Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

An Information Retrieval System for Expert and Consumer Users

Rena Peraki, Euripides G.M. Petrakis, Angelos Hliaoutakis
Department of Electronic and Computer Engineering
Technical University of Crete (TUC)
Chania, Crete, Greece, GR-73100
Email: renaperaki@yahoo.gr, {euripides, angelos}@intelligence.tuc.gr

***Abstract*—The requirements for the design and implementation of a medical information retrieval system for expert and consumer users are discussed. The proposed approach is based on the integration of state-of-the-art tools and methodologies of systems design and document management. Relying on term extraction by natural language processing, the system supports automatic categorization as well as, indexing and retrieval of medical documents by user profile (novice users and experts). This is achieved by mapping document terms to external lexical resources such as WordNet, and MeSH (the medical thesaurus of NLM). Evaluation results of all methods are presented and discussed as well.**

Keywords*—*document categorisation; health informatics; term extraction

I. Introduction

In today's networked world, substantial amounts of medical information are daily becoming available by health-care and research organizations, companies as well as by individuals. The advent of the World Wide Web (WWW) has also generated additional interest in methods and tools supporting efficient management (i.e., storage, retrieval) and communication of this information.

The users of the medical domain can be either health care professionals (experts) or consumers (novice users). Expert users are familiar with the type and content of medical resources (such as dictionaries and databases) and use medical terminology for their searches. However, the spread and availability of medical information over the Web have made this information available also to consumer (i.e., novice) users. Unlike experts, consumers are usually unfamiliar with the content and type of specialized medical text resources and typically use the Web for their searches. They are often uncertain as to the exact type of information they are looking for and they do simple searches using natural language (rather than domain specific) terms. Ensuring reliability of the acquired information over the internet is a challenging and difficult task to deal with. This can be achieved mainly by checking the provenance of information on the internet. Trustworthily information is mainly acquired by Health Care Organizations such as the U.S. National Library of Medicine[1] (N.L.M). Recently, the task of accrediting medical information has been undertaken by the Health On the Net[2] (H.O.N) foundation whose purpose is to promote and guide the deployment of useful and trustworthy information on the WWW.

In our earlier work [1], we investigate on potential improvements to the problem of term extraction related to document representation and indexing in large document collections such as Medline[3], the premier bibliographic database of the U.S. National Library of Medicine (NLM). Using term extraction methods such as AMTEx[1] and MMTx[4], document representation are semantically compact and more efficient, being reduced to a limited number of meaningful multi-word terms (phrases), rather than large vectors of single-words, part of which may be void of distinctive content semantics.

In traditional document management [2], 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 and are often used in indexing medical documents. Multi-word terms carry important classificatory content information, since they comprise of modifiers denoting a specialisation of the more general single-word, head term. For example, the compound term "heart disease" denotes a specific type of disease. In the present work, we show how this information can be used for the automatic categorization of medical documents by user profile (novice users and experts).

An almost orthogonal issue is speed of search. Document indexing and document categorization (i.e., by subject or topic) might not only increase the speed of access to the huge amounts of medical information, but also make this information usable and easily accessible by subject (topic of interest). Without an indexing process, a search for medical information would scan every document in the corpus, which would require considerable time and computing power. Satisfying user requirements while providing fast access to up-date and accredited resources of medical information are prerequisites for the successful implementation of medical information systems. Nevertheless, modeling the design prior to implementation is an essential part for every application

This work is dedicated to the memory of Rena Peraki.

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

[2] http://www.hon.ch

[3] http://www.nlm.nih.gov/databases/databases Medline.html

[4] http://mmtx.nlm.nih.gov

development. System modeling is considered to be successful when the following requirements are met: the system's functionality is formally described and correct, the end-user needs are met, scalability and extensibility are supported, the system's design can be visualized, checked for errors and edited before implementation starts.

Most existing medical information systems aren't really expandable and don't adapt easily to new requirements. In practice, architectural descriptions are informal documents, usually expressed as block diagrams. They are not supported by software tools allowing programmers to make changes to an already running implementation.

This work suggests that the design requirements of medical information systems be described formally using the Unified Modeling Language[5] (UML), the current state-of-the-art language for system design. UML provides a family of diagrammatic notations by which a software system can be described at a high level of abstraction that enforces system modularization, by splitting the system's functionality into a collection of connected components. Each component is a replaceable part of the overall system that fulfils a clear function, evolves independently, can be reused and, later, be replaced. When all of these components have been implemented they are integrated together to form the complete system. An additional benefit of using UML lays in the decoupling of the design from the implementation, allowing certain parts of the system to be redesigned and replaced at a later stage without interfering with the existing running software (e.g., new features are added by adding new code).

As a use case, we consider the design of a medical information system for targeted audiences such as consumer and expert users. To meet this requirement, the system categorizes medical documents by user type (i.e., in documents suitable for consumer or expert users) by assigning to each one a weight representing the belief that the document belongs to each one of the two categories.

The system allows the user to enter free text queries in plain English (most systems place limits on query length). This is particularly convenient especially for consumer users who may not be familiar with the medical terminology. Also, the system supports browsing the database via citation or classification links, a functionality already supported by bibliographic information systems such as CiteSeerX[6] or Publish or Perish[7].

Data and algorithmic resources such as the text extraction methods considered in this work (AMTE$_X$ and MMTx), UMLS and Medline are presented in Sec. II. Architectural design requirements including issues related to document categorization are presented and discussed in Sec. III. Evaluation results are presented in Sec. IV followed by conclusions and issues for further research in Sec. V.

II. Background and Resources

The extraction of terms for the medical, biological and biomedical domain has greatly motivated research for both indexing, as well as knowledge extraction purposes [3], [4]. Automatic indexing and categorisation of medical documents relies mainly on term extraction for the identification of discrete content indicators, namely index terms. Traditional indexing techniques (e.g., the $tf \cdot idf$ method) ignore multi-word and compound terms, which are split into isolated single word index terms. However, compound and multi-word terms are very common in the biomedical domain [5], [6] and are often used in indexing medical documents. Multi-word terms carry important classificatory content information, since they comprise of modifiers denoting a specialisation of the more general single-word, head term. For example, the compound term "heart disease" denotes a specific type of disease.

MedLine documents are currently indexed by human experts by assigning to each one, a number (typically 10 to 12) of terms, based on a controlled list of indexing terms, deriving from a subset of the UMLS[8] (Unified Medical Language System) Metathesaurus, the MeSH[9] (Medical Subject Headings) thesaurus. The automatic mapping of biomedical documents to UMLS term concepts has been undertaken by U.S. National Library of Medicine with the development of MMTx[10] (MetaMap Transfer tool).

The limitations of MMTx in term extraction have been analyzed in detail by Divita et al. [7] The experiments with the MMTx application on MedLine documents have shown that the MMTx not only fails to extract all domain terms, but it also over-generates terms by producing general terms, which diffuse the document concept leading to inaccurate retrieval of MedLine documents. The latter reflects an inherent limitation of MMTx, which was not designed by default to focus on MeSH terms, whereupon MedLine indexing has been based. Additionally, the variant generation process of MMTx is found to account for the over-generation problem for retrieval purposes.

AMTE$_X$ [1] aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC-value method [8]. AMTE$_X$ is based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS Meta-thesaurus mapping of MMTx. It is therefore more selective resulting in more compact document representations than MMTx. In the following, the performance of AMTE$_X$ is compared against the current state-of-the-art, the MetaMap Transfer (MMTx) method using two types of corpora: a subset of Medline (PMC) full document corpus and a subset of Medline (OHSUMED) abstracts. Subsequently, the experimental results demonstrate that AMTEx performs better in indexing in 50% of the processing time compared to MMTx.

Our approach relies on popular knowledge and algorithmic

[5] http://www.uml.org

[6] http://citeseerx.ist.psu.edu

[7] http://www.harzing.com/pop.htm

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

[9] http://www.nlm.nih.gov/mesh

[10] http://mmtx.nlm.nih.gov

resources namely, the UMLS MeSH and Semantic Network[11] for document indexing, MMTx and our AMTE$_X$ extension for term extraction and the Medline collection of biomedical articles for testing their performance. All methods are implemented and their performance is discussed in Sec. IV.

III. System Architecture

Fig. 1 illustrates the generic architecture of a medical information system. It consists of several modules, the most important of them being the document management module.

A. Document Management

The document management module supports automatic analysis, storage indexing and categorization of medical documents. Document analysis relies on term extraction for the identification of discrete content indicators, namely index terms. Documents are indexed, based on a controlled list of indexing terms, deriving from a subset of the UMLS Metathesaurus, the MeSH thesaurus. We focus our attention on multi-word MeSH terms (phrases), and term extraction by AMTEX or MMTx for reducing document representations to a limited number of meaningful multi-word MeSH terms. This information is subsequently used for filtering medical information by user profile.

Both, AMTE$_X$ and MMTx has been shown to be more suitable than single-word term extraction methods not only for document indexing and retrieval [1], but also, for the general concept description and ontology construction tasks [9]. AMTE$_X$ in particular, has been shown to be more selective than the MetaMap Transfer (MMTx) method of NLM which maps arbitrary text to concepts in the UMLS Metathesaurus (equivalently, it discovers Metathesaurus concepts in text).

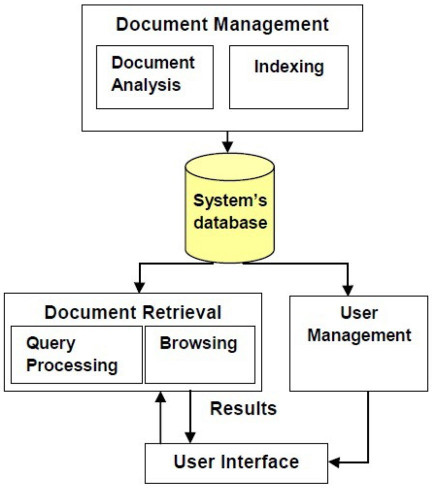

Fig. 1. System Architecture.

In the following, we show how term extraction methods (such as MMTx and AMTEx) can be used for filtering medical information for targeted audiences such as experts and novice users. An obvious application of this filtering operation will be retrieval of medical information by user profile. This approach is automatic and relies on the categorisation of medical terms to terms comprehendible by novice users and to more involved terms typically used by experts (e.g., medical doctors, practitioners etc). This is made possible with the aid of WordNet[12], a thesaurus for natural language terms of the English language. It is based on the observation that up to 30% of the terms participating in MeSH vocabulary are general terms (terms that can be found in WordNet as well) while, the remainder 70% are domain specific UMLS terms that do not belong to WordNet. The performance of the method is assessed using the OHSUMED subset of MedLine based on relevance assessments provided by naive users and experts.

B. Document Categorization by User Profile

This approach is also automatic and relies on the categorization of medical terms to terms comprehendible by novice users and to more involved terms typically used by experts (e.g., medical doctors, practitioners etc). More specifically, MeSH terms can be 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. Fig. 2 illustrates the respective categorization of Medline documents and MeSH terms.

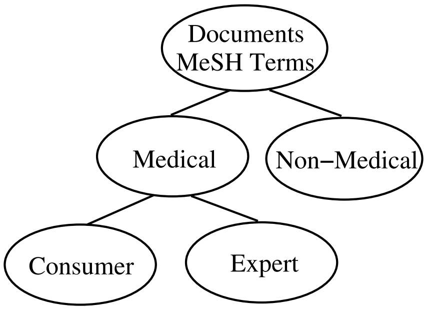

Fig. 2. Categorisation of Medline documents and MeSH terms.

Term and document categorization is realized with the aid of WordNet. By combining information from WordNet and MeSH the following three term vocabularies are constructed:

- *Vocabulary of General Terms (VGT):* These are terms that belong to WordNet but not to MeSH:

$$VGT = (WordNet) - (MeSH)$$

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

- *Vocabulary of Consumer Terms (VCT):* These are medical terms that belong to both, WordNet and MeSH:

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

  It follows that VCT contains 7,165 consumer (MeSH) terms.

[11] http://semanticnetwork.nlm.nih.gov/

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

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

  $$VET = (MeSH) - (WordNet)$$

  It follows that VET contains 16,719 consumer (MeSH) terms.

In this work, document categorisation by user profile relies on the idea of computing the percentage of expert (VET) and consumer (VCT) terms in a document term vector. For example, a document with VET% = 0.62 has 62% probability of being a document suitable for experts. Notice that, general terms are not used in document categorization and that, a non-general term can be either an expert or a consumer term that appear in documents of only one category.

C. Information Retrieval by User Profile in Medline

In the following we design an information retrieval method capable of both i) ranking documents by similarity with a query, and ii) bringing documents matching a given user profile higher in the ranked list of similar documents.

Documents are represented by term vectors [2] extracted by AMTE$_X$ or MMTx respectively. Each term in such a vector is represented by its weight. The term frequency-inverse document frequency model is used for computing the weight of each multi-word term: The weight d_i of a term i in a document is computed as $d_i = tf_i \cdot idf_i$, where tf_i is the frequency of term i in a document and idf_i is the inverse document frequency of i in the whole document collection.

As it is typical in information retrieval (IR), the similarity between a query q and a document d is computed by matching their term vectors according to Vector Space Model (VSM) [2]:

$$Document - Similarity = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2 \sum_i d_i^2}}, \qquad (1)$$

where i denote terms in the query and the document and q_i and d_i are their $tf \cdot idf$ weights in their respective vector representations. More specifically, the query is matched against all Medline documents and the returned list of documents is ranked by decreasing similarity. For ranking query results by user profile we distinguish between the following two cases:

- *Known user profile:* The user identifies her/himself as an expert (or consumer) prior to issuing a query. The similarity score by VSM is multiplied by its percentage of VET (or VCT) terms that is, its probability of being a document for experts (or consumer users respectively).
- *Unknown user profile:* The system determines her/his profile from the query. If the query contains at least one expert term, the user is considered to be an expert (a consumer otherwise). Retrievals are then processed similar to the previous case.

D. Document categorization by subject

A document is also indexed by the two-layered indexing structure of Fig. 3 by mapping the terms of its document vector to their semantic categories of the Semantic Nework[13] (SN) at the top level and then, to their MeSH topics at the lower level. The Semantic Network may be viewed as an upper level ontology of the biomedical domain. The purpose of the Semantic Network is to provide a consistent categorisation of all concepts represented in the Metathesaurus (and therefore in its MeSH subset) and a set of useful relationships among these concepts. Every concept in the Metathesaurus is assigned to at least one semantic type in the Semantic Network. Two high semantic level hierarchies are defined, one for entities related to pathology and one for events (treatment for diseases). This layered index is especially useful for browsing the document collection (e.g., to find documents containing certain terms or even similar or more general terms).

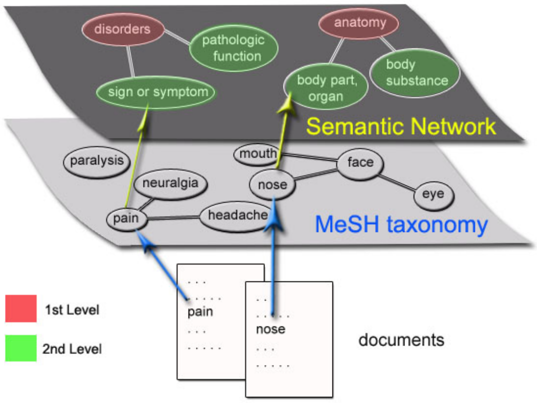

Fig. 3. A two-layered indexing structure for Medline documents.

E. User Management and User Interface

User Management: This module manages the user's database and supports functionality such a user login and user registration. During login the system validates the user (administrator, consumer or expert). User Interface Module: handles system-user interaction. Provides a graphical user interface for performing most user tasks such as query formulation, results presentation and database browsing.

F. Implementation

To demonstrate and objectively asses the quality of the design, a demonstrator medical information system is implemented and discussed in [10]. An early version of the implementation is available on the Web[14]. UML case, package, activities and class diagrams are used to model users, data entities, system functionalities and their inter-relationships. UML diagrams were created using EclipseUML[15]. This tool comes with an automatic code generator for the UML class diagrams, which saves considerable programming effort during the development stage. Data store and access mechanisms are implemented using Lucene[16], a free and open-source

[13] http://semanticnetwork.nlm.nih.gov

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

[15] http://www.ejb3.org

[16] http://lucene.apache.org

information retrieval library written in Java that supports text indexing and searching capabilities. System functionality is also implemented in Java. The performance of the document categorization and retrieval methods has been studied extensively in [11].

IV. Evaluation and Experiments

We conducted two groups of experiments. The first set of experiments is designed to demonstrate the relative effectiveness of AMTE_X and MMTx methods in indexing medical documents. The second group of experiments is designed to demonstrate the categorisation effectiveness of both AMTE_X and MMTx in retrieving medical documents according to user profile. The main data sources used in the experiments are:

- *PMC:* A corpus of 5,819 full documents from PubMed[17] indexed in MedLine selected out of 60 Journals. The documents were selected on the basis of having an identification (UID) number, which was used to retrieve their respective Medline index sets. This index set for each document is manually assigned by Medline experts.
- *OHSUMED:* It is a standard TREC[18] collection of 348,566 medical document abstracts from Medline, published between 1988-1991. OHSUMED is commonly used in benchmark evaluations of IR applications. OHSUMED provides 64 queries and the relevant answer set (documents) for each query. The correct answers were compiled by the editors of OHSUMED and are also available from TREC. For the evaluations, we applied all 64 queries available.

Both, data store and access mechanisms are implemented using Lucene[19]. A document is indexed by the two-layered indexing structure of Fig. 3 by mapping the terms of its document vector to their semantic categories of the Semantic Network at the top level and then, to their MeSH topics at the lower level.

A. Term extraction experiment

In the following, AMTE_X and MMTx are evaluated in terms of precision and recall against the Medline provided MeSH index terms which constitute the ground truth. Because MMTx is slow, a subset of of the OHSUMED TREC collection is selected for this experiment consisting of 10% of OHSUMED (i.e., 34,000 documents). Because it is not possible for a statistically-based term extraction method such as AMTE_X to work for abstracts, we treat the totality of the corpus as a single document during the extraction step. Subsequently, the extracted terms are associated with their respective source abstract. We also run the same experiment using the PMC full document corpus. Table I and Table II below summarize these results.

OHSUMED	MMTx	AMTE_X
Number of Terms	40	8
Precision	0.089	0.125
Recall	0.336	0.101
Time (hours)	14.516	7.383

TABLE I
Average precision and recall of AMTE_X and MMTx on OHSUMED.

PMC	MMTx	AMTE_X
Number of Terms	72	25
Precision	0.033	0.034
Recall	0.162	0.062
Time (hours)	2.727	1.387

TABLE II
Average precision and recall of AMTE_X and MMTx on PMC.

For OHSUMED, AMTE_X demonstrates improved precision and a reasonable recall compared to MMTx by merely a fifth of the average term output of MMTx. For PMC, the results are similar. Notice that MMTx is tuned towards higher recall (by revealing more indexing terms through an exhaustive variant generation phase). In both cases, AMTE_X performs much faster than MMTx. This is due to the algorithmic simplicity of AMTE_Xcompared to MMTx, especially in regards to the variant generation phase of MMTx.

B. Indexing by User Profile

In the following experiment we evaluate our document categorisation method of Sec. III-B. We run a retrieval experiment on the full OHSUMED dataset using VSM [2]. Retrieval using vectors of AMTE_X and MMTx terms is compared with retrieval using vectors of Medline provided MeSH terms (i.e., these terms are used as ground truth). For this experiment, the results were evaluated against all 64 TREC provided queries and answers; 15 out of the 64 queries contain no expert terms and suitable for consumer users. The remaining queries are suitable for experts.

The objective is to measure the ability of a method in retrieving information for consumer and expert users respectively. We run this experiment twice, once for experts and once for consumer users. A method is deemed successful if it retrieves documents suitable for the particular type of users under consideration.

Each method retrieves the best 20 answers for each TREC query, so that each plot below contains exactly 20 points (each method is represented by a curve). The top-left point of a curve corresponds to the average precision/recall values for the best answer or best match (which has rank 1), while the right-most point corresponds to the average precision/recall values for the entire answer set.

Fig. 4 illustrates the relative performance of the three retrieval methods examined for the consumer retrieval task. Retrievals with the manually assigned MeSH terms performs better than any other method. This result reveals a tendency

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

[18] http://trec.nist.gov/data/t9_filtering.html

[19] http://lucene.apache.org

of the human indexers to assign simpler terms for the indexed documents achieving precision close to 65% for small answer sets with up to 10 answers. Both AMTE$_X$ and MMTx perform similarly (the AMTE$_X$ method performs better than MMTx for small answer sets).

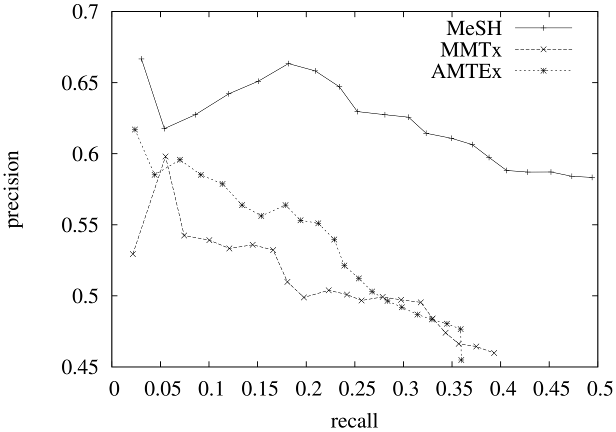

Fig. 4. Average precision-recall of AMTE$_X$ and MMTx for the consumer users retrieval task.

Fig. 5 illustrates results for the retrieval experiment for expert users. AMTE$_X$ outperforms all other methods achieving precision up to 75% for small answer sets (i.e., with up to 3 answers). This experiment demonstrates the selective ability of AMTE$_X$ towards extracting complex medical terms which can be found in the majority of Medline documents. It also reveals a weakness of manually assigning MeSH terms to documents as the human indexers may be not familiar with the content and complexity of domain specific medical publications in Medline.

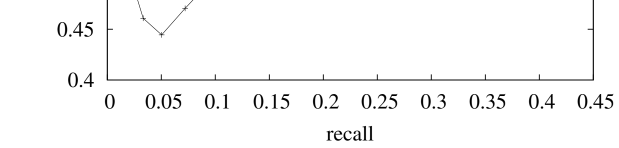

Fig. 5. Average precision-recall of AMTE$_X$ and MMTx for the expert users retrieval task.

V. Conclusions

Issues related to the design and implementation of medical information retrieval systems are discussed. As a case study we introduce to the research community the problem of automatic categorization of medical information by user profile by investigating two common types of users of medical information (i.e., consumers and experts). We also investigate on potential improvements to the problem of indexing medical documents using AMTE$_X$ and we compare its performance against MMTx (the state-of-the-art method of the U.S. NLM). Based on our experiments, we conclude that AMTE$_X$ś selective term output is very well suited for both problems, performing faster than MMTx. However, MMTx's increased recall can be well suited in some retrieval cases, where the small document size is prohibitive for the optimal application of our AMTE$_X$ statistical term extraction process.

More elaborate experimentation is needed for confirming the performance of AMTE$_X$ for general medical collections, such as the Web. For the categorisation of medical documents by user profile problem, future work involves investigation of more elaborate classification methods such as machine learning, fuzzy clustering and document classification.

Acknowledgment

This research has received funding from project YPERTHEN: "Computational Tools and Services for Diagnosis and Personalized Treatment of Cancer", Funding Agency: INTERREG Greece-Cyprus (2011-2013).

References

[1] A. Hliaoutakis, K. Zervanou, and E. Petrakis, "The AMTEx Approach in the Medical Document Indexing and Retrieval Application," *Data and Knowledge Engineering*, vol. 68, no. 3, pp. 380–392, March 2009.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley Longman, 1999.

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

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

[5] D. Maynard and S. Ananiadou, "TRUCKS: A Model for Automatic Multi-Word Term Recognition," *Journal of Natural Language Processing*, vol. 8, no. 1, pp. 101–125, 2000.

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

[7] G. Divita, T. Tse, and L. Roth, "Failure Analysis of MetaMap Transfer (MMTx)," in *Medinfo: Proc. of the 11th World Congress on Medical Informatics*. San Francisco: IOS Press, Sept. 2004, pp. 763–765.

[8] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-Word Terms: The C-Value/NC-value Method," *International Journal of Digital Libraries*, vol. 3, no. 2, pp. 117–132, 2000.

[9] E. Drymonas, K. Zervanou, and E. Petrakis, "Unsupervised Ontology Acquisition from Plain Texts: the OntoGain System," in *15th Intern.l Conf. on Applications of Natural Language to Information Systems (NLDB'2010)*. Cardiff, Wales, UK: Springer, LNCS 6117, Jun 2010, pp. 277–287.

[10] R. Peraki, "Requirements Analysis for Medical Information System Design," Chania, Crete, Greece, April 2008. [Online]. Available: http://www.intelligence.tuc.gr/publications.php?pub_author=11&pub_type=9&pub_subject=All

[11] A. Hliaoutakis, "Automatic Term Indexing in Medical Text Corpora and its Applications to Consumer Health Information Systems," Chania, Crete, Greece, December 2009. [Online]. Available: http://www.intelligence.tuc.gr/publications.php?pub_author=4&pub_type=9&pub_subject=All