de PubMed Local, foi utilizada uma ferramenta para extração de dados, seguindo uma especificação para geração de tópicos e associações a serem armazenados em um repositório.

Após esses processos, tem-se como resultado a estrutura representada pela Figura 3.1, onde ainda não se tem uma interface na qual o usuário-final possa fazer suas pesquisas. Dessa

forma, primeiramente, foi feito o mapeamento dos dados para o banco de dados e, em seguida, utilizado o *Metamorphosis* (LIBRELOTTO; RAMALHO; HENRIQUES 2006), para transformar os dados no banco de dados em uma rede semântica representada de acordo com a norma ISO 13250 Topic Maps (BIEZUNSKY; BRYAN; NEWCOMB, 1999), para que os mesmos sejam mostrados em uma visualização semântica. Cada um dos passos mostrados será detalhado a seguir.

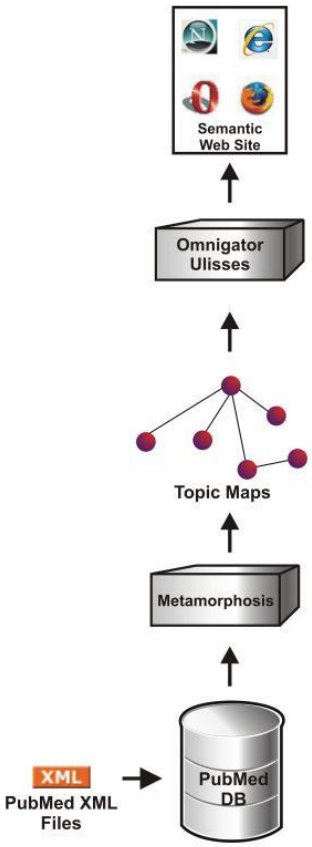


Figura 3.1. PubMed Local antes da implementação da interface.

Devido ao fato de o PubMed apresentar alguns problemas em suas consultas e, inicialmente, o PubMed Local não possuir uma interface onde se possa realizar consultas de forma simples, projetou-se uma interface para o PubMed Local que permite ao usuário a realização de consultas no banco de dados, assim como a visualização semântica do artigo. Após a interface ter sido integrada ao sistema, a estrutura do PubMed Local passou a ser como apresentado na Figura 3.2, onde a interface realiza uma consulta diretamente ao banco de dados e mostra ao usuário a rede semântica do artigo que está visualizando.

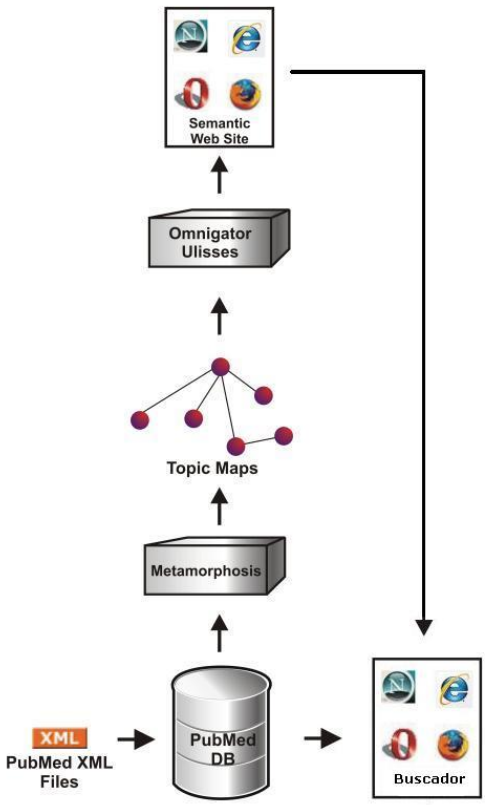


Figura 3.2. Estrutura do PubMed Local.

3.1. CRIAÇÃO DO MODELO ENTIDADE-RELACIONAMENTO PARA O PUBMED LOCAL

Inicialmente, foi feita a estruturação e criação de um banco de dados, para armazenar os mais de 50 GB de dados gerenciados pelo PubMed. Para a criação desse banco de dados, foi necessário efetuar os *downloads* dos arquivos disponíveis, através de um protocolo que está firmado entre o Centro Universitário Franciscano e a NLM. Após o *download* dos dados, mapeou-se a estruturação dos arquivos utilizados pelo NLM para uma estrutura relacional, onde seria possível a importação dos dados para o banco de dados criado.

A Figura 3.3 exibe o diagrama Entidade-Relacionamento (ER) das tabelas mais importantes da conversão dos arquivos XML para uma base de dados relacional. A ilustração mostra as principais tabelas criadas para armazenar a base de dados do PubMed. Nessa base foram removidos os itens duplicados, tornando o processo de pesquisa consistente e melhorando o tempo de resposta das consultas.

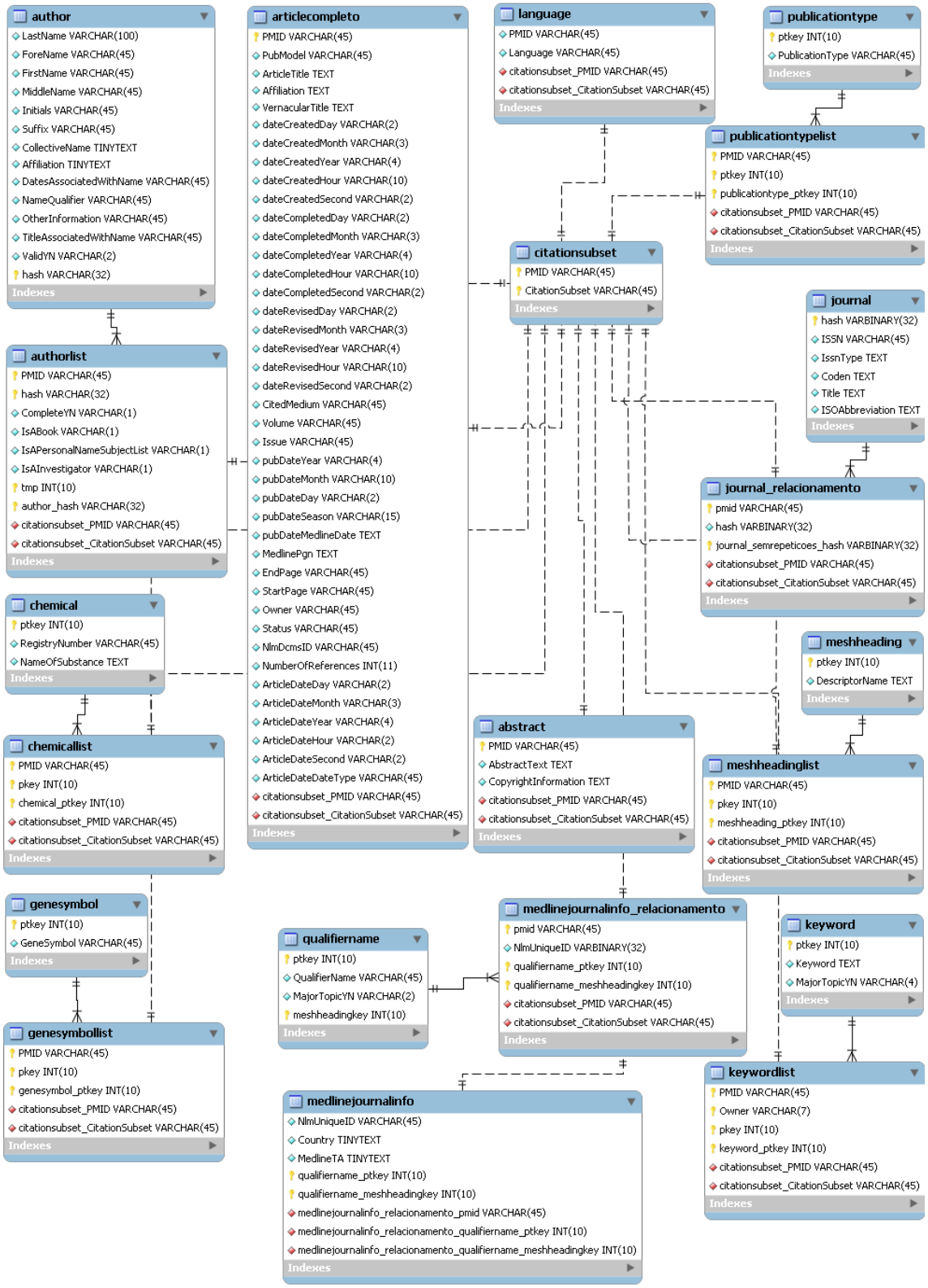


Figura 3.3. Diagrama ER do PubMed Local

Para a construção do banco de dados relacional do PubMed, construiu-se um *parser* XML em Java, para transformar os arquivos XML do PubMed em base de dados Relacional,

tornando assim, as consultas na base local mais simples e rápidas. Esse *parser* é descrito a seguir.

3.2. PARSER PARA IMPORTAR XML DO PUBMED

Devido à limitação do Exult, onde o banco de dados gerado não era satisfatório, foi necessária a construção de um *parser* específico para importar os arquivos do PubMed para uma base de dados local, otimizando as consultas.

O *parser* foi construído em Java e utiliza a biblioteca SAX (SAX PROJECT, 2008) para a leitura dos arquivos XML, devido à velocidade com que lê arquivos XML (DEVX, 2007). Foram construídas classes em Java, representando cada conjunto de *tags* do PubMed, conforme mostra o XML Schema, gerado através do DTD pelo Altova XML Spy (ALTOVA, 2007). Através das relações contidas no DTD, foi construído o modelo entidade-relacionamento, que origina a base de dados do PubMed Local.

A Figura 3.4 mostra o tempo gasto para serem inseridos alguns arquivos XML no banco de dados. Esses arquivos são denominados "medline07n0001" até "medline07n0521", onde cada um deles é constituído de 30.000 citações do PubMed. No início do processo, a importação dos arquivos XML era mais rápida do que no final, devido às consultas que eram geradas para garantir a integridade da base de dados, assegurando assim a inexistência de informações duplicadas (autores, termos MeSH, entre outras).

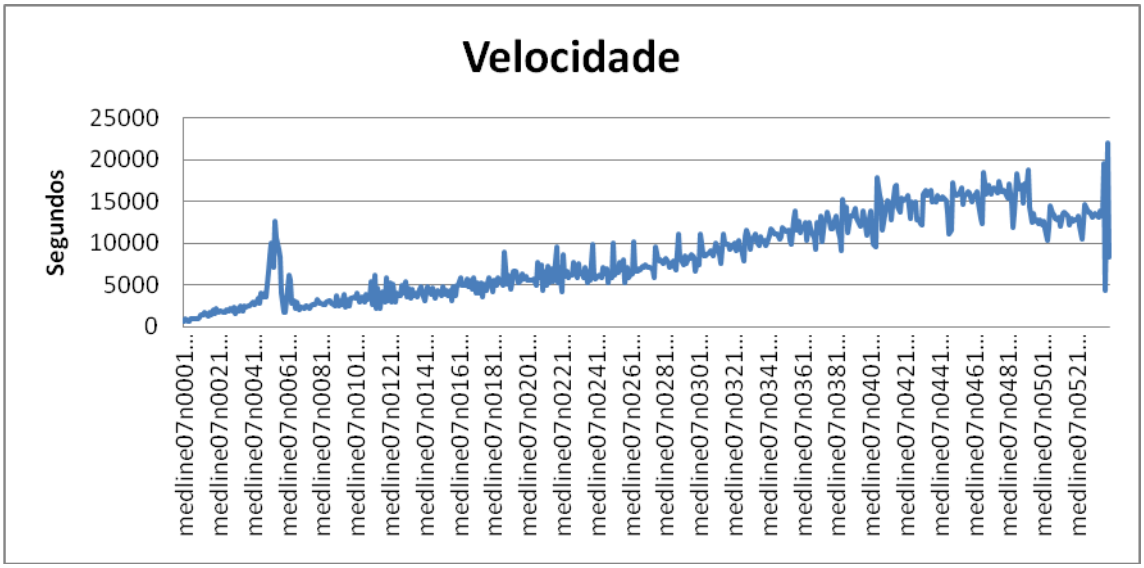


Figura 3.4. Gráfico do tempo gasto para inserir os dados ao PubMed Local.

Para a geração do banco de dados foi executado um processo do *parser*, composto de 4 *threads* simultâneas, cada uma carregando um arquivo XML por vez. O processo de inserção demorou aproximadamente 198 horas, em um computador Intel Core 2 Duo 2 6400 com 3 GB de memória RAM. O tempo apresentado foi obtido através do Firebug (FIREBUG, 2007). O tamanho total do banco de dados resultou em 37 GB.

3.3. SUMÁRIO DO CAPÍTULO

Neste capítulo, apresentou-se a estrutura do PubMed Local, mostrando os passos de forma detalhada, desde o *download* dos arquivos até o mapeamento dos arquivos XML para o banco de dados local. Até o momento, se tem pronto a base de dados que será utilizada pelo PubMed Local para a realização das consultas através da interface que será apresentada no capítulo 5. No próximo capítulo, detalha-se como ocorre a criação do Topic Map para o domínio do PubMed.

4. REPRESENTAÇÃO DO CONHECIMENTO DO DOMÍNIO DO PUBMED

As redes semânticas são bastante utilizadas na área de Inteligência Artificial (IA). Essas redes possuem objetos que são representados como nodos de um grafo, onde as relações entre objetos são representadas por arcos (LIBRELOTTO, 2005). Esses nodos são organizados em uma estrutura taxonômica, e os arcos representam as relações binárias. Tudo que pode ser expresso em lógica de primeira ordem (DALEN, 1994); (HODGES, 1997) pode também ser expresso como uma rede semântica. Um exemplo de uma rede semântica é a WordNet (FELLBAUM, 1999).

As redes semânticas têm como objetivo descrever uma estrutura capaz de representar o conhecimento através da organização de conceitos (LIBRELOTTO, 2005). Para melhor entendimento da visualização da rede semântica, a seção 4.1 apresenta o conceito de Topic Maps (BIEZUNSKY; BRYAN; NEWCOMB, 1999), enquanto que a seção 4.2 detalha a ferramenta *Metamorphosis* (LIBRELOTTO, 2005). Por fim, será apresentado o *Ulisses* (LIBRELOTTO *et al*, 2007).

4.1. TOPIC MAPS

Conforme a ISO (*International Organization for Standardization*) e a IEC (*International Electrotechnical Commission*), Topic Maps (TM) é uma norma internacional unificada para descrever estruturas de conhecimento e formalizar a sua associação com recursos de informações (BIEZUNSKY; BRYAN; NEWCOMB, 1999).

TM é uma tecnologia habilitada para a representação e manipulação do conhecimento, propiciando também uma poderosa forma de navegação sobre recursos de informação. O conceito Topic Maps pode ser definido como uma descrição de um ponto de vista sobre uma coleção de recursos, organizado formalmente por tópicos e pela ligação de partes relevantes do conjunto de informação aos tópicos apropriados. Um mapa de tópicos expressa a opinião de alguém sobre o que os tópicos são, e quais as partes do conjunto de informação que são relevantes para cada tópico. Falar sobre Topic Maps é falar sobre estruturar o conhecimento.

Os principais objetivos de Topic Maps são: (a) estruturar recursos de informação não estruturados, com mecanismos externos aos recursos; (b) permitir procuras que recuperem a

informação requisitada; e (c) criar visões diferentes para usuários ou finalidades específicas, filtrando a informação.

Um *topic map* pode ser visto como um conjunto organizado de tópicos (representação formal de temas de um universo de discurso), contendo: (a) uma estrutura hierárquica de tópicos (definido pelas relações é-um ou contém); (b) vários nomes para cada tópico (ou tema de um índice); (c) ponteiros (ocorrências) entre tópicos e documentos externos (conectando os temas aos recursos de informação); e (d) relacionamentos semânticos (associações) entre tópicos. Um Topic Map é, portanto, composto por tópicos associados que dão origem a uma rede semântica estruturada que agrupa informações relacionadas a certo domínio.

4.2. METAMORPHOSIS – UM AMBIENTE PARA EXTRAÇÃO AUTOMÁTICA DE TOPIC MAPS

A construção de uma ontologia pode ser efetuada de uma forma manual ou usando uma ferramenta para gestão de ontologias, como o Protégé (PROTÉGÉ, 2005). Além disso, existem processos automatizados para a especificação de extração de topic maps baseados em ontologias.

O *Metamorphosis* (LIBRELOTTO; RAMALHO; HENRIQUES 2006) é uma ferramenta que permite a construção automática de Topic Maps a partir de dados extraídos de diferentes recursos de informação, permitindo uma navegação semântica sobre tais recursos. A motivação principal do *Metamorphosis* é aproximar os usuários e o paradigma Topic Maps.

A figura 4.1 mostra o cenário de uso proposto. Ela ilustra algumas das interações entre os componentes do sistema, recursos de informação e usuários:

1. *Metamorphosis Repository* (MMRep): é o componente central que se responsabiliza pelo armazenamento e gerenciamento dos Topic Maps. Todos os outros componentes interagem com o MMRep;

2. *Topic Map Discovery* (TMDiscovery): é o navegador dirigido por Topic Maps que permite ao usuário navegar sobre os topic maps armazenados no MMRep;

3. *Topic Map Extractor* (Oveia): automatiza a tarefa de geração de Topic Maps. Ele habilita ao usuário a capacidade de expressar a extração de tarefas e gerar topic maps de acordo com a sintaxe XTM, a qual pode ser carregada para o MMRep. *Oveia* implementa alguns mecanismos de extração com os quais é possível popular uma ontologia automaticamente;

4. *Information resources*: os recursos de informação que servirão como fonte para a construção da ontologia;

5. *Web interface*: a partir de um topic map armazenado no MMRep, obtém-se a visão da rede semântica encontrada na ontologia em questão.

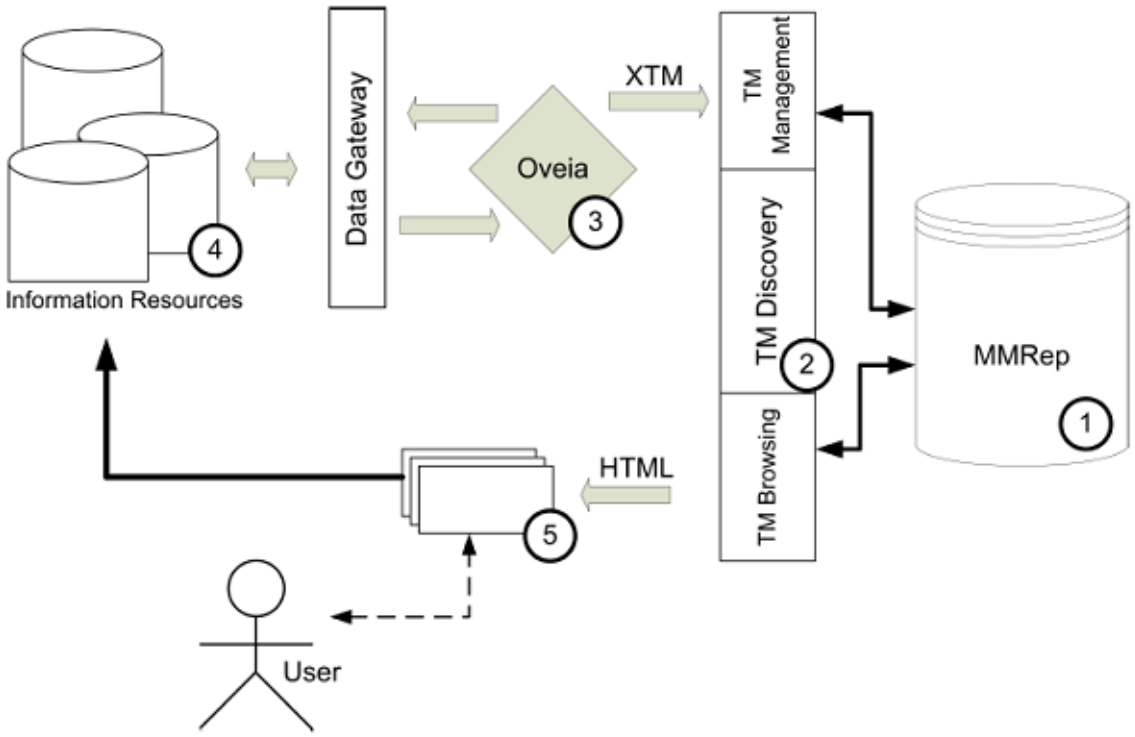


Figura 4.1. Arquitetura do *Metamorphosis* (LIBRELOTTO; RAMALHO; HENRIQUES 2006)

4.3. EXTRAÇÃO DE TOPIC MAPS

A fim de obter uma rede semântica de dados do PubMed, dividiu-se essa tarefa em algumas partes, como mostra a figura 4.2.

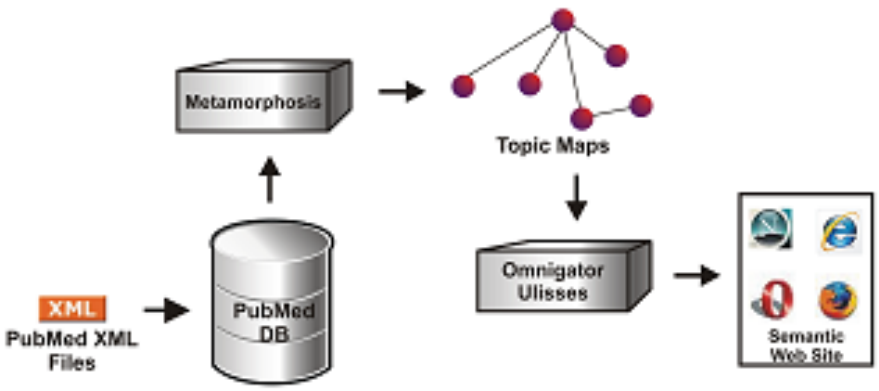


Figura 4.2. Arquitetura do sistema

No primeiro, criou-se um banco de dados relacional (PubMed Local) para armazenar todo o conteúdo dos arquivos XML obtidos do PubMed. Esta base de dados é gerada através do DTD do PubMed usando um *parser* criado. Um *script* SQL processa o resultado da base de dados para remover dados redundantes e apagar tabelas desnecessárias.

Para extrair dados a partir desta base de dados usou-se o *Metamorphosis*. O *Metamorphosis* possui mecanismos para consultar a base de dados local do PubMed, seguindo uma especificação de uma ontologia (XS4TM). Além disso, existe uma interface Web para fazer uma consulta sobre a base de dados. Esta interface possui um campo texto para o usuário realizar as suas consultas. Após a submissão da consulta, o *Metamorphosis* processa o resultado procurando termos MeSH que descreve as publicações desejadas. Estes termos estão estruturados em arquivos RDF (MARK; ALISTAIR; GUUS M, 2007).

Usando esses termos MeSH como palavras-chave, o *Metamorphosis* busca artigos que correspondam com a consulta do usuário. Esse processo de busca inclui diversos campos, como título do artigo, *abstract*, palavras-chave, as substâncias químicas e termos MeSH. Quando um artigo satisfaz a consulta, será mapeado para um tópico, bem como as suas principais áreas, criando associações entre eles.

Quando o sistema recebe uma solicitação, os dados requeridos serão coletados a partir da base de dados e selecionados no momento da execução. Em seguida, eles irão continuar a serem processados e convertidos em dados relevantes semanticamente pelo *Metamorphosis*. Os dados resultantes têm o formato padrão XTM. Uma das vantagens desta abordagem é que nenhum novo banco de dados será criado e não serão produzidos dados redundantes.

Após o fim do processo, o *Metamorphosis* tem todos os tópicos e as associações armazenadas em seu repositório. Os documentos XTM gerados podem ser processados e exibidos para o usuário, mediante a camada de apresentação. Desta forma, qualquer navegador de Topic Maps é capaz de navegar na rede semântica composta por esses conceitos.

4.3.1. Definindo os conceitos Topic Maps para as citações do PubMed

Para a extração de Topic Maps do PubMed, inicialmente, é necessário definir quais são os conceitos principais, ou seja, quais são os tópicos desse domínio em estudo. Desta forma, os tópicos escolhidos foram:

***Article* :** cada artigo é armazenado em uma *tag* chamada <*MedlineCitation*>*;*

Author: os autores do artigo são declarados em <*Author*>;

Keyword: as palavras-chaves são termos MeSH. Elas são definidas em <*MeshHeading*>;

Publication year: este metadata está no caminho //*PubDate/Year;*

Journal: todos os jornais são encontrados na *tag* <*Journal*>;

Language: as linguagens dos artigos estão definidas em <*Language*>;

Chemical substances: todas as substâncias químicas citadas em cada artigo são referenciadas em <*Chemical*>;

A partir da escolha dos tópicos a serem extraídos do sistema, parte-se, para a definição das características de cada tópico (abaixo se apresentam as principais):

(1) Artigo: identificador (PMID), título, paginação, abstract, afiliação e DOI;

(2) Autor: iniciais, nome, nome do meio e sobrenome;

(3) Palavras-chave: termos descritores e termos qualificadores;

(4) Jornais: identificador (ISSN), título, abreviação, volume, série, data de publicação;

(5) Substâncias químicas: o número de registro e o nome da substância.

Na fase seguinte, especificam-se quais são as associações que deverão ser encontradas no *topic map* gerado: (a) autor escreve artigo; (b) artigo é descrito por suas palavras-chaves; (c) artigo foi publicado em um ano de publicação; (d) jornal é composto por artigos; (e) jornal referencia palavras-chave; (f) artigo é escrito em um idioma; (g) artigo cita substâncias químicas; (h) jornal é publicado em um ano; (i) autor publica em um ano; (j) artigo é publicado em um jornal; e (k) autor escreve em um idioma.

Olhando para um TM pode-se pensar como tendo duas partes distintas: uma ontologia e um catálogo de objetos. A ontologia é definida por aquilo que designa como tipo de tópico, tipo de associação, e regras de associação. O catálogo é composto por conjunto de informações objetos que estão presentes em recursos de informações (um objeto pode ter múltiplas ocorrências nos recursos de informações) e que estão vinculadas à ontologia.

O *topic map* do PubMed definido acima pela ontologia (tipo de tópico, regras e tipos de associação) e as características do tópico são mapeadas para a especificação XS4TM.

A especificação XS4TM descrevendo o cenário do PubMed foi definida em um editor Web XS4TM. A figura 4.3 mostra uma visão da presente especificação, com sete tipos de tópicos definidos, nove tipos de associações, e dezoito tipos de regras.

No lado esquerdo, XS4TM apresenta a árvore XML extraída do XML Schema do PubMed. Os tipos de tópicos deste estudo de caso são mostrados na janela ao centro. Para criar um novo tipo de tópico, o usuário só precisa fazer um simples arrastar e soltar a partir da

árvore XML. As características do tópico são definidas na primeira coluna e as características da associação são definidas na última coluna.

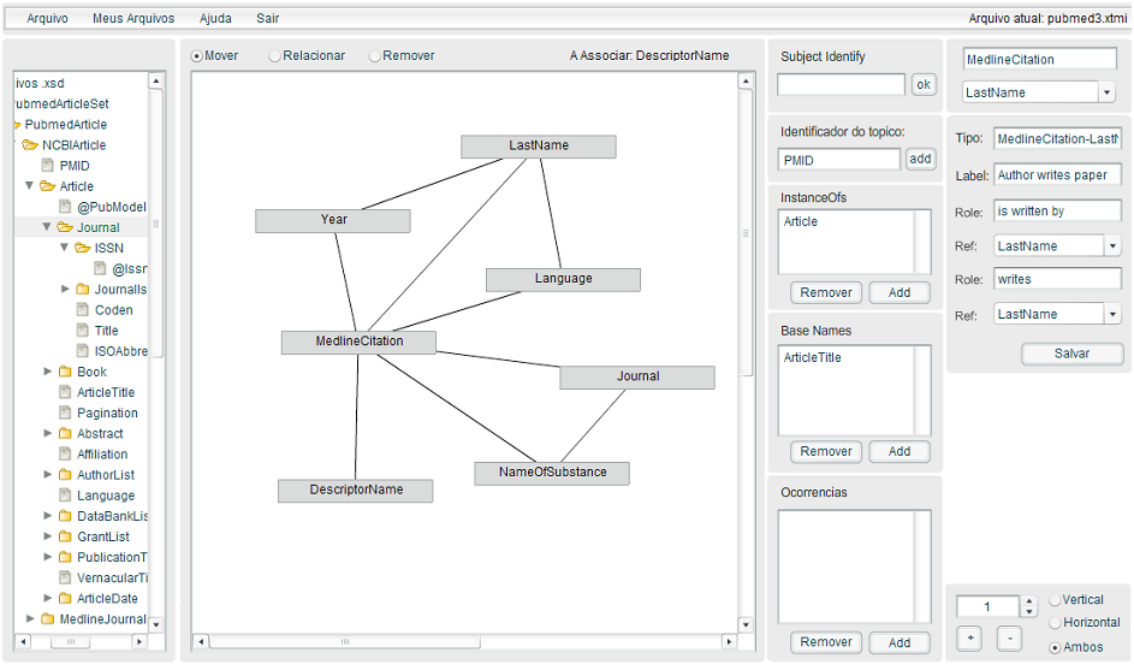


Figura 4.3. Especificação XS4TM do PubMed.

Com a especificação XS4TM completa, o Oveia[1] pode processá-la. Seu comportamento pode ser descrito em quatro partes: lê a especificação XS4TM, extrai os temas e as associações do resultado da consulta definida, cria o *topic map*, e armazena-o no repositório.

4.4. NAVEGADOR DO TOPIC MAP

A navegação dos topic maps baseada na ferramenta Ulisses (LIBRELOTTO *et al*, 2007). A idéia sobre a qual se baseou o Ulisses é navegação conceitual, que pode ser descrita como: quando se está posicionado sobre um determinado conceito, a ferramenta de navegação mostrará as informações associadas a este conceito em particular; se for escolhido algum dos outros conceitos relacionados, a navegação muda para a visão desse novo conceito; se for escolhido algum dos recursos de informação, o sistema mostrará o conteúdo do próprio recurso (LIBRELOTTO *et al*, 2007). No capítulo 5 será apresentada, a interface do PubMed Local, já com a opção de visualização da rede semântica integrada a ela.

[1] Oveia é um módulo do *Metamorphosis*.

4.5. SUMÁRIO DO CAPÍTULO

Neste capítulo apresentou-se o processo de criação do *topic map*, bem como as ferramentas utilizadas para a sua extração e navegação. Com isso, se tem a definição da ontologia para o PubMed que serve para realizar a extração dos Topic Maps, utilizando a ferramenta *Metamorposhis.* Já no capítulo a seguir, será apresentada a interface do PubMed Local para a realização das consultas já integrada com a visualização da rede semântica.

5. INTERFACE PARA CONSULTAS PERSONALIZADAS NO PUBMED LOCAL

Como mostrado no Capítulo 2.3, o PubMed apresenta diversos problemas em seus métodos de busca. Esses problemas foram levados em consideração para a criação da interface de busca para o PubMed Local.

No PubMed Local, assim como no PubMed, é possível executar diferentes formas de buscas, nas quais podem ser definidas busca padrão, busca com filtros, busca com prioridades e, além dos métodos de busca citados anteriormente, é possível visualizar a rede semântica do artigo que se está visualizando.

5.1. UTILIZAÇÃO DA INTERFACE PARA CONSULTA NO PUBMED LOCAL

Na busca padrão, o usuário insere os termos que deseja efetuar sua busca, e recebe como resposta os artigos que possuem o maior número de referências ao termo buscado, sem se preocupar em que parte do artigo está contido o termo escolhido. Essa forma de busca nem sempre retorna o resultado desejado, por se tratar de uma busca genérica, podendo fazer com que o usuário tenha que verificar dezenas de artigos, antes de achar o que realmente é de seu interesse. A Figura 5.1 ilustra a busca padrão, utilizando os termos "*dna gene expression*" como parâmetro de busca.

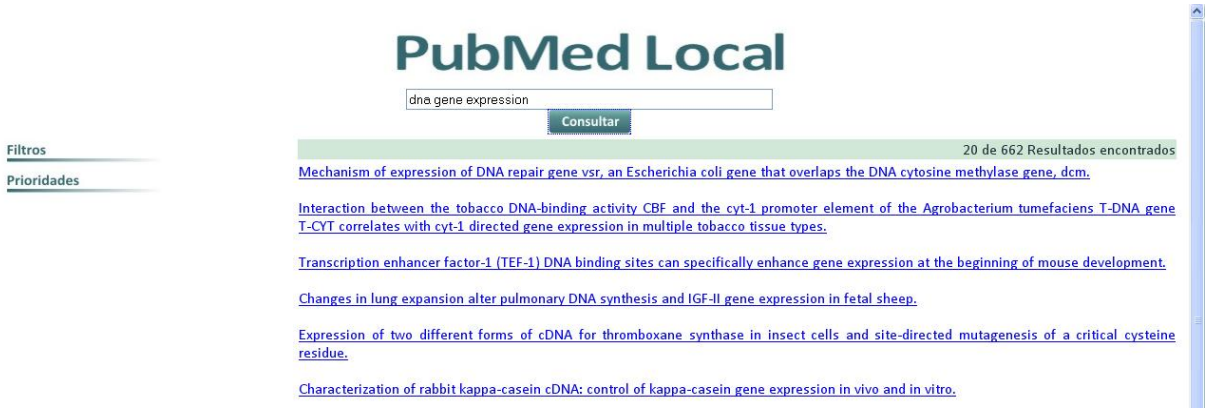


Figura 5.1: Busca padrão, onde o usuário apenas digita os termos para busca e manda consultar.

Na busca com filtros, o usuário pode definir filtros que serão aplicados na busca principal. Dessa forma é possível obter um resultado mais adequado, com um menor número de resultados retornados. Os filtros que o usuário pode definir são: ano de publicação, mês de publicação, termos MeSH, autor e origem. A Figura 5.2 ilustra a mesma busca feita

anteriormente, usando os termos "*dna gene expression*" como parâmetro de busca, e utilizando um filtro para melhorar o resultado da busca. O filtro adotado no campo origem foi o país "*England*". Nesse caso, obtém-se 523 resultados a menos do que foi obtido na Figura 5.1.

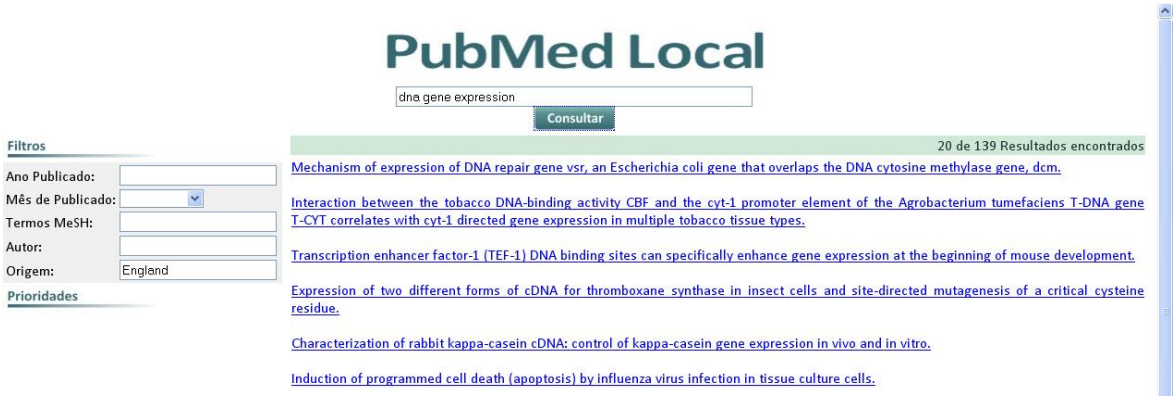


Figura 5.2: Busca com filtros.

A busca com prioridades é o que se definiu como busca avançada, pois com ela o usuário pode definir qual é o parâmetro mais importante, qual o menos importante, e quais podem ser ignorados, no momento da busca. Dessa forma, o usuário consegue obter um melhor resultado, tendo que alterar um menor número de vezes os termos de sua busca. As prioridades que podem ser definidas são: autor, titulo, palavras-chave, termos MeSH, local de publicação e termos químicos. A Figura 5.3 ilustra a busca avançada, utilizando, como nos exemplos anteriores, os termos "*dna gene expression*" como parâmetro de busca. Nesse caso, foi definido que as palavras chaves dos artigos seriam o mais importante a ser buscado, ignorando-se o resto dos parâmetros. Logo, o resultado retornado foi alterado em relação à primeira busca, onde todas as prioridades foram definidas com o mesmo peso.

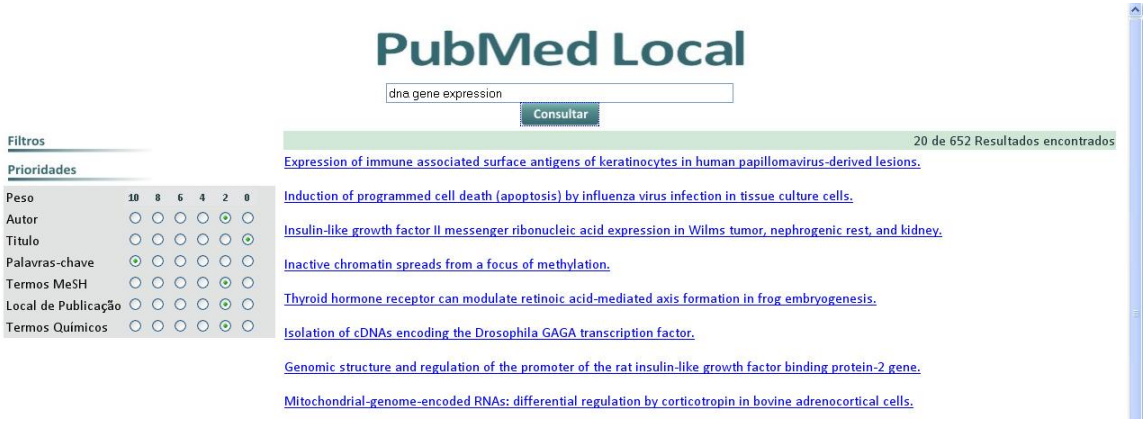


Figura 5.3: Busca avançada.

Na Figura 5.4 é mostrado um exemplo, ilustrando todos os tipos de busca que podem ser utilizados. Como nos exemplos anteriores, os termos "*dna gene expression*" foram adotados como parâmetro de busca. O filtro "*England*" foi adotado como local de origem, e as prioridades foram deixadas como padrão, com exceção das palavras-chave, às quais foram dadas prioridade máxima, e o titulo, que foi deixado com prioridade mínima, fazendo com que não fosse executada busca dos mesmos.

Figura 5.4: Busca avançada com filtros.

Quanto ao resultado obtido em cada uma das buscas já citadas, mostra-se o título do artigo, que contém o determinado termo. Ao clicar no título do artigo, é mostrado, ao usuário, o início do resumo, da mesma forma que, quando o artigo não possui um resumo cadastrado no banco de dados, é mostrada a mensagem de que o resumo não está disponível. Também é mostrado um *link* para que o usuário possa ver mais informações sobre o artigo. A Figura 5.5 ilustra o que foi apresentado nesse parágrafo.

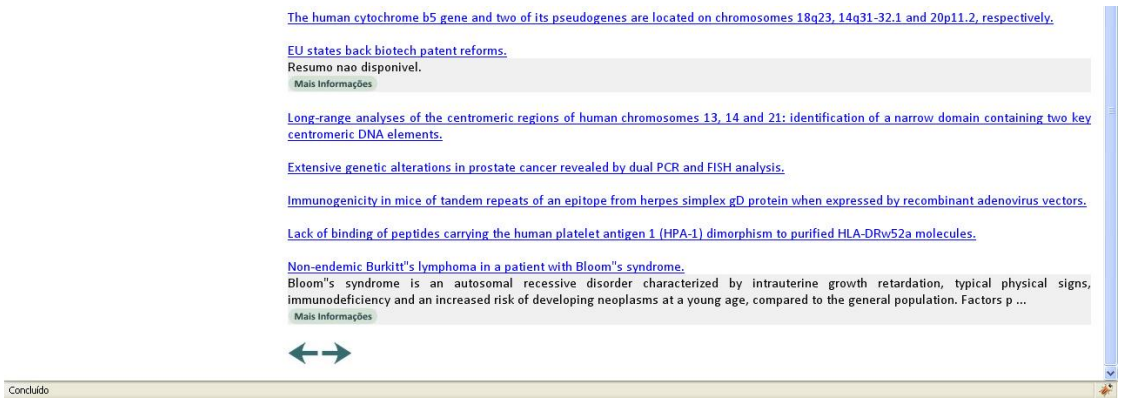


Figura 5.5: Resultados obtidos.

Ao acessar as informações adicionais do artigo, clicando no *link* “Mais informações”, serão mostradas informações detalhadas sobre o artigo, em uma nova página. Dessa forma, o usuário é capaz de ver os autores, local de publicação, resumo completo e o ano de publicação do mesmo, além de disponibilizar mais duas opções para o usuário.

Primeiro, o usuário pode acessar a rede semântica, visualizando quais são os artigos relacionados ao que ele está vendo no momento. Essa rede permite ao usuário ter a visualização gráfica da relação do artigo selecionado com os artigos relacionados.

Além disso, está disponível para o usuário o *link* para ele acessar diretamente o artigo que está visualizando, no momento, no PubMed. Dessa forma, caso o PubMed esteja disponibilizando esse artigo para visualização completa, além do seu resumo, o usuário terá acesso ao mesmo, sem ter que executar uma consulta diretamente no PubMed, somente para visualizar o artigo que achou anteriormente no PubMed Local.

Essa página, além de mostrar as informações sobre o artigo, também mostra ao usuário artigos que estão relacionados ao artigo que está sendo visualizado. A Figura 5.6 ilustra o que foi apresentado neste parágrafo.

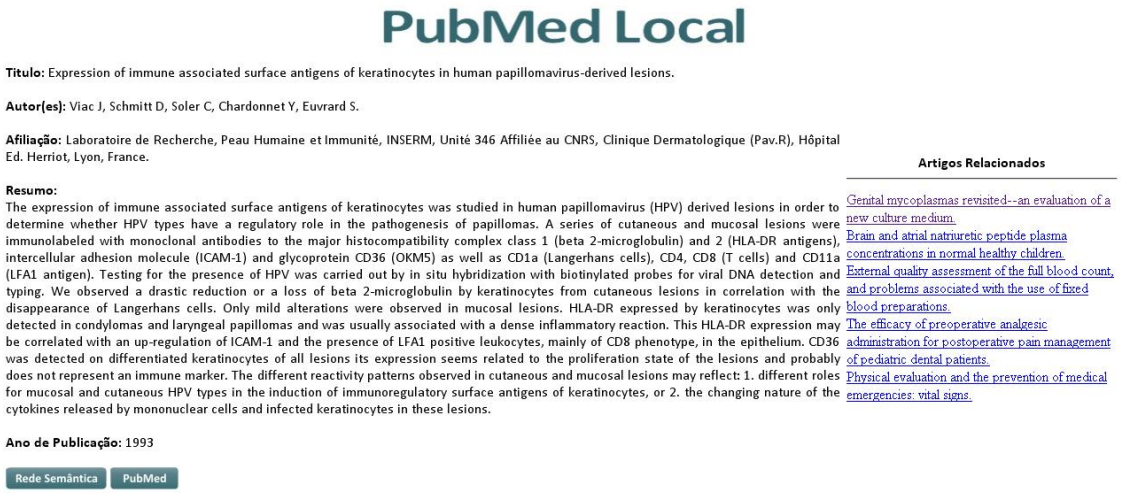


Figura 5.6: Página para visualização das informações adicionais do artigo.

Conforme citado anteriormente, ao clicar no link “Rede Semântica” será mostrada ao usuário a visualização gráfica do artigo selecionado, com os artigos relacionados à sua volta. A visualização semântica foi integrada à interface através do Ulisses. A Figura 5.7 exemplifica a rede semântica.

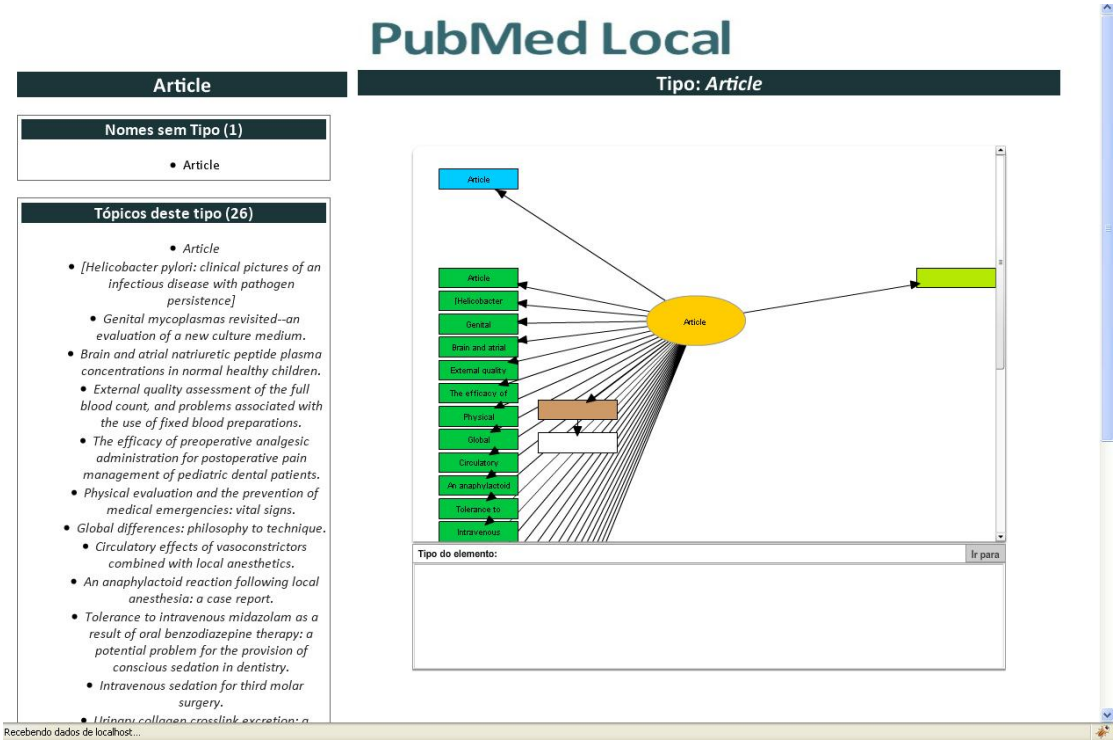


Figura 5.7: Visualização da rede semântica.

Por último, o *link* para o PubMed, como citado anteriormente, permite ao usuário que com apenas um clique seja direcionado para o PubMed, onde será mostrado o conteúdo apresentado no PubMed Local, além de também poder apresentar ao usuário a opção para *download* do artigo completo, caso o artigo em questão esteja disponível para *download*, como pode ser visto na Figura 5.8.

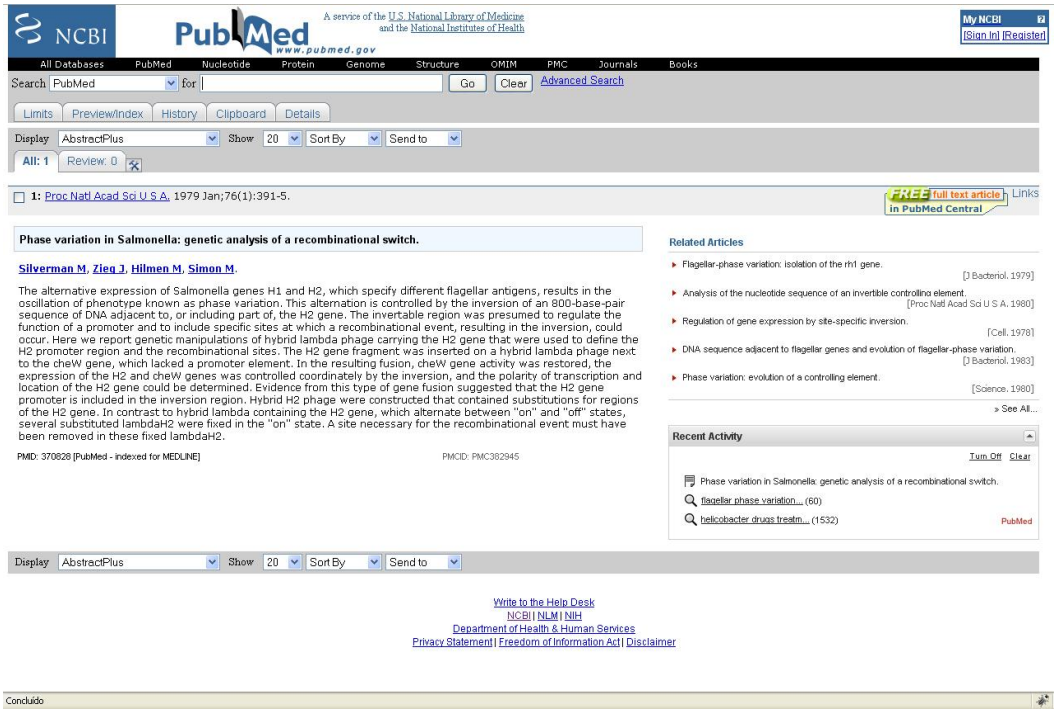


Figura 5.8: Artigo contendo link para visualização completa do mesmo. (NCBI, 2007)

5.2. CONSTRUÇÃO DA INTERFACE PARA CONSULTA NO PUBMED LOCAL

Esta seção mostra como essa interface foi construída e quais ferramentas foram utilizadas. Para construção da interface foi utilizado PHP (*Hypertext Preprocessor*) (PHP, 2008), uma linguagem de programação amplamente utilizada no desenvolvimento *web*, que pode ser mesclada dentro do código HTML (*HyperText Markup Language*), que por sua vez é uma linguagem de marcação, utilizada na criação de páginas para a internet (W3SCHOOLS, 2008) e Linguagem SQL (*Structured Query Language* ou Linguagem de Consulta Estruturada**)** para a realização das consultas.

Além disso, como suporte ao HTML e PHP, foram utilizadas funções em JavaScript, que é uma linguagem de *scripts* para a *web.* Com ela é possível adicionar diversas funcionalidades aos *websites*, validar formulários, detectar qual o navegador está sendo usado pelo usuário, entre outras (JAVASCRIPT, 2008). Uma das funcionalidades usadas é dar o efeito dos menus de “Filtros” e “Prioridades”, como pode ser visto na Figura 5.9.

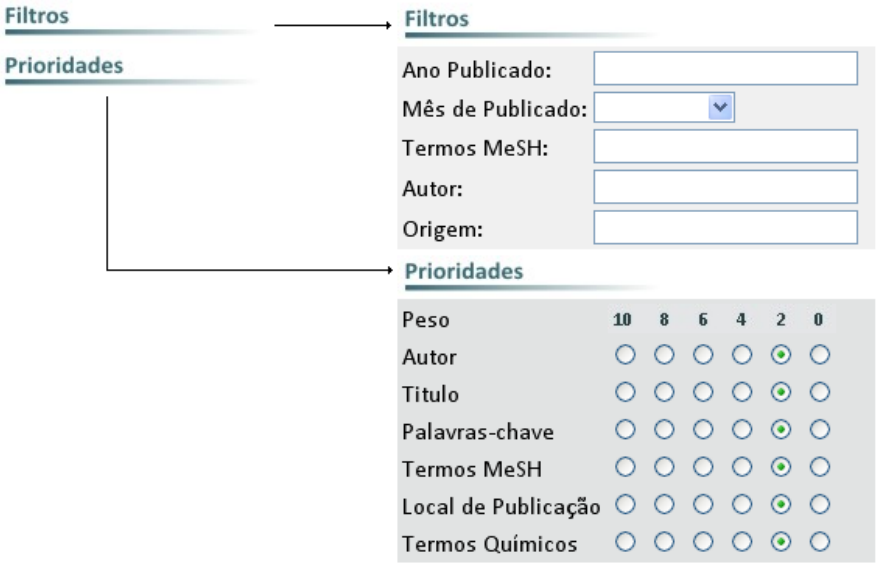


Figura 5.91: Exemplo da utilização do JavaScript.

Além da utilização do JavaScript em sua forma padrão, também foi utilizado o conceito de AJAX (*Asynchronous JavaScript And XML*) (AJAX, 2008), que se constitui em uma nova forma de utilizar o JavaScript e o XML. Com ele é possível desenvolver *websites* melhores, mais velozes, e com uma interface mais *amigável* (AJAX, 2008). O AJAX foi utilizado para que seja possível a execução da página de resultados dentro de uma célula da tabela, em que a interface é montada. Dessa forma, é utilizado o código gratuito disponibilizado (DINAMIC, 2008).

Primeiramente, é criada uma tabela temporária, que tem como objetivo armazenar o resultado gerado das consultas. A tabela *consulta* não faz parte do ER do banco de dados; ela é apenas uma tabela temporária, onde o resultado de cada consulta executado tem seus valores inseridos. Após a execução de todas as consultas, essa tabela é utilizada para mostrar o resultado obtido.

Para as consultas, utiliza-se a idéia das prioridades, independente do método de busca utilizado. Quando uma busca é efetuada, realiza-se o seguinte procedimento:

1. caso exista a tabela *consulta*, onde é armazenado o resultado das consultas já realizadas, ela é removida para que não sobrecarregue o banco de dados com consultas já realizadas.
2. cria-se a tabela temporária *consulta*.

As consultas foram projetadas utilizando o MySQL 5 (MYSQL, 2008) O MySQL é um banco de dados completo, robusto e extremamente rápido, com todas as características existentes nos principais bancos de dados disponíveis no mercado. Uma de suas

peculiaridades são suas licenças para uso gratuito, tanto para fins estudantis como para realização de negócios, possibilitando que empresas o utilizem livremente.

Inicialmente, é feito uma consulta do PMID (*PubMed Identificator*) e do peso na tabela em questão. O peso é um campo criado para que seja possível definir o quanto os artigos são referenciados. Para exemplificar, adota-se a tabela *keywordlist*, como se pode ver na Figura 3.3, apresentada na seção 3.1, pois essa é a tabela que contém o PMID e a chave necessária para ligar as consultas à tabela *keyword*, que contém a palavra-chave buscada.

Para fazer a ligação das duas tabelas, utilizou-se o *left join* (DUBOIS, 2008), que mostrou melhor resultado nos testes feitos. Após o *join* ter sido executado, fez-se o *where*, que vai chamar a função para verificar se o termo recebido como parâmetro possui espaços e, após a verificação, utilizado um *insert*, para inserir o PMID e o peso do termo buscado.

Após o *insert* ser executado, ele busca pelo termo *insert* e, então, caso possua mais de um *insert* para ser executado, ele atribui os mesmos para as variáveis temporárias. Isso é necessário pelo fato da função *mysql_query* não suportar múltiplos *inserts* (DUBOIS, 2008). Essa distribuição nas variáveis temporárias faz com que cada termo inserido para a busca execute uma consulta separadamente e, dessa forma, é feita uma busca mais ampla. Mas, ao mesmo tempo, as prioridades passam a ter mais importância, pois cada vez que o mesmo artigo for referenciado, sua prioridade vai aumentar. Um exemplo da consulta é apresentado na Figura 5.10.

```
if ( $_GET['palavrachave'] != 0 ) {
    $sql = "INSERT INTO consulta SELECT pmid AS PMID, ".$_GET['palavrachave']."
            AS PESO FROM keywordlist
            LEFT JOIN keyword on keywordlist.pkey = keyword.ptkey
            WHERE ".VerificaEspaco2("keyword", $_GET['query'],
            $_GET['palavrachave'], "keywordlist", "keyword", "pkey", "ptkey")."
            ORDER BY pmid;";
    $tmp = explode("INSERT", $sql);
    foreach($tmp as $sql ) {
        if ( strlen($sql) > 10 ){
            $sql = "INSERT " . $sql;

            mysql_query($sql) or die ('Error Keyword: '.mysql_error ());
        }
    }
}
```

Figura 5.10: Exemplo da consulta SQL utilizada nas consultas por prioridades ao banco de dados.

A função *VerificaEspaco*, assim como a *VerificaEspaco2*, buscam por espaço no termo que o usuário definiu para a busca. Dessa maneira, se o usuário definir duas ou mais

palavras para sua busca, a função irá dividir o termo de busca em duas partes, e executar o SQL como se houvessem sido definidos dois termos para a busca. Para que a busca por espaços seja executada, utilizou-se a função do PHP chamada *str_replace* (WELLING; THOMSON, 2008), que substitui o espaço por "*keyword LIKE '%câncer%' INSERT INTO consulta SELECT pmid AS PMID, "2" AS PESO FROM keywordlist LEFT JOIN keyword ON keywordlist.pkey = keyword.ptkey WHERE keyword LIKE '%human%*"", para onde os termos buscados sejam "*câncer human*". A Figura 5.11 mostra as funções completas, sendo a primeira para consultas simples e a segunda para as consultas onde é necessária a ligação de duas tabelas.

```
function VerificaEspaco ($chave, $txt, $peso, $tabela) {
    $sqls = $chave." LIKE '%".str_replace(' ', "%'; INSERT INTO consulta
    SELECT pmid AS PMID, ".$peso." AS PESO FROM ".$tabela."
    WHERE ".$chave." LIKE '%", $txt)."%'";
    return $sqls;
}

function VerificaEspaco2 ($chave, $txt, $peso, $tabela,
$tabela2, $chave2, $chave3) {
    $sqls = $chave." LIKE '%".str_replace(' ', "%'; INSERT INTO consulta
    SELECT pmid AS PMID, ".$peso." AS PESO FROM ".$tabela." LEFT JOIN ".$tabela2."
    ON ".$tabela.".".$chave2." = ".$tabela2.".".$chave3."
    WHERE ".$chave." LIKE '%", $txt)."%'";
    return $sqls;
}
```

Figura 5.11: função que verifica espaços no termo utilizado para a consulta.

Após todas as consultas terem sido processadas, e todos os PMID (juntamente com seus pesos) estarem inseridos na tabela consulta, seleciona-se qual a condição a ser executada para mostrar o resultado.

Caso nenhum filtro tenha sido definido, será apenas realizada uma consulta na tabela consulta, e usado o *left join* novamente, para ligar a tabela *consulta* à tabela *articlecompleto*. A partir desse resultado, os pesos dos artigos que têm o mesmo PMID são somados e mostrados em ordem decrescente, para que o usuário visualize primeiro os artigos que foram referenciados o maior número de vezes nas buscas executadas. A Figura 5.12 ilustra a situação apresentada.

```
if ( $_GET['fano'] == NULL AND $_GET['ftermomesh'] == NULL
AND $_GET['fautor'] == NULL AND $_GET['forigem'] == NULL
AND $_GET['fmes'] == NULL ) {
    $sql = "SELECT consulta.pmid, SUM(PESO), articletitle
            FROM consulta
            LEFT JOIN articlecompleto
            ON consulta.pmid = articlecompleto.pmid
            GROUP BY pmid ORDER BY SUM(PESO) DESC, PMID ASC LIMIT 0, 20;";
$resposta = mysql_query($sql) or die ('Error R: '.mysql_error ());
}
```

Figura 5.12: Consulta que gera o resultado quando não foi definido nenhum filtro.

Quando analisada a construção das consultas, a diferença entre a busca padrão e a busca por prioridades é muito pequena. A busca por prioridades recebe como parâmetro do usuário qual o peso que ele julga mais importante para sua consulta, enquanto a busca padrão tem as prioridades definidas como padrão, com o peso 2.

O SQL apresentado na Figura 5.9 é um SQL por prioridades, enquanto o SQL apresentado na Figura 5.13 é um SQL padrão. Quando executada uma busca padrão, ambos os SQL são executados. Caso seja executada uma busca avançada, o SQL da Figura 5.10 pode não ser executado, se sua prioridade for definida como zero.

```
$sql = "INSERT INTO consulta SELECT pmid AS PMID, \"2\" AS PESO
        FROM journal_relacionamento
        LEFT JOIN journal_semrepeticoes
        ON journal_relacionamento.hash = journal_semrepeticoes.hash
        WHERE ".VerificaEspaco2("title", $_GET['query'], "2",
        "journal_relacionamento", "journal_semrepeticoes", "hash", "hash")."
        ORDER BY pmid;";
$tmp = explode("INSERT", $sql);
    foreach($tmp as $sql ) {
        if ( strlen($sql) > 10 ){
            $sql = "INSERT " . $sql;
            mysql_query($sql) or die ('Error Keyword: '.mysql_error ());
        }
    }
```

Figura 5.13: SQL de uma busca padrão

Quanto às buscas com filtros, as consultas e inserções dos dados na tabela consulta são realizadas da mesma maneira que a consulta padrão e a consulta com prioridades; o que diferencia a consulta com filtros é o SQL necessário para gerar o resultado. Para isso, primeiro é testado para ver qual ou quais os filtros foram definidos pelo usuário. Após isso, é gerado o resultado, baseado nos filtros definidos. Para gerar o resultado, a consulta é feita na tabela *consulta* como nos outros tipos de buscas. Entretanto, quando os filtros estão definidos,

além de fazer a busca pelo PMID para mostrar o título do artigo, é utilizado o *where* para limitar a busca nos artigos que possuam o determinado filtro aplicado, como pode ser visto na Figura 5.14.

```
else if ( $_GET['fano'] != NULL AND $_GET['ftermomesh'] == NULL
AND $_GET['fautor'] == NULL AND $_GET['forigem'] == NULL
AND $_GET['fmes'] == NULL ) {
    $sql = "SELECT consulta.pmid, SUM(PESO), articletitle
            FROM consulta
            LEFT JOIN articlecompleto
            ON consulta.pmid = articlecompleto.pmid
            WHERE pubdateyear like '%".$_GET['fano']."%'
            GROUP BY pmid ORDER BY SUM(PESO) DESC, PMID ASC LIMIT 0, 20;";
$resposta = mysql_query($sql) or die ('Error R: '.mysql_error ());
}
```

Figura 5.14: Exemplo da busca com filtros quando o campo dos termos MeSH foi preenchido.

Após o resultado ser gerado, começa o processo para mostrar ao usuário o resultado da sua consulta. Para o usuário-final, são mostrados os títulos dos artigos que contenham o termo que foi buscado, além de ser dado ao usuário à opção para selecionar o título do artigo para obter mais informações sobre o mesmo. Nesse caso, mostram-se os primeiros 250 caracteres do resumo do artigo escolhido.

Para facilitar a navegação do usuário na interface, foi desenvolvido um algoritmo para que sejam mostrados apenas 20 resultados por página, como visto na Figura 5.5.

5.3. SUMÁRIO DO CAPÍTULO

Neste capítulo foi mostrado como ocorrem as consultas no PubMed Local, realizando buscas padrão, utilizando filtros e usando prioridades nas buscas para tornar o resultado da pesquisa o mais satisfatório possível. Também foi abordado como a interface foi desenvolvida, e ainda explicando o funcionamento da interface, desde uma consulta padrão até a visualização da rede semântica. Neste ponto, o PubMed Local está completo, possuindo a base de dados, a interface para consulta, contendo a opção de utilização de filtros e prioridades, e visualização da rede semântica.

No capítulo seguinte, será apresentada, a mineração de dados realizada com alguns dados contidos no PubMed, sendo utilizadas técnicas de regras de associação e redes de dependência.

6. DESCOBERTA DE CONHECIMENTO NO PUBMED

O grande volume de dados e informações disponível no PubMed, tem desafiado a habilidade dos seres humanos em interpretar e compreender esses dados e informações. Assim, uma nova geração de técnicas e ferramentas computacionais que, automaticamente e inteligentemente, processam e analisam os dados estão sendo desenvolvidas para atender esse desafio. Essas técnicas derivam de diversas áreas, sendo partes de um processo denominado de *Knowledge Discovery in Databases* (KDD) definido por (FAYYAD *et al*, 1996) como: "*o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis presentes nos dados."*

O processo de KDD é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Esse conjunto é composto de etapas: seleção de dados; pré-processamento e limpeza; transformação; *Data Mining*; interpretação (GIMENES, 2008).

O processo de Extração de Conhecimento de Bases de Dados inicia-se com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (Base de Dados da qual se pretende extrair conhecimento). (REZENDE, 2005).

A seleção dos dados é o próximo passo. Consiste em selecionar e coletar o conjunto de dados ou variáveis necessárias.

A etapa da limpeza vem a seguir, visando adequá-los às técnicas. Isso se faz através da integração de dados heterogêneos, eliminação de incompletude, correção de erros, eliminação de dados redundantes, eliminação de valores não pertencentes ao domínio, etc.

Os dados pré-processados devem ainda passar por uma transformação que os formata adequadamente, visando facilitar o uso das técnicas de mineração de dados. Prosseguindo, chega-se à fase de mineração de dados, que começa com a escolha das técnicas a serem aplicadas com fins de encontrar padrões nos dados.

Para tanto serão utilizados regras de associação e redes de dependência. Regras associativas encontram registros de dados que de alguma maneira estão ou devem estar relacionados. As redes de dependência são modelos gráficos que representam distribuições conjuntas para um conjunto de variáveis.

Após a interpretação das informações obtidas será avaliado o conhecimento gerado, compreendendo e validando a relevância do mesmo para a pesquisa científica biomédica.

Para este trabalho foram aplicadas técnicas de *data mining*, regras de associação (AGRAWAL; SRIKANT 1995) e redes de dependência (HECKERMAN *et al*, 2000), para os dados fornecidos pelo PubMed, incidiu sobre o período de 2003 a 2005.

Essa análise irá ajudar a compreender melhor as características da investigação referenciada neste banco de dados. Este trabalho descreve alguns resultados de investigação que visa o conhecimento a partir da extração de dados PubMed.

6.1. SELEÇÃO DOS DADOS

Um arquivo XML do PubMed tem grande massa de informação. Neste estudo, vamos nos concentrar num determinado conjunto de informações que fornece uma descrição geral de uma investigação dos artigos.

É constituído por: identificador PubMed, ano de publicação, termos MeSH e nome dos Autores. Um *parser* lê vários arquivos XML e gera novos arquivos texto onde cada linha refere para uma entrada de cada citação. O exemplo abaixo mostra uma entrada sobre os dados gerados por uma citação no PubMed. O artigo entrada tem cinco palavras-chave e oito autores. Os nomes dos autores da citação aparecem em letras maiúsculas.

> 17191901, 2004, Erythrina, Plant Extracts, Plant Roots, chemistry, isolation purification, TANAKA_H, HIRATA_M, ETOH_H, SATO_M, MURATA_J, MURATA_H, DARNAEDI_D, FUKAI_T

6.2. PRÉ-PROCESSAMENTO DOS DADOS

Nesta etapa, os dados de entrada correspondem a três dados obtidos, respectivamente, para o ano de 2003, 2004 e 2005 e de acordo com o procedimento descrito no capítulo 6.1. O pré-processamento e análise dos dados foram realizados em uma etapa do *SQL Server 2005 Database* (COMMUNITY, 2007). Várias consultas SQL foram aplicadas para obter dados instantâneos de distribuição, bem como para realizar transformação de dados para posterior análise. A primeira operação consistiu na remoção do ruído a partir de dados, tais como caracteres irregulares. Um segundo passo permitiu organizar os dados para acesso mais eficiente. Uma dessas transformações realizadas nesta última etapa consiste na geração de dois *datasets* separados com palavras-chave e dados dos autores. Ambos os *datasets* foram

concebidos para representar uma relação um-para-muitos e foram extraídos de forma independente. A etapa de extração de dados foi realizada com o *SQL Server Business Intelligence (BI) Development Studio.*

A Tabela 6.1 resume os tamanhos dos dados. O conjunto de dados (dsPM) corresponde à entrada dos dados obtidos através da análise dos arquivos XML do PubMed. Para a análise foram apenas consideradas as 5000 palavras-chave mais freqüentes (Top-5-KW) e os 1000 mais frequentes autores (Top-1-A).

Tabela 6.1. Número de artigos em conjuntos de dados para os diferentes anos. KW para os *datasets* de palavras-chave e A, para os *datasets* de autores.

Dataset	2003-KW	2004-KW	2005-KW	2003-A	2004-A	2005-A
dsPM	46.588	76.913	117.380	74.856	44.787	113.833
Top-5-KW/1-A	45.584	76.896	117.368	11.505	7.919	18.028

É importante mencionar que essa redução de dados obtidos através da seleção do Top-K elementos, não deteriora a análise desde o início 5000 palavras-chave corresponde a aproximadamente 99% dos artigos.

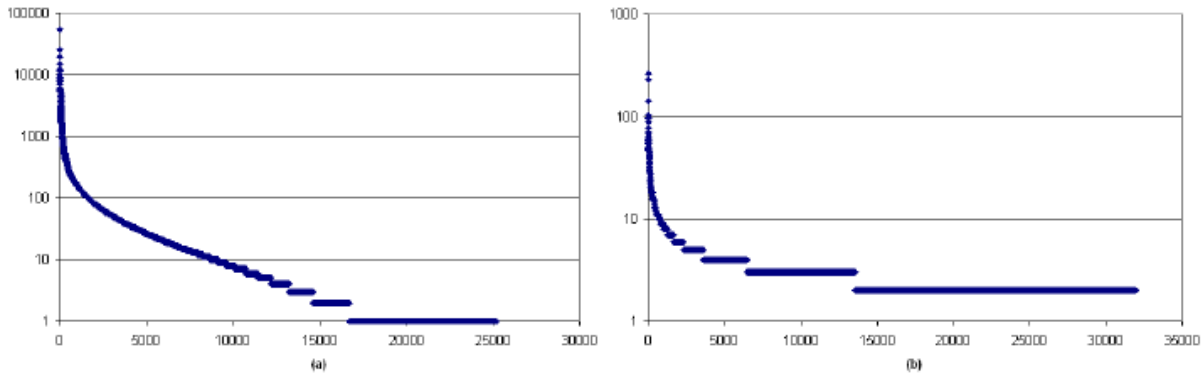


Figura 6.1. Eixo horizontal: Top-K elementos; Eixo vertical: Frequência dos elementos, (a) Distribuição de Palavras-chave para 2003; (b) Distribuição de Autores para 2003; Gráfico plotados em escala logarítmica.

A análise Top-K nos diz que a maioria das ocorrências estão concentradas em poucas palavras-chave e autores, como demonstrado pelas figuras 6.1 (a) e 6.1 (b).Já nas Figuras 6.2 (a) e 6.2 (b) mostram as vinte palavras-chave e autores mais frequentes para o ano de 2005.

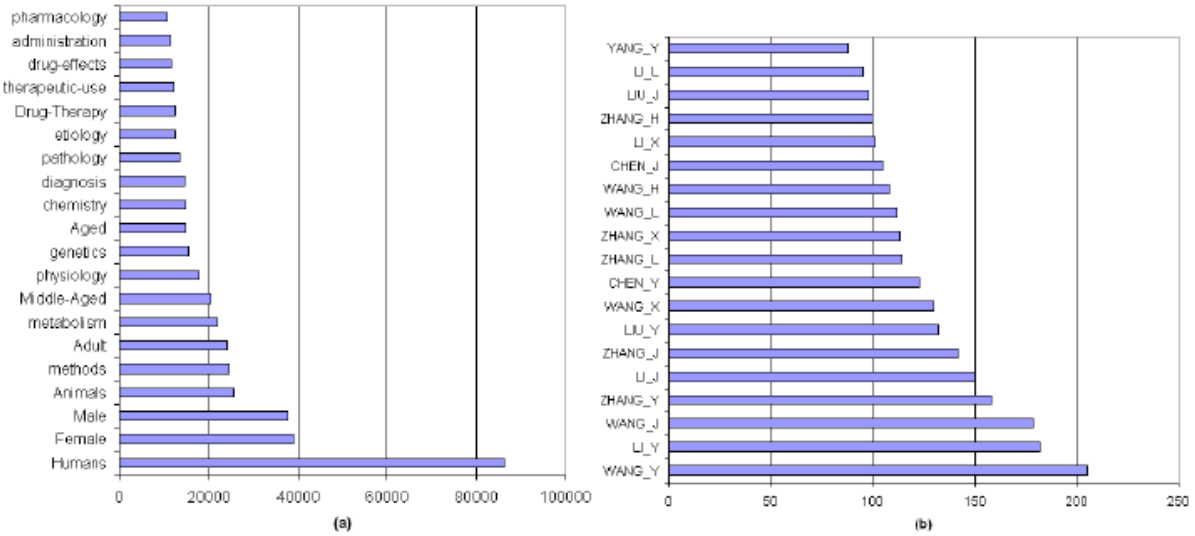


Figura 6.2. (a) Vinte palavras-chave mais frequentes para o ano de 2005, (b) Vinte autores mais frequentes para o ano de 2005;

6.3. MINERAÇÃO DE DADOS DO PUBMED

Uma regra de associação (AR) é uma expressão da forma A → C, onde A e C, respectivamente, denominados antecedente e consequente, são conjuntos de elementos (também chamados de itens) (AGRAWAL; SRIKANT 1995). Para cada AR está associado um conjunto de dois ou mais valores que descrevem suas propriedades com relação ao banco de dados D a partir de onde foi obtido.

O primeiro chama-se *support*, e corresponde ao número de entradas de dados em que esta regra aparece. O segundo valor é chamado de *confidence* e descreve a probabilidade de que uma entrada em *D* que contém *A* conterá também *C*. Se *sup(A* $\cup$ *C)* denota *support* da regra, *confidence* é dado por $conf(A \rightarrow C) = \frac{\text{sup}\ (A \cup C)}{\text{sup}\ (A)}$. Vários algoritmos têm sido propostos para extrair eficientemente RAs sob um *support* mínimo e *confidence* (HIPP; GUNTZER, NAKHAEIZADEH, 2000). Outra importante medida RA de interesse é *lift* , que é dada por $lift(A \rightarrow C) = \frac{conf(A \rightarrow C)}{\text{sup}\ (A \rightarrow C)}$.

Para *lift* valores superiores a 1.0, as entradas de dados que contêm *A* tendem para conter *C* com mais freqüência do que as entradas que não contêm *A.*

As Redes de Dependência (DN) são modelos gráficos que representam distribuições conjuntas para um conjunto de variáveis, de forma semelhante à Markov e modelos Bayesianos (HECKERMAN *et al*, 2000). DN são particularmente úteis para aprender e descrever relações probabilísticas em dados. Para um conjunto de variáveis X_1; ... ; X_n, uma

DN consiste de um par (*G, P*), onde *G* corresponde a um gráfico direcionado e *P* um conjunto de distribuições probabilísticas. Para cada variável X_i corresponde um nó $g_i \in G$ que tem uma distribuição probabilística associada condicionada sobre as demais variáveis (NEVILLE; JENSEN, 2004).

6.3.1. Procurando regras de associação

Para os dados descritos na seção anterior, fizemos várias execuções em diferentes definições sobre o *support* e *confidence* mínimos. A Tabela 6.2 enumera algumas das regras sobre o *dataset* das palavras-chave encontradas para o ano de 2005.

Tabela 6.2. Exemplos de regras de associação encontradas no *dataset* de palavras-chave para o ano de 2005, para um *support* mínimo de 0,003% (272 citações de artigos).

Conf.	Lift	Rule
1.00	3.16	$HeartFailure, Humans \rightarrow Congestive$
1.00	2.60	$Heavy \rightarrow Metals$
1.00	2.53	$Automated, ArtificialIntelligence, Algorithms \rightarrow PatternRecognition$
0.85	2.48	$Local, Pathology \rightarrow NeoplasmRecurrence$
1.00	2.12	$Drugs, Rats, Animals \rightarrow ChineseHerbal$
1.00	1.34	$TumorCells, Pharmacology, Humans \rightarrow Cultured$
1.00	1.11	$SpragueDawley, Methods, Animals, Male \rightarrow Rats$
1.00	1.98	$AminoAcid, MolecularSequenceData, Genetics, Animals \rightarrow SequenceHomology$
1.00	1.97	$ChineseHerbal, Drugs, Plants, Isolation \rightarrow Medicinal$
1.00	1.64	$DiseaseModels, Pathology \rightarrow Animal$
1.00	2.50	$ElectrosprayIonization, Spectrometry \rightarrow Mass$

6.4. DESCOBRINDO REDES DE DEPENDÊNCIA

Analisando dados dos autores relativos ao ano de 2005, foi obtida uma classificação baseada no número de publicações por autor para este ano. Esta classificação é mostrada na Figura 6.2 (b). Esta rede é baseada nas informações de co-autoria de cada artigo. As figuras 6.3 (a) e 6.3 (b) mostram uma sub-rede da DN de autores. Ambos os exemplos foram preparados para que o mesmo conjunto de autores possua o autor alvo no centro da DN. Estas DN apresentam algumas relações de co-autoria interessantes. Por exemplo, pode-se observar que o autor WANG J tem uma grande rede de colaboradores (não totalmente ilustrado na Figura 6.3 (a)). Uma vez que tem um grande número de publicações compartilhado com vários autores, tais relações aparecem na DN como conexões fracas.

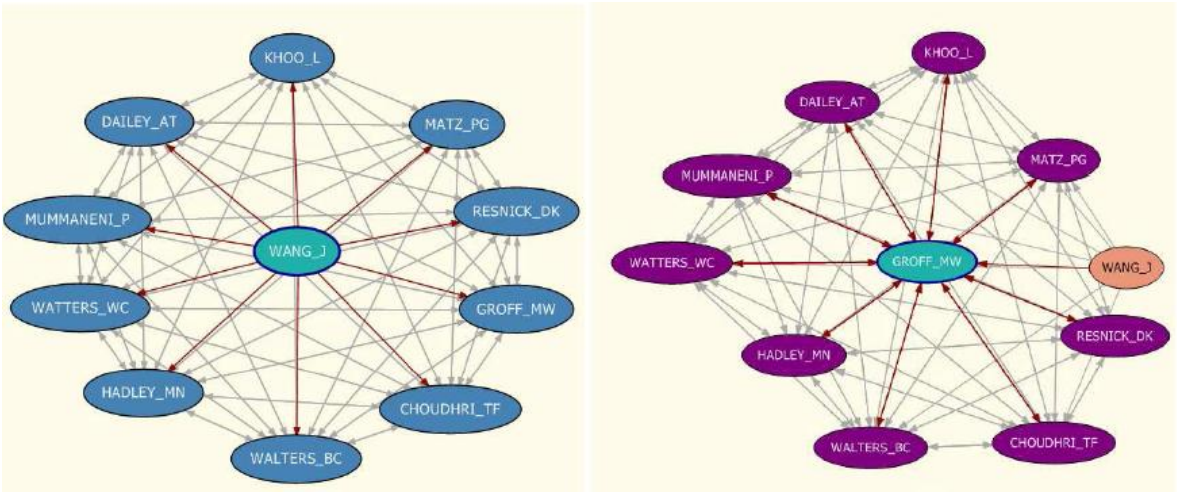


Figura 6.3. Sub-rede da rede de dependência para citações de artigo do PubMed do ano de 2005 ; (a) autor WANG J autor, (b) autor Groff MW.

Outros autores, como Groff MW (ver Figura 6.3 (b)), tem uma forte rede conectada porque ele conserva a maior parte dos seus co-autores. Assim, a sua aparição em uma publicação representa também uma elevada probabilidade de ocorrência de um dos seus co-autores na mesma publicação e vice-versa. Estas e outras observações são diretamente apoiadas pela DN (as instruções sobre as ligações das figuras 6.3 (a) e (b)) e as regras de associação na tabela 6.3.

Tabela 6.3. Exemplos de regras de associação para os autores WANG_J e GROFF_FW.

Id	Conf.	Lift	Rule
1	1.000	2.01	GROFF_MW → WANG_J
2	0.095	3.24	WANG_J → GROFF_MW
3	1.000	4.23	WATTERS_WC, MUMMANENI_P, RESNICK_DK, HADLEY_MN → GROFF_MW
4	1.000	4.23	WATTERS_WC, WANG_J → GROFF_MW
5	1.000	4.23	WATTERS_WC, WALTERS_BC, KHOO_L, WANG_J → GROFF_MW
6	0.944	1.99	RESNICK_DK → WANG_J
7	0.944	1.99	HADLEY_MN → WANG_J
8	1.000	2.01	WATTERS_WC, MATZ_PG, CHOUDHRI_TF, RESNICK_DK → WANG_J
9	1.000	2.01	DAILEY_AT, CHOUDHRI_TF → WANG_J
10	1.000	2.01	WATTERS_WC, WALTERS_BC, MUMMANENI_P, GROFF_MW → WANG_J
11	1.000	2.01	GROFF_MW, CHOUDHRI_TF, RESNICK_DK, HADLEY_MN → WANG_J

Da Tabela 6.3, as regras 1 e 2 descrevem duas situações diferentes. Na regra 1, a presença do autor Groff MW no antecedente resulta em uma alta probabilidade de ocorrência de co-autor WANG J (valor elevado de *confidence*). Por outro lado, a regra 2 tem um valor muito baixo de *confidence*, o que significa que Groff MW só ocorre em uma pequena fração

dos artigos onde ocorre WANG J. O alto valor de *lift* nesta regra resulta que quando WANG J ocorre, há uma alta probabilidade de Groff MW também ocorrer.

As regras 3, 4 e 5 têm um valor alto de *lift*, portanto, alto valor de previsibilidade.

Isto significa que, quando o conjunto de autores no antecedente destas regras ocorrer, existem grandes chances de que Groff MW ocorre também como co-autor. As regras de 6 a 11 têm alto valor de *confidence* e, portanto, autor WANG J ocorre na maioria dos casos em que o antecedente ocorrer. Para valores baixos de *lift* para autores nos antecedentes possuem menos chance de ocorrer junto com WANG_J.

6.5. SUMÁRIO DO CAPÍTULO

Este capítulo descreveu os resultados sobre a mineração de dados obtidos a partir da investigação das citações de artigo do PubMed. Aplicando técnicas de mineração de dados como regras de associação e redes de dependência, relações interessantes puderam ser encontradas. Em particular, temos demonstrado que a co-relações de diferentes magnitudes foram descobertas entre a investigação entre os temas e autores. Os números apresentados neste capítulo, bem como outros complementares valores podem ser encontrados em (IBM, 2007). Através da utilização de técnicas de mineração de dados aplicadas, foi possível encontrar padrões nas publicações.

O capítulo seguinte mostra alguns casos de estudo e resultados utilizando a interface do PubMed Local, realizando um comparativo com os resultados do PubMed.

7. CASOS DE ESTUDOS E RESULTADOS

De forma a validar o trabalho realizado, optou-se por mostrar uma busca realizada no PubMed Local, onde é feita a busca-padrão, através da qual o usuário apenas insere os termos necessários para sua busca e manda consultar. Após isso, é feita uma comparação com as buscas, utilizando diferentes prioridades.

Foram solicitadas aos professores da área das Ciências Biológicas do Centro Universitário Franciscano algumas palavras-chave, sobre as quais deveriam retornar artigos relacionados com as suas áreas de pesquisa. Os termos analisados foram:

- "*helicobacter drugs treatment*"

Para ilustrar o objetivo descrito, serão analisados os 10 primeiros resultados de cada pesquisa apresentada.

A Figura 7.1 é o resultado da busca padrão, onde se obteve como resultado os títulos dos seguintes artigos:

1. A substance-abuse research-treatment clinic: effective procedures and systems.
2. The Beech Hill Hospital Eating Disorders Treatment Program for Drug Dependent Females: program description and case analysis.
3. Enhancing addiction treatment through psychoeducational groups.
4. Role of Helicobacter pylori serology in evaluating treatment success.
5. Six-month maintenance treatment of duodenal ulcer with sucralfate: influence on antral gastritis and Helicobacter pylori antral colonization--a prospective uncontrolled study.
6. Subjective reports of withdrawal among cocaine users: recommendations for DSM-IV.
7. [Hemoptysis as manifestation of a Helicobacter pylori infection]
8. [Mechanisms in hypergastrinemia in autoimmune atrophic gastritis and Helicobacter pylori infection]
9. Comparison of the urease test and of direct smear examination in the control of treatment of Helicobacter pylori-induced infection.
10. *Clinical aspects of infection with Helicobacter pylori.*

Figura 7.1: Pesquisa padrão utilizando os termos enviado por professores da área.

De posse do resultado-padrão, são mostrados alguns resultados, gerados a partir da mudança de prioridades.

Para ilustrar a primeira mudança nas prioridades, o usuário deseja realizar a mesma consulta mostrada na Figura 7.1, sendo que para ele as palavras-chaves são a parte mais importante de sua consulta. Assim sendo, definiram-se palavras-chave com prioridade 10. As outras prioridades não são importantes para ele; então, definiram-se as prioridades restantes com peso 0. O resultado obtido, como mostra a Figura 7.2, foi o seguinte:

1. Field tests for rational drug use in twelve developing countries.
2. Clinical aspects of hepatitis B virus infection.
3. Serious drug interactions.
4. Association between HIV and tuberculosis: technical guide.
5. The effect of dose of mifepristone and gestation on the efficacy of medical abortion with mifepristone and misoprostol.
6. A prospective study of the risk of tuberculosis among HIV-infected patients.
7. Prescribing pattern in acute diarrhoea in three districts in Bangladesh.
8. Costs of diarrhoeal diseases and the savings from a control programme in Cebu, Philippines.
9. Onchocerciasis and other eye problems in developing countries: a challenge for optometrists.
10. Global malaria control strategy.

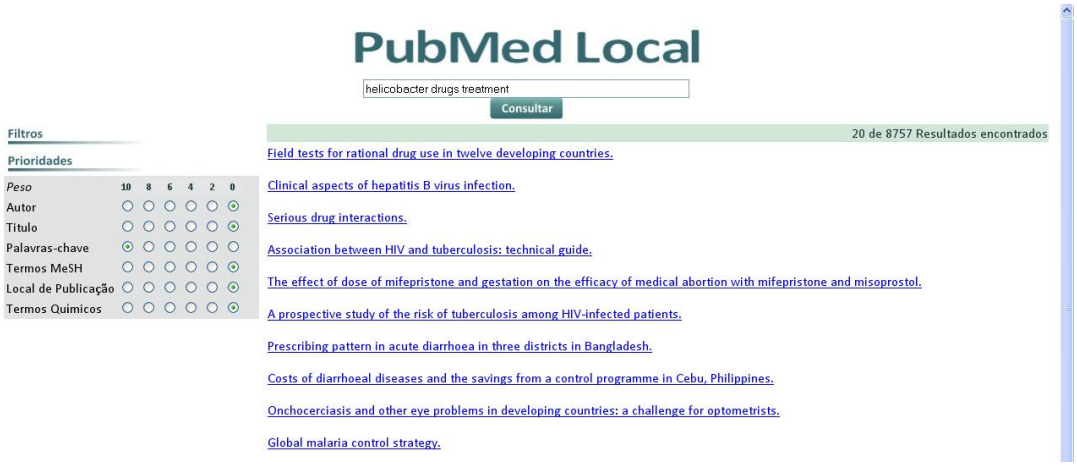


Figura 7.2: Busca avançada onde as palavras-chave são os termos mais importantes.

Ao comparar o resultado obtido entre a Figura 7.1 e a Figura 7.2, é possível perceber a mudança em relação aos 10 primeiros resultados obtidos: nenhum deles retornou artigos encontrados na busca padrão.

No próximo caso apresentado, o usuário deseja ter o título dos artigos como a prioridade mais importante, definindo-os com o peso 10; e que as palavras-chave sejam a segunda prioridade mais importante, definindo-os com peso 8. Para os termos MeSH o usuário definiu peso 0, enquanto as outras prioridades foram deixadas com seus valores iniciais, peso 2.

O resultado obtido, como mostra a Figura 7.3, foi o seguinte:

1. Chronic nicotine treatment potentiates behavioral responses to dopaminergic drugs in rats.
2. Six-month maintenance treatment of duodenal ulcer with sucralfate: influence on antral gastritis and Helicobacter pylori antral colonization--a prospective uncontrolled study.
3. Consumption of buprenorphine and other drugs among heroin addicts under ambulatory treatment: results from cross-sectional studies in 1988 and 1990.
4. [Recent development of endocrine treatment for breast cancer--new drugs and new treatment methods using tamoxifen]
5. Role of Helicobacter pylori serology in evaluating treatment success.
6. [Expectation of new antituberculous drugs and targeting therapy for treatment of mycobacterial infections]

7. Evaluation of new anti-infective drugs for the treatment of vascular access device-associated bacteremia and fungemia. The European Working Party of the European Society of Clinical Microbiology and Infectious Diseases.
8. Comparison of the urease test and of direct smear examination in the control of treatment of Helicobacter pylori-induced infection.
9. The effect of long-term treatment with antidepressant drugs on the hippocampal mineralocorticoid and glucocorticoid receptors in rats.
10. Antiepileptic drugs for the treatment of panic disorder.

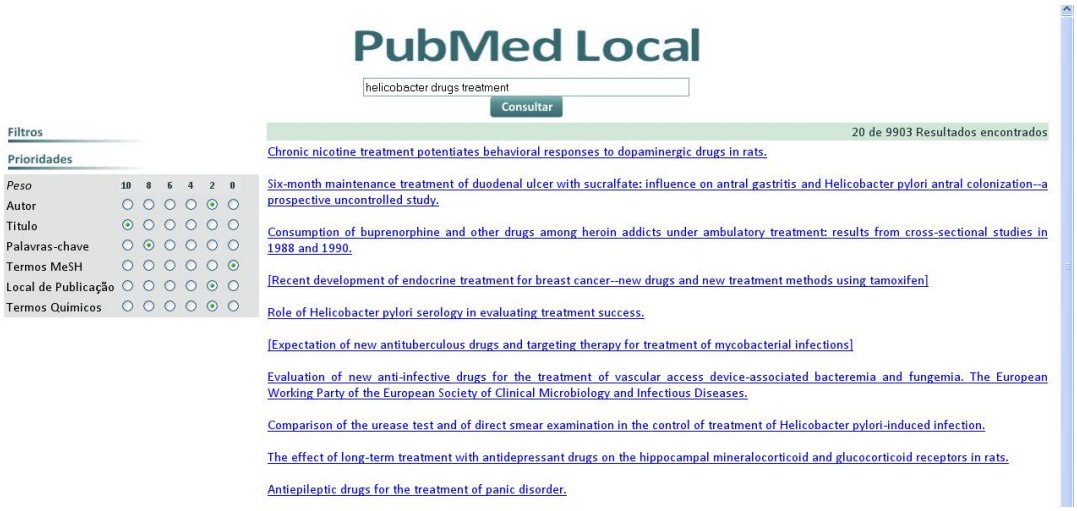


Figura 7.3: busca avançada onde o titulo é o termo mais importante, seguido pelas palavras-chaves, onde os termos mesh são ignorados.

Comparando os resultados obtidos entre a Figura 7.1 e a Figura 7.3, novamente é possível visualizar uma mudança. Nesse caso, os artigos selecionados também são mostrados na busca-padrão, mas todos eles em posições diferentes, que é o caso dos seguintes artigos:

1. Six-month maintenance treatment of duodenal ulcer with sucralfate: influence on antral gastritis and Helicobacter pylori antral colonization--a prospective uncontrolled study. Este retorna na busca padrão como o quinto artigo mostrado, enquanto na busca avançada ele retorna como o segundo resultado.
2. Role of Helicobacter pylori serology in evaluating treatment success. Nesse caso sua posição na busca padrão é a quarta, enquanto na busca avançada é o quinto.
3. Comparison of the urease test and of direct smear examination in the control of treatment of Helicobacter pylori-induced infection. Neste ultimo caso, este artigo é mostrado na nona posição na busca padrão e na oitava posição na busca avançada.

Enquanto os 3 artigos apresentados aparecem em ambas buscas, os outros 7 são diferentes.

Os testes realizados, e apresentados neste capítulo, foram executados em um fragmento do PubMed. O PubMed Local possui mais de 38 GB de dados, enquanto os testes apresentados foram realizados em um banco de dados que dispõe de 158 MB de dados. Não foi possível realizar os testes mostrados no PubMed Local, pois ao tentar executar as consultas, o mesmo apresentou falta de memória, após 5 horas de pesquisas. A configuração da máquina utilizada é Core 2 Duo 6400, com 3 GB de memória RAM. Para utilizar a totalidade dos dados do PubMed Local necessita-se de uma configuração mais robusta, que suporte todo o processamento.

Para obter um nível de satisfação entre profissionais da área, é necessário que o sistema fique disponível a usuários-finais, pois dessa forma pode-se realizar um grande numero de testes, e, então, analisar os resultados obtidos, para que se obtenha um resultado confiável.

7.1. SUMÁRIO DO CAPÍTULO

Neste capítulo foram apresentados alguns casos de estudo e seus resultados, demonstrando a eficiência da interface em relação ao sistema Entrez. No capítulo seguinte, serão abordados alguns trabalhos relacionados realizando um comparativo com cada um deles.

8. TRABALHOS RELACIONADOS

8.1. GOPUBMED

O *Biotechnological Centre* – BIOTEC da *Technische Universität Dresden* desenvolveu uma máquina de busca em ciências da vida, e, em particular, em biologia molecular, GoPubMed (DOMS; SCHROEDE, 2005), baseada no serviço PubMed. Os resultados das buscas são classificados de acordo com GeneOntology (GO) (GENE ONTOLOGY CONSORTIUM, 2004), que é um vocabulário controlado para descrição do gene e dos atributos de seus produtos em qualquer organismo.

O GoPubMed oferece algumas vantagens em relação ao PubMed, que facilitam encontrar resultados mais precisos e de forma ágil. Inicialmente, ele mostra os resultados da pesquisa por categorias, de acordo com o GO, permitindo aos usuários navegar rapidamente através dos *abstracts* por categoria. Em segundo lugar, ele mostra automaticamente termos gerais da ontologia relacionados com a pesquisa original, que muitas vezes nem sequer aparecem diretamente no *abstract*. Também está disponível para os usuários dados estatísticos sobre os resultados da pesquisa, como os termos, os autores, países, jornais que mais aparecem dentre os resultados da pesquisa.

Além da busca-padrão oferecida pelo GoPubMed, também pode-se realizar busca por assunto, autor, lugar da publicação, ou pela data de publicação. O retorno da pesquisa é informado classificado em quatro categorias: "*What*", "*Who*", "*Where*" e "*When*". Essas categorias mostram os conceitos pessoas, instituições, locais, revistas e datas que são relevantes para a pesquisa dos usuários.

Na categoria *What* são listados os conceitos relevantes da biomedicina que foram encontrados durante a pesquisa realizada. Esses conceitos pertencem ao *Gene Ontology* e ao MeSH (*Medical Subject Headings*), sendo que os termos da GO descrevem genes e produtos genes de diferentes organismos, e são divididos em três áreas básicas: a função molecular de produtos genes, o seu papel nos processos biológicos e sua localização nos componentes celulares.

Já os termos MeSH descrevem anatomia, Ciências Biológicas, química e drogas, doenças, cuidados de saúde, Ciências Naturais, organismos, psiquiatria e psicologia, técnicas e equipamentos, nomes de grupos, tecnologia, indústria e agricultura.

A idéia do GoPubMed de utilizar o *Gene Ontology* para pesquisa e visualização dos dados do PubMed tiveram dois problemas para serem resolvidos (DELFS, 2004): primeiro, como extrair os termos GO dos *abstracts* do PubMed; segundo, como construir uma sub-ontologia relevante do GO.

8.1.1. Comparativo com o PubMed Local

A grande diferença do GoPubMed em relação a esse trabalho está na forma de efetuar as consultas, enquanto nesse trabalho se pode utilizar pesquisas definindo prioridades e filtros, o GoPubMed após a consulta ter sido realizada, disponibiliza para o usuário a navegação por categorias. Outra vantagem em relação ao GoPubMed é a opção de visualização gráfica que o PubMed Local disponibiliza ao usuário.

8.2. ALIBABA

O Departamento de Ciência da Computação de *Humboldt-Universität zu* Berlin desenvolveu uma ferramenta chamada AliBaba (PLAKE *et al*., 2006). AliBaba é uma ferramenta interativa que mostra os resultados de pesquisas no PubMed de forma gráfica, utiliza métodos avançados de *text mining* para extrair vários tipos de informações biomédicas a partir do resultado de uma consulta no PubMed e apresenta estas informações em uma forma gráfica.

AliBaba permite aos usuários acesso a literatura subjacente ao navegar neste gráfico de células, doenças, drogas, proteínas, espécies, e de tecidos, bem como as respectivas associações. Diversas opções de filtros podem ser utilizadas para realizar a pesquisa, tornando a pesquisa mais focada.

8.2.1. Comparativo com a Mineração sobre o PubMed Local

Enquanto a ferramenta AliBaba utiliza técnicas de *text mining,* esse trabalho utiliza técnicas de *data mining* como regras de associação e redes de dependência para extrair o conhecimento contigo no PubMed. Também utiliza da mesma forma valor de *confidence* para saber a probabilidade de um/uns termo(s) aparecer(em) junto com outro(s).

8.3. SUMÁRIO DO CAPÍTULO

Neste capítulo foram abordados alguns projetos relacionados ao PubMed, onde foi feito uma breve descrição de cada um, e um comparativo com esse trabalho. No próximo capítulo serão apresentas as conclusões do trabalho.

9. CONCLUSÃO

Com o objetivo principal de representar o conhecimento contido no PubMed, a presente dissertação apresentou uma arquitetura para visualizar os resultados da pesquisa de forma gráfica através de uma rede semântica.

A estruturação e criação do banco de dados local com as informações contidas no PubMed serviu para entender como as publicações são armazenadas, se tornando a base de dados utilizada pela arquitetura que foi desenvolvida neste trabalho.

O PubMed Local, como foi chamada a arquitetura desenvolvida neste trabalho, foi idealizada para suprir uma carência encontrada no PubMed. Conforme foi visto ao longo desta dissertação, o PubMed possibilita pesquisas estáticas, permitindo apenas a utilização de filtros para refinar a consulta. Desta forma, a interface para consultas proposta para o PubMed Local permite que as pesquisas sejam dinâmicas, podendo definir prioridades no momento da pesquisa, tornando o resultado da pesquisa mais eficiente e intuitivo para o pesquisador.

Outra carência do PubMed é a forma de visualização dos dados, não possuindo uma interface gráfica. De forma a suprir esta carência, optou-se pela aplicação de ontologias representadas pela norma ISO 13250 Topic Maps para a geração de uma rede semântica, a partir das consultas realizadas.

Outro objetivo dessa dissertação foi a utilização de técnicas de *data mining* para representar o conhecimento contido no PubMed, já que a grande quantidade de informação disponível para usuário torna difícil achar padrões, como por exemplo, co-autorias mais frequentes e palavras-chave mais utilizadas em conjunto. Por isso, utilizou-se regras de associação e redes de dependência para buscar esse conhecimento que não era conhecido.

Através deste estudo, conseguiu-se representar parte do conhecimento encontrado nesse domínio, utilizando técnicas de mineração de dados conforme foi descrito no capítulo 7, e também através da própria rede semântica que é gerada através da interface do PubMed Local.

Cabe aqui ressaltar que este trabalho não teve por objetivo a comparação em relação ao tempo de execução das consultas nos dois bancos de dados. Isso porque os testes no

âmbito deste trabalho foram realizados em uma única máquina, ficando limitados, enquanto o PubMed possui um recurso computacional superior.

Os trabalho a seguir, foram publicados durante o desenvolvimento dessa dissertação.

LIBRELOTTO, Giovani R. ; MARTINS, M O. ; MACHADO, Henrique Tamiosso; VIZZOTTO, J. K. ; RAMALHO, José Carlos ; HENRIQUES, Pedro. Rangel. . *Generating Semantic Networks to the PubMed*. In: XATA 2008 - XML: Aplicações e Tecnologias Associadas, 2008, Évora. XATA 2008: 6ª Conferência Nacional, 2008, 2008. v. 6. p. 94-103.

LIBRELOTTO, Giovani R. ; MACHADO, Henrique Tamiosso ; BERNARDI, G. ; TURCHETTI, R. . *Applying association rules and dependency networks in the PubMed database*. In: Simpósio Brasileiro de Banco de Dados (SBBD) e Simpósio Brasileiro de Engenharia de Software (SBES), 2008, Campinas-SP. SBBD e SBES 2008, 2008.

LIBRELOTTO, Giovani R. ; FERREIRA, P. G. D. ; MARTINS, M O. ; MACHADO, Henrique Tamiosso ; RAMALHO, José Carlos ; HENRIQUES, Pedro. Rangel. . *Topic Maps applied to PubMed*.. In: *Extreme Markup Languages* 2007, 2007, Montreal. *Extreme Proceedings* 2007. Montreal, 2007.

LIBRELOTTO, Giovani R. ; MACHADO, Henrique Tamiosso ; MARTINS, M O. ; FERREIRA, P. G. D. ; RAMALHO, José Carlos ; HENRIQUES, Pedro. Rangel. . *Comparing Topic Maps Constraint Specification Languages*. In: *International Conference on Topic Maps Research and Applications*, 2007, Leipzig. *Proceedings of International Conference on Topic Maps Research and Applications*, 2007.

R. P. de Azevedo ; MACHADO, Henrique Tamiosso ; MARTINS, M O. ; LIBRELOTTO, Giovani R. . Criando Redes Semânticas a partir de consultas no PubMed.. In: IV Workshop de Trabalhos de Iniciação Científica, 2007, Gramado, RS. WebMedia - *XIII Brazilian Symposium on Multimedia and the Web*, 2007.

LIBRELOTTO, Giovani R. ; MARTINS, M O. ; MACHADO, Henrique Tamiosso ; SARAIVA, J. ; TURCHETTI, R. ; SILVA, F. L. ; AUGUSTIN, I. . Aplicando Topic Maps a Ambientes Sensíveis ao Contexto. In: *I Workshop on Pervasive and Ubiquitous Computing*, 2007, Gramado, RS. Anais do I *Workshop on Pervasive and Ubiquitous Computing*, 2007., 2007.

LIBRELOTTO, Giovani R. ; MACHADO, Henrique Tamiosso ; CABRAL, H. C. B . *Finding association rules and dependency networks in article's metadata*. In: Escola Brasileira de Bioinformática (EBB), 2008, Santo André, RS. EBB 2008, 2008.

MACHADO, Henrique Tamiosso ; LIBRELOTTO, Giovani R. ; R. P. de Azevedo ; VIZZOTTO, J. K. . Representaçao do Conhecimento do Domínio do PubMed. In: III Workshop em Nanociencias, 2008, Santa Maria - RS. Livro de Resumos do III Workshop em Nanociencias, 2008.

MACHADO, Henrique Tamiosso ; LIBRELOTTO, Giovani R. ; VIZZOTTO, J. K. . Uma arquitetura para a navegação sobre o conhecimento do PubMed. In: XII Simpósio de Ensino, Pesquisa e Extensão, 2008, Santa Maria. Anais SEPE 2008, 2008.

MACHADO, Henrique Tamiosso ; R. P. de Azevedo ; VIZZOTTO, J. K. ; LIBRELOTTO, Giovani R. . Representação do conhecimeno na bioinformática. In: II Workshop em Nanociências, 2007, Santa Maria. Livro de Resumos do II Workshop em Nanociências, 2007.

MACHADO, Henrique Tamiosso ; DAMROSIO, G. ; MARTINS, M O. ;

VIZZOTTO, J. K. ; LIBRELOTTO, Giovani R. . Definição de um modelo relacional para o PubMed.. In: 22ª Jornada Acadêmica Integrada da UFSM, 2007, Santa Maria, RS. Anais da 22ª Jornada Acadêmica Integrada, 2007.

DAMROSIO, G. ; MACHADO, Henrique Tamiosso ; MARTINS, M O. ; LIBRELOTTO, Giovani R. . Ambiente para Processamento Ontológico do PubMed.. In: 1ª Mostra de Projetos de Pesquisa e Extensão dos Cursos de Ciência da Computação e Sistemas de Informação da Unifra, 2007, Santa Maria, RS. Anais do VI SIRC/RS, 2007.

MACHADO, Henrique Tamiosso ; LIBRELOTTO, Giovani R. ; VIZZOTTO, J. K. . Uma arquitetura para a navegação sobre o conhecimento do PubMed. 2008. (Apresentação de Trabalho/Simpósio).

MACHADO, Henrique Tamiosso ; MARTINS, M O. ; DAMROSIO, G. ; VIZZOTTO, J. K. ; LIBRELOTTO, Giovani R. . Definição de um modelo relacional para o PubMed. 2007. (Apresentação de Trabalho/Outra).

MACHADO, Henrique Tamiosso ; MARTINS, M O. ; DAMROSIO, G. ; LIBRELOTTO, Giovani R. . Banco de Dados Relacional para o PubMed. 2007. (Apresentação de Trabalho/Simpósio).

MARTINS, M O. ; MACHADO, Henrique Tamiosso ; LIBRELOTTO, Giovani R. ; VIZZOTTO, J. K. . Resolvendo Inconsistência de Dados em Banco de Dados ER gerados a partir de arquivos XML. 2007. (Apresentação de Trabalho/Simpósio).

9.1. TRABALHOS FUTUROS

Para trabalhos futuros, projeta-se a implementação do banco de dados do PubMed Local em um *grid* computacional, ou um *cluster*. Desta forma, espera-se melhorar o desempenho das consultas. Também, pretende-se realizar a implementação da consulta com termos MeSH, onde quando o usuário insere um termo para efetuar sua busca, o sistema deve primeiro consultar cada um desses termos na tabela *meshheadings* e, então, utilizar os sinônimos encontrados como parâmetros de busca adicional. Isso irá gerar mais resultados, mas também é importante porque o usuário não sabe exatamente o que está escrito no artigo; logo, buscando por sinônimos, terá como retorno artigos tratando do mesmo assunto, mas para isso os artigos devem estar indexados pelos termos MeSH, o que hoje muitos dos artigos contidos no PubMed não estão.

Além disso, estudos referentes à otimização das consultas e utilização de outros SGBDs são uma forma de tornar as pesquisas viáveis para os usuários.

BIBLIOGRAFIA

AGRAWAL R. and SRIKANT, R. **Fast algorithms for mining association rules**. Procedings of 20th International Conference Very Large Data Bases. - 1995. - pp. 487–499.

AJAX. **AJAX Tutorial**. 2008. – Disponível em: http://www.w3schools.com/ajax/default.asp. Acesso em: Novembro de 2008.

ALTOVA. **Altova XMLSpy.** 2007. Disponível em: http://www.altova.com/products/xmlspy/xml_editor.html. Acesso em: Dezembro de 2007.

BIEZUNSKY M., BRYAN M., NEWCOMB S**. ISO/IEC 13250 - Topic Maps.** ISO/IEC JTC 1/SC34. 1999. Disponível em: http://www.y12.doe.gov/sgml/sc34/document/0129.pdf. Acesso em: December 2007.

CANESE K. **PubMed Celebrates its 10th Anniversary!** NLM Tech Bull. - Sep-Oct 2006. - p. 352:e5.

COMMUNITY S. S. D. M**. Sqlserverdatamining**. 2007. Disponível em: http://www.sqlserverdatamining.com/DMCommunity/. Acesso em: Dezembro de 2007.

CORLAN, A. Medline (pubmed) trend. 2007. Disponível em: http://dan.corlan.net/medline-trend.html. Acesso em: Outubro de 2008.

DELFS Ralfh. **GoPubMed**: ontology-based literature search applied to Gene Ontology and PubMed. German Conference on Bioinformatics. - 2004. - Vols. 169-178.

DEVX. **DevX**. 2007. Disponível em: http://www.devx.com/xml/Article/16922/1954. Acesso em: Novembro 2007.

DINAMIC. **Dinamic Drive**. 2008. Disponível em: http://www.dynamicdrive.com/dynamicindex17/ajaxcontent.htm. Acesso em: Novembro 2008.

DOMS Andreas; SCHROEDE Michel. **GoPubMed**: Exploring PubMed with the GeneOntology. Nucleic Acid Research. - 2005. - W783-W786. - Vol. 33.

DUBOIS Paul**. MySQL**. Paperback. 2008.

ENTREZ MEDICINE U. N. L.. **Fact Sheet.** Disponível em: http://www.ncbi.nlm.nih.gov/Database/index.html. Acesso em: Julho 2008.

FAYYAD U. M., et al. **Advances in Knowledge Discovery and Data Mining**. AAAI Press / The MIT Press. - California : [s.n.], 1996.

FELLBAUM Christiana**. WORDNET**: an electronic lexical database and some of its applications. Cambridge, MA : MIT Press, 1999.

FIREBUG**.** **Firebug**. 2007. Disponível em: https://addons.mozilla.org/en-US/firefox/addon/1843. Acesso em: Dezembro de 2007.

GENE ONTOLOGY CONSORTIUM. **The Gene Ontology (GO) database and informatics resource.** Nucleic Acids Research. - 2004. - Vols. D258-D261.

GIBAS Cynthia; JAMBECK Per. **Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia**. Rio de Janeiro : Campus, 2001.

GIMENES, Eduardo. Data Mining – Data Warehouse, A Importância da Mineração de Dados em Tomadas de Decisões. Disponível em: <http://br.geocities.com/dugimenes/>. Acesso em: 17 de setembro de 2008.

HECKERMAN D., et al**. Dependency networks for inference, collaborative filtering and data visualization.** Journal of Machine Learning Research. - 2000. - 1. - pp. 49–75.

HIPP J., GUNTZER, U., NAKHAEIZADEH, G. **Algorithms for association rule mining a general survey and comparison.** SIGKDD Explorations. - 2000. - Vol. 2. - pp. 58–64.

HODGES Wilfrid. A Shorter Model Theory. [s.l.] : Cambridge University Press, 1997.

IBM. **Many eyes for shared visualization and discovery.** 2007. Disponível em: http://services.alphaworks.ibm.com/manyeyes/browse/topichubs. Acesso em: Fevereiro de 2008.

JAVASCRIPT. **JavaScript Tutorial.** 2008. Disponível em: http://www.w3schools.com/js/default.asp. Acesso em: Novembro 2008.

LIBRELOTTO G. R, RAMALHO J. C., HENRIQUES P. R.. **Metamorphosis** - A Topic Maps Based Environment to Handle Heterogeneous Information Resources. Lecture Notes in Computer Science. - [s.l.] : Springer-Verlag GmbH, 2006. - Vol. 3873. - pp. 14–25.

LIBRELOTTO G R., et al. **Navegando na Rede Semântica dos Topic Maps com o Ulisses.** XATA2007 – XML: Aplicações e Tecnologias Associadas. - Lisboa : [s.n.], 2007.

LIBRELOTTO Giovani Rubert. **Representação de Conhecimento na Semanti Web**. Portugal : Departamento de Informática. Universidade do Minho, 2005.

MARK A. V. M., ALISTAIR M., GUUS S. **A Method to Convert Thesauri to SKOS**. 2007. Disponível em: http://thesauri.cs.vu.nl/eswc06. Acesso em: Novembro de 2008.

MEDICINE U. N. L. **PUBMED.** 2007. Disponível em: http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed. Acesso em: Abril de 2007.

_____. **MEDLINE** – Fact Sheet. 2006. Disponível em: http://www.nlm.nih.gov/pubs/factsheets/medline.html. Acesso em: Abril de 2007.

MENEZES. P. B.. **Linguagens Formais e Autômatos**. Sagra. Porto Alegre : Sagra Luzzatto, 2000.

MYSQL**.** **MySQL 5** – Reference Manual. 2008. Disponível em: http://dev.mysql.com/doc/refman/5.0/en/index.html. Acesso em: Novembro 2008.

NCBI**.** **Information National Center for Biotechnology.** 2007. Disponível em: http://www.ncbi.nlm.nih.gov. Acesso em: 10 de Junho de 2007.

NEVILLE J., JENSEN, D**.** **Dependency networks for relational data.** Proceedings of The 4th IEEE International Conference on Data Mining.. 2004.

NOVIXYS. **Exult**.2007. Disponível em: http://www.novixys.com/Exult/index.html. Acesso em: Agosto de 2007.

PEPPER Steve, GRAHAM Moore. **XML Topic Maps (XTM) 1.0** - Annex D: XTM 1.0 Document Type Declaration (Normative). 2001. TopicMaps.Org Specification. Disponível em: http://www.topicmaps.org/xtm/1.0/#dtd. Acesso em: August 2007.

PHP**.** **PHP Manual**. 2008. Disponível em: http://br.php.net/manual/pt_BR/preface.php. Acesso em: Novembro 2008.

PLAKE Conrad et al.. **ALIBABA:** PubMed as a graph. BIOINFORMATICS APPLICATIONS NOTE. - Berlin : [s.n.], 2006. - 19. - Vol. 22. - pp. 2444-2445.

PROTÉGÉ. **The protégé ontology editor and knowledge acquisition system**. 2005. Disponível em: http://protege.stanford.edu. Acesso em: Agosto de 2007.

REZENDE, Solange. **Mineração de Dados,** EM: ENIA - Encontro Nacional de Inteligência Artificial. Porto Alegre – RS, 25-29 Julho, 2005.

SAX PROJECT. **Sax Project**. Sax Project, 2008. Disponível em: http://www.saxproject.org/. Acesso em: Novembro 2008.

TIDIA - Ae system. **First module of the "On-line Course on PubMed Search"**. TIDIA - Aprendizagem Eletrônica. 2006. Disponível em: http://tidia-ae.incubadora.fapesp.br. Acesso em: 12 de Julho 2007.

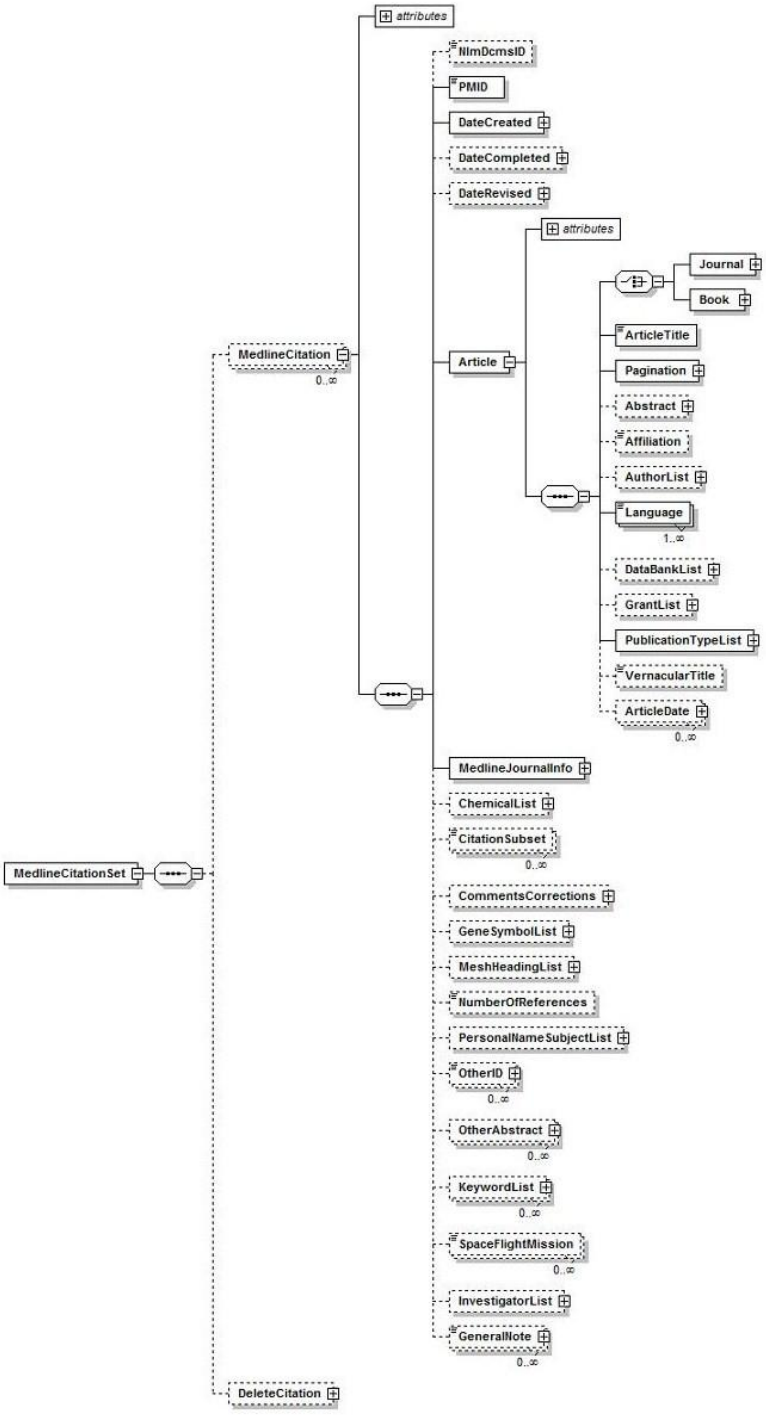
DALEN Dirk Van**.** Logic and Structure. [s.l.] : Springer, 1994. - Vols. ISBN: 3-540-57839-0.

WELLING Luke, THOMSON Laura. **PHP and MySQL Web Development**. [s.l.] : Paperback, 2008.

W3SCHOOLS. **Introduction to HTML.** 2008. Disponível em: http://www.w3schools.com/html/html_intro.asp. Acesso em: Novembro 2008.

ANEXO A

Schema XML do PubMed gera no *software Altova XMLSpy* (ALTOVA , 2007)



ANEXO B

Exemplo de um documento em XML disponibilizado pelo PubMed.

```
<PubmedArticle>
    <MedlineCitation Status="Publisher" Owner="NLM">
        <PMID>19197936</PMID>
        <DateCreated>
            <Year>2009</Year>
            <Month>2</Month>
            <Day>6</Day>
        </DateCreated>
        <Article PubModel="Print-Electronic">
            <Journal>
                <ISSN IssnType="Electronic">1521-4141</ISSN>
                <JournalIssue CitedMedium="Internet">
                    <PubDate>
                        <Year>2009</Year>
                        <Month>Feb</Month>
                        <Day>5</Day>
                    </PubDate>
                </JournalIssue>
                <Title>European journal of immunology</Title>
                <ISOAbbreviation>Eur. J. Immunol.</ISOAbbreviation>
            </Journal>
            <ArticleTitle>Signaling events leading to the curative effect of cystatin on experimental visceral
            leishmaniasis: Involvement of ERK1/2, NF-kappaB and JAK/STAT pathways.</ArticleTitle>
            <Pagination>
                <MedlinePgn/>
            </Pagination>
            <Abstract>
                <AbstractText>Curative effect of cystatin, a natural cystein protease inhibitor, on
                experimental visceral leishmaniasis was associated with strong upregulation of iNOS. The
                transductional mechanisms underlying this cellular response was investigated in the murine
                macrophage cell line RAW 264.7 and in the BALB/c mouse model of visceral leishmaniasis.
                Cystatin synergizes with IFN-gamma in inducing ERK1/2 phosphorylation and NF-kappaB
                DNA-binding activity. Pretreatment of cells with specific inhibitors of NF-kappaB or
                ERK1/2 pathway blocked the cystatin plus IFN-gamma-inducible NF-kappaB activity and
                markedly reduced the expression of iNOS at both mRNA and protein levels. Silencing of
                mitogen- and stress-activated protein kinase 1 significantly reduced cystatin-mediated NF-
                kappaB-dependent iNOS gene transcription suggesting the involvement of mitogen- and
                stress-activated protein kinase 1 activation in ERK1/2 signaling. DNA binding as well as
                silencing experiments revealed the requirement of IFN-gamma-mediated JAK-STAT
                activation even though cystatin did not modulate this signaling cascade by itself. In the in
                vivo situation, key steps in the activation cascade of NF-kappaB, including nuclear
                translocation of NF-kappaB subunits, IkappaB phosphorylation and IkappaB kinase, are all
                remarkably enhanced in Leishmania-infected mice by cystatin. Understanding the molecular
                mechanisms through which cystatin modulates macrophage effector responses will
                contribute to better define its potential for macrophage-associated diseases, in
                general.</AbstractText>
            </Abstract>
            <Affiliation>Molecular Cell Biology Laboratory, Infectious Diseases and Immunology Division,
            Indian Institute of Chemical Biology, Kolkata, India.</Affiliation>
            <AuthorList>
                <Author>
                    <LastName>Kar</LastName>
                    <FirstName>Susanta</FirstName>
```

```
                    <Initials>S</Initials>
                </Author>
                <Author>
                    <LastName>Ukil</LastName>
                    <FirstName>Anindita</FirstName>
                    <Initials>A</Initials>
                </Author>
                <Author>
                    <LastName>Das</LastName>
                    <FirstName>Pijush K</FirstName>
                    <Initials>PK</Initials>
                </Author>
            </AuthorList>
            <Language>ENG</Language>
            <PublicationTypeList>
                <PublicationType>JOURNAL ARTICLE</PublicationType>
            </PublicationTypeList>
            <ArticleDate DateType="Electronic">
                <Year>2009</Year>
                <Month>2</Month>
                <Day>5</Day>
            </ArticleDate>
        </Article>
        <MedlineJournalInfo>
            <MedlineTA>Eur J Immunol</MedlineTA>
            <NlmUniqueID>1273201</NlmUniqueID>
        </MedlineJournalInfo>
    </MedlineCitation>
    <PubmedData>
        <History>
            <PubMedPubDate PubStatus="entrez">
                <Year>2009</Year>
                <Month>2</Month>
                <Day>7</Day>
                <Hour>9</Hour>
                <Minute>0</Minute>
            </PubMedPubDate>
            <PubMedPubDate PubStatus="pubmed">
                <Year>2009</Year>
                <Month>2</Month>
                <Day>7</Day>
                <Hour>9</Hour>
                <Minute>0</Minute>
            </PubMedPubDate>
            <PubMedPubDate PubStatus="medline">
                <Year>2009</Year>
                <Month>2</Month>
                <Day>7</Day>
                <Hour>9</Hour>
                <Minute>0</Minute>
            </PubMedPubDate>
        </History>
        <PublicationStatus>aheadofprint</PublicationStatus>
        <ArticleIdList>
            <ArticleId IdType="doi">10.1002/eji.200838465</ArticleId>
            <ArticleId IdType="pubmed">19197936</ArticleId>
        </ArticleIdList>
    </PubmedData>
</PubmedArticle>
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo