Keyword Extraction Based on Sequential Pattern Mining

Jiajia Feng[1]
[1] College of Computer Science and Info. Eng. Hefei University of Tech. Hefei, China
fengjia602@163.com

Fei Xie [2, 1]
Xuegang Hu[1]
[2] College of Computer Science and Info. Eng. Hefei University of Tech. Hefei, China
xiefei9815057@sina.com

Peipei Li[1], Jie Cao[3]
[3] Jiangsu Provincial Key Laboratory of E-business Nanjing University of Finance and Economics Nanjing, China
peipeili_hfut@163.com

Xindong Wu[1, 4]
[4] Department of Computer Science University of Vermont Burlington, USA
xwu@cs.uvm.edu

ABSTRACT
Keyword extraction is to automatically extract keywords that capture the main topic discussed in a given document. In this paper, a new keyword extraction algorithm based on sequential patterns is proposed. By preprocessing, a document is represented as sequences of words where a sequential pattern mining algorithm is applied on, and important sequential patterns are mined that reflect the semantic relatedness between words. Both statistical features and pattern features within words are used to build the keyword extraction model. The algorithm is independent of languages and does not need the help of a semantic dictionary to get the semantic features. Experimental results on Chinese journal articles show that the proposed algorithm always outperforms the baseline method KEA.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms
Algorithms, Performance, Measurement, Experimentation

Keywords
Keyword Extraction; Sequential Pattern Mining; Pattern Features

1. INTRODUCTION
With the rapid development of the Internet, more and more document messages, such as web news, e-mails, e-books and other documents are being emerged. When dealing with these documents, people hope to quickly get the key elements of each document, where keywords serve as a condensed summary, and the readers can quickly browse and read the main points of the given documents. Furthermore, keywords have been successfully used in the following tasks: automatic indexing, document summarization, text categorization, and text clustering.

ICIMCS'11 August 5-7, 2011, Chengdu, Sichuan, China

Existing keyword extraction methods are based on analyzing words' various features in a document, which use supervised learning or unsupervised learning to extract keywords. Main features within words in a document can be divided into two categories: statistical features and semantic features.

Methods based on statistical features are simple, such as word frequency statistics [1], TFIDF [2], word co-occurrence [3], and the small world network model [4]. However, the quality of extracted keywords is not high.

In recent years, researchers have focused on keyword extraction based on semantic features, which have improved the quality of extracted keywords, such as methods based on lexical chains [5, 6], KEA++ based on the professional thesaurus [7], and keyword extraction based on web mining and statistical methods [8]. However, they are dependent on languages, and usually require the help of a semantic dictionary to obtain the semantic features within words.

In this paper, a keyword extraction algorithm is proposed to overcome the disadvantages of the above methods, which applies sequential pattern mining to mine word patterns with wildcards from document sequences to obtain semantic features within words, which is independent of languages and does not need the help of a semantic dictionary. Experiments demonstrate that pattern features obtained by sequential pattern mining enable improving the quality of extracted keywords.

2. RELATED WORK
2.1 Keyword Extraction
Keywords are defined as a list of words that capture the main topic discussed in a given document and serve as a condensed summary. Due to the importance of keywords, many research efforts have been made on keyword extraction. Existing keyword extraction methods can be divided into unsupervised learning methods and supervised learning ones.

Unsupervised learning methods are to extract keywords from a single document using heuristic rules. Barker and Cornacchia [9] extracted keywords by analyzing the noun phrases' length, word frequency, and the frequency of the first word of a phrase. Steier and Belew [10] extracted keywords using the mutual information between two words. Mihalcea and Tarau [11] proposed the graph-ranking model and applied it to keyword extraction. Wang et al. [12] used the Page-Rank algorithm to score phrases for keyword extraction. Matsuo and Ishizuka [13] used the word co-occurrence statistics to extract keywords. Li et al. [5] used lexical chains built by HowNet to extract keywords. Those methods have a common advantage that the keywords are

extracted from a single document without training data. However, these methods are heuristic, and the quality of extracted keywords is relatively low.

Supervised learning methods consider the keyword extraction as a classification problem that can be solved by machine learning methods. Supervised learning methods extract keywords by a classification, and need to establish a training set. Therefore, these methods are more complex, and the quality of extracted keywords is improved.

GenEx [14] and KEA [15] are two typical supervised keyword extraction systems. GenEx used the genetic algorithm and the C4.5 decision tree to extract keywords. KEA used a Naïve Bayes classifier to extract keywords. The features used in GexEx and KEA are mainly statistical features, such as frequency, location and TFIDF value.

Studies show that the performance can be improved by adding semantic features. Turney [8] presented a method based on web mining to obtain the correlation between two candidate keywords that greatly improved the quality of extracted keywords. Medelyan and Witten [7] proposed KEA++ based on KEA by adding semantic features acquired from a professional thesaurus. Ercan and Cicekli [6] obtained semantic features from lexical chains built by WordNet. These methods are restricted by languages and the domain that needs the help of a semantic dictionary to acquire semantic features, such as WordNet and HowNet.

The algorithm proposed in this paper is a supervised learning method based on semantic features without the help of other semantic analysis tools, which is independent of languages. By sequential pattern mining from a document sequence, word patterns with wildcards are acquired by the analysis of which we can obtain the semantic features within words.

2.2 Sequential Pattern Mining

Sequential pattern mining is an important research task in many domains that was first introduced by Agrawal and Srikant [16]. Since then, many improved algorithms have been proposed. The search strategies can be divided into the breadth-first search and the depth-first search. The breadth-first searching methods generate a large number of candidate patterns, such as GSP [17], and SPADE [18]. The depth-first searching methods are based on the divide and conquer method that iteratively partitions the original data set to reduce the size of the data, such as PrefixSpan [19] and SPAM [20].

In recent years, the problem definition of sequential pattern mining has been extended. Ji et al. [21] proposed sequential pattern mining with gap constraints. Yan et al. [22] studied the closed pattern mining problem and proposed the CloSpan algorithm. A sequential pattern is closed if there does not exist a super pattern that has the same support with it. Zhang et al. [23] studied the sequential pattern mining from a single long sequence with gap constraints.

Sequential pattern mining has been widely applied in many domains, such as text mining and bioinformatics. In text mining, a document is represented as a sequence of words. Sequential patterns are mined in a single document and a document collection. Sequential pattern mining has been successfully applied to question answering [24] and authorship attribution extraction [25]. In this paper, we apply sequential pattern mining to keyword extraction that enables improving the quality of extracted keywords.

3. SEQUENTIAL PATTERN MINING

First, we briefly explain what sequential pattern mining is. Then, the sequential pattern mining algorithm SPAM [20] adopted in this paper is described.

3.1 Problem Definition

Let $I = \{i_1, i_2, ...i_n\}$ be a set of all items. An itemset is a subset of all items. A sequence $S =< s_1, s_2, ..., s_l >$ is an ordered list of itemsets. Here $s_j = (x_1, x_2, ...x_m)$, $1 \le j \le n$, is an itemset that is also called an element of the sequence. For brevity, the brackets are omitted if an element has only one item. That is, element (*x*) is written as x. An item can occur at most once in an element of a sequence. But it can occur multiple times in different elements of a sequence. The number of items in a sequence is called the length of the sequence. A sequence with length *l* is called an *l*-sequence. A sequence $a =< a_1, a_2, ..., a_n >$ is called a subsequence of another sequence $b =< b_1, b_2, ..., b_m >$, if there exist integers $1 \le j_1 < j_2 < ... < j_n \le m$, such that $a_i \subseteq b_{j_i}$ for each $1 \le i \le n$.

A sequence database D is a set of tuples <sid, s>, where sid is a sequence id and s is a sequence. A tuple <sid, s> is said to contain sequence b, if b is a subsequence of s. The support of sequence b is the number of tuples in the sequence database containing b. Given a positive integer *min_sup* as the support threshold, a sequence is called a frequent sequential pattern if the sequence is contained in at least *min_sup* tuples in the sequence database. The task of sequential pattern mining is to find all frequent sequential patterns in a database.

3.2 The SPAM Algorithm

In this paper, we adopt the SPAM algorithm [20] for sequential pattern mining. SPAM uses a vertical bitmap representation for efficient support counting. A brief introduction of the SPAM algorithm is as follows.

The algorithm utilizes a depth-first traversal of the search space combined with a vertical bitmap representation. The candidate patterns are generated in two steps: S-step and I-step as shown in Figure 1. The S-step refers to an element extension, while the I-step means an item extension. SPAM uses a vertical bitmap representation of the database for efficient support counting that greatly reduces the running time in the process of mining.

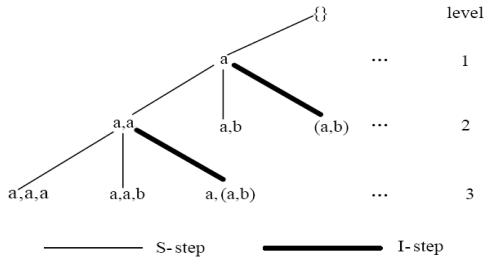

Figure 1. The Lexicographic Sequence Tree

There are a large number of candidate patterns produced by SPAM, and a pruning strategy is hence very important to reduce

the number of candidate patterns. First of all, SPAM satisfies the Apriori property that any super-sequence of a non-frequent sequence cannot be frequent. We can prune the candidate patterns whose sub-sequential patterns are not frequent. The S-step pruning strategy and I-step pruning strategy are also adopted in SPAM.

4. KEYWORD EXTRACTION

4.1 Preprocessing

In this paper, both pattern features obtained by sequential pattern mining from a document and statistical features are used to extract keywords. The preprocessing step includes the word segmentation, statistical feature extraction, and generation of a document sequence database.

The aim of word segmentation is to transform a document into several sequences of words. A document is split into sentences separated by stop marks. The document title and section titles are also included as sentences. We use ICTCLAS [26] to split a Chinese document into words. All English words are stemmed using the iterated Lovins [27] approach. Figure 2 shows the mapping between a document and a sequence database.

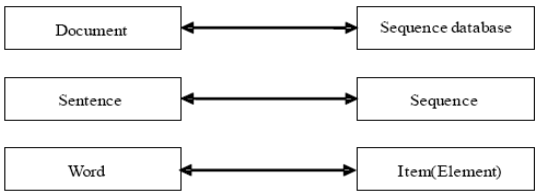

Figure 2. The relationship between a document and a sequence database

4.2 Feature Extraction

In this paper, we divide the features within words into two categories: basic statistical features and pattern features shown in Table 1. Basic statistical features are obtained in the preprocessing stage, while pattern features are acquired by sequential pattern mining from the sentence sequence database.

Table 1. Word's features

Category	Feature	Description
Basic statistical feature	Word_POS	part of speech of a word
	First_loc	Location of the first occurrence
	Word_freq	Frequency of a word
	Sentence_num	Sentence number of the first occurrence
Pattern feature	Seq_length	Length of the longest pattern which the word is in
	Seq_sup	Support of the longest pattern which the word is in
	Seq_sup_len	seq_sup* seq_length

Feature selection is a key step in the supervised keyword extraction that directly affects the quality of keywords. The basic features of words are extracted in preprocessing stage by simple statistical methods while pattern features are obtained by sequential pattern mining.

4.3 Algorithm Description

The following pseudo-code is a brief description of our keyword extraction method based on sequential pattern mining proposed in this paper.

Input: training documents with keywords labeled by experts; minimal support threshold *min_sup*; a test document

Output: extracted keywords of the test document

1: For each training document *d* do
2: Document *d* is split into sentences according to the punctuation symbols.
3: Each sentence is segmented into words, and the basic features within words are counted.
4: *d* is transformed into a sequence database consisting of sentence sequences.
5: Mine sequential patterns from the sequence database using the SPAM algorithm where *min_sup* is set to 3.
6: Calculate the pattern features within words.
7: Each word is represented as a vector with basic statistical features and pattern features. The class label is also assigned according to whether the word is a keyword.
8: Train the classifier using a machine learning algorithm.
9: The test document is preprocessed, and the words are represented as vectors with basic statistical features and pattern features.
10: Each word in the test document is decided whether it is a keyword by the trained classifier.

In this paper, the J48 decision tree [28] classification algorithm is used to train the classifier for keyword extraction. In our experiments different classification algorithms are used to train the classifier. The experimental results show that the J48 algorithm has the best performance compared to other algorithms.

5. EXPERIMENTAL RESULTS

In this paper, we select 200 Chinese journal papers downloaded from the Chinese natural language processing open platform [29] to test the performance of our proposed algorithm. The keywords in the documents are assigned by the authors.

To evaluate the performance of the keyword extraction method, the precision P, recall R and F1-measure are adopted as the evaluation metrics. They are defined as: P=n/N, R=n/M, F1=2*P*R/(P+R), where n is the number of correct extracted keywords, N is the number of extracted keywords, and M is the number of labeled keywords.

Experiment 1 In this experiment, we choose different feature sets to test the impact of the features in the performance of the keyword extraction method. Table 2 shows the performances of keyword extraction with different feature sets on the Chinese journal articles.

Group I: Word_POS + First_loc + Sentence_num;

Group II: Word_POS + First_loc + Sentence_num + Seq_sup_len;
Group III: Word_POS + First_loc + Word_freq + Sentence_num;
Group IV: Word_POS + First_loc + Word_freq + Sentence_num + Seq_sup_len;

Table 2. The impact of different feature sets in the performances of keyword extraction

Feature sets	Precision	Recall	F
Group I	0.568	0.127	0.208
Group II	0.556	0.376	0.449
Group III	0.344	0.421	0.378
Group IV	0.557	0.298	0.393

There are several observations seen from Table 2. First, Group II has the best performance. It indicates that the quality of extracted keywords is improved if we use the pattern features. Second, Group IV with all features presents the worse performance than Group II. This shows that there is conflict between Seq_sup_len and Word_freq, and Seq_sup_len is more useful than the Word_freq feature.

Experiment 2 In this experiment, we study the influence of the gap on the performance of three groups. Figure 3 shows the performances of keyword extraction when the gap width changes from 3 to 10.

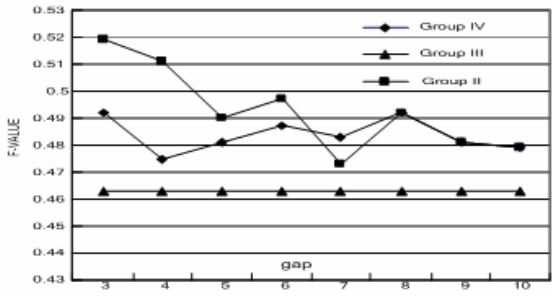

Figure 3. The relationship between gap and F-value

It can be seen from Figure 3 that Group II and Group IV perform better than Group III. It indicates that the pattern features are useful for improving the performance. Figure 3 also shows that, if the value of gap is bigger than 8, Group II and Group IV have the same F-values.

Experiment 3 In this experiment, we compare our method against the KEA method. Figure 4 reports the experimental results. It can be seen that our method improves the quality of keywords extracted.

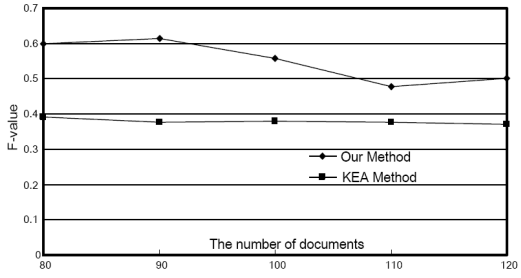

Figure 4. The comparison results between our method and KEA

6. Conclusion

In this paper, a sequential pattern mining algorithm is designed for keyword extraction. By sequential pattern mining from a document sequence, semantic features within words are obtained, and are used to train the keyword extraction model. This method is not limited by the language and the thesaurus. Experimental results demonstrate that pattern features acquired by sequential pattern mining enable improving the quality of extracted keywords. Word patterns obtained by sequential pattern mining contain fruitful semantic information among words. In the future, we will apply sequential patterns to other text mining tasks and video annotation [30].

7. ACKNOWLEDGMENTS

This research has been supported by the National Natural Science Foundation of China (NSFC) under awards 60828005, 61005044, 71072172, and 60975034, the National 973 Program of China under award 2009CB326203, the Fundamental Research Funds for the Central Universities under award 2011HGZY0003, the Jiangsu Provincial Key Laboratory of E-business, Nanjing University of Finance and Economics under award JEB1103, and the Research Foundation of Education Bureau of Anhui Province under grants KJ2010B168 and 2010SQRL148.

8. REFERENCES

[1] Luhn, H. P. 1957. A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development*.1 (4): 309-317.

[2] Li, J., Fan, Q. and Zhang, K. 2007. Keyword extraction based on TF/IDF for Chinese news document. *Wuhan University Journal of Natural Sciences*.12 (5): 917-921.

[3] Ma, Y., Wang, Y., Su, G. and Zhang, Y. 2003. A novel Chinese text subject extraction method based on character co-occurrence. *Journal of Computer Research and Development*. 40(6): 874-878.

[4] Zhao, P., Cai Q., Wang, Q. and Geng, H. 2007. An automatic keyword extraction of Chinese document algorithm based on complex network features. *Pattern Recognition and Artificial Intelligence*. 20(6): 827-831.

[5] Li, X., Wu, X., Hu, X., Xie, F. and Jiang, Z. 2008. Keyword extraction based on lexical chains and word co-occurrence for Chinese news web pages. In *Proceedings of ICDM Workshops 2008* (December 15-19, 2008). Pisa, Italy, 744-751.

[6] Ercan, G. and Cicekli, I. 2007. Using lexical chains for keyword extraction. *Information Processing and Management: An International Journal.* 43(6): 1705-1714.

[7] Medelyan, O. and Witten, I.H. 2006. Thesaurus based automatic keyphrase indexing. In *Proceedings of the Joint Conference on Digital libraries*. 296-297.

[8] Turny, P.D. 2003. Coherent keyphrase extraction via web mining. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*. Acapulco, Mexico. 434-439.

[9] Barker, K. N. and Cornacchia, N. 2000. . Using noun phrase heads to extract document keyphrases. In *Canadian Conference on Artificial Intelligence*. 40-52.

[10] Steier, A.M., Belew, R.K. 1993. Exporting phrases: a statistical analysis of topical language, In *Proceedings of Second Symposium on Document Analysis and Information Retrieval*. 179-190.

[11] Mihalcea, R. and Tarau, P. 2004. TextRank: Bringing order into texts, In *Proceedings of EMNLP*. Barcelona, Spain. 404–411.

[12] Wang, J., Liu, J. and Wang, C. 2007. Keyword extraction based on PageRank. In *Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Nanjing, China. 857-864.

[13] Matsuo, Y. and Ishizuka, M. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*. 13(1): 157-169.

[14] Turney, P.D. 1999. Learning to extract keyphrases from text. *NRC Technical Report ERB-1057*. National Research Council, Institute for Information Technology, 1-43, Canada.

[15] Frank, E., Paynter, G.W. and Witten, I.H. 1999. Domain-Specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, Morgan Kaufmann. 668-673.

[16] Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In *Proc. IEEE ICDE'95*(Mar. 1995). Taipei, Taiwan. 3-14.

[17] Srikant, R. and Agrawal, R. 1996. Mining sequential pat terns: Generalizations and performance improvements. In *Proc. of the 5th International Conference on Extending Database Technology*. Avignon.

[18] Zaki, M.J. 2001. SPADE: An efficient algorithm for mining frequent sequences. In *International Conference on Machine Learning*. vol. 42, 31-60.

[19] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M. 2001. PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. In *ICDE 2001*. 215-226, Heidelberg, Germany.

[20] Ayres, J., Flannick, J., Gehrke, J. and Yiu, T. 2002. Sequential PAttern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (July 23-26, 2002). Edmonton, Alberta, Canada.

[21] Ji X., Bailey, J. and Dong, G. 2005. Mining minimal distinguishing subsequence patterns with gap constraints. In *Proc. IEEE ICDM* (Nov, 2005). Houston, Texas, USA. 194-201.

[22] Yan, X., Han, J., and Afshar, R. 2003. CloSpan: mining closed sequential patterns in large datasets. In *Proceedings of SIAM International Conference on Data Mining*, San Francisco, CA, USA.166-177.

[23] Zhang, M., Kao, B., Cheung, D. and Yip, K. 2005. Mining periodic patterns with gap requirement from sequences. In *Proc. ACM SIGMOD'05*. Baltimore Maryland.

[24] Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L. and García-Hernández, R. 2006. A text mining approach for definition question answering. *Natural Language Processing Lecture Notes in Computer Science*. Volume 4139/2006, 76-86.

[25] Coyotl-Morales, R., Villaseñor-Pineda L., Montes-y-Gómez, M. and Rosso, P. 2006. Authorship attribution using word sequences. In *Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes in Computer Science.* Volume 4225/2006, 844-853.

[26] Zhang, H., Yu, H., Xiong, D. and Liu, Q. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *proceedings of 2nd SigHan Worksho*p. 184-187.

[27] Lovins, J.B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics 11*. 22–31.

[28] Haykin, S. 1999. Neural networks: a comprehensive foundation [J]. *Computaciony Sistemas*. 4(2): 188-190.

[29] http://www.nlp.org.cn/categories/default.php?cat_id=16

[30] Wang, M., Hua, X., Tang, J., and Hong, R. 2009. Beyond distance measurement: constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, vol. 11, no. 3, 2009.