2005

Extracting Conceptual Terms from Medical Documents

Quanzhhi Li
New Jersey Institute of Technology, ql23@njit.edu

Yi-Fang Brook Wu
New Jersey Institute of Technology, wu@njit.edu

Xin Chen
New Jersey Institute of Technology, xc7@njit.edu

Ravzan Stefan Bot
New Jersey Institute of Technology, rsb2@njit.edu

Extracting Conceptual Terms from Medical Documents

Quanzhi Li
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
QL23@njit.edu

Yi-Fang Brook Wu
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
Wu@njit.edu

Xin Chen
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
xc7@njit.edu

Razvan Stefan Bot
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
rsb2@njit.edu

ABSTRACT

Automated biomedical concept recognition is important for biomedical document retrieval and text mining research. In this paper, we describe a two-step concept extraction technique for documents in biomedical domain. Step one includes noun phrase extraction, which can automatically extract noun phrases from medical documents. Extracted noun phrases are used as concept term candidates which become inputs of next step. Step two includes keyphrase extraction, which can automatically identify important topical terms from candidate terms. Experiments were conducted to evaluate results of both steps. The experiment results show that our noun phrase extractor is effective in identifying noun phrases from medical documents, so is the keyphrase extractor in identifying document conceptual terms.

Keyword

Noun phrase, Noun Phrase Extraction, Conceptual Term, Keyphrase Extraction, Medical document

BACKGROUND

The pervasion of medical information via the WWW has created a continuously growing need for the development of techniques to assist medical professionals and researchers to access and share medical information. The concepts in textual documents are usually described by noun phrases, and they carry the primary information of documents. Evidences have shown that noun phrases help readers to understand, organize, access and share information of a document. Many studies pertaining applications of phrases focus on retrieval system and browsing interface (Croft, Turtle and Lewis, 1991; Edgar, Nichols, Paynter, Thomson and Witten, 2003; Fagan, 1989); some others explore their applicatuions on document classification and clustering (Zamir and Etzioni, 1999; Larkey, 1999).

Many biomedical document analysis or retrieval studies have used documents from MEDLINE, the premier bibliographic database of the National Library of Medicine (NLM). Blake and Pratt (2002) use noun phrase as the concept terms to detect the connections among medical literature. Their study based on MEDLINE shows that using noun phrases would be more effective in finding complementary literature. Kumar et al (2004) describe an approach called BioMap, to using noun phrases extracted from the abstracts of MEDLINE collections to build a knowledge base for biomedical literature.
Croft et al. (1991) propose a method where phrases identified in natural language queries are used to build structured queries for a probabilistic retrieval model. Their experimental results show that retrieval performance can be improved by using phrases this way, and phrases extracted automatically from a natural language query perform nearly as well as manually selected phrases. Fagan's (1989) study shows that phrase-based automatic indexing helps to improve the precision of the overall document retrieval. Larkey (1999) develops a system for searching and classifying U.S. patent documents, based on INQUERY. The system includes a "phrase help" facility, which can help users find and add phrases and terms related to those in their query. The phrases are built from historical patent text, using a set of heuristics.

Several noun phrase extraction techniques have been introduced in previous studies. Some of them are described below. Chen et al. (1997) developed a noun phrase extraction system called FastNPE. It mainly relies on concatenation of adjacent tokens to identify phrases. Later, they revised the above system and developed a new system, which is called AZ Phraser (Bennett et al, 1999). AZ Phraiser's part-of-speech tagging is based on earlier work of Brill. Their tagger is divided into two main phases of operation - lexical analysis and contextual analysis. The lexicon mostly comprises the Wall Street Journal

corpus and the Brown corpus. The contextual analysis uses several contextual rules. The contextual analysis phase is to ensure that the part-of-speech tags are disambiguated. NPtool (Voutilainen, 1993) is a commercial noun phrase extraction program. After preprocessing the documents, it has the following three steps: morphological analysis, constraint grammar parsing, NP-hostile and friendly finite state parsing, and NP extraction.

In this paper, we will introduce two systems. The first one is a generic noun phrase extractor. The main differences between our noun phrase extractor and other systems are that our system uses a WordNet lexical database (Fellbaum, 1998) as our lexicon to do the part-of-speech tagging, and our system does not need any training data. It is a generic noun phrase extractor, and can be used in any domain without modification or fine-tuning. We will describe its algorithm and the evaluation in the next section.

One limitation with a generic noun phrase extractor is that it extracts noun phrases regardless of the domain of knowledge within which a particular document is situated. When extracting noun phrases from a document, it simply extracts all the noun phrases from documents, without indicating a term's degree of relevance to the main topics of the document. This might be reasonable, should a general document collection, such as newspaper articles be used. However, when a specialized document collection is presented, a more domain-oriented concept extraction would be desirable. If we integrate the domain knowledge with the noun phrase extractor, we can extract the concepts which are semantically relevant to the topical theme of a document. This is the goal of our second system, which is called Keyphrase Identification Program (KIP) (Wu et al, 2004). It is built on top of the noun phrase extractor. After all the noun phrases of a document are identified, KIP will rank all the noun phrases in terms of their degree of relevance to the main theme of the document, and select only the important ones. KIP fulfills this task by considering the domain of a document. We call the most important topical terms for a document as "keyphrases." Document keyphrases provide a concise summary of a document's content, offering semantic metadata summarizing and characterizing a document.

Previous studies have shown that document keyphrases can be used in a variety of applications, such as retrieval engines (Li et al, 2004; Jones and Staveley, 1999), browsing interfaces (Gutwin et al, 1999), thesaurus construction (Kosovac et al, 2000), and document classification and clustering (Jonse and Mahoui, 2000). Several automatic keyphrase extraction techniques have been proposed in previous studies. Turney (2000) is the first person who treats the problem of phrase extraction as supervised learning from examples. Turney uses nine features to score a candidate phrase; some of the features are the location of the first occurrence of the phrase in the document and whether or not the phrase is a proper noun. Keyphrases are extracted from candidate phrases based on examination of their features. Turney's program is called Extractor. Kea, a keyphrase extraction program developed by Frank et al (1999), uses a machine learning algorithm which is based on naïve Bayes' decision rule. It has some pre-built models. A model is used to identify the keyphrases within a document. The model is learned from the training documents with exemplar keyphrases and corresponds to a specific corpus containing the training documents. Each model consists of a Naive Bayes classifier and two supporting files, which contain phrase frequencies and stopped words.

To extract keyphrase from documents, KIP's algorithm considers the composition of a noun phrase. To analyze a noun phrase and assign a score for it, KIP uses a glossary database, which contains pre-identified medical keyphrases and keywords, to calculate scores of noun phrases in a document. The noun phrases having higher scores will be extracted as keyphrases.

In the following sections, we first describe our generic noun phrase extractor and the experiments based on medical domain. Then we describe our second system, KIP, and its evaluation based on medical documents.

EXTRACTING NOUN PHRASES FORM MEDICAL DOUMENTS

In this section, we first describe the algorithm of our generic noun phrase extractor. Then we present two experiments used to evaluate its performance with medical documents. The noun phrase extractor has two main components: a part-of-speech tagger and a noun phrase extraction component.

Part of Speech Tagger

After being loaded into the system, a document is first parsed into sentences. Then sentences are tokenized to obtain the atom units, each of which could be a punctuation mark or a word. Each word is assigned with an initial part-of-speech (POS) tag. To assign the right tag, we use a WordNet lexical database (Fellbaum, 1998), which contains words divided into four categories (noun, verb, adjective, and adverb) and the number of senses of each word used in the categories it belongs to. If a word is found in more than one category, it is marked as a multi-tag word. The initial POS tag for a word is determined by the category having the maximum number of senses of this word. The next step is multi-tag disambiguation. For every multi-tag word, the sequence of the POS tags of the proceeding n tokens (n ranges from 2 to 4) is examined against a list of predefined syntactic rules. For example, "hit" can be either a noun or a verb. If the proceeding word is a determiner (the, a,

this, etc), it will be tagged as a noun rather than a verb and the multi-tag mark is removed. If a word is not found in any of the categories and its POS tag cannot be solved by the syntactic rules, some heuristics are used to determine its POS tag. For instance, if a word is not found in the lexical database, but ends with "tion", it is tagged as a noun.

Noun Phrase Extraction

In general, a noun phrase means a sequence of words that usually gives us very useful information. People mostly use noun phrase as concept terms. After tagging the text, the noun phrase extractor extracts noun phrases by selecting the sequence of POS tags that are of interests. The current sequence pattern is defined as [A] {N}, where A refers to Adjective, N refers to Noun, [ ] means optional, and { } means repetition. A set of exceptional rules is used as well. The system has a system parameter to set the minimum and maximum numbers of words of a noun phrase. By changing the parameter value user can get noun phrases with different length. The system can extract noun phrase with length of one word to 8 words. The default setting is to extract all the noun phrases with length equal to or more than two words. At this stage, the system produces a list of noun phrases for the processed document.

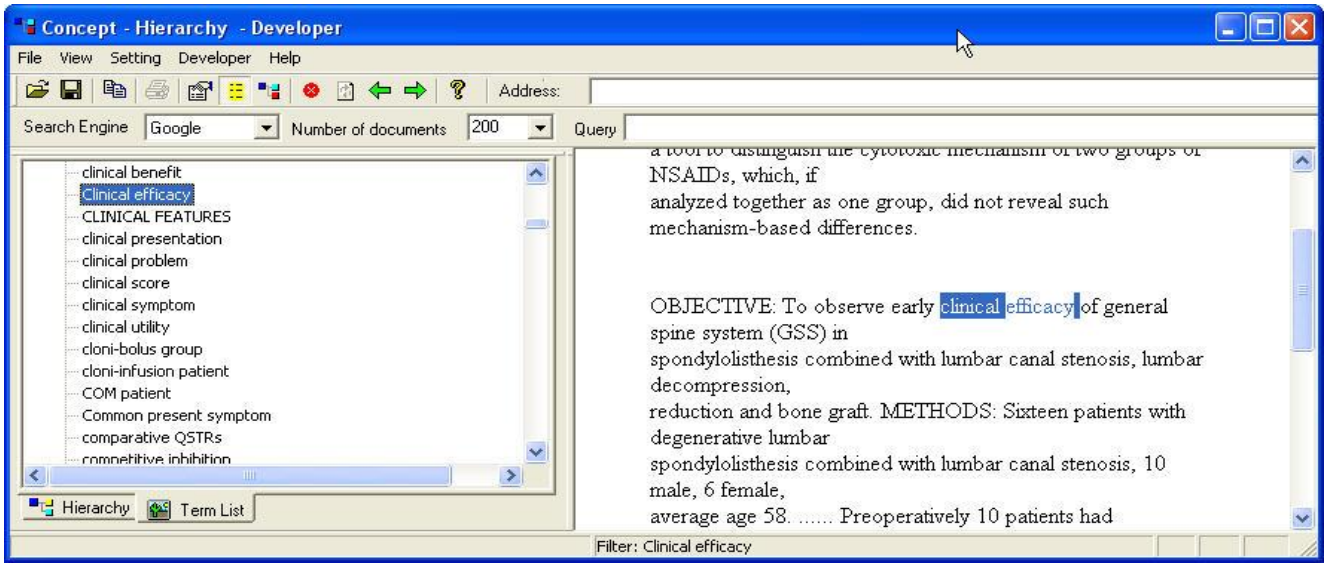

Figure 1. A Screenshot of the Noun Phrase Extractor

Figure 1 shows a screenshot of the noun phrase extractor. The extracted noun phrases are displayed in the left frame, and the related paragraph is displayed in the right frame. The generated noun phrases are also automatically sent to a file, with related information, like the phrase frequency in a document. This program also has other functions, but in this paper we just describe the function of extracting noun phrases from documents.

Experiments

We have tested the noun phrase extractor in other domains, but we have not tested it with medical documents. To test its performance in medical domain, we performed two small experiments with medical documents. In experiment 1 we calculated the precision and recall based on a small document collection which contained 60 medical documents. In experiment 2, we computed only the precision based on 500 medical documents.

Experiment 1

In this experiment, we assessed the effectiveness of the system by computing its precision and recall. Precision is the number of phrases correctly identified by the noun phrase extractor, divided by the total number of system-identified noun phrases. Recall is defined as the number of noun phrases correctly identified by the system, divided by the total number of noun phrases manually identified by human experts. Many previous studies have used precision and recall to evaluate the performance of noun phrase extraction systems (Tolle and Chen, 2000; Bennet et al, 1999).

We recruited two biomedical professionals from a large pharmaceutical company to identify the noun phrases for our test documents. The test documents were collected from the website of National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). *Entrez* is an integrated, text-based search and retrieval system used at NCBI for the

major medical databases. We used *Entrez* to collect the test documents. The steps of collecting the test documents are as follows: first we did a search using term "pain relief" at *Entrez*, and 18,937 hits were returned; among the returned hits we randomly selected 60 documents, and each of these 60 documents contained only the abstract and title. All the 60 documents were loaded in our system and the noun phrases were output to a file as well as the user interface, as displayed in Figure 1. In order to calculate the recall, all the noun phrases of the test documents should be pre-identified. Because manually identifying noun phrases from documents is time-consuming, so in this experiment we used only a small test collection containing only 60 documents. Each of the two experts was asked to identify all the noun phrases for 30 documents.

The total number of noun phrases identified by the human experts was 4,045 for the 60 documents. The total number of noun phrases extracted by the noun phrase extractor was 3,638. Among these extracted noun phrases, 3,526 of them were correctly identified, which means they were real noun phrases. The result is displayed in Table 1. There is usually a trade-off between precision and recall, and either of them alone does not paint a complete picture of system effectiveness. Therefore, the F measure was invented to show the combined results. The formula is: F = 2*precision*recall / (recall + precision).

Total number of noun phrases identified by human experts	Total number of noun phrases extracted by the noun phrase extractor	Total number of noun phrases correctly identified by the noun phrase extractor	Precision	Recall	F
4045	3,638	3,526	0.97	0.87	0.91

Table 1. Precision and Recall of the Noun Phrase Extractor

In order to compare the performance of the three noun phrase extractors (FastNPE, AZ Phraser and Chopper), Bennett et al. (1999) did an experiment based on 40 medical documents abstracts. The reported results are as follows: for FastNPE, the precision is 0.80 and the recall is 0.50; for Chopper, which parses a text by breaking it down into constituent sentences or phrases, the precision is 0.90 and the recall is 0.97; and for AZ Phraser, the precision is 0.86 and the recall is 0.92. Because our experiment used a different data set from theirs (although they are all from medical domain), so we do not directly compare our system to those systems reported in their paper. It might be interesting to mention that the F value for FastNPE, AZ Phraser and Chopper is 0.62, 0.93, and 0.88, respectively.

Experiment 2

As mentioned in experiment 1, in order to calculate the recall, all the noun phrases of the test documents should be pre-identified manually by human experts. Manually identifying noun phrase from documents is time-consuming and costly. Because we needed to calculate the recall in experiment 1, so we used only 60 documents in experiment 1. Compared to computing recall, computing precision requires less human effort, since it does not require manually identifying noun phrases from documents. So, it is possible to use a larger document collection to compute the precision. In this experiment, we used 500 documents to calculate the precision of our noun phrase extractor. Each of the 500 documents contained only the abstract and title. They were randomly selected from MEDLINE. The result is shown in Table 2. This value was the same as the result of experiment 1. This result also shows that the performance of our program is consistent when dealing with the medical documents. When being applied to the medical documents, the system could identify noun phrases effectively.

Number of Documents	Total number of noun phrases extracted by the noun phrase extractor	Total number of noun phrases correctly identified by the noun phrase extractor	Precision
500	36,345	35,248	0.97

Table 2. Precision of the Noun Phrase Extractor based on 500 Medical Documents

EXTRACTING KEYPHRASES FROM MEDICAL DOCUMENTS

In last section, we have discussed a generic noun phrase extractor and its evaluation using medical documents. One limitation with applying the noun phrase extractor in medical domain is that it does not consider the context of a document. If the domain knowledge can be integrated with the generic noun phrase extractor, we can extract the concepts which are really semantically relevant to the main topical theme of the document. To extract the topical concepts (keyphrases) from

documents, we developed a system called Keyphrase Identification Program (KIP) (Wu et al, 2004). In the following subsections, we describe KIP's algorithm first, then we present its evaluation.

KIP's Algorithm

KIP is based on the noun phrase extractor described in last section. After the noun phrase extractor has extracted all the noun phrases from a document, KIP will assign scores to these phrases, rank them and extract the ones with higher scores. KIP is a domain-specific keyphrase extraction program. Its algorithm is based on the logic that a noun phrase containing domain-specific keywords and/or keyphrases is likely to be a keyphrase of the document. The more keywords/keyphrases it contains and the more significant the keywords/keyphrases are, the more likely that this noun phrase is a keyphrase. The pre-identified domain-specific keywords and keyphrases are stored in a glossary database, which is used to calculate scores of noun phrases. Here a pre-defined domain-specific keyword means a single term word, and a pre-defined domain-specific keyphrase means a phrase containing one or more words. KIP operations can be summarized as follows. KIP first get a list of keyphrase candidates, which are noun phrases generated by the noun phrase extractor. Then it examines the composition of a keyphrase candidate (a noun phrase) and assigns a score to it. The score of a noun phrase is determined mainly based on three factors: its frequency of occurrence in the document, its composition (what words and sub-phrases it contains), and how specific these words and sub-phrases are in the domain of the document. To calculate scores of noun phrases, readily available pre-identified domain-specific keyphrases are parsed to form a glossary database. Finally, the noun phrases with higher scores are selected as keyphrases of the document.

In order to calculate the scores for noun phrases, we use a glossary database containing domain-specific keyphrases and keywords, which provide initial weights for the words and sub-phrases of a candidate keyphrase. In the following paragraphs, we will first describe how to build this database, then how to calculate a noun phrase's score, and finally how the keyphrases are extracted.

The glossary database has two lists (tables): (a) a keyphrase list and (b) a keyword list. A keyphrase is an entry in the pre-defined keyphrase list, and it could contain one or more words; and a keyword means a single word parsed from list (a). Before using KIP, users will need a corresponding glossary database from a particular domain. When the system is applied to a new domain, the only thing required is to build or change to a new database specific to the domain. In this study, we are going to extract keyphrases for medical documents. So we use MeSH to build our glossary database. MeSH is NLM's controlled vocabulary thesaurus. It consists of a lot of medical terms in a hierarchical structure.

The keyphrase list was generated adding all the MeSH terms to it. The keyword list was automatically generated from the keyphrase list. To obtain the keywords, all keyphrases (MeSH terms) were split into individual words and added as keywords to the keyword list. The glossary database has two tables, one for keyphrases and another for keywords. The keyphrase table and keyword table all have two columns (keyphrases/keywords and weights). The weights for keyphrases and keywords are assigned automatically. The rationale behind the method of assigning weights is that it reflects how specific a keyword or keyphrase is in a specific domain. KIP will use the weights of keyphrases and keywords in the database to calculate the scores of noun phrases in a document.

A noun phrase's score is defined by multiplying a factor F by a factor S. F is the frequency of this phrase in the document, and S is the sum of weights of all the individual words and all the possible combinations of adjacent words within the noun phrase (we call a combination of adjacent words a "sub-phrase" of this noun phrase). So we have the following equation:

The score of a noun phrase = $F \times S$.

The following example is used to explain how a noun phrase's score is calculated. Assume there is a noun phrase "ABC," where A, B and C are three words. The possible combinations of adjacent words are AB, BC and ABC. The score for noun phrase "ABC" will be the frequency of "ABC" in this document multiplied by the summation of weights of A, B, C, AB, BC, and ABC. The motivation for including the weights of all possible sub-phrases into the phrase score, in addition to the weights of individual words, is to find out if a sub-phrase is a keyphrase in the glossary database. If it is, this phrase is expected to be more important. KIP will lookup the keyphrase table to obtain the weights for all the sub-phrases of the noun phrase. If a sub-phrase is found, the corresponding weight in the keyphrase table is assigned to this sub-phrase; otherwise, a predefined low weight will be assigned to this sub-phrase. Similarly, KIP obtains the weight of a word by looking up the keyword table. If it finds the word from the table, the corresponding weight in the keyword table will be the weight of the word. Otherwise, a predefined weight will be assigned to it.

All the scores of noun phrases are normalized to range from 0 to 1 after they are calculated. Noun phrases in the document are then ranked in descending order by their scores. The keyphrases of a document can be extracted from the ranked noun

phrase list. In order to be as flexible as possible, the KIP system has a set of parameters to let the users decide how many keyphrases they want.

Experiment

Usually, there are two ways to evaluate the effectiveness of a keyphrase extraction system. One is to use human judgment, asking domain experts to rate the keyphrases generated by the system. The second way is to measure how well the system-generated keyphrases match the author-provided keyphrases. We chose the second approach and assessed KIP's effectiveness with medical documents by computing its precision and recall using author-provided keyphrases for documents. In this experiment, precision means the proportion of the extracted keyphrases that match the keyphrases assigned by a document's author(s). Recall means the proportion of the keyphrases assigned by a document's author(s) that appear in the set of keyphrases generated by the keyphrase extraction system. Measuring precision and recall against author keyphrases is easy to carry out, since it does not involve human subjects. Previous studies have used this measure and found it is an appropriate method to measure the effectiveness of a keyphrase extraction system (Jones and Paynter, 2002; Turney, 2000; Frank et al, 1999). We used 100 medical papers as the test documents in this evaluation. Seventy of them were from journal of Medical Informatics & the Internet in Medicine 2002-2003, and thirty of them were from Journal of Applied Clinical Medical Physics 2004. All these 100 papers have author-assigned keyphrases. Author-assigned keyphrases were removed from the papers before the documents were processed by KIP. The average length of these papers was 14 pages. The average number of author-assigned keyphrases for these papers was 4. We calculated the precision and recall when the number of extracted keyphrases was 5, 10 and 15, respectively. The result is shown in Table 3.

Number of extracted keyphrases	Average Precision ± Standard Deviation	Average Recall ± Standard Deviation
5	0.26±0.14	0.34±0.19
10	0.19±0.08	0.50±0.22
15	0.14±0.06	0.56±0.23

Table 3. Precision and Recall of KIP with Medical Documents

We need to point out that some author-provided keyphrases may not occur in the document they are assigned to. In experiments reported by Turney (2000), about only 75% of author-provided keyphrases appear somewhere in the document. That means the highest recall for a system could only be 0.75. In his experiment, Turney (2000) reports that his system's precision is 0.24 and 0.13 when the number of extracted keyphrases is 5 and 15, respectively. In Jones and Paynter's experiment (2002), Kea's precision and recall is 0.24 and 0.32, respectively, when the number of extracted keyphrases is 15. Because the data set are different, we do not directly compare our system to theirs. We report some of their results here for reference.

FUTURE STUDY AND CONCLUSION

Biomedical professionals have diverse methods and perspectives to solving problems. The terminology used by different professionals and applications make the communications among them difficult. To solve this problem, the unified medical language system (UMLS) of NLM has attempted to integrate a number of medical terminologies into a unified knowledge source. The UMLS contains three knowledge sources: the semantic network, the metathesaurus, and the Specialist lexicon. The Specialist Lexicon is an English language lexicon. The main difference between Specialist lexicon and other lexicons is that it contains many biomedical terms. Right now our noun phrase is mainly based on WordNet, which contains all the general English words. We will investigate how the system performs when using Specialist Lexicon, which is more suitable than WordNet. Our future work includes integrating Specialist Lexicon into our system, so that our system will be more effective in identifying biomedical words and noun phrases. Another future work will be to use human subjects to evaluate the generated keyphrases from medical documents by KIP.

In this paper, we describe a generic noun phrase extractor and a keyphrase extractor. We also report our experimental results based on medical documents. The experiment results show that the noun phrase extractor is effective in identifying noun phrase for medical documents, and the keyphrase extractor can extract topical phrases for medical documents.

REFERENCE

1. Arampatzis, A. T., Tsoris, T., Koster, C. H. A., and Weide, T. (1998). Phrase-based Information Retrieval, *Information Processing & management*, Vol. 34, No. 6, pp.693-707.
2. Bennett NA, He Q, Powell K, Schatz BR (1999). Extracting noun phrases for all of MEDLINE. *Proc AMIA Symp.* 671-5.
3. Blake, C. & Pratt, W. (2002).A Semantic Approach to Identify Candidate Treatments from Existing Medical Literature. *In AAAI Symposium on Knowledge-based Approaches*, Stanford, CA
4. Chen, H., T.D. Ng, J. Martinez, B.R. Schatz (1997), A Concept Space Approach to Addressing the Vocabulary Switching Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *JASIS*, 48(1), pp. 17-31.
5. Croft, B., Turtle, H, & Lewis, D. (1991). The use of phrases and structured queries in information retrieval. *Proceeding of SIGIR'91: The 14th Annual International Conference on Research and Development in Information Retrieval*, Philadelphia, PA, ACM Press.
6. Edgar, K. D., Nichols, D. M., Paynter, G., Thomson, K., & Witten, I. H. (2003). A User Evaluation of Hierarchical Phrase Browsing. *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*, Trondheim, Norway.
7. Fagan, J. L. (1989). The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document. Retrieval. *Journal of the American society for information Science* 40(2): 115-132.
8. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MIT Press.
9. Frank, E., Paynter, G., Witten, I. H., Gutwin, C., & Nevill-Manning, C. (1999). Domain-specific Keyphrase Extraction. *Proceeding of the sixteenth international joint conference on artificial intelligence*, San Mateo, CA.
10. Gutwin, C., Paynter, G., Witten, I. H., Nevill-Manning, C., & Frank, E. (1999). Improving Browsing in Digital Libraries with Keyphrase Indexes. *Journal of. Decision Support Systems* 27(1-2): 81-104.
11. Jonse, S., & Mahoui, M. (2000). Hierarchical document clustering using automatically extracted keyphrase. *Proceeding of the third international Asian conference on digital libraries* Seoul, Korea. 113-120.
12. Jones, S., & Paynter, G. W. (2002). Automatic extraction of Document Keyphrases for Use in Digital Libraries: Evaluation and Applications. *Journal of the American Society for Information Science and Technology* 53(8): 653-677.
13. Jones, S., & Staveley, M. (1999). Phrasier: A system for interactive document retrieval using keyphrases. *Proceedings of SIGIR'99*, Berkeley, CA.
14. Kosovac, B., Vanier, D. J., & Froese, T. M. (2000). Use of keyphrase extraction software for creation of an AEC/FM thesaurus. *Electronic Journal of Information Technology in Construction* 5: 25-36.
15. Kumar, K. , Palakal, M. , Mukhopadhyay, S., Stephens, M. & Li, H (2004). BioMap: Toward the Development od a Knowledge Base of Biomedical Literature. *2004 ACM Symposium on Applied Computing*, Nicosia, Cyprus. pp121-127.
16. Larkey, L. S. (1999). A Patent Search and Classification System. *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, Berkeley, CA, ACM Press.
17. Li, Q., Wu, Y .B., Bot, R. S., Chen, X. (2004). Incorporating Document Keyphrases in Search Results. *Proceedings of the Tenth Americas Conference on Information Systems*, New York, New York.
18. Tolle, K. M. and Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools, *Journal of the American Society for Information Science*, 51(4), pp. 352-370.
19. Turney, P. D. (2000). Learning algorithm for keyphrase extraction. *Information Retrieval* 2(4): 303-336.
20. Voutilainen, A. (1993). NPtool: A Detector of English Noun Phrases. Proc. *Workshop on Very Large Corpora, Columbus*, OH, June 22.
21. Wu, B. Y., Li, Q., Bot, R. S., Chen, X. (2004) KIP: A Keyphrase Identification Program with Learning Functions. *Proceedings of International Conference on Information Technology: Coding and Computing (ITCC'04)*, Las Vegas, NV, Volume 2, 450-455.
22. Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. *Computer Networks and ISDN Systems*. 31(11-16): 1361-1374.