Data Integration Approaches for Supporting Retrieval of Medical Information in the Web

Stamatis Andrianakis

Master Thesis

Technical University of Crete
School of Electronic and Computer Engineering

Intelligent Systems Laboratory

January 2015

Περίληψη

Το διαδίκτυο αποτελεί πλέον τον βασικότερο τρόπο ανάκτησης επιστημονικής πληροφορίας. Ειδικότερα για δεδομένα που αφορούν τον ιατρικό τομέα έχουν αναπτυχθεί *Ιατρικά Πληροφοριακά Συστήματα* τα οποία αποθηκεύουν και οργανώνουν την ιατρική πληροφορία και τελικά την καθιστούν διαθέσιμη στους χρήστες είτε μέσω εξειδικευμένων μηχανών αναζήτησης που παρέχονται από αυτά, είτε μέσω μηχανών αναζήτησης γενικού σκοπού όπως η Google.

Ο μεγάλος όγκος των ιατρικών δεδομένων που διατίθενται, καθώς και το πλήθος και η ανομοιογένεια των Ιατρικών Πληροφοριακών Συστημάτων ως προς τον τρόπο ταξινόμησης και αναζήτησης που προσφέρουν, καθιστούν την αναζήτηση αξιόπιστης ιατρικής πληροφορίας χρονοβόρα και κάποιες φορές δύσκολη διαδικασία. Στην παρούσα εργασία περιγράφεται μία μέθοδος ενοποιημένης αναζήτησης και ανάκτησης ιατρικής πληροφορίας από το διαδίκτυο και παρουσιάζεται το MIIDLE, ένα σύστημα που παρέχει τη δυνατότητα ενοποιημένης αναζήτησης ιατρικής πληροφορίας στο διαδίκτυο από ετερογενείς πηγές πληροφορίας.

Το MIIDLE χρησιμοποιεί το ελεγχόμενο λεξιλόγιο *MeSH* του National Library of Medicine ως κοινή βάση. Με τη χρήση της μεθόδου AMTE$_X$ πραγματοποιεί εξαγωγή ιατρικών όρων από τα ανακτώμενα ιατρικά δεδομένα από τις διαφορετικές πηγές ενώ παράλληλα διευρύνει το ερώτημα από το οποίο προέκυψαν τα δεδομένα με όρους που ανήκουν στο *MeSH*. Συνδυάζοντας τους ιατρικούς όρους που εξήχθησαν και το διευρυμένο ερώτημα και με χρήση της μεθόδου Vector Space Model επιστρέφει στον χρήστη τα αποτελέσματα ταξινομημένα ως προς τη σχετικότητά

τους με το αρχικό ερώτημα.

Τα αποτελέσματα του MIIDLE αξιολογούνται από χρήστες και η αποδοτικότητά του συγκρίνεται με την αποδοτικότητα των πηγών από τις οποίες άντλησε τα αποτελέσματα.

Abstract

In recent years, the World Wide Web has become the basic source of scientific information. Especially for medical information, several *Medical Information Systems* have been developed in order to store and organize medical data and make it available to users through specialized search engines provided by them, or through general purpose search engines like Google.

Vast amount of medical data available in the Web and the large number of Medical Information Systems make the search process of reliable medical information a time-consuming and sometimes difficult process. This work presents an integration method for search and retrieval of medical data and MIIDLE, an integration system for search and retrieval of medical information from heterogeneous sources.

MIIDLE utilizes *MeSH*, the National Library of Medicine's controlled vocabulary thesaurus as a common vocabulary for the integration process. Using AMTE$_X$ method, it extracts medical terms from the retrieved data, and it expands the query used for the retrieval with MeSH terms. Combining the extracted terms with the expanded query it ranks the results with respect to their relevance using Vector Space Model.

MIIDLE results are evaluated by users and its performance is compared with the performance of the sources that it accesses.

Acknowledgements

I would like to thank Prof. Euripidis G.M. Petrakis for his guidance and support. I would also like to thank Angelos Hliaoutakis for the technical support and recommendations he provided to me and Dr K. Tavernaraki, Dr A. Moustris, Dr G. Papaefstathiou, Dr M. Grigoraki, Dr A. Kovalevskagia, I. Papaefstathiou, I. Andrianaki and N. Leontaris for their help in the evaluation of the results of this work. Finally my heartful thanks go to my family for their endless love and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Science requires knowledge and knowledge requires information. The easiest way to retrieve scientific information nowadays is to issue queries in the *World Wide Web (WWW)*. The dramatic growth of the WWW with more than one billion Web sites[1], a large part of which concerns scientific data, makes it a good source for this purpose. The problem is that information in the web are scattered among multiple data sources with different semantics each one and no explicit relation between their contents (i.e., different Web sites may share similar content on the same subject using similar or different terminology without linking one another or sharing common page links to other reference content resources).

In medicine, large content repositories (e.g., Medline[2]) are in every day use by experts and naive users. Each one of these systems are organized using a different index vocabulary (some without any index like content published by small organization or individuals). Their main operation is to issue queries on a subject of interest to a Web search engine (e.g., Google) or the search engine provided by the content provider (e.g., PubMed[3], HON[4]). However, there is still a need for tools and mechanisms for unifying and filtering all query results from

[1] http://www.internetlivestats.com/
[2] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[3] http://www.ncbi.nlm.nih.gov/pubmed
[4] http://www.hon.ch

all sources by virtue of intent meaning and presenting these to the user in a single form. Today, in order to retrieve medical information[5] health professionals and naive users have to search each digital library individually and filter the results manually, which is very time consuming process.

In this work we develop *MIIDLE - Information Integration in Medical Digital Libraries*, an integration system for medical sources. The ranking process is based on vector-space model using MeSH[6], the medical vocabulary of the US NLM (National Library of Medicine). In MIIDLE, MeSH medical terms are extracted from documents by applying AMTE$_X$ [19] method (Section 2.3.1).

MIIDLE has been tested on medical information available in three sources, two of them are medical data sources: PubMed[7] and MedlinPlus[8] and the third is the Web and results retrieved by issuing the same queries using Google search engine. Query results are evaluated by users and average evaluation scores over 12 queries are presented and discussed.

The characteristics of the data sources as well as related work of integration systems and methods are presented in chapter 2. The architecture of MIIDLE is presented in details in chapter 3. The evaluation of the results and a comparison between the results sets of the three data sources that MIIDLE accesses and results set that MIIDLE returns are presented in chapter 4. Finally conclusions are discussed in chapter 5.

[5]http://www.pewinternet.org/fact-sheets/health-fact-sheet/
[6]http://www.nlm.nih.gov/mesh/
[7]http://www.ncbi.nlm.nih.gov/pubmed
[8]http://www.nlm.nih.gov/medlineplus/

Chapter 2

Background and Related Work

In this chapter is presented an overview of medical information systems and the sources of medical data over the Web. A description of data integration follows along with some technics for medical term extraction that have been used in this work.

2.1 Medical Information Systems

The amount of medical data available on the Internet is huge, and it is growing every day. In order to be easily accessible this information is organized in *Medical Information Systems* such as digital libraries, portals etc. Medical information systems deal with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

National Library of Medicine[1] (NLM) is the world's largest biomedical library and the developer of electronic information services that deliver vast amount of medical data (printed and electronic) to millions of users every day. It has been founded in 1836 and it is located on the campus of the National Institutes of Health in Bethesda, Maryland. It is a department of the *National Institutes of Health* (NIH), which in turn is a part of the *United States Department of*

[1] http://www.nlm.nih.gov

Health and Human Services. NLM provides to expert users the *MEDLINE* which includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. It also covers much of the literature in biology and biochemistry. An important division of NLM is *National Center for Biotechnology Information*[2] (NCBI). It houses a series of databases relevant to biotechnology and biomedicine and provides access to biomedical and genomic information. NCBI has developed **PubMed**[3] (see section 2.2.1) which is a free search engine for expert users, accessing primarily the MEDLINE database and other selected life sciences journals and online books. It comprises more than 23 millions citations and abstracts for biomedical literature, and links to full-text content if it is available. While PubMed is mostly for experts, **MedLinePlus**[4] (see section 2.2.2), is the National Institutes of Health's web site intended to be used mostly by consumers.It is produced by NLM and brings information about diseases, conditions, and wellness issues in language that can be understood by non-expert users.

Health On the Net Foundation - HON[5] founded in 1996 under the auspices of the Geneva Ministry of Health and based in Geneva, Switzerland. HON accredited in 2002 by the Economic and Social Council of the United Nations, with operational support provided by the Geneva Health Ministry and the National French Health Authority with additional project funding from the European Union. It is a non-profit, non-governmental organization which promote and guide the deployment of useful and reliable online medical and health information, and its appropriate and efficient use. HON proposes solutions to the two main obstacles: the accessibility of the information and the trustworthiness of medical and health information on the Internet. To do this, the Foundation has issued the *HONcode*© certification. HONcode is an ethical standard aimed

[2]http://www.ncbi.nlm.nih.gov
[3]http://www.ncbi.nlm.nih.gov/pubmed
[4]http://www.nlm.nih.gov/medlineplus
[5]http://www.hon.ch

at offering quality health information. It demonstrates the intent of a website to publish transparent information which will improve the usefulness and objectivity of the information and the publishment of correct data and guides site managers in setting up a minimum set of mechanisms to provide quality, objective and transparent medical information tailored to the needs of the audience. Health On the Net foundation also provides a large variety of tools in order to facilitate the search process of both experts and consumer users: *HONcodeHunt*© searches for reliable information in HON databases for consumers[6] and for expert[7] users, *MedHunt*© that aims to provide access to reliable medical pages that crawled from the web for consumers[8] and experts[9] and *HONselect*©, a user-friendly directory of selected medical resources forming an encyclopedia of 33,000 medical terms in 7 languages, again for consumers[10] and expert[11] users.

WRAPIN[12] (Worldwide online Reliable Advice to Patients and Individuals) is an "ad-hoc" search engine of health/medical documents. The documents indexed by WRAPIN come from a human selection of the best trustworthy medical databases (PubMed, MedHunt, HONcodeHunt, OESO (Medical scientific articles about Oesophagus diseases), URO France (Medical scientific articles about urology), ClinicalTrials (Clinical trials from the U.S. National Library of Medicine), FDA (U.S Food and Drug Administration), etc), and it can translate the query into five languages (English, French, German, Spanish and Portuguese). WRAPIN enables the comparison of the documents in several formats (Text, HTML, PDF, etc.) and any length, to discover if the information exists in the published literature and provide a summary conclusion of the ideas contained in order to determine the reliability of documents by checking these ideas against

[6]http://www.hon.ch/HONsearch/Patients/hunt.html
[7]http://www.hon.ch/HONsearch/Pro/hunt.html
[8]http://www.hon.ch/HONsearch/Patients/medhunt.html
[9]http://www.hon.ch/HONsearch/Pro/medhunt.html
[10]http://www.hon.ch/HONsearch/Patients/medhunt.html
[11]http://www.hon.ch/HONsearch/Pro/honselect.html
[12]http://www.wrapin.org/

established benchmarks, and enable users to determine the relevance of a given document from a page of search results.

WebMD is a corporation which provides health information services for both experts and consumers. The WebMD Health Network operates WebMD portal[13] and other health-related sites including: **Medscape**[14], **MedicineNet**[15], **eMedicine**[16], **eMedicineHealth**[17], **RxList**[18], **theheart**[19], **Medscape Education**[20], etc. These sites provide similar services to those of WebMD. MedicineNet is an online media publishing company, while Medscape offers up-to-date information for physicians and other healthcare professionals. RxList offers detailed information about pharmaceutical information on generic and name-brand drugs and eMedicineHealth is a consumer site offering similar information to that of WebMD.

Healthline Networks[21] is a privately owned provider of health information and technology solutions. Healthline provides some useful tools like SymptomChecker[22], a comprehensive tool with more than 1000 diseases and conditions and 4500 symptom choices which, through guided search, results filters and related symptoms, helps users check the likely cause of their symptoms more quickly and accurately, and BodyMaps[23], which is an interactive visual search tool that allows users to explore the human using 3D rotatable models.

Finally, **iMedisearch**[24] is an extension of the features of pharmacists' clinical resources website *RPhWorld*[25] which purpose is to enable pharmacists, and

[13]http://www.webmd.com
[14]http://www.medscape.com
[15]www.medicinenet.com
[16]http://emedicine.medscape.com
[17]http://www.emedicinehealth.com
[18]http://www.rxlist.com
[19]http://www.medscape.com/cardiology?t=1
[20]http://www.medscape.org
[21]http://www.healthline.com/
[22]http://www.healthline.com/symptom-checker
[23]http://www.healthline.com/human-body-maps
[24]http://www.imedisearch.com
[25]http://www.rphworld.com

other healthcare professionals, to access free and reliable clinical resources on the internet. iMedisearch is a search engine of reliable online sourses, mainly for pharmacists.

2.2 Data Sources

This work refers to the integration of heterogeneous online data sources over the Internet. The data sources that have been used for the implementation of MIIDLE are:

- *PubMed*, the search engine for expert users that accesses MEDLINE and other online medical data sources,
- *MedlinePlus*, the National Library of Medicine's web site for consumer health information,
- *Google* Search Engine

The features of the above search engines (search/retrieval methods and vocabularies) are presented in this section.

2.2.1 Pubmed

PubMed is a free resource, developed and maintained by National Center for Biotechnology (NCBI) at the National Library of Medicine (NLM). It is a database of bibliographic information drawn primarily from the life sciences literature and it is intended to be used by expert users. PubMed provides free access to MEDLINE, the NLM's database of citations and abstracts in the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical sciences and contains links to full-text articles at participating publishers' web sites as well as links to other third party sites such as libraries and sequencing centers. In addition to MEDLINE citations, PubMed also contains: in-process citations

that provide a record for an article before it is indexed with Medical Subject Headings terms (MeSH)(see section 2.2.5) and added to MEDLINE or converted to out-of-scope status, citations that precede the date that a journal was selected for MEDLINE indexing, some OLDMEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status, citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE, citations to some additional life science journals that submit full-text articles to PubMedCentral[26] and receive a qualitative review by NLM, citations for the majority of books and book chapters available on the NCBI Bookshelf.

A strong feature of PubMed is the automatic translation of a simple query into MeSH terms and subheadings. When a user puts a simple query into PubMed's search field, the system translates it and automatically, adds field names, relevant MeSH terms, synonyms, Boolean operators, and 'nests' the resulting terms appropriately, enhancing the search formulation significantly, in particular by routinely combining (using the OR operator) textwords and MeSH terms. For following example, shows this linking for the query: "*Breast cancer treatment*":

```
("breast neoplasms"[MeSH Terms]
  OR ("breast"[All Fields] AND "neoplasms"[All Fields])
  OR "breast neoplasms"[All Fields]
  OR ("breast"[All Fields] AND "cancer"[All Fields])
  OR "breast cancer"[All Fields]) AND ("therapy"[Subheading]
  OR "therapy"[All Fields]
  OR "treatment"[All Fields]
  OR "therapeutics"[MeSH Terms]
  OR "therapeutics"[All Fields])
```

[26]http://www.ncbi.nlm.nih.gov/pmc/

another example for the simple query "*birth control pills* ":

```
"contraceptives, oral"[Pharmacological Action]
OR "contraceptives, oral"[MeSH Terms]
OR ("contraceptives"[All Fields] AND "oral"[All Fields])
OR "oral contraceptives"[All Fields]
OR ("birth"[All Fields] AND "control"[All Fields] AND "pills"[All Fields])
OR "birth control pills"[All Fields]
```

The retrieval system of PubMed is part of NCBI's vast retrieval system, known as **Enterz** [4]. The basic characteristics of Enterz retrieval system are presented in the following section.

Enterz

The Entrez Global Query Cross-Database Search System is a powerful federated search engine, or Web portal that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information website. It is an integrated search and retrieval system that provides access to more that 30 databases contain over that 690 million records with a single query string and user interface and can efficiently retrieve related sequences, structures, and references. Some of the NCBI's databases that Enterz is accessing are: PubMed, PubMed central, NLM catalog, Taxonomy, BioSample, Protein, Genome, etc [3, 37].

As mentioned above, the common way to access data stored at NCBI databases is through a Web browser using the Web interface that NCBI Enterz system offers to users, where they can search, retrieve for display and download in a selected format if the document is available. Another way is bypassing the Web interface using *Enterz Programming Utilities (E-utilities)*[2]. E-utilities constitute the Application Programming Interface (API) for the Enterz system. This

API provides nine server-side programs that support a uniform set of parameters in order to access the databases for a variety of operations. A fixed URL syntax translates these parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. Each data record has as primary key an integer number called *UID* (unique identifier). A piece of software in any computer language that can send a URL to E-utilities (Java, Perl, Python, C++, etc), can access these data by posting an E-utility URL to NCBI and receive the results in XML format. The fix part of the URL is: *http://eutils.ncbi.nlm.nih.gov/entrez/eutils/*. A brief description of the nine E-utilities follows:

- **EInfo for database statistics**

  URL: *eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi*

  Provides the number of records indexed in each field of a given database, the date of the last update of the database, and the available links from the database to other Entrez databases.

- **ESearch for text searches**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi

  Responds to a text query with the list of matching UIDs in a given database (for later use in ESummary, EFetch or ELink), along with the term translations of the query.

- **EPost: UID uploads**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi

  Accepts a list of UIDs from a given database, stores the set on the History Server, and responds with a query key and Web environment for the uploaded dataset.

- **ESummary: document summary downloads**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi

Responds to a list of UIDs from a given database with the corresponding document summaries.

- **EFetch: data record downloads**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

  Responds to a list of UIDs in a given database with the corresponding data records in a specified format.

- **ELink: Entrez links**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi

  Responds to a list of UIDs in a given database with either a list of related UIDs (and relevancy scores) in the same database or a list of linked UIDs in another Entrez database; checks for the existence of a specified link from a list of one or more UIDs; creates a hyperlink to the primary LinkOut provider for a specific UID and database, or lists LinkOut URLs and attributes for multiple UIDs.

- **EGQuery: global query**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi

  Responds to a text query with the number of records matching the query in each Entrez database.

- **ESpell: spelling suggestions**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/espell.fcgi

  Retrieves spelling suggestions for a text query in a given database.

- **ECitMatch: batch citation searching in PubMed**

  URL: eutils.ncbi.nlm.nih.gov/entrez/eutils/ecitmatch.cgi

  Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.

The E-utilities that have been used for this work are: *EGQuery, Esearch*

and *EFetch*. The EGQuery returns the number of records retrieved in all Enterz databases. The only required parameter is *term*, which contains the query with spaces replaced by '+' signs. The results by default are in HTML format or in XML format using the parameter *retmode=xml*. A sample URL for the query *breast cancer treatment* for XML results is: "*http://eutils.ncbi.nlm.nih.gov/gquery ?term=breast+cancer+treatment&retmode=xml*". Esearch returns a list of the UIDs of results matching the query. Required parameters are *term* and *db* which indicates to Enterz system the database the user wants to search. Optional parameters are *retstart:* Sequential index of the first UID in the retrieved set to be shown in the XML output (default=0), *retmax:* Total number of UIDs from the retrieved set to be shown in the XML output (default=20), *rettype:* indicates if the retrieval XML will the list of UIDs or just the <Count> tab, etc. A sample URL for the query *birth control pills* in order to retrieve the first 100 UIDs is from PubMed is: "*http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed &term=birth+control+pills&retstart=0s&retmax=100*". The retrieval of the corresponding data records can be performed using eFetch. This E-utility returns the formatted records for a list of input UIDs. It requires the name of database with the *db* parameter and comma-separated list of UIDs with the *id* parameter. For example the URL for the retrieval of data records with UIDs 24720068 and 24717251 from PubMed is: "*http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi? db=pubmed&id=24720068,24717251*" (for detailed list of all E-utilites parameters see [1]).

The DTD files of XML results for the three E-utilities can be found in appendices A.1, A.2, A.3.

2.2.2 MedlinePlus

In addition to PubMed, National Library of Medicine offers the *MedlinePlus*, a Web portal for consumer health information. MedlinPlus brings up-to-date

information about diseases, conditions, and wellness issues in simple language, understandable for non-expert users, for free. It includes more than 900 health topic pages (many of these are also available in Spanish), information from more than 1000 organizations and over 35000 links to authoritative health information. Besides the Web interface[27] where a user can set a query, MedlinePlus offers a search-based Web service that provides access to MedlinePlus health topic data in XML format. Using this service, software developers can build applications that utilize MedlinePlus health topic information. The service accepts keyword searches as requests and returns relevant health topics in ranked order. The returned XML files contain complete health topic records. The access can be performed using a fixed URL part: "*http://wsearch.nlm.nih.gov/ws/query*", followed by parameters for the details of the search. Parameter *db* specifies the database, currently only *healthTopics* and *healthTopicsSpanish* are available, and parameter *term* specifies the query. The DTD files of XML results for the three E-utilities can be found in appendix A.4

2.2.3 Google

Google is the most well-known and most used search engine in the World Wide Web. The main function of Google search is to look for text based on a user's query, in Web pages that offered by Web servers and return a list of ranked results mainly based on the *PageRank* algorithm. PageRank assumes that if a Web page is linked from many other *important* Web pages is more likely to be important. That is a hyperlink from an important page counts as a vote of support. The PageRank of a page is counted recursively using the weighted sum of PageRanks of all the pages that link to it. The greater the number of pages with large PageRank pointing to a page, the greater the PageRank of this page. In addition to PageRank, Google has added many other secret criteria for determining the

[27] http://www.nlm.nih.gov/medlineplus/

ranking of pages in the result lists.

Google offers the *Custom Search JSON/Atom API*[28], in order to access the search process programmatically. Using this API, developers can send search requests to Google Search engine and receive the results in JSON or Atom format.

2.2.4 Unified Medical Language System (UMLS)

The **Unified Medical Language System© (UMLS)©** is a compendium of many controlled vocabularies in the biomedical sciences. It has been created in 1986 and its goal is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. It provides a mapping structure among the vocabularies and thus allows one to translate among the various terminology systems. In order to achieve this goal, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs) for use by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research. By design, the UMLS Knowledge Sources are are multi-purpose. They can be applied in systems that perform a range of functions involving one or more types of information, e.g., patient records, scientific literature, guidelines, public health data. There are three UMLS Knowledge Sources: the *Metathesaurus*©, the *Semantic Network*, and the *SPECIALIST Lexicon*, and they are distributed with the associated UMLS software tools: *Lexical Tools* and the *MetamorphoSys* that assist developers in customizing or using the UMLS Knowledge Sources for particular purposes.

[28]https://developers.google.com/custom-search/json-api/v1/overview

UMLS Metathesaurus

The UMLS© Metathesaurus© is a very large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names and their relationships. It is updated twice a year and contains over 150 source vocabularies: electronic versions of classifications, code sets, thesauri, and lists of controlled terms in the biomedical domain, used in electronic health records, patient health record, natural language processing and automated indexing research, linking between different clinical or biomedical vocabularies, information retrieval. Metathesaurus is structured as a set of relational files organized by concept or meaning, and it links alternative names and views of the same concept from different source vocabularies and identifies useful relationships between different concepts that are categorized through *Semantic Network*.

UMLS Semantic Network

Another UMLS knowledge source that is used to support Metathesaurus is the Semantic Network. It provides a consistent categorization of all UMLS Metathesaurus concepts. It consists of:

1. a set of broad subject categories, or *semantic types*, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and

2. a set of useful and important relationships, or semantic relations, that exist between semantic types.

At least one semantic type is assigned to each concept in the Metathesaurus. Semantic types include anatomical structure, biological function, chemical, disease or syndrome, laboratory or test result, medical device, and organism. There are 133 semantic types and 54 semantic relationships. Major groupings of semantic types include:

- organism

- anatomical structure
- biologic function
- chemical
- physical object
- idea or concept

The linking between the semantic types, that provides the structure of the network is created by the relationships in the biomedical domain. The primary link between the semantic types is the "isa" link. The isa link establishes the hierarchy of types within the Semantic Network and facilitates the assignment of the most specific semantic type available for a Metathesaurus concept. There are also 5 major categories of non-hierarchical relationships:

- physically related to
- spatially related to
- temporally related to
- functionally related to
- conceptually related to

SPECIALIST Lexicon

The *SPECIALIST Lexicon* has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System (NLP)[29]. It

[29]http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html

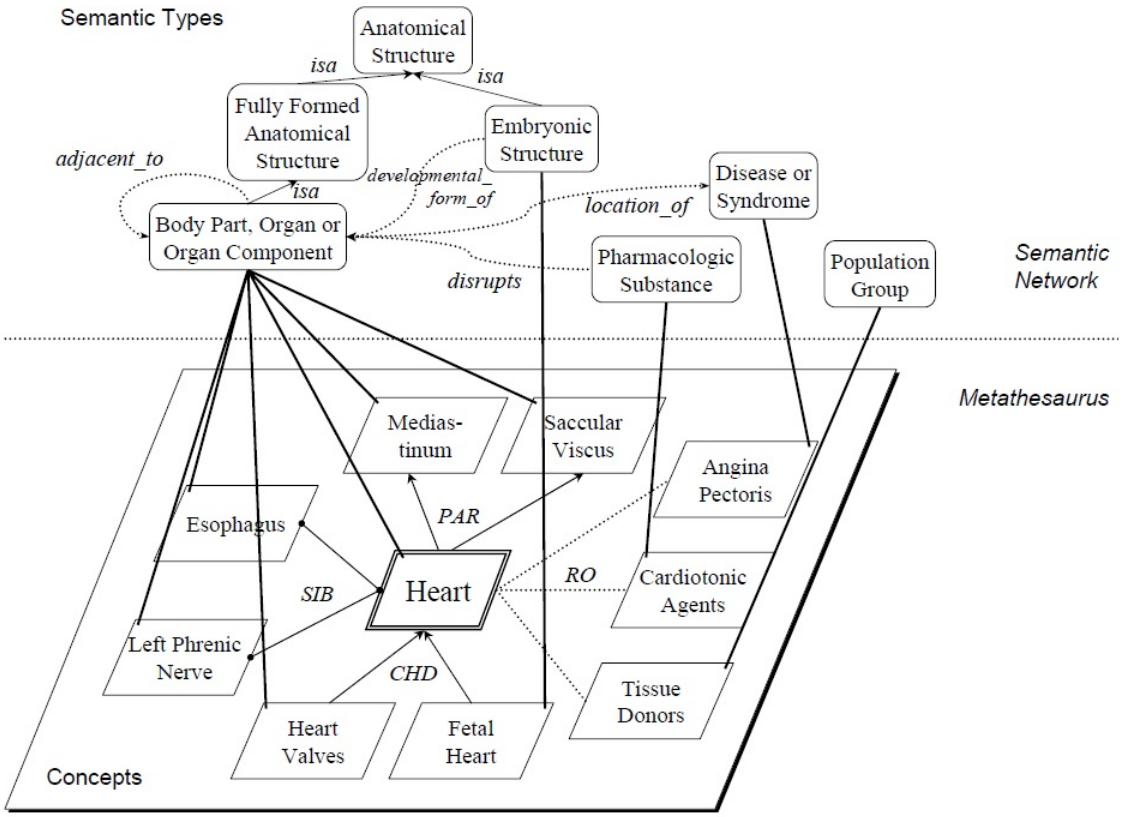


Figure 2.1: Metathesaurus with Semantic Network relations

is a large syntactic lexicon of biomedical and general English. The Lexicon entry for each word or term records the syntactic, morphological, and orthographic information. The Lexicon consists of a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech. Lexical items may be multi-word terms.

2.2.5 Medical Subject Headings (MeSH)

Medical Subject Headings (MeSH) is the NLM's controlled vocabulary thesaurus. It is structured as a taxonomy hierarchy of medical and biological terms only, which represent a subset of UMLS Metathesaurus and used for the purpose of indexing journal articles and books as well as for searching in the life sciences' literature, in other words the goal of MeSH is *"...to provide a reproducible par-*

tition of concepts relevant to biomedicine for purposes of organization of medical knowledge and information..." [31]. It is used by the MEDLINE (see 2.1 in page 4) and PubMed (see 2.2.1). Every term (node) that occurs in MeSH may be thought of as representing a concept.

The structure of MeSH is a hierarchical tree with most general terms higher in the taxonomy than most specific terms. One term can be appear in more than one subtree. There are 16 tree hierarchies (subtrees) which are identified with by letter: A. Anatomy, B. Organisms, C. Diseases, D. Chemical and Drugs, etc[30]. MeSH Records major component in MEDLINE/PubMed are:

- **MeSH headings [MH]**: Also called *"Main Headings"* or *"Descriptors"*. They represent concepts and topics found in the biomedical literature. They are used to index citations in MEDLINE database, for cataloging of publications and other databases, and are searchable in PubMed as [MH]. Most Descriptors indicate the subject of an indexed item, such as a journal article, that is, what the article is about. There are 27,149 descriptors in 2014 MeSH.

- **Subheadings [SH]**: Also called *Qualifiers*[31]. They are attached to MeSH Headings to describe a specific aspect of a concept in conjunction with Descriptors. For example, *Liver/drug effects* indicates that the article or book is not about the liver in general, but about the effect of drugs on the liver. Subheadings are searchable in PubMed as MeSH Subheadings [SH].

- **Supplementary Concept Records - SCR [NM]**: Also called *Supplementary Chemical Records*. They are used to index chemicals, drugs, and other concepts such as rare diseases for MEDLINE and are searchable by Substance Name [NM] in PubMed. SCR are updated weekly, unlike Descriptor and Qualifier records, which are generally updated on an annual

[30] http://www.nlm.nih.gov/bsd/disted/meshtutorial/meshtreestructures/index.html
[31] http://www.nlm.nih.gov/mesh/topsubscope.html

basis. There are currently over 200,000 SCR records with over 505,000 SCR terms.

- **Publication Characteristics [PT]**: They indicate what the indexed item is. They are data are data about the data, rather than being about the content and they are searchable in PubMed as Publication Type [PT].

Other important data objects in MeSH records are:

- **Entry Terms**: They are synonyms, alternate forms, and other closely related terms in a given MeSH record that are generally used interchangeably with the preferred term for the purposes of indexing and retrieval. They are used as pointers to the MeSH Headings. In other words the set of terms that points to a specific MeSH Heading are the terms that represent the concept introduced by the Heading.

- **MeSH tree Number**: Indicates the places within the MeSH hierarchies the MeSH Heading occurs. In the first place appears a letter that indicates one of 16 subtrees which is followed by numbers that indicate the position in this subtree. For example the MeSH Heading *Heart Failure* has the tree number *C14.280.434*.

- **MeSH Scope Note**: A short piece of free text provides a type of definition, in which the meaning of the MeSH Heading is circumscribed. Frequently some other MeSH Headings appear in Scope Note, indicating relationships, which are often very important, but which may not otherwise be represented in the MeSH structure.

The MeSH vocabulary and structure can be accessed through the *MeSH Browser*[32] and the MeSH tree navigation[33].

[32]http://www.nlm.nih.gov/mesh/MBrowser.html
[33]https://www.nlm.nih.gov/cgi/mesh/2014/MB_cgi

2.3 Term Extraction

Term extraction is the process of automatic identification of terms or linguistic expressions that are relevant to a specific domain or concept. Terms can be words or multi-word expressions that are referred to a specific field. The output of a term extraction system must offer not just an unordered list of related terms but a sorted list where each term receives a score which shows the degree of relevance to the specific domain. The term extraction method that is used in this work is the $AMTE_X$ method [19, 21, 20, 18].

2.3.1 AMTE$_X$ - Automatic Term Extraction in Medical Document Collections

AMTE$_X$ [19] is a medical document indexing method, specifically designed for the automatic indexing of documents in large medical collections, such as MEDLINE. The input is the document under consideration and utilizing the MeSH Thesaurus together with the C/NC-value method [12] for term extraction, AMTE$_X$ returns a list of the extracted terms with the corresponding score for each one of them. The processing steps of the AMTE$_X$ method are presented below:

1. *Multi-word Term Extraction:* The C/NC-value method is used for term extraction. During term extraction in AMTE$_X$ the document text is parsed, using the C/NC-value part-of-speech tagger and linguistic filters.

2. *Term Ranking:* Extracted terms are evaluated and the final candidate sorted list is produced. The more important terms, which are more likely to be included in the final list of extracted terms, appear higher in the list.

3. *Term Mapping:* Candidate terms are mapped to terms of the MeSH Thesaurus, in order to filter the non-medical terms. The list of terms now contains only MeSH terms.

4. *Single-word Term Extraction:* For the multi-word terms which do not fully match MeSH, their single word constituents are used for matching. If mapped to a single word MeSH term, the mapped term is added to the term list.

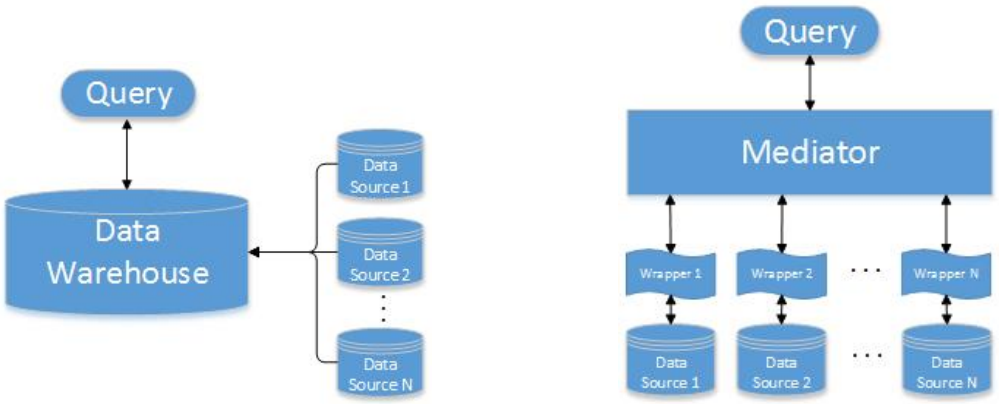
5. *Term Variants:* Term variants are included in the candidate term list. MeSH itself is used for locating variant terms, based on the MeSH term, Entry Terms property. However, only the stemmed term-forms are used in AMTEX since the full list of Entry Terms may contain terms, which often are not synonymous.

6. *Term Expansion:* Each term in the list is expanded using its position in the MeSH tree hierarchy. The semantic similarity of the neighbour terms are examined, either higher or lower in the hierarchy, and if the similarity is greater than a threshold T, they are also included in the list.

2.4 Data Integration

Data integration is a process which combines, in a unified view, heterogeneous data residing in different sources and stored and retrieved using various technologies. Its objective is to offer an incorporate form and structure of these data. There are two general approaches for the process of data integration. First the *Data Warehouse* approach where the data are extracted from multiple sources and transformed in a new database into a single common schema in order to be compatible with each other. Using this approach, the process of integrated search and retrieve is easy and quick because the data are stored locally in a common schema. The disadvantages are the demanding of large storage space and also the data freshness. The update is performed at scheduled times which means that the data are not always up-to-date [24, 39, 40, 35, 7]. The general architecture of Data Warehouse approach can be seen in Figure 2.2(a).

Second the *Mediation* or *Virtual Data Integration* approach [38] accesses the data in real time using a query interface over a single global schema. Through the interface a user can pose a query to multiple sources using the global schema. There are no actual data contained in the mediator. When it receives a query it identifies the relevant data sources and the relevant data in them. Usually the communication to the sources is performed through *Wrappers* which are responsible to translate the query and the results into the appropriate schema [34, 33, 28]. The disadvantage of this approach is that it often needs a lot of time to perform the schema translation and the real time download of the data. The general architecture of virtual data integration approach can be seen in Figure 2.2(b).



(a) Integration Using Data Warehouse (b) Integration Using Mediation

Figure 2.2: Data Integration Approaches

The mapping between the global schema and the schemata of the sources can be done by mapping the sources' elements to the global schema (known as *Local as View - LAV* approach) or by mapping the global schema to each source element (known as *Global As View - GAV* approach) [25, 28]. In LAV approach the query processing is more difficult that in GAV approach. On the other hand the insertion of new sources in the integration system that uses the LAV approach is easier comparing to GAV systems. An attempt to combine the benefits of LAV

and GAV approaches is introduced in *GLAV* mediation language [13].

Data Integration Methodologies and Systems

Since the issue of data integration is not simple to resolve, various methodologies have been proposed and several systems have been developed. One of the first approaches to the data integration issue was the *Information Manifold* project [27, 23, 26]. The main contribution of the Information Manifold was the way it described the contents of the data sources it knew about which was Horn rules and Classic Description Logic. It also proposed the LAV approach. The *Stanford-IBM Manager of Multiple Information Sources - TSIMMIS* [14, 9] was the first project that illustrate the benefits of semi-structured data in data integration. TSIMMIS uses multiple simple mediators which include the knowledge that is necessary for processing specific type of information. They communicate with the sources through wrappers (translators in TSIMMIS) which are responsible for the logical conversion of the queries (which are in a common information model) into requests that the underlying sources can execute, and the conversion of the results that the sources return to the common model. One of the goals of TSIMMIS is the automated or semi-automated generation of mediators from high level descriptions of the information processing they need to do. For this task it provides a module called *mediator generator*. TSIMMIS introduced the GAV approach. Mediators and translators are not required to produce objects with a fixed schema or type. The schema of objects is determined in terms of the processed query and the accessed sources. *Nimble* [10, 11] is a general purpose integration tool that can be used in both internal and external data of an enterprise (relational DBs, data warehouses, legacy systems, web pages, etc). It is based on an XML-like data model and it follows the GAV approach. *Tukwila* [22] is an integration system that involves runtime adaptivity into its core. It includes an advanced query optimizer and adaptive query processing features that allow it

to incrementally re-optimize queries in the middle of execution. *Infomaster* [15] tries to integrate various technologies including Z39.50, SQL databases and web based sources. Its core is a facilitator that determines which sources contain the information necessary to answer each query. It follows the LAV approach. The facilitator designs a strategy for answering the query and performs translations to convert sources' information to a common form. *MedMaker* [32] is an integration system that follows the GAV approach and tries to integrate sources that do not have a well defined static schema. It provides a high level language the Mediator Specification Language, that allows the declarative specification of mediators. When it receives a query, a module called Mediator Specification Interpreter collects and integrates the necessary information from the sources. *PICSEL* [16] is an information integration system that follows the LAV approach. It defines an information server as a knowledge-based mediator in which CARIN [29, 30] is used as the core logical formalism to represent both the domain of applications and the contents of information sources relevant to that domain. *Garlic* [8] is a data integration project that tried to build a multimedia information system capable of integrating not only text but also multimedia data such as images, video, audio, etc. Since much of these data are modeled by objects, the systems provides an object-oriented schema to applications and uses object queries. It uses the GAV approach. The *Mediator EnvirOment for Multiple Information Sources - MOMIS* [6, 5] is a framework for information extraction and integration of heterogeneous information sources. It implements a semi-automatic methodology for the integration and follows the GAV approach. Finally, *KARMA* [36, 17] is an integration platform where a user can describe the kind of data sources he wants to have access by entering examples of the data he want to see in a table with the attributes of the data he wants to retrieve. Once the user provide these sample data, the system will translate them into queries that retrieve this kind of data from multiple sources.

Chapter 3

Information Integration in Web-Based Medical Digital Libraries (MIIDLE)

Data integration for search and retrieval is the automatization of the process that a human would perform in order to obtain data residing in different sources. An integration system send the queries to the selected data sources, and combine the results that are returned in order to provide users with a unified view of information. In this chapter is presented an integration system for medical information sources.

3.1 The *MIIDLE* System

MIIDLE is an integration system for web-based medical digital libraries. It uses specific medical vocabulary in order to perform tasks like term extraction and expansion (described later in this chapter). The basic idea of MIIDLE is to extract terms from the results using the AMTE_X algorithm (see section 2.3.1) and calculate the relevance of each result combining the query and some expansion

of it with the extracted terms. The basic steps of the MIIDLE are:

1. The system sends user's query to each medical source. The duplicates are removed and the results are placed together in a list.

2. The results are merged into one text where term extraction is applied using AMTE$_X$. The idea here is to produce a term representation for each text using terms from a common vocabulary (MeSH).

3. The query is expanded with terms that are synonyms to the user's query terms, in order to enhance it and to support the vocabulary of previous step.

4. The expanded query is compared with the list of term-represented documents of second step, and each document is replaced with its terms that are also in the expanded query.

5. A score for each result is calculated applying vector space model on it. The sorted list is returned to the user.

Figure 3.1 shows the high level architecture of the MIIDLE. Each of the modules is described in details in the following sections.

3.1.1 Query Disseminator

Search process on the Web may be in the form of a specific question, or in the form of a general subject search. Specific questions such as finding out who wrote a paper or locating a citation from a known author or title is a simple and straightforward process. An integration system wouldn't help someone who wishes to perform such a search. He can easily use attributes like *author, title, etc* in a general purpose search engine like Google, or in a topic-specific search engine,

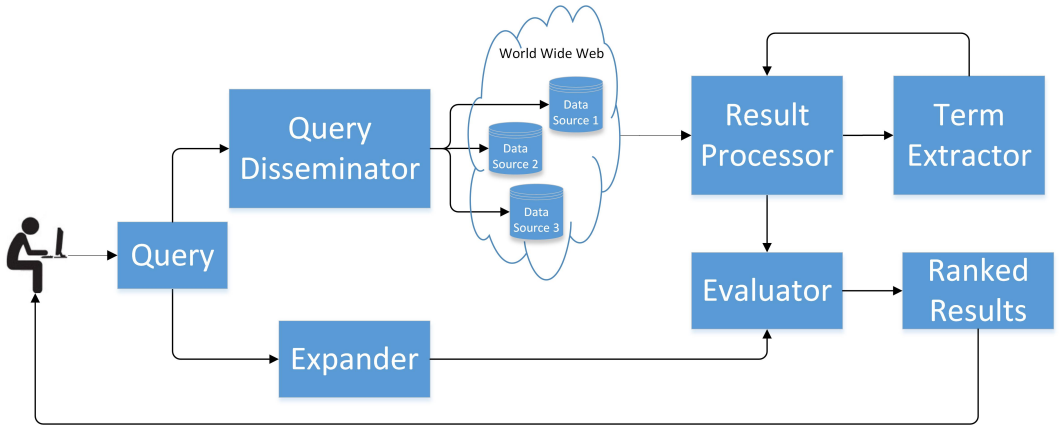


Figure 3.1: High Level Architecture of MIIDLE

and find the information he looks for. On the other hand, a search integration system would offer great help to someone who wants to search for a wide-ranging topic using free text queries. For this reason, and considering that almost every search engine supports free text search, it has been decided the usage of free text for the queries. Thereby queries can be single words or small phrases that directly posed to the sources without any process on it.

We access three sources: PubMed (see section 2.2.1), MedLinePlus (see section 2.2.2) and Google (see section 2.2.3). These sources were selected for two reasons. First they provide an Application Programming Interface (API) in order to access their data, and second they are different from each other as to the type of data they offer. PubMed is mostly a medical data source for experts, google is a general purpose search engine and MedlinePlus is a medical portal mostly for consumer (non-expert) information. The integration process has been designed based on these three types, that can be considered that covers almost the entire search process of medical information offered by the Web.

In order to access PubMed, MIIDLE uses *egquery, esearch* and *efetch* e-utilities (see section 2.2.1). With egquery gets the total number of results that PubMed returns for the query, esearch returns the unique ID for each result and efetch returns the result. The retrieved results from PubMed using eutilities are

in XML format, that contains exact the same information which is in the web page of the corresponding results. Thus, it is easy to isolate the useful for the system information like Abstract and Title for each result. The results that MedlinePlus returns are also in XML format and following the same procedure the results are obtained. On the other hand, Google returns the search results in JavaScript Object Notation (JSON) format where is possible to isolate the external links of the google search. The final results are of course in html format where the only process we have made to them is to remove the tags of html.

One question that had to be answered is the number of the results that MIIDLE should download from each source in order to perform the integration, according to the limitations of the sources, and the constraints of the system. While PubMed and MedlinePlus have no limitation to the number of the results that they return, Google sets a limit of 100 results that can be downloaded using the free API that offers. On the other hand, the extraction process that is based on statistical method (see section 3.1.2), needs a quite big text in order to return accurate results. Finally, we know that the first results of each search engine are the most relevant to the query, and users usually prefer to open the first 10-20 results [1,2]. For the above reason, and for fairness between the three sources, we choose to download the first 100 results from each source for each query.

3.1.2 Term Extractor and Result Mapping

As mentioned above, the term extraction of the integration system is performed using AMTE$_X$. The extraction process of AMTE$_X$ based on the C/NC-value method which combines statistical and linguistic information, and requires big text in order to perform correct extraction. That's why it has been decided to place all the results together in one corpus before pass them to it.

[1]http://searchenginewatch.com/article/2049695/Top-Google-Result-Gets-36.4-of-Clicks-Study

[2]http://searchenginewatch.com/article/2276184/No.-1-Position-in-Google-Gets-33-of-Search-Traffic-Study

The initial idea for the implementation of the extraction process of the integration system was to use the AMTE$_X$ algorithm to extract the MeSH terms from the results, as seen in step 3 (page 20), of AMTE$_X$ algorithm. These terms is difficult to be found in non-expert documents, which are a big part of the results we have retrieved. For example, for the query: *"breast cancer treatment"*, the output terms from AMTE$_X$ algorithm is:

breast neoplasm, carcinoma ductal, mammaplasty, health personnel, therapeutics, clinical trial, carcinoma ductal breast, aromatase inhibitors, risk factors, cells, mass screening, lymph nodes, prostatic neoplasms, carcinoma lobular, education medical, dissection, chemotherapy adjuvant, clinical trials, education medical graduate, mastectomy segmental

It is easy to understand that it is not possible for non-expert documents to contain terms like these, and the trials we've made have shown this. Based on this, we changed the AMTE$_X$ output in order to return the entry terms of MeSH terms. For the above query the extracted terms are:

breast cancer, ductal carcinoma, health care provider, breast reconstruction, treatment, clinical trial, aromatase inhibitors, risk factors, mammary ductal carcinoma, cells, screening, prostate cancer, lobular carcinoma, weight loss, medical education, dissection, adjuvant chemotherapy, lymph nodes, graduate medical education, clinical trials

The number of the extracted terms that were found in the results for MeSH Terms and Entry Terms extraction for the previous query can be seen in Table 3.1.

After the extraction process, each result is replaced with the MeSH terms of the entry terms founded in it, in order to compare it with the expanded query for the evaluation and ranking of the results.

	Extraction Method	
♯ of terms found	**Entry Terms**	**Mesh Terms**
5	2	0
4	7	4
3	38	18
2	96	50
1	85	95
0	6	67
At least one term	228	167

Table 3.1: Number of terms found in results for the query "breast cancer treatment"

3.1.3 Query Expander

Results are replaced with MeSH terms found in them as described in the previous paragraph. The queries that usually users set are single words or small phrases that may not constitute of terms from this vocabulary. In order to create a common basis between the term-represented documents of the previous step and the query, MIIDLE expands it with terms that are synonyms with the original ones and belong to MeSH vocabulary. This expansion also helps to generalize the query in the same search field, without changing its meaning. For example, for the query *"breast cancer treatment"* the expansion process produces the following terms:

human mammary neoplasms, carcinoma human mammary, mammary carcinoma human, cancer of breast, mammary neoplasms human, neoplasms human mammary, breast tumor, breast tumors, mammary carcinomas human, human mammary carcinoma, neoplasm human mammary, mammary neoplasm human, cancer breast, neoplasms breast, tumors breast, carcinomas human mammary, breast cancer, human mammary neoplasm, neoplasm breast, tumor breast, breast neoplasm, breast neoplasms, cancer of the breast, human mammary carcinomas, therapeutics, treatment, therapeutic, treatments

Although the expansion of the query augment its capability, it is used only in the process of ranking in combination with the search results. It is not used in the primary search on the medical sources, because it reduces the results (almost to 0) even in the google search engine.

3.1.4 The Ranking Process

After the replacement of the results' text with the MeSH terms, the integration system finds the common terms between them and the terms of the expanded query and performs the evaluation which leads to the final ranking of the documents. The evaluation of the results in combination with the expanded query, is performed using Vector Space Model. If we consider the query and the replaced with extracted terms results as vectors:

$d_i = (t_1, t_2, \ldots, t_k)$ (The i result with k terms)

$q = (q_1, q_2, \ldots, q_n)$ (The expanded query with n terms)

we can find the similarity between them, by calculating the cosine between these vectors:

$$Sim(\overrightarrow{q}, \overrightarrow{d_i}) = \frac{\overrightarrow{q}\,\overrightarrow{d}}{|\overrightarrow{q}||\overrightarrow{d}|} = \frac{\sum_{i=1}^{M} w_{iq} w_{id}}{\sqrt{\sum_{i=1}^{M} w_{iq}^2 \sum_{i=1}^{M} w_{id}^2}}$$

Although AMTE$_X$ evaluates the extracted terms and the final list is ranked with a score attached to each one of the them, it has been decided not to use

these scores for the final evaluation of the results, because their values are not representative for the terms as they come from the big corpus that AMTE$_X$ used for the extraction. For the calculation of the w_ij weights, we used the tf-idf weighting scheme:

$$w_{ij} = \frac{freq_{ij}}{maxfreq_{lj}} \log \frac{N}{n_i}$$

Where:

- w_{ij}: weight of term t_i associated with document d_j
- $freq_{ij}$: frequency of term t_i in document d_j
- $maxfreq_{li}$: maximum frequency over all terms in d_j
- n_i: number of documents where term t_i occurs

As the process of the expansion gives only synonyms of the original query (see section 3.1.3), we consider that the weights of the terms of the query equal 1. Using the scores of the above calculations, the results are sorted and returned to the user. Figure 3.2 shows MIIDLE in details.

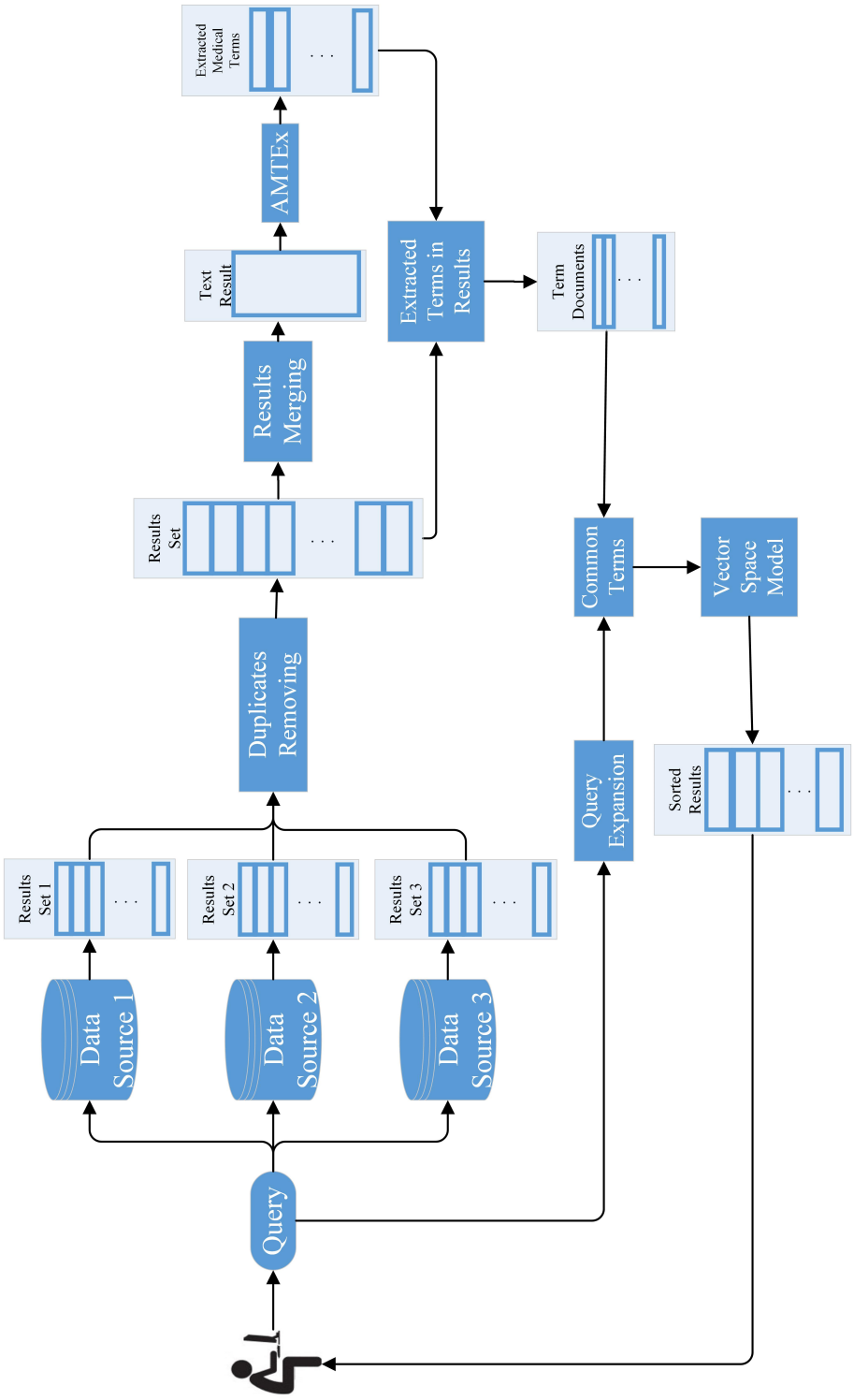


Figure 3.2: MIIDLE Integration System

Chapter 4

Evaluation

Integration attempts to simulate what a user would do in order to search in different data sources. Therefore, it is a process that involves subjectivity and the results may fulfil the intent meaning of the query issued by the user. In the following, the accuracy of MIIDLE is assessed by users and the results of this evaluation are presented and discussed.

4.1 Evaluation Setup

The users that were asked to evaluate MIIDLE are twelve people who are either physicians and health professionals with experience in searching in medical data bases or naive users. In order to perform the evaluation, we randomly chose queries from the set of queries we used for the implementation and debugging of MIIDLE. These queries are simple phrases from a variety areas of medical science. User were given the twenty first results from each one of the three sources that MIIDLE accessed and the first twenty ranked results from the output of MIIDLE and they were asked to rate them in terms of their relevance to the query on a scale from 0 (completely irrelevant) to 4 (completely relevant). Table 4.1 shows the type of users and the corresponding queries.

Users	Queries
Paediatrician	1. Asthma Treatment 2. Viral Infections
Neurologist	1. Brain Injury 2. Insomnia Treatment 3. Brain Cancer Treatment
Gastroenterologist	Kidney Failure
Pharmacist	Alternative Medicine
Pathologist-Nutritionist	Obesity and Weight Loss
Orthopaedist	Low Back Pain
Naive Users	1. Breast Cancer Treatment 2. Breast Cancer Risk During Hormone Therapy 3. Cancer Chemotherapy

Table 4.1: Users and Queries for the evaluation of MIIDLE

Users were asked to rate the results both in terms of their relevance to the query, and their position in the result set returned by each source, given that higher position means that the result should be more relevant to the query.

4.2 Evaluation of the Results

The evaluation of the results using the ratings of the users is performed using three measures. The sum of the scores for each query and each source, the total number of each level (0 to 4) for each query and each source as well as the score of each result in a particular position in the set of result of each source and finally the precision for each source. Unfortunately it is not possible to calculate the recall for each query because the results are not locally based and the *False Negative* results are unknown.

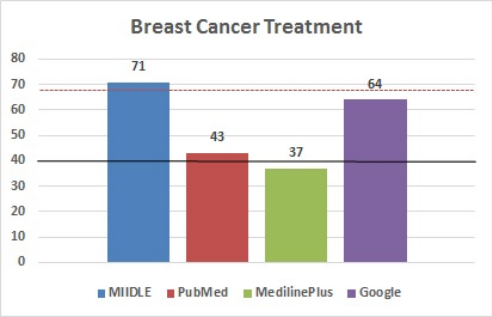
4.2.1 Sum of Scores for Each Source

Considering that the highest rate is 4 and there were given the first 20 results to the users, we have put two thresholds for the scores. The first is 40 which is the half of the highest sum that can be achieved and the second is 68 which is

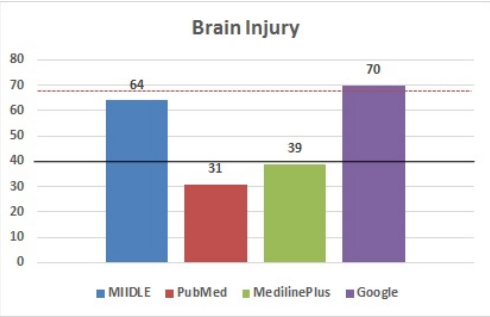
the mean value of the scores of two sources that gave the best results: MIIDLE and Google, in order to compare them. The results can be seen in Table 4.2 and graphically in figures 4.1(a)-4.1(m).

	Sum of Scores			
Query	**MIIDLE**	**PubMed**	**MedlinePlus**	**Google**
Breast cancer treatment	71	43	37	64
Brain Injury	64	31	39	70
Breast Cancer Risk During Hormone Therapy	63	46	4	64
Cancer chemotherapy	60	27	45	67
Viral Infections	62	43	42	56
Low Back Pain	78	48	38	73
Insomnia Treatment	74	41	8	73
Obesity and Weight Loss	65	42	36	72
Asthma Treatment	65	40	24	74
Alternate Medicine	74	12	24	74
Kidney Failure	63	37	28	73
Brain Cancer Treatment	75	29	18	46
Total Sum	814	439	343	806

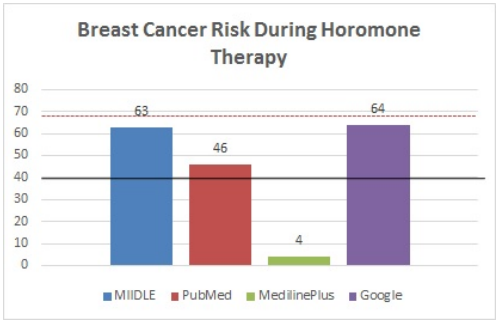
Table 4.2: Sum of scores for the queries for each data source



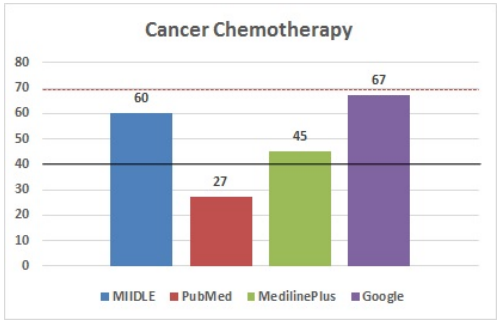
(a) Query: Breast Cancer Treatment



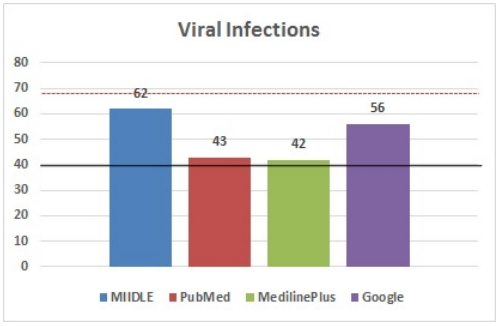
(b) Query: Brain Injury



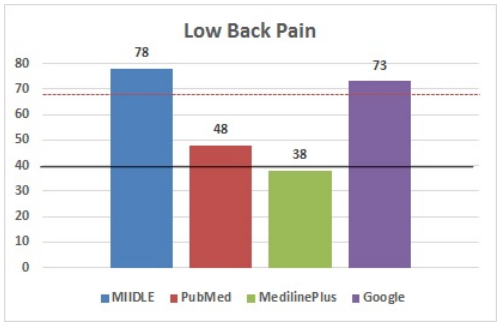
(c) Query: Breast Cancer Risk During Hormone Therapy



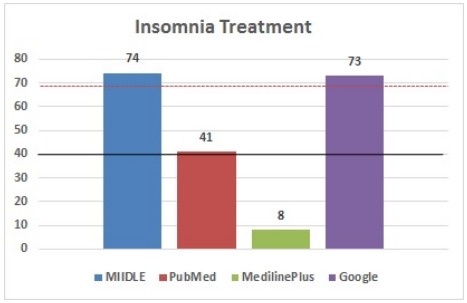
(d) Query: Cancer Chemotherapy



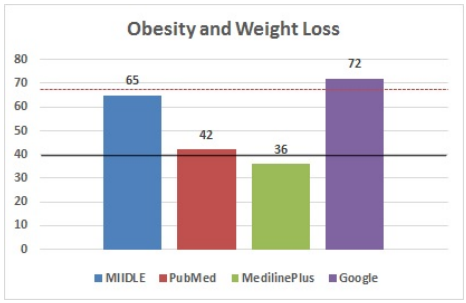
(e) Query: Viral Infections



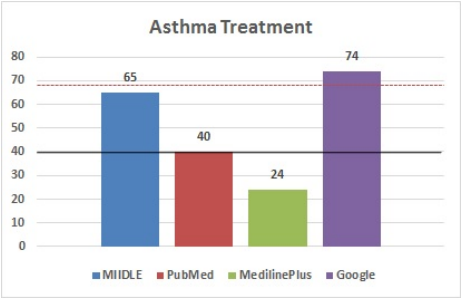
(f) Query: Low Back Pain



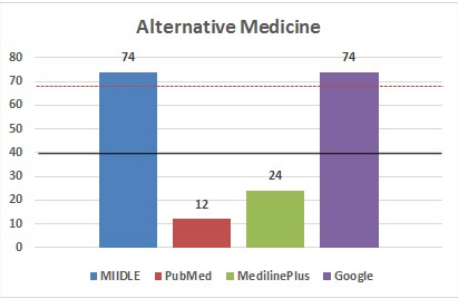
(g) Query: Insomnia Treatment



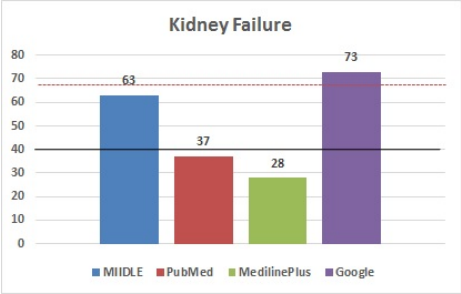
(h) Query: Obesity and Weight Loss



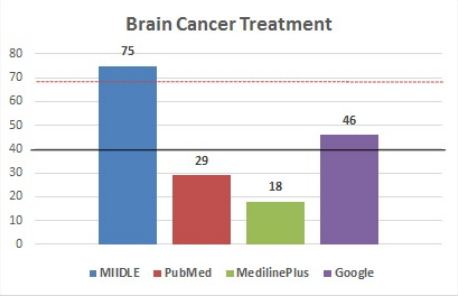
(i) Query: Asthma Treatment



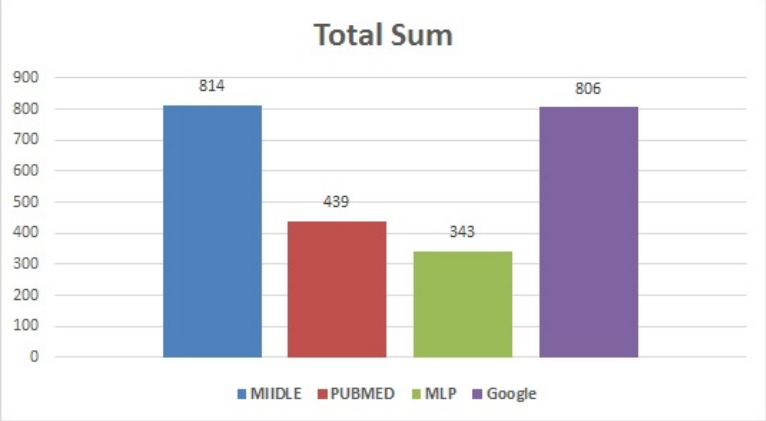
(j) Query: Alternative Medicine



(k) Query: Kidney Failure



(l) Query: Brain Cancer Treatment



(m) Sum of All Results

Figure 4.1: Sum of scores for the queries for each data source

It is obvious that the sum of the grades for both MIIDLE and Google for all the queries are above 40 while seven queries from PubMed and only two from

Performance	MIIDLE	PubMed	MedlinePlus	Google
Equal or above 40	12	7	2	12
Equal or above 68	5	0	0	7

Table 4.3: Performance of Sources According to Their Sum Scores

MedlinePlus satisfy this threshold. Finally, for four queries for the MIIDLE the score is above 68 while the same threshold is achieved for seven queries for Google (Table 4.3).

From the above figures it is apparent that, according to users' opinion, the results of MIIDLE and Google are much more relevant to the submitted queries than the results of PubMed and MedlinePlus with a little better results for Google. The mean values of the results are shown in figure 4.2.

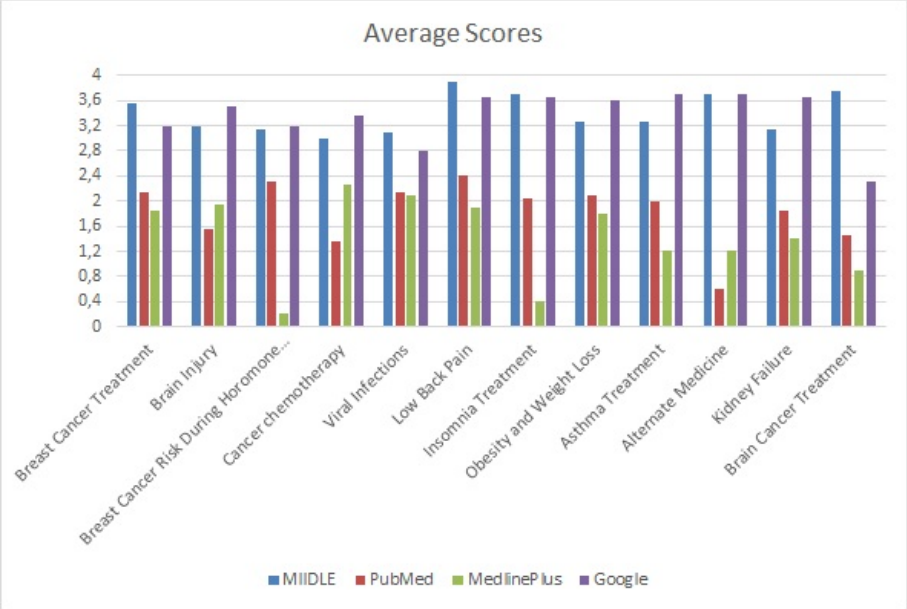


Figure 4.2: Mean Value of all the scores for each query

4.2.2 Scoring Levels

As have been mentioned, there are five level of grading in the evaluation process: 0-4. The number of responses for each level of relevance for each query, as well as the score of each result in each query is shown in the Tables 4.4-4.15. In Figures

4.3 - 4.14 are presented the level of scores of each result in the position that they appear.

	Breast Cancer Treatment			
	MIIDLE	**PM**	**MP**	**G**
4	13 (65%)	3 (15%)	3 (15%)	10 (50%)
3	5 (25%)	7 (35%	4 (20%)	5 (25%)
2	2 (10%)	4 (20%)	3 (15%)	4 (20%)
1	0	2 (10%)	7 (35%)	1 (5%)
0	0	4 (20%)	3 (15%)	0

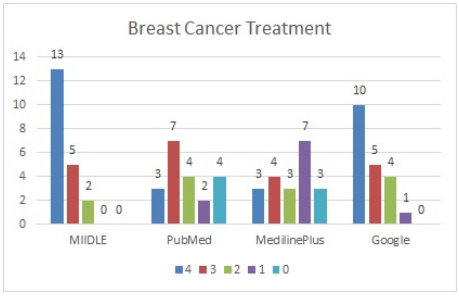


Table 4.4: Number of scores for each level for the query "Breast Cancer Treatment"

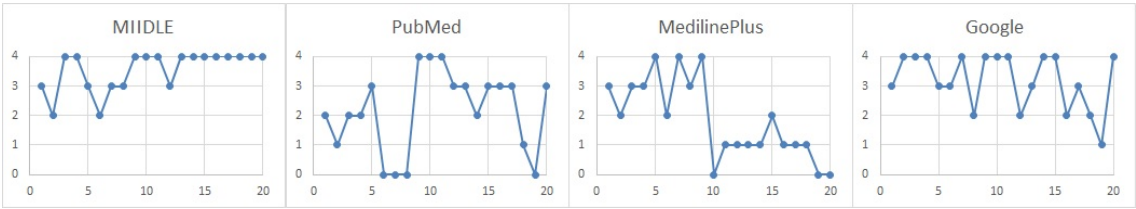


Figure 4.3: Score in each level for the query "Breast Cancer Treatment"

	Brain Injury			
	MIIDLE	**PM**	**MP**	**G**
4	9 (45%)	1 (5%)	2 (10%)	11 (55%)
3	8 (40%)	3 (15%)	2 (10%)	8 (40%)
2	1 (5%)	6 (30%)	9 (45%)	1 (5%)
1	2 (10%)	6 (30%)	7 (35%)	0
0	0	4 (20%)	0	0

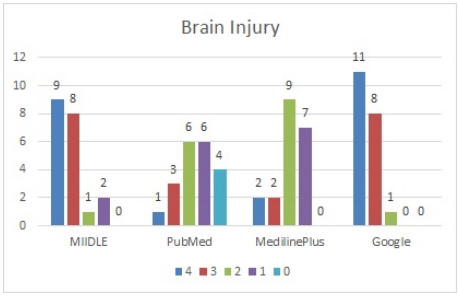


Table 4.5: Number of scores for each level for the Query "Brain Injury"

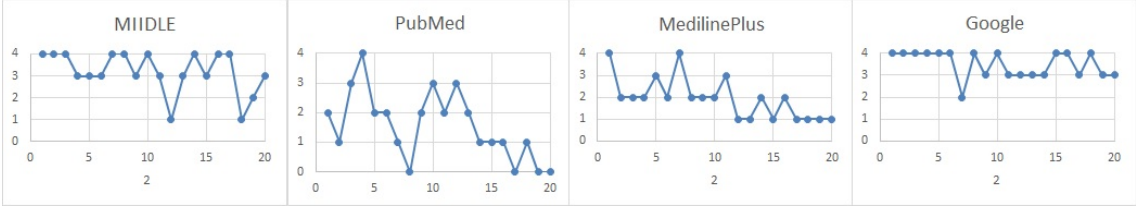


Figure 4.4: Score in each level for the query "Brain Injury"

	Breast Cancer Risk During Hormone Therapy			
	MIIDLE	**PM**	**MP**	**G**
4	11 (55%)	3 (15%)	1 (5%)	11 (55%)
3	2 (10%)	7 (35%	0	4 (20%)
2	6 (30%)	6 (30%)	0	3 (15%)
1	1 (5%)	1 (5%)	0	2 (10%)
0	0	3 (15%)	19 (95%)	0

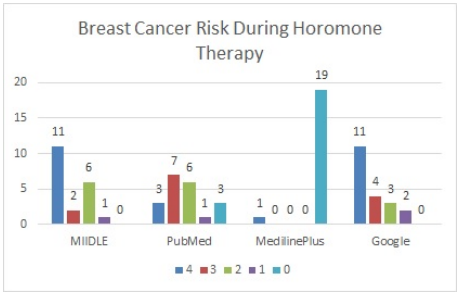


Table 4.6: Number of scores for each level for the Query "Breast Cancer Risk During Hormone Therapy"

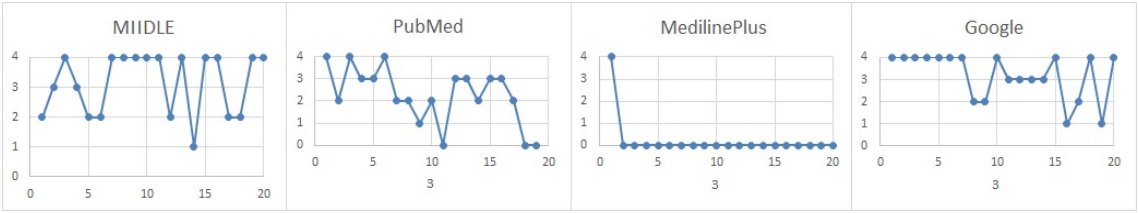


Figure 4.5: Score in each level for the query "Breast Cancer Risk During Hormone Therapy"

	Cancer Chemotherapy			
	MIIDLE	**PM**	**MP**	**G**
4	5 (25%)	0	1 (5%)	10 (50%)
3	11 (55%)	3 (15%)	5 (25%)	7 (35%)
2	3 (15%)	5 (25%)	12 (60%)	3 (15%)
1	1 (5%)	8 (40%)	2 (10%)	0
0	0	4 (20%)	3 0	0

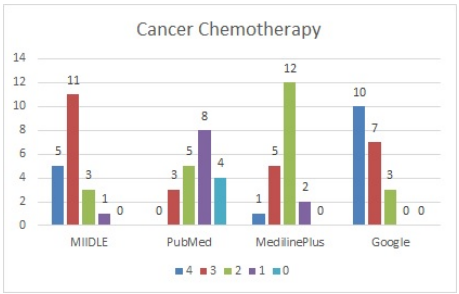


Table 4.7: Number of scores for each level for the Query "Cancer Chemotherapy"

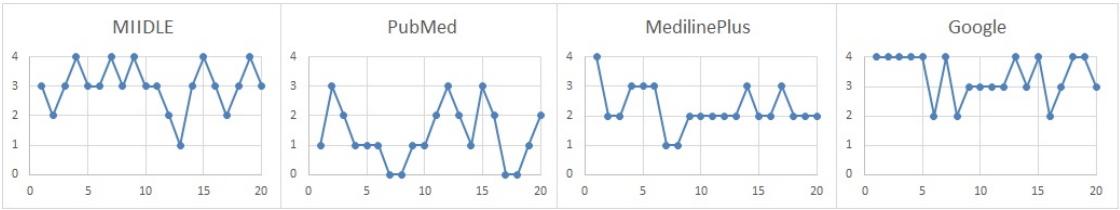


Figure 4.6: Score in each level for the query "Cancer Chemotherapy"

	Viral Infections			
	MIIDLE	**PM**	**MP**	**G**
4	9 (45%)	0	3 (15%)	6 (30%)
3	5 (25%)	8 (40%	4 (20%)	7 (35%)
2	5 (25%)	7 (35%)	3 (15%)	5 (25%)
1	1 (5%)	5 (25%)	7 (35%)	1 (5%)
0	0	0	3 (15%)	1 (5%)

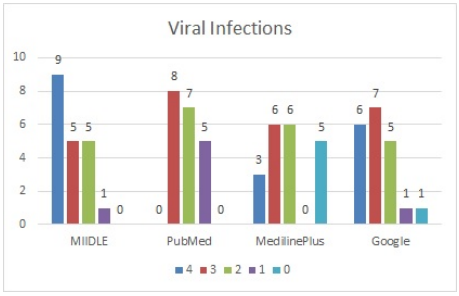


Table 4.8: Number of scores for each level for the Query "Viral Infections"

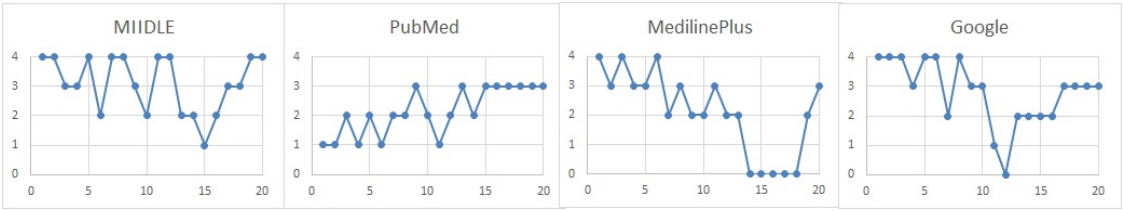


Figure 4.7: Score in each level for the query "Viral Infections"

	Low Back Pain			
	MIIDLE	**PM**	**MP**	**G**
4	18 (90%)	1 (5%)	1 (5%)	14 (70%)
3	2 (10%)	10 (50%	5 (25%)	5 (25%)
2	0	6 (30%)	7 (35%)	1 (5%)
1	0	2 (10%)	5 (25%)	0
0	0	1 (5%)	2 (10%)	0

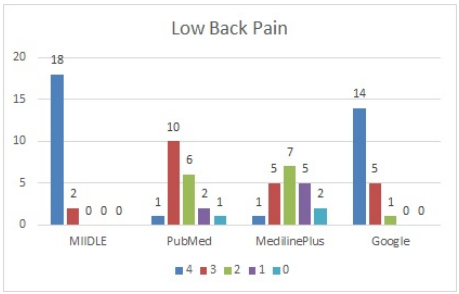


Table 4.9: Number of scores for each level for the Query "Low Back Pain"

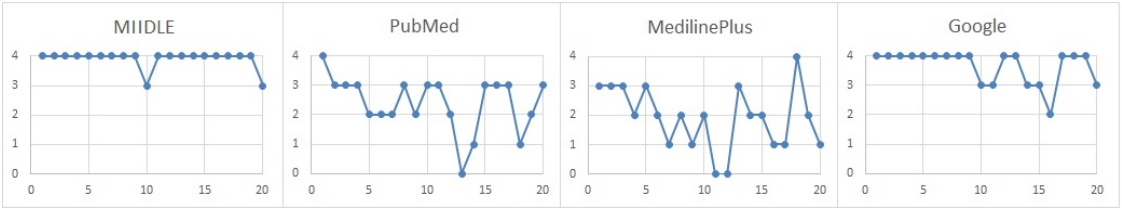


Figure 4.8: Score in each level for the query "Low Back Pain"

	Insomnia Treatment			
	MIIDLE	**PM**	**MP**	**G**
4	15 (75%)	1 (5%)	0	13 (65%)
3	4 (20%)	5 (25%	2 (10%)	7 (35%)
2	1 (5%)	10 (50%)	1 (5%)	0
1	0	2 (10%)	0	0
0	0	2 (10%)	17(85%)	0

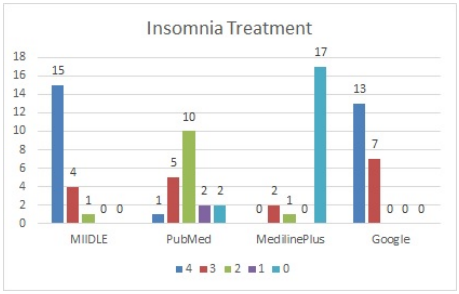


Table 4.10: Number of scores for each level for the Query "Insomnia Treatment"

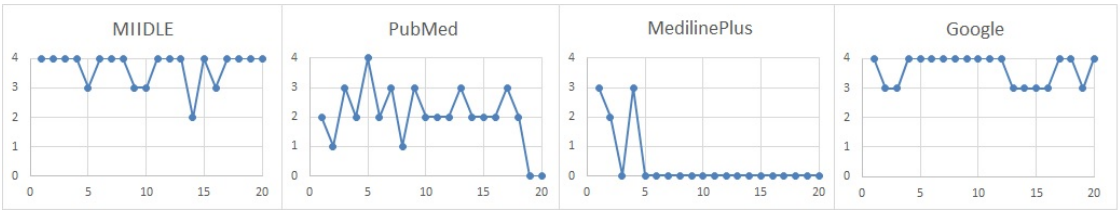


Figure 4.9: Score in each level for the query "Insomnia Treatment"

	Obesity and weight Loss			
	MIIDLE	**PM**	**MP**	**G**
4	9 (45%)	1 (5%)	4 (20%)	13 (65%)
3	7 (35%)	4 (20%)	3 (15%)	6 (30%)
2	4 (20%)	12 (60%)	4 (20%)	1 (5%)
1	0	2 (10%)	3 (15%)	0
0	0	1 (5%)	6 (30%)	0

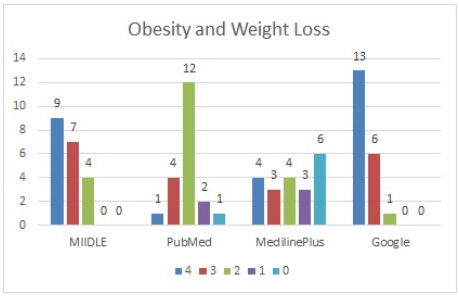


Table 4.11: Number of scores for each level for the Query "Obesity and Weight Loss"

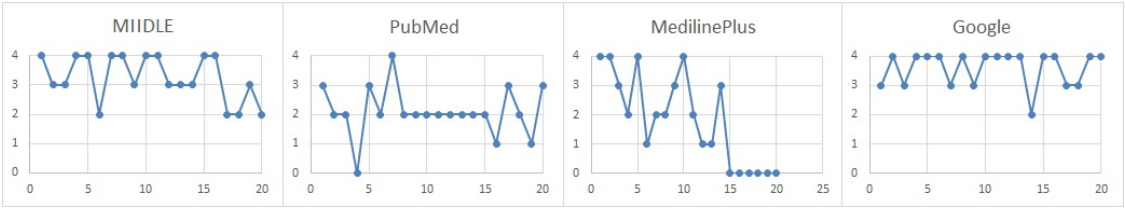


Figure 4.10: Score in each level for the query "Obesity and Weight Loss"

	Asthma Treatment			
	MIIDLE	**PM**	**MP**	**G**
4	11 (55%)	3 (15%)	1 (5%)	15 (75%)
3	4 (20%)	3 (15%	3 (15%)	4 (20%)
2	4 (20%)	6 (30%)	3 (15%)	1 (5%)
1	1 (5%)	7 (35%)	5 (25%)	0
0	0	1 (5%)	8 (40%)	0

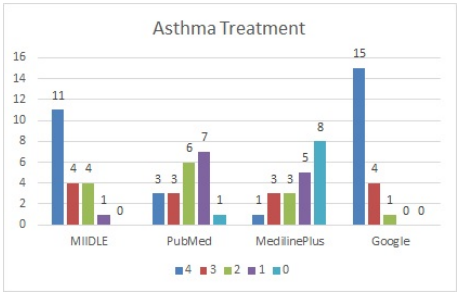


Table 4.12: Number of scores for each level for the Query "Asthma Treatment"

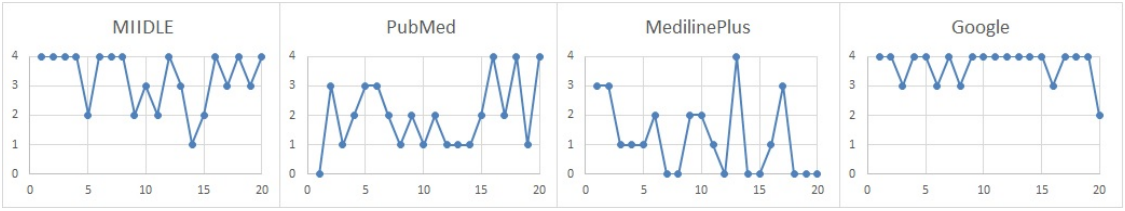


Figure 4.11: Score in each level for the query "Asthma Treatment"

	Alternative Medicine			
	MIIDLE	**PM**	**MP**	**G**
4	15 (75%)	1 (5%)	6 (30%)	16 (80%)
3	4 (20%)	2 (10%)	0	2 (10%)
2	1 (5%)	1 (5%)	0	2 (10%)
1	0	0	0	0
0	0	16 (80%)	14 (70%)	0

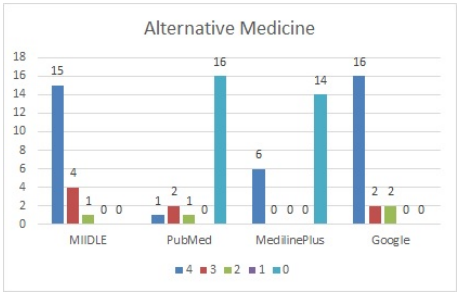


Table 4.13: Number of scores for each level for the Query "Alternative Medicine"

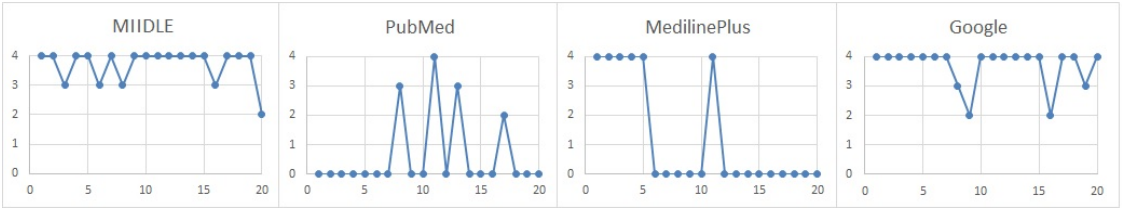


Figure 4.12: Score in each level for the query "Alternative Medicine"

	Kidney Failure			
	MIIDLE	**PM**	**MP**	**G**
4	9 (45%)	1 (5%)	3 (15%)	15 (75%)
3	6 (30%)	6 (30%)	3 (15%)	4 (20%)
2	4 (20%)	4 (20%)	2 (10%)	0
1	1 (5%)	7 (35%)	3 (15%)	1 (5%) (5%)
0	0	2 (10%)	9 (45%)	0

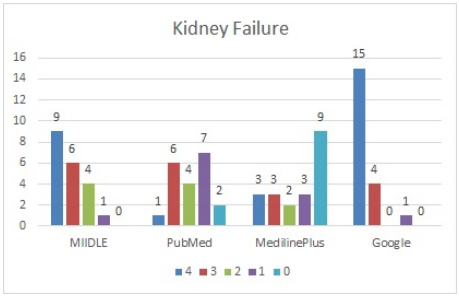


Table 4.14: Number of scores for each level for the Query "Kidney Failure"

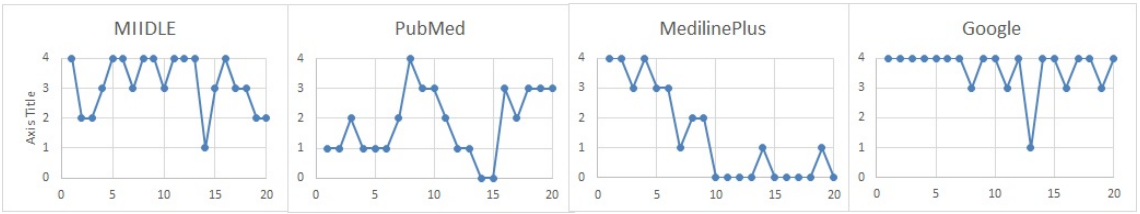


Figure 4.13: Score in each level for the query "Kidney Failure"

	Brain Cancer Treatment			
	MIIDLE	**PM**	**MP**	**G**
4	15 (75%)	3 (15%)	1 (5%)	5 (25%)
3	5 (25%)	2 (10%)	1 (5%)	2 (10%)
2	0	4 (20%)	2 (10%)	7 (35%)
1	0	3 (15%)	7 (35%)	6 (30%)
0	0	8 (40%)	9 (45%)	0

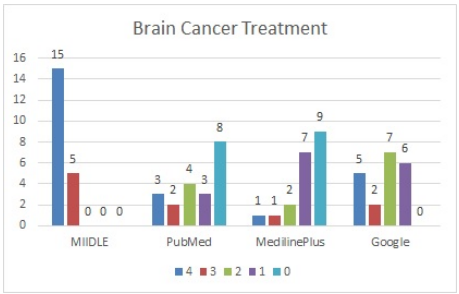


Table 4.15: Number of scores for each level for the Query "Brain Cancer Treatment"

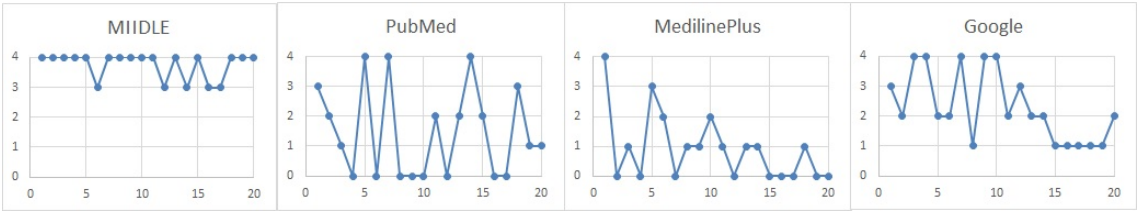


Figure 4.14: Score in each level for the query "Brain Cancer Treatment"

Observing these results we calculate some interesting measures about the the score levels for each source. We consider that documents with 3 or 4 are the

most relevant to the query and that if a source has 80% of its result with this score level is acceptable. The above is satisfied for seven queries for MIIDLE, for eight queries for Google and for none for PubMed and MedlinePlus. For the levels 2 or 3 or 4 the corresponding documents for each query are for MIIDLE eleven, for PubMed four, for MedlinePlus one and for Google eleven. Also the minimum percentage of the results that has level of score 3 or 4 is 65% for MIIDLE (for the query "Breast Cancer Risk During Hormone therapy"), 65% for Google (for the query "Viral Infections"), 15% for PubMed (for the queries "Alternative Medicine" and "Cancer Chemotherapy") and 5% for MedlinePlus(for the query "Breast Cancer Risk During Hormone Therapy"), while for the score levels 2 or 3 or 4 is 90% for MIIDLE and Google, 20% for PubMed and 5% for MedlinePlus. The percentage of documents that has score level 3 or 4 are shown in Figure 4.15, and for score level 2 or 3 or 4 in Figure 4.16

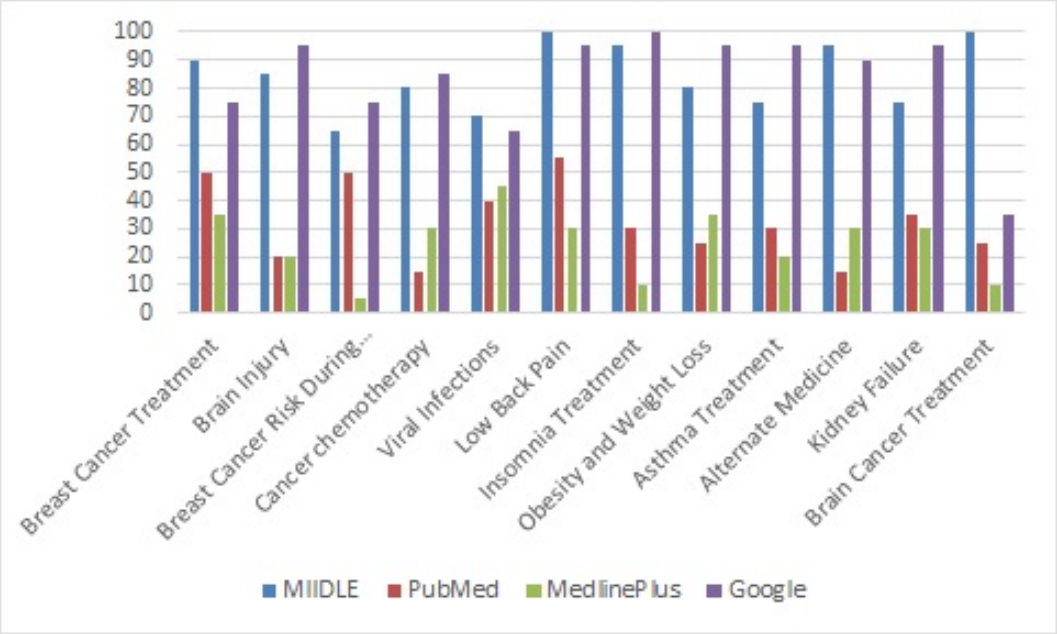


Figure 4.15: Percentage of documents in each result that has score level 3 or 4

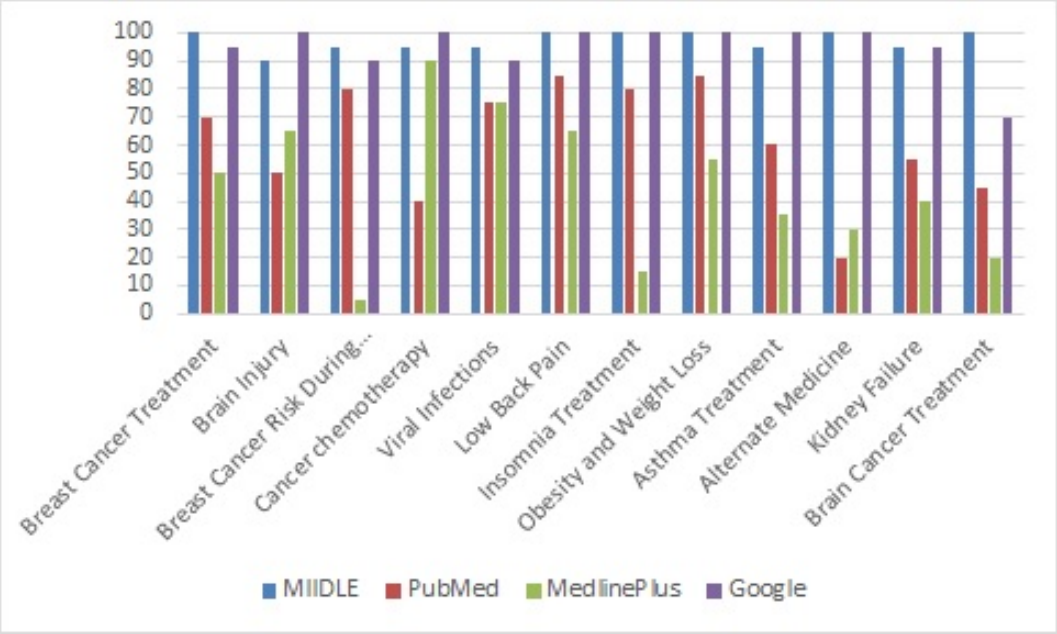


Figure 4.16: Percentage of documents in each result that has score level 2 or 3 or 4

4.2.3 Precision

Precision for each query and each source is the percentage of the correctly retrieved documents compared to the total number of the retrieved documents. Since the scores that users have asked to give to the results are not boolean, we use the two thresholds of the previous section in order to classify the documents into the set of correctly retrieved (true positive) and false retrieved (false positive) results: th_1=2 which means that the results that took score 2 or more are considered as true positive and th_2=3 which means that the results that took 3 or 4 are the true positive. Using the above scores and these thresholds we can calculate the precision for each query and each result. Figure 4.17 shows the precision for th_1 and Figure 4.18 the precision for th_2.

Figure 4.17: Precision for th_1=*2*

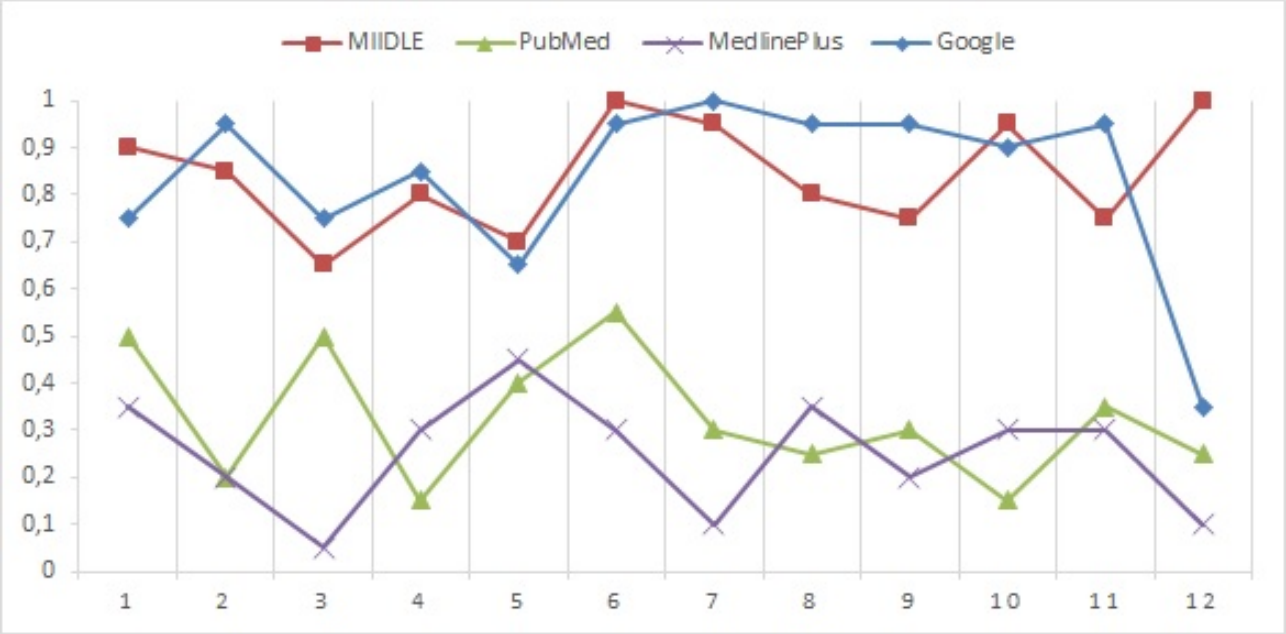


Figure 4.18: Precision for th_2=*3*

It is clear that for $th_1 = 2$ precision is quite stable for MIIDLE and Google, between 0.9 and 1, while for PubMed varies between 0.2 to 0.8 and for Medline-

Plus between 0.05 to 0.9. For $th_2 = 3$ precision for MIIDLE and Google varies from 0.65 to 1, for PubMed from 0.15 to 0.55 and for MedlinePlus form 0.05 to 0.45.

Finally, in the following table is presented the mean values of the precision for each source for thresholds 2 and 3:

	MIIDLE	**PubMed**	**MedlinePlus**	**Google**
$th_1 = 2$	97,00%	62.08%	45.00%	95.00%
$th_2 = 3$	84.17%	35.50%	25%	83.33%

Table 4.16: Mean Value of Precision for $th_1 = 2$ and $th_2 = 3$

Chapter 5

Concluding Remarks

We presented MIIDLE, an approach for medical data integration. MIIDLE retrieves medical documents from the sources that tries to integrate and uses AMTE_X method in order to extract the MeSH terms from these documents. The ranking process is performed by applying vector-space model on them using the extracted MeSH terms and an expanded version of the query used for the retrieval. The results are ranked according to their relevance to the query. The evaluation of the ranked results is performed by users who were asked to rate them. For the experiments we integrate three sources: PubMed, MedlinePlus and results that retrieved by posing the queries to Google search engine. The evaluation of ranked results from MIIDLE and the results from the sources it accessed showed that, according to the users' opinion, MIIDLE and Google are by far better than the other two sources, with MIIDLE a little better than Google.

The general idea of the above system is that it performs ranking of documents in terms of a query using a thesaurus of a specific scientific field, and an extraction method that uses statistical and linguistic information in order to extract terms of the thesaurus from the documents. Using an appropriate thesaurus this method can be extended to other scientific fields.

Bibliography

[1] The e-utilities in-depth: Parameters, syntax and more. http://www.ncbi.nlm.nih.gov/books/NBK25499/.

[2] Introduction to enterz e-utilities: sample applications, parameters, syntax. http://www.ncbi.nlm.nih.gov/books/NBK25497/ - http://www.ncbi.nlm.nih.gov/books/NBK25501/.

[3] National center for biotechnology information, NCBI databases. http://www.ncbi.nlm.nih.gov/guide/all/.

[4] The ncbi handbook, 2nd edition. http://www.ncbi.nlm.nih.gov/books/NBK143764/.

[5] Domenico Beneventano and Sonia Bergamaschi. The momis methodology for integrating heterogeneous data sources. In *18 th IFIP World Computer Congress*. Kluwer, 2004.

[6] Domenico Beneventano, Sonia Bergamaschi, Silvana Castano, Alberto Corni, R. Guidetti, G. Malvezzi, Michele Melchiori, and Maurizio Vincini. Information integration: The MOMIS project demonstration. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 611–614, 2000.

[7] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. A principled approach to data integration and reconcil-

iation in data warehousing. In *In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99*, 1999.

[8] Michael J. Carey, Laura M. Haas, Peter M. Schwarz, Manish Arya, William F. Cody, Ronald Fagin, Myron Flickner, Allen Luniewski, Wayne Niblack, Dragutin Petkovic, Joachim Thomas II, John H. Williams, and Edward L. Wimmers. Towards heterogeneous multimedia information systems: The garlic approach. In *RIDE-DOM*, pages 124–131, 1995.

[9] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The tsimmis project: Integration of heterogenous information sources. In *Information Processing Society of Japan (IPSJ 1994)*, 1994.

[10] D. Draper, A.Y. Halevy, and D.S. Weld. The nimble xml data integration system. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 155–160, 2001.

[11] Denise Draper, Alon Y. Halevy, and Daniel S. Weld. The nimble integration engine. *SIGMOD Rec.*, 30(2):567–568, May 2001.

[12] K.T. Frantzi and S. Ananiadou. The C/NC value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–180, 1999.

[13] Marc Friedman, Alon Levy, and Todd Millstein. Navigational plans for data integration. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 67–73. AAAI Press/The MIT Press, 1999.

[14] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and Jennifer Widom. Integrating and accessing heterogeneous information sources in tsimmis. In *Proceedings of the AAAI Symposium on Information Gathering*, pages 61–64, March 1995.

[15] Michael R. Genesereth, Arthur M. Keller, and Oliver M. Duschka. Infomaster: An information integration system. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 539–542, New York, NY, USA, 1997. ACM.

[16] Francois Goasdoue, Veronique Lattes, and Marie-Christine Rousset. The use of CARIN language and algorithms for information integration: The PICSEL system. *International Journal of Cooperative Information Systems*, 9:383–401, 2000.

[17] Andreas Harth, Craig Knoblock, Steffen Stadtmüller, Rudi Studer, and Pedro Szekely. On-the-fly integration of static and dynamic sources. In *Proceedings of the Fourth International Workshop on Consuming Linked Data (COLD2013)*, 2013.

[18] A. Hliaoutakis. Automatic Term Indexing in Medical Text Corpora and its Applications to Consumer Health Information Systems. Master's thesis, Department of Electronic and Computer Engineering, Technical University of Crete, Greece, 2009.

[19] Angelos Hliaoutakis, Kaliope Zervanou, and Euripides G. M. Petrakis. The AMTEx approach in the medical document indexing and retrieval application. *Data and Knowledge Engineering (DKE)*, 68(1):380–392, 2009.

[20] Angelos Hliaoutakis, Kaliope Zervanou, Euripides G. M. Petrakis, and Evangelos E. Milios. Automatic document indexing in large medical collections. In *ACM International Workshop on Health Information and Knowledge Management (HIKM 2006)*, Arlington, VA, USA, 2006.

[21] Angelos Hliaoutakis, Kalliopi Zervanou, and Euripides G. M. Petrakis. Medical document indexing and retrieval: AMTEx vs. NLM MMTx. In *12th*

International Symposium for Health Information Management Research, Sheffield, UK, 2007.

[22] Zachary Ives, Daniela Florescu, Inria Roquencourt, Marc Friedman, Alon Levy, and Daniel Weld. An adaptive query execution system for data integration. pages 299–310, 1999.

[23] Thomas Kirk, Alon Y. Levy, Yehoshua Sagiv, and Divesh Srivastava. The information manifold. In *In Proceedings of the AAAI 1995 Spring Symp. on Information Gathering from Heterogeneous, Distributed Enviroments*, pages 85–91.

[24] Wilburt J. Labio, Yue Zhuge, Janet L. Wiener, Himanshu Gupta, Héctor García-Molina, and Jennifer Widom. The whips prototype for data warehouse creation and maintenance. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 557–559, New York, NY, USA, 1997. ACM.

[25] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA, 2002. ACM.

[26] Alon Levy, Anand Rajaraman, and Joann Ordille. Querying heterogeneous information sources using source descriptions. pages 251–262, 1996.

[27] Alon Y. Levy. The information manifold approach to data integration. *IEEE Intelligent Systems*, 13:12–16, 1998.

[28] Alon Y. Levy. Logic-based techniques in data integration. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 575–595. Kluwer Academic Publishers, Dordrecht, 2000.

[29] Alon Y. Levy and Marie-Christine Rousset. Carin: A representation language combining horn rules and description logics. In *ECAI*, pages 323–327. John Wiley and Sons, Chichester, 1996.

[30] Alon Y. Levy and Marie-Christine Rousset. Combining horn rules and description logics in {CARIN}. *Artificial Intelligence*, 104(1–2):165 – 209, 1998.

[31] Stuart J. Nelson, W. Douglas Johnston, and Betsy L. Humphreys. Relationships in medical subject headings (MeSH). In Carol A. Bean, editor, *Relationships in the Organization of Knowledge*. Kluwer Academic Publishers, The Netherlands, 2001.

[32] Yannis Papakonstantinou, Hector Garcia-Molina, and Jeffrey Ullman. Medmaker: A mediation system based on declarative specifications. pages 132–141, 1996.

[33] Yannis Papakonstantinou, Ashish Gupta, Hector Garcia-molina, and Jeffrey Ullman. A query translation scheme for rapid implementation of wrappers (extended version). pages 161–186, 1995.

[34] Mary Tork Roth and Peter M. Schwarz. Don't scrap it, wrap it! a wrapper architecture for legacy data sources. In *VLDB*, pages 266–275. Morgan Kaufmann, 1997.

[35] Dimitri Theodoratos, Spyros Ligoudistianos, and Timos K. Sellis. Designing the global data warehouse with spj views. In *CAiSE'99*, pages 180–194, 1999.

[36] Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock. Building data integration queries by demonstration. In *Proceedings of the International Conference on Intelligent User Interface*, January 2007.

[37] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Lewis Y. Geer, Yuri Kapustin, Oleg Khovayko, David L, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Vadim Miller, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Re Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, pages 13–21, 2008.

[38] Gio Wiederhold. Mediators in the architecture of future information systems. *Computer*, 25(3):38–49, March 1992.

[39] J.L. Wiener, H. Gupta, W.J. Labio, Y. Zhuge, H. Garcia-Molina, and J. Widom. A System Prototype for Warehouse View Maintenance. In *ACM Workshop on Materialized Views: Techniques and Applications*, pages 26–33, June 1996.

[40] Gang Zhou, Richard Hull, Roger King, and Jean-Claude Franchitti. Data integration and warehousing using h2o. *IEEE Data Eng. Bull.*, 18(2):29–40, 1995.

Appendix A

Data sources' DTD Files

A.1 EGQuery DTD File

```
<<!--
        This is the Current DTD for Entrez eGSearch
        $Id: egquery.dtd 39250 2004-05-03 16:19:48Z yasmax $
-->
<!-- ============================================================ -->

<!ELEMENT       DbName          (#PCDATA)>      <!-- .+ -->
<!ELEMENT       MenuName        (#PCDATA)>      <!-- .+ -->
<!ELEMENT       Count           (#PCDATA)>      <!-- \d+ -->
<!ELEMENT       Status          (#PCDATA)>      <!-- .+ -->
<!ELEMENT       Term            (#PCDATA)>      <!-- .+ -->

<!ELEMENT       ResultItem      (
                                    DbName,
                                    MenuName,
                                    Count,
```

```
                                    Status
                                )>
<!ELEMENT       eGQueryResult   (ResultItem+)>

<!ELEMENT       Result          (Term, eGQueryResult)>
```

A.2 ESearch DTD File

```
<!--
                This is the Current DTD for Entrez eSearch
$Id: eSearch_020511.dtd 85163 2006-06-28 17:35:21Z olegh $
-->
<!-- ============================================================= -->

<!ELEMENT       Count           (#PCDATA)><!-- \d+ -->
<!ELEMENT       RetMax          (#PCDATA)><!-- \d+ -->
<!ELEMENT       RetStart        (#PCDATA)><!-- \d+ -->
<!ELEMENT       Id              (#PCDATA)><!-- \d+ -->

<!ELEMENT       From            (#PCDATA)><!-- .+ -->
<!ELEMENT       To              (#PCDATA)><!-- .+ -->
<!ELEMENT       Term            (#PCDATA)><!-- .+ -->

<!ELEMENT       Field           (#PCDATA)><!-- .+ -->

<!ELEMENT        QueryKey       (#PCDATA)><!-- \d+ -->
```

```
<!ELEMENT        WebEnv          (#PCDATA)><!-- \S+ -->

<!ELEMENT       Explode          (#PCDATA)><!-- (Y|N) -->
<!ELEMENT       OP               (#PCDATA)><!-- (AND|OR|NOT|RANGE|GROUP) -->
<!ELEMENT       IdList           (Id*)>

<!ELEMENT       Translation      (From, To)>
<!ELEMENT       TranslationSet   (Translation*)>

<!ELEMENT       TermSet (Term, Field, Count, Explode)>
<!ELEMENT       TranslationStack        ((TermSet|OP)*)>

<!-- Error message tags  -->

<!ELEMENT        ERROR                  (#PCDATA)><!-- .+ -->

<!ELEMENT        OutputMessage     (#PCDATA)><!-- .+ -->
<!ELEMENT        QuotedPhraseNotFound   (#PCDATA)><!-- .+ -->
<!ELEMENT        PhraseIgnored          (#PCDATA)><!-- .+ -->
<!ELEMENT        FieldNotFound          (#PCDATA)><!-- .+ -->
<!ELEMENT        PhraseNotFound         (#PCDATA)><!-- .+ -->
<!ELEMENT        QueryTranslation       (#PCDATA)><!-- .+ -->

<!ELEMENT        ErrorList      (PhraseNotFound*,FieldNotFound*)>
<!ELEMENT        WarningList    (PhraseIgnored*,
QuotedPhraseNotFound*,
OutputMessage*)>
<!-- Response tags -->
```

```
<!ELEMENT       eSearchResult   (((
                                Count,
                                    (RetMax,
                                    RetStart,
                                    QueryKey?,
                                    WebEnv?,
                                    IdList,
                                    TranslationSet,
                                    TranslationStack?,
                                    QueryTranslation
                                    )?
                                ) | ERROR),
ErrorList?,
WarningList?
)>
```

A.3 EFetch DTD File for PubMed

```
  <!-- NLM MedlineCitationSet DTD

     This is the DTD which NLM has written for Internal and External Use.
     May 1, 2013

**THIS IS THE CURRENT DTD FOR 2013 CURRENTLY IN USE.
```

```
     SEE http://www.nlm.nih.gov/databases/dtd/nlmmedlinecitationset_140101.dtd
     FOR THE FORTHCOMING NLMMedlineCitationSet DTD DATED JANUARY 1, 2014 FOR
     FUTURE USE.**

    NOTE:  The use of "Medline" in a DTD or element name does not mean the record
    represents a citation from a Medline-selected journal.  When the NLM DTDs and
    XML elements were first created, MEDLINE records were the only data exported.
    Now NLM exports citations other than MEDLINE records using these tools. To
    minimize unnecessary disruption to users of the data and tools, NLM has
    retained the original DTD and element names (e.g., NLMMedlineCitationSet,
    MedlineTA, MedlineJournalInfo)).

    NOTE:  StartPage and EndPage in Pagination element are not currently used;
           are reserved for future use.


       * = 0 or more occurrences (optional element, repeatable)
       ? = 0 or 1 occurrences (optional element, at most 1)
       + = 1 or more occurrences (required element, repeatable)
       | = choice, one or the other but not both
       no symbol = required element
-->
<!-- ============================================================= -->
<!--             Revision Notes Section

  The following changes were made:

  a. Changed nlmmedlinecitationset_130101.dtd to nlmmedlinecitationset_130501.dtd.
  b. Added new AbstractText NlmCategory attribute valid value UNASSIGNED.
```

```
  c. Added new Article PubModel attribute valid value Electronic-eCollection.

See http://www.nlm.nih.gov/databases/dtd/history_dtd_nlmmedline.html for historic
Revision Notes for previous versions of NLMMedlineCitationSet DTD.
-->
<!-- ================================================================== -->
<!-- ================================================================== -->
<!ELEMENT MedlineCitationSet (MedlineCitation*, DeleteCitation?)>
<!ELEMENT MedlineCitation (PMID, DateCreated, DateCompleted?, DateRevised?,
                Article, MedlineJournalInfo, ChemicalList?,SupplMeshList?,
                CitationSubset*, CommentsCorrectionsList?, GeneSymbolList?,
                MeshHeadingList?,NumberOfReferences?, PersonalNameSubjectList?,
                OtherID*, OtherAbstract*, KeywordList*, SpaceFlightMission*,
                InvestigatorList?, GeneralNote*)>
<!ATTLIST MedlineCitation
                Owner  (NLM | NASA | PIP | KIE | HSR | HMD | NOTNLM) "NLM"
Status (Completed | In-Process | PubMed-not-MEDLINE |
                        In-Data-Review | Publisher | MEDLINE |
                        OLDMEDLINE) #REQUIRED
                VersionID CDATA #IMPLIED
                VersionDate CDATA #IMPLIED>
<!ELEMENT Abstract (AbstractText+,CopyrightInformation?)>
<!ELEMENT AbstractText (#PCDATA)>
<!ATTLIST       AbstractText
                Label CDATA #IMPLIED
                NlmCategory (UNLABELLED | BACKGROUND | OBJECTIVE | METHODS |
                             RESULTS | CONCLUSIONS | UNASSIGNED) #IMPLIED>
<!ELEMENT AccessionNumber (#PCDATA)>
<!ELEMENT AccessionNumberList (AccessionNumber+)>
```

```
<!ELEMENT Acronym (#PCDATA)>
<!ELEMENT Affiliation (#PCDATA)>
<!ELEMENT Agency (#PCDATA)>
<!ELEMENT Article (Journal,ArticleTitle,((Pagination, ELocationID*) |
                        ELocationID+),Abstract?, Affiliation?, AuthorList?,
                        Language+, DataBankList?, GrantList?,PublicationTypeList,
                        VernacularTitle?, ArticleDate*)>
<!ATTLIST Article
                PubModel (Print | Print-Electronic | Electronic |
                          Electronic-Print | Electronic-eCollection) #REQUIRED>
<!ELEMENT ArticleDate (Year,Month,Day)>
<!ATTLIST ArticleDate DateType CDATA  #FIXED "Electronic">
<!ELEMENT ArticleTitle (#PCDATA)>
<!ELEMENT Author (((LastName, ForeName?, Initials?, Suffix?) |
                         CollectiveName),Identifier*)>
<!ATTLIST Author ValidYN (Y | N) "Y">
<!ELEMENT AuthorList (Author+)>
<!ATTLIST AuthorList CompleteYN (Y | N) "Y">
<!ELEMENT Chemical (RegistryNumber,NameOfSubstance)>
<!ELEMENT ChemicalList (Chemical+)>
<!ELEMENT CitationSubset (#PCDATA)>
<!ELEMENT CollectiveName (#PCDATA)>
<!ELEMENT CommentsCorrections (RefSource,PMID?,Note?)>
<!ATTLIST       CommentsCorrections
                RefType (CommentOn | CommentIn | ErratumIn | ErratumFor |
                PartialRetractionIn | PartialRetractionOf | RepublishedFrom |
                RepublishedIn | RetractionOf | RetractionIn | UpdateIn |
                UpdateOf | SummaryForPatientsIn | OriginalReportIn |
                ReprintOf | ReprintIn | Cites) #REQUIRED >
```

```
<!ELEMENT CommentsCorrectionsList (CommentsCorrections+)>
<!ELEMENT CopyrightInformation (#PCDATA)>
<!ELEMENT Country (#PCDATA)>
<!ELEMENT DataBank (DataBankName, AccessionNumberList?)>
<!ELEMENT DataBankList (DataBank+)>
<!ATTLIST DataBankList CompleteYN (Y | N) "Y">
<!ELEMENT DataBankName (#PCDATA)>
<!ELEMENT DateCompleted (Year,Month,Day)>
<!ELEMENT DateCreated (Year,Month,Day)>
<!ELEMENT DateRevised (Year,Month,Day)>
<!ELEMENT Day (#PCDATA)>
<!ELEMENT DescriptorName (#PCDATA)>
<!ATTLIST DescriptorName
                MajorTopicYN (Y | N) "N"
                Type (Geographic) #IMPLIED>
<!ELEMENT ELocationID (#PCDATA)>
<!ATTLIST ELocationID EIdType (doi | pii) #REQUIRED
                ValidYN  (Y | N) "Y">
<!ELEMENT       EndPage (#PCDATA)>
<!ELEMENT ForeName (#PCDATA)>
<!ELEMENT GeneSymbol (#PCDATA)>
<!ELEMENT GeneSymbolList (GeneSymbol+)>
<!ELEMENT GeneralNote (#PCDATA)>
<!ATTLIST GeneralNote Owner (NLM | NASA | PIP | KIE | HSR | HMD) "NLM">
<!ELEMENT Grant (GrantID?, Acronym?, Agency, Country)>
<!ELEMENT GrantID (#PCDATA)>
<!ELEMENT GrantList (Grant+)>
<!ATTLIST GrantList CompleteYN (Y | N) "Y">
<!ELEMENT       Identifier (#PCDATA)>
```

```
<!ATTLIST       Identifier
                Source CDATA #REQUIRED >
<!ELEMENT ISOAbbreviation (#PCDATA)>
<!ELEMENT ISSN (#PCDATA)>
<!ATTLIST ISSN IssnType  (Electronic | Print) #REQUIRED>
<!ELEMENT       ISSNLinking (#PCDATA)>
<!ELEMENT Initials (#PCDATA)>
<!ELEMENT Investigator (LastName,ForeName?, Initials?,Suffix?,Identifier*,
            Affiliation?)>
<!ATTLIST Investigator ValidYN (Y | N) "Y">
<!ELEMENT InvestigatorList (Investigator+)>
<!ELEMENT Issue (#PCDATA)>
<!ELEMENT Journal (ISSN?, JournalIssue, Title?, ISOAbbreviation?)>
<!ELEMENT JournalIssue (Volume?, Issue?, PubDate)>
<!ATTLIST JournalIssue CitedMedium (Internet | Print) #REQUIRED>
<!ELEMENT Keyword (#PCDATA)>
<!ATTLIST Keyword MajorTopicYN (Y | N) "N">
<!ELEMENT KeywordList (Keyword+)>
<!ATTLIST KeywordList Owner (NLM | NLM-AUTO | NASA | PIP | KIE | NOTNLM | HHS) "NLM">
<!ELEMENT Language (#PCDATA)>
<!ELEMENT LastName (#PCDATA)>
<!ELEMENT MedlineDate (#PCDATA)>
<!ELEMENT MedlineJournalInfo (Country?, MedlineTA, NlmUniqueID?,ISSNLinking?)>
<!ELEMENT       MedlinePgn (#PCDATA)>
<!ELEMENT MedlineTA (#PCDATA)>
<!ELEMENT MeshHeading (DescriptorName, QualifierName*)>
<!ELEMENT MeshHeadingList (MeshHeading+)>
<!ELEMENT Month (#PCDATA)>
```

```
<!ELEMENT NameOfSubstance (#PCDATA)>
<!ELEMENT NlmUniqueID (#PCDATA)>
<!ELEMENT Note (#PCDATA)>
<!ELEMENT NumberOfReferences (#PCDATA)>
<!ELEMENT OtherAbstract (AbstractText+,CopyrightInformation?)>
<!ATTLIST OtherAbstract Type (AAMC | AIDS | KIE | PIP |
                                    NASA | Publisher) #REQUIRED
                               Language CDATA "eng">
<!ELEMENT OtherID (#PCDATA)>
<!ATTLIST OtherID Source (NASA | KIE | PIP | POP | ARPL | CPC |
                                IND | CPFH | CLML | NRCBL | NLM) #REQUIRED>
<!ELEMENT PMID (#PCDATA)>
<!ATTLIST       PMID Version CDATA #REQUIRED>
<!ELEMENT Pagination ((StartPage, EndPage?, MedlinePgn?) | MedlinePgn)>
<!ELEMENT PersonalNameSubject (LastName,ForeName?, Initials?,Suffix?)>
<!ELEMENT PersonalNameSubjectList (PersonalNameSubject+)>
<!ELEMENT PubDate ((Year, ((Month, Day?) | Season)?) | MedlineDate)>
<!ELEMENT PublicationType (#PCDATA)>
<!ELEMENT PublicationTypeList (PublicationType+)>
<!ELEMENT QualifierName (#PCDATA)>
<!ATTLIST QualifierName MajorTopicYN (Y | N) "N">
<!ELEMENT RefSource (#PCDATA)>
<!ELEMENT RegistryNumber (#PCDATA)>
<!ELEMENT Season (#PCDATA)>
<!ELEMENT SpaceFlightMission (#PCDATA)>
<!ELEMENT       StartPage (#PCDATA)>
<!ELEMENT Suffix (#PCDATA)>
<!ELEMENT       SupplMeshList (SupplMeshName+)>
<!ELEMENT       SupplMeshName (#PCDATA)>
```

```
<!ATTLIST       SupplMeshName Type (Disease | Protocol) #REQUIRED>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT VernacularTitle (#PCDATA)>
<!ELEMENT Volume (#PCDATA)>
<!ELEMENT Year (#PCDATA)>
<!ELEMENT DeleteCitation (PMID+)>
```

A.4 MedlinePlus DTD File

```
<!--
    Description:

    This DTD defines the health topics in MedlinePlus.
    ============================================================

-->

<!ELEMENT health-topics (health-topic)+>
<!ATTLIST health-topics
  date-generated CDATA #REQUIRED
  total CDATA #REQUIRED>

<!ELEMENT health-topic (also-called*,full-summary,group+,language-mapped-topic?,
                        mesh-heading*,other-language*,primary-institute?,
                        related-topic*,see-reference*,site+)>
<!ATTLIST health-topic
```

```
  id CDATA #REQUIRED
  date-created CDATA #REQUIRED
  language (English | Spanish) #REQUIRED
  title CDATA #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT full-summary (#PCDATA)>

<!ELEMENT group (#PCDATA)>
<!ATTLIST group
  id CDATA #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT language-mapped-topic (#PCDATA)>
<!ATTLIST language-mapped-topic
  id CDATA #REQUIRED
  language (English | Spanish) #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT mesh-heading (descriptor, qualifier*)>

<!ELEMENT other-language (#PCDATA)>
<!ATTLIST other-language
  vernacular-name CDATA #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT primary-institute (#PCDATA)>
<!ATTLIST primary-institute
  url CDATA #REQUIRED>
```

```
<!ELEMENT related-topic (#PCDATA)>
<!ATTLIST related-topic
  id CDATA #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT see-reference (#PCDATA)>

<!ELEMENT site (information-category+,organization*,standard-description*)>
<!ATTLIST site
  language-mapped-url CDATA #IMPLIED
  title CDATA #REQUIRED
  url CDATA #REQUIRED>

<!ELEMENT also-called (#PCDATA)>

<!ELEMENT descriptor (#PCDATA)>
<!ATTLIST descriptor
  id CDATA #REQUIRED>

<!ELEMENT qualifier (#PCDATA)>
<!ATTLIST qualifier
  id CDATA #REQUIRED>

<!ELEMENT information-category (#PCDATA)>

<!ELEMENT organization (#PCDATA)>

<!ELEMENT standard-description (#PCDATA)>
```