Text Retrieval Technology Based on Keyword Retrieval

Text Retrieval Technology Based on Keyword Retrieval

Chunhao Huang，Zhiyuan Zhu and Jing Guo*

College of Electronic and Information Engineering Southwest University Chongqing, China

*Corresponding author：poem24@swu.edu.cn

Abstract: In such an era of informatization and big data where the Internet has greatly enriched the sources of information of people and changed the way they get information, search engine is widely used and it results from the combination of text retrieval and Web. Based on the research on Boolean model, vector model and probability model, this paper deduces a text retrieval algorithm using keyword retrieval, which can realize keyword retrieval of text documents in text database.

1．Introduction

Information retrieval based on text content includes text retrieval, image retrieval and picture retrieval, etc. The earliest and most typical text retrieval is the book index used in libraries. The librarian indexes some key information of the book such as the name of the book, the author of the book, the information of the publisher and the date of publication so that the reader or the librarian can quickly apply the index to find the location of the book. With the emergence of computers, people can give the retrieval task of index to computers so as to manage more documents more conveniently and more quickly. In order to make book search faster, the first generation of text retrieval technology appeared, that is, keyword retrieval was used to return the matched document information to the user as the result.

As researchers developed text retrieval, they began to integrate neural network, fuzzy system, evolutionary computation and other soft computing technologies into text retrieval. So far, great achievements have been made in the development of domestic text retrieval technologies, such as the concept space-based extended search [1] realized by Institute of Computing Technology, Chinese Academy of Sciences, concept retrieval [2] realized by Beijing University of Posts and Telecommunications in a specific field, associative retrieval system [3] jointly developed by University of Science and Technology of China and Institute of Computing Technology, Chinese Academy of Sciences, Web information retrieval based on conceptual relevance feedback mentioned in Literature [4], concept-based Chinese search engine mentioned in Literature [5] and intelligent fuzzy information retrieval system mentioned in Literature [6]. All the above achievements are systems developed by researchers through a large number of studies on text retrieval. Text retrieval technology has great research value, and scientists are studying it further.

2．TEXT RETRIEVAL

Information retrieval based on text content includes text retrieval, image retrieval and picture retrieval, etc. The earliest and most typical text retrieval is the book index used in libraries. The librarian indexes some key information of the book such as the name of the book, the author of the book, the information of the publisher and the date of publication so that the reader or the librarian can quickly apply the index to find the location of the book. With the emergence of computers, people can give the retrieval

task of index to computers so as to manage more documents more conveniently and more quickly. In order to make book search faster, the first generation of text retrieval technology appeared, that is, keyword retrieval was used to return the matched document information to the user as the result. This paper introduces three traditional retrieval models: Boolean model, vector model and probability model [7].

2.1. Boolean Model

Boolean model [8] is a widely used matching model derived from set theory and Boolean algebra. The query request of this model utilizes the common operators including AND, OR and NOT. The result is a document with a return value of ‘‘true’’ at query time. In the model, a query word including keywords and operators is a Boolean expression. Its text is represented by a set of binary-state variables [9], usually consisting of an entry or phrase in the training text. True will be given if the entry or phrase is contributed to the text while False will be given if not [10]. At the time of retrieval, the text can be divided into two parts including matching set and unmatching set according to the conditions submitted by the user. However, the Boolean model has some defects. Firstly, only the documents and conditions can be completely matched when retrieving, so too many results will be returned and the retrieval function will be insufficient as a result. Secondly, the conditions provided by users are often difficult to be converted into the required Boolean expressions for retrieval so the retrieval task can’t be completed [11]. In addition to these defects, the Boolean model is still the main model adopted by most web search sites such as Yahoo and InfoSeek.

2.2. Vector Space Model

Vector space model [12] uses weights and vectors to represent users’ query items and information. When retrieving, it only needs to calculate the similarity between vectors to get the desired results. The most important methods of vector space model are weight calculation method and similarity calculation method. The most famous weight calculation method is IDF weight calculation method [13], which is a combination of the frequencies of the index entries in the document. The similarity is usually calculated by using two methods of the vector inner product or the included angle cosine [14]. The result is a number between 0 and 1, which is consistent with the property of similarity percentage. Included angle cosine calculation is to divide the product of each vector inner product by the inner product of vectors. When the calculated cosine value is 0, it means that there is no coincidence between the retrieved query item and the file.

The above vector space model method turns the text retrieval problem into a similarity comparison between two vectors, and the problem becomes simpler.

2.3. Probability Model

Every entry is related to each other, but the Boolean model and the vector model ignore this and treat each word as an independent item. However, the probability model takes the internal relationship [15] between the entry and text into account and takes advantage of such relationship.

Probability model [16] was proposed by Robertson and Sparck Jones in 1973. Its basic idea is to divide documents into two parts according to query items: sets related to query items and sets unrelated to retrieval items. The index entries of document sets of the same class have similar or same distribution, but those of different classes have different distributions. Therefore, the distribution of index entries in the document can be calculated, and the relevancy can be determined according to the calculated results. The binary independent probability model proposed by Robertson and Sparck Jones, a new probabilistic retrieval model [17] proposed by Chengxiang Zhai combined with language model, the probability-based model proposed by Kacprzyk are all typical probability retrieval models [18].

Retrieval technology has been improved as it develops up to now. Scientific researchers have also been studying related technologies in order to find more suitable retrieval system. In addition to the three kinds of traditional retrieval model introduced above, P norm model [19], language model [20], fuzzy logical model and probabilistic inference network [21] extended from the Boolean model are now

popular text retrieval models. Some mature retrieval systems such as parallel information retrieval system, deductive information retrieval system, hypertext-based information retrieval system, distributed retrieval system and intelligent retrieval system also represent the development direction of retrieval technology.

2.4. Keyword Retrieval

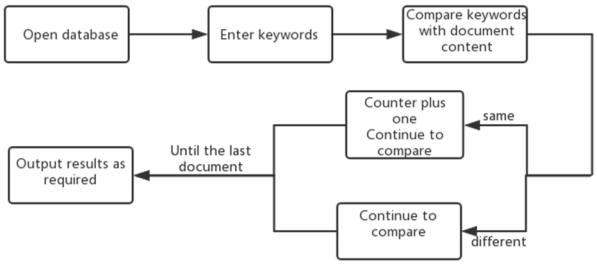

Figure 1. Flow chart of text retrieval

Keyword-based retrieval technology is the key technology of search engine and web text retrieval. If users search, they only need to enter keywords. Through specific search software, the search engine can access the information in the information resources contained in the keywords, find it and return it to the users.

The keyword matching retrieval pattern is usually only relevant if a text contains the same keywords as the user entered, otherwise it is irrelevant. This correlation matching is surface-based matching. Keyword-based retrieval is to form a logical expression through word segmentation of the input query statement and then make a correlation match with the text in the database to return the text documents the similarity of which is greater than a given value or ranks high.

The solutions to the following three problems are the key to the model in speech recognition.

2.5. Experimental Results

Text retrieval uses keyword retrieval technology. The keywords to be retrieved are compared with the characters in the text document in the text database for many times. The content of the document and the title of the document can be retrieved. Finally, the text document information with high relevance ranking can be obtained.

The following figure shows the library txt document numbers from mydoc001 to mydoc300 for text retrieval.

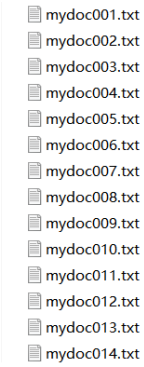

Figure 2. The text retrieval library

The first experiment is to retrieve the content of 300 documents.The number of keywords is 5,The key words are ‘China’, ‘people’, ‘Jintao Hu’, ‘environment’ and ‘politics’.Experimental result: The first is the 274 document "should we cancel the monthly telephone rental ?" the number of key words is 63 times;The second is the 150 document "What behaviors do foreigners dislike?" the number of key words is 52 times;The third is the 79 document "Another life is possible" the number of key words is 37

times.

The second experiment is to retrieve the titles of 300 documents.The key word is "human".The first is the 69th document "what are fewer Chinese than Koreans?";The second is the 201 document "people who love nature are good people";The third is the 276 document "how money destroys and benefits people".

3. Conclusion

Information access is becoming more and more convenient, and text retrieval technology is also used more and more frequently. This paper gives a brief description of the development history of speech recognition technology. In terms of text retrieval, three mainstream text retrieval models are introduced in this paper: Boolean model, vector model and probability model. In the experiment, a text searcher was designed to simulate document retrieval in a pile of txt documents. The keywords retrieval and the document content retrieval by author or title in 300 txt documents were successfully realized.

At the same time, there are also some shortcomings in this paper. This paper applies the keyword retrieval method with high time complexity. Once the amount of data in the database is large, the retrieval speed will be greatly reduced. Therefore, more efficient retrieval methods can be used in the subsequent improvement.

Acknowledgments

The research is supported by Fundamental Research Funds for the Central Universities (Grant No. SWU019040)

References

[1] Zheng Yi, & Wu Bin, & Shi Zhongzhi (2002). A concept space based text retrieve system. Comp uter Engineering and Applications, 38(12), 67-69.
[2] Li Lei, & Wang Nan, & Zhong Yixin(2000). Semantic network based concept retrieval. Journal of the China Society for Scientific and Technical Information ,19(5), 525-531.
[3] Li Yuan, & He Qing, & Shi Zhongzhi(2001). Association retrieve based on concept semantic spa ce. Journal of University of Science Technology Beijing, 23(6), 577-580.
[4] Chia-Hui Chang, & Ching-Chi Hsu(1999). Enabling Concept-Based Relevance Feedback for Info rmation Retrieval on the WWW. IEEE TRANSACTIONS ON KNOWL EDGE ANDDATA ENG INEERNG, 11(4), 595-609.
[5] Hong Zhang, & Yanhong Ma, & Qinyu Zhang, & Pengshou Xie (2005). Study and Design of Ch inese Concept Based Search Engine. Proceedings of ISCIT, 38-41.
[6] Yih Jen Homg, & Shyi Ming Chen, & Chia Hoang Lee (2001). Automatically Constructing Multi Relationship Fuzzy Concept Networks in Fuzzy Information Retrieval Systems. IEEE Intern ational Fuzzy Systems Conference, 606-609.
[7] S. E. Robertson & K. Sparck Jones(1976). Relevance weighting of search terms. Journal of the A merican Society for Information Science, 27(3),129-146.
[8] W.M Shaw Jr, & R.Burgin, & P.Howell (1997). Performance standards and evaluations in IR test collections:Cluster based retrieval models. Information Processing&Management, 33(1), 1-1 4.
[9] Cooper, &Getting,W.S. (1988). beyond Boole. Information Processing and Management, 24, 225-243.
[10] O.Cordo'n
a. E. Herrera Viedma LequeM(2006). Improving the learning of Boolean queries by means o f a multi objective IQBE evolutionary algorithm. Information processing and management, 6 15-632.
[11] Choi J, & Kim M(2006). Adaptive relevance feedback method of extended Boolean model using hierarchical clustering techniques. Information Processing and Management, 331-349.

[12] Salton, G., &Wong,A.& Yang, C.S. (1973). On the specification of term values in automatic Journal of Documentation,29(4), 351-372.
[13] April Kontostathisa & William M(2006). A Framework for Understanding Latent,Semantic Index ing (L SI) Performance. Information Processing and Management 2006,56-73.
[14] Miles Efron(2008). Query Expansion and Dimensionality Reduction: Notions of Optimality in Ro cchio Relevance Feedback and Latent Semantic Indexing.Information Processing and Manag ement, 163-180.
[15] Callan, J. P., & Croft, W. B., & Harding, S. M. (1992). The INQUERY retrieval system. IEEE Transactions on Knowledge and Data Engineering, 4(5), 487-502.
[16] Robertson, S. E., &Sparck, J. K. (1976).Relevance weighting of search term.Journal of the American Society for Information Science, 27(3), 129-146.
[17] ChengXiang Zhai(2006). A risk minimization framework for information retrieval. Information P rocessing and Management, 31-35.
[18] Janusz Kacprzyk, & Katarzyna Nowacka & SlawomirZadrozny(2006),A Possibilistic Logic Bas ed Information Retrieval Model with Various Term-Weighting Approaches. Artificial Intelli gence and Soft Computing -ICAISC 2006. Lecture Notes in Computer Science, 2006, Volum e 4029/2006,1120-1129.
[19] Salton, G., & Fox, E. A., & Wu, H. (1983). Extended Boolean Information Retrieval. Communica tion of the ACM, 36(11), 1022-1036.
[20] Ponte J M, & Croft W B(1998). A Language Modeling Approach to Information Retrieval. Proce edings of 21" International ACM SIGIR Conference on Research and Development in Infor mation Retrieval, 275-281.
[21] FAN Xiao zhong, & LI Hong qiao, & LI Liang fu(2003). Hybrid Chinese Information Retrieval Model Based on the Combination of Keyword and Concept. Journal of Beijing institute of te chnology, 12, 120-123.