Automatic Term Indexing in Medical Text Corpora and its Applications to Consumer Health Information Systems

Angelos Hliaoutakis

December 3, 2009

Contents

List of Figures **iii**

Abstract **v**

Acknowledgements **vii**

1 Introduction **1**

2 Background and Related Work **6**

2.1 Medical Information Systems 6

2.2 Data Resources 9

2.2.1 MedLine 9

2.2.2 OHSUMED filtering track collection 10

2.2.3 PubMed Central Database 11

2.2.4 Unified Medical Language System (UMLS) 12

2.3 Term Extraction 18

2.4 Algorithmic Resources 20

2.4.1 The MMTx Aprroach 20

2.4.2 The C/NC–Value Method 24

3 Automatic Term Extraction in Medical Document Collections: The (AMTE$_X$) method **27**

3.1 The AMTE$_X$ Algorithm 30

3.2 Refining the AMTE$_X$Method 33

4 MedHealth: A Medical Information System for Consumer and Expert Users 36

4.1 A Dictionary of Medical Terms for Consumer and Expert users 36

4.2 Document Classification 39

4.2.1 Document Classification by user profile 39

4.2.2 Document Classification by Topic 40

4.3 System Architecture 43

5 Experiments and Evaluation 47

5.1 Experimental Setup 47

5.2 Evaluation of AMTE$_X$ 48

5.2.1 MMTx vs AMTE$_X$ Method 49

5.3 Evaluation of MedHealth 57

5.3.1 Document Indexing evaluation for Consumer and Expert Users 57

5.3.2 Information Retrieval evaluation for Consumer and Expert Users 60

5.4 MedHealth 62

6 Conclusions 66

Bibliography 68

Appendix 72

A 72

A.1 Lucene 72

A.2 PMC Full Document Retrieval Experiment Queries 72

A.3 MeSH DTD File 73

A.4 Semantic Network Categories 77

List of Figures

2.1 MedLine document structure 10

2.2 Location of term “Acids” in MeSH taxonomy 14

2.3 Semantic Network – Metathesaurus Structure 17

3.1 Term expansion using MeSH. 29

3.2 AMTE$_X$ algorithm. 30

3.3 AMTE$_X$ algorithm example. 32

4.1 Document’s Expert-Consumer Categorization 37

4.2 Document Categorization by Topic 41

4.3 Architecture of the System 45

5.1 AMTE$_X$ with/without single-word extraction and $T_{FinalList}$ in PMC dataset. 49

5.2 Precision/Recall of AMTE$_X$ vs MMTx on OHSUMED dataset retrieval task 54

5.3 Precision/Recall of AMTE$_X$ vs MMTx on PMC dataset retrieval task 56

5.4 Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset consumer indexing task 58

5.5 Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset expert indexing task 59

5.6 Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset consumer retrieval task 60

5.7 Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset expert retrieval task . . . 61

5.8 Login on MedHealth . . . 62

5.9 Register on MedHealth . . . 63

5.10 Browsing MedHealth . . . 63

5.11 Searching criteria on MedHealth . . . 64

5.12 Searching in MedHealth . . . 65

Abstract

A large amount of medical information is currently becoming from many sources, such us journals, books, databases and lately the World Wide Web (WWW). Users of the medical domain can be either health care professionals (experts) or consumers (novice users). Consumers searching for medical information usually issue a natural language query and need to retrieve medical documents on a topic of interest that is easy to read and comprehend (e.g. medical information that do not contain complex medical terminology). On the other hand, experts usually do more specialized searches (involving complex terminology) looking for state-of-the-art documents to fulfil their information needs. This thesis deals with issues related to the design and implementation of medical information systems to fulfill the need of both types of users. Medical information is acquired by Web sources or by data repositories available (e.g. Medline). Acquired medical documents from authoritative (validated) sources are analyzed (though text analysis) and are automatically indexed by author, type, content and association (e.g. link) information. Indexing is based-upon the idea of keeping the most characteristic medical terms of a document. This information is usually extracted by term extraction. Term extraction relates to extracting the most characteristic or important terms (words or phrases) from a document. This information allows for faster and better understanding of the contents of a document collection without first browsing through the contents of its documents. In this thesis is presented AMTE$_X$ an automatic term extraction method, specifically designed for the automatic indexing of documents in large medical collections such as MedLine, the premier bibliographic database of the U.S. National Library of Medicine (NLM). AMTE$_X$ combines MeSH, the terminological thesaurus resource of NLM, with a well-established method for extraction of domain terms, the C/NC-value method. The performance evaluation of AMTE$_X$ in the indexing as well as in the retrieval task is measured against the current state-of-the-art method, the MMTx method which is suggested by the National Library of Medicine (U.S. NLM).

By associating information extracted using AMTE$_X$ with well established lexical resources such WordNet, NLM's Semantic Network and MeSH, medical information (documents) is furthered classified, both, by user profile (i.e., for consumer and expert users) and by topic. Document indexing is implemented based on this categorization. Building upon this categorization, a medical information system capable of serving both types of users (experts or consumer users) is finally constructed. Evaluation results of all methods implemented, are presented and discussed.

Acknowledgements

This research was supported by the Intelligent Systems Laboratory of the Technical University of Crete. I would like to thank Prof. Petrakis for the advice, encouragement and support he provided to me in supervising this thesis. I would also like to thank Prof. Millios, for his recommendations. Finally, I feel grateful to Dr.Zervanou Kelly and Rena Peraki, for their collaboration and valuable comments.

Chapter 1

Introduction

New technological developments in communications have not only increased our ability to disseminate information in electronic form, but also the amount of the communicated information. The availability of large medical collections, such as MedLine[1] (Medical Literature Analysis and Retrieval System Online), poses new challenges to information and knowledge management. MedLine constitutes the primary medical repository of the U.S. National Library of Medicine, including (today) over 16 million computer-readable records and is expanding rapidly. It is a rich resource of medical, biological and biomedical information, requiring efficient management and retrieval.

Typically, medical information systems such as MedLine, are designed to serve health care professional users (expert users in general such as clinical doctors, medical researchers). Typically, expert users are familiar with the type and content of the medical resources (such as the NLM dictionaries and databases) they are using and use medical terminology for their searches. However, the spread and availability of medical information on the web have made this information available to consumer (i.e naive) users as well. Unlike expert users, consumers are usually unfamiliar with the content and type of specialized medical resources, and typically use the Web for their searches. They are often uncertain as to the exact type of information they are looking, they do simple searches using natural language (rather than domain specific) terms. A medical information system targeted for both types of user must be capable

[1] http://www.nlm.nih.gov/databases/databases_MedLine.html

of providing dedicated, domain specific or, simple, easily comprehensible answers to expert and consumer users accordingly. An almost orthogonal issue is speed of search. Indexing might not only increase the speed of access to the huge amounts of medical information, but also make this information usable and easily accessible indexing medical documents by subject (topic of interest). Without an indexing process, the search engine of a medical information system would scan every document in the corpus, which would require considerable time and computing power.

MedLine documents are currently indexed by human experts, based on a controlled list of indexing terms, deriving from a subset of the UMLS[2] (Unified Medical Language System) Metathesaurus, the MeSH[3] (Medical Subject Headings) thesaurus. The automatic mapping of biomedical documents to UMLS term concepts has been undertaken by U.S. National Library of Medicine with the development of MMTx[4] (MetaMap Transfer tool). MMTx was originally developed to improve retrieval of bibliographic material, such as MedLine citations [9]. Its applications also include semi-automatic and fully automatic indexing, hierarchical indexing and text mining for various medical and biological concept and relation extraction [9].

The limitations of MMTx in term extraction and in the UMLS Metathesaurus mapping have been analyzed in detail in a pilot study by Divita et al. [20]. The experiments with the MMTx application on MedLine documents have shown that the MMTx output suffers, not only in recall (as noted by [20]), failing to extract all domain terms, but it also over-generates by producing general terms, which diffuse the document concept leading to inaccurate retrieval of MedLine documents. The latter reflects an inherent limitation of MMTx, which was not designed by default to focus on MeSH terms, whereupon MedLine indexing has been based. Additionally, the variant generation process of MMTx is found to account for the over generation

[2] http://www.nlm.nih.gov/research/umls
[3] http://www.nlm.nih.gov/mesh
[4] http://mmtx.nlm.nih.gov

problem for retrieval purposes.

In the first part of this work we investigate approaches (including AMTEx and MMTx) for indexing and retrieval in medical document collections. In this study, MMTx is briefly reviewed and an alternative method, the Automatic MeSH Term Extraction method (AMTE$_X$) is proposed. AMTE$_X$ aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC-value method [21]. Additionally, AMTE$_X$ aims at improving indexing and retrieval of medical documents, based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS Metathesaurus mapping of MMTx. The performance evaluation of two AMTEx configurations is measured against the current state-of-the-art, the MetaMap Transfer (MMTx) method in four experiments, using two types of corpora: a subset of MEDLINE (PMC) full document corpus and a subset of MEDLINE (OHSUMED) abstracts, for each of the indexing and retrieval tasks respectively. The experimental results demonstrate that AMTEx performs better in indexing in 20-50% of the processing time compared to MMTx, while for the retrieval task, AMTEx performs better in the full text (PMC) corpus.

In the second part of the work we show how indexing methods (such as MMTx and AMTEx) can be use for filtering medical information for targeted audiences such as experts and naive users. An obvious application of this filtering operation will be retrieval on medical information by user profile. Corresponding to individual background and interests, different users prefer to select different medical documents. Existing approaches for consumer/expert classification, such as machine learning [47], have reported accuracy around 78%. Others try to develop controlled consumer vocabularies by assessing terms from nursing informatics. Nurses terminology corresponds to that of naive users in this study as opposed to medical experts or doctors terminology. [48].

In this work, we propose a classification method for medical documents based on the level of specificity of the terms used in the description of medical information they contain. Given that medical information is typically described by terms belonging to a medical dictionary (such as MeSH), the distinction is based on the classification of the dictionary terms to those comprehendible by naive users and to more involved terms typically used by experts (eg. medical doctors, practitioners etc). This distinction is automatic, and is made possible with the aid of WordNet, a thesaurus for natural language term of the English language, and is based on the observation that up to 30% of the terms participating in MeSH vocabulary are general terms (terms that can be found in Wordnet) and the remainder 70% are medical terms (more specific UMLS terms that do not belong to Wordnet). The performance of the method is assessed using MedLine and based on the relevance assessments provided by naive users and experts. More specifically, the performance of the proposed term classification method is measured by the accuracy of retrievals in experiments conducted using two sets of queries addressing experts and naive users respectively. The experimental results demonstrated that retrievals using term classification succeeded in returning the proper document type (expert or consumer) to its corresponding users issuing the queries (experts and naive users respectively) while maintaining up 75% precision and up to 50% recall in indexing experiments.

Related work and the resources used in this thesis are discussed in Chapter 2. These include Medline, the OHSUMED data-set of TREC-9 filtering track collections, Pubmed Central (PMC) database which contains a free digital archive of biomedical and life sciences journal literature, the MeSH, UMLS Metathesaurus and the UMLS Semantic network. Related work in the field of term extraction and, in particular, approaches to the extraction of medical terminology for indexing purposes are presented as well. Then, algorithmic resources such as MMTx, and the C/NC-value method to term extraction are discussed.

In Chapter 3 we present the AMTE$_X$ approach and a case study on automatic consumer-expert medical term classification in Chapter 4. Our proposed document classification method categorize documents by user profile based on the probability value to contain terms of each type (consumer or expert terms). The application of above ideas into MedHealth, a document information system for consumer and naive users, is discussed.

Finally, Chapter 5 presents the experimental results followed by conclusions and issues for future research in Chapter 6.

Chapter 2

Background and Related Work

An overview of medical information systems and of medical data sources used in this work, are surveyed in this chapter. An introduction to term extraction follows as it forms the basis for the implementation of a medical information system.

2.1 Medical Information Systems

The amount of health data accessible on the Web is increasing and Internet has become a major source of health information. Many medical information systems such as search engines, portals, meta search engines, digital libraries etc are currently becoming available the most important of them being:

Medscape [1]is a free Web site for health professionals and interested consumers. Practice-oriented information is peer-reviewed and edited by thought-leaders in AIDS, infectious diseases, urology, and surgery. Medical experts are interested in searching bibliographic databases e.g. MEDLINE through **PubMed**. *MEDLINE*(see section 2.2.1) is a database of over 15 million medical and scientific articles (most of them in English) indexed from the sixties to date, created and maintained by the U.S. National Library of Medicine (NLM). PubMed [2] is developed and maintained by the National Center for Biotechnology Information (NCBI) at NLM. It provides access to MEDLINE and to articles in selected life sciences journals not included in

[1] http://www.medscape.com/home

[2] http://www.ncbi.nlm.nih.gov/PubMed/

MEDLINE through an easy to use free Internet site. MEDLINE indexers describe the content of biomedical articles by assigning to each one, a number (typically 10 to 12 per article) of MeSH terms (see section 2.2.4). PubMed uses them for retrieval and the search strategy is enhanced (e.g.the query “bad breath” is mapped automatically to the MeSH term “halitosis”). Within PubMed, Consumer Health link leads to **MedlinePlus** [3] which is intended to be used mostly by consumers and also provides information that is authoritative and up to date. MedlinePlus provides quality medical information for health professionals and consumers, from the US National Library of Medicine. It includes an extensive Health Topics section, as well as dictionaries and a medical encyclopedia. The **National electronic Library for Health** [4](NeLH) provides health care professionals with knowledge and know-how to support health care related decisions.

Other medical information systems that are used mainly by consumers include, **MedicineNet.com** [5], which provides authoritative medical information only for consumers. These are articles are written by a network of health professionals. Health and medical information can also be provided by portals, such as **WebMD** [6], **MEDNETS** [7], **Healthline** [8] etc, by metasearch engines such as **OmniMedicalSearch.com** [9] that also offers a special link for the experts.

Also there exist medical search engines used by both experts and consumers. Some of them are listed below: **MedHunt** [10], developed and maintained by the Health On the Net Foundation (HON - a not-for-profit organization that aims to provide access to reliable sources of online medical information). MedHunt retrieves medical information either from HONs accredited sites or from medical pages crawled

[3] http://medlineplus.gov

[4] http://www.library.nhs.uk/Default.aspx

[5] http://www.medicinenet.com

[6] http://www.webmd.com

[7] http://www.mednets.com

[8] http://www.healthline.com

[9] http://www.omnimedicalsearch.com

[10] http://www.hon.ch/MedHunt

from the Web. **HON - Health On the Net Foundation** [11] has become one of most respected not-for-profit portals to medical information on the Internet. HON co-operates closely with the University Hospitals of Geneva and the Swiss Institute of Bioinformatics. Besides MedHunt© there are other widely-used medical search tools, the HONselect©, and the HON Code of Conduct (HONcode©) for the provision of authoritative, trustworthy Web-based medical information.

WRAPIN [12] (Worldwide online Reliable Advice to Patients and Individuals), uses medical trustworthy sources (NLMs PubMed, HONs MedHunt, U.S. Food and Drug Administration (FDA)), supports different types of query from a few keywords to entire web pages (specified by their URL). It maps both query and documents to MeSH terms(the HONMeSHMapper module is used) subsequently used for indexing and retrieval. Finally, **BioMedNet** [13] BioMedNet is an Internet-based club for researchers, clinicians, and students in Biology and Medicine providing a wide range of information services, accessed through an integrated software package. BioMedNet provides a Collaborative Work Environment where members can hold real-time discussions while sharing documents from the The BioMedNet Library - an extensive full-text library of journals, books, and databases.

The work referred to above share the same interests with us, that is granting access to medical information according to users profile (i.e., consumer and experts in our case). However, they rely solely on the manual categorization of information, a solution which requires intervention by human experts and therefore is slow and does not scale up for large document collections. This is exactly the problem the present work is dealing with: MedHealth supports automatic categorization, indexing and retrieval of medical information as targeted to consumer and expert users.

[11] http://www.hon.ch/
[12] http://www.wrapin.org/
[13] http://www.bmn.com/

2.2 Data Resources

This thesis is about medical information extraction and term classification, which is depended to data resources like *MedLine*, a large medical document collection that is continuously updated, and the *UMLS* Knowledge Sources such as *MeSH*, a subset of the *UMLS Metathesaurs* and the *Semantic Network*. Experimental results where based on two main medical document collections, the *OHSUMED* collection and the *PubMed Central* database.

2.2.1 MedLine

MedLine database is a collection of biomedical articles. It consists of abstract of medical publications together with metadata, that is information on the organization of the data, the various data domains, and the relations between them. Publications in the MedLine database are manually indexed by NLM using MeSH terms, with typically 10-12 descriptors assigned to each publication by human experts. Hence, the MeSH annotation defines for each publication a highly descriptive set of features. Over 16 million publications (in MedLine 2008) that contain abstracts are currently indexed and used in the retrieval system prototype MedSearch [14]. The articles stored in MedLine have both Descriptive and Semantic Metadata. So , MedLines documents have more information than the simple article reference. Figure 2.1 shows the structure of a MedLines document.

In the developed system [7] the main fields needed to find relevant information for a query are, the **TI** field, which describes the title of a document, the **AB** field, the abstract, and the **MH** field, including the manually assigned MeSH terms related to the document. **PMID** is the unique identifier number for the Pubmed system online, **UID** is a unique number for each document in MedLine, **AU** is the authors of the document, **LA** the language of the document's publication, **PT** is the publication

[14] http://www.intelligence.tuc.gr/medsearch/

PMID	PubMed Identifier
UID	Unique Identifier
TI	The article's title
AU	The article's authors
LA	Language of publication
MH	MeSH terms related
PT	Publication Type
DA	Date of acceptance
DP	Date of publication
AB	Abstract
SO	Source of publication

Figure 2.1: MedLine document structure

type (e.g book, article, e.t.c), **DA** the date of acceptance to MedLine collection, **DP** the publication date, and **SO** the source of publication.

2.2.2 OHSUMED filtering track collection

The OHSUMED test collection is a set of 348,566 references from MedLine, the on–line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five–year period (1987–1991). The OHSUMED collection is part of the data in TREC(Text REtrieval Conference) filtering track, TREC-9(2000). This document collection does not only includes documents, but also topics (queries), and relevance judgements. The available fields are title, abstract, MeSH indexing terms (MeSH thesaurus shall be described later on), author, source, and publication type. The OHSUMED document collection was obtained by William Hersh and colleagues for the experiments described in [27] and [26].

The field definitions are:

.I sequential identifier

.U MedLine identifier (UI)

.M Human–assigned MeSH terms (MH)

.T Title (TI)

.P Publication type (PT)

.W Abstract (AB)

.A Author (AU)

.S Source (SO)

Some abstracts are truncated to 250 words and some references have no abstracts at all (titles only). There is no access to the full text of the documents.

The topic statements (queries) are provided in the standard TREC format and consist of <title> and <desc> (= description) fields only. The meaning of these fields is slightly different for each query type.

The test collection was built as part of a study assessing the use of MedLine by physicians in a clinical setting (Hersh and Hickam, above). Novice physicians using MedLine generated 106 queries. Only a subset of 63 of these queries were used in the TREC–9 filtering track. Before they searched, they were asked to provide a statement of information about their patient as well as about their information need.

2.2.3 PubMed Central Database

MedLine is a huge medical abstract document collection, but collection to medical abstract, but access to the full texts is not freely available. PubMed Central (PMC) was created to allow and encourage free access to the full–text of articles from life sciences journals. PMC [15] is the National Library of Medicine's digital archive of free full–text journal literature. Traditionally, journals deposit material in PMC on a voluntary basis. Articles may be retrieved either by browsing a table of contents for a specific journal or by searching the database.

Now, PubMed Central includes nearly 2 million articles from more than 800 Journals. Participating journals range from small new journals like Evidence–based Com-

[15] http://www.pubmedcentral.gov

plementary and Alternative Medicine [16] to standards like Proceedings of the National Academy of Sciences of the USA [17], the Journal of Clinical Investigation and the journals of the American Society for Microbiology. PubMed Central was started in 2000 as a project for the National Center for Biotechnology Information (NCBI), a center in the National Library of Medicine (NLM) [14].

2.2.4 Unified Medical Language System (UMLS)

The purpose of NLM's Unified Medical Language System (UMLS) [18] is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) 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 multi-purpose. They are not optimized for particular applications, but 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, and public health data. There are three UMLS Knowledge Sources: the *Metathesaurus*, the *Semantic Network*, and the SPECIALIST Lexicon.

UMLS Metathesaurus

MedLine and Pubmed Central make use of a thesaurus that can provide terms for indexing and retrieval purposes. UMLS contains a very large, multi–purpose and multi–lingual thesaurus concerning biomedical and health related concepts. In particular, it contains information about over 1 million biomedical concepts and 2.8

[16] http://web.pubmedcentral.nih.gov/tocrender.fcgi?action=archive&journal=241
[17] http://web.pubmedcentral.nih.gov/tocrender.fcgi?action= archive&journal=2
[18] http://www.nlm.nih.gov/research/umls/

million concept names from more than 100 controlled vocabularies and classifications (some in multiple languages) used in patient records, administrative health data, bibliographic and full–text databases and expert systems. Furthermore, all the names and meanings are enhanced with attributes and inter–term relationships.

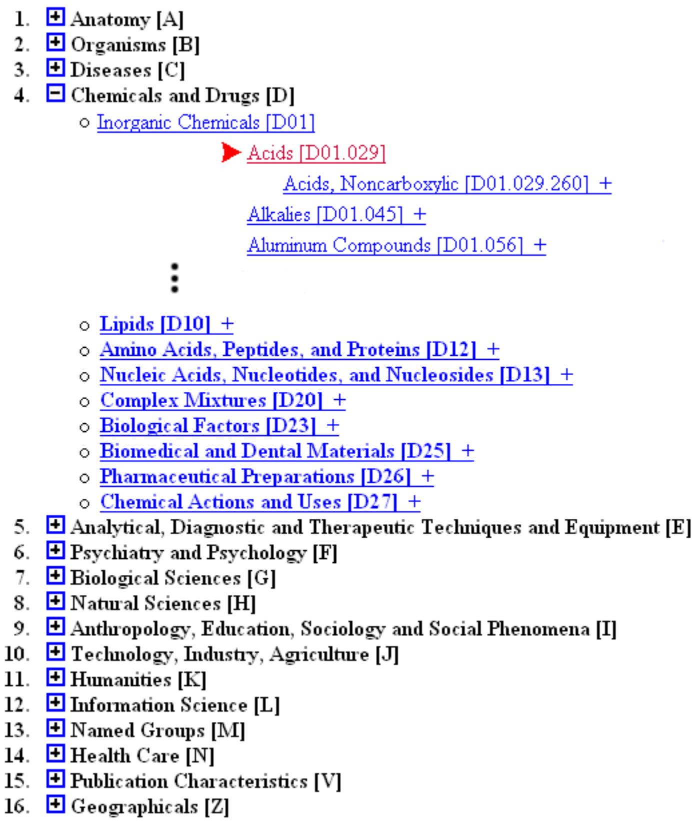
UMLS includes other meta thesaurus source vocabularies, such as Medical Subject Headings (MeSH) that is the National Library of Medicine's vocabulary thesaurus [36]. MeSH consists of sets of terms naming *descriptors* in a hierarchical structure. The Gene Ontology (GO), which is a structured network of defined terms that describe gene proteins and concerns all organisms. Another meta source is the Spatial Data Transfer Standard (SDTS), which contains an ontology used to describe the underlying conceptual model and the detailed specifications for the content, structure, and format of spatial data, their features and associated attributes. Concepts in SDTS are commonly used on topographic quadrangle maps and hydrographic charts.

MeSH

MeSH (Medical Subject Headings) [35, 40] is a taxonomic hierarchy of medical and biological terms only, suggested by the U.S National Library of Medicine (NLM) [19]. Those terms represent a subset of the UMLS metathesaurus. NLM has adopted the Extensible Markup Language (XML) [20] as the description langauge for MeSH. The MeSH vocabulary file is available in XML format. All terms in MeSH are organized in a hierarchy with most general terms (e.g "Chemicals and Drugs") higher in the taxonomy than most specific terms (e.g "Aspirin"). There are 24,767 main headings, termed descriptors, in MeSH (2008). Moreover, the structure of MeSH is a hierarchical tree, where a term can appear in different subtrees. There are 16 tree hierarchies (subtrees) in the MeSH ontology (see Figure 2.2), of ISA kind of relationship between nodes (concepts) in each subtree.

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

[20] http://www.w3.org/XML/



1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊞ Diseases [C]
4. ⊟ Chemicals and Drugs [D]
   - ○ Inorganic Chemicals [D01]
     - ➤ Acids [D01.029]
       - Acids, Noncarboxylic [D01.029.260] +
     - Alkalies [D01.045] +
     - Aluminum Compounds [D01.056] +

⋮

   - ○ Lipids [D10] +
   - ○ Amino Acids, Peptides, and Proteins [D12] +
   - ○ Nucleic Acids, Nucleotides, and Nucleosides [D13] +
   - ○ Complex Mixtures [D20] +
   - ○ Biological Factors [D23] +
   - ○ Biomedical and Dental Materials [D25] +
   - ○ Pharmaceutical Preparations [D26] +
   - ○ Chemical Actions and Uses [D27] +
5. ⊞ Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. ⊞ Psychiatry and Psychology [F]
7. ⊞ Biological Sciences [G]
8. ⊞ Natural Sciences [H]
9. ⊞ Anthropology, Education, Sociology and Social Phenomena [I]
10. ⊞ Technology, Industry, Agriculture [J]
11. ⊞ Humanities [K]
12. ⊞ Information Science [L]
13. ⊞ Named Groups [M]
14. ⊞ Health Care [N]
15. ⊞ Publication Characteristics [V]
16. ⊞ Geographicals [Z]

Figure 2.2: Location of term "Acids" in MeSH taxonomy

MeSH concepts correspond to MeSH objects which are described with terms of several properties (see Appendix A.3 in page 73), the most important of them being:

MeSH Headings (MH): These are term names or identifiers. They are used in MedLine as the indexing terms for documents. Every document in MedLine have some MeSH terms that are indexed with. A MH term belongs to a concept, and is preferred to label the meaning that the corresponding concept reflects; its use indicates the topic discussed by the document.

Entry Terms: These terms are used as pointers to the MH, there are mostly the synonym terms of the MH, naming the same concept, the MH. "Mostly" is because there is not quite a synonymy relation in those terms with the MH. In most cases it is, but there can be terms that designate the MH in an opposite way like “anions” and “cations”. They are also referred to as *quasi–synonyms*. The set of entry terms that points to a MH are the terms that represent the concept introduced by the MH. So, an admission is made in this study that all entry terms are synonyms with the MH.

MeSH Tree Number: The tree numbers indicate the positions of the terms in the MeSH taxonomy. For example D is the code name of the ”Chemical and drugs” subtree (1 of 15) and the term ”Acids” has a tree number D01.029, meaning that ”Acids” belongs to D subtree (see Figure 2.2).

MeSH Scope Note: Mainly the text descriptions of the MeSH terms. This short piece of free text provides a type of definition, in which the meaning of the MH is circumscribed.

Main Headings (descriptor records) are distinct in meaning from other Main Headings in the thesaurus (ie. their meanings do not overlap). Moreover, descriptor names reflect the broad meaning of the concepts involved. The hierarchical relationships can be intellectually accessible by users of MeSH (e.g., clinician, librarian, and indexer). An indexer is able to assign a given Main Heading to an article and a clinician can find a given Main Heading in the tree hierarchy. The relationship between entry terms and main headings is one of the most essential in the thesaurus.

UMLS Semantic Network

The UMLS Semantic Network is another UMLS Knowledge Source developed as part of the Unified Medical Language System project. The network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus and respectiverly on MeSH.

The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts. All information about specific concepts is found in the Metathesaurus. The Network provides information about the set of basic semantic types, or categories, which may be assigned to these concepts, and it defines the set of relationships that may hold between the semantic types. The Semantic Network contains 135 semantic types and 54 relationships [33].

The semantic types are the nodes in the Network, and the relationships between them are the links. There are major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. The current scope of the UMLS semantic types is quite broad, allowing for the semantic categorization of a wide range of terminology in multiple domains.

The Metathesaurus consists of terms from its source vocabularies. The meaning of each term is defined by its source, explicitly by definition or annotation; by context (its place in a hierarchy); by synonyms and other stated relationships between terms; and by its usage in description, classification, or indexing. Each Metathesaurus concept is assigned at least one semantic type (see figure 2.3).

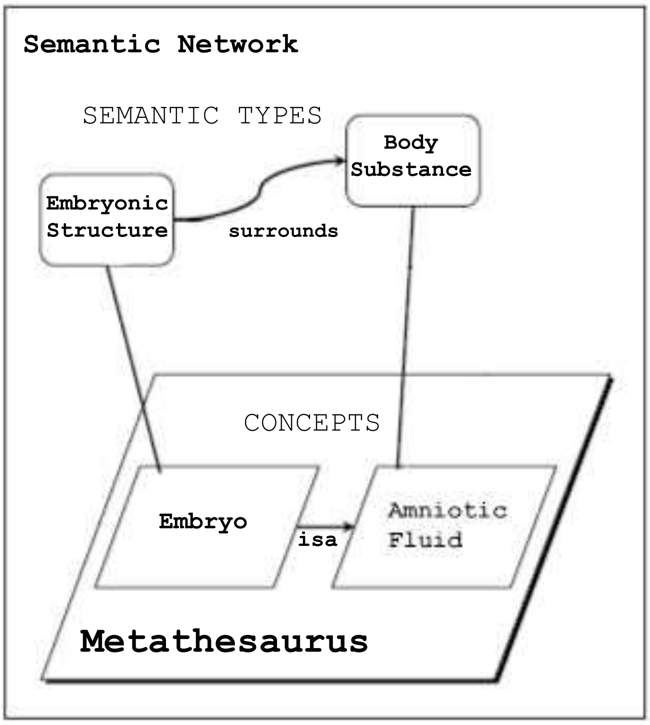


Figure 2.3: Semantic Network – Metathesaurus Structure

Results in [18] shows a 13% inconsistency in the relationships between the *Semantic Network (SN)* and the *Metathesaurus*. Inconsistency means an inaccurate/missing SN relation, or an inaccurate categorization on the SN or an inaccurate Metathesaurus relation, for example the Metathesaurus concept "Toad Licking" is represented in the SN as "Pharmacologic Substance", which is a wrong hierarchical relation. In reverse, the links that are expressed between *MeSH* terms are, with a few exceptions, reflected in the Semantic Network. That is, if two MeSH terms are linked by a certain relation, then that link is expressed in the Network as a link between the semantic types that have been assigned to those MeSH terms. For example, "Amniotic Fluid", which is a "Body Substance", is a child of "Embryo", which is an "Embryonic Structure". The labeled relationship between "Amniotic Fluid" and its parent "Embryo" is "surrounds". This is allowable, since amniotic fluid surrounds the embryo and the relation "Body Substance surrounds Embryonic Structure" is indeed represented in the Network [15].

The UMLS Semantic Network is provided in two formats: a relational table format and a unit record format, in this thesis both of them were used depending on the application.

2.3 Term Extraction

Term Extraction aims at the identification of linguistic expressions denoting specialized concepts, namely domain or scientific terms. Terms are word or multi–word expressions, which, contrary to general language words, are deliberately created within a scientific or technical linguistic community not only for concept naming purposes, but also for specialized concept distinction and classification purposes [2]. The automatic identification of terms is of particular importance in the context of information management applications, because these linguistic expressions are bound to convey the informational content load of a document. In early approaches, terms have been sought for indexing purposes, using mostly $tf \cdot idf$ counts [31]. Term extraction approaches largely rely on the identification of term formation patterns (e.g. [4, 19, 24]). Statistical techniques may also be applied to measure the degree of unithood or termhood of the candidate multi–word terms (e.g. [13]). Later and current approaches tend to follow a hybrid approach combining both statistical and linguistic techniques (e.g. [22, 32, 29]). The extraction of terms for the medical, biological and biomedical domain has greatly motivated research for both indexing, as well as knowledge extraction purposes [24, 44, 43, 45]. In the specific context of term extraction for indexing purposes, the main objective of the term extraction process is the identification of discrete content indicators, namely index terms. A traditional technique for automatic indexing has been the $tf \cdot idf$ method [31]. Although terms (domain terms [21]) may be discovered in such a process, neither all terms are useful index terms, nor

[21] At this point a distinction is needed to be made between (a) the notion of *term* which, depending on the scientific community, may refer to the terminologically acceptable notion of *domain or*

all index terms are terms. For example, a valid term appearing very frequently in a document collection is useless for the retrieval of a specific document. Moreover, query and document representations traditionally ignore multi–word and compound terms, which may perform quite efficiently split into isolated single word index terms. However, compound and multi–word terms are very common in the biomedical domain [29] and are often used in indexing medical documents. Multi–word terms carry important classificatory content information, since they comprise of modifiers denoting a specialization of the more general single–word, head term [19]. For example, the compound term *heart disease* denotes a specific type of *disease*. A recent study by Milios et al. [34] of the extraction of multi–word terms for retrieval purposes suggests that multi–word term methods may complement other methods to improve results. Currently machine learning techniques are also applied for indexing, such as the Naïve Bayes learning model implemented in the KEA (Automatic Keyphrase Extraction, [42]). Comparative experiments of $tf \cdot idf$, KEA and the C/NC–value term extraction methods by Zhang et al. [46] show that C/NC–value significantly outperforms both $tf \cdot idf$ and KEA in a narrative text classification task using the extracted terms.

Since term extraction is primarily based on surface term form patterns, it inherently suffers from two problems: ambiguity and variation. Ambiguity relates to the semantic interpretation of a given term form and it arises when this form can be interpreted in more than one way. Variation is generally defined as the alteration of the surface term form of a terminological concept. According to Jacquemin [29], variation is more specifically defined as a transformation of a controlled multi–word term and can be of three types: morphological, syntactic or semantic. Many approaches, such as [32], [29] and [24], including MMTx and the AMTE$_X$ method that will be

scientific term, as defined in the beginning of this section; and (b) the notion of *index term*, namely a key concept, word or phrase, which semantically labels and conceptually categorizes the content of a document for information management purposes, such as retrieval. In the rest of this study, the notion of term refers mainly to index terms, though in the C/NC–value approach used in the method, the design objective is *domain term* extraction, rather than indexing.

discussed further later on, attempt to resolve the problems of ambiguity and variation in terminological concepts by combining simple text normalization techniques, statistics, or more elaborate rule–based, linguistic techniques, with existing thesaurus and lexicon information. In a previous work, MedSearch was implemented [28], a retrieval system that discovers semantically similar terms in documents and queries based on the computation of semantic similar terms in different taxonomies using the SSRM statistical method [41].

2.4 Algorithmic Resources

2.4.1 The MMTx Aprroach

MMTx is is developed at the National Library of Medicine (NLM) to map biomedical text to UMLS Metathesaurus concepts. This approach uses the UMLS Metathesaurus and the SPECIALIST Lexicon as its lexicographic resources. In this section is briefly presented the UMLS knowledge sources and then an outline of the MMTx approach.

The UMLS Medical Knowledge Resources

As it was mentioned earlier, the *Unified Medical Language System (UMLS)* is a source of medical knowledge developed and maintained by the U.S. National Library of Medicine. UMLS consists of the Metathesaurus, the Semantic Network and the SPECIALIST lexicon.

UMLS Metathesaurus is a large, multi–purpose, and multi–lingual vocabulary database. It integrates about 800.000 concepts from 50 families of vocabularies. In the Metathesaurus, equivalent terms are clustered into unique concepts. Thus, the Metathesaurus on its own does not have a hierarchical structure, and it does not fulfills ontological requirements (see section 2.2.4, page 12).

Semantic Network consists of 135 semantic types categorizing the Metathesaurus concepts. The Semantic Network may be viewed as an upper level ontology of the biomedical domain. In this perspective, the Metathesaurus entities constitute the properties of the semantic network concepts (i.e. they can be inherited by concepts related by an IS–A relationship). Thus, the Semantic Network of UMLS provides a basis for an ontology of the biomedical domain (see section 2.2.4, page 15). Finally, the

SPECIALIST lexicon is intended to be a general English lexicon which includes many medical and biomedical terms. The lexicon entry for each word or term records the syntactic, morphological and orthographic information of the respective lemma.

MMTx (MetaMap Transfer)

MMTx uses the UMLS Metathesaurus and SPECIALIST lexicon knowledge resources during the term extraction process. This process maps arbitrary text to Metathesaurus term concepts and performs the following steps [9]:

1. **Parsing:** The document text is parsed, using the Xerox part-of-speech tagger and the SPECIALIST minimal commitment parser to perform a shallow syntactic analysis of the text. A simple linguistic filter of the form (Adj|Noun)$^+$Noun isolates noun phrases [8]. The SPECIALIST parser provides information on the internal syntactic structure of the noun phrase, identifying the head and modifier components of the phrase. For example, the term *"ocular complications"* is analysed as:

   ```
   [mod(ocular), head(complications)]
   ```

   where *complications* is the head, namely the term that is being modified/specialised and *ocular* is the modifier, namely the concept specialising the term *complica-*

tions.

2. **Variant Generation:** Variant generation is performed in an iterative manner. First, the multi–word term phrase is split into *generators*. A variant generator is considered any meaningful subsequence of words in the phrase. That is either a single word or a term existing in the SPECIALIST lexicon [12]. For example, the term *"liquid crystal thermography"* would be split into the generators: *"liquid crystal thermography"*, *"liquid crystal"*, *"liquid"*, *"crystal"* and *"thermography"* [8]. In the second phase, for each of the generators, all possible semantic (synonyms, acronyms and abbreviations) and derivational variants are identified using the SPECIALIST lexicon and a supplementary database of synonyms. At this stage, please note that, although the process was started of variant generation of a noun phrase, it may has derivational and semantic variants belonging to other parts-of-speech, such as verbs. All these variants are in turn used as generators and their respective variants are recomputed. Finally, inflectional and spelling variants are generated based on all word–forms found in the previous processes.

3. **Candidate Retrieval:** At this stage, the candidate set of all Metathesaurus term mappings is retrieved. The main criterion of the retrieval is that the Metathesaurus term string should contain at least one of the variants found during the variant generation process [10]. The mapping process may vary [8]. It may have:

   simple match where, for example, *intensive care unit* maps to *Intensive Care Units*;

   complex match where *intensive care medicine* maps to *Intensive Care* and *Medicine*;

   partial match – gapped where *ambulatory monitoring* maps to *Ambulatory*

Cardiac Monitoring;

normal and overmatch where *application* maps to *Job Application*, *Heat/Cold Application* and *Medical Informatics Application*.

The normal partial match is assumed as a good matching for correctness, where at least one word of either the noun phrase or the Metathesaurus string (or both) does not participate in the matching (e.g. *liquid crystal thermography* maps to *Thermography*, where the mapping does not involve *liquid crystal*).

4. **Candidate Evaluation:** The candidate set of Metathesaurus mappings is evaluated. The evaluation process computes the mapping strength between the candidate Metathesaurus string and the text string. The mapping strength weight is calculated by a linguistically principled function consisting of a weighted average of four criteria [11]:

Centrality indicates whether the Metathesaurus string involves the *head* of the text phrase and its value is 1 (yes) or 0 (no);

Variation is the distance score between the phrase and its variants (this is computed during variant generation);

Coverage denotes the length of the text phrase and the Metathesaurus candidate string participating in the match.

Cohesiveness is similar to coverage and denotes the continuous words of the text phrase and the Metathesaurus term participating in the match.

The weight for the last two criteria, coverage and cohesiveness, is doubled in the scoring function and their measures are normalised to a value between 0 and 1,000.

Summarizing, based on the above functions and abilities of the MMTx approach, the following can be observed:

- During the variant generation stage, the iterative expansion of the initial text phrase to all possible variants is quite exhaustive. MMTx extracts terms not only from terms in the original phrase, but also from their derivative terms.

- By default MMTx extracts general Metathesaurus terms not just MeSH terms.

- Term selection in based on a scoring function (for evaluating the importance of all candidate terms) using the SPECIALIST lexicon as an outside source. Moreover, the scoring function is rather arbitrarily of empirically defined making it plausible for unrelated terms to be included in the list of extracted terms.

2.4.2 The C/NC–Value Method

The C/NC–value method [22] is a hybrid method for term extraction. C/NC–value is domain–independent and combines statistical and linguistic information for the extraction of multi–word and nested terms. While MMTx is focalized in medical domain, the C/NC–value approach is a general term extraction method. In this method, the text is first tokenised and tagged by a part-of-speech tagger. Subsequently, a set of rules and linguistic filters is used to identify in text candidate term phrases. The three filters available are:

N^+N

$(A|N)^+N$

$((A|N)^+|((A|N)^*(N\ \ P)?)(A|N)^*)N$

where N is a noun, A is an adjective and P stands for a preposition. Obviously, the linguistic filters used have an impact on the precision and recall of the system. Using a rather closed filter, such as the first one, will result in increased precision and decreased recall, whereas an open filter, such as the last one will increase recall and decrease precision [21]. The current implementation of C/NC–value in this thesis uses all three linguistic filters. The generated list of candidate noun phrases is then filtered through a stoplist.

The statistical part defining the termhood of the candidate phrases aims to get more accurate terms than those obtained by the pure frequency of occurrence method, especially terms that may appear as nested within longer terms, such as the term *enzyme inhibitors* nested in *Angiotensin-converting enzyme inhibitors.* The measurement used for this estimation is C–value. C–value is defined as the relation of the cumulative frequency of occurrence of a word sequence in the text, with the frequency of occurrence of this sequence as part of larger proposed terms in the same text. Depending on whether the term is nested or not C–value is defined as:

$$C\text{-}value = \{\, l\ og_2|a|f(a), log_2|a|(f(a) - \frac{1}{P(T_a)}\sum_{b\in T_a} f(b)). \tag{2.1}$$

In the above, the first C–value measurement is for non-nested terms and the second for nested terms, where a denotes the word sequence that is proposed as a term, $|a|$ is the length of this term in words, $f(a)$ is the frequency of occurrence of this term in the corpus (both as an independent term and as a nested term within larger terms). T_a denotes the set of extracted terms that contain a and $P(T_a)$ denotes the number of these terms. The C–value algorithm produces a list of proposed terms ranked with decreasing term likelihood. The NC-value takes into account the context of each term and assigns weights to specific verbs, adjectives and nouns that appear in candidate term context. The weight factor of a context word w is higher for the respective words that tend to appear with terms and is computed as

$$weight(w) = \frac{t(w)}{n} \tag{2.2}$$

where $t(w)$ is the number of terms the word w appears with and n is the number of all terms. Finally, the NC–value is defined by

$$NC{-}value(a) = 0.8 \cdot C{-}value(a) + 0.2 \cdot CF(a) \tag{2.3}$$

Here a is the proposed term, $C{-}value(a)$ is calculated as shown in Eq.2.1, and $CF(a)$ is computed as

$$CF(a) = \sum_{w \in C_a} f_a(w) \cdot weight(w), \tag{2.4}$$

where C_a is the set of context words of term a, w is a context word in C_a, $weight(w)$ is the weight of w and $f_a(w)$ is its frequency as context word of a.

C/NC-value has been successfully tested in various domains, such as molecular biology (nuclear receptors [5]), eye pathology medical records [21], biomedical business newswire texts [45] and computer science papers [34].

Chapter 3

Automatic Term Extraction in Medical Document Collections: The (AMTE$_X$) method

Based on the study of the MMTx algorithm and resources in section 2.4.1 at page 20, we make the following observations:

- During the variant generation stage, the iterative expansion of the initial text phrase to all possible variants is quite exhaustive. MMTx extracts term variants, not only based on the terms found in the original text phrase, but also from their variant terms. This is due to an obvious attempt to increase recall of Metathesaurus mappings, a known limitation of MMTx as discussed in [20]. However, this process also results in term over-generation and increased term ambiguity, which diffuse the original term concept, leading to inaccurate indexing.
- MMTx extracts general Metathesaurus terms, not MeSH terms. Although MMTx was originally developed to improve retrieval of bibliographic material, such as MEDLINE citations [9], MMTx mappings were not based on the MeSH Thesaurus, which contains the controlled list of MEDLINE indexing terms. This design option broadens the application domain of MMTx, but it also affects its accuracy in the MEDLINE indexing task, as shown in the experiments in section 5.
- Term selection is based on a scoring function, for evaluating the importance of all

candidate terms, using the SPECIALIST lexicon as an external lexical resource. Moreover, the scoring function, though partly based on valid linguistic principles, such as the centrality criterion, it is arbitrarily and empirically defined, making it possible for unrelated terms to be included in the list of extracted terms. The C/NC-value scoring functions are especially tuned to multi-word terms, taking into consideration nested terms and term context words. Additionally, C/NC-value has been proven to extract up to 98% of correct terms [5], [21], [45], [34] in various application domains. Finally, WordNet and MeSH can be used as additional lexical resources, if needed, for both general and medical terms.

Based on the above observations, some basic changes are proposed towards the development of an improved term extraction method that could substitute MMTx:

1. **Step 1** and **Step 2** of MMTx can be replaced by C-Value (or the complete C-Value/NC-value) method. This method is corpus independent, does not need a lexicon and has been proven to be particularly effective in term extraction in medical and general document collection.

2. **Step 3** of MMTx by default is using UMLS Metathesaurus terms. Substituting Metathesaurus with the MeSH thesaurus is proposed. MeSH can also be used for locating semantically similar and variant terms.

3. **Step 4** can be replaced with the NC-value ranking method or by a mix of NC-value with the MMTx method. Both the MMTx and the NC-Value formula, evaluate extracted terms based on linguistic and statistical criteria.

4. **Term expansion**. The list of terms is augmented by hyponyms and hypernyms which are semantically similar to terms already in the list. Discovering semantically similar terms using MeSH and the semantic similarity method by

[30] is proposed. The evaluation of the semantic similarity methods [38] indicated that this method is particularly effective, achieving up to 73% correlation with results obtained by humans.

Each term is represented by its MeSH tree hierarchy. Fig. 4 illustrates this process: The neighborhood of the term is examined and all terms with similarity greater than threshold T are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term depending on the value of T.

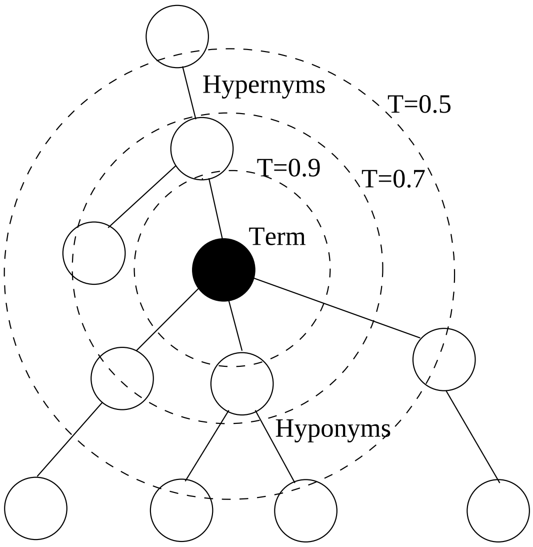


Figure 3.1: Term expansion using MeSH.

An important observation and a desirable property of most semantic similarity methods is that they assign higher similarity to terms which are close together (in terms of path length) and lower in the hierarchy (more specific terms), than to terms which are equally close together but higher in the hierarchy (more general terms). Therefore, expanding with threshold T will introduce new terms depending also on the position of the terms in the taxonomy: More specific terms (lower in the taxonomy) are more likely to expand than more general terms (higher in the taxonomy). Notice finally that expansion with low threshold values T (e.g., $T = 0.5$) is likely to introduce many new terms and diffuse the topic of the query (topic drift). In this work work $T = 0.9$ (the query is expanded only with very similar terms). As shown in [6] expansion with lower values of T (e.g., $T = 0.6$) demonstrated an increase in recall (more correct terms are revealed) but at the same time a decrease in precision (the expansion step introduced some unrelated terms as well).

Because no synonymy relation in defined in MeSH, expansion to MeSH terms with Entry Terms was not applied. Entry terms also include stemmed MH terms and are sometimes referred to as quasi-synonyms (they are not always exactly synonyms). The specification of T requires further investigation (e.g., appropriate threshold values can be learned by training). Word sense disambiguation [37] can also be applied to detecting the correct sense to expand (rather than expanding the most common sense of each term).

3.1 The AMTE$_X$ Algorithm

An outline of the AMTE$_X$ procedure is illustrated in Fig. 3.2. In particular, the AMTE$_X$ method has the following processing stages:

Input: Document d, MeSH taxonomy.

Output: MeSH terms t.

1. **Multi-word Term Extraction:** C/NC-value method
2. **Term Ranking:** C-value ranking
3. **Term Mapping:** Only MeSH terms are retained.
4. **Single-word Term Extraction:** Single-word MeSH terms are added.
5. **Term Variants:** Stemmed terms are added.
6. **Term expansion:** Semantically similar terms from MeSH

Figure 3.2: AMTE$_X$ algorithm.

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 candidate terms are evaluated, by C-value. The final candidate term list is ranked by decreasing term likelihood. Top ranked terms are more important than terms ranked lower in the list and are more likely to be included in the final list of extracted terms.

3. *Term Mapping:* Candidate terms are mapped to terms of the MeSH Thesaurus, (by simple string matching) by complete, full string matching. 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 AMTE$_X$ since the full list of Entry Terms may contain terms, which often are not synonymous.

6. *Term Expansion:* The list of terms is augmented with semantically similar terms from MeSH. Fig. 4 illustrates this process: a term is represented by its MeSH tree hierarchy (hypernyms/hyponyms). The neighbourhood of the term is examined and all terms with similarity greater than threshold T are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term depending on the value of T.

An example of how AMTE$_X$ works is shown in Fig. 3.3.

Input: A non-surgical approach to the management of lumbar spinal stenosis: A prospective observational cohort study...the remaining patients, LSS was established by the presence of low back pain and leg pain in an older individual with a clear history of neurogenic claudication...was responsible for statistical analysis, helped with design and presentation, and contributed to the writing of the manuscript.(full article)

Output: lower_back_pain shoulder_pain odds_ratios neck_pain public_health

1. **Compute C/NC–Value multi-word terms:** year_olds lifestyle_survey adolescent_health lower_back lower_back_pain shoulder_pain salminen_jj neckshoulder_pain bmj_volume year_olds_group odds_ratios past_half virtanen_sm neck_pain public_health

2. **Compute single terms :** year old lifestyle survey adolescent health lower back pain shoulder salminen jj neckshoulder pain bmj volume group odds ratios past half virtanen sm neck public health

3. **Mapping:** For all terms in candidate list, map each one to terms from the MeSH taxonomy

4. **Final Mappings:** odds_ratio data_collection low_back_pain shoulder_pain adolescent public_health life_style neck_pain health

Figure 3.3: AMTE$_X$ algorithm example.

AMTE$_X$ in its current state does not include a syntactic parser, such as the SPECIALIST minimal commitment parser used in MMTx. This is due to the fact that AMTE$_X$ uses an alternative, well established method for term extraction, the C/NC-value, which relies on linguistic filtering rules and where the head/modifier information is indirectly inferred through the statistical measures, namely the nested term estimations. In AMTE$_X$ v2 presented here, the estimated head of a multi-word term is successfully used for the refinement of the Single-word Term Extraction process.

AMTE$_X$ approach to Term Variant generation is more limited than MMTx. This constrains the term recall to terms that are closer to the original term in text. As it is observed in the results of the experiments in section **??**, AMTE$_X$ managed to achieve better precision in a fraction of the processing time taken for MMTx. This is partly due to the fact that AMTE$_X$ outperforms MMTx in suggesting candidate terms. It is also due to the fact the AMTE$_X$ approach to variant generation is limited to MeSH and does not operate iteratively, generating variants out of already found variants, thus avoiding the diffusion of the original concept to unrelated concepts.

In Term Expansion, the method used in AMTE$_X$ for discovering semantically similar terms, is based on the semantic similarity method by Li et al. [30]. The evaluation of the semantic similarity methods indicated that this method is particularly effective, achieving up to 73% correlation with results obtained by humans [6]. An important observation and a desirable property of this method is that it tends to assign higher similarity to terms which are close together (in terms of path length) and lower in the hierarchy (more specific terms), than to terms which are equally close together but higher in the hierarchy (more general terms). Therefore, expanding with threshold T will introduce new terms depending also on the position of the terms in the taxonomy: More specific terms (lower in the taxonomy) are more likely to expand than more general terms (higher in the taxonomy). Figure 4 illustrates this process for various values of the threshold T.

Because no synonymy relation is defined in MeSH, in this work expansion was not applied to the Entry Terms of terms. Word sense disambiguation [37] can also be applied for detecting the correct sense to expand (here, expansion is applied to the most common sense of each term).

3.2 Refining the AMTE$_X$Method

In order to determine the optimal set of indexing terms, namely one increasing recall and precision, there exist three thresholds in the AMTE$_X$ process that could be refined:

i) C-Value threshold (T_{Cvalue}) for the term extraction, which in the initial experiments presented in [6] was set to its recommended value ($T_C value = 1.5$) to limit output to the most valid terms;

ii) Term expansion threshold ($T_{Expansion}$), whereupon we have experimented in our pilot small scale experiments with AMTE$_X$ [6] ;

iii) Final list threshold ($T_{FinalList}$), which determines the minimum value a mapped to MeSH candidate index term must have to be included in the final index term list. In the experiments presented in [6], all candidate terms were retained.

The optimal value for each of these thresholds is not easy to determine, as each of these affects recall at different stages of the AMTE$_X$ process [6]. A simple approach to this optimization problem would be to consider only the threshold applied at the end of the process, the $T_{FinalList}$. Moreover, precision or recall alone should not determine an optimal threshold, since an increase in precision for example, simultaneously affects recall. A balanced measure such as an F-measure, where recall and precision are equally weighted (shown on Equ. 3.1 below), would provide us a better indicator for the final threshold.

$$F = \frac{2 * precision * recall}{precision + recall} \tag{3.1}$$

Thus, in AMTE$_X$ v2, we have chosen to be exhaustive with both T_{Cvalue} (i.e. $T_{Cvalue} = 0$) and $T_{Expansion}$ (i.e. $T_{Expansion} = 0.5$) thresholds and use the maximum F-measure to determine the $T_{FinalList}$. Moreover, in the Term Expansion step, the semantic similar terms ($T_{Expansion} = 0.5$) added to the candidate list are assigned a weight, as shown on Equ. 3.2 below:

$$weight(w) = sim * weight(s) \tag{3.2}$$

where a term w, semantically similar to term s, has ranking weight, $weight(w)$, combining its semantically similar term weight, $weight(s)$, and the similarity value, *sim*, by which w is similar to s. In this way, in AMTE$_X$ v2 the final candidate list ranks accordingly terms which are added to it by the Term Expansion process. In AMTE$_X$ v1, these terms were merely assigned the $weight(s)$ of Equ. 2.2.

In the pilot experiments with AMTE$_X$ v1 [6], in the Single-word Term Extraction step, an attempt was made to find partial matches in MeSH, for all word constituents of an unmatched multi-word term. It was observed that single term insertion in the candidate list through that process produced worse results. In AMTE$_X$ v2, we have chosen to conceptually limit the search for single-word mappings using only the head word of the multi-word term. The experiments presented in chapter 5 of this study show that this type of Single-word Term Extraction slightly improves both recall and precision. Regarding ranking weight for these terms, we consider it equal to its source, i.e. the original multi-word term weight.

In the next chapter we introduce a term classification study, for indexing and for discriminating documents between those suitable for expert and consumer users respectively. The proposed tools are all implemented and integrated into an online health information system, with indexing, retrieval and browsing capabilities, where system operations and system results (e.g. query results) are automatically classified based on user type.

Chapter 4

MedHealth: A Medical Information System for Consumer and Expert Users

MedHealth approach that supports extraction, categorization and retrieval of medical information by user profile is presented. It works in stages, the most important of them being the construction the dictionaries for medical and consumer terms respectively and the document classifier.

4.1 A Dictionary of Medical Terms for Consumer and Expert users

In order to achieve a categorization of terms into consumer and expert terms, the following data and algorithmic resources are needed:

- MeSH thesaurus. A taxonomy of medical and biological terms and concepts suggested by the U.S National Library of Medicine.
- Wordnet [1] thesaurus. A large lexical database of English terms (alternatively the SPECIALIST Lexicon of the UMLS can be used instead).
- A method for extracting MeSH terms from medical documents. AMTE$_X$ or MMTx discussed in Chapter 3 and section 2.4.1 are used in this work.
- Score function. A function denoting the probability of a document to belong into one of the two categories (i.e., consumer or expert document).

[1] http://wordnet.princeton.edu

MeSH thesaurus contains *medical* terms. Some of them are general (more abstract) and some are more specific and are used mainly by experts. Wordnet thesaurus is a *general* domain vocabulary containing general English terms as well as common medical terms, easy to comprehend by naive users (consumers). Based on this observation, medical terms are categorized into non-medical terms, medical terms for experts and medical terms for consumers. There may be terms common in two, or in all the three categories above. This is equivalent to creating three new vocabularies. As we shall show below, medical documents can also be categorized as non-medical, consumer and expert documents respectively. This is also illustrated in Figure 4.1 below.

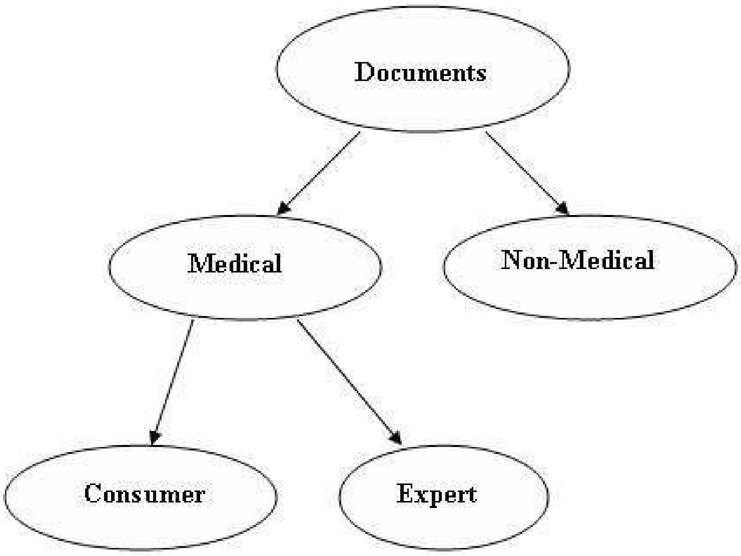


Figure 4.1: Document's Expert-Consumer Categorization

Wordnet thesaurus (2006) contains 127.361 terms, while MeSH (2006) contains 23.884 terms. The three new vocabularies are constructed by combining their terms are follows:

- **Vocabulary of General Terms (VGT):** these are terms that belong to Wordnet vocabulary and not in MeSH:

$$(Wordnet) - (MeSH) \tag{4.1}$$

  It follows that VGT contains **105.675** general (Wordnet) terms.

- **Vocabulary of Consumer Terms (VCT):** these are terms that belong to Wordnet and also in MeSH:

$$(Wordnet) \cap (MeSH) \tag{4.2}$$

  It follows that VCT contains (a subset of the MeSH terms) **7.165** consumer (MeSH) terms.

- **Vocabulary of Expert Terms (VET):** these are MeSH terms that do not belong to WordNet:

$$(MeSH) - (Wordnet) \tag{4.3}$$

  It follows that VET contains **16.719** consumer (MeSH) terms.

Notice that, *consumer* and *expert* terms are *only* MeSH terms (their intersection is the MeSH vocabulary). Notice also that the 70% of the MeSH terms are *expert* terms, while only 30% of MeSH terms are *consumer* terms. Documents are represented by term vectors produced by AMTE$_X$ (v2.0) and MMTx respectively. Each term in this vector is represented by its weight. The weight of a term is computed as a function of its frequency of occurrence in the document collection and can be defined in many different ways. The term frequency-inverse document frequency model is used for computing the weight: The weight d_i of a term i in a document is computed as

$d_i = tf_i * idf_i$, where tf_i is the frequency of term i in the document and idf_i is the inverse document frequency of i in the whole document collection.

$$document_i = \{0.62, 0.38\} \tag{4.4}$$

$document_i$ consists of 62% consumer terms and 38% expert terms. These numbers presents the probabilities of a document to belong in each category. In retrievals, these probabilities are also combined with the documents similarity score computed by VSM in response to queries to produce the ranking of documents which are provided as answers to consumer or expert users.

4.2 Document Classification

4.2.1 Document Classification by user profile

For each document both, its vector representation and its score probability pair are computed. Based on this information, document categorization is determined by machine learning or heuristics such as those discussed below.

- **Machine Learning by decision trees**. Let the system decide which category a document belongs. Creating a *Decision Tree* with 100-200 consumer and expert documents and after the training process, the system can decide in which category a input document belongs to.

- **Heuristic Categorization**. MeSH terms that belong to VCT may be regarded as *expert* ones (i.e *pain*, alzheimer, e.t.c). Consequently, a document is regarded as one suitable for expert users, if its corresponding concept vector contains *at least* 1 expert term from the VET. Otherwise it is regarded as a *consumer* document.

  Likewise a document is assigned a weight of *belief score* representing its probability of belonging into one of the two categories. This score is computed according to Formula 4.4 above.

In retrievals, and fore ranking query answers according to user profile we distinguish between the following two cases:

- If *user profile is known* (e.g., the user identifies him/herself as expert or consumer) then, the document score computed by VSM[23] is multiplied by the document belief score (Equation 4.4) that the document matches his/her profile.
- The *user profile in unknown* then, the system determines his/her profile based on the query. If the query contains at least one expert term, the user is considered to be an expert. Retrievals are then processed similarly to the first case above.

4.2.2 Document Classification by Topic

A Consumer-Expert Health Information System is described below whose purpose is to classify medical document by topic, in order to help potential users to to browse the document collection and find what they are searching for. To achieve such categorization, examination of the semantics of terms contained in each document is needed. For this categorization, the UMLS Semantic network (SN) is used.

As denoted in [16] and further in [17] SN suffers from semantic type assignment errors. More specifically, there is a 13% inconsistency between the UMLS Metathesaurus and Semantic Network relationships, these researches were based in SN relations to prove that there are some errors in the semantic categorization. Adding more relations to SN, or splitting existing ones was also suggested as a resolution to this problem, meaning that there is a good percentage of consistency in the network as it is [39]. These observations are based on comparisons between SN and the UMLS Metathesurus and not the MeSH thesaurus that is used in this study. The Semantic Network consists of broad categories, and it is often presented as the overarching knowledge structure, while UMLS (or MeSH) contains mostly essentially finer-grained concepts (at a lower semantic level). Therefore, we believe that the SN is the appropriate source for the categorization of medical information as we are based on MeSH rather than on UMLS Metathesaurus. The Semantic Network (version 2008AA) consists of 135 semantic types, 54 relationships and 15 semantic groups (further categorization of the 135 semantic types).

In this study the Semantic Network is used as a layer above the documents. This means that depending on the terms in the vector representation of a document, the document is classified by topic, by simply mapping the Vector terms to their semantic categories-groups on the Semantic Network (see figure 4.2).

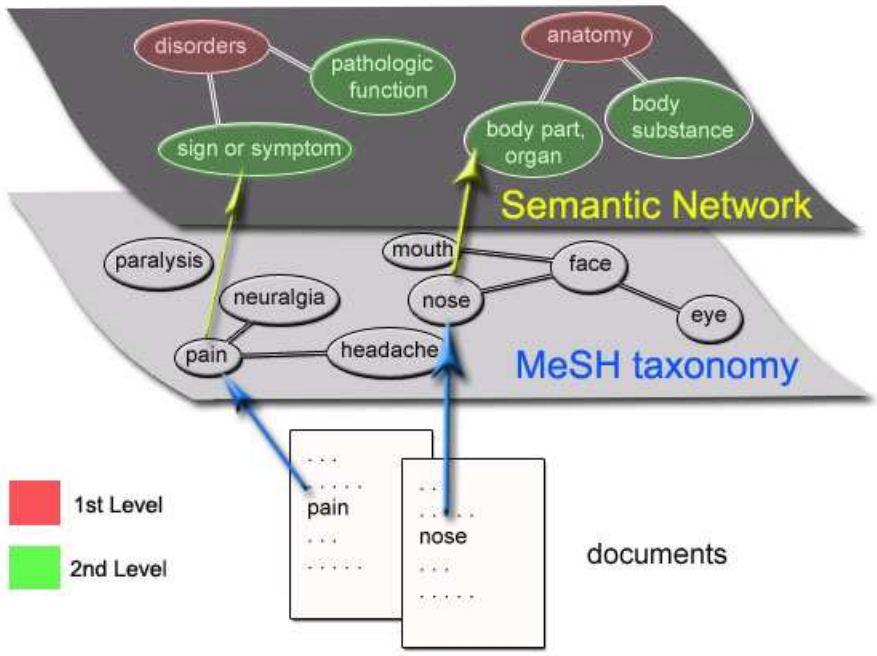


Figure 4.2: Document Categorization by Topic

Notice that, document categorization can be done using either of the two semantic (generalization) levels in SN. There are 15 major or 135 minor categories in the

Semantic Network. If we choose the top (first) level consisting of 15 major categories the document in the example denoted in figure 4.2 has different topics derived from the Semantic Network (i.e *"Anatomy"*, *"Disorders"*, e.t.c), than when choosing the lower (second) level with 135 semantic types.

In the next section the implementation of a consumer health information system is supporting the above functionality is presented.

4.3 System Architecture

Medical information systems such as medical web portals, have the same basic functionality. In particular, they are represented by four main modules. The *System Management* module, which parses each document from the medical collection and adds them to the system's database, the *Document Retrieval* module, which retrieves documents from the system's database in respect of the user's query, the *User Management* module, which manages the user's database and the *User Interface* module which provides the system's functionality to the the user. The system consists of four main components (see Figure 4.3)

Management Module: The system parses documents in the medical document collection, analyzes and indexes its content. The medical collection that is used here for demonstrating system's functionality is the OHSUMED collection. Before a document is added to the systems database, it must be processed by the **Document Analysis subsystem** which parses the document, extracts its (semantic and lexicographic) terms, categorizes it to Semantic Network MeSH categories (see Appendix A.4) and to consumer/expert categories with a weight of belief that it belongs to each category. **The Indexing subsystem** builds the documents terms indexes. The above process is shown in Figure 4.3.

Retrieval Module: This module retrieves documents from the database either in respect to the users query or by browsing the medical topics. This module consists of the following subsystems (see Figure 4.3): The **Query Handling** subsystem which parses the users query, extracts its (semantic and lexicographic terms), retrieves and ranks a list of documents relevant to the query and finally it suggests a list of terms for the query expansion process, if it is necessary [28]. For retrieval and **Browsing**, the system uses one of the: (a) lexicographic terms or (b) MMTx terms or (c) AMTEx terms representation vectors for both the query and the documents and produces a ranked list of retrieved documents, depending what option the user demanded.

User management: This module manages the users database. The user registers to the system, creating a profile whether he is an expert or consumer user, and an optional expression of favorite categories. In user login, the system validates the user (administrator, consumer or expert). In case where he has not registered, the system forces him to register, otherwise only browsing options is available.

User Interface: This module provides the systems functionality to its users. The user uses the system's functionality according to his type (administrator, expert, consumer). The user may enter a query, retrieve relevant documents and may reformulate the query, expanded with new terms, if the results are inefficient. Results presentation is part of this module.

Therefore, the user has the following options:

- *Login and registration*: The user inserts personal information to the system, and declare his profile (consumer/expert). The expert user can also indicate any desired favorite categories from the UMLS Semantic Network.
- *Term extraction method selection*: The user can select the term extraction method (AMTE$_X$ MMTx, or manually assigned index MeSH terms) before his search.
- *Browsing the system*: by selecting categories from the UMLS Semantic Network.
- *Searching the system*: The user has the option to search the system with a user defined query. The system automatically rank the results depending on the type of the user.

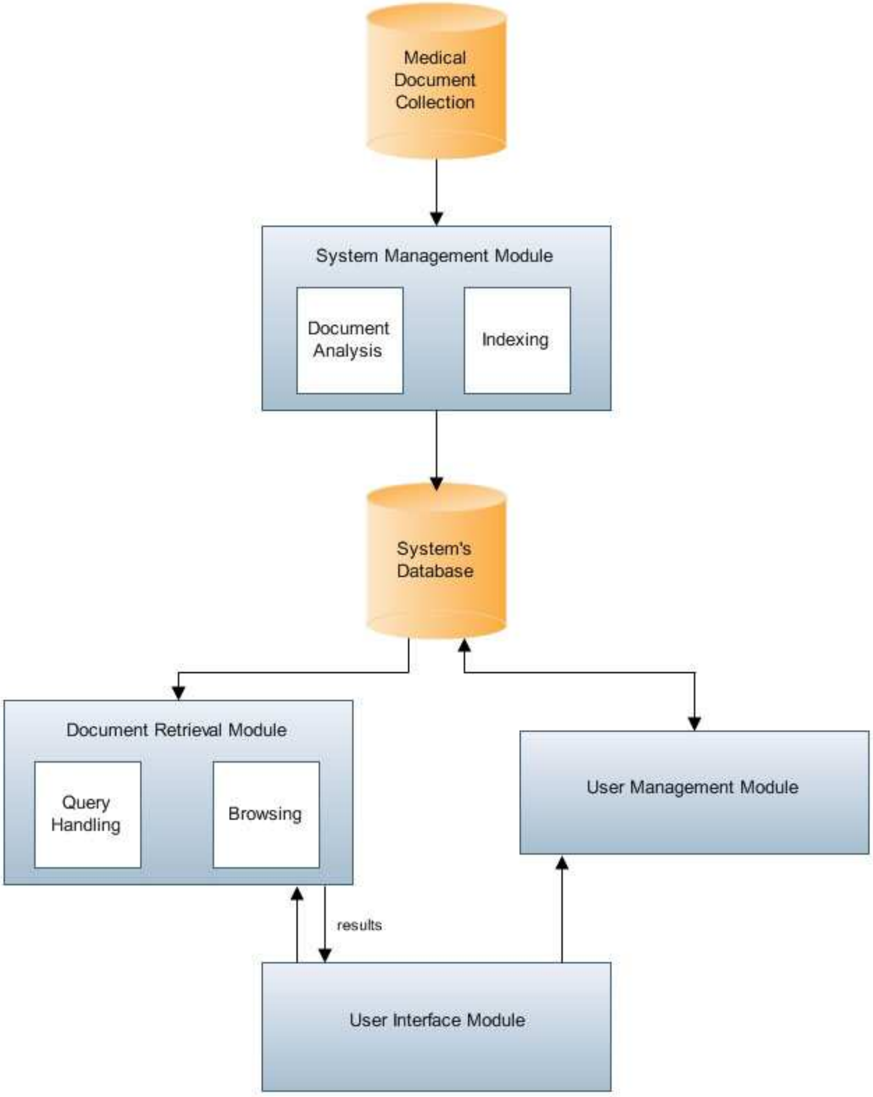


Figure 4.3: Architecture of the System

A complete working interface and API of the system, having most of the above characteristics, can be found at the web at **http://147.27.14.6:6900/medhealth**. In addition to the above, the Web user can find information about the term extraction methods, medical knowledge sources and vocabularies that the online system uses.

In section 5.4 you can find usage samples of the software along with some technical details. The system has a key advantage that has to be mentioned. It has plug and play features from the developers view, and further a friendly user interface from the consumer/expert user view. It is implemented in Java. Many features such as the information sources, or the term extraction methods, can be considered as functions, that can be added, moved or changed. In other words it is a *it's plugable architecture*, allowing for expansion with minimum effort. One can write a new term extraction method perhaps in Java and just plug the Class produced in the system. Samples and images of the demo interface are presented as well. Indexing and retrieval tasks of the system were implemented using Apache Lucene (see Appendix A.1).

Chapter 5

Experiments and Evaluation

There are two main categories of experiments. The goal of the information extraction experiments is to evaluate AMTE$_X$ and MMTx methods in indexing and information retrieval applications for abstract (OHSUMED) and full medical documents (PMC).In term classification experiments we evaluate our proposed information categorization methods in a retrieval task on OHSUMED and PMC datasets: The categorization method is deemed successful if it succeeds to retrieve medical information for the particular type of user issuing the query (consumer or expert user). Finally, the above functionality is integrated into a prototype medical information system which concludes the presentation of the results in this section.

5.1 Experimental Setup

The main data sources used in all experiments are listed below.

- **PMC** a corpus of 5,819 full PMC documents selected out of 60 Journals. The documents were selected on the basis of having an UID number, which was used to retrieve their respective MEDLINE index sets. This index set for each document is manually assigned by MEDLINE experts. The corpus was indexed with the Lucene (see Appendix A.1), creating a document database of 552Mb. For the main PMC database see Chapter 2.2.3.
- **OHSUMED** standard TREC collection corpus [1]. OHSUMED is a collection

of MEDLINE document abstracts used for benchmarking information retrieval systems evaluation. For more information about OHSUMED see Chapter 2.2.2. Besides the TREC collection corpus was indexed with the Lucene java library (see Appendix A.1), creating a document database of 1.321Mb.

- **Queries**. All abstract document (OHSUMED collection) experimental results were evaluated against the 64 TREC provided queries and answers of the standard TREC collection corpus.

5.2 Evaluation of AMTE$_X$

AMTE$_X$ v2, (see section 3.2) attempts to modify the Single-word Term Extraction process, using only the head term constituent for MeSH mapping. Nevertheless, we needed to ascertain that the single-word term extraction step significantly contributes to AMTE$_X$ performance, rather than unnecessarily complicating the AMTE$_X$ algorithm. Thus, a second experiment (using AMTEx v2 in addition to AMTEx) was conducted on the same dataset, where the single-word term extraction step was not included in the process.

In order to determine the $T_{FinalList}$ (see section 3.2), we have experimented with the PMC corpus. The MEDLINE index set for each document is used in this experiment as the ground truth. As for the evaluation, precision is the percentage of correctly retrieved terms compared to the total number of retrieved terms, and recall is the percentage of correctly extracted terms compared to the MeSH terms appearing in the respective MEDLINE document index. In this experiment F-measure of equally weighted precision and recall is used, as shown on Equ. 3.1 in section 3.2.

The comparative results in Figure 5.1 show clearly that Single-word Term Extraction improves AMTEx performance.

The peak of a curve in Figure 5.1 indicates the optimal F-measure performance. It is observed that the optimal F-measure performance is reached before the 20th point of a curve. Thus, in AMTE$_X$ v2, the $T_{FinalList}$ is set to the 20 top terms in the list.

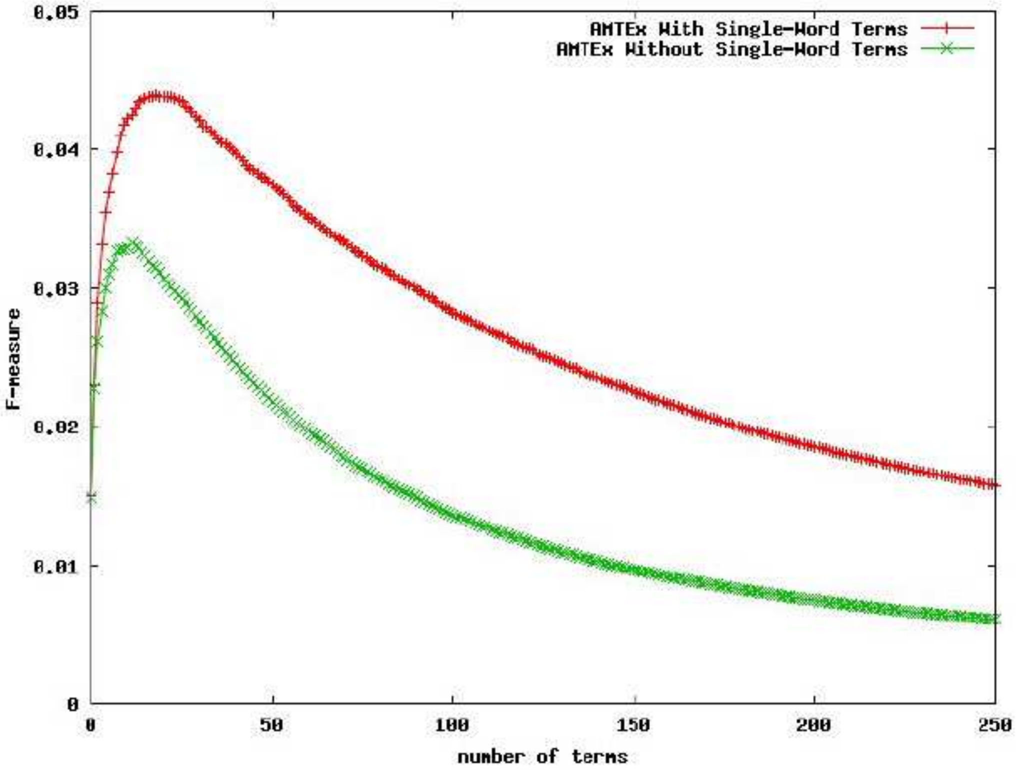


Figure 5.1: AMTE$_X$ with/without single-word extraction and $T_{FinalList}$ in PMC dataset.

5.2.1 MMTx vs AMTE$_X$ Method

In the pilot experiments presented in [6], a comparison of the first AMTE$_X$ version performance towards MMTx, which is considered the benchmark method, using a small set of 61 full documents. In this work, a series of comparative experiments were conducted to test the AMTE$_X$ approach in:

- a significantly larger corpus of full documents,
- a corpus of document abstracts,
- using both versions of AMTE$_X$ v1 and v2,

- for indexing and retrieval tasks,
- against MMTx, v24B.

For this reason, four experiments were conducted, comparing AMTE$_X$ v1 and v2 to MMTx v2.4B: the first two experiments assess the performance of AMTE$_X$ vs MMTx in the indexing task on a corpus of document abstracts (Abstract Indexing experiment) and on a large dataset of full documents (Full Document Indexing experiment). The other two experiments compare the performance of AMTE$_X$ vs. MMTx v2.4B in the retrieval task using again the respective document abstract (Abstract-based Retrieval experiment) and full document (Full Document-based Retrieval experiment) datasets.

It should be noted that in MMTx term ranking is less rigorous than AMTE$_X\dot{\text{I}}$n MMTx valid term output has mostly a weight value of 1000, whereas in AMTE$_X$ each term is ranked based on its individual weight. Thus, the evaluation score value of the 10th or 100th best answer of MMTx is not particularly adequate, since all its results may be equally weighted. This fact makes hard any controlling processes for the over-generated extracted MMTx terms.

Also it must be considered that for the indexing experiments, we thought it to be fair for MMTx to restrict its term mapping process to MeSH, rather the full UMLS, similarly to the AMTEx, since the ground truth consists of the MEDLINE provided index sets, which are based on MeSH.

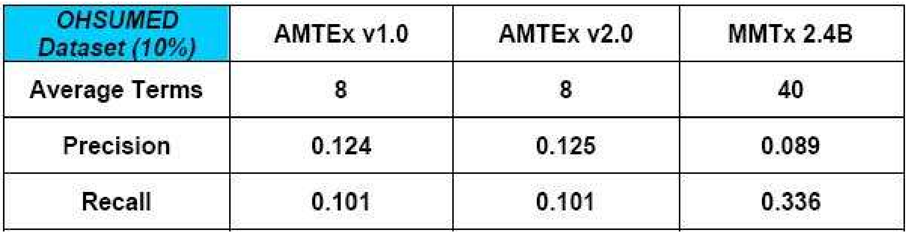
Abstract Indexing Experiment

This first experiment is conducted to test the performance of the three systems in the indexing task in a document abstracts corpus. The problems related to processing document abstracts were first identified in the pilot experiments with AMTE$_X$ [6]. These relate to the abstract size, which is quite limited to be used as input to a method using statistics, such as AMTE$_X$. Moreover, the content of the abstract has not been found to contain all necessary textual information for accurately indexing the full document. We have concluded at the time that we needed to enforce the AMTE$_X$ approach before embarking on such an experiment.

For the Abstract Indexing experiment presented here, a corpus subset of the OHSUMED standard TREC collection corpus wis selected. The selected subset is consisted of 10% of OHSUMED, i.e. 30,000 document abstracts (because MMTx is slow, processing of the entire OHSUMED was not feasible).These were again evaluated in terms of precision and recall against the MEDLINE provided MeSH index term sets.

For processing of document abstracts, AMTE$_X$ algorithm is slightly modified to respond to the problems of document limited size and content that was identified. Thus, both AMTE$_X$ versions first treat the totality of the corpus as a single document input during the term extraction step. Subsequently the extracted terms are associated to their respective source document by string matching. This modification of the AMTE$_X$ process has been thought necessary, since AMTE$_X$ term extraction is not only linguistic but also statistically based.

Table 5.1 demonstrates the comparative performance of AMTE$_X$ v1 and v2 against MMTx v2.4B in terms of average document precision and recall. It is observed that AMTE$_X$ shows improved precision compared to MMTx, and a reasonable recall by merely a fifth of the average term output compared to MMTx.



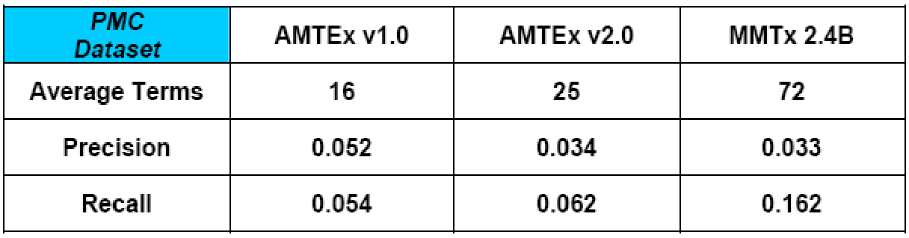
OHSUMED Dataset (10%)	AMTEx v1.0	AMTEx v2.0	MMTx 2.4B
Average Terms	8	8	40
Precision	0.124	0.125	0.089
Recall	0.101	0.101	0.336

Table 5.1: AMTE$_X$ vs. MMTx performance on the OHSUMED data set

Full Document Indexing Experiment

In the second experiment we have assessed the performances of the two versions of AMTE$_X$ against the MMTx v.2.4B in the indexing task using a full document dataset, the 5,819 PMC full document corpus. The results were evaluated for precision and recall, against the ground truth, i.e. the MEDLINE document index set (assigned manually by the experts). All methods process single document input during the term extraction step.

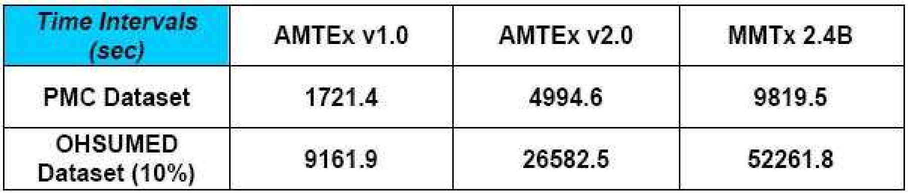
The results in Table 5.2 show average term output, precision and recall for each document, for all three systems. It is observed that AMTE$_X$ v1, shows a precision result that is higher than MMTx, whereas the average extracted terms are much less. AMTE$_X$ v2 demonstrates the best recall of the two AMTE$_X$ systems, for a fraction of the average MMTx term output.



PMC Dataset	**AMTEx v1.0**	**AMTEx v2.0**	**MMTx 2.4B**
Average Terms	**16**	**25**	**72**
Precision	**0.052**	**0.034**	**0.033**
Recall	**0.054**	**0.062**	**0.162**

Table 5.2: AMTE$_X$ vs. MMTx performance on the PMC data set

Finally, Table 5.3 illustrates the comparative results of all systems, in both full document PMC and OHSUMED document abstracts indexing experiments in terms of time efficiency. It is observed that the time taken for OHSUMED processing was longer in all systems. Nevertheless, both AMTE$_X$ systems are shown to perform much faster than MMTx. This is due to the algorithmic simplicity of AMTE$_X$ compared to MMTx especially with regards to variant generation and term expansion processes (even though MMTx was tested using MeSH rather than the full UMLS).



Time Intervals (sec)	AMTEx v1.0	AMTEx v2.0	MMTx 2.4B
PMC Dataset	1721.4	4994.6	9819.5
OHSUMED Dataset (10%)	9161.9	26582.5	52261.8

Table 5.3: Time intervals (in seconds) of AMTE$_X$ and MMTx for PMC & OHSUMED data set

Abstract-based Retrieval Experiment

In the third experiment we attempted to test AMTE$_X$ performance in the medical document retrieval task based on the document abstracts dataset. Documents are represented by term vectors produced by AMTE$_X$ (v2.0) and MMTx respectively. Document matching is performed by Vector Space Model (VSM, [23]). Both methods (i.e., retrieval by AMTE$_X$ and MMTx vectors) are compared against retrieval using vectors of MEDLINE provided index term sets, i.e. the terms used as ground truth in the indexing experiments. The OHSUMED standard TREC collection corpus subset used in the indexing experiment is used here as well. However, for this task the results were evaluated against 64 TREC provided queries and answers [1]. These constituted the ground truth for all systems performance.

Figure 5.2 illustrates the performance of AMTE$_X$ v2.0 compared to MMTx and the MEDLINE provided index term sets. Each method is represented by a precision/recall curve. For each query, the best 100 answers were retrieved. Precision and recall values are computed after each answer (from 1 to 100) and therefore, each curve contains exactly 100 points. Each point in a curve is the average precision/recall over 64 queries. The top-left point of a curve corresponds to the average precision/recall values for the best answer or best match (which has rank 1), while the right-most point corresponds to the average precision/recall values for the entire answer set.

It is observed that for this retrieval task based on the OHSUMED document abstracts dataset AMTE$_X$ approaches the performance of the manually assigned MeSH

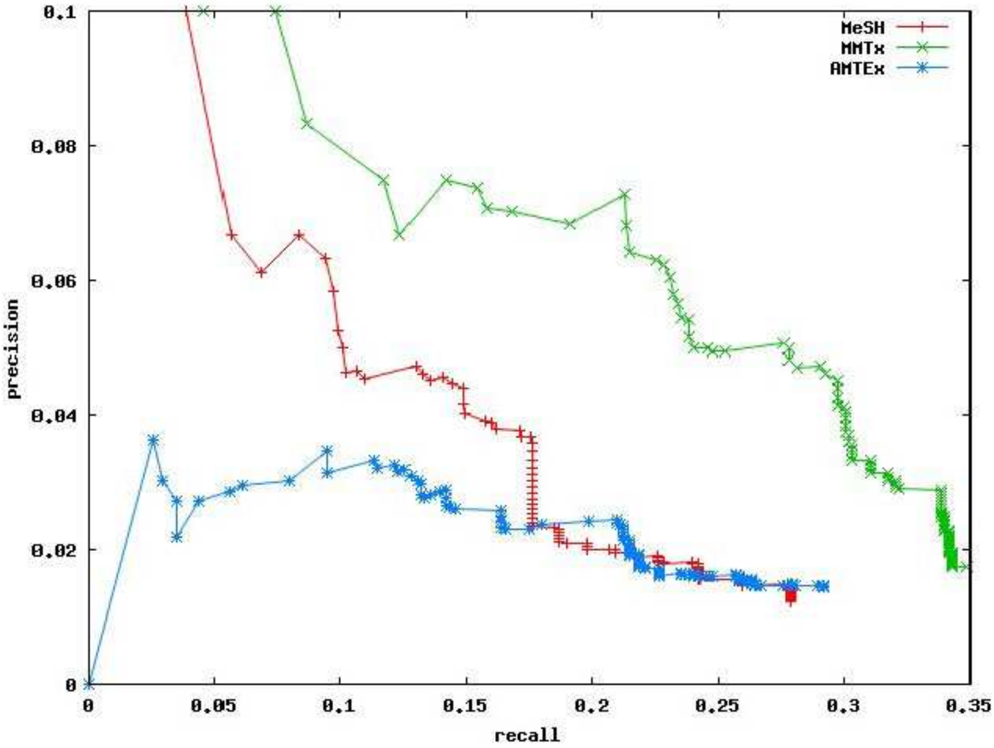


Figure 5.2: Precision/Recall of AMTE$_X$ vs MMTx on OHSUMED dataset retrieval task

terms as we gradually reach the entire answer set, while the increased recall of MMTx results in significantly better precision than both the manually assigned MeSH terms and AMTE$_X$.

The poor performance of AMTE$_X$ is due on the combined effect of two reasons. First, given the nature of the corpus, namely the OHSUMED document abstracts, AMTE$_X$ method due to its statistical part for term extraction, is slightly modified to treat the whole OHSUMED collection as a single document, rather than processing the very small individual document abstracts. The term results of this process were subsequently mapped to individual documents. At this stage the MMTx has the advantage of extracting few terms, even for small document size, which can be subsequently expanded, thus increasing MMTx term recall.

Secondly, this effect is further supported in retrieval, due to the nature of the

Vector Space Model (VSM) [23] in document matching. In particular, document matching relies on comparison of term vectors and in VSM partial matching is supported, i.e. for two documents to be similar the terms of one vector may be a subset of the terms of another vector. Thus, VSM clearly favors representation with many terms, without any regard to excessive terms, while AMTE$_X$ output incorporates semantic similarity of terms from the 6th step, not suitable for strictly string matching retrieval results.

As we shall see in the fourth and last experiment, using the PMC full document dataset the combined effect of these two factors is overcome and AMTE$_X$ performs clearly better when a full document rather than a document abstract is provided.

Full Document-based Retrieval Experiment

In the fourth and last full document retrieval experiment the 5,819 PMC full document corpus was also used as for the indexing task. In this experiment AMTE$_X$ method (v2.0) is again compared to MMTx, which is considered the benchmarking method for this task and to the retrieval results of the manually assigned MeSH terms. However, for this task the results were evaluated against 15 TREC provided queries (for PMC, there are no relevance judgments available by TREC or elsewhere). Relevance judgements on the first 25 answers retrieved by all the three competitive methods (AMTEx, MMTx and manually assigned MeSH terms) for all the 15 queries were provided by a domain expert (it was impossible to evaluate answers for the entire set of the 64 TREC queries as in the previous experiment as this would require 64x20x3 = 3,840 relevance judgments by the domain expert). The queries used for this experiment are presented in section A.2 at page 72. Figure 5.3 shows AMTE$_X$ (v2.0) clearly outperforming MMTx and nearing the performance of the manually assigned MeSH index terms.

Notice that although MMTx is tuned towards higher recall (by revealing more

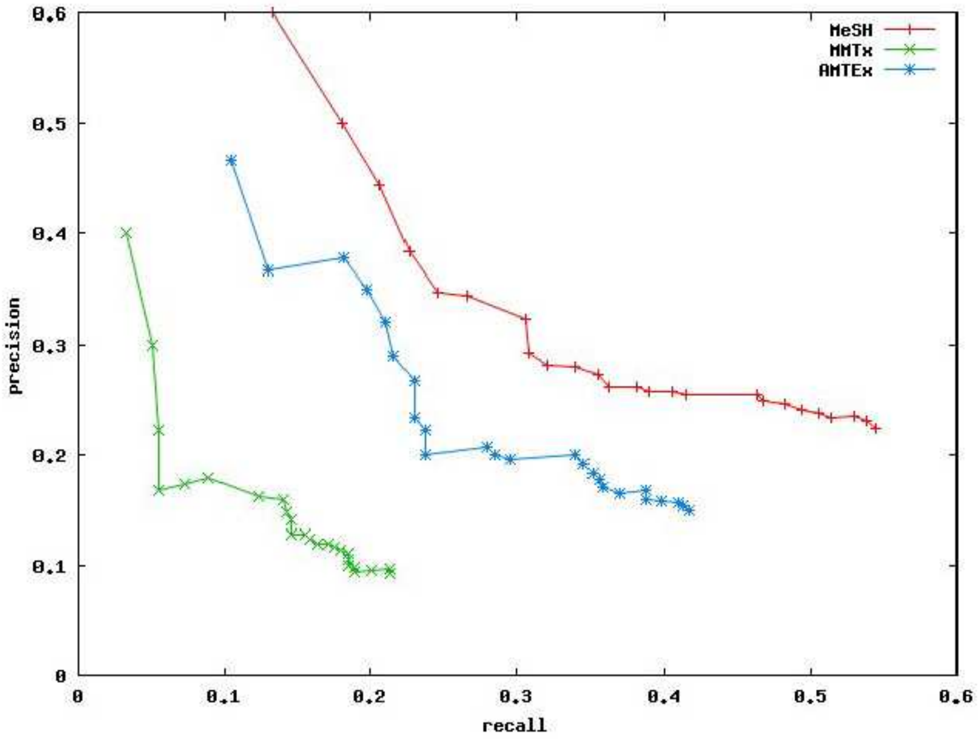


Figure 5.3: Precision/Recall of AMTE$_X$ vs MMTx on PMC dataset retrieval task

indexing terms) this does not always lead to improved retrieval performance. A possible explanation could lie in the fact that the document indexing is based on manual assignment of MeSH terms based on the document conceptual classification done by human experts, whereas retrieval is based on string matching (on terms or term parts found in the document) and the evaluation of these results.

Based on all four experiments we conclude that the AMTE$_X$ selective term output method is very well suited for both indexing and retrieval, performing faster and providing a better and concise term output, whereas MMTx increased recall can be well suited in some retrieval cases, where the small document size is prohibitive for the optimal application of AMTE$_X$statistical term extraction process.

5.3 Evaluation of MedHealth

Similarly to the indexing experiments, the performance of the proposed term and document categorization method is evaluated by a series of retrieval experiments. In particular, we run a series of experiments addressing consumer or experts users. In the first series of experiments we measure the capability of the proposed indexing method in categorizing the documents as expert or consumer documents. In the second series of experiments , a method is successful if it succeeds in returning relevant documents according to user's profile. For all experiments only the OHSUMED standard TREC collection corpus is used, because relevant judgements are given with the collection. AMTE$_X$ method (v2.0) is again compared to MMTx and to the manually assigned MeSH index terms. The queries used for this experiment are presented in section A.2 at page 72.

5.3.1 Document Indexing evaluation for Consumer and Expert Users

For the indexing task the results were evaluated against 15 TREC provided queries. Relevance judgements on the first 20 answers retrieved by all the three competitive methods (AMTEx, MMTx and manually assigned MeSH terms) for all the 15 queries were provided by the members of the Intelligence Systems Laboratory. Each human judged the answers to a number of queries (the same for all methods), by assessing if is an answer is a consumer document (simply by understanding what the document subject is about) or expert document (by not understanding what the document subject is about).

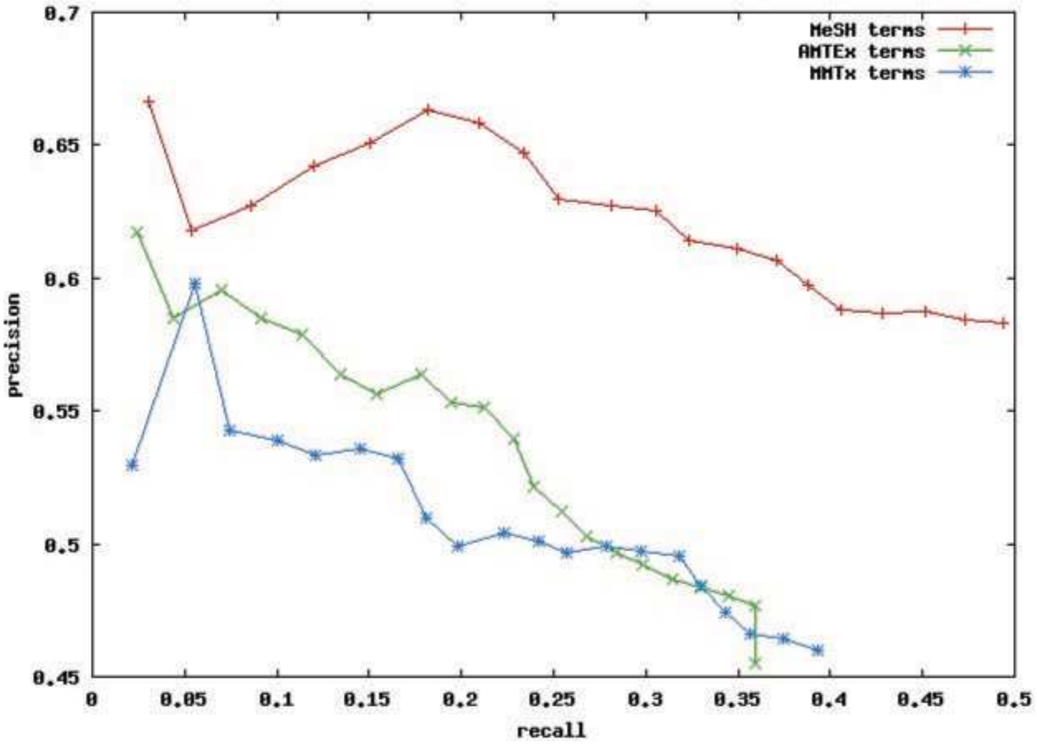


Figure 5.4: Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset consumer indexing task

In the consumer classification indexing experiment, documents that have higher similarity score, weighted with the consumer probability value are placed at the top of the results list. As shown in figure 5.4 the AMTE$_X$extracted terms curve draws closer to the performance of the manually assigned MeSH index terms (which is used as ground trouth).

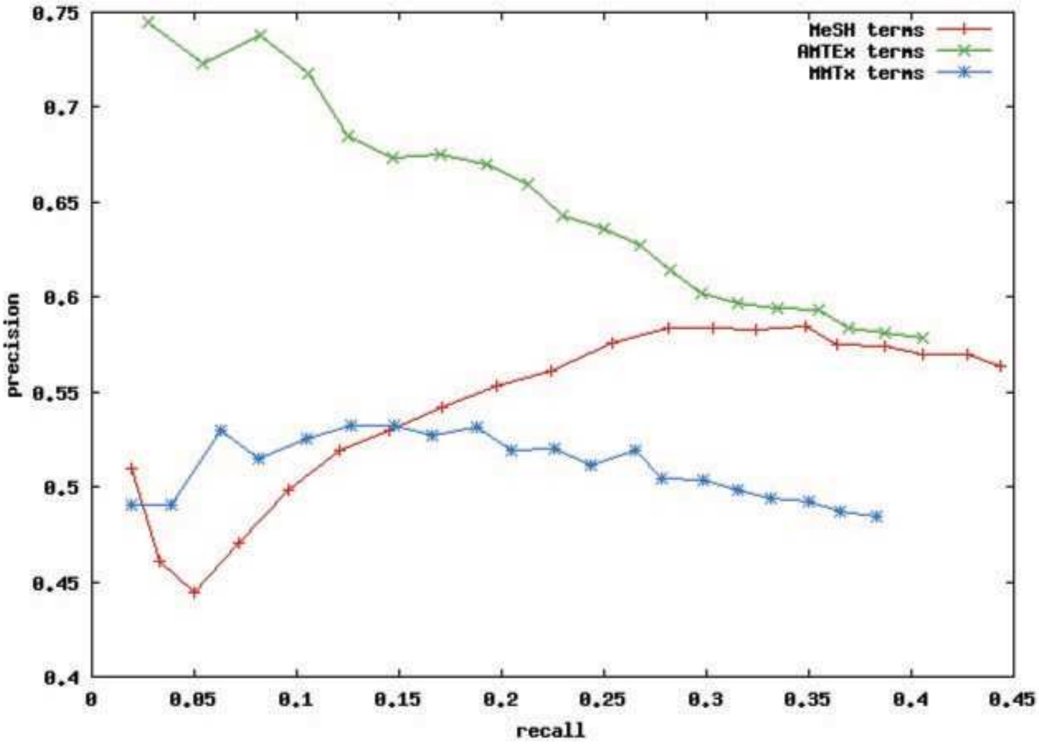


Figure 5.5: Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset expert indexing task

The users that evaluate the results noted that the OHSUMED dataset contains mostly medical expert documents, than consumer ones. According to that, AMTE$_X$ shows its selective ability to extract medical terms, outperforming even the manual assigned index MeSH terms, as shown in figure 5.5.

5.3.2 Information Retrieval evaluation for Consumer and Expert Users

For the retrieval task the results were also evaluated against 15 TREC provided queries as in the previous experiments, with the additional requirement that an answer is deemed relevant if it is both similar to the query (i.e., it is contained in the set of relevant answers provided for this query) and is also correctly categorized as an answer appropriate for expert or consumer users respectively.

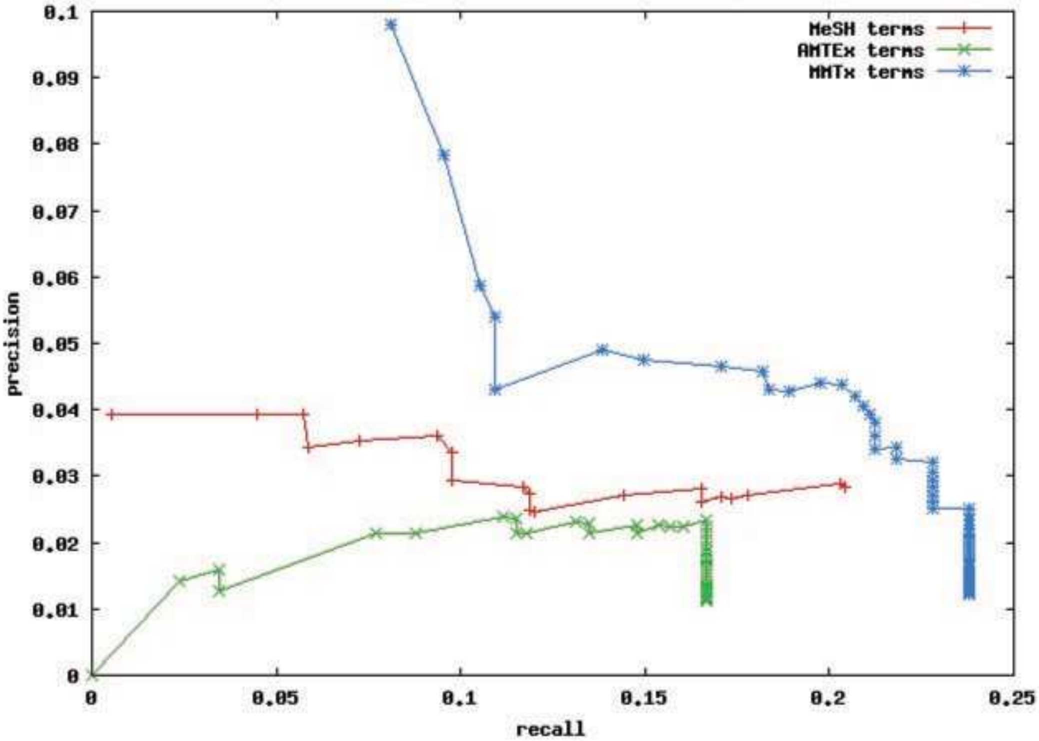


Figure 5.6: Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset consumer retrieval task

Both in the consumer and expert retrieval experiment as shown in figures 5.6 and 5.7, MMTx retrieval performance outperforms AMTE$_X$ and MesH index terms. As it was mentioned in section 5.2.1, this effect is further supported in retrieval, due to the nature of the Vector Space Model (VSM) [23] in document matching. Thus, VSM favors longer vectors (meaning more info), and the statistical analysis part of AMTE$_X$ term extraction method is aggrieved.

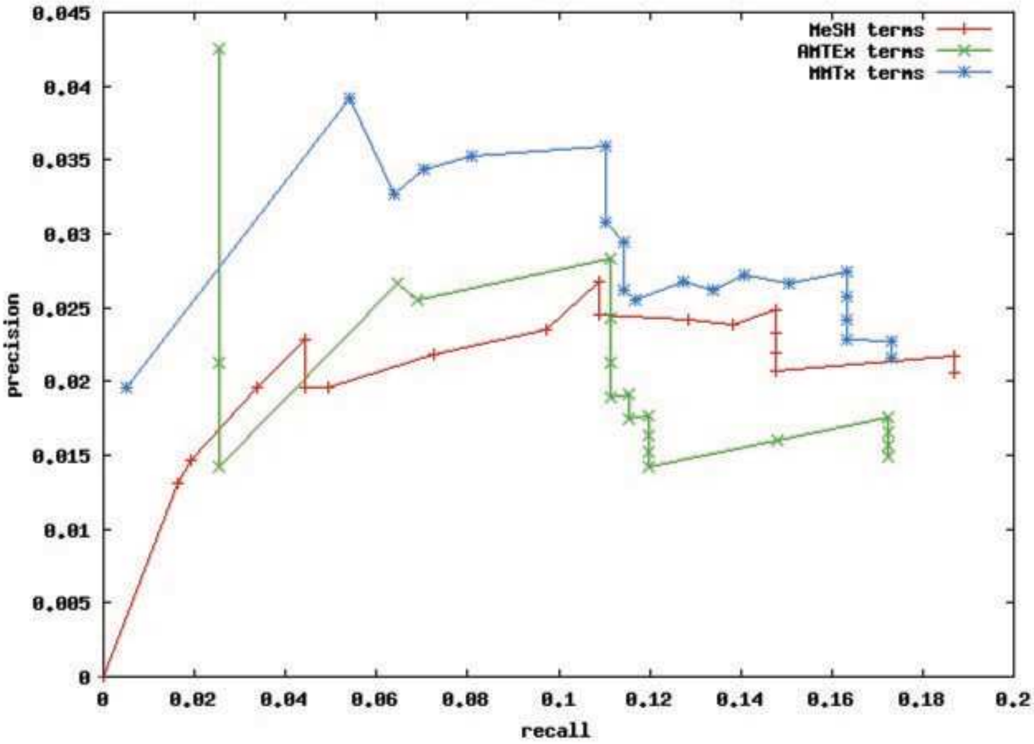


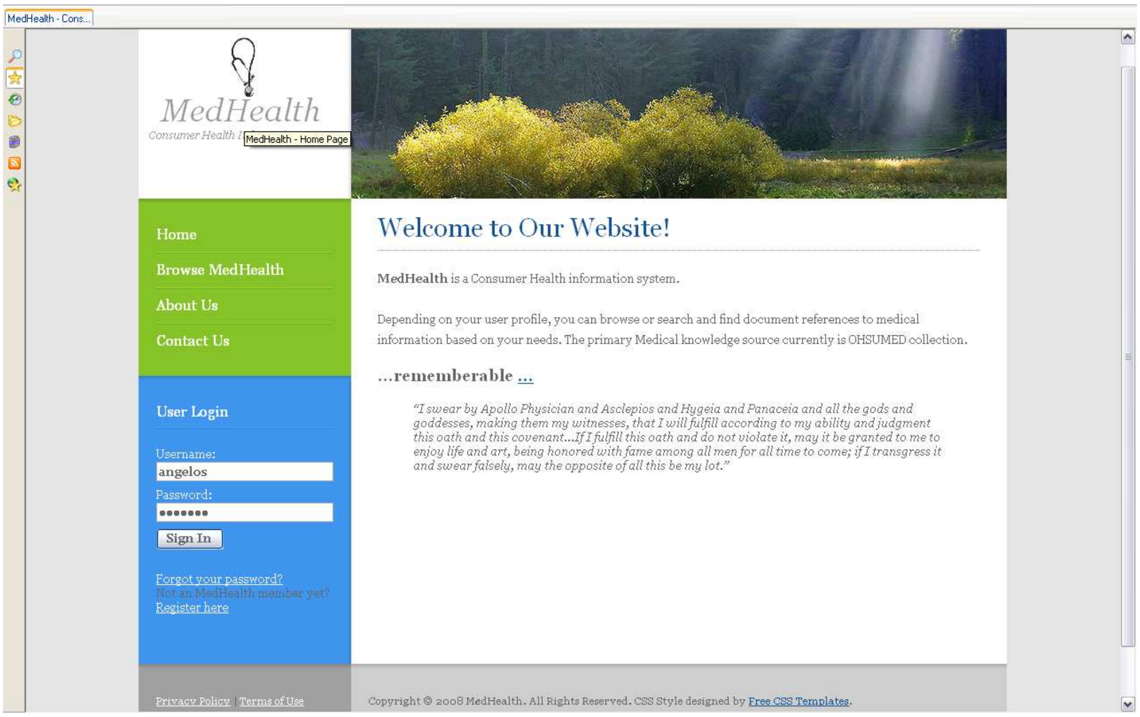
Figure 5.7: Precision/recall of AMTE$_X$ vs MMTx and MeSH index terms on OHSUMED dataset expert retrieval task

As it was mentioned before OHSUMED corpus is consisted mostly of medical *expert* documents (papers, medical records, diagnoses and diseases, materia medica and courses, e.t.c) while MeSH terms represent consumer terminology in a percentage of 30% and *expert* terminology at 70% (see section 4.1). This is the main reason that in the consumer experiment results are slightly different with the same experiment presented in section 5.2.1 without consumer/expert classification, while in the expert experiment, precision of AMTE$_X$ is better than all the other methods.

5.4 MedHealth

The Consumer Health Information System from the developer side of view, uses Java programming language, Apache softwares such as Tomcat server, HTML mark up language, java server pages, and several databases implemented in mysql and postgresql.

MedHealth system is currently in a demo stage. Its main functionality is discussed below:



Welcome to Our Website!

MedHealth is a Consumer Health information system.

Depending on your user profile, you can browse or search and find document references to medical information based on your needs. The primary Medical knowledge source currently is OHSUMED collection.

...rememberable ...

"I swear by Apollo Physician and Asclepios and Hygeia and Panaceia and all the gods and goddesses, making them my witnesses, that I will fulfill according to my ability and judgment this oath and this covenant...If I fulfill this oath and do not violate it, may it be granted to me to enjoy life and art, being honored with fame among all men for all time to come; if I transgress it and swear falsely, may the opposite of all this be my lot."

Username: angelos

Password: •••••••

Forgot your password?
Not an MedHealth member yet?
Register here

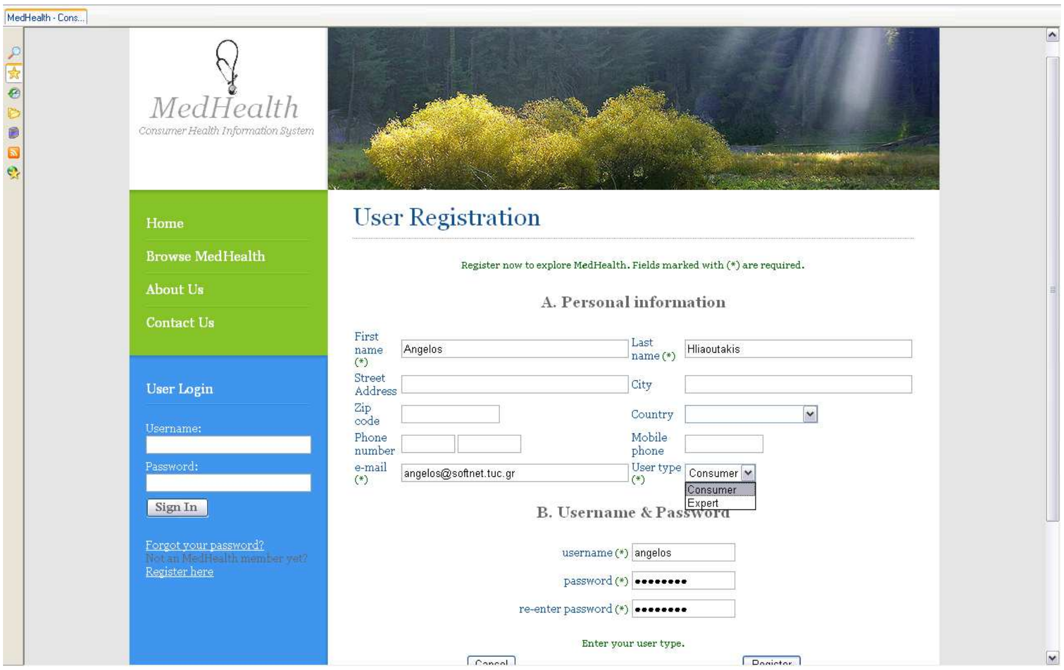
Privacy Policy | Terms of Use

Copyright © 2008 MedHealth. All Rights Reserved. CSS Style designed by Free CSS Templates.

Figure 5.8: Login on MedHealth

Login Users can login to MedHealth (Fig. 5.8), in order to access additional operations such as searching MedHealth or editing their own user profile.

Register Users can register on MedHealh sytem (Fig. 5.9), provide their personal information, and most important denote their type of user (consumer /expert). While they are logged in, consumer or expert discrimination is important for the system's browsing and searching operations.



User Registration

Register now to explore MedHealth. Fields marked with (*) are required.

A. Personal information

First name (*)	Angelos	Last name (*)	Hliaoutakis
Street Address		City	
Zip code		Country	
Phone number		Mobile phone	
e-mail (*)	angelos@softnet.tuc.gr	User type (*)	Consumer Consumer Expert

B. Username & Password

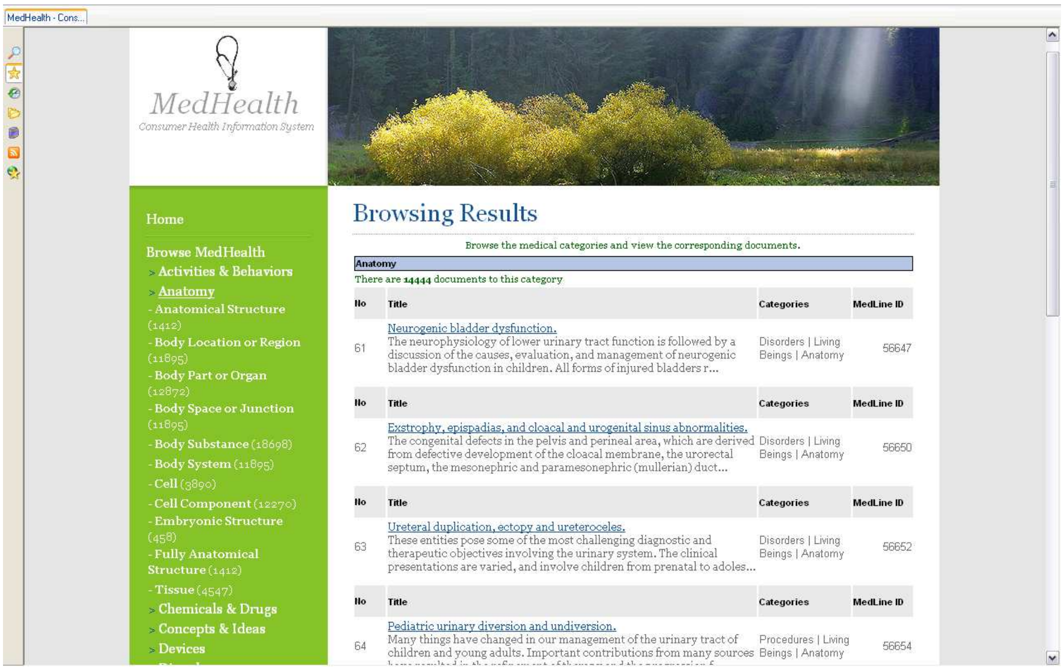
username (*) angelos

password (*) ●●●●●●●●

re-enter password (*) ●●●●●●●●

Enter your user type.

Figure 5.9: Register on MedHealth



Browsing Results

Browse the medical categories and view the corresponding documents.

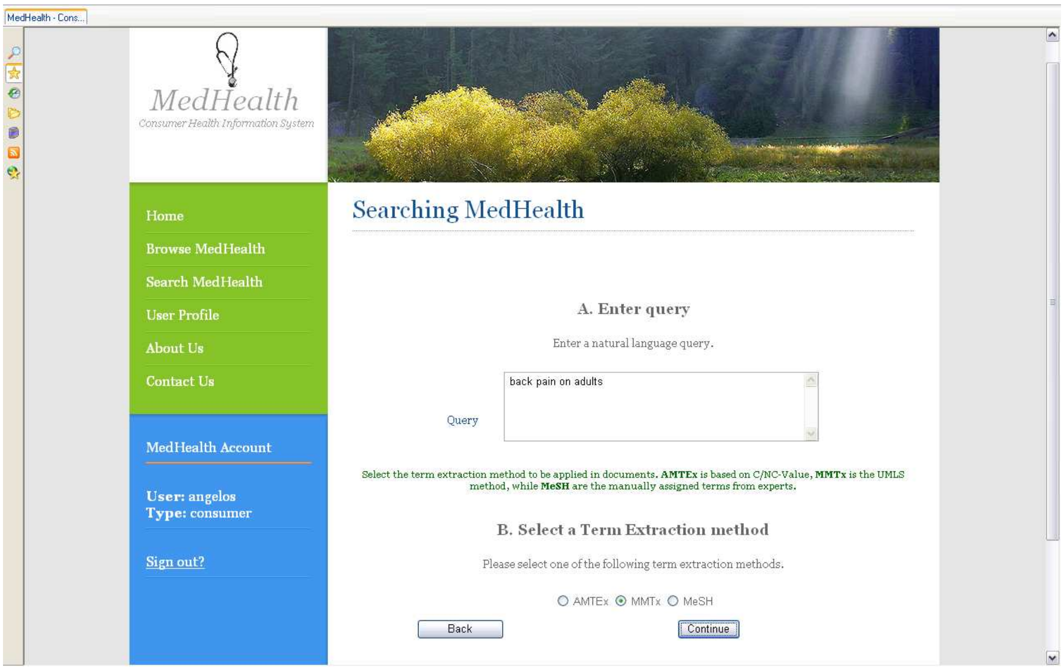
Anatomy

There are **14444** documents to this category

No	Title	Categories	MedLine ID
61	Neurogenic bladder dysfunction. The neurophysiology of lower urinary tract function is followed by a discussion of the causes, evaluation, and management of neurogenic bladder dysfunction in children. All forms of injured bladders r...	Disorders | Living Beings | Anatomy	56647
62	Exstrophy, epispadias, and cloacal and urogenital sinus abnormalities. The congenital defects in the pelvis and perineal area, which are derived from defective development of the cloacal membrane, the urorectal septum, the mesonephric and paramesonephric (mullerian) duct...	Disorders | Living Beings | Anatomy	56650
63	Ureteral duplication, ectopy and ureteroceles. These entities pose some of the most challenging diagnostic and therapeutic objectives involving the urinary system. The clinical presentations are varied, and involve children from prenatal to adoles...	Disorders | Living Beings | Anatomy	56652
64	Pediatric urinary diversion and undiversion. Many things have changed in our management of the urinary tract of children and young adults. Important contributions from many sources	Procedures | Living Beings | Anatomy	56654

Figure 5.10: Browsing MedHealth

Browsing Browsing operation can be performed from everyone, even if the user is not logged in MedHealth (Fig. 5.10). Although this may occurs, there are still advantages for the users that have been registered on the system. A registered user is a consumer or expert user. This means that while logged in the system, browsing action takes place accordingly to his nature (consumer/expert), while a user not logged in MedHealth (guest user) browsing results are not classified.



Searching MedHealth

User: angelos
Type: consumer

A. Enter query

Enter a natural language query.

Query: back pain on adults

Select the term extraction method to be applied in documents. **AMTEx** is based on C/NC-Value, **MMTx** is the UMLS method, while **MeSH** are the manually assigned terms from experts.

B. Select a Term Extraction method

Please select one of the following term extraction methods.

○ AMTEx ● MMTx ○ MeSH

Back | Continue

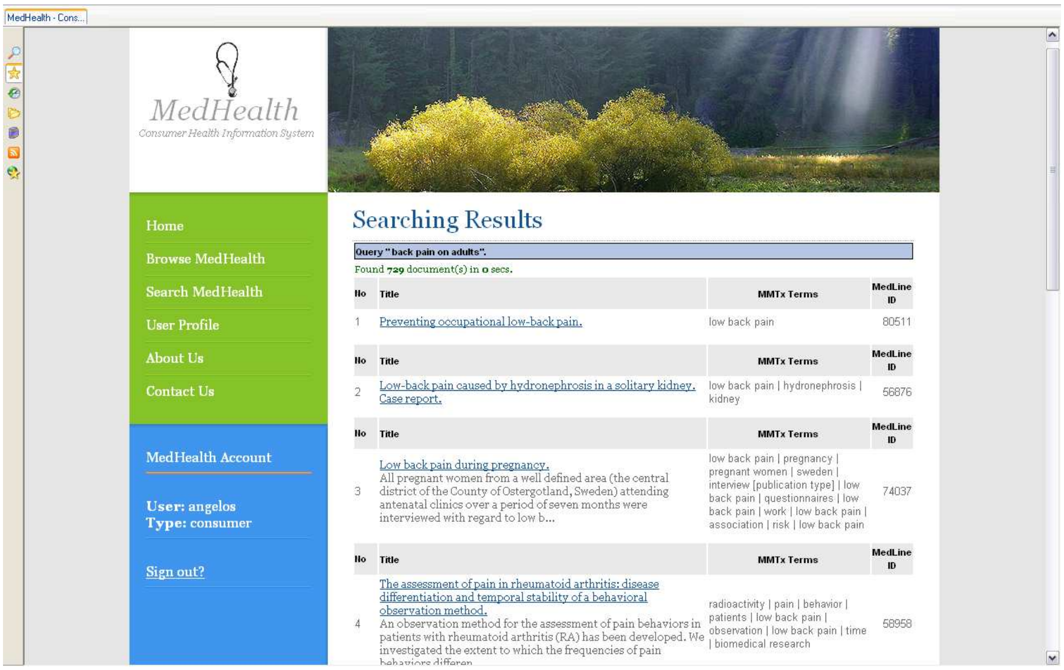
Figure 5.11: Searching criteria on MedHealth

Searching Users may search MedHealth (Fig. 5.12) only if they are registered users. They have to specify two basic things for the searching operation:

- A user defined natural language query and
- A term extraction method that will specify the document's field to be compared with (Fig. 5.11)

There are three implemented term extraction method selections. AMTE_X MMTx and the manually assigned MeSH terms. For example, selecting MMTx, means that the specified user query will be compared with the MMTx extracted terms field of every document in the MedHealth database (medical storage).

Note that MedHealth's searching method is the state-of-the-art Vector Space Model [23].



User: angelos
Type: consumer

Searching Results

Query "back pain on adults".

Found **729** document(s) in **0** secs.

No	Title	MMTx Terms	MedLine ID
1	Preventing occupational low-back pain.	low back pain	80511

No	Title	MMTx Terms	MedLine ID
2	Low-back pain caused by hydronephrosis in a solitary kidney. Case report.	low back pain | hydronephrosis | kidney	56876

No	Title	MMTx Terms	MedLine ID
3	Low back pain during pregnancy. All pregnant women from a well defined area (the central district of the County of Ostergotland, Sweden) attending antenatal clinics over a period of seven months were interviewed with regard to low b...	low back pain | pregnancy | pregnant women | sweden | interview [publication type] | low back pain | questionnaires | low back pain | work | low back pain | association | risk | low back pain	74037

No	Title	MMTx Terms	MedLine ID
4	The assessment of pain in rheumatoid arthritis: disease differentiation and temporal stability of a behavioral observation method. An observation method for the assessment of pain behaviors in patients with rheumatoid arthritis (RA) has been developed. We investigated the extent to which the frequencies of pain behaviors differen	radioactivity | pain | behavior | patients | low back pain | observation | low back pain | time | biomedical research	58958

Figure 5.12: Searching in MedHealth

Chapter 6

Conclusions

We presented approaches to the problems of automatic term extraction and automatic categorization of medical information according to user's profile (i.e., consumer and expert users). The term extraction problem for the automatic indexing of documents in large medical collections was presented. Existing approaches to this problem were also presented, focusing on the MMTx method. Building upon existing work on term extraction the AMTEx method is proposed, aiming at providing more accurate and concise terms while being more efficient in terms of processing speed. AMTEx is specifically designed for the automatic indexing of MEDLINE documents, using the MeSH Thesaurus resource and a well-established method for extraction of domain terms, the C/NC-value method. As a case study we consider the further classification of medical documents, to documents appropriate for naive and expert users. The performance of all methods is assessed by a series of experiments. AMTE$_X$ has been also compared against MMTx in the indexing and the retrieval tasks, with and without consumer–expert classification criteria. More evaluation experimental results must confirm the performance of AMTE$_X$ and MMTx methods, but in practice it is quite hard to find domain experts to handle the large load that the evaluation process entails.

Experimental results, both in abstract and full document collection, showed that AMTE$_X$ in more selective in the extraction of medical terms, indicating that is a useful automatic indexing method. Although MMTx shows weak performance in the indexing task, is most suitable for the retrieval of medical information, represented with longer term vectors. Results show that AMTEx performs very well in both tasks, with its average term output being 20 to 50% less than MMTx and its processing speed 3 to 5 times faster than MMTx. MMTxs increased recall may present better results in the small size document retrieval task, where the small document size is prohibitive for the optimal application of AMTEx statistical term extraction process.

Although AMTE$_X$ is a term extraction method for medical domain, it is in fact a general purpose one. Notice that one of the two main knowledge resources of AMTE$_X$ is the C/NC–value, a general domain term extraction method. The other resource is the MeSH thesaurus. Replacing the MeSH thesaurus with other thesaurus in other domains, (such as electronic engineering, or economics) AMTE$_X$ may become a term extraction method for different domains as well.

For the consumer–expert classification problem this work introduced a simple and easy way to discriminate documents into these categories, but combinational work from research fields such us fuzzy clustering, classification, may result in a more elaborate and accurate classification method.

Bibliography

[1] TREC:Text REtrieval Conference TREC-9 Filtering Track Collections: OHSUMED., March 2007. http://trec.nist.gov/data/t9_filtering.html.

[2] ISO 704. Principles and Methods of Terminology. Technical report, Intern. Organization for Standardization, Geneva, Switzerland, 1986.

[3] E. Voutsakis Euripides G.M. Petrakis E. Milios A. Hliaoutakis, G. Varelas. Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue of Multimedia Semantics*, Vol. 3, No. 3:55–73, July/September, 2006.

[4] S. Ananiadou. A Methodology for Automatic Term Recognition. In *Proc. of COLING-94*, pages 1034–1038, Kyoto, 1994.

[5] S. Ananiadou, S. Albert, and D. Schuhmann. Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline. *Genome Informatics Series*, 11, 2000.

[6] Euripides G. M. Petrakis Evangelos Milios. Angelos Hliaoutakis, Kalliopi Zervanou. Automatic Document Indexing in Large Medical Collections. In *ACM International Workshop on Health Information and Knowledge Management (HIKM 2006)*, Arlington, VA, USA., November 11, 2006. Kluwer Academic Publishers.

[7] Euripides G.M. Petrakis Evangelos Milios Angelos Hliaoutakis, Giannis Varelas. MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. In *10th ECDL European Conference on Research and Advanced Technology for Digital Libraries (ECDL'2006)*, pages 512–515, Alicante, Spain, September 2006.

[8] A. R. Aronson. MetaMap: Mapping Text to the UMLS® Metathesaurus®, March 1996. http://skr.nlm.nih.gov/papers.

[9] A. R. Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus®: The MetaMap Program. In *Proceedings of AMIA 2001*, pages 17–21, 2001.

[10] A. R. Aronson. MetaMap Candidate Retrieval, July 2001. http://skr.nlm.nih.gov/papers.

[11] A. R. Aronson. MetaMap Evaluation, May 2001. http://skr.nlm.nih.gov/papers.

[12] A. R. Aronson. MetaMap Variant Generation, May 2001. http://skr.nlm.nih.gov/papers.

[13] Daille B, E. Gaussier, and J. Lange. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proc. of COLING-94*, pages 515–521, Kyoto, 1994.

[14] Jeff Beck. PubMed Central. XML-based archive of life sciences literature at the NLM. 2005.

[15] Bethesda. *UMLS Reference Manual.* National Library of Medicine (US), NCBI, 2009.

[16] O. Bodenreider and McCray AT. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6):414–432, December 2003.

[17] O. Bodenreider and A. Burgun. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. *MEDINFO 2004 IMIA*, 11, Issue Pt.1, 2004.

[18] Olivier Bodenreider. Consistency between Metathesaurus and Semantic Network. In *Workshop on The Future of the UMLS Semantic Network NLM.* Lister Hill National Center for Biomedical Communications Bethesda, Maryland - USA, April 8, 2005.

[19] D. Bourigault, I. Gonzalez-Mullier, and C.Gros. LEXTER, a Natural Language Tool for Terminology Extraction. In *EURALEX '96: Proc. I-II, Part II – Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg*, pages 771–779, Göteborg University, Göteborg, Sweden, 1996.

[20] G. Divita, T. Tse, and L. Roth. Failure Analysis of MetaMap Transfer (MMTx). *Medinfo*, pages 763–767, 2004.

[21] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: The C-Value/NC-value Method. *International Journal of Digital Libraries*, 3(2):117–132, 2000.

[22] K.T. Franzi 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.

[23] Salton G. Automatic text processing: the transformation analysis and retrieval of information by computer. Reading MA: Addison-Wesley;. 1989.

[24] R. Gaizauskas, G. Demetriou, and K. Humphreys. Term Recognition in Biological Science Journal Articles. In *Workshop on Computational Terminology for Medical and Biological Applications, (NLP 2000)*, pages 37–44, Patras, 2000.

[25] Erik Hatcher and Otis Gospodnetic. *Lucene in Action.* 2004. 456 pages, ISBN: 1932394281.

[26] Hickam DH Hersh WR. Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, pages 382–389, 1994.

[27] Leone TJ Hickam DH Hersh WR, Buckley C. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pages 192–201, 1994.

[28] A. Hliaoutakis, G. Varelas, E. G.M. Petrakis, and E. Milios. MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. In *Proc. of the* 10^{th} *ECDL European Conference on Research and Advanced Technology for Digital Libraries (ECDL'2006)*, pages 512–515, Alicante, Spain, September 17-22 2006.

[29] C. Jacquemin. *Spotting and Discovering Terms through Natural Language Processing.* MIT Press, Cambridge, MA, USA, 2001.

[30] Y. Li, Z. A. Bandar, and D. McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):871–882, July/Aug. 2003.

[31] C. Manning and H. Schüzte. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA, June 18 1999.

[32] D. Maynard and S. Ananiadou. TRUCKS: A Model for Automatic Multi-Word Term Recognition. *Journal of Natural Language Processing*, 8(1):101–105, 2000.

[33] National Library Of Medicine. 2008AA Documentation. Section 3, Semantic Network. 2008.

[34] E. Milios, Y. Zhang, B. He, and L. Dong. Automatic Term Extraction and Document Similarity in Special Text Corpora. In *Proc. of the* 6^{th} *Conf. of the Pacific Association for Computational Linguistics*, pages 22–25, Halifax, Aug 2003.

[35] S.J. Nelson, D. Johnston, and B.L. Humphreys. Relationships in Medical Subject Headings. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 171–184. Kluwer Academic Publishers, New York, 2001.

[36] S.J. Nelson, T. Powell, and B.L. Humphreys. The Unified Medical Language System (UMLS) Project. In A. Kent and C.M. Hall, editors, *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, Inc., New York, 2002.

[37] S. Patwardhan, S. Banerjee, and T. Petersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Intern. Conf. on Intelligent Text Processing and Comutational Linguistics*, pages 17–21, Mexico City, 2003.

[38] E. G.M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In 4^{th} *Workshop on Multimedia Semantics (WMS'06)*, pages 44–52, Chania, Crete, Greece, 1998.

[39] Anand Smith, Barry Kumar and Steffen Schulze-Kremer. Revising the UMLS Semantic Network, in M. Fieschi, et al. (eds.). *Medinfo, IOS Press*, 1700, 2004.

[40] W. Douglas Johnston Stuart J. Nelson and Betsy L. Humphreys. Relationships in Medical Subject Headings (MeSH). In *National Library of Medicine, Bethesda, MD, USA*, 2002.

[41] G. Varelas, E. Voutsakis, P. Raftopoulou, E.G.M. Petrakis, E., and Milios. Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In *Proc. of the* 7^{th} *ACM Intern. Workshop on Web Information and Data Management(WIDM 2005)*, pages 10–16, Bremen, Germany, 2005.

[42] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. KEA: Practical Automatic Keyphrase Extraction. In *Proc. of the* 4^{th} *ACM Conference on Digital Libraries*, pages 254–255, Berkeley, CA, USA, Aug. 1999.

[43] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event Extraction from Biomedical Papers using a Full Parser. In *Proceedings of the sixth Pacific Symposium on Biocomputing (PSB 2001)*, pages 408–419, Hawaii, U.S.A., 2001.

[44] H. Yu, V. Hatzivassiloglou, A. Rzhetsky, and W.J. Wilbur. Automatically Identifying Gene/Protein Yerms in MEDLINE Abstracts. *Journal of Biomedical Informatics*, 35:322–330, 2002.

[45] K. Zervanou and J. McNaught. A Domain-Independent Approach to IE Rule Development. In *Proc. of the* 4^{th} *Intern. Conf. on Language Resources and Evaluation (LREC 2004)*, pages 745–748, Lisbon, Portugal, May 2004.

[46] Y. Zhang, E. Milios, and N. Zincir-Heywood. Narrative Text Classification and Automatic Key Phrase Extraction in Web Document Corpora. In *7th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005)*, pages 51–58, Bremen, German, Nov. 5 2005.

[47] Watters C Zheng W, Milios E. Filtering for medical news items using a machine learning approach. In *Proceedings AMIA Symposium*, pages 949–53, 2002.

[48] Rita D. Zielstorff. Controlled vocabularies for consumer health. *Journal of Biomedical Informatics: Building Nursing Knowledge through Informatics: From Concept Representation to Data Mining*, 36(4-5):326–333, August-October 2003.

Appendix A

A.1 Lucene

Lucene is a free/open source information retrieval library, originally implemented in Java. It is supported by the Apache Software Foundation and is released under the Apache Software License. Lucene has been ported to programming languages including Delphi, Perl, C#, C++, Python, Ruby and PHP.

While suitable for any application which requires full text indexing and searching capability, Lucene has been widely recognized for its utility in the implementation of Internet search engines and local, single-site searching. Lucene itself is just an indexing and search library and does not contain crawling and HTML parsing functionality. The Apache project Nutch is based on Lucene and provides this functionality; the Apache project Solr is a fully-featured search server based on Lucene.

At the core of Lucene's logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene's API to be independent of file format. Text from PDFs, HTML, Microsoft Word documents, as well as many others can all be indexed so long as their textual information can be extracted [25].

A.2 PMC Full Document Retrieval Experiment Queries

1. Menopausal woman without hormone replacement therapy

2. Woman with advanced metastatic breast cancer

3. Woman with back pain
4. Patient with hypothermia
5. Male with pericardial effusion
6. Patient with fever or lymphadenopathy
7. Man with cystic fibrosis
8. Carcinoid tumors of the liver
9. Female with urinary retention
10. Stroke and systolic hypertension
11. Female with lactase deficiency
12. Female some months pregnant
13. Man with sickle cell disease
14. Adult respiratory distress syndrome
15. Young man diabetic

A.3 MeSH DTD File

```
<!-- MESH DTD file for descriptors desc2008.dtd -->

<!ENTITY  % DescriptorReference "(DescriptorUI, DescriptorName)">
<!ENTITY  % normal.date "(Year, Month, Day)">
<!ENTITY  % ConceptReference "(ConceptUI,ConceptName,ConceptUMLSUI?)">
<!ENTITY  % QualifierReference "(QualifierUI, QualifierName)">
<!ENTITY  % TermReference "(TermUI, String)">
<!ELEMENT DescriptorRecordSet (DescriptorRecord*)>
<!ELEMENT DescriptorRecord (%DescriptorReference;,
                            DateCreated,
                            DateRevised?,
                            DateEstablished?,
                            ActiveMeSHYearList,
                            AllowableQualifiersList?,
                            Annotation?,
                            HistoryNote?,
                            OnlineNote?,
                            PublicMeSHNote?,
                            PreviousIndexingList?,
                            EntryCombinationList?,
                            SeeRelatedList?,
                            ConsiderAlso?,
                            RunningHead?,
                            TreeNumberList?,
                            RecordOriginatorsList,
                            ConceptList) >
<!ATTLIST DescriptorRecord DescriptorClass (1 | 2 | 3 | 4)  "1">
<!ELEMENT ActiveMeSHYearList (Year+)> <!ELEMENT
AllowableQualifiersList (AllowableQualifier+) > <!ELEMENT
AllowableQualifier (QualifierReferredTo,Abbreviation )> <!ELEMENT
Annotation (#PCDATA)> <!ELEMENT ConsiderAlso (#PCDATA) > <!ELEMENT
Day (#PCDATA)> <!ELEMENT DescriptorUI (#PCDATA) > <!ELEMENT
DescriptorName (String) >
<!ELEMENT DateCreated (%normal.date;) >
<!ELEMENT DateRevised (%normal.date;) >
<!ELEMENT DateEstablished (%normal.date;) >
<!ELEMENT DescriptorReferredTo (%DescriptorReference;) >
<!ELEMENT EntryCombinationList (EntryCombination+) > <!ELEMENT
EntryCombination     (ECIN,ECOUT)> <!ELEMENT ECIN
(DescriptorReferredTo,QualifierReferredTo) > <!ELEMENT ECOUT
(DescriptorReferredTo,QualifierReferredTo? ) > <!ELEMENT HistoryNote
(#PCDATA)> <!ELEMENT Month (#PCDATA)> <!ELEMENT OnlineNote
```

```
(#PCDATA)> <!ELEMENT PublicMeSHNote (#PCDATA)> <!ELEMENT
PreviousIndexingList(PreviousIndexing)+> <!ELEMENT PreviousIndexing
(#PCDATA) > <!ELEMENT RecordOriginatorsList
                                (RecordOriginator,
                                 RecordMaintainer?,
                                 RecordAuthorizer? )>
<!ELEMENT RecordOriginator (#PCDATA)> <!ELEMENT RecordMaintainer
(#PCDATA)> <!ELEMENT RecordAuthorizer (#PCDATA)> <!ELEMENT
RunningHead (#PCDATA)>
<!ELEMENT QualifierReferredTo (%QualifierReference;) >
<!ELEMENT QualifierUI (#PCDATA) > <!ELEMENT QualifierName (String)>
<!ELEMENT Year (#PCDATA)> <!ELEMENT SeeRelatedList
(SeeRelatedDescriptor+)> <!ELEMENT SeeRelatedDescriptor
(DescriptorReferredTo)> <!ELEMENT TreeNumberList (TreeNumber)+>
<!ELEMENT TreeNumber (#PCDATA)> <!ELEMENT ConceptList (Concept+)>
<!ELEMENT Concept (%ConceptReference;,
                   CASN1Name?,
                   RegistryNumber?,
                   ScopeNote?,
                   SemanticTypeList?,
                   PharmacologicalActionList?,
                   RelatedRegistryNumberList?,
                   ConceptRelationList?,
                   TermList)>
<!ATTLIST Concept PreferredConceptYN (Y | N) #REQUIRED > <!ELEMENT
ConceptUI (#PCDATA)> <!ELEMENT ConceptName (String)> <!ELEMENT
ConceptRelationList (ConceptRelation+)> <!ELEMENT
          ConceptRelation (Concept1UI,
                           Concept2UI,
                           RelationAttribute?)>
<!ATTLIST ConceptRelation RelationName (NRW | BRD | REL) #IMPLIED>
<!ELEMENT Concept1UI (#PCDATA)> <!ELEMENT Concept2UI (#PCDATA)>
<!ELEMENT ConceptUMLSUI (#PCDATA)> <!ELEMENT CASN1Name (#PCDATA)>
```

```
<!ELEMENT PharmacologicalActionList (PharmacologicalAction+)>
<!ELEMENT PharmacologicalAction (DescriptorReferredTo)> <!ELEMENT
RegistryNumber (#PCDATA)> <!ELEMENT RelatedRegistryNumberList
(RelatedRegistryNumber+)> <!ELEMENT RelatedRegistryNumber (#PCDATA)>
<!ELEMENT RelationAttribute (#PCDATA)> <!ELEMENT ScopeNote(#PCDATA)>
<!ELEMENT SemanticTypeList (SemanticType+)> <!ELEMENT SemanticType
(SemanticTypeUI, SemanticTypeName)> <!ELEMENT SemanticTypeUI
(#PCDATA)> <!ELEMENT SemanticTypeName (#PCDATA)> <!ELEMENT TermList
(Term+)>
<!ELEMENT Term (%TermReference;,
                DateCreated?,
                Abbreviation?,
                SortVersion?,
                EntryVersion?,
                ThesaurusIDlist?)>
<!ATTLIST Term    ConceptPreferredTermYN (Y | N) #IMPLIED
                  IsPermutedTermYN (Y | N) #IMPLIED
                  LexicalTag (ABB|ABX|ACR|ACX|EPO|LAB|NAM|NON|TRD) #IMPLIED
                  PrintFlagYN (Y | N) #IMPLIED
                  RecordPreferredTermYN (Y | N)  #IMPLIED>
<!ELEMENT TermUI (#PCDATA)> <!ELEMENT String (#PCDATA)> <!ELEMENT
Abbreviation (#PCDATA)> <!ELEMENT SortVersion (#PCDATA)> <!ELEMENT
EntryVersion (#PCDATA)> <!ELEMENT ThesaurusIDlist(ThesaurusID+)>
<!ELEMENT ThesaurusID (#PCDATA)>
```

A.4 Semantic Network Categories

15 main Categories and 135 Subcategories

Activities & Behaviors

Activity

Behavior

Daily or Recreational Activity

Event

Governmental or Regulatory Activity

Individual Behavior

Machine Activity

Occupational Activity

Social Behavior

Anatomy

Anatomical Structure

Body Location or Region

Body Part, Organ, or Organ Component

Body Space or Junction

Body Substance

Body System

Cell

Cell Component

Embryonic Structure

Fully Formed Anatomical Structure

Tissue

Chemicals & Drugs

Amino Acid, Peptide, or Protein

Antibiotic

Biologically Active Substance

Biomedical or Dental Material

Carbohydrate

Chemical

Chemical Viewed Functionally

Chemical Viewed Structurally

Clinical Drug

Eicosanoid

Element, Ion, or Isotope

Enzyme

Hazardous or Poisonous Substance

Hormone

Immunologic Factor

Indicator, Reagent, or Diagnostic Aid

Inorganic Chemical

Lipid

Neuroreactive Substance or Biogenic Amine

Nucleic Acid, Nucleoside, or Nucleotide

Organic Chemical

Organophosphorus Compound

Pharmacologic Substance

Receptor

Steroid

Vitamin

Concepts & Ideas

Classification

Conceptual Entity

Functional Concept

Group Attribute

Idea or Concept

Intellectual Product

Language

Qualitative Concept

Quantitative Concept

Regulation or Law

Spatial Concept

Temporal Concept

Devices

Drug Delivery Device

Medical Device

Research Device

Disorders

Acquired Abnormality

Anatomical Abnormality

Cell or Molecular Dysfunction

Congenital Abnormality

Disease or Syndrome

Experimental Model of Disease

Finding

Injury or Poisoning

Mental or Behavioral Dysfunction

Neoplastic Process

Pathologic Function

Sign or Symptom

Genes & Molecular Sequences

Amino Acid Sequence

Carbohydrate Sequence

Gene or Genome

Molecular Sequence

Nucleotide Sequence

Geographic Areas

Geographic Area

Living Beings

Age Group

Alga

Amphibian

Animal

Archaeon

Bacterium

Bird

Family Group

Fish

Fungus

Group

Human

Invertebrate

Mammal

Organism

Patient or Disabled Group

Plant

Population Group

Professional or Occupational Group

Reptile

Rickettsia or Chlamydia

Vertebrate

Virus

Objects

Entity

Food

Manufactured Object

Physical Object

Substance

Occupations

Biomedical Occupation or Discipline

Occupation or Discipline

Organizations

Health Care Related Organization

Organization

Professional Society

Self-help or Relief Organization

Phenomena

Biologic Function

Environmental Effect of Humans

Human-caused Phenomenon or Process

Laboratory or Test Result

Natural Phenomenon or Process

Phenomenon or Process

Physiology

Cell Function

Clinical Attribute

Genetic Function

Mental Process

Molecular Function

Organ or Tissue Function

Organism Attribute

Organism Function

Physiologic Function

Procedures

Diagnostic Procedure

Educational Activity

Health Care Activity

Laboratory Procedure

Molecular Biology Research Technique

Research Activity

Therapeutic or Preventive Procedure