Improving Index Term Extraction for Chinese Books with Professional Score

Shu'qi Lv[a], Ning Li[b] and Ying'ai Tian[c]
Computer School, Beijing Information Science and Technology University, Beijing, China
[a]Shuqi150915@163.com, [b]ningli.ok@163.com, [c]tianyingai149@hotmail.com

Keywords: Back-of-the-book Index; Index Term Extraction; Wikipedia; PageRank

Abstract. The current situation of the index term extraction for Chinese books was investigated. Aiming to improve performance of traditional key phrase extraction methods for extracting index terms, we propose a novel feature named professional score to evaluate the importance of each candidate. Wikipedia is used to identify whether candidates are meaningful keywords in the domain of the book. Then, we quote the idea of PageRank algorithm to calculate the professional score of candidates by fully utilizing the category structure and citing relationships in Wikipedia. To evaluate the performance of our proposed feature in improving the index term extraction for Chinese books, the traditional TF-IDF and the combination method of TF-IDF and our proposed professional score are conducted. It is found that the precision, recall and F-measure obtained by the combining method are respectively higher 54%, 35% and 46% than those obtained by the traditional TF-IDF.

Introduction

There is no doubt that book is one of the most traditional forms for preserving and inheriting information. While how much knowledge can readers learn from a book has been in separable from the way it provides information. Hence, the crucial way of using book properly is to reorganize the knowledge that can allow readers to quickly obtain the useful information. The back-of-the-book index is one of the important means to make effective use of books, which typically consists of the important phrases and their locations generally in alphabetical order, providing an overview of the books to guide readers to locate the information they most need to consult. Fig. 1 shows a representative back-of-the-book index snippet, which contains all the three essential elements of book index: index term, semantic relationship, and location.

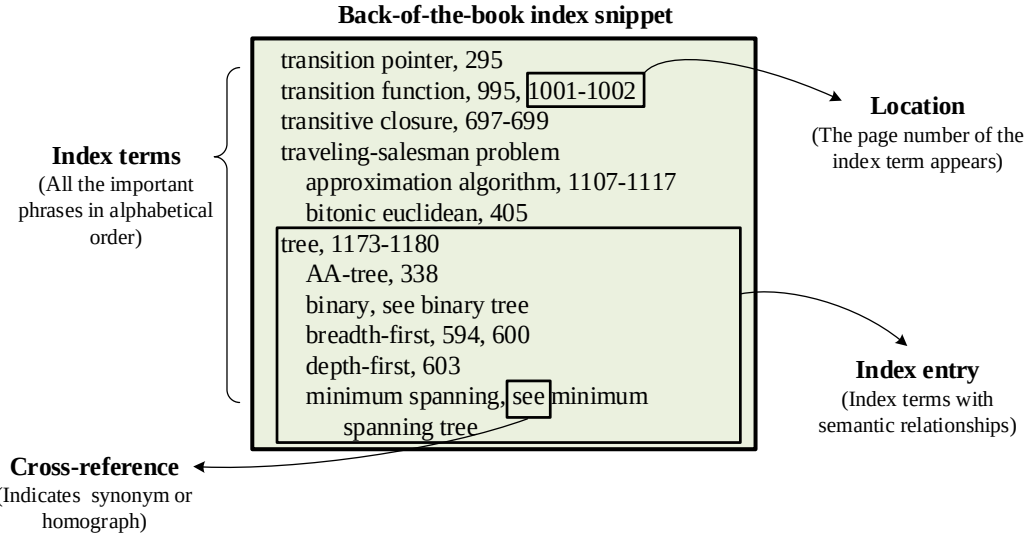

Fig.1 The back-of-the-book index snippet

The research on the back-of-the-book establishment starts overseas, and already has some indexing tools, such as CINDEX, Macrex, and DEXter [2]. Though these indexing tools couldn't fully relieve the manual indexing work, they do provide great help for professional indexers, and also promote the popularity of back-of-the-book index aboard. However, there are few people devote themselves to automatic establish book index for Chinese books, and the domestic indexing tools: Indexing star and the embedded indexing module in the Microsoft Word even couldn't extract index terms automatically. And the investigation [3] shows that nearly ninety percent of Chinese books are lack of back-of-the-book index. Whereas, very little attention has been drawn to automatic book indexing by taking advantage of the online information resource.Hence, it's quite vital for us to explore the back-of-the-book index establishment for Chinese books.

In this paper, we mainly focus on the first and major task of back-of-the-book index establishment, index term extraction, and provide a novel feature to select the proper index terms for Chinese books. The standard *Guideline for Establishment of Indexes (General)* GB/T 22466-2008 [1] points out that index terms should be the concepts or features, not just key phrases, which normally in the form of noun and noun phrase or adjective if necessary, covering the altogether book contents. Moreover, the books from various domains are different in expression. Hence, the extraction work needs to be built on the full analysis of books. Csomai and Mihalce, the authorities in the field, evaluated various of informativeness measurements such as TF-IDF, x^2 and language model for selecting index terms from the book [5], and provided a supervised learning framework for index term extraction based on discourse comprehension, syntactic features and encyclopedic features [6]. Even though their work like other researches [7,8], is closer to key phrases extraction for long document rather than build a real back-of-the-book index, they achieve the goal of automatic extract key phrases from English books. However, most of the features they introduced belong to inner informations of the book. And for the encyclopedia feature, they only took advantage of the article page and hyper links to calculate the probablity of being key words in the Wikipedia database, and did not explore the category structure and citing relationship of Wikipedia. Hence, to make better use of Wikipedia and make improvement in the index term selection for Chinese books, we propose a novel external feature named professional score, exploiting Wikipedia and PageRank algorithm to evaluate the importance of each candidate in the domain of the book. In the experiments, we extract index terms repectively using the traditional key phrase extraction method TF-IDF and the combination method of TF-IDF and our proposed professional score. The results show that the precision, recall and F-measure obtained by the novel method with professional score are respectively higher 54%, 35% and 46% than those obtained by the traditional TF-IDF.

Related Work

IndexTerm Identification Method. Though the studies on the index term extraction for Chinese book are scant, as the closest work, key phrase extraction has been considerable improved. The typical approaches can be divided into supervised extraction method and unsupervised extraction method [9]: the former requires the availability of a training corpus with features extracted from documents to select key phrases [10-12]; the latter requires unlabeled data and relies on TF-IDF or other similar measures to score the candidate phrases [13-15]. Probably many of these approaches can be used to extract index terms from Chinese books. Nonetheless, these approaches only utilize the inner information of the documents or need to train the corpus before extraction. Moreover, books are professional sources for readers to learn useful information, authoritative index terms

should be provided. Hence, external encyclopedia should be fully considered to improve the extraction method and ensure the professionalism of index terms.

PageRank Algorithm. PageRank is the key technology of Google,which utilizes the voting mechanism to evaluate the relative importance of nodes indifferent classes of networks, including social network, citation network, biological network, and road network [16]. The closet work for us is the study on citation network for bibliographic evaluation: Nykl and Campr utilized the PageRank algorithm to rank authors of scientific publications in the citation network [17]; Ding evaluated author citation networks based on the weighted PageRank algorithm in terms of the number of individual author citations [18]. Inspired by these, we consider the PageRank algorithm as a measure of professional importance for phrases to help identify professionalism index terms for Chinese books.

Wikipedia. As the biggest online encyclopedia for retrieval available today, Wikipedia consists of millions of articles from all sorts of field,whose titles namely entries are usually well-formed keywords. Each entry describes a topic and belongs to at least one category of Wikipedia. Wikipedia being a structured resource has been widely used in Lexicon acquisition, information extraction and text categorization [19-20]. And the category structure of Wikipedia can also be used for entity ranking [21], which is quite close to index term identification. Hence, as an open and structured resource, Wikipedia can be used to measure whether the phrases are professional in the domain. Combined with PageRank algorithm, it can be more effective.

Professional Score Evaluation

In this section, we present the major part of our paper, professional score evaluation, using PageRank algorithm to explore the external information in Wikipedia to optimize the performance of index term extraction for Chinese book.

Given the candidates of the book, we can use Wikipedia as the background network library to identify whether the phrases are meaningful keywords in the domain of the book. Then we utilize the category structure of Wikipedia graph and citing relationships between entries, quoting the idea of PageRank algorithm to rank the importance of candidates, in order to evaluate the candidates based on the professional Web knowledge for professional scores.

To insure the PageRank algorithm work best in Wikipedia, there are two issues we should further discuss: domain and initial value.

Domain. Books from various domains have difference in their own emphasis. And category information has been proved to be a highly effective source of information for improving in retrieval performance [21]. Therefore, the category of the book should be considered to construct the PageRank computation for specific domain of the book. We utilize the category to select the domain-related candidates, removing the contribution of candidates which are in the disparate domain for the evaluation. The whole process is relying on the category structure of Wikipedia.

While the structure of Wikipedia category is a directed acyclic graph and each entry may have multiple categories simultaneously [22], which actually cause the difficulty in extracting domain-relevant categories. Hence, in order to simplify this problem, we analyze the structure of Wikipedia category graph and transform it into hierarchical tree-structured taxonomy. The page “Category: Categories” is the most central node in the Wikipedia category system, and all the other categories can be searched from here. Hence, at first, we use “Category: Categories” page as the first level node of our category structure, and select its twenty subcategories to be the second level nodes. All these categories are classified based on subject according to the Dewey Decimal System and the Chinese Library Classification(CLC). Next, we traverse all these twenty nodes for

descendants to get the third level nodes by searching category inlinks and outlinks. Do the same process to the following categories to complete the tree-structured taxonomy. For those categories have already been obtained, we still re-record and label them. Thus, in this hierarchical tree-structured taxonomy, there is the containment relationship between the neighboring levels. The categories of the lower layer are the refinement of the upper level.

With all the previous preparations being completed, we use the domain of the book and our tree-structured taxonomy to identify the domain-related candidates. Chinese books provide CLC in the front page, which shows the domain of the book. We utilize the numbers and names of CLC to match with the categories of Wikipedia, to select the optimal categories for the pending book, to determine target categories. For the N phrases which belong to the identical target category or descendants of the target categories, we consider these candidates to be domain relevant.

Initialvalue. Each node in PageRank algorithm has been assigned an initial value. In the citation network, Ding used the number of citations received by this node or the number of publications where this node acts as a first author, to be the initial weight of author, in order to rank the authors. Therefore, in order to distinguish the importance of candidates and improve the accuracy of calculation, we do the same process to the candidates. To evaluate the professional score of the candidates, the initial weight should be assigned based on the number of citations and statistics information. We make use of the observation that the phrase with higher gram and cited by more entries tend to have more chance be a professional term. Hence, for the candidates in the target domain, we combine phrase length and the number of citations in same category to be the initial value. While for those candidates we could not match to the target categories in the Wikipedia, the initial weight is set as zero, which means these phrases doesn't have any contribution to other candidates in the PageRank calculation.

Formally, **C** is the candidate set, **D** is the target category set. At first, we search the candidate *c* in the Wikipedia to get a Wikipedia matching entries set **E**. For every *c* in the set **E**, we extract all the citation relations between *c*, and match its categories with the target category set **D** to get the domain-relevant entries set **K**. After that, we use all the *N* candidates *c* in **K** and all the citation relations to generate N-dimensional transfer matrix $\mathbf{M_G}$. Next, we use the formula $w_p(c_i) = l(c_i) \times \frac{n(c_i)}{\sum_{k=1}^{N} n(c_k)}$, in which $l(c_i)$ is the phrase length and $n(c_i)$ denotes the number of citations for c_i in the same categories, $\sum_{k=1}^{N} n(c_k)$ represents the sum of citations number of all the candidates, to assign the initial weight to all the candidates *c* in **K** to structure the initial vector **p**. Ultimately, iteratively execute the following formula *x* times until **p** gradually converge to a steady value.

$$\mathbf{P}_{i+1} = (1-\delta)\mathbf{M}_{\mathrm{G}} \times \mathbf{P}_i + \delta \mathrm{v}(1)$$

Where, δ is the damping factor, normally assigned as 0.85, and for the teleportation vector v, all the elements are $\frac{1}{N}$. For candidates from the target categories, the value of the corresponding elements from **p** will be the PageRank value.

While, PageRank value is relative, which represents the relative importance of each node. And when a large number of candidates are computed, the value of each node is rather small. Hence, in order to better represent the professionalism, we use the formula $p(c_i) = \frac{\mathbf{p}(i)}{\mathbf{p}(j)}$ to denote the professional score of candidate c_i, in where p(*i*) is the PageRank value of candidate c_i and p(*j*) provides the highest PageRank value in the evaluation. While, for those candidates which don't match to any entries or don't belong to the target categories in the Wikipedia, the professional score will be zero.

The pseudo code of the aforementioned procedureis given in *Algorithm I*.

*Algorithm I*Professional score evaluation

Input: Candidate set **C**, Target category set **D**
Output: Professional score$p(c)$
1: **for**$c \in$ **C do**
2: Search c in Wikipedia to get matching entry set **E**;
3: **for**$c \in$ **E do**
4: Check the category of c with target category set **D** to get domain-relevant entry set **K**;
5: **end for**
6: **end for**
7: Generate the transfer matrix $\mathbf{M_G}$ based on $c \in$ **K**;
8: Get initial value $w_p(c)$ ($c \in$ **K**) according to section 3;
9: Iteratively calculate **p** according to (1), until it converges to a steady value;
10: **for**$c \in$ **C do**
11: **if**$c \in$ **K then**
12: Calculate $p(c)$ according to section 3;
13: **else if**
14: $p(c) = 0$;
15: **end if**
16: **end for**

Experiment

In the previous section, we introduce the professional score evaluation of candidates by utilizing Wikipedia and PageRank algorithm. In order to evaluate the performance of our proposed feature, some index term extraction experiments respectively using traditional TF-IDF and the combination method ofTF-IDF and our proposed professional score are conducted, and the results are compared in terms of precision, recall and F-measure in this part.

For the combination method of TF-IDF and our provided feature, we introduce adjustment factor α to combine TF-IDF with professional score, and obtain a new value $V(c)$ of candidate c with the formula given as

$$V(c) = \alpha P(c) + (1 - \alpha)T(c) \quad (2)$$

Where $p(c)$ denotes the professional score of candidate c, and $T(c)$ represents the normalized TF-IDF value of c.

Dataset. The experiment dataset is composed of Chinese books and labeled index terms. The original books are domain-specific text book in pdf format. We extract the paged body text and save them in the word documents, and introduce the following traditional steps to generate all the possible index terms [13]:1)Word segmentation by using HMM method, then 2)Filter stop words with needless part-of-speech, 3) Collect all the phrases forcandidates. For index terms, we invite four indexers to help us manually label them, according to the standard *Guideline for Establishment of Indexes (General) GB/T 22466-2008*. And all four indexers' work are analyzed to get average results for the standard set.

Result Analysis. Fig. 2 depicts precision, recall and F-measure obtained by the combination method of TF-IDF and our proposed professional score versus different adjustment factor α, i.e. revealing how the different adjustment factors affect the overall performance of the combined method. As shown in the figure, we can observe that the larger is the adjustment factor of professional score, the better is the performance of the combining method, which also indicates our

proposed feature can effectively identify the index terms for Chinese books.The precision and F-measure of the combined method both reach the best when the adjustment factor of professional score is around 0.7.

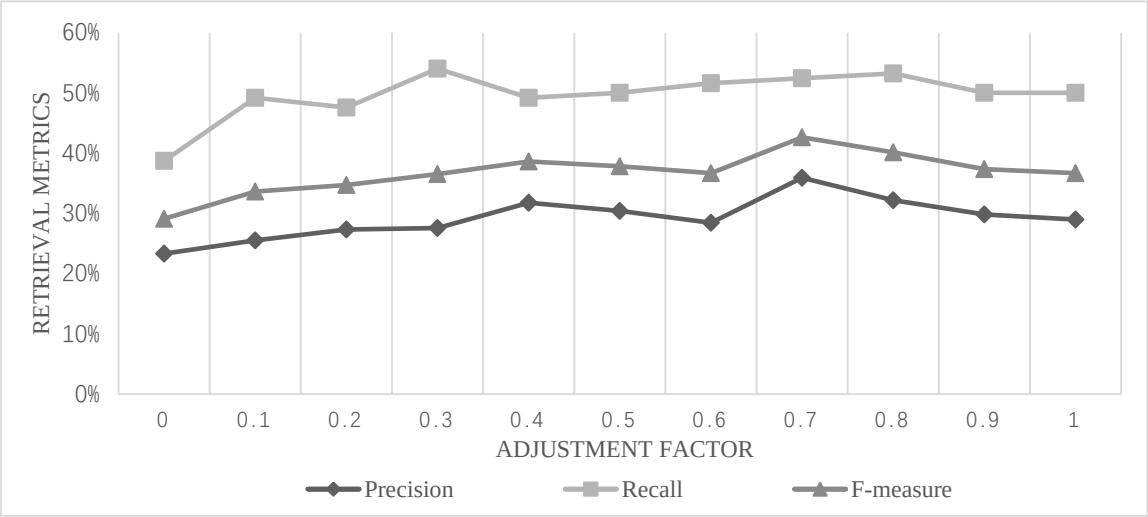

Fig.2 Overall performance

According to the results of Fig.2, we set adjustment factor α as 0.7 for the best performance of the combining method to compare with the traditional TF-IDF. The difference between the two aforementioned methods in terms of precision, recall and F-measure are shown in Fig. 3. It found that the precision, recall and F-measure obtained by the combination method of TF-IDF and our proposed feature are respectively higher 54%, 35% and 46% than those obtained by the traditional TF-IDF. The analysis of the improvement are as follows. Firstly, our proposed feature can identify the domain-related keywords of the Chinese book and reduce the influence of the phrases in disparate domain. Secondly, professional score is used to evaluate all the candidates between each other, not just calculating the value based on statistical features. Accordingly, our proposed professional score can improve the traditional key phrase extraction methods to satisfy the index term extraction method for Chinese books.

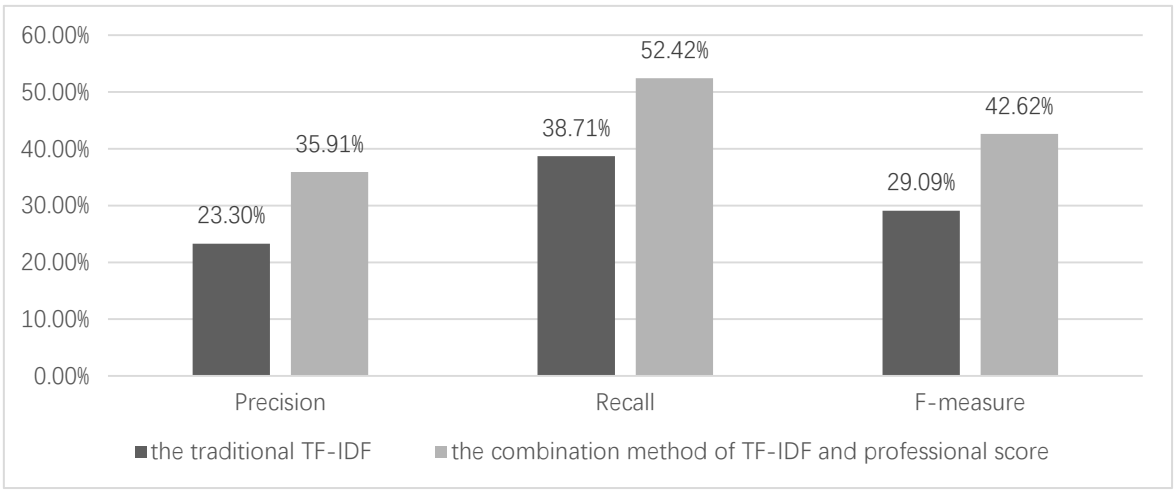

Fig.3 Result of contrastiveexperiment

Conclusion

In this paper, we propose a novel feature to improve index term extraction for Chinese books, which provides help for selecting authoritative index terms for Chinese books via fully utilizing the Web knowledge in the Wikipedia and reference mechanism of PageRank algorithm. In the contrastive

experiments, the traditional TF-IDF and the combination method of TF-IDF and our proposed feature are conducted for evaluating the performance of our proposed feature. The results indicate that the precision, recall and F-measure obtained by the combining method are respectively higher 54%, 35% and 46% than those obtained by the traditional TF-IDF.

However, we mainly focus on introducing Wikipedia feature to index term extraction, while there are more features of Chinese books which could be used to optimize index term extraction for back-of-the-book index establishment, such as statistical feature and structural feature. In the future, we will continue to improve index term extraction for Chinese books by considering more features combined with professional score.

Acknowledgement

This work is partly supported by the National NaturalScience Foundation of China (No. 61672105), the National High-tech R&D Program (863 Program No. 2015AA015403), the General Program of Science and Technology Development Project of Beijing Municipal Education Commission (No. KM 201511232013).

References

[1] Guideline for Establishment of Indexes (General) GB/T 22466-2008.

[2] Y. Kang. The Comparison of Foreign Book Indexing Software. Journal of The China Society of Indexers, pages 18-23, 2009.

[3] J. P. Qiu and Q. Yang. The research on the development of back-of-the-book indexes in the digital publishing situation. Journal of The China Society of Indexers, pages 12-17, 2014.

[4] L. Su. The Comparison of Indexing Star and Microsoft Word Indexing Module. Journal of The China Society of Indexers, pages 6-11, 2006.

[5] A. Csomai and R. Mihalcea. Investigations in unsupervised back-of-the-book indexing. Association for the Advancement of Artificial Intelligence, pages 211-216, 2007.

[6] A. Csomai and R. Mihalcea. Linguistically motivated features for enhanced back-of-the-book indexing. Proceedings of Association for Computational Linguistics, pages 932-940, 2008.

[7] A. Nazarenko and TAE Mekki. Building back-of-the-book indexes. Terminology, pages 199-224, 2004.

[8] Z. H. Wu, Z. H. Li, P. Mitra and C. L. Giles. Can back-of-the-book indexes be automatically created. Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages 1745-1750, 2013.

[9] K.S. Hasan and V. Ng. Automatic Keyphrase Extraction: A Survey of the State of the Art. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1262-1273, 2014.

[10] I. H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. Acm Conference on Digital Libraries, pages 254-255, 2010.

[11] D. X. Wang, X. Y. Gao and P. Andreae. DIKEA: Exploiting Wikipedia for keyphrase extraction. Web Intelligence, pages 153-165, 2015.

[12] C. Wang and S. J. Li. CoRankBayes: Bayesian learning to rank under the co-training framework and its application in keyphrase extraction. Acm International Conference on Information & Knowledge Management, pages 2241-2244, 2011.

[13] Y. J. Chen, X. D. Shi, C. L. Zhou, and C. Su. Automatic Keyphrase Extraction from Chinese Books. Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pages 92-97, 2007.

[14] Z. Y. Liu, C. Liang, and M. S. Sun. Topical word trigger model for keyphrase extraction. Proceedings of COLING, pages 1715-1730, 2012.

[15] A. Bougouin, F. Boudin, and B. Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. International Joint Conference on Natural Language Processing, pages 543-551, 2013.

[16] P. Garg, K. Kothapalli. STIC-D: Algorithmic techniques for efficient parallel PageRank computation on Real-World Graphs. International Institute of Information Technology, 2016.

[17] M. Nykl, M. Campr and K. Jezek. Author ranking based on personalized PageRank. Journal of Informetrics, pages 777-799, 2015.

[18] Y. Ding. Applying weighted PageRank to author citation networks. Journal of the American Society for Information Science and Technology, pages 236-245, 2011.

[19] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Joint Conference on Emnlp & Cnll, pages 708-716, 2010.

[20] E. Gabrilovich, S. Markovitch. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. National Conference on Artificial Intelligence & the Eighteenth Innovative Applications of Artificial Intelligence Conference, pages 1301-1306, 2015.

[21] R. Kaptein and J. Kamps. Exploiting the category structure of Wikipedia for entity ranking. Artificial Intelligence, pages 111-129, 2013.

[22] T. Y. Shi, S. D. Jiao, J. Q. Hou and M. L. Li. Improving keyphrase extraction using Wikipedia semantics. Second International Symposium on Intelligent Information Technology Application, pages 42-46, 2008.