Technical University of Crete

Diploma Thesis

Medical Document Classification based on user profile

Author:
Serafeim Koutsos

Supervisor:
Prof. Euripides G.M Petrakis

INTELLIGENT SYSTEMS LABORATORY
Department of Electronic and Computer Engineering

Dissertation Thesis Committee
Professor Euripides G.M Petrakis.
Professor Michalis Zervakis.
Associate Professor Alexandros Potamianos.

July 2013

TECHNICAL UNIVERSITY OF CRETE

Abstract

Intelligent Systems Laboratory
Department of Electronic and Computer Engineering

Undergraduate Student

Medical Document Classification based on user profile

by Serafeim Koutsos

The Internet is used as one of the major sources in health information. The number of related pages available on the Internet almost doubles every year. This is also the case for medical information, which is now available from a variety of sources. Users of the medical domain can be either health care professionals (experts) or consumers (novice users). The use of automated information classification methods is essential for both experts and consumers.In this thesis,we investigate the classification of separate medical documents into two classes;consumers and experts,using machine learning methods and more specifically Decision Tree analysis,multiple criteria decision analysis (MCDA) and readability formulas.The medical documents are represented by terms extracted from AMTEx,a medical document indexing method,MMTx,the method being developed by U.S National Library of Medicine,or the MeSH method,under which documents are indexed by human experts.Decision Trees and MCDA are applied to these term vectors in order to classify medical documents into the aforementioned classes.In this respect,we are trying several readability formulas which are subsequently proven ineffective.The readability formulas usually measure difficulty of writing style instead of difficulty of content.Incorporating MCDA analysis tools,the categorization ability is vastly improved in all of the document representation approaches.

Acknowledgements

I would like to thank Prof.Petrakis for the advice,encouragement and support he provided to me in supervising this thesis. Also,I would like to thank all the members of Intelligent Systems Laboratory and specially Dr.Klio Lakiotaki and Dr.Angelos Hliaoutakis for their collaboration and valuable comments.Most of all,I would like to thank my family and my siblings (George, Manolis and Poppi) for their continuous support.

Contents

Abstract ii

Acknowledgements iii

List of Figures vi

List of Tables vii

Abbreviations viii

1 Introduction 1

2 Background and Related Work 4
- 2.1 Medical Information Systems . . . 4
- 2.2 Data Resources: . . . 6
  - 2.2.1 Medline . . . 6
  - 2.2.2 OHSUMED filtering track collection . . . 6
  - 2.2.3 Unified Medical Language System (UMLS) . . . 9

3 Algorithmic Resources 15
- 3.1 Term Extraction . . . 15
- 3.2 The MMTx Approach . . . 16
- 3.3 The C/NC Value for Term Extraction . . . 19
  - 3.3.1 Linguistic Processing . . . 20
  - 3.3.2 The statistical Part . . . 21
- 3.4 Automatic Term Extraction in Medical Document Collection:The AMTEx Method . . . 23
- 3.5 Decision Tree Classifiers . . . 25
  - 3.5.1 Basics of Decision Trees . . . 25
  - 3.5.2 Splitting Criteria . . . 26
  - 3.5.3 Overfitting and Pruning . . . 29
  - 3.5.4 Complexity of decision tree induction . . . 30
- 3.6 Readability Formulas . . . 31
  - 3.6.1 Formulas used in health care . . . 31
- 3.7 Multicriteria Decision Analysis . . . 33

3.7.1 The UTA method 33

4 Medical Document Classification by User Profile **36**
4.1 Data Retrieval and expert term extraction 36
4.2 Data representation and modeling 39
4.3 Medical Document Classification 41

5 Experiments and Evaluation **42**
5.1 Experimental Setup 42
5.2 Decision Tree 43
5.3 Readability Formulas 45
5.4 Multicriteria Decision Analysis:UTASTAR algorithm 51

6 Conclusions **56**

A Appendix A **57**
A.1 PubMed Health Abstract Retrieval Experiment Queries 57
A.2 MEDLINE/PubMed Data Elemet (Field) Descriptions 59
A.3 Semantic Network Categories 64
A.4 Applying MCDA:Detailed Steps 68

Bibliography **70**

List of Figures

2.1 Semantic Network - Metathesaurus Structure. 13

3.1 MetaMap system diagram. 17
3.2 Variant Generation. 18
3.3 Comparison among Splitting Criteria 28
3.4 Overfitting 29
3.5 Error of Prediction 30

4.1 Document's Expert-Consumer Categorization 37

5.1 QQ-plot Consumer OHSUMED data set 47
5.2 QQ-plot Expert OHSUMED data set 47
5.3 QQ-plot Consumer PubMEd data set 48
5.4 QQ-plot Expert PubMed data set 48
5.5 Histograms Consumer OHSUMED data set 48
5.6 Histograms Expert OHSUMED data set 48
5.7 Histograms Consumer PudMed data set 48
5.8 Histograms Expert PubMed data set 48
5.9 Overlay Histograms Readability Scores-OHSUMED data set 49
5.10 Overlay Histograms Readability Scores-PubMed dataset 50
5.11 ROC curves for AMTEx,MMTx and MeSH approaches. 52

List of Tables

2.1 MedLine Document Structure 7
2.2 MedLine Document Structure 8

3.1 The AMTEx Algorithm 23
3.2 Flesch Reading Ease Formula 32

4.1 Multicriteria Input matrix 40

5.1 OHSUMED Expert-Consumer Evaluation Decision Tree Results 43
5.2 OHSUMED dataset-weights by expert terms.Decision Tree Results 44
5.3 Decision Tree Results Test Set from PubMed 44
5.4 Decision tree Pubmed Data set 45
5.5 Decision tree Four Categories Pubmed Data set 45
5.6 Decision tree Three Categories Pubmed Data set 45
5.7 Statistical Description for each Readability formulas-OHSUMED data set 46
5.8 Statistical Description for each Readability formulas-PubMed data set . . 46
5.9 Shaphiro-Wilk test-OHSUMED data set 47
5.10 Shaphiro-Wilk test-PubMed data set 47
5.11 Decision tree Readability formulas 51
5.12 UTA significance weight terms for AMTEx,MMTx and MeSH 54
5.13 UTA classification evaluation measures for AMTEx,MMTx and MeSH . . 55
5.14 UTA results on PubMed Data set 55

Abbreviations

MEDLINE	**M**edical **L**iterature **A**nalysis and Retrieval **S**ystem Online
NLM	**N**ational **L**ibrary of **M**edicine
UMLS	**U**nified **M**edical **L**anguage **S**ystem
MeSH	**M**edical **S**ubject **H**eadings
AMTEx	**A**utomatic **M**esh **T**erm **E**xtraction
MMTx	**M**eta **M**ap **T**ransfer
MCDA	**M**ultiple **C**riteria **D**ecision **A**nalysis

Dedicated to my family and my friends...

Chapter 1

Introduction

Both medical professionals and laypersons have an interest in current medical information now largely available in electronic form, including electronic newspapers. Many health organizations, including hospitals, universities and government departments, are now providing validated medical information for use. Overall, there has been a growing mass of medical information and news and related data that users of all levels of sophistication can access. One of the most commonly used medical databases is MedLine MedLine[1](Medical Literature Analysis and Retrieval System Online), which constitutes the primary medical repository of the U.S. National Library of Medicine (NLM), including (as of today) approximately 20 million computer-readable records, rapidly expanding. It is a rich resource of medical, biological and biomedical information, requiring efficient management and retrieval, therefore it poses new challenges to information and knowledge management.

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). Generally, 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 using natural language terms. Existing medical information systems such as MedScape[2], MedLIne-Plus[3], Wrappin[4] and MedHunt[5](maintained by HON, the Health on Net Foundation, a

[1]http://www.nlm.nih.gov/pubs/factsheets/medline.html
[2]http://www.medscape.com/
[3]http://www.nlm.nih.gov/medlineplus/
[4]http://www.wrapin.org/
[5]http://www.hon.ch/HONsearch/Patients/medhunt.html

non-profit organization) are capable of providing dedicated, domain specific answers to experts or simple, easily comprehended answers to novice users respectively.

All systems referred to above 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. PubMed[6] of NLM is of particular interest to us. It provides free access to MedLine document abstracts and to articles in selected life sciences journals not included in MedLine.

An automatic system able to characterize medical articles as "consumer specific" or "expert specific" and thus appropriately recommend it could prove highly valuable in increasing the time efficiency of this type of categorization, by reducing and possibly eliminating human interference in this process. Ultimately this automation could lead to timely informed medical databases, assisting consumers in managing their personal health information and experts in significantly reducing their effort on information seeking tasks in the optimal way.

In this thesis, we propose a correspondence method for medical documents based on the terms used in the description of medical information using machine learning techniques (decision tree), readability formulas and Multiple Criteria Decision Analysis. Medical information is typically described by terms belonging to a medical dictionary (such as MeSH), extracted by humans or methods such as AMTEx and MMTx. MMTx is developed at the NLM to map biomedical text to UMLS Metathesaurus concepts and improve retrieval of bibliographic material, such as MedLine citations. Its applications also include semi-automatic and fully automatic indexing, hierarchical indexing and text mining for various medical and biological concepts and relation extractions. AMTEx aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC value method. Additionally, AMTEx aims at improving indexing and retrieval of medical documents, based on the extraction and mapping of document terms to the MeSH Thesaurus.

Readability is considered as one of the quality criteria for health information, and readability formulas are used to measure the reading level of health information in some studies [1]. However, most readability formulas were developed for general educational purposes and only measure one aspect of text difficulty, concentrating on the measurement of difficulty of writing style but not on the difficulty of the content [2]. Thus, readability formulas are not optimal for accurate document categorization.

[6] http://www.ncbi.nlm.nih.gov/pubmed

We studied the classification performance of the three medical document representation methods by applying Decision Tree analysis, Multicriteria decision analysis and readability formulas. Decision trees are a simple but powerful form of multiple attribute analysis and achieve high classification accuracy. Multiple criteria analysis improves the classification accuracy by 7% reaching 90% in the average case and 96% in the best case for the AMTEx method, while MMTx and MeSH perform 56% and 47% respectively.

Related work and resources used in this thesis are discussed in Chapter 2. These include MedLine, the OHSUMED data-set of TREC-9 filtering track collections, PubMed Health database which contains plain language summaries to research papers for naive users, the MeSH, UMLS Metathesaurus and the UMLS Semantic Network. Related work in the field of Multicriteria Decision Analysis is presented as well. In Chapter 3 we present the algorithms used in this thesis. Presented also are approaches to the extraction of medical terminology for indexing purposes such as MMTx, the C/NC-value method and AMTEx, Decision tree analysis, Multicriteria decision analysis and readability formulas. Our method on document categorization, data representations by user profile, is presented in Chapter 4. Finally, Chapter 5 presents the experimental results and evaluation followed by conclusions and issues for future work in Chapter 6.

Chapter 2

Background and Related Work

An overview of medical information systems and medical data sources used in this work,are described in this chapter.An introduction to UTA follows as a part of our work.

2.1 Medical Information Systems

As the volume of medical information on the Web continues to increase,there is growing interest for medical systems which are helping people better find,filter and manage these resources.Many medical information systems (search engines,portals,etc) are currently becoming available,the most important of them being:

Medscape [1] is a free Web site for health professionals and interested consumers.It features peer-reviewed original medical journal articles, CME (Continuing Medical Education), daily medical news, major conference coverage, and drug information.

MEDLINE[2] is the National Library of Medicine (NLM) premier bibliographic database that contains over 20 million references to journal articles in life sciences with a concentration on biomedicine. The MEDLINE database is directly searchable from NLM as a subset of the PubMed database as well as through other numerous search services that license the data. In addition to the comprehensive journal selection process, what sets MEDLINE apart from the rest of PubMed is the added value of using the NLM controlled vocabulary, Medical Subject Headings (MeSH), to index citations.MEDLINE indexers describes the content of the biomedical article by assigning to each one,a number (typically 10 to 12 per article) of MeSH Terms (see section 2.2.3). **Pubmed**[3] is a

[1] http://www.medscape.com/
[2] http://www.nlm.nih.gov/
[3] http://www.ncbi.nlm.nih.gov/pubmed

free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM).Pubmed has been available since 1996. Its over 22 million references include the MEDLINE database plus the following types of citations:in-process citations,citations to articles that are out of scope.MEDLINE is the largest database of PubMed. Pubmed uses the MeSH terms for retrieval and the search strategy is enhanced (e.g the query "bad breath" is mapped automatically to the MeSH term "halitosis").

For consumer users,the National Library of Medicine (NLM) provides **MedlinePlus**[4] and **Pubmed Health**[5].MedlinePlus is the site for patients.It brings information for diseases,conditions and wellness issues in a language that common people can understand.It includes an extensive Health Topics section,Drug information ,medical dictionary,health news,interactive health tutorials and more.Respectively,Pubmed Health provides information for consumers and clinicians on prevention and treatment of diseases and conditions.Generally,it specializes in reviews of clinical effectiveness research, with easy-to-read summaries for consumers as well as full technical reports.The reviews were generally published or updated from 2003. There is also information for consumers and clinicians based on those reviews.

Also there exist medical search engines used by both expert and consumers.Some of them are listed below:**MedHunt**,developed and maintained by the Health On Net Foundation(HoN)[6].MedHunt[7] retrieves medical information either from HoNs accredited sites or from medical pages crawled from the Web (such as MEDLINE,Mayo Clinic[8],U.S. Food and Drug Administration[9]). HoN is a not-for-profit organization founded in 1995 under the auspices of the Geneva Ministry of Health and based in Geneva, Switzerland and aims to provide access to reliable sources of medical information. HoN has become one of the most respected not-for-profit portals for medical information on the Internet.HON co-operates closely with the University Hospitals of Geneva and the Swiss Institute of Bioinformatics are two widely-used medical search tools, MedHunt,HONselect and the HON Code of Conduct (HONcode) for the provision of authoritative,trustworthy Web-based medical information.

WRAPIN[10](Worldwide online Reliable Advice to Patients and Individuals) uses medical trustworthy sources (NLMs Pubmed,HoNs MedHunt and US Food and Drug Administration),supports different types of query from a few keywords to entire web pages(specified

[4]http://www.nlm.nih.gov/medlineplus/
[5]http://www.ncbi.nlm.nih.gov/pubmedhealth/
[6]http://www.hon.ch/
[7]http://www.hon.ch/HONsearch/Patients/medhunt.html
[8]http://www.mayoclinic.com/
[9]http://www.fda.gov/
[10]http://www.wrapin.org/

by their URL). It maps both query and documents to MeSH Terms (the HoNMeSHMapper module is used) subsequently used for indexing and retrieval. Also, **Medworm**[11] is a medical RSS feed provider as well as a search engine built on data collected from RSS feeds. MedWorm collects updates from thousands of authoritative data sources via RSS feeds. From the data collected,MedWorm provides new outgoing RSS feeds on various medical categories that the user can subscribe to.

2.2 Data Resources:

Automatic Indexing and categorization of medical documents relies mainly on term extraction from large medical document collection,like MedLine and the UMLS Knowledge Sources such as MeSH,a subset of the *UMLS Metathesaurus* and the *Semantic Network*. Experimental results here based on two main medical document collections,the *OHSUMED* collection and the PubMed Health 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. It now provides over 20 million references to biomedical and life sciences journal articles back to 1946 (in MedLine 2013). The articles stored in MedLine have both Descriptive and Semantic Metadata. So,MedLine's documents have more information than a simple article reference. There are 81 tags providing information for each document in MedLine,usually NLM indexers using a range 35-40 tags (see Appendix A.2). The tables 2.1 and 2.2 below show an example from a random MedLine document more specific.

2.2.2 OHSUMED filtering track collection

The OHSUMED test collection is a set of 348,566 references from MEDLINE, the online 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

[11] http://www.medworm.com

collection does not only include documents,but also topics(queries) and relevance judgments.some abstracts are truncated at 250 words and some references have no abstracts at all (titles only). There is no access to the full text of the documents.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 experiment described in [3] and [4].

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)

Table 2.1: MedLine Document Structure

PMID	Unique number assigned to each PubMed citation
OWN	Organization acronym that supplied citation data
STAT	Status Tag
DA	Date Created
DCOM	Completion Date
IS	ISSN -International Standard Serial Number of the journal
VI	Volume number of the journal
IP	The number of the issue in which the article was published
DP	Publication Date
TI	The title of the article
PG	The full pagination of the article
AB	English language abstract taken directly from the published article
AD	Institutional affiliation and address of the first author
FAU	Full Author Name
AU	Authors
LA	The language in which the article was published
PT	The type of material the article represents
PL	Journal's (country only) or books place of publication
TA	Standard journal title abbreviation
JT	Full journal title from NLM cataloging data
JID	Unique journal ID in the NLM catalog of books, journals, and audiovisuals
SB	Journal or citation subset values representing specialized topics
MH	NLM Medical Subject Headings (MeSH) controlled vocabulary
EDAT	The date the citation was added to PubMed
MHDA	The date MeSH terms were added to the citation
CRDT	The date the citation record was first created
PST	Publication status
SO	Composite field containing bibliographic information

.S Source (SO)

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 contains 106 queries that were generated by actual physicians in the course of patient care.Only a subset of 63 of these queries were used in the TREC-9 filtering track.Before they searched,they were asked to provide an information about their patient as well as about their information need.

Table 2.2: MedLine Document Structure

PMID	23488026
OWN	NLM
STAT	MEDLINE
DA	20130315
DCOM	20130401
IS	0029-6570 (Print)
IS	0029-6570 (Linking)
VI	27
IP	14
DP	2012 Dec 5-11
TI	Back pain:pathogenesis,diagnosis and management.
PG	49-56; quiz 58
AB	Back pain is a common problem that may have physical and psychological ...
AD	Nottingham University Hospital, Nottingham. jennie.walker@nottingham.ac.uk
FAU	Walker, Jennie
AU	Walker J
LA	eng
PT	Journal Article
PL	England
TA	Nurs Stand
JT	Nursing standard (Royal College of Nursing (Great Britain) : 1987)
JID	9012906
SB	N
MH	Back Pain/diagnosis/etiology/nursing
MH	Education, Nursing, Continuing
MH	Great Britain/epidemiology
MH	Humans
MH	Prevalence
EDAT	2013/03/16 06:00
MHDA	2013/04/02 06:00
CRDT	2013/03/16 06:00
PST	ppublish
SO	Nurs Stand. 2012 Dec 5-11;27(14):49-56; quiz 58.

NLM has agreed to make the MedLine references in the test database available for experimentation,restricted to the following conditions:

- The data will not be used in any non-experimental clinical,library or other setting.
- Any human users of the data will explicitly be told that the data is incomplete and out-of-date.

2.2.3 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS)[12] is a comprehensive list of biomedical terms for developing computer systems capable of understanding the specialized vocabulary used in biomedicine and health care. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs). Developers use the Knowledge Sources and tools to build or enhance systems that create, process, retrieve, and integrate biomedical and health data and information. The Knowledge Sources are multi-purpose and are used in systems that perform diverse functions involving information types such as patient records, scientific literature, guidelines, and public health data. The associated software tools assist developers in customizing or using the UMLS Knowledge Sources for particular purposes. The Lexical Tools work more effectively in combination with the UMLS Knowledge Sources, but can also be used independently. There are three UMLS Knowledge Sources:the *Metathesaurus*,the *Semantic Network* and the *SPECIALIST Lexicon* [5] .

UMLS Metathesaurus

The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts, their various names, and the relationships among them. It is built from the electronic versions of numerous thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing biomedical literature, and/or basic, clinical, and health services research. In this documentation, these are referred to as the "source vocabularies" of the Metathesaurus. In the Metathesaurus, all the source vocabularies are available in a common, fully-specified database format.

The Metathesaurus is organized by concept or meaning. In essence, it links alternative names and views of the same concept and identifies useful relationships between different concepts. All concepts in the Metathesaurus are assigned at least one Semantic

[12] http://www.nlm.nih.gov/research/umls/

Type from the Semantic Network to provide consistent categorization at the relatively general level represented in the Semantic Network. Many of the words and multi-word terms that appear in concept names or strings in the Metathesaurus also appear in the SPECIALIST Lexicon. The Lexical Tools are used to generate the word, normalized word, and normalized string indexes to the Metathesaurus. MetamorphoSys is used to install the UMLS Knowledge Sources and customize the Metathesaurus.

The scope of the Metathesaurus is determined by the combined scope of its source vocabularies. Many relationships (primarily synonymous), concept attributes, and some concept names are added by the NLM during Metathesaurus creation and maintenance, but essentially all the concepts themselves come from one or more of the source vocabularies. Generally, if a concept does not appear in any of the source vocabularies, it will also not appear in the Metathesaurus.

In particular,it contains information million biomedical concepts and 5 million concept names, all of which stem from the over 100 incorporated controlled vocabularies and classification systems.Some examples of the incorporated controlled vocabularies are ICD-10, MeSH, SNOMED CT, DSM-IV, LOINC, WHO Adverse Drug Reaction Terminology, UK Clinical Terms, RxNorm, Gene Ontology, and OMIM

MeSH:Medical Subject Headings

MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.Those terms represent a subset of the UMLS metathesaurus.NLM has adopted the Extensible Markup Laanguage(XML)[13] as the description language for MeSH.The MeSH vocabulary file is available in an XML format.MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders." More specific headings are found at more narrow levels of the twelve-level hierarchy, such as "Ankle" and "Conduct Disorder." There are 26,853 descriptors in 2013 MeSH. There are also over 213.000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid." In addition to these headings, there are more than 214.000 headings called Supplementary Concept Records (formerly Supplementary Chemical Records) within a separate thesaurus.

MeSH descriptors are organized in 16 categories,of ISA kind relationship between nodes (concepts): category A for anatomic terms, category B for organisms, C for diseases, D for drugs and chemicals, etc. Each category is further divided into subcategories.

[13]http://www.w3.org/TR/REC-xml/

Within each subcategory, descriptors are arrayed hierarchically from most general to most specific in up to twelve hierarchical levels. These trees should not be regarded as representing an authoritative subject classification system but rather as arrangements of descriptors for the guidance and convenience of persons who are assigning subject headings to documents or are searching for literature. The trees are not an exhaustive classification of the subject matter but contain only those terms that have been selected for inclusion in this thesaurus. Their structure frequently represents a compromise among the views and needs of particular disciplines and users, in the absence of any single universally accepted arrangement.The categories are:

1. Anatomy[A]

2. Organisms[B]

3. Diseases [C]

4. Chemical and Drugs [D]

5. Analytical,Diagnostic and Therapeutic Techniques and Equipment [E]

6. Psychiatry and Psychology [F]

7. Phenomena and Processes [G]

8. Disciplines and Occupations [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]

Mesh concepts corresponds to MeSH objects which are described with terms of several properties,the most importand of them being:

- **MeSH Headings(MH):**
  These are term names or identifiers.This is the term used in the MEDLINE database as the indexing term. Every document in MedLine have some MeSH terms that are indexed with.The term reflects a meaning; its use indicates the topics discussed by the work cited.

- **Entry Terms:**
  These terms are used as pointers to the MH,they are considered to be synonymous, or close in scope to the MeSH term.The presence of an entry term in the record is an indication that this topic should be indexed by the given MH. The set of entry terms that points to a MH are the terms that represent the concept introduced by the MH.PubMed recognizes these terms so that, if we search with an Entry Term, the appropriate MeSH term will be included in the search.So,an admission is made that all entry terms are synonymous with the MH.

- **MeSH Tree Number:**
  The tree numbers indicate the places within the MeSH hierarchies, also known as the Tree Structures, in which the MH appears. Thus, the numbers are the formal computable representation of the hierarchical relationships.For example D is the code name of the "Chemical and Drugs" subtree and the term "Lipids" has a tree number D10,meaning that "Lipids" belongs to D subtree.

- **MeSH Scope Note:**
  This short piece of free text provides a type of definition, in which the meaning of the MH is circumscribed. Other MHs frequently appear in scope notes, usually in ALL CAPS. These represent relationships, which are often very important, but which may not otherwise be represented in the MeSH structure.

Main Headings (descriptor records) are distinct in meaning from other Main Headings in the thesaurus (i.e their meaning 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 clinicians,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 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 133 semantic types and 54 relationships [5](see Appendix A.3).

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 2.1). In all cases, the most specific semantic type available in the hierarchy is assigned to the concept.

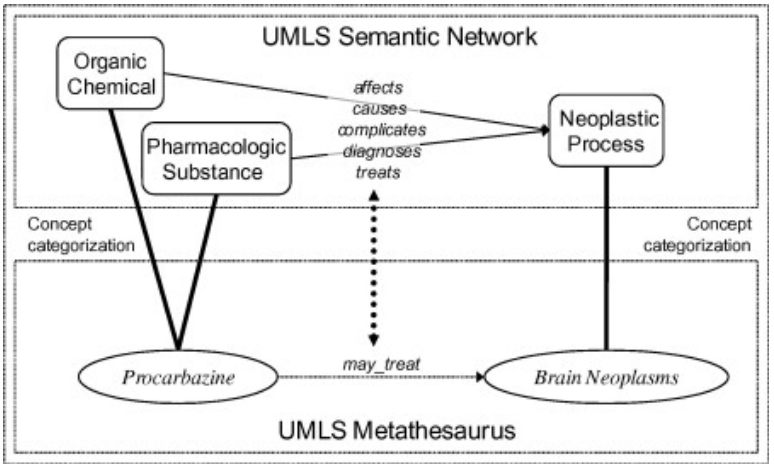


FIGURE 2.1: Semantic Network - Metathesaurus Structure.

Results in [6] 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 the relation "Body Substance surrounds Embryonic Structure" is represented in the Network [5] . 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.

SPECIALIST Lexicon

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

Chapter 3

Algorithmic Resources

Algorithmic Resources such as term extraction methods,decision tree classifiers,readability formulas and multicriteria decision analysis considered in this work are presented in this chapter.

3.1 Term Extraction

Term Extraction relates to identifying the most characteristic or important terms in a corpus.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 destination and classification purposes [7] .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 principal informational content of a document.In early approaches,terms have been sought for indexing purposes,using mostly *tf idf* [8].Term extraction approaches largely rely on the identification of term formation patterns (e.g [9–11]).Statistical techniques may also be applied to measure the degree of unithood or termhood of the candidate multi-word terms [12].Later and current approaches tend to follow a hybrid approach combining both statistical and linguistic techniques (e.g. [13–15]).

The extraction of terms for the medical,biological and biomedical domain has greatly motivated research for both indexing,as well as knowledge extraction purposes [11, 16, 17].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 idf* method [8].In

traditional indexing techniques,query and document representations 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 [14] 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 [10] .For example,the compound term *"heart disease"* denotes a specific type of disease.A study by Milios et al.[18] of the extraction of multi-word terms for retrieval purposes shows that multi-word term methods may complement other methods to improve results.Currently,machine learning techniques are also applied for indexing,such as the Naive Bayes learning model implementation in the KEA (Automatic Keyphrase Extraction [19]).Comparative experiments of *tf idf*,KEA and the C/NC value term extraction methods by Zhang et al.[20] show that C/NC value significantly outperforms both *tf idf* and KEA in a narrative text classification task using extracted terms.

3.2 The MMTx Approach

MetaMap is widely available program providing access to the concepts of the *UMLS Metathesaurus* (see section 2.2.3,page 9) from biomedical text.MetaMap arose in the context of an effort to improve biomedical text retrieval,specifically the retrieval of MEDLINE/PubMed citations [21].It provided a link between the text of biomedical literature and the knowledge,including synonymy relationships,embedded in the Metathesaurus.A system diagram showing MetaMap processing is shown in figure 3.1.The following steps are followed:

1. **Lexical/Syntactic Analysis**
   In MetaMap, input text undergoes a lexical/syntactic analysis consisting of a first analysis in which tokens, sentence boundaries and acronyms or abbreviations are identified and each token is assigned to a part of speech tagger [22] . Input words are mapped to the SPECIALIST lexicon using lexical lookup and then the SPECIALIST minimal commitment parser identifies phrases and their lexical heads.The SPECIALIST minimal commitment parser produces a high-level syntactic analysis.The parser optionally uses the Xerox Part-of-speech tagger which assigns syntactic labels to all textual items.The parser is very good at determining the simple noun phrases in text;and the errors it does make are normally inconsequential to MetaMap [23].The tagger also improves parsing results.For example,the term *"ocular complications"* is analyzed as:

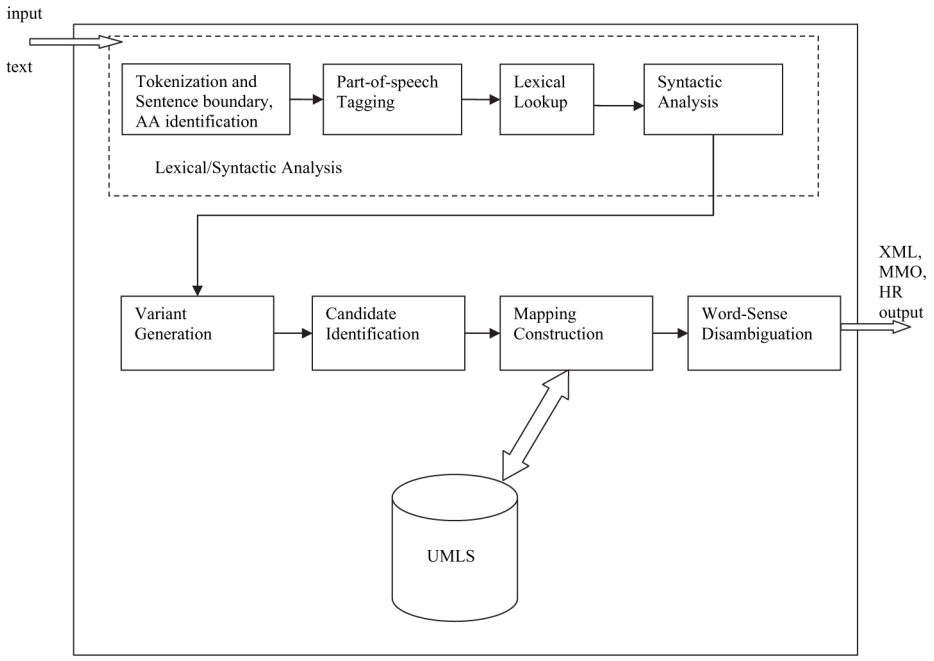


FIGURE 3.1: MetaMap system diagram.

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

where *complications* is the head,namely the term that is being modified/specialized and *ocular* is the modifier,namely the concept specializing the term complications.Each phase found by this analysis is further analyzed by the following process:

2. **Variant Generation**
   Variants of all phrase words are determined.Variant generation is performed in 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 [24].The approach taken in computing variants is a canonicalization approach .This simply means that a variant represents not only itself but all of its inflectional and spelling variants [1].The computation for each generator procceeds as follows:

   i. Compute all acronyms,abbreviations and synonyms of the generator.This results in the three sets Generator,Acronyms/Abbreviations and Synonyms (figure 3.2)

[1] A spelling variant of a word is just a variant having the same principal part as the word.For example,*haemorrhaged* is a spelling variant of *hemorrhaged*

ii. Augment the elements of the three sets by computing their derivational variants and the synonyms of the derivational variants.

iii. For each member of the Acronyms/Abbreviations set,compute synonyms;and

iv. For each member of the Synonyms set,compute acronyms/abbreviations.

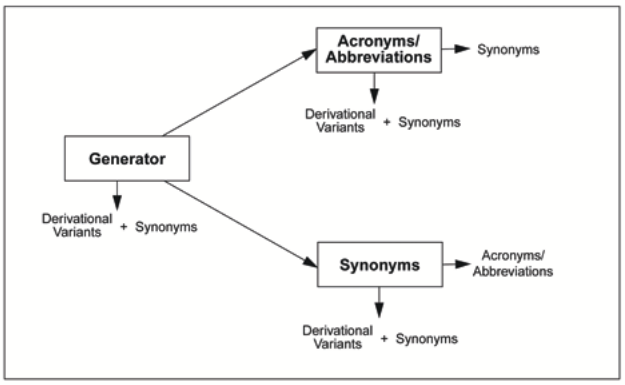


FIGURE 3.2: Variant Generation.

Note that,acronyms and abbreviations are not recursively generated since doing so,almost always produces incorrect results.Derivational variants and synonyms are recursively generated since this often produces meaningful variants.For example,the variant generators for the noun phase of *"liquid crystal thermography"* are *"liquid crystal thermography"*,*"liquid crystal"*,*"liquid"*,*"crystal"* and *"thermography"*.(prepositions,determiners,conjunctions,auxiliaries,modals,pronouns and punctuation are ignored)

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 [24].The mapping process may vary [23].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 *applications* maps to *Job Applications,Heat/Cold Application* and *Medical Information Application.*

4. **Candidate Evaluation**
   The evaluation performed on both the candidates and final mappings is a linear combination of four linguistically measures,listed below.The evaluation process begins by focusing on the association,or mapping,of input text words to words of the candidates.The four linguistic measures are [25]:

   - **Centrality:**is a Boolean value which is 1 if the string involves the head of the phrase and 0 otherwise.
   - **Variation:**is the average of the variation between all text words and their matching candidate words.It estimates how much the variants in the Meta string differ form the corresponding words in the phrase.
   - **Coverage:**indicates how much of the Meta string and the phrase are involved in the match.
   - **Cohesiveness:**is similar to coverage value but emphasizes the importance of maximal sequence of continuous words participating in the match. Coverage and cohesiveness measure how much of the input text is involved in the mapping (coverage) and in how many chunks of contiguous text(cohesiveness).

   The four measures are combined linearly giving coverage and cohesiveness twice the weight of centrality and variation,and the result is normalized to a value between 0 and 1000.

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

- By default MMTx extracts general Metathesaurus terms,not just MeSH terms.
- Term Selection is based on a scoring function(for evaluating the importance of all candidate terms) using 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.
- 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.

3.3 The C/NC Value for Term Extraction

The C/NC value method is a hybrid domain independent method combining linguistic and statistical (with emphasis on the statistical part) for the extraction of multi-word

and nested terms (i.e. terms that appear within other longer terms,and may or may not appear by themselves in the corpus).This method takes as input a corpus and produces a list of candidate multi-word terms,ordered by the likelihood of being valid terms,namely their C-value measure.

The C/NC value method comprises of two main parts:

i. The linguistic.

ii. The statistical,which consist of:

   a. C-value.

   b. NC-value.
   NC-value is an enhancement to C-value.It incorporates contextual information,aiming at improving the ranking of candidate multi-word term list extracted by C-value.

3.3.1 Linguistic Processing

The Linguistic processing applies the following parts:

- **Part of Speech (POS) Tagging of the corpus:** Part of Speech (POS) Tagging is the process of assigning a grammatical category tag (such as noun,verb,adjective,adverb or preposition) to each word.In C/NC value,POS is applied prior to linguistic filters that exact noun phrases.

- **The Linguistic Filter:** Terms consists mostly of noun and adjectives [26] and sometimes prepositions [27].The statistical information,without any linguistic filtering,is not enough to produce useful results.Without any linguistic information,undesirable strings such as *"of the","is a","etc"*,would also be extracted.The linguistic filter is used to extract noun phrases that constitute multi-word terms, discarding such undesirable strings.The three filters available are:

$$\mathrm{N}^{+}N$$
$$(\mathrm{Adj}|\ N) + N$$
$$((\mathrm{Adj}|\ N) + ((Adj \mid N)^{*}(NP)?)(Adj \mid N)^{*})N$$

  where N is a noun,Adj is an adjective and P stands for preposition.

- **The Stop-List:**
  A stop-list is a list of words,which are not expected to occur as term words in

that domain.It is used to avoid the extraction of strings that are unlikely to be terms,improving the precision of the output list.The stop list is manually constructed based on domain observation.

3.3.2 The statistical Part

1. **C-value**

   The C-value constitutes a measure of the importance of each candidate term extracted in the previous steps.The higher the C-value measure,the most likely it is the candidate term to be a valid term. The C-value of a term is computed as follows:

$$C_{value} = \begin{cases} \log_2 |a| \times f_(a) & a \text{ is non-nested term,} \\ \log_2 |a| \times (f_a - \dfrac{1}{P(T_a)} \sum_{b \in T_a} f(b) & a \text{ is a nested term.} \end{cases} \tag{3.1}$$

   where:

   - $\boldsymbol{a}$ is the candidate string.
   - $\boldsymbol{f(.)}$ is the frequency of occurrence of this term in the corpus.
   - $\boldsymbol{T_a}$ denotes the set of extracted terms that contain a.
   - $\boldsymbol{P(T_a)}$ is the number of these terms.

   The negative effect on the candidate string a being a substring of other longer candidate terms is reflected by the negative sign '-' in the formula above.The independence of a from this longer candidate terms is given by P_{T_a}.The greater this number,the bigger its independence (and the opposite) is reflected by having P_{T_a} as the denominator of a negatively signed fraction.The measure is built using several statistical characteristics of the candidate string.These are:

   i. The total frequency of occurrence of the candidate string in the corpus.
   ii. The frequency of candidate string as part of other longer candidate terms.
   iii. The number of these longer candidate terms.
   iv. The length of the candidate string (in number of words).

   The higher the number of distinct longer terms that our string appears as nested in,the more certain we can be about its independence(i.e. that the candidate term extracted is a real term).The fact that a longer string appears X times is more important,than that of a shorter string appear again X times [28].

2. **NC-Value**

   NC-Value is an enhancement in C-Value that is computed based on context information.

   Firstly,NC-Value creates a list of important term context words.Term context words are words that appears in the vicinity of terms in texts.These will be ranked according to their *"importance"* when appearing with terms.The criterion for the extraction of a word as term context word is the number of terms it appears with.The higher this number is,the higher the likelihood that the word is *"related"* to terms(it occurs with other terms in the same corpus).

   Each candidate term in the C-Value list appears in the corpus with a set of context words. From these context words, the nouns, adjectives and verbs are retained for each candidate term. NC-Value provides a method for the extraction of term context words (words that tend to appear with terms) and incorporates this information (from term context words) into the term extraction process. This above criterion is more formally expressed as [28]:

$$weight(w) = \frac{t(w)}{n} \tag{3.2}$$

   where:

   - $\boldsymbol{w}$ is the context word (noun,verb or adjective).
   - $\boldsymbol{weight(w)}$ the assigned weight to the word w.
   - $\boldsymbol{t(w)}$ the number of terms the word w appears with.
   - $\boldsymbol{n}$ is the number of all terms.

   The purpose of the denominator n is to express this weight as a probability (the probability that the word w might be a term context word). The NC-value measure is then computed as :

$$NC_{Value}(a) = 0.8 \times C - value(a) + 0.2 \times \sum_{w \in C_a} f_a(w) \times weight(w) \tag{3.3}$$

   where:

   - $\boldsymbol{a}$ is the candidate term.
   - $\boldsymbol{C_a}$ is the set of distinct context words of term a.
   - $\boldsymbol{w}$ is a context word in C_a.
   - $\boldsymbol{weight(w)}$ is the weight of w as a term context word of term a.
   - $\boldsymbol{f_a(w)}$ is the frequency of w as context word of a.

The two factors of NC-value, i.e. C-value and the context information factor, have been assigned the weights 0.8 and 0.2 respectively. These have been chosen among others after experiments and comparisons of the results [28].

C/NC-value has been successfully tested in various domains, such as molecular biology (nuclear receptors [29]), eye pathology medical records [28], biomedical business newswire texts [17] and computer science papers [30].

3.4 Automatic Term Extraction in Medical Document Collection:The AMTEx Method

AMTEx is a medical document indexing method,specifically designed for automatic indexing of documents in large medical collection,such as MEDLINE.AMTEx combines MeSH (the terminological thesaurus resource of NLM,section 2.2.3 at page 9) with a well-established method for extraction of terminology,the C/NC-value method(3.3)

Based on the observations we made on the MMTx algorithm at section 3.2 at page 16,we proposed two basic changes towards the development of an improved term extraction method that could substitute MMTx:

i. Term extraction based on a well-established method,the C/NC value.

ii. Use of MeSH Thesaurus as lexical resource,both for (limited) term variant retrieval and candidate term mapping.

Input: Document d_i, MeSH Thesaurus.
Output: MeSH terms t.
1. *Multi-word Term Extraction:*C/NC value method.
2. *Term Ranking:*NC value ranking (3.3)
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.

Table 3.1: The AMTEx Algorithm

The AMTEx Algorithm

An outline of the AMTEx procedure is illustrated in table 3.4.In particular the AMTEx method has the following processing stages:

1. *Multi-word Term Extraction:*The C/NC value method (section 3.3) is used for term extraction.This method is domain independent,does not require any lexical resources and has been proven to be particularly effective in multi-word

and nested term extraction both in medical and general document collections.During term extraction in AMTEx,the document text is annotated for part-of-speech tagger and linguistic filters.

2. *Term Ranking:*Extracted candidate terms are ordered,first by C-value and subsequently by NC-value score.The final candidate term list is ranked by decreasing term likelihood (3.3).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 applying simple string matching).The list of terms now contains only MeSH terms.

4. *Single-word Term Extraction:*For this 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,this is also added to the candidate term list, retaining its original C/NC ranking value.

5. *Term Variants:*Term variants are included in the candidate term list.The C/NC-value implementation in *AMTEx* includes inflectional variants of the extracted terms.Also,MeSH itself can be used for locating variant terms,based on the MeSH term,Entry Terms property.However,only the stemmed term-forms are used in *AMTEx* since the full list of Entry Terms may contain terms,which often are not synonymous.

6. *Term Expansion:*The list of terms is augmented with semantically (conceptually) similar terms from MeSH,figure X,illustrates this process:A term is represented by its MeSH tree hierarchy.The neighborhood of the term is examined and all terms with similarity greater than threshold $T_{expansion}$ are also included in the query vector.This expansion may include terms more than one level higher or lower than original term depending on the value of $T_{expansion}$

AMTEx 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 AMTEx 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 AMTEx v2 presented here,the estimated head of multi-word term is successfully used for the refinement of Single-word Term Extraction process.

Our approach to *Term Variant* generation is more limited than MMTx.This constrains our term recall to terms that are closer to the original term in text.The

AMTEx 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 AMTEx for discovering semantically similar terms,is based on the semantic similarity method by Li et al. [31].The evaluation of the semantic similarity methods indicated that this method is particularly effective,achieving up to 73% correlation with results obtained by humans [30].An important observation and a desirable property of this method is that it tends to assign higher similarity to terms which are closer 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 $T_{expansion}$ 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 taxonomy).Figure X,illustrates this process for various values of the threshold $T_{expansion}$.

Because no synonymy relation is defined in MeSH,we did not apply expansion to the Entry Terms of terms.Word sense disambiguation [32] can also be applied for detecting the correct sense to expand(here,expansion is applied to the most common sense of each term).

3.5 Decision Tree Classifiers

Decision tree is a classification scheme which generates a tree and a set of rules from a given data set.

3.5.1 Basics of Decision Trees

A *"divide-and-conquer"* approach to the problem of learning from a set of independent instances leads naturally to a style of representation called decision tree.Nodes in a decision tree involve testing a particular attribute.Usually,the test at a node compares an attribute value with a constant.However,some trees compare two attributes with each other,or utilize some functions of one or more attributes.Leaf nodes give a classification that applies to all instances that reach the leaf,or a set of classifications,or a probability distribution over all possible classifications [33].

The task of constructing a tree from the training set has been called tree induction,tree building and tree growing.Most existing induction tree algorithms proceed

a top down fashion.Starting with an empty tree and the entire training set,some variants of the following algorithm is applied until no more splits are possible [34].

1. If all the training examples at the current node t belong to category c,create a leaf node with the class c.
2. Otherwise ,score each one of the set of possible splits S,using a *"goodness measure"*.
3. Choose the best split s^* as the test at the current node.
4. Create as many child nodes as there are distinct outcomes of s^*.Label edges between the parent and child nodes with outcomes of s^*,and partition the training data using s^* into the child nodes.
5. A child node t is said to be pure if all the training samples at t belong to the same class.Repeat all the previous steps on all impure child nodes.

An object X is classified by passing it through the tree starting at the root node. The test at each internal node along the path is applied to the attributes of X, to determine the next arc along which X should go down. The label at the leaf node at which X ends up is output as its classification.An object is *misclassified* by a tree if the classification output by the tree is not the same as the object's correct class label.The proportion of objects correctly classified by a decision tree is known as its *accuracy*,whereas the proportion of misclassified object is the *error*.

3.5.2 Splitting Criteria

Most common algorithms (ID3,C4.5 and other),learn decision tree by constructing them top-down,beginning with the question *"which attribute should be tested at the root of the tree"*?.To answer this question,each instance attribute is evaluated using a statistical test to determine how well it alone classifies the training examples. *"The best split"* is defined by how well the variable splits the set into homogeneous subsets that have the same value of the target variable. Different algorithms use different statistical test (formula) for measuring "best". Below,we present a few of the most common formula. These formula are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split.

i. **Gini Index**
   Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.To compute Gini impurity for a set of items, suppose

i takes on values in $1, 2, ..., m$, and let p_i,the fraction of items belonging to class i at a given node t [35].

$$GINI(t) = 1 - \sum_i [p(i|t)]^2$$

- Maximum $(1 - \frac{1}{n_c})$ when records are equally distributed among all classes implying least interesting information.
- Minimum (0.0) when all records belong to one class,implying most interesting information.

When a node t is split into k partitions (children),the quality of the split is computed as:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where n_i is the number of records at child i,and n is the number of records at node t.

ii. **Entropy**

Measures the homogeneity of a node.The entropy at a given note t is:

$$Entropy(t) = -\sum_i p(i|t) \log_2 p(i|t)$$

where $p(i|t)$ denotes the fraction of records belonging to a class i,at a given node t.

- Maximum $\log n_c$ when records are equally distributed among all classes, implying least information.
- Minimum (0.0) when all records belong to one class,implying most information.

Information gain measures how well a given attribute separates the training examples according to their target classification using entropy measure impurity.

$$Info_GAIN_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i))$$

where parent node p is split into k partitions,and n_i is the number of records in partition i.We compare the degree of impurity of the parent node *before splitting* with the degree of impurity of the child node *after splitting*.The larger the difference,the better the condition(maximize GAIN).This used by ID3 and C4.5 algorithms.The disadvantage of this approach is that tends to prefer splits that result in large number of partitions,each being small but

pure. In order to avoid this disadvantage,the above algorithms use Gain Ratio.Gain Ratio solves this problem;it adjusts the information gain by the entropy of the partitioning and this in effect penalizes the cases that have a large number of small partitions.

$$GainRATIO = \frac{GAIN_{split}}{SplitINFO}$$

where

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

iii. **Classification Error**

Classification error measures the misclassification error made by a node.

$$Error(t) = 1 - max_i P(i|t)$$

- The maximum value $1 - \frac{1}{n_c}$ occurs when there is an equal number of records distributed among all the classes,thus giving the least information.
- The minimum value (0.0) occurs when all records belong to one class implying the most information.

The comparison between the above splitting criteria is shown in figure 3.3.

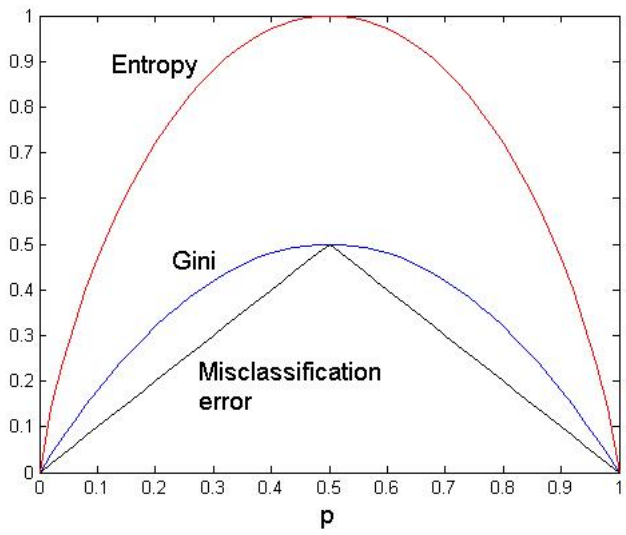


FIGURE 3.3: Comparison among Splitting Criteria.
For a 2-class problem.

3.5.3 Overfitting and Pruning

Real data are commonly affected by *noise* arising from misclassification or incorrect measurement or recording of attribute values(figure 3.4).These cause the divide-and-conquer algorithm to generate complex trees that attempt to model the discrepancies.Overfitting and Underfitting are the two main problems during the construction of the tree [36]. Model Overfitting occurs when the DT (in an

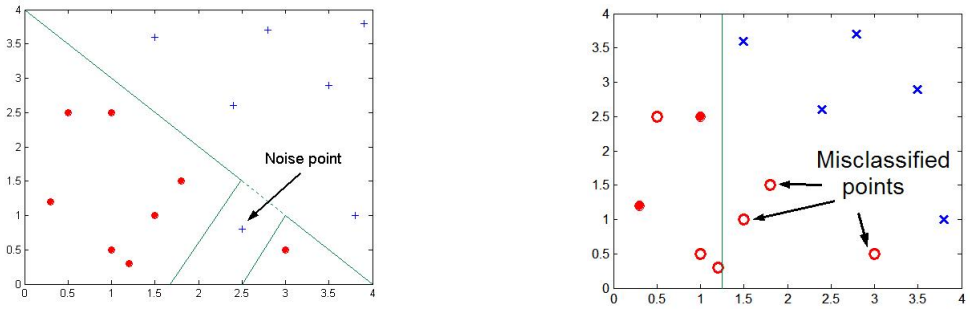


FIGURE 3.4: Overfitting due to Noise and Insufficient Example

attempt to reduce the training errors that have been mentioned before) has a lot of nodes that fits the training data extremely well,but this in return creates a high generalization error.Generalization error are generated by using test or previously unseen records on the new model that has been developed.On the other hand,model underfitting occurs when the algorithm performs too much of the generalization and oversimplifies the model,usually due to the lack of training data.As a result,the model is too simplistic to pick up on the important features of the unseen data,and may work well on small training sets but as the amount of training data increases,its performance suffers because it underfits the data.

Such overfitting is usually addressed by *pruning* the initial tree [37].The *pruning* technique,identifies subtrees that contribute little to predictive accuracy and replacing each by a leaf.There are two strategies of pruning;the postpruning and prepruning.Prepruning would involve trying to decide during the tree building process when to stop developing subtrees.Postpruning doing the prunning after the tree has been constructed.There are two types of postpruning that are generally used:subtree replacement and subtree raising.Subtree replacement goes back to the tree once it is created and attempts to remove branches by replacing them with leaf nodes.In subtree raising [2],a node may be moved upwards towards the root of the tree,replacing other nodes along the way.Pruning techniques are based on cost-complexity models,pessimistic accuracy estimates,or minimizing the length

[2]Subtree raising is more complex,takes more time and it is not clear that it is always worthwhile.Although,it is used by C4.5.

of a message describing the tree and the data [33, 35].The optimal prediction is obtained, when the remaining interference error and the estimation error balance each other (figure 3.5).

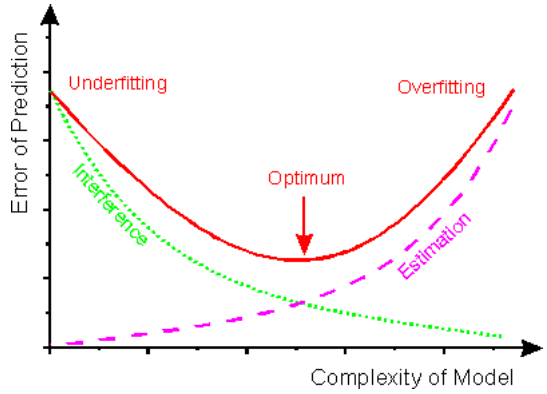


FIGURE 3.5: Error of prediction as a function of the complexity

3.5.4 Complexity of decision tree induction

Computing optimal decision tree is known to be NP-complete [38].The algorithms presented so far use a greedy,top-down,recursive partitioning strategy to induce a reasonable solution.A decision tree consists of two parts:building phase and pruning phase.

The training data contains n instances and m attributes.The depth of the tree is on the order of $\log n$.Thus,the computational cost of building the tree in the first place is:

$$O(mn \log n)$$

The cost of pruning,it depends on what pruning technique is used(pre or post pruning).We consider post-pruning by subtree replacement.Each node needs to be considered for replacement.The tree has at most n leaves,one of each instance.Thus the complexity of subtree replacement is: $O(n)$. During subtree lifting process,each instance may have to be reclassified at every node between its leaf and the root,that is,as many as $O(\log n)$ times.That makes the total number of reclassifications $O(n \log n)$.And reclassification is not a single operation:one that occurs near the root will take $O(\log n)$ operations.Thus the total complexity of subtree lifting is:

$$O(n(\log n)^2)$$

Thus,the full complexity of decision tree induction is:

$$O(mn \log n) + O(n(\log n)^2)$$

3.6 Readability Formulas

Readability is how easily the written materials can be read and understood. McLaughlin offered a more theoretical definition;*"A readability formula is simply a mathematical equation derived by regression analysis which best expresses the relationship between two variables,which in this case are a measure of the difficulty experienced by people reading a given text,and a measure of the linguistic characteristics of that text"*[39].

Construction of formulas

Readability formulas are multiple regression equations which predict the reading ability required to understand a given piece of text.The equations usually involve one or more of the following [40]:

- Average word length in syllables.
- Average sentence length in words.
- Proportion of common words used.
- Proportion of words with three or more syllables in them.
- Proportion of words which are monosyllabic.

3.6.1 Formulas used in health care

Several of the formulas to be described are included in various computer word-processing programs.All of them,can be fairly easily calculated with the SMOG Grading being the easiest.

i. *Flesh Reading Ease*(Flesh,1948)

In the Flesch Reading Ease test, higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The formula for the Flesch Reading Ease Score (FRES) test is

$$FRES = 206.835 - 1.015(\frac{total\ words}{total\ sentences}) - 84.6(\frac{total\ syllables}{total\ words})$$

In the *Art of Readable Writing* [41],Flesch described his Reading Ease scale in this way (Tale 3.2):

ii. *Flesch-Kincaid Grade Level*

The *Flesch-Kincaid Grade Level Formula* translates the 0-100 score to a U.S. grade level, making it easier for teachers, parents, librarians, and others to judge the readability level of various books and texts.It can also mean the

Score	Description	Reading Grade	Typical Magazine
90-100	Very Easy	5th	Comics
80-89	Easy	6th	Pulp fiction
70-79	Fairly Easy	7th	Slick fiction
60-69	Standard	8th to 9th	Digests
50-59	Fairly Difficult	10th to 12th	Quality
30-49	Difficult	13th to 16th	Academic
0-29	Very Difficult	College graduate	Scientific

TABLE 3.2: Flesch Reading Ease Formula

number of years of education generally required to understand this text, relevant when the formula results in a number greater than 10. The grade level is calculated with the following formula:

$$Flesch_Kincaid_Grade = 0.39(\frac{total\ words}{total\ sentences})+11.8(\frac{total\ syllables}{total\ words})-15.59$$

The result is a number that corresponds with a grade level. For example, a score of 8.2 would indicate that the text is expected to be understandable by an average student in year 8 in the United Kingdom (usually around ages 12-14 in the USA).This formula was based on Navy training manuals that ranged in difficulty from 5.5 to 16.3 in reading grade level. The score reported by this formula tends to be in the mid-range of the 4 scores. Because it is based on adult training manuals rather than school book text, this formula is probably the best one to apply to technical documents.

iii. *Gunning Fog Index*(1973)

The fog index is commonly used to confirm that text can be read easily by the intended audience. Texts for a wide audience generally need a fog index less than 12. Texts requiring near-universal understanding generally need an index less than 8.The formula is:

$$FogIndex = 0.4\left[(\frac{words}{sentences}) + 100(\frac{complex\ words}{words})\right]$$

iv. *McLaughlin's SMOG Grading*(1969)

SMOG grading implicitly makes two claims: that counting polysyllabic words in a fixed number of sentences gives an accurate index of the relative difficulty of various texts; and that the formula for converting polysyllable counts into reading grades gives acceptable results.The formula is:

$$SMOG_grade = 1.043\sqrt{number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}}+3.1291$$

v. *Coleman-Liau Index*(1975) Coleman-Liau formula relies on characters instead of syllables per word. Although opinion varies on its accuracy as compared to the syllable/word and complex word indices, characters are more readily and accurately counted by computer programs than are syllables.Its output approximates the U.S. grade level thought necessary to comprehend the text.The formula is:

$$Coleman_Liau_grade = 0.0588L - 3S - 15.8$$

where L:is the average number of letters per 100 words and S:is the average number of sentences per 100 words.

vi. *Automated Readability Index*(1967)
The Automated Readability Index (ARI), based on text from grades 0 to 7, was derived to be easy to automate. The formula is:

$$ARI_grade = 4.71(\frac{characters}{words}) + 0.5(\frac{words}{sentences}) - 21.43$$

where *characters* is the number of letters, numbers, and punctuation marks, *words* is the number of spaces, and *sentences* is the number of sentences.ARI tends to produce scores that are higher than Kincaid and Coleman-Liau but are usually slightly lower than Flesch.

3.7 Multicriteria Decision Analysis

Multicriteria Decision Analysis(MCDA) is both an approach and a set of techniques,with a goal of providing an overall ordering options,from the most preferred to the least preferred option. The options may differ in the extent to which they achieve several objectives, and no one option will be obviously best in achieving all objectives.MCDA is a way of looking at complex problems by breaking the problem into more manageable pieces to allow data and judgements to be brought to bear on the pieces, and then of reassembling the pieces to present a coherent overall picture to decision makers [42]. The purpose is to serve as an aid to thinking and decision making, but not to take the decision. As a set of techniques, MCDA provides different ways of disaggregating a complex problem, of measuring the extent to which options achieve objectives, of weighting the objectives, and of reassembling the pieces (see Appendix A.4).

3.7.1 The UTA method

The UTA (UTilites Additives) method proposed by Jacquet-Lagreze and Siskos (1982) aims an inferring one or more additive value functions from a given ranking

on a reference set A_R.In this work,we adopt the UTASTAR method which is an improved version of the original UTA [43].

In abstract,the UTASTAR algorithm ,considers as input a weak-order preference structure on a set of alternatives (here medical documents) together with the performances of all the alternatives on all attributes (here SN categories),and returns as output a set of additive value functions based on multiple criteria,in such a way that the resulting structure would be as consistent as possible with the initial structure given by the user.This is accomplished by means of special linear programming techniques in four basic steps.The UTASTAR algorithm aims at estimating additive utilities of the form:

$$U_{(g)} = \sum_{i=1}^{m} u_i(g_i) \tag{3.4}$$

subject to the following constrains:

$$\begin{cases} u_i(g_{i^*}) = 0, \ \forall i \\ \\ \sum_{i=1}^{m} u_i(g_i^*) = u_1(g_1^*) + u_2(g_2^*) + \ldots + u_m(g_m^*) = 1 \end{cases} \tag{3.5}$$

where $u_i(g_i)$ $i = 1, 2, \ldots, m$ are non-decreasing real valued functions,named marginal utility functions.

Step 1:Express the global value of reference actions $u[g(a_k)]$ $k = 1, 2, \ldots, m$,first in terms of marginal values $u_i(g_i)$,and then of variables w_{ij} according to the equation 3.7.1 .The transformation of the global value of reference actions into weights values expression is made according to equation 3.6 :

$$w_{ij} = u_i(g_i^{j+1}) - u_i(g_i^j) \geqslant 0, \ \forall i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, a_i - 1 \tag{3.6}$$

$$\begin{cases} u_i(g_i^1) = 0, \ \forall i = 1, 2, \ldots, n \\ \\ u_i(g_i^j) = \sum_{t=1}^{j-1} w_{it}, \ \forall i = 1, 2, \ldots, n \text{ and } j = 2, 3, \ldots, a_{i-1} \end{cases} \tag{3.7}$$

Step 2:Introduce two error functions σ^+ and σ^- on A_R (reference set of alternatives) by writing for each pair of successive actions in the given ranking the equation 3.8:

$$\begin{aligned} \Delta(a_k, a_{k+1}) = u[g(a_k)] - \sigma^+(a_k) + \sigma^-(a_k) \\ - u[g(a_{k+1})] + \sigma^+(a_{k+1}) - \sigma^-(a_{k+1}) \end{aligned} \tag{3.8}$$

Step 3:Solve the linear program (LP):

$$\begin{cases} [min]z = \sum_{k=1}^{m}[\sigma^{+}(a_k) + \sigma^{-}(a_k)] \\ \text{subject to} \\ \quad \left.\begin{array}{ll} \Delta(a_k, a_{k+1}) \geqslant \delta & \text{if } a_k \succ a_{k+1} \\ \Delta(a_k, a_{k+1}) = 0 & \text{if } a_k \succ a_{k+1} \end{array}\right\} \forall k \\ \sum_{i=1}^{n}\sum_{j=1}^{a_i-1} w_{ij} = 1 \\ w_{ij} \geqslant 0,\ \sigma^{+}(a_k) \geqslant 0,\ \sigma^{-}(a_k) \geqslant 0 \quad \forall i, j \text{ and } k \end{cases} \tag{3.9}$$

Step 4 (Stability Analysis):Check the existence of multiple or near optimal solutions of the linear program 3.9.In case of non-uniqueness,find the mean additive value function of those (near) optimal solutions which maximize the objective function 3.10,on the polyhedron of the constrains of the LP 3.9 bounded by the new constraint 3.11,where z^* is the optimal value of the LP in the Step 3 and ε is a very small positive number.

$$u_i(g_i^*) = \sum_{j=1}^{a_j-1} w_i j \quad \forall i = 1, 2, \ldots, n \tag{3.10}$$

$$\sum_{k=1}^{m}[\sigma^{+}(a_k) + \sigma^{-}a_k] \leqslant z^* + \varepsilon \tag{3.11}$$

UTASTAR's output involves the value functions associated to each criterion,approximated by linear segments,as well as the criteria significance weights (trade-offs among the criteria values).

In this work,we advocate that the characterization of a medical document as *"for experts"* of *"for consumers"* depends on the expert terms that the document contains,which in turn are mapped to Semantic Network sub-category terms.The latter,represent the criteria on which the overall percentage of expert terms in a document term vector depends.

Chapter 4

Medical Document Classification by User Profile

We follow a three phase methodological framework,as described below (figure x):

i. Data Retrieval and expert term Extraction.

ii. Data Representation and Modeling.

iii. Medical Document Classification.

4.1 Data Retrieval and expert term extraction

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.
- A method for extracting MeSH terms from medical documents.AMTEX or MMTx discussed in Chapter 3,in section 3.4 and 3.2 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 expert or consumer document).

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 distinguished into:

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

i. general medical terms expressing known concepts (e.g "pain","headache") which are easily conceived by all users.

ii. domain specific terms which are used mainly by experts.

iii. general-non medical terms.

The more expert terms a document contains,the higher its probability to be a document for experts [44].Figure 4.1 illustrates the respective categorization of Medline documents and MeSH terms. Wordnet thesaurus [2] contains 127.361 terms,while

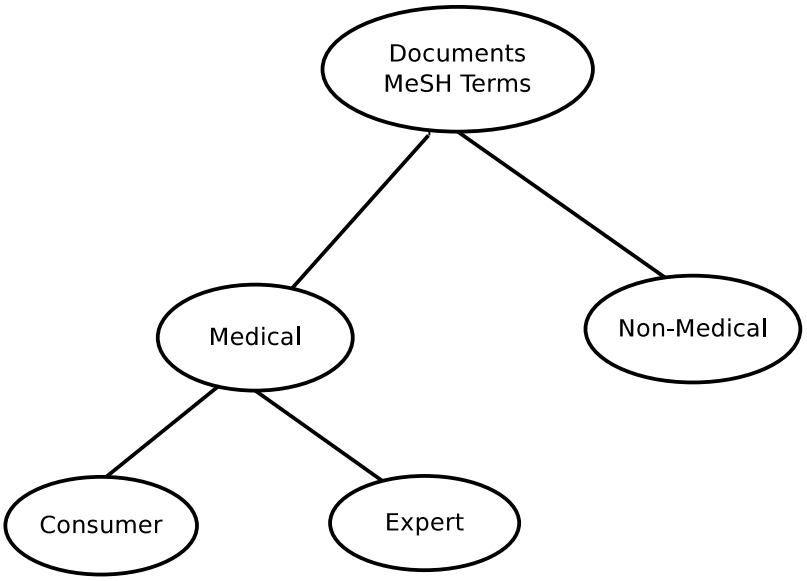


FIGURE 4.1: Categorization of Medline documents and MeSH terms.

MeSH (2013) 26.853 terms.The three vocabularies are constructed by combining their terms as follows:

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

$$VGT = (Wordnet) - (MeSH)$$

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

– **Vocabulary of Consumer Terms (VCT):**these are terms that belong to both;Wordnet and MeSH.

$$VCT = (Wordnet) \cap (MeSH)$$

It follows that VCT contains 7.165 consumer (MeSH) terms.

[2]Latest released version is 3.0 on December 2006.

– **Vocabulary of Expert Terms:**these are MeSH terms that do not belong to Wordnet:

$$VET = (MeSH) - (Wordnet)$$

It follows that VET contains 16.719 expert (MeSH) terms.

Notice that,*consumer* and *expert* terms are *only* MeSH terms (their intersection is the MeSH vocabulary).Notice,also that the 70% os the MeSH terms are *expert* terms,while only 30% of MeSH terms are *consumer* terms.Documents are represented by term vectors [45] produced by AMTEx(v2.0),MMTx and MeSH [3] respectively.Each term in this vector 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 of each multi-word term: The weight d_i of term i in a document is computed as $d_i = tf_i \times idf_i$, where tf_i is the frequency of term i in a document and idf_i is the inverse frequency of i in the whole document collection.

In Multiple Criteria Decision Analysis, a criterion is considered as a mean for evaluating and comparing potential actions (in this case medical documents), according to a point of view which must be as well-defined as possible. This evaluation must take into account all the attributes (in this case the SN categories) linked to the point of view considered for each action. These are called the criteria performances. To ensure that in our experiments all the actions are evaluated on the same basis of criteria, only expert terms are considered in calculating the criteria performances. Besides that, consumer terms convey less classificatory information as they tend to appear in both document types with equal probability.

Since we are trying to evaluate the medical documents on their "comprehension difficulty level" for a consumer,we adapt a unified measurement scale that reflects the degree of "difficulty" of each attribute. In other words,by considering only expert terms in the criteria performances we determine whether they are targeted to experts or not.

Thus,since the amount of expert terms in a document is low,even for expert documents,we ignore consumer terms during modeling process in our experiments and we represent all document based on expert terms.Consequently,we assume mutual exclusion,meaning that any medical document that is not expert is presumably a consumer document.

[3]MeSH document vector contains the terms that are assigned by the stuff of NLM during the indexing process.

4.2 Data representation and modeling

In decision tree analysis,a set of x attributes defines an x-dimensional *description space* in which each instance is a point.Also,the Disaggregation-Aggregation approach of MCDA has mainly focused on the development of comprehensive decision models from small data sets. The total number of expert terms found in a document is too large in order to consider all the initial MeSH terms extracted from the document as attributes in the decision tree analysis or criteria in MCDA process.For these reasons,after the term extraction process,every term originating from either AMTEx or MMTx or MeSH is mapped by two-layered indexing structure of figure X to the UMLS Semantic Network category terms.These subcategories are considered as criteria (MCDA) and attributes (DT).MeSH terms,as given by Medical Subject Headings Section staff for every document are also similarly mapped.

Only expert terms,as described in section 4.1 count in this process and the simple term frequency measure is applied.Hence,a document in the data set is represented by 130-dimensional vector of expert term frequency as:

$$d_i = tf_1, tf_2, \ldots tf_n \text{ where } n = 1, 2, \ldots 130$$

and

$$tf_i = \frac{\sum_{j=1}^{k} t_j \rightarrow t_j \in VET}{N}$$

where

- k is the number of expert terms that belong to the ith SN category and
- N is the total number of expert terms in d_i

For example,consider that for a document d_i five different expert MeSH terms are extracted by AMTEx,two of which belong to the sub-category *"Molecular Function"*,one to the sub-category *"Cell"* and the remaining two to *"Disease or Syndrome"*.Then,the value of sub-categories *"Molecular Function"* and *"Disease or Syndrome"* will be 2/5,while *"Cell"* 1/5.Therefore ,the smaller the number of a sub-category,the less this sub-category contributes to the classification of d_i,as expert document.A zero value here means that no expert term from this SN subcategory was extracted.

MCDA methods usually assume that only a small reference set is available,since it is difficult for the decision makers to express their global preferences on too many alternatives.Therefore,during the 10-fold cross validation,the average values of the estimated UTASTAR parameters(solutions of equations 3.9,3.10,3.11 of section

doc_id	$\mathbf{P}_e$	C_1	C_2	...	C_{130}
67914	**0.62**	0.0	0.1	...	0.07
69631	**0.41**	0.1	0.0	...	0.0
69966	**0.83**	0.2	0.08	...	0.0
57296	**0.41**	0.0	0.04	...	0.0
⋮		⋮	⋮	⋮	⋮
67019	**0.77**	0.03	0.0	...	0.0

doc_id	**Ranking Order**	C_1	C_2	...	C_{130}
67914	**3**	0.0	0.1	...	0.07
69631	**4**	0.1	0.0	...	0.0
69966	**1**	0.2	0.08	...	0.0
57296	**4**	0.0	0.04	...	0.0
⋮		⋮	⋮	⋮	⋮
67019	**2**	0.03	0.0	...	0.0

TABLE 4.1: Initial data form before and after (Multicriteria input matrix) transforming global rating into ranking order.

3.7.1) were calculated.By applying the so called UTASTAR algorithm in the training sets,a vector of significance weights for the UMLS Semantic Network category terms is calculated,indicating the different role of each category in the characterization of a MEDLINE document as *"for consumers"*or *"for experts"*.In out experiments,the probability of a document to be considered as *"expert"* or *"consumer"* is calculated as the number of expert terms that the specific document contains divided by the total number of terms extracted from the specific document.More specifically,this probability is calculated as the percentage of terms that belong to the VET.For example,a document with VET%=0.62 has 62% probability of being a document suitable for experts.Therefore,we assume that this probability represents the global estimation of a document as expert and based on that,we transform the initial global ratings into a ranking order for the training set.

To compute the readability score for each document,we used STYLE program. STYLE calculates the readability,sentence length variability,sentence type,word usage and sentence openers at a rate of about 400 words per second and runs on UNIX Operating System [46].Running a perl script on Linux command prompt,STYLE reads all documents (expert and consumer) and prints the readability indices for each document.

4.3 Medical Document Classification

For each document both,its vector representation,its score probability,the average values of UTASTAR parameters and its readability grade are computed.During the classification phase,a medical document is labeled as either expert or consumer.Based on this information,document categorization is determined by machine learning,Multicriteria decision analysis and readability formulas.

- **Machine Learning by decision trees**:Let the system decide which category a document belongs.Creating a decision tree with consumer and expert documents and after the training process,the system can decide in which category an input document belongs to.
- **Multicriteria decision Analysis**:Make a choice based upon the utility score calculated based on the final solution that corresponds to the marginal value functions (criteria weights) presented in Table 5.12.
- **Readability Formulas**:Using a threshold, defined by user in relation to the interpretation of readability formulas, lets the system decide if a document belongs to consumer or expert class according to its readability score.

Chapter 5

Experiments and Evaluation

We designed a series of experiments whose purpose is twofold: First, to study and compare the effectiveness of AMTEx, MMTx and MeSH, in classifying medical documents and second, to prove that Semantic Network categories contribute differently in classifying medical documents.

5.1 Experimental Setup

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

- **OHSUMED**: Standard TREC collection corpus. OHSUMED is a collection of MEDLINE document abstracts used for benchmarking information retrieval systems evaluation. Fore more information about OHSUMED see Chapter 2.2.2 at page 6.
- **PubMed** : Provides free access to MEDLINE, NLM's database of citations and abstracts in the fields of medicine, nursing, dentistry etc. The documents were selected on the basis of having a unique PMID number, which was used to retrieve their respective MEDLINE index sets. This index set for each document is manually assigned by MEDLINE experts.

Initially, documents are retrieved from a subset of the OHSUMED TREC collection consisting of 10% of OHSUMED, i.e 34.0000 document abstracts (because MMTx is slow,processing of the entire OHSUMED document collection was not feasible). Both data storing and access mechanisms are implemented using Lucene. Relevance judgments on the first 20 answers retrieved by all three competitive methods(AMTEx, MMTx and manually assigned MeSH terms) for all 15 queries were provided by the members of the Intelligence Systems Laboratory. Each subject judged the results to a number of queries (the same for all methods), by assessing

if a result is a consumer document (by understanding the document subject) or expert document (by not understanding the document subject).

Also, we collected a new data set from PubMed which includes 580 expert and 572 consumer medical documents as result to 50 different queries(see Appendix A.1). To separate consumer from expert documents we exploited PubMed Health[1], which specializes in reviews of clinical effectiveness research, with easy-to-read summaries for consumers (plain language summary).

5.2 Decision Tree

As mentioned before, the retrieved documents were evaluated manually by users and considered as the ground truth for our experiment. Subsequently, we extracted only expert documents that contain at least 2 Semantic Network category terms resulting in a subset of 237 different expert documents. To avoid any inconsistencies originating from our data set, the same number of consumer documents is selected. Therefore, our experimental data set consists of the above 237 expert and 237 consumer documents and is used in all our experiments.

After the manual evaluation of the abstract of the documents, we created a decision tree. We used Weka [33, 47] and J48 classifier as the tool for training the decision tree with default parameter values (confidence threshold for pruning at 0.25). J48 algorithm is a Java reimplementation of the C4.5 algorithm. Ten-fold cross validation was chosen as the evaluation method to compare the effect of different extracting methods.

The documents are represented as term vectors and the attributes are all the MeSH terms extracted from each method. Also we map each term extracted to the respective Semantic Network category that it belongs. This mapping to the SN category reduces the dimensional description space of the decision tree and provides better results in relation to the other decision trees.

	AMTEx	MMTx	MeSH	All_Terms	SN_category
Accuracy(%)	59.5	58	63.7	60.1	68
TruePositiveRate Expert(%)	63	61.3	60.3	60.3	69
TruePositiveRate Consumer(%)	55.4	54.5	67.5	59.9	67

Table 5.1: OHSUMED Expert-Consumer Evaluation Decision Tree Results

Table 5.1 shows the accuracy of each method. *All_Terms* is a combination of the three methods. The accuracy is near to statistical probability. This is due to the enormous vectors we have created, which use the extracted MeSH terms as

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

attributes. For example we have 1,410 , 2,418, 2,625, 4,622 attributes for AMTEx, MeSH, MMTx,All_Terms respectively. Mapping each term extracted to the respective Semantic Network creates significantly smaller vectors (range 130) which leads to better results.

Subsequently, during the modeling process we ignore consumer terms and we represent all documents based on expert terms as we mentioned at the beginning of this section. The data set contains 237 consumer and 237 expert documents. Table 5.2 shows the results of decision tree analysis on this document collection.

	AMTEx	MMTx	MeSH
Classification Accuracy(%)	83.75	78.9	65.82
TruePositiveRate Expert(%)	73.4	71.7	60.8
TruePositiveRate Consumer(%)	94.1	86.1	70.9

TABLE 5.2: OHSUMED dataset-weights by expert terms.Decision Tree Results

The new data retrieved from MEDLINE and PubMed Health were initially used for testing our first decision tree model and subsequently training a new decision tree with these new data.We collected 1152 documents. We considered the "plain language summary" from PubMed Health as consumer level and the documents retrieved from MEDLINE as expert. Table 5.3, shows the results to the new unseen data for the AMTEx method.

	AMTEx
Classification Accuracy(%)	68.2
TruePositiveRate Expert(%)	62.1
TruePositiveRate Consumer(%)	74.3

TABLE 5.3: Decision Tree Results Test Set from PubMed

Subsequently, we used these new data as the training set for a new Decision Tree. Table 5.4 shows the performance of the decision tree classifier in these data. The classification accuracy of AMTEx method is lowest (Table 5.4). This is due to the new ground truth we assumed. Although the "plain language summary" is considered to be simpler to read than the abstract, in fact it does not seem to be more easily comprehended. The mere division of medical documents into two classes of "consumer" and "expert" seems absolute and therefore problematic. The validity of this hypothesis becomes apparent once the document classes increase, initially to four classes and then to three classes according to the probability thresholds (Table 5.5 and Table 5.6). When consumer probability for a document equates 1 it corresponds to the first category (consumer) otherwise it is assigned to the second category (Consumer_ expert). When expert probability is higher than 0.1

it corresponds to the expert category, otherwise it is assigned to the third category (expert_ consumer).

	AMTEx
Classification Accuracy(%)	63.02
TruePositiveRate Expert(%)	75.2
TruePositiveRate Consumer(%)	50.6

TABLE 5.4: Decision tree Pubmed Data set

	AMTEx
Classification Accuracy(%)	76
TruePositiveRate Consumer(%)	1
TruePositiveRate Consumer_Expert(%)	71.8
TruePositiveRate Expert_Consumer(%)	0.13
TruePositiveRate Expert(%)	82.1

TABLE 5.5: Decision tree Four Categories Pubmed Data set

	AMTEx
Classification Accuracy(%)	78.3
TruePositiveRate Consumer(%)	88.8
TruePositiveRate Expert_Consumer(%)	16.3
TruePositiveRate Expert(%)	79.4

TABLE 5.6: Decision tree Three Categories Pubmed Data set

5.3 Readability Formulas

As we mentioned in section 3.6.1,we compute six different readability formulas:

i. *Flesch Reading Ease.*

ii. *Flesch Kincaid Grade.*

iii. *Gunning Fog Index.*

iv. *McLaughlin's SMOG Grading.*

v. *Coleman-Liau Index.*

vi. *Automated Readability Index.*

At a glance the statistical description of Readability formulas (table 5.7) shows that the scores for each formula characterize all documents as very difficult to read. In order to find if all these formulas follow the normal (gaussian) distribution and find a threshold we create histograms, q-q plots and run normality tests such as the Shaphiro-Wilk test.

	Consumer			Expert		
	mean	std	median	mean	std	median
Flesch Reading Ease	34.31	13.52	35.85	29.45	13.64	30.1
Flesch Kincaid Grade	13.22	2.3	13.1	14.45	2.37	14.4
Gunning Fog Index	16.78	2.67	16.6	18.11	2.71	18
McLaughlin's SMOG Grading	14.17	1.8	14	15.12	1.81	15
Coleman-Liau Index	17.73	2.8	17.55	17.78	2.94	17.7
Automated Readability Index	14.92	2.73	14.6	16.09	2.86	16

TABLE 5.7: Statistical Description for each Readability formulas-OHSUMED data set

	Consumer			Expert		
	mean	std	median	mean	std	median
Flesch Reading Ease	40.17	13	40.25	37.03	13.93	36.4
Flesch Kincaid Grade	13.18	2.73	13.2	12.54	2.59	12.4
Gunning Fog Index	16.48	3.11	16.4	16.2	3	16
McLaughlin's SMOG Grading	14.1	2.1	14.1	13.84	2	13.7
Coleman-Liau Index	16.49	2.12	16.5	17.61	2.56	17.7
Automated Readability Index	15.51	3.24	15.3	14.23	3.02	14

TABLE 5.8: Statistical Description for each Readability formulas-PubMed data set

Normality test

For accuracy purposes and in order to ensure that our data follow a normal distribution we ran normality tests. Normality tests are used to determine whether or not a data set is well-modeled by a normal distribution. It is a formal way to conclude if a distribution is normally distributed, however it is very strict and almost never shows that the data are following the characteristic pdf of normal distribution. If the p-value is above 0.05, then the data are normally distributed. The table 5.9 and 5.10 shows the results of the Shaphiro-Wilk normality test for both data sets.

QQ plot

As we stated above normality tests and q-q plots constitute the formal way to conclude the normality of our data. A Q-Q plot is a probability plot, namely a graphical method of comparing two probability distributions by plotting their quantiles against each other. A normal QQ plot compares randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are

	Consumer	Expert
	p-value	p-value
Flesch Reading Ease	5.906676e-06	0.001886487
Flesch Kincaid Grade	0.005681392	**0.07570962**
Gunning Fog Index	**0.06163741**	**0.295823**
McLaughlin's SMOG Grading	**0.3019587**	**0.3452352**
Coleman-Liau Index	6.093783e-06	0.005836703
Automated Readability Index	0.0015	**0.3866164**

TABLE 5.9: Shaphiro-Wilk test-OHSUMED data set

	Consumer	Expert
	p-value	p-value
Flesch Reading Ease	0.0004793388	0.001456995
Flesch Kincaid Grade	1.494351e-09	0.004602622
Gunning Fog Index	3.464652e-08	0.001953796
McLaughlin's SMOG Grading	2.204126e-06	0.002383906
Coleman-Liau Index	0.03707768	**0.07620615**
Automated Readability Index	1.130511e-10	0.0003780495

TABLE 5.10: Shaphiro-Wilk test-PubMed data set

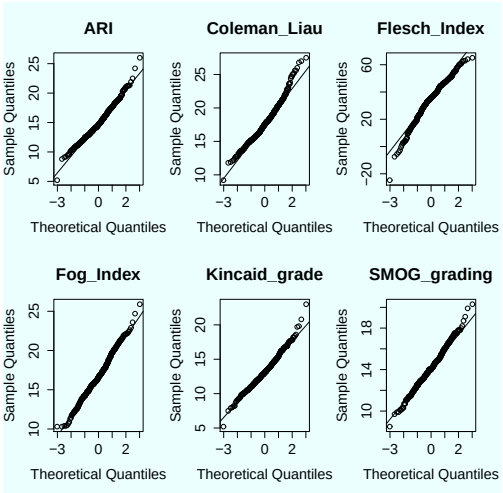


FIGURE 5.1: QQ-plot Consumer OHSUMED data set

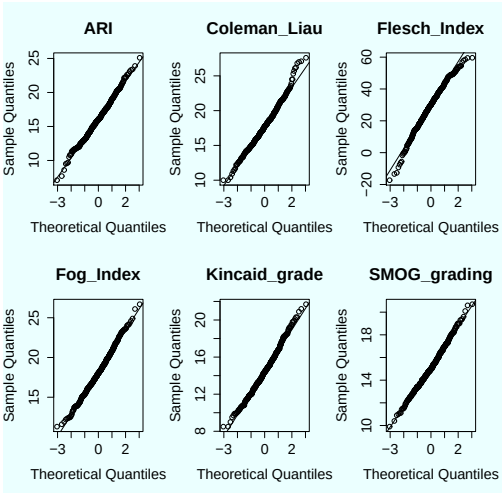


FIGURE 5.2: QQ-plot Expert OHSUMED data set

normally distributed (figures 5.1,5.2,5.3.5.4). The tails of some indexes in PubMed data seems to discrepant, but apparently follow the normal distribution as long as their histograms have the desirable symmetry.

Additionally, the histograms indicate the normality for the Ohsumed and Pubmed data set.

The overlay histograms 5.9 and 5.10 indicate that the overlay is very large and it is impossible to find a threshold to separate documents according to their readability

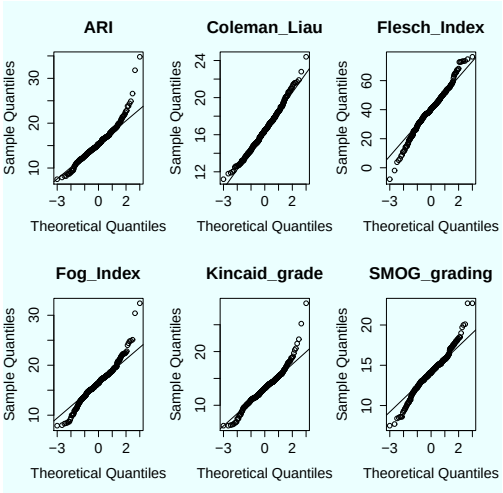


FIGURE 5.3: QQ-plot Consumer PubMEd data set

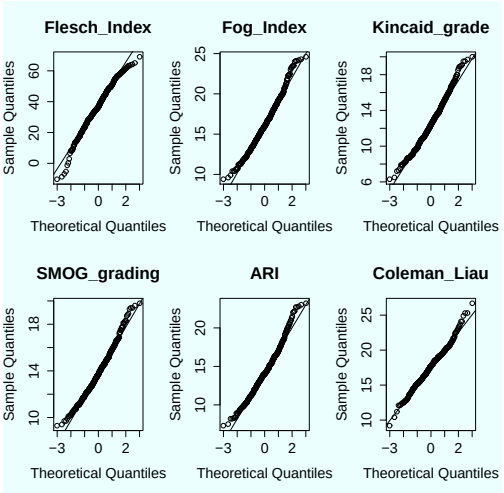


FIGURE 5.4: QQ-plot Expert PubMed data set

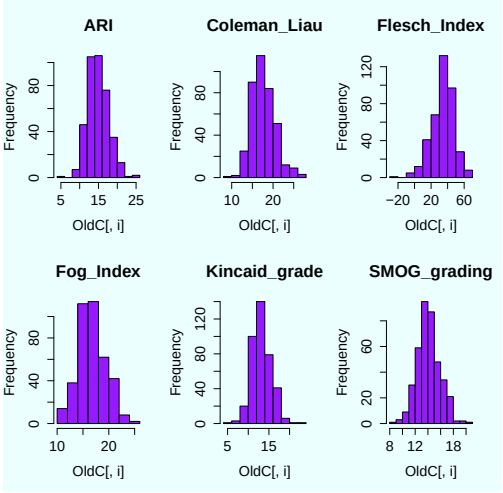


FIGURE 5.5: Histograms Consumer OHSUMED data set

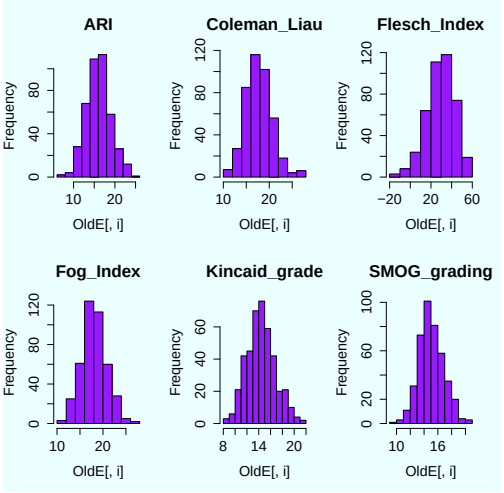


FIGURE 5.6: Histograms Expert OHSUMED data set

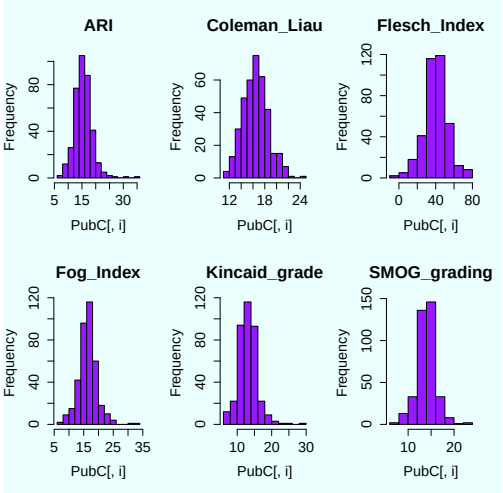


FIGURE 5.7: Histograms Consumer PudMed data set

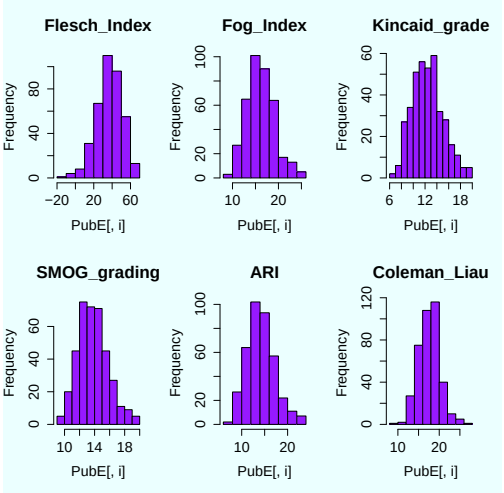
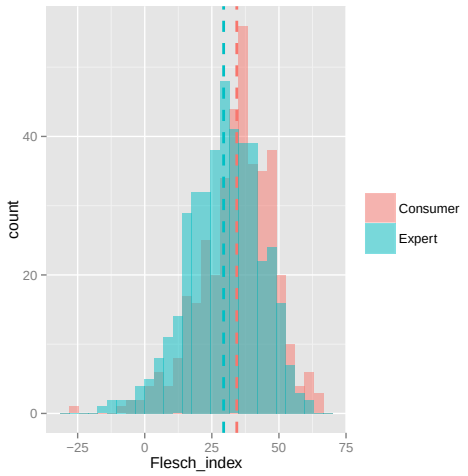
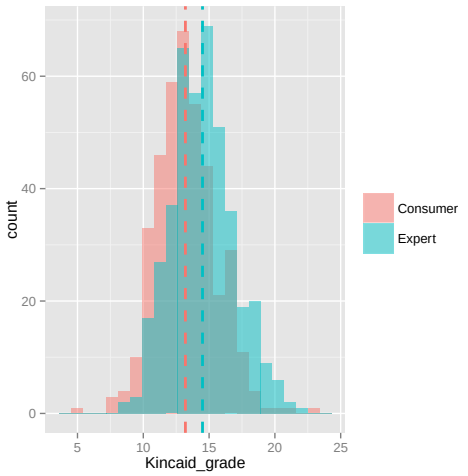


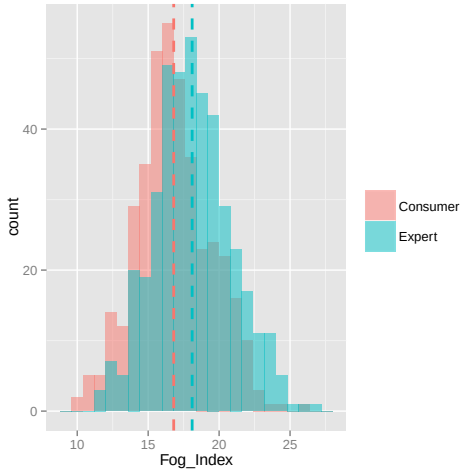
FIGURE 5.8: Histograms Expert PubMed data set



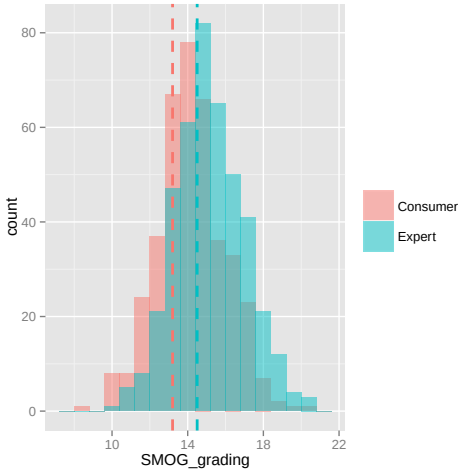
(a) Flesch Reading Ease



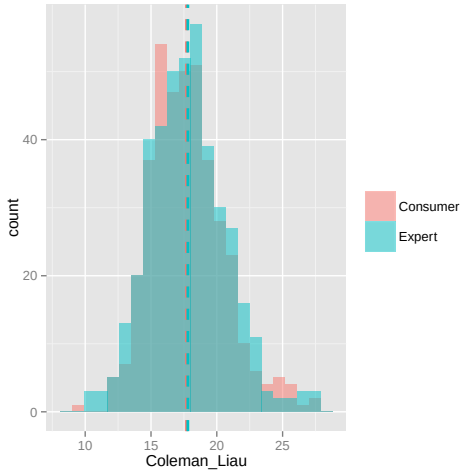
(b) Flesch Kincaid Grade



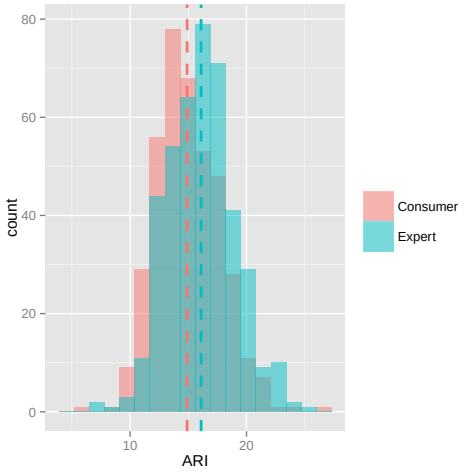
(c) Gunning Fog Index



(d) SMOG Grading

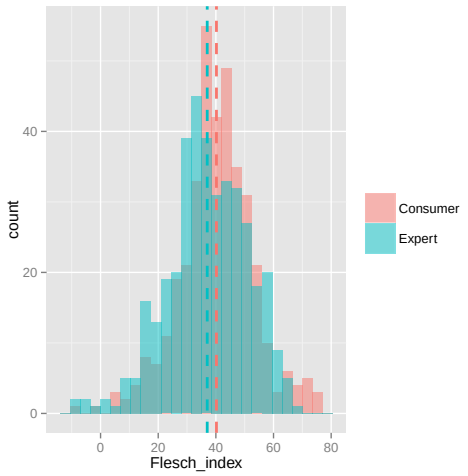


(e) Coleman-Liau Index

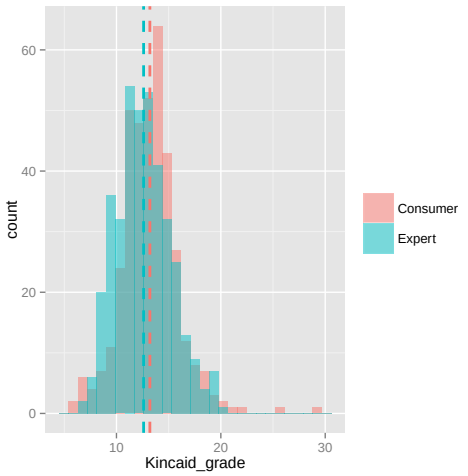


(f) Automated Readability Index

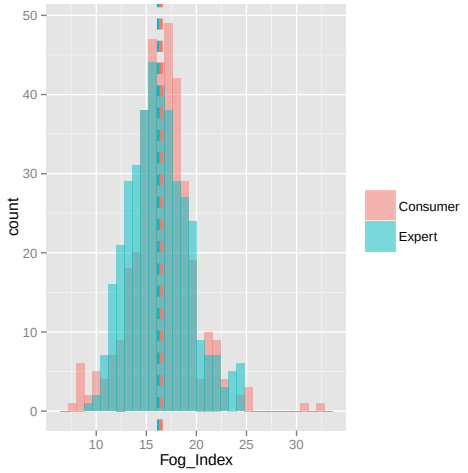
FIGURE 5.9: Overlay Histograms Readability Scores-OHSUMED data set



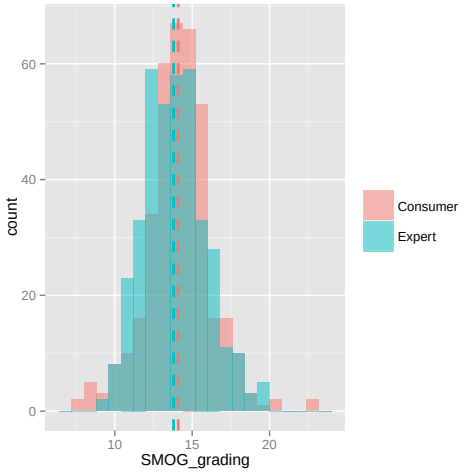
(a) Flesch Reading Ease



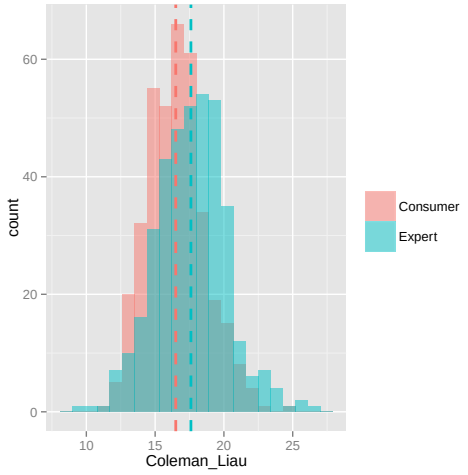
(b) Flesch Kincaid Grade



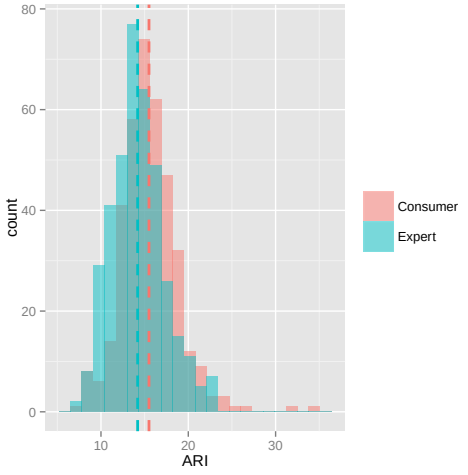
(c) Gunning Fog Index



(d) SMOG Grading



(e) Coleman-Liau Index



(f) Automated Readability Index

FIGURE 5.10: Overlay Histograms Readability Scores-PubMed dataset

scores. This is because the traditional readability formulas measure the difficulty of writing style for general purpose of education. When two sets of articles have similar linguistic features in terms of difficulty of style, then the features that are able to capture the difficulty of content become important. Readability scores do not reflect other factors that affect comprehension such as frequency and explanation of medical terminology, writing style or use of culturally specific information [1].

The above results are also confirmed with decision tree analysis. We used the readability formulas as attributes in the training process and the classification accuracy presented in Table 5.11.

	OHSUMED	PubMed
Classification Accuracy(%)	56.72	68.56
TruePositiveRate Expert(%)	68.8	68
TruePositiveRate Consumer(%)	43.7	69.2

TABLE 5.11: Decision tree Readability formulas

5.4 Multicriteria Decision Analysis:UTASTAR algorithm

As previously stated, we assumed that not all the Semantic Network sub-categories contribute identically to the classification of medical documents as expert or consumer, instead we identify different significance weights for those sub-categories. Table 5.12 shows the results in terms of significance weights, for AMTEx, MMTx and MeSH term respectively. A SN category is included in Table 5.12, if its weight is greater than **0.0077**,which results from:

$$\sum_i \frac{w_i}{n}$$

where n is the number of SN sub-categories and represents the baseline of equal significance.

The significance weights do not follow the frequency appearance of the sub-categories in the document corpus, meaning that the most frequent sub-category is not necessarily the most important. The last column of table 5.12, SNPE, shows the sub-category expert probability calculated as the fraction of expert terms appearing in the document corpus and belongs to the specific sub-category divided by the total MeSH terms.

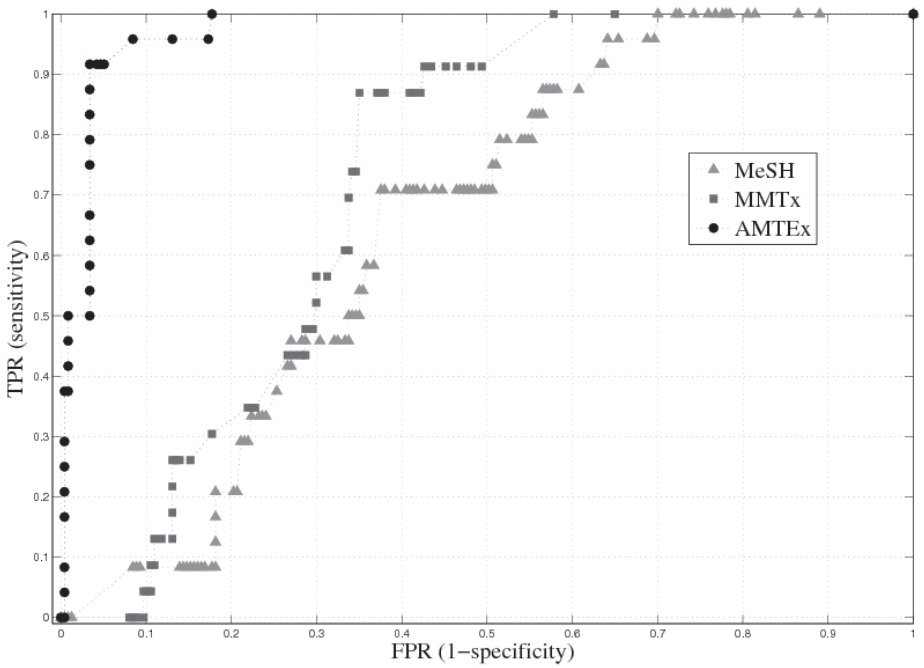


FIGURE 5.11: ROC curves for AMTEx,MMTx and MeSH approaches.

Table shows 5.13 the results of best-case, worst-case and average case values of all the evaluation measures calculated on AMTEx,MMTx,and MeSH term vectors from 10-folds. It is obvious from table 5.13 that AMTEx outperforms all the other methods in the majority of performance measures.

Additionally, figure 5.11 illustrates the ROC curves for the best precision values in all cases allowing us to compare the classification performance of our model built based on AMTEx terms, MMTx terms and MeSH terms. Receiver Operating Characteristics analysis (ROC) is a useful technique for organizing classifiers and visualizing their performance. It is related to cost-benefit analysis in decision making and has been widely used in medicine. ROC graphs are two-dimensional graphs in which True Positive rate is plotted versus False Positive. The diagonal line of a ROC graph represents the case of randomly guessing a class. Furthermore, the Area Under the Curve (AUC) has been shown to be an accurate evaluation measure.The best classifier is considered the one that is closest to (0,1) and furthest from diagonal.

Results from the application of our weight model (as calculated by the UTA) in 1041 pubmed documents in order to study our models generalization ability (Table 5.14).

Differences in the results between the two datasets, both in UTA and Decision Tree, may arise either from an overfitting of the model, or from the different ground truth considered and are mostly due to the lower precision achieved. A low precision here means that our model misclassifies actual expert documents as consumer.

At this point it is important to make the observation that in supervised machine learning, a preclassified data set (in this case medical documents) is available. Preclassification is done by human experts, thus liable to a certain level of subjective judgement. The preclassified data set (training set) is used in the learning process, that of the automatic formulation of classification criteria based on the aforementioned training set, in order to correctly classify new data items. It is practically impossible for the learned classifiers to always be correct. It is relatively difficult to predetermine the target group of any given article. This evaluation is one that falls on a continuum and largely into the *"it depends"* or *"don't know"* class [48].

SN sub-category	**AMTEx**	**MMTx**	**MeSH**	**SNPE**
Mammal	0.0053	0.0074	**0.0085**	0.37
Body Part,Organ or Organ Component	**0.0314**	0.0074	0.0055	0.46
Cell	0.0057	**0.0082**	**0.0098**	0.69
Organism Attribute	0.0052	**0.0087**	0.0056	0.71
Finding	**0.1032**	**0.0090**	0.0058	0.57
Physiologic Function	**0.0195**	0.0074	0.0055	0.67
Organism Function	0.0055	0.0071	**0.0086**	0.49
Cell Function	**0.0139**	**0.0086**	0.0507	0.77
Molecular Function	**0.0120**	**0.0081**	**0.0122**	0.83
Genetic Function	0.0056	0.0071	**0.0101**	0.87
Pathologic Function	0.0055	**0.0146**	0.0072	0.64
Disease or Syndrome	**0.0146**	**0.0313**	**0.0276**	0.72
Laboratory Procedure	**0.0140**	0.0082	**0.0159**	0.81
Diagnostic Procedure	0.0053	0.0072	**0.0080**	0.78
Therapeutic or Preventive Procedure	0.0055	**0.0133**	**0.0081**	0.75
Natural Phenomenon or Process	0.0053	0.0073	**0.0089**	0.43
Amino Acid Sequence	0.0071	0.0071	**0.0102**	1.00
Organic Chemical	0.0052	**0.0113**	**0.0061**	0.96
Amino Acid, Peptide, or Protein	**0,0092**	0,0075	0,0060	0.94
Pharmacologic Substance	0.0053	**0.0120**	**0.0142**	0.801
Biologically Active Substance	**0.0973**	**0.0120**	**0.0243**	0.87
Enzyme	**0.0137**	**0.0128**	0.0064	0.91
Immunologic Factor	**0.0314**	0.0071	**0.0205**	0.94
Hazardous or Poisonous Substance	0.0064	0.0075	**0.0360**	0.82
Intellectual Product	0.0057	**0.0088**	0.0057	0.68
Neoplastic Process	0.0054	**0.0085**	**0.0258**	0.85
Receptor	0.0055	0.0073	**0.0693**	0.97
Inorganic Chemical	0.0055	0.0071	**0.0120**	0.61

Table 5.12: UTA significance weight terms for AMTEx,MMTx and MeSH

		AMTEx	**MMTx**	**MeSH**
F-measure	Best-case	**0.8000**	0.3175	0.2615
	Worst-case	**0.5333**	0.2581	0.1977
	Average-case	**0.6527**	0.2814	0.2261
Precision	Best-case	**0.7097**	0.1942	0.1604
	Worst-case	**0.3636**	0.1481	0.1149
	Average-case	**0.5085**	0.1660	0.1313
Recall	Best-case	**1.000**	**1.000**	**1.000**
	Worst-case	**0.9130**	0.8696	0.6957
	Average-case	**0.9449**	0.9322	0.8391
AUC	Best-case	**0.9732**	0.7293	0.6578
	Worst-case	**0.9528**	0.6407	0.5359
	Average-case	**0.9629**	0.6882	0.6047
Accuracy	Best-case	**0.9580**	0.6705	0.6336
	Worst-case	**0.8397**	0.4733	0.3359
	Average-case	**0.9037**	0.5656	0.4743

Table 5.13: UTA classification evaluation measures for AMTEx,MMTx and MeSH

	PubMed(%)
Accuracy	70,73
Precision	64,18
Recall	96,81
F-measure	0,77
AUC	76,22

Table 5.14: UTA results on PubMed Data set

Chapter 6

Conclusions

We presented approaches to the problems of automatic term extraction and automatic categorization of medical information according to user profile (i.e. consumer and expert users). We adopted ideas from document information management, machine learning techniques, readability formulas and Recommender Systems and demonstrated how these can be applied for effectively classifying medical documents according to user profile. Medical documents were represented by term vectors extracted from three different approaches (AMTEx, MMTx and MeSH). Mapping the terms to their corresponding Semantic Network sub-category reduces the attributes of decision tree and can act as criteria to MCDA for the categorization providing remarkably high accuracy results, in contradiction to readability formulas which alone may not be good indicators of text difficulty of consumer health information due to their not providing information about the content but only about the writing style.

Using the OHSUMED medical collection, we trained the decision tree and the estimated accuracy was high. Testing the model with new unseen data retrieved from PubMed proved that our model probably overfits the data or the data affected by noise. We proved that the separation to two classes (expert-consumer) is absolute and we introduced a new class which represents the middle ground of uncertainty regarding the target group of a given medical document.

Possible future work may contain experiments using n-grams (a continuous sequence of n items from a given sequence of text) in order to model the data in both data sets. Some studies [48, 49] show that n-grams and more specifically unigrams achieve high accuracy. Also, Latent Semantic Analysis (LSA) is used in some studies although it is controversial due to computational cost [50]. Additionally, subject based document categorization may prove very useful in related applications in the future.

Appendix A

Appendix A

A.1 PubMed Health Abstract Retrieval Experiment Queries

1. Pneumonia in children.
2. Breast cancer Treatment.
3. Brain injury.
4. Breast cancer risk with hormone therapy.
5. Treatment of the cough in children.
6. Chinese herbal medicine.
7. Diet for diabetes mellitus.
8. Diabetes and Pregnancy.
9. Bone Infections.
10. Infectious Diseases.
11. Cancer Chemotherapy.
12. Abdominal Pain.
13. AIDS and Infections.
14. Birth Control.
15. Cardiac Diseases.
16. Dental Health.
17. Low Blood Pressure.
18. Lung Cancer.
19. Parkinson's Disease.

20. Viral Infections.
21. Low Back Pain
22. Acute and Chronic Pain
23. Bipolar Disorder
24. Diabetes Treatment
25. Insulin resistance
26. Crohn's Disease
27. Inflammatory Bowel Disease
28. Rheumatoid Arthritis
29. Ovarian Cancer
30. High Blood Pressure
31. Stroke Prevention
32. Hormone Replacement Therapy
33. Osteoporosis Prevention
34. Insomnia Treatment
35. Obesity and Weight Loss
36. Asthma Treatment
37. Smoking Cessation
38. Sleep Problems
39. Alternative Medicine
40. Multiple Sclerosis
41. Blood Disorders
42. Birth Control Pills
43. Chronic Kidney Disease
44. Congestive Heart Failure
45. Digestive Disorders
46. Immune System Disorders
47. Kidney Failure
48. Cardiovascular disease
49. Mental Health
50. Sickle Cell Disease

A.2 MEDLINE/PubMed Data Elemet (Field) Descriptions

Tag	Name	Description
AB	Abstract	English language abstract taken directly from the published article
AD	Affiliation	Institutional affiliation and address of the first author
AID	Article Identifier	Article ID values supplied by the publisher may include the pii (controlled publisher identifier), doi (digital object identifier), or book accession
AU	Author	Authors
BTI	Book Title	Book Title
CI	Copyright Information	Copyright statement provided by the publisher
CIN	Comment In	Reference containing a comment about the article
CN	Corporate Author	Corporate author or group names with authorship responsibility
CON	Comment On	Reference upon which the article comments
CP	Chapter	Book chapter
CRDT	Create Date	The date the citation record was first created
CRF	Corrected and republished from	Final, correct version of an article
CRI	Corrected and republished in	Original article that was republished in corrected form
CTDT	Contribution Date	Book contribution date
CTI	Collection Title	Collection Title
DA	Date Created	Used for internal processing at NLM
DCOM	Completion Date	NLM internal processing completion date
DEP	Date of Electronic Publication	Electronic publication date
DP	Publication Date	The date the article was published
DRDT	Date Revised	Book Revision Date
EDAT	Entrez Date	The date the citation was added to PubMed;the date is set to the publication date if added more than 1 year after the date published

Continued on next page

Table A.1 – continued from previous page

Tag	Name	Description
EFR	Erratum For	Cites the original article needing the correction
EIN	Erratum In	Reference containing a published erratum to the article
ED	Editor	Book editors
EN	Edition	Book edition
FAU	Full Author Name	Full Author Names
FED	Full Editor Name	Full Editor Names
FIR	Full Investigator	Full investigator or collaborator name
FPS	Full Personal Name as Subject	Full Personal Name of the subject of the article
GN	General Note	Supplemental or descriptive information related to the document
GR	Grant Number	Research grant numbers, contract numbers, or both that designate financial support by any agency of the US PHS or other funding agencies
GS	Gene Symbol	Abbreviated gene names (used 1991 through 1996)
IP	Issue	The number of the issue, part, or supplement of the journal in which the article was published
IR	Investigator	Investigator or collaborator
IRAD	Investigator Affiliation	Affiliation investigator or collaborator
IS	ISSN	International Standard Serial Number of the journal
ISBN	ISBN	International Standard Book Number
JID	NLM Unique ID	Unique journal ID in the NLM catalog of books, journals, and audiovisuals
JT	Full Journal Title	Full journal title from NLM cataloging data
LA	Language	The language in which the article was published
LID	Location ID	The pii or doi that serves the role of pagination

Continued on next page

Table A.1 – continued from previous page

Tag	Name	Description
LR	Modification Date	Citation last revision date
MH	MeSH Terms	NLM Medical Subject Headings (MeSH) controlled vocabulary
MHDA	MeSH Date	The date MeSH terms were added to the citation. The MeSH date is the same as the Entrez date until MeSH are added
OAB	Other Abstract	Abstract supplied by an NLM collaborating organization
OCI	Other Copyright Information	Copyright owner
OID	Other ID	Identification numbers provided by organizations supplying citation data
ORI	Original Report In	Cites the original article associated with the patient summary
OT	Other Term	Non-MeSH subject terms (keywords) either assigned by an organization identified by the Other Term Owner, or generated by the author and submitted by the publisher
OTO	Other Term Owner	Organization that may have provided the Other Term data
OWN	Owner	Organization acronym that supplied citation data
PB	Publisher	Publishers of Books & Documents citations
PG	Pagination	The full pagination of the article
PHST	Publication History Status Date	Publisher supplied dates regarding the article publishing process
PL	Place of Publication	Journal's (country only) or books place of publication
PMCR	PMC Release	Availability of PMC article
PMID	PubMed Unique Identifier	Unique number assigned to each PubMed citation
PRIN	Partial Retraction In	Partial retraction of the article
PROF	Partial Retraction Of	Article being partially retracted

Continued on next page

Table A.1 – continued from previous page

Tag	Name	Description
PS	Personal Name as Subject	Individual is the subject of the article
PST	Publication Status	Publication status
PT	Publication Type	The type of material the article represents
RF	Number of References	Number of bibliographic references for Review articles
RIN	Retraction In	Retraction of the article
RN	EC/RN Number	Includes chemical, protocol or disease terms. May also a number assigned by the Enzyme Commission or by the Chemical Abstracts Service.
ROF	Retraction Of	Article being retracted
RPF	Republished From	Article being cited has been republished or reprinted in either full or abridged form from another source
RPI	Republished In	Article being cited also appears in another source in either full or abridged form
SB	Subset	Journal or citation subset values representing specialized topics
SFM	Space Flight Mission	NASA-supplied data space flight/mission name and/or number
SI	Secondary Source Identifier	Identifies secondary source databanks and accession numbers of molecular sequences discussed in articles
SO	Source	Composite field containing bibliographic information
SPIN	Summary For Patients In	Cites a patient summary article
STAT	Status Tag	Used for internal processing at NLM
TA	Journal Title Abbreviation	Standard journal title abbreviation
TI	Title	The title of the article
TT	Transliterated Title	Title of the article originally published in a non-English language, in that language
UIN	Update In	Update to the article

Continued on next page

Table A.1 – continued from previous page

Tag	Name	Description
UOF	Update Of	The article being updated
VI	Volume	Volume number of the journal
VTI	Volume Title	Book Volume Title

A.3 Semantic Network Categories

15 main Categories and 133 Subcategories.

1. Activities Behaviors
   Activity
   Behavior
   Daily or Recreational Activity
   Event
   Governmental or Regulatory Activity
   Individual Behavior
   Machine Activity
   Occupational Activity
   Social Behavior

2. 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

3. 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

4. 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

5. Devices
   Drug Delivery Device
   Medical Device
   Research Device

6. 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

7. Genes Molecular Sequences
   Amino Acid Sequence
   Carbohydrate Sequence
   Gene or Genome
   Molecular Sequence
   Nucleotide Sequence

8. Geographic Areas
   Geographic Area

9. 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

10. Objects
    Entity
    Food
    Manufactured Object
    Physical Object
    Substance

11. Occupations
    Biomedical Occupation or Discipline
    Occupation or Discipline

12. Organizations
    Health Care Related Organization
    Organization
    Professional Society
    Self-help or Relief Organization

13. Phenomena
    Biologic Function
    Environmental Effect of Humans
    Human-caused Phenomenon or Process
    Laboratory or Test Result

Natural Phenomenon or Process
Phenomenon or Process

14. Physiology
    Cell Function
    Clinical Attribute
    Genetic Function
    Mental Process
    Molecular Function
    Organ or Tissue Function
    Organism Attribute
    Organism Function
    Physiologic Function

15. Procedures
    Diagnostic Procedure
    Educational Activity
    Health Care Activity
    Laboratory Procedure
    Molecular Biology Research Technique
    Research Activity
    Therapeutic or Preventive Procedure

A.4 Applying MCDA:Detailed Steps

1. **Establish the decision context.**
   1.1. Establish aims of the MCDA,and identify decision makers and other key players.
   1.2. Design the socio-technical system for conducting the MCDA.
   1.3. Consider the context of the appraisal.
2. **Identify the options to be appraised.**
3. **Identify Objectives and Criteria.**
   3.1. Identify criteria for assessing the consequences of each option.
   3.2. Organize the criteria by clustering them under high-level and lower-level objectives in a hierarchy.

4. **"Scoring".Assess the expected performance of each option against the criteria.Then assess the value associated with the consequences of each option for each criterion.**
   4.1. Describe the consequences of the options.
   4.2. Score the options of the criteria.
   4.3. Check the consistency of the scores of each criterion.
5. **"Weighting".Assign weights for each of the criterion to reflect their relative importance to the decision**
6. **Combine the weights and scores for each option to derive an overall value.**
   6.1. Calculate overall weighted scores at each level in the hierarchy.
   6.2. Calculate overall weighted scores.
7. **Examine the results.**
8. **Sensitivity analysis.**
   8.1. Conduct a sensitivity analysis:do other preferences or weights affect the overall ordering of the options.
   8.2. Look at the advantage and disandantage of selected options,and compare pairs of options.
   8.3. Create possible new options that might be better than those originally considered.
   8.4. Repeat the above steps until a "requisite" model is obtained.

Bibliography

[1] Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *JAMA : the journal of the American Medical Association*, 287(20):2691–700, 2002. ISSN 0098-7484. URL http://www.ncbi.nlm.nih.gov/pubmed/12020305.

[2] George R. Klare. Readable computer documentation. *ACM J. Comput. Doc.*, 24(3):148–168, August 2000. ISSN 1527-6805. doi: 10.1145/344599.344645. URL http://doi.acm.org/10.1145/344599.344645.

[3] W Hersh, Chris Buckley, TJ Leone, and D Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. *SIGIR'94*, 1994. URL http://link.springer.com/chapter/10.1007/978-1-4471-2099-5_20.

[4] W Hersh and D Hickam. Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, 82(4):382–9, October 1994. ISSN 0025-7338. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=225961&tool=pmcentrez&rendertype=abstract.

[5] Bethesda (MD): National Library of Medicine (US). Umls reference manual, 2009 Sep. URL http://www.ncbi.nlm.nih.gov/books/NBK9676/.

[6] 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, April 8,2005.

[7] ISO 704. Principles and methods of terminology. Technical report, International Organization for Standardization, Geneva,Switzerland 1986.

[8] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing.* MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

[9] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, pages 1034–1038, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: 10.3115/991250.991317. URL http://dx.doi.org/10.3115/991250.991317.

[10] Didier Bourigault, Isabelle Gonzalez-Mullier, and Cécile Gros. Lexter, a natural language processing tool for terminology extraction. In *Proceedings of 7th EURALEX International Congress*, 1996.

[11] Robert Gaizauskas, George Demetriou, and Kevin Humphreys. Term recognition and classification in biological science journal articles. In *In Proc. of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*, pages 37–44, 2000.

[12] Batrice Daille, Eric Gaussier, and Jean-Marc Lang. Towards automatic extraction of monolingual and bilingual terminology, 1994.

[13] K.T. Frantzi and S. Ananiadou. The C-Value/NC-Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6:145–179, 1999.

[14] Diana Maynard and Sophia Ananiadou. Trucks: a model for automatic multiword term recognition, 2000.

[15] C. Jacquemin. *Spotting and discovering terms through natural language processing.* The MIT Press, 2001.

[16] Akane Yakushiji, Yuka Tateisi, Yusuke Miyao, and Jun ichi Tsujii. Event extraction from biomedical papers using a full parser. In *Pac. Symp. Biocomput*, pages 408–419, 2001.

[17] Kalliopi Zervanou and John McNaught. A domain-independent approach to ie rule development. In *LREC*. European Language Resources Association, 2004.

[18] E Milios, Y Zhang, B He, and L Dong. Automatic term extraction and document similarity in special text corpora. *Proceedings of the sixth conference of Pacific Association for Computational Linguistics*,

2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.3239&rep=rep1&type=pdf.

[19] I Witten, GW Paynter, and Eibe Frank. KEA: Practical automatic keyphrase extraction. *Proceedings of the fourth ACM Conference on Digital Libraries*, 1999. URL http://dl.acm.org/citation.cfm?id=313437.

[20] Yongzheng Zhang, Nur Zincir-Heywood, and Evangelos Milios. Narrative text classification for automatic key phrase extraction in web document corpora. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, pages 51–58, New York, NY, USA, 2005. ACM. ISBN 1-59593-194-5. doi: 10.1145/1097047.1097059. URL http://doi.acm.org/10.1145/1097047.1097059.

[21] Alan R Aronson and François-michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36, May 2010. ISSN 1527-974X. doi: 10.1136/jamia.2009.002733. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995713&tool=pmcentrez&rendertype=abstract.

[22] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA Annual Symposium. AMIA Symposium*, pages 17–21, January 2001. ISSN 1531-605X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243666&tool=pmcentrez&rendertype=abstract.

[23] Alan R Aronson. MetaMap : Mapping Text to the UMLS . 1996.

[24] Alan R Aronson. MetaMap Variant Generation. pages 1–5, 2001. URL http://skr.nlm.nih.gov/papers/references/mm.variants.pdf.

[25] Alan R Aronson. MetaMap Evaluation. pages 1–12, 2001. URL http://skr.nlm.nih.gov/papers/references/mm.evaluation.pdf.

[26] Juan C. Sager. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia, 1990. URL http://ontology.csse.uwa.edu.au/reference/browse_paper.php?pid=233281729.

[27] John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 2 1995. ISSN 1469-8110. doi: 10.1017/S1351324900000048. URL http://journals.cambridge.org/article_S1351324900000048.

[28] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms: the c-value/nc-value method, 2000.

[29] Sophia Ananiadou, Sylvie Albert, and Dietrich Schuhmann. Evaluation of automatic term recognition of nuclear receptors from medline, 2000.

[30] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, and Evangelos E. Milios. Information retrieval by semantic similarity. *Int. J. Semantic Web Inf. Syst.*, 2(3):55–73, 2006.

[31] Yuhua Li, Z.A. Bandar, and D. Mclean. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):871–882, 2003. ISSN 1041-4347. doi: 10.1109/TKDE.2003.1209005.

[32] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2588 of *Lecture Notes in Computer Science*, pages 241–257. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-00532-2. doi: 10.1007/3-540-36456-0_24. URL http://dx.doi.org/10.1007/3-540-36456-0_24.

[33] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[34] Sreerama K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Min. Knowl. Discov.*, 2(4):345–389, 1998.

[35] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[36] J. R. Quinlan. Learning decision tree classifiers. *ACM Computing Surveys*, 28(1):71–72, 1996. URL http://portal.acm.org/citation.cfm?doid=234313.234346.

[37] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[38] Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is np-complete. *Inf. Process. Lett.*, 5(1):15–17, 1976.

[39] Harry G. McLaughlin. SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646, 1969.

[40] P. Ley and T. Florio. The use of readability formulas in health care. *Psychology, Health Medicine*, 1(1):7–28, 1996.

[41] WH DuBay. The principles of readability. 2004. *Costa Mesa: Impact Information*, pages 1–70, 2008. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Principles+of+Readability#1.

[42] JS Dodgson, M Spackman, A Pearman, and LD Phillips. *Multi-criteria analysis: a manual*. 2009. ISBN 9781409810230. URL http://eprints.lse.ac.uk/12761/1/Multi-criteria_Analysis.pdf.

[43] Ehrgott M. Figueira J., Greco S. *Multiple Criteria decision analysis:state of the art surveys*. Springer, 2005. URL http://www.springer.com/business+%26+management/operations+research/book/978-0-387-23067-2.

[44] Rena Peraki, Euripides G. M. Petrakis, and Angelos Hliaoutakis. An information retrieval system for expert and consumer users. *2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE)*, 0:145–150, 2012. doi: http://doi.ieeecomputersociety.org/10.1109/BIBE.2012.6399664.

[45] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999. ISBN 0-201-39829-X.

[46] LL Cherry and W Vesterman. *Writing tools: The STYLE and DICTION programs*. 1981. URL http://art.abc6.com/it/MckMISC/gnuwin32/doc/UnixV7/32dic.pdf.

[47] Remco R Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse. WEKA Manual for Version 3-6-9. 2013. URL http://www.cs.waikato.ac.nz/~ml/weka/.

[48] Carolyn Watters, W. Zheng, and E. Milios. Filtering for medical news items. *Proceedings of the American Society for Information Science and Technology*, 39(1):284–291, 2002. URL http://onlinelibrary.wiley.com/doi/10.1002/meet.1450390131/full.

[49] Yunli Wang. Automatic recognition of text difficulty from consumers health information. In *CBMS*, pages 131–136, 2006.

[50] I. Kuralenok and I. Nekrest'yanov. Automatic document classification based on latent semantic analysis. *Programming and Computer Software*, 26(4): 199–206, 2000. ISSN 0361-7688. doi: 10.1007/BF02759469. URL http://dx.doi.org/10.1007/BF02759469.