2009 International Conference on Electronic Commerce and Business Intelligence

Using Citation-KNN for Automatic Keyword Assignment

Chengzhi ZHANG
Department of Information Management, Nanjing University of
Science & Technology. Nanjing, China
Institute of Scientific & Technical Information of China. Beijing, China
e-mail: zhangchz@istic.ac.cn

Hongjiao XU
Institute of Scientific & Technical
Information of China. Beijing, China
e-mail: xuhongjiao_1111@163.com

Abstract—**Currently, the automatic keywords extraction method can only extract keywords appeared in the articles and it cannot extract the implicit keyword which does not appear in the articles. It is a difficult work to extract implicit keywords in an article in the task of automatic keywords extraction. This work can also be called automatic keyword assignment. In this paper, an automatic keyword assignment method based on Citation-KNN (Citation-K Nearest Neighborhood) is proposed. Experimental results show that the proposed method can not only improve the precision and recall of keyword extraction, but also extract implicit keyword which does not appear in the articles efficiently.**

Keywords- ***Keyword Extraction; Keyword Assignment; mplicit Keyword Extraction; Citation-KNN***

I. INTRODUCTION

Keywords can reflect the subject or content of the articles effectively. They are words, phrases or terms extracted from article which can reveal content of articles in order to satisfy the needs of text mining and information retrieval.

Luhn did automatic indexing experiments in late 1950s [1, 2]. In 1963, Chemical Abstracts (CA) started to compile keywords indexing using computers and offered rapid ways to search topics of documents. The pure statistical method is applied in automatic keywords extraction early [3, 4, 5]. In early 1970s, Earl started to use linguistic methods such as syntactic analysis [6] to extract keywords. At the middle of 1970s, vector space model was applied in automatic keywords extraction by Salton [7]. In late 1990s, genetic algorithm [8, 9] and Bayes [10] methods were employed by Turney and Frank respectively. In 2001, Anjewierden & Kabel put forward automatic indexing based on ontology [11]. In 2003, Tomokiyo & Hurst extracted keywords based on language model [12]. Hulth extracted keywords through Bagging algorithm [13]. In 2004, Li proposed keywords extraction method based on maximum entropy [14]. In 2007, Ercan & Cicekli put forward automatic indexing method based on lexical chain [15].

According to the results of Turney, about 65% to 90% of manual annotated keywords can be found in full text of an article [16]. In this paper, we call keywords that do not appear in the article 'implicit keyword'. The extraction of implicit keywords is very difficult and all available automatic keywords extraction algorithms can not extract them efficiently. Usually, the automatic implicit keywords extraction methods are dependent on some exterior resources, such as thesaurus, ontology and so on, to transform process of automatic implicit keywords to process of classification of keywords or transform keywords of text to implicit keywords. In this paper, we'll use an automatic implicit keywords extraction (also called automatic keyword assignment) method based on Citation-KNN method to extract the implicit keywords of the article. The experiment shows that this method is effective.

The rest of this paper is organized as follows. In section 2, a detailed description of the proposed approach is presented. Subsequently in section 3, the authors report experiment results that evaluate the proposed approach. The paper is concluded with summary and future work directions.

II. AUTOMATIC KEYWORD ASSIGNMENT ALGORITHM BASED ON CITATION-KNN

A. Description of Citation-KNN Algorithm

1) KNN Algorithm: K nearest neighborhood method (KNN) is a lazy learning algorithm based on statistics. It was put forward by Cover & Hart in 1968[17] and has been applied in many fields. In the field of automatic text categorization, KNN was proved to be one of the best methods [18]. Test samples are classified according to major classes in K nearest neighborhood.

$$y'_i\text{=argmax}\sum_{j=1}^{K} I(v = y_j) \qquad (1)$$

Where, v is class label, y'_i is class label of a nearest neighborhood, I (·) is indicator function and when its parameter is true, then returns 1, else return 0.

For each nearest neighborhood may have different influence in classification, the nearest neighborhood are weighted according to similarity between test sample and each nearest neighborhood x_i [19], and the nearest neighborhood with higher the similarity degree will be given higher weight. If weigh equals to the similarity between test sample and nearest neighborhood, absolute weight and relative weight will be calculated as follows respectively:

$$w_i\text{= Sim}(x'_i, x_j) \qquad （2）$$

978-0-7695-3661-3/09 $25.00 © 2009 IEEE
DOI 10.1109/ECBI.2009.25

$$w_i = \mathrm{Sim}(x_i, x_j) \Big/ \sum_{j=1}^{K} \mathrm{Sim}(x_i, x_j) \qquad (3)$$

In view of weight of the nearest neighborhood, classification decision function is given as follows:

$$y'_i = \operatorname{argmax} \sum_{j=1}^{K} w_i \cdot I(v = y_j) \qquad (4)$$

2) Citation-KNN: Citation-KNN [20] was firstly put forward by Wang & Zucker to solve the problem of multi-instance learning. Citation-KNN is an enhancing algorithm based on traditional KNN algorithm. Its main principle is that using the thought of citation and quotation in literature metrology, when we decide the classification of test sample x'_i, we should take the class of the training sample (which is the quotation of test sample) which x'_i is one of its K nearest neighborhood the into account, as well as the class of K training sample of the nearest neighborhood (which is the citation of test sample). Figure 1 show the citation and quotation relationship in Citation-KNN algorithm. In figure 1, $R_1, R_2, \ldots, R_k$ are K articles cited by sample A, $w_{r1}, w_{r2}, \ldots, w_{rk}$ are their corresponding weights. $C_1, C_2, \ldots, C_q$ are the Q articles quotation of sample A, $w_{C1}, w_{C2}, \ldots, w_{Cq}$ are their corresponding weight.

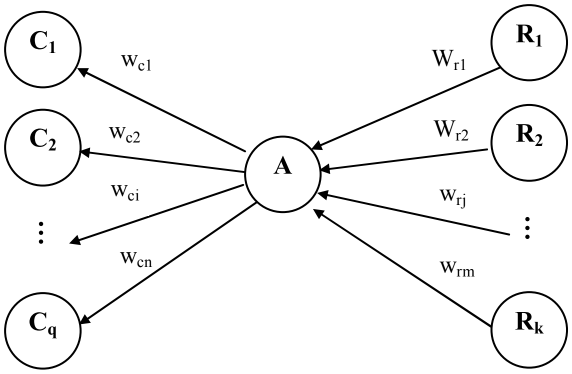

Figure 1. Schematic diagram of citation and quotation in Citation-KNN

When we make classification decision, we should consider class label of the citation and quotation. At the same time, we can give different weight based on how the citation and quotation influence classification. Then the classification decision function is:

$$y'_i = \operatorname{argmax}\left(\sum_{i=1}^{K} (\alpha \cdot w_{r_i} \cdot I(v = R_i)) + \sum_{j=1}^{Q} (\beta \cdot w_{c_j} \cdot I(v = C_j)) \right) \qquad (5)$$

Where, α and β is the weight of citation and quotation in Citation-KNN, and α+β=1. In this paper, we assume α=β=0.5. Figure 2 shows the description of Citation-KNN algorithm.

B. Neighborhood Weighted Method

In this paper, when we use Citation-KNN in automatic implicit keyword extraction, the decision function is weighted according to the features of different neighborhood. The main weighting methods include determining weight based on similarity and features of sample (for example PageRank value, quotation frequency). And the similarity-based weighting method is given in formula (3). Similarity is the cosine of documents vector angle [21]. The definition and calculation method of the PageRank value of sample, quotation frequency can be found in [22].

```
Algorithm: Description of Citation-KNN algorithm
Input: Test set $\{(x'_1,y'_1),\ldots,(x'_n,y'_n)\}$, and $x'_i \in X'$, $y'_i \in Y$,
       training set $\{(x_1,y_1),\ldots,(x_m,y_m)\}$, and $x_j \in X$, $y_j \in Y$
Output: The corresponding class label ($y'_i$) of $x'_i$ in test
        dataset
Steps:
  Set the number of nearest neighborhood, and
    assume α=β=0.5;
  For i=1 to N
    Select the nearest K training samples to $x'_i$ in
      training set and form the set $\mathbf{x_1}$, then calculate the
      weight ($w_{ri}$) of each sample, 1<i<K;
    Select training sample which $x'_I$ is one of its K
      nearest neighborhoods and form the training
      sample test $\mathbf{x_2}$, and calculate the weight ($w_{cj}$) of
      each sample, 1<j<Q;
    $y'_i = \operatorname{argmax}\left( \sum_{i=1}^{K} (\alpha \cdot w_{r_i} \cdot I(v = R_i)) + \sum_{j=1}^{Q} (\beta \cdot w_{c_j} \cdot I(v = C_j)) \right)$
  End For
```

Figure 2 Description of Citation-KNN Algorithm.

C. Automatic Implicit Keyword Extraction Algorithm based on Citation-KNN

Automatic implicit keyword abstraction is transformed to classification task in this paper. As shown in Figure 2, The nearest K training samples to document x'_i in training set are selected and the weight (w_{ri}) of each sample is computed. Then, training samples which x'_i is one of its K nearest neighborhoods are selected, and the weight (w_{cj}) of each sample is computed. We can combine the weight of each sample document. According to the classification decision function, we can get the result of implicit keyword extraction.

III. Experiment Result

A. The Evaluation Method

1) Dataset: Dataset used in this experiment is Chinese academic periodical full-text database (URL: http://www.cnki.net). We select more than 100 thousand academic papers in the field of economy. The keywords are given by the authors of these papers, and they can be used as training set of K nearest neighborhood keyword extraction. And the test set contains 600 documents whose keywords are given by their authors.

2) Evaluation method: The method put forward by Turney will by used in the application of evaluation of experimental result. Precision, Recall and F-measure will be employed to evaluate the performance of the algorithm. In Turkey's method, only when the keywords extracted by the automatic extraction system are in full accordance with that given by human, these keywords are matching [16].

$$P=a/b \tag{6}$$

$$R= a/c \tag{7}$$

$$\text{F-measure}= 2*P*R/(P+R) \tag{8}$$

Where, a is the number of keywords that extracted by the automatic extraction system in full accordance with that given by human, b is the number of keywords extracted by the automatic extraction system, c is the number given by human.

B. Experiment Result and Analysis

The authors compare automatic implicit keyword based on KNN and Citation-KNN. Set cosine value of documents vector angle to be the similarity between papers. In Table I, the first 10 similar Chinese documents of the Chinese document"*现代网络银行发展中的金融监管思考* (*Reflections on Financial Supervision in the Development of Modern Internet Banking*)" are shown. This document has the keywords such as "*网络银行*(*internet banking*), *国际经验* (*international experience*), *金融监管* (*financial supervision*)". According to KNN method, the keywords "*网络银行* (*network-bank*), *监管*(*supervision*), *发展战略* (*development strategy*), *启示*(*inspiration*)" can be assigned to the given document. According to Citation-KNN method, we can get the keywords "*网络银行*(*network bank*), *监管* (*supervision*)" of the same document.

In this article, we employ Precision, Recall and F1 value to evaluate the extraction result of automatic implicit keyword. In this experiment, we must at first extract words from keywords given by authors which do not appear in the documents and then use these words to evaluate the performance of automatic implicit keyword extraction.

Table II shows the result of two automatic implicit keyword words extraction. According to table II, we can find that automatic implicit keyword words extraction based on KNN or Citation-KNN is effective to some extent. And automatic implicit keyword words extraction based on Citation-KNN is superior to the one based on KNN. It illustrates that the reliability of the algorithm based on Citation-KNN is better than the one based on KNN. We can also see from Table II that precision and recall of both methods is under 50%. So we'll try to find ways to improve the performance of automatic implicit keyword words extraction based on KNN and Citation-KNN in future.

TABLE I. SIMILAR CHINESE DOCUMENTS (TOP-10) OF A GIVEN CHINESE DOCUMENT

Title of documents	**Keywords of documents**
网络银行发展中的问题及其对策 (The Problems of Network-Bank During the Developing Proceeding and the Countermeasures)	网络银行(network bank), 金融电子化(financial electrification), 金融监管 (financial supervision)
网络银行理论及其在我国的实践 (Theory of Network Bank and Its Practice in China)	网络银行(network bank), 理论依据(theoretical basis), 发展前景(prospect)
全球网络银行的发展与中国网络银行发展战略(The Strategic Development of Global and Chinese Internet Bank)	网络银行(internet bank), 生成机理(producing mechanism), 制约因素(restricted factor), 发展战略 (developing strategy)
对我国网络银行发展与监管问题的研究(A Study on the Development and the Supervision of Internet banking in China)	网络银行(internet banking), 监管(supervision)
网络银行的竞争优势探析(An Analysis of the Competitive Advantage in the Internet Bank)	网络银行(internet bank), 竞争优势(competitive advantage), 阻碍因素(hindering factors), 政策建议(policy suggestions)
网络银行的安全性分析(Security analysis of Internet Bank)	网络银行(internet bank), 安全性(security)
国外网络银行发展模式的启示 (The Inspiration from the Model of Developing Internet Bank in Foreign Countries)	网络银行(internet bank), 网络安全(internet security), 发展模式(development model), 启示 (inspiration)
西方网络银行的发展战略及启示 (Inspiration of the Western Internet Bank's Development Strategy)	网络银行(internet bank), 发展战略(development strategy), 启示(inspiration)
我国网络银行集约化经营之策略 (The Strategies to Manage the Chinese Banks intensively in the internets)	网络银行(internet bank), 集约化经营(intensive management), 网上支付(online payment), 便利服务(facilitation services), 网络顾客(network customers)
我国发展网络银行所面临的问题与对策(Problems of Our Developing Net Bank and the Countermeasures)	网络银行(internet bank), 创新 (innovation), 对策 (countermeasures)

TABLE II. RESULT OF AUTOMATIC IMPLICIT KEYWORD WORDS EXTRACTION

Model	**P**	**R**	**F_1**
KNN	0.2586	0.4804	0.3362
Citation-KNN	0.3577	0.4795	0.4097

IV. CONCLUSION AND FUTURE WORK

In this paper, automatic implicit keyword words extraction based on Citation-KNN is put forward which is based on KNN. The experiment result shows that this method is effective to complete the task of automatic implicit keyword words extraction.

The problem of automatic implicit keyword extraction is that effect of this method is dependent on the scale of test dataset. Documents which are similar to the given document can only be found when the test dataset is big enough. Then the reliability of automatic implicit keyword extraction is ensured. Meanwhile, it's also a key problem to compute the similarity of document which has vital influence on the performance of automatic implicit keyword extraction.

The future wok includes: get a large scale dataset the documents in which have keywords; improve the reliability of automatic implicit keyword words extraction based on Citation-KNN; put forward a reliable evaluation method to evaluate automatic implicit keyword words extraction; explore methods to compute similarity between documents.

ACKNOWLEDGEMENTS

This research was supported by Project of the Education Ministry's Humanities and Social Science (No.08JC870007), Youth Research Support Fund of Nanjing University of Sci. & Tech. (No.JGQN0701), Scientific Research Foundation of Nanjing University of Sci. & Tech. (No.AB41123), and National Key Project of Scientific and Technical Supporting Programs (No.2006BAH03B02).

REFERENCES

[1] A. Rauber and D. Merkl. SOMLib: A Digital Library System Based on Neural Networks. Proceedings of the Fourth ACM conference on Digital Libraries, Berkeley, CA, USA, 1999: 240~241.

[2] H. P.Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1957, 1(4): 309~317

[3] H. P.Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 1958. 2(2): 159~165.

[4] H. P. Edmundson, V. A. Oswald. Automatic Indexing and Abstracting of the Contents of Documents. Planning Research Corp, Document PRC R-126, ASTIA AD No. 231606, Los Angeles, 1959: 1~142.

[5] H. P. Edmundson. New Methods in Automatic Abstracting Extracting. Journal of the Association for Computing Machinery, 1969, 16(2): 264~285.

[6] L. E. Lois. Experiments in Automatic Indexing and Extracting. Information Storage and Retrieval, 1970, 6: 313~334.

[7] G. Salton, A. Wong, C. S. Yang. A Vector Space Model for Automatic Indexing. Communications of ACM, 1975, 18(11): 613~620.

[8] P. D. Turney. Learning to Extract Keyphrases from Text. NRC Technical Report ERB-1057, National Research Council, Canada. 1999: 1~43.

[9] Turney P D. Learning Algorithms for Keyphrase Extraction. Information Retrieval. 2000, 2:303~336

[10] E . Frank, G. W. Paynter, I. H. Witten, et al.. Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), California: Morgan Kaufmann, 1999, 668~673

[11] A . Anjewierden, S. Kabel. Automatic Indexing of Documents with Ontologies. In: Proceedings of the 13th Belgian/Dutch Conference on Artificial Intelligence (BNAIC-01), Amsterdam, Neteherlands, 2001: 23~30.

[12] T . Tomokiyo, M . Hurst. A language Model Approach to Keyphrase Extraction. In: Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition & Treatment, Sapporo, Japan, 2003: 33~40.

[13] A. Hulth. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216~223.

[14] S. J. Li, H. F. Wang, S. W. Yu, C. S. Xin. Research on Maximum Entropy Model for Keyword Indexing. Chinese Journal of Computers, 2004, 27 (9): 1192~1197.

[15] G . Ercan, I . Cicekli. Using Lexical Chains for Keyword Extraction. Information Processing and Management, 2007, 43(6): 1705~1714.

[16] P. D. Turney. Extraction of Keyphrase from Text: Evaluation of Four Algorithms. Technical Report ERB-1051, National Research Council, Institute for Information Technology, 1997

[17] T. M. Cover, P. E. Hart. Nearest neighborhood pattern classification. IEEE Transactions on Information Theory, 1968, IT-13 : 21~27

[18] Y. Yang, X. Liu. A Re-examination of Text Categorization Methods. In: Proceedings of 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), Berkeley, CA, USA, 1999: 42~49.

[19] P. Tan, M . Steinbach, V. Kumar. Introduction to Data Mining. Boston: Addison-Wesley, 2006: 225.

[20] J . Wang, J. D. Zucker. Solving the Multiple-instance Problem: A Lazy Learning Approach. In: Proceedings of 17th International Conference on Machine Learning (ICML2000). San Francisco: Morgan Kaufmann Publishers, 2000: 1119~1125.

[21] R. Baetz-Yates, B. Ribeiro-Neto. Modern Information Retrieval. New York: Association for Computing Machine (ACM) Press, 1999: 27~30.

[22] C. Z. Zhang, X. N. Su, D. M. Zhou. Document Clustering Using Sample Weighting. In: He YX, Xiao GZ, Sun MS eds. Recent Advance of Chinese Computing Technologies Singapore: Chinese and Oriental Languages Information Processing Society, 2008, 3: 260~265.