Automatically Finding Significant Topical Terms from 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

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

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

ABSTRACT

With the pervasion of digital textual data, text mining is becoming more and more important to deriving competitive advantages. One factor for successful text mining applications is the ability of finding significant topical terms for discovering interesting patterns or relationships. Document keyphrases are phrases carrying the most important topical concepts for a given document. In many applications, keyphrases as textual elements are better suited for text mining and could provide more discriminating power than single words. This paper describes an automatic keyphrase identification program (KIP). KIP's algorithm examines the composition of noun phrases and calculates their scores by looking up a domain-specific glossary database; the ones with higher scores are extracted as keyphrases. KIP's learning function can enrich its glossary database by automatically adding new identified keyphrases. KIP's personalization feature allows the user build a glossary database specifically suitable for the area of his/her interest.

Keywords

Keyphrase Extraction, Document keyphrase, Document Metadata, Text Mining, Glossary Database.

INTRODUCTION

Document keyphrases are the most important topical phrases for a given document. They address the main topics of the document and provide semantic metadata which can summarize the document. One of the keys to successful text mining applications is the ability of extracting important document features upon which the mining algorithm is performed to discover interesting patterns or relationships. However, many text mining applications do not have adequate natural language processing ability beyond simple keyword indexing, and as a result, there are too many textual elements (words) included in the analysis. Previous studies have shown that, in some text mining related applications, keyphrases as textual elements are better suited and could provide more discriminating power than single words (Jonse and Mahoui, 2000; Witten, 1999; Gutwin et al., 1999). Document keyphrases can also be used in many other applications, such as automatic text summarization, development of search engines, document clustering, document classification, browsing interface, and thesaurus/glossary construction.

Most documents do not have author-assigned keyphrases; only a few, mostly scholarly papers, have a list of keyphrases provided by authors. Keyphrases can also be assigned manually by experts or professional indexers. However, manually assigning keyphrases to documents is costly and tedious, and the results may not be consistent. So there is a need for automatic keyphrase generation techniques. There are two methods for automatic keyphrase generation: keyphrase assignment and keyphrase extraction. The keyphrase assignment method chooses phrases that best describe a document from a controlled vocabulary. Keyphrase extraction techniques choose phrases from the document text as keyphrases. The problem with keyphrase assignment is that most controlled vocabularies are not updated frequently enough, and such controlled vocabularies might not even be available in some domains. Therefore, automatic keyphrase extraction method is desirable.

In this paper, a keyphrase identification program (KIP) is described. KIP is a keyphrase extraction technique. This algorithm considers the composition of a noun phrase. To analyze a noun phrase and assign a score to it, KIP uses a glossary database, which contains pre-identified domain-specific keyphrases and keywords. The noun phrases having higher scores will be extracted as keyphrases. In this paper, we will also discuss KIP's two special features: the learning function and the personalization feature. The learning function can enrich the glossary database by automatically adding new keyphrases extracted from documents to the database. The personalization feature will let a user build a glossary database specifically tailored for the area of his/her interest. Thereafter, using this personalized glossary database, KIP can extract keyphrases more effectively from the documents of the user's interest. Following a brief review of previous studies, the design of KIP, its learning function and personalization feature are described in details; finally, the experiment and result are presented.

RELATED WORK

Document keyphrases can be used in many applications, such as browsing interface (Gutwin et al., 1999; Jones and Paynter, 1999), retrieval engine (Jones and Staveley, 1999; Li et al., 2004), document classification and clustering (Jonse and Mahoui, 2000; Witten, 1999), and thesaurus construction (Kosovac et al, 2000). Several automatic keyphrase extraction techniques have been proposed in previous studies.

A keyphrase extraction method based on modeling documents as weighted undirected and weighted bipartite graphs is proposed by Zha (2002) In this approach, spectral graph clustering algorithms are used for partitioning sentences of the documents into topical groups. Within each topical group, the mutual reinforcement principle is used to compute keyphrase and sentence saliency scores. The keyphrases and sentences are ranked according to their saliency scores. Then keyphrases are selected based on their scores. In this approach, a phrase's frequency in the document is the dominant factor of its score. The paper does not report any evaluation result.

Krulwich and Burkey (1996) use some heuristics to extract significant topical phrases from a document. The heuristics are based on documents' structural features, such as the presence of phrases in document section headers and the use of italics. This approach is not difficult to implement, but the limitation is that not every document has explicit structural features.
Kea is a machine learning algorithm based on naïve Bayes' decision rule (Frank et al., 1999; Witten et al., 1999). Kea has some pre-built models. A model is used to identify the keyphrases for a document. The model is learned from the training documents with exemplar keyphrases. Each model consists of a naive Bayes classifier and two supporting files containing phrase frequencies and stopped words. Once it is learned from the training documents, a model can be used to identify keyphrases from other documents.

Turney (2000) treats the problem of phrase extraction as supervised learning from examples. Nine features are used to score a candidate phrase, such as the location of the first occurrence of the phrase in the document. Keyphrases are extracted from candidate phrases based on examination of their features. Turney's system, GenEx, has two components: Extractor and Gentor. Extractor processes a document and produces a list of phrases based on the setting of 12 parameters. In the training stage, Gentor is used to tune the parameter setting to get the optimal performance. Once the training process is finished, Extractor alone can extract keyphrases. In Extractor's formula to calculate a phrase's score, the dominant factors are the frequency of the phrase, the frequencies of words within it, and the location of its first occurrence.

Extractor and Kea use supervised machine learning approaches. They all use some corpora to train the program. For each document in the corpus, there must be a target set of keyphrases provided by authors or generated by experts. In some applications, there is no appropriate document set that can be used to train the algorithm. In our study, we are looking for a method which can identify real keyphrases now and also be able to gradually and automatically adapt to the new development of the domain of documents it tries to derive keyphrases from.

THE KIP ALGORITHM

KIP is a domain-specific keyphrase extraction program. The following aspects were considered when designing KIP: first, it is a keyphrase extraction program, rather than a keyphrase assignment program; second, the program has to be able to learn to adapt to the new development for a chosen domain; and third, it can be personalized to effectively extract keyphrases from documents of a user's specific interest. The algorithm is described in this section, and its learning function and personalization feature are presented in next section.

In this paper, we distinguish these two concepts: keyword and keyphrase. A keyword is a single-term word; a keyphrase is a single-term or multi-term phrase. A keyphrase generated by KIP can be a single-term keyphrase or a multiple-term keyphrase. KIP is based on the logic that a noun phrase containing domain-specific keywords and/or keyphrases is likely to be a keyphrase. The more domain-specific keywords/keyphrases a noun phrase contains and the more significant these

keywords/keyphrases are, the more likely that this noun phrase is a keyphrase. KIP operations can be summarized as follows. KIP first extracts a list of keyphrase candidates from documents, which are noun phrases. Then it examines the composition of each candidate 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/sub-phrases it contains), and how specific these words/sub-phrases are in the domain of the document. To calculate scores of noun phrases, a glossary database, which contains domain-specific keywords and keyphrases, is used. Finally, the noun phrases with higher scores are chosen as keyphrases of the document. In this paper, two kinds of keyphrases are mentioned. One is the pre-defined keyphrase, which is stored in the glossary database; another one is the keyphrase automatically generated for a document. The first kind of keyphrase is used to calculate the score for the second one.

KIP has the following main components: a tokenizer, a part-of-speech (POS) tagger, a noun phrase extractor, a keyphrase extraction tool, a learning function, and the personalization feature. The first four components are introduced in this section, and the learning function and the personalization feature will be discussed in next section. KIP's main components are outlined in Figure 1.

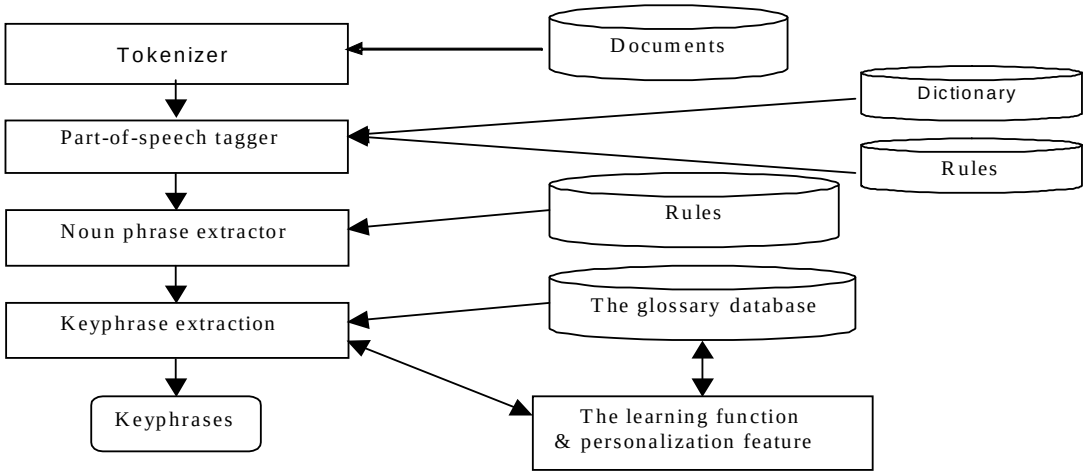

Figure 1. KIP's Main Components

Tokenizer, Part-of-speech Tagger and Noun Phrase Extractor

Tokenizer is to separate all the words, punctuation marks and other symbols from document text to obtain the atom units. Each word is assigned an initial POS tag. We use the WordNet lexical database v2.0 (Fellbaum, 1998) to assign the right POS tag to each word. This database contains words which are divided into four categories (noun, verb, adjective, and adverb) and the number of senses for each word used in categories it belongs to. A word is marked as a multi-tag word if it appears in more than one category. A word's initial POS tag is determined by the category having the maximum number of senses of this word. For every multi-tag word, the sequence of the POS tags of the proceeding n tokens (n is between 2 to 4) is examined against a list of predefined syntactic rules to determine its right POS tag. KIP's noun phrases extractor (NPE) 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 repetition, and [ ] means optional. A set of optional rules is also used. Phrases satisfying the above sequence patterns or the optional rules will be extracted as noun phrases. Users may choose to obtain noun phrases of different length by changing system parameters. At this stage, KIP produces a list of noun phrases, which will be used in next stage, keyphrase extraction.

Extracting Keyphrases

In this stage, every candidate keyphrase will be assigned a score and ranked. The phrases with higher scores will be extracted as this document's keyphrases. In order to assign a score to a noun phrase, a glossary database containing domain-specific keyphrases and keywords is used. This database provides initial weights for the words and sub-phrases of a candidate keyphrase. In the following subsections, we will describe how to build this database, how to calculate a noun phrase's score, and how the keyphrases are extracted.

Building the Glossary Database

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 glossary database. To build this database, we need to find a human-developed glossary or thesaurus for the domain of interest. It could be as simple as users manually inputting some of the known keyphrases, or it could be as elaborated as those from published sources. The glossary database has two lists (tables): (a) a keyphrase list and (b) a keyword list. We use the Information Systems (IS) domain as an example to illustrate how a glossary database is built. For IS domain, both lists were generated from two main sources: (1) author keyphrases from an IS abstract corpus, and (2) "Blackwell Encyclopedic Dictionary of Management Information Systems" by Davis (1997).

Keyphrase List. The keyphrase list was generated as follows. First, 3,000 abstracts from IS related journals were automatically processed, and all keyphrases provided by original authors were extracted to form an initial list. Second, this list was further augmented with keyphrases extracted from the Blackwell encyclopedic dictionary. *Keyword List.* The keyword list was automatically generated from the keyphrase list. To obtain the keywords, all keyphrases were split into individual words and added as keywords to the keyword list.

The keyphrase table has three columns (keyphrases, weights, and sources) and the keyword table has two columns (keywords and weights). The first column of the tables represents keyphrases/keywords. The second column contains the weights of keyphrases/keywords. The third column of the keyphrase table represents the sources of keyphrases. Keyphrases in keyphrase table may come from up to three sources. Initially, they are all manually identified by the way described above. During KIP's learning process, the system may automatically learn new phrases and add them to the keyphrase table. During the learning process, Users may also choose phrases from the processed documents and add them to the glossary database. The keyphrases/keywords weights in the database are automatically assigned by following steps:

(1). Assigning weights to keywords. A keyword can be in one of three conditions: (A) the keyword itself alone is a keyphrase and is not part of any keyphrase in the keyphrase table, (B) the keyword itself alone is not a keyphrase but is only part of one or more keyphrases in the keyphrase table, and (C) the keyword itself alone is a keyphrase and also is part of one or more keyphrases in the keyphrase table. Each keyword in the keyword table will be checked against the keyphrase table to see which condition it belongs to. The weights are automatically assigned to keywords differently in each condition. The rationale behind the method of assigning weights to a keyword is that it reflects how specific a keyword is in the domain. The more specific a keyword is, the higher weight it has. (2). Assigning weights to keyphrases. The weight of each word in the keyphrase is found from the keyword table, and then all the weights of the words in this keyphrase are added together. The sum is the weight of this keyphrase.

Calculating Noun Phrase Scores

The score of a noun phrase is defined by multiplying a factor *F*, which is the frequency of this phrase in the document, by a factor *S*, which is the sum of weights of all the individual words and all the possible combinations of adjacent words within the noun phrase (we call the combination of adjacent words a "sub-phrase" of this noun phrase):

The score of a noun phrase = $F \times S$. The sum of weights S is defined as: $S = \sum_{i=1}^{N} w_i + \sum_{j=1}^{M} p_j$, where w_i is the weight of a word within this noun phrase, and p_j is the weight of a sub-phrase within this noun phrase. 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 pre-defined keyphrase in the glossary database. If it is, this phrase is expected to be more important. To obtain the weights for all the sub-phrases of the noun phrase, KIP will lookup the keyphrase table. 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.

Generating Keyphrases

Noun phrases are ranked in descending order by their scores. The keyphrases are extracted from the ranked noun phrase list. In order to be as flexible as possible, KIP has a set of parameters to let the users decide how many keyphrases they want. The number of extracted keyphrases for a document can be defined in three ways: (1) asking a specific number of keyphrases, (2) specifying the percentage of noun phrases to be extracted, and (3) setting a score threshold for keyphrases to be extracted. KIP contains all the above options, as well as possible combinations of these options.

THE LEARNING FUNCTION AND PERSONALIZATION FEATURE

KIP's Learning Function

Many keyphrase research efforts and keyphrase applications rely on keyphrases identified by human beings as positive examples. However, sometimes, such examples are not up to date or even not available. Therefore, an adaptation and learning function is necessary for KIP. So the glossary database grows as the field of documents advances. This function is optional and it can be enabled or disabled. With it enabled, whenever the system identifies a new keyphrase ("new" means this keyphrase is not in the database's keyphrase table, and it satisfies the inclusion requirements), this keyphrase will be automatically added to the keyphrase table and the contained words will be added to the keyword table. The inclusion requirement can be modified by the user in a similar way to defining how many keyphrases will be extracted from a document, as described at the end of last section. With the learning function enabled, the database will grow gradually, and the system performance will be improved. It will benefit future keyphrase extraction for new documents. The learning function is especially useful when KIP is used in a domain where there are very few or none existing domain-specific keyphrases and keywords. When it is applied to such a domain, KIP can automatically learn new keyphrases, and finally build a glossary for this domain.

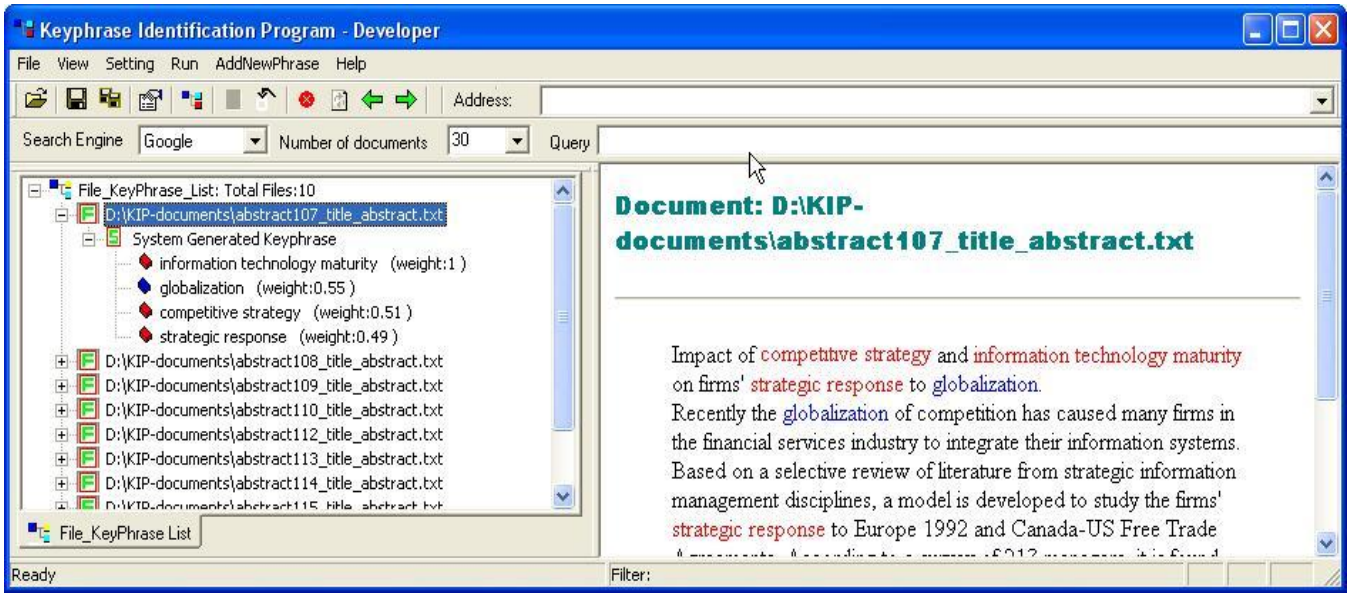

Figure 1. An Example for the Learning Function

Figure 1 is used to explain how this function works. For the document displayed, KIP extracts four keyphrases for it, and they are shown in the left frame. Three of them are marked with a red cube in front of each of them, and one is marked with a blue cube ("globalization"). The three keyphrases with a red cube are new to the glossary database, which means they are not contained in the keyphrase table. The one with a blue cube is already in the database. With the automatic learning function enabled, the system will add these three new keyphrases to the database automatically. To better control the quality of the new keyphrases added to the database, KIP has some parameters allowing the user to adjust the inclusion requirements for adding new keyphrases to the database. The system also has an option to let the user exclude some new keyphrases from being added to the database if the user thinks they are not qualified.

KIP's Personalization Feature

The learning process can be automatic without user involvement, and it can also be user-involved. If the automatic option is disabled, the user can decide if she/he wants a new identified keyphrase to be added to the database. In this way, the user can control the quality of new keyphrases added to the database. Only the new identified keyphrases satisfying the user can be added to the database. Another useful feature is that if the user thinks a phrase is good keyphrase and needs to be added to the database, but it is not identified as a keyphrase by the system, the user can highlight this phrase in the document text. Then the system will add this phrase to the database automatically. The personalization feature is based on KIP's learning function. Figure 2 is used to explain how it works. In this figure, the system extracts four keyphrases for the document displayed.

Initially, three of them are new to the glossary database and marked with a red cube ("business negotiation," "artificial adaptive agent," and "machine-learning approach"), and one is marked with a blue cube ("electronic commerce"), which means it already exists in the database. The three new keyphrases are supposed to be added to the database. However, in this example, suppose the user is not satisfied with the phrase "artificial adaptive agent" and does not want it to be added to the database. The user can exclude this phrase by right clicking the phrase and choosing the corresponding option from the popup menu. After that, the icon in front of this phrase changes to an "X" from a cube. If the user does not think a phrase is an appropriate keyphrase, the user can delete this phrase from the phrase list, and this phrase will be removed by the system from the keyphrase list. In this example, the user thinks that the phrase "automated negotiation" is a good keyphrase for this document and it should be added to the glossary database, though it is not extracted by the system. The user can highlight this phrase in the right frame, and this phrase will be added to this document's keyphrase list. KIP will also add this phrase to the glossary database automatically.

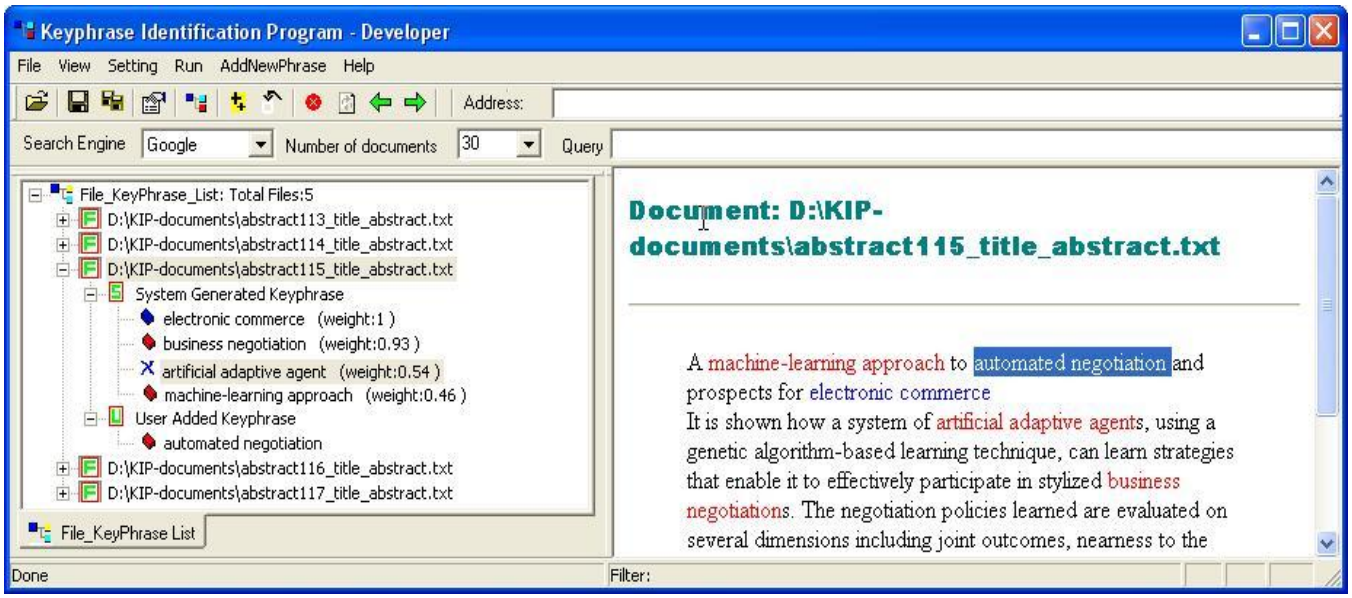

Figure 2. An Example for the Personalization Feature

From KIP's personalization feature, the user can control the quality of the glossary database and the direction of its growth flexibly and easily. After a period of time, the glossary database will be personalized to this user's interest area, and KIP will be more effective in identifying keyphrases from documents in the user's interest area. With the same starting glossary database, different users with different research interests within a domain will eventually have different evolved glossary databases, and, as a result, KIP will gradually be more and more effective and personalized for the user. For example, if one user's research area is data and text mining, and another user's area is human computer interaction, after a period of time, even with the same starting glossary database, these two users could gradually build up two different glossary databases independently. The user may be a single person, a research group, or an organization specializing in a certain area.

EXPERIMENT

Generally, there are two ways to evaluate the effectiveness of a keyphrase extraction system. The first one uses the standard information retrieval measure: precision and recall. The precision and recall are calculated using author-assigned keyphrases for documents against the system-generated keyphrases. One problem with this method is that some author-assigned keyphrases may not occur in the documents they are assigned to. In experiments reported by Turney (2000), about only 75% of author-assigned keyphrases appear somewhere in the document. The second method uses human judges to assess the quality of system-generated keyphrases. Several previous studies have used human assessment to evaluate system-generated keyphrases (Turney, 2000; Barker and Cornacchia, 2000; Jones and Paynter, 2002). Human evaluation reflects how human readers feel about the keyphrases when dealing with them in the real world. In this experiment, we used human judges to assess the quality of the keyphrases generated by KIP. We used the IS domain to perform this evaluation. The process of building an IS glossary database has been formerly described.

We used twenty short papers as the test documents. They were from AMCIS'01, 02 and 03. Their average length was three pages. Ten IS researchers were recruited as the domain experts. Each expert was given all of the 20 documents and the extracted keyphrase list for each document. They were asked to read a document fully first and then go over the extracted keyphrase list for that document. For each keyphrase, the subject rated the quality of the keyphrase, in terms of "how well it represented major issues in that document," using a five-point scale ranging from 1 to 5. 1 means worst, 5 means best and 3 means neutral. For each document, KIP extracted 15 keyphrases. The judges did not know how these phrases were generated and what system they were evaluating before finishing the experiment.

The inter-judge agreement is important for experiments involving human judgment. Kendall Coefficient of Concordance W (Siegel and Castellan, 1988) is good at measuring the agreement between subject's relative rankings of keyphrases. The average W value for all the 20 documents is 0.57, which means a good agreement among subjects. Table 1shows the average scores assigned by the judges for the 20 documents when the number of extracted keyphrase is 3, 6, 9, 12 and 15. Table 1 also shows that the mean scores are all statistically ($p < 0.01$) greater than the midpoint 3. So, on average, the keyphrases were rated positively by the subjects.

	Number of extracted keyphrases				
	3	**6**	**9**	**12**	**15**
Mean Score	3.75	3.61	3.50	3.35	3.26
Standard Deviation	0.32	0.27	0.23	0.21	0.20

Table 1. Human Evaluation Result of KIP

From Table 1, we can also see that when the number of extracted keyphrases decreases, the mean score increases. This is what we expected, because KIP outputs the keyphrases in descending order of their importance to the document. The results in Table 1 also show that the system is effective in ranking the phrases. We can see this trend from Figure 4. We used a paired t-test to test the significance of this trend. The result shows that there is a significant difference between any two evaluation points (e.g. when the number of extracted keyphrases is 3 and when it is 6) at the $p < 0.01$ level. This is especially useful and important when only a limited number of keyphrases are required, because we will be confident that the extracted set of keyphrases are the best ones among all the candidate keyphrases.

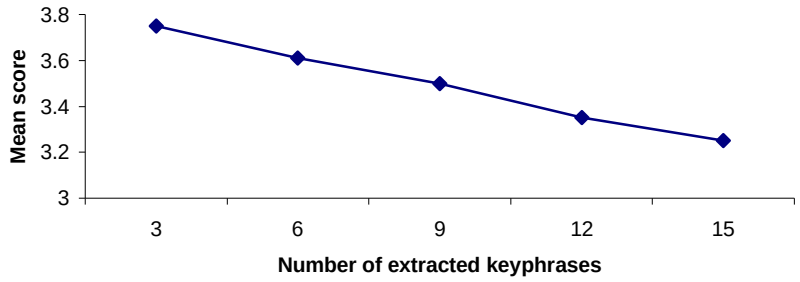

Figure 4. The Relationship Between Received Scores and Number of Extracted Keyphrases

CONCLUSION

Document keyphrases are the most important topical phrases for a given document. They can provide semantic metadata which can characterize documents and produce an overview of the content of a document. In this paper, a new keyphrase extraction algorithm is introduced and the human evaluation result about its effectiveness is reported. KIP's learning function and personalization feature make it easier to apply KIP to different domains. The features and performance of KIP will make it useful for a variety of applications, such as document clustering, document classification, retrieval engines, and browsing interface.

REFERENCES

1. Barker, K. and Cornacchia, N. (2000) Using noun phrase heads to extract document keyphrases, *Proceedings of the thirteenth Canadian conference on artificial intelligence*, Montreal, Canada, 40-52.
2. Davis, G. B. (1997). Blackwell Encyclopedic Dictionary of Management Information System. Malden, MA, Blackwell Publishing.
3. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MIT Press.
4. Frank, E., Paynter, G., Witten, I. H., Gutwin, C. and Nevill-Manning, C. (1999). Domain-specific keyphrase extraction. *Proceeding of the sixteenth international joint conference on artificial intelligence*, San Mateo, CA.
5. Gutwin, C., Paynter, G., Witten, I. H., Nevill-Manning, C. and Frank, E. (1999). Improving Browsing in Digital Librarieswith Keyphrase Indexes. Journal of. *Decision Support Systems* 27(1-2): 81-104.
6. Jones, S. and Mahoui, M. (2000). Hierarchical document clustering using automatically extracted keyphrase. *Proceeding of the third international Asian conference on digital libraries* Seoul, Korea. 113-120.
7. Jones, S. and Paynter, G. W. (1999) Topic-based browsing within a digital library using keyphrases, *Proceedings of digital libries'99: The fourth ACM conference on digital libraries*, ACM Press, Berkeley, CA, 114-121
8. Jones, S. and 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.
9. Jones, S. and Staveley, M. (1999). Phrasier: A system for interactive document retrieval using keyphrases. *Proceedings of SIGIR'99: The 22nd international conference on research and development in information retrieval*, Berkeley, CA.
10. Kosovac, B., Vanier, D. J. and 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.
11. Krulwich, B. and Burkey, C. (1996). Learning user information interests through the extraction of semantically significant phrases. *AAAI 1996 Spring Symposium on Machine Learning in Information Access*, California, AAAI Press.
12. Li, Q., Wu, Y .B., Bot, R. S., and Chen, X. (2004). Incorporating Document Keyphrases in Search Results. *Proceedings of the Tenth Americas Conference on Information Systems*, New York, New York.
13. Siegel, S. and Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd Ed.). New York, McGraw Hill College Div.
14. Turney, P. D. (2000). Learning algorithm for keyphrase extraction. *Information Retrieval* 2(4): 303-336.
15. Witten, I.H. (1999) Browsing around a digital library, *Proceeding of Australasian Computer Science Conference*, Auckland, NewZealand, 1-14
16. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C. and Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction. *Proceedings of the Fourth ACM Conference on Digital Libraries.*
17. Zha, H. (2002). Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland.