A Re-ranking Method Based on Cloud Model

Maoyuan Zhang Zhenxia Lou Jan Wan
Department of Computer Science and Technology
Central China Normal University
Wuhan, China
zhangmy@mail.ccnu.edu.cn louzx@yahoo.cn
wanjian731@126.com

Jinguang Chen
Engineering & Research Center for Information Technology on Education
Central China Normal University
Wuhan, China
cjg2003@hutc.zj.cn

***Abstract*—By introducing cloud model, this paper presents a re-ranking method which improves the accuracy of the IR (information retrieval) while recall is preserved. It is rare in traditional Chinese information retrieval to consider uncertainty while calculating the related degree of the query and each document in the result set. This paper researches IR in a perspective of uncertainty by introducing cloud model, measures the relevance between the query and document by the uncertainty degree that using document represents the query, and then re-ranks the result set. Experiments on NTCIR-5 (the 5th NII Test Collection for IR Systems) document collection for SLIR (Single Language IR) show that this method achieves an 18.08% and 26.50% improvement comparing to the initial retrieval method without any re-ranking.**

Keywords-Information retrieval; Re-ranking; Cloud model; Uncertainty

I. INTRODUCTION

In modern society, how to accurately retrieve the information that user needs from the massive continuous growing document resources becomes more and more important. Many retrieval models provide a great convenience for information retrieval. Ideal search model for IR is that the most relevant document ranks top of the result set. So the user can find the relevant documents fast as to improve the efficiency of information retrieval. How the document set of the retrieval results has been ranked embodies the retrieval precision. Re-ranking is to change the documents' order of the initiative retrieval results so as to improve the accuracy of the IR while recall is preserved.

In order to improve the retrieval precision, researches on a variety of different sorting algorithms and re-ranking methods have also become a hot topic of information retrieval. Two kinds of methods are mainly used for re-ranking: statistics-based method and semantic-based method. As for statistics-based method， Kyung- Soon Lee et al. expanded the original document set and added the document clustering to traditional method instead of query expansion to re-rank documents [1]. Jaroslaw Balinski and Czeslaow Danilowicz proposed a method based on inter-document distances [2]. HE Ting-ting et al. used the distribution of the topic word pairs which are composed of two correlated words respectively selected from the original query words and documents [3]. Dong Zhou and Vincent Wade proposed a novel document re-ranking method based on Latent Dirichlet Allocation (LDA) which exploits the implicit structure of the documents with respect to original queries [4]. The other kind of method is semantic-based method. ZHANG Min et al. introduced semantic relations between words in WordNet for query expansion and replacement in order to re-rank documents [5]. Hongbo Deng, Michael R. Lyu, and Irwin King proposed an approach that incorporates the content with other link information in a latent space graph together, and then performed the re-ranking algorithm on the graph [6]. Zhou Bo, Cen Rongwei et al. proposed a re-ranking method based on document similarity which used both the relevant documents and irrelevant documents in the feedback information [7].

All the methods above estimate the relevance of each document to the query and rank the documents accordingly. However, such an approach ignores the uncertainty associated with the estimates of relevancy for the original initiative retrieval is natural language-related and natural language has uncertainty. C J van Rijsbergen indicated a document is retrieved if it logically implies the request. However, there is always a measure of uncertainty associated with such an implication [8]. In order to do a further research on the uncertainty between the query and the document, this paper explained information retrieval from the perspective of uncertainty, and re-ranked the document set of the retrieval results by introducing the cloud model theory.

The rest of the paper is organized as follows. Section Ⅱ introduces the retrieval system based on cloud model, section Ⅲ introduces cloud model theory and our proposed re-ranking method based on it, section Ⅳ presents experiments and evaluation results, and sectionⅤ makes a conclusion of the whole work.

II. RETRIEVAL SYSTEM BASED ON CLOUD MODEL

We have successfully developed a retrieval system based on cloud model. The system is composed of three levels, which are the query level, the retrieval level, and the re-ranking level, as shown in Fig. 1:

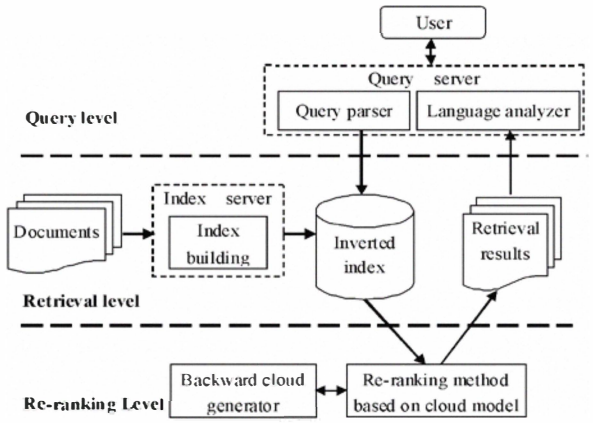

Figure 1. The retrieval system based on cloud model.

A. The Query Level

The query level includes two components: the query parser and the language analyzer. The query parser converts each keyword into an instance of term class and compares them with the terms in index. The language analyzer performs the syntax analysis using dictionary and analysis grammar. For example, it will perform word segmentation on the user's query, because word segmentation is the fundamental task of Chinese information retrieval.

B. The Retrieval Level

The retrieval level is where inverted index is built by index server for the full texts of all documents. The inverted index is widely used in the information retrieval field. It maps each word into a list of documents in which it appears. Each word appearing one or more times in the documents has a corresponding entry in the inverted index. The index server would find matches against a set of documents that could satisfy a query by decoding compressed information in the inverted index.

C. The Re-ranking Level

The re-ranking level is the core part of this system. It performed the re-ranking method based on cloud model. Specifically, it uses backward cloud generator to get the uncertainty degree of each document representing the query based on the initial retrieval results generated by the retrieval level. Each document will get a score for its uncertainty degree to represent the query. After processed by the re-ranking method based on the cloud model, the final retrieval results will be presented to the user.

III. RE-RANKING METHOD BASED ON CLOUD MODEL

A. Cloud Model Theory

Cloud model [9], evolved from entropy and fuzzy set theory, was first proposed in 1990s [10]. It integrates randomness and fuzziness together to research the uncertainty of the concept in natural language. It is a conversion model with uncertainty using the quantity number expression of a quality concept which is expressed by natural language to express the uncertainty within it.

The concept in natural language is qualitative, but it also contains too much uncertainty. For example, "stable" is a qualitative concept, but the uncertainty about it is there are no standards to measure it. Different people may have different understanding about "stable" under various circumstances.

Detailed description about cloud model theory can be found in [9]. We present a brief introduction as written in that reference for the benefit of readers to better understand our work. In cloud model, cloud drop $x_{i\,(i=1,\ldots,n)}$ is a number that randomly realizes the concept *C*. The numerical characteristics of cloud model are expressed with Expectation *Ex*, Entropy *En* and Super-entropy *He*, and they reflect the whole characteristics of the quality conception *C*. *Ex* is the most classical sample of cloud drop while qualifying the concept *C*; *En* measures the uncertainty degree of the qualitative concept *C*; *He* measures the uncertainty degree of *En*. Their definition of computation formulas are introduced as following:

$$Ex = \bar{X} \tag{1}$$

$$En = \sqrt{\frac{\pi}{2}} * \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Ex\right| \tag{2}$$

$$He = \sqrt{S^2 - En} \tag{3}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2$.

B. Definitions

The uncertainty between the document and the query has rarely been taken into consideration in traditional Chinese information retrieval, but cloud model can measure the uncertainty degree of a qualitative concept expressed by a quantity of precise data which gives us much inspiration in information retrieval and re-ranking. Therefore, we researched IR and re-ranking from the perspective of uncertainty, and proposed a method that using the uncertainty between the document and the query to re-rank documents by introducing the cloud model theory.

Our paper considers the query as a qualitative concept, a document related to the query as a cloud drop that randomly realizes the qualitative concept, and then uses cloud model to calculate the uncertainty degree between the query and document by the precise data of query's keywords' distribution in the document. The lower the uncertainty degree of the document expresses the query, the more relevant they are. So the distribution of the keywords in the document will reflect the relevance between the document and the query on uncertainty, as also affect the accuracy rate of IR.

Some IR models also covered the distribution of the query's keywords in the document. Take the free/open source information retrieval software library Apache Lucene for example, it simply considered the proportion of the keywords appearing in the document. Its basic similarity scoring formula [11] is introduced as following:

$$Score(q,d) = \sum_{t\ in\ q} \begin{pmatrix} tf(t\ in\ d) \times idf(t)^2 \times boost(t.field\ in\ d) \\ \times lengthNorm(t.field\ in\ d) \times coord(q,d) \\ \times queryNorm(q) \end{pmatrix} \quad (4)$$

It computed the score for each document d matching each keyword t in a query q. The impact factor *coord(q,d)* in Lucene's scoring system represents the proportion of the keywords appearing in a document. It is one of the important factors which will impact the accuracy of the IR, but it is too vague, and it cannot express the uncertainty comprehensively and precisely. So our method made some definitions as following:

Definition 1: Suppose q represents the query, d represents a document, and then the score of d gets for its uncertainty degree to express its relevance to q is defined as

$$QDU(q,d) = \alpha \bullet Ex - \beta \bullet En - \lambda \bullet He \quad (5)$$

where $\alpha + \beta + \lambda = 1$, and $\alpha > \beta > \lambda \geq 0$.

In (5) *Ex* is the expectation of q's keywords' distribution in d, *En* measures the uncertainty degree of d expressing q, and *He* reflects the uncertainty degree of *En*. They can be calculated by (1), (2), (3) separately. Since *Ex* normally has greater importance than *En* and *He*, and *En* has greater importance than *He*. This paper makes $\alpha > \beta > \lambda \geq 0$.

Definition 2: Suppose q represents the query, d represents a document, *QDUmax* is the maximum score that a document in the retrieval results can get for its lowest uncertainty to present q ,and then the normalized value of *QDU(q,d)* is defined as

$$QDUNorm(q,d) = \frac{QDU(q,d)}{QDU\max} \quad (6)$$

A document with higher *Ex*, lower *En* and *He* will show a lower uncertainty degree of the relevance between the document and the query. Therefore, the higher *QDUmax(q,d)* is, the more relevance the document has to the query, and the toper position it gets in the information retrieval result set.

The re-ranking method is based on the initial results of the information retrieval which doesn't need to run IR for a second time. It reordered the document set of the results. In this paper, we applied Lucene to get the document set of the IR's results. In order to re-rank the documents while take into consideration the uncertainty of a document logically implies the query, the method we proposed in this paper takes the score of a document got for its uncertainty degree expressing its relevance to the query as a factor. This factor can affect the document's order in the result set, as well as the accuracy of the result set. The method's basic scoring formula is defined as following:

Definition 3: Suppose q represents the query, d represents a document, and then the score d gets based on the method we proposed is defined as:

$$QDUScore(q,d) = QDUNorm(q,d) \times Score(q,d) \quad (7)$$

This factor *QDUNorm(q,d)* is present by the quantitive distribution of the keywords appearing in a document. It can be calculated by (6). It is much more accurate and comprehensive than the one in Lucene's scoring system (i.e., *coord(q,d)*) which only considered the proportion of the keywords appearing in a document. *Score(q,d)* has been calculated by Lucene scoring system based on (4). This method fulfilled reconstructing the IR in the respect of uncertainty.

C. The Re-ranking Algorithm

The re-ranking algorithm is presented as follows.

Input: The initial retrieval results.

Output: Retrieval results which has been re-ranked.

Steps:

Step 1. The value of *QDUmax* is set to 0;

Step 2. While the document set of the initial retrieval results has not been processed by step 2 completely, calculate *QDU*(*q*,*d*) of the document according to (5); If all the documents has been processed, go to step 4;

Step 3. If *QDU*(*q*,*d*) is bigger than *QDUmax*, the value of *QDUmax* is set as *QDU*(*q*,*d*);Go to step 2;

Step 4. While the document set of the initial retrieval results has not been processed by step 4 completely, calculate *QDUScore*(*q*,*d*) of the document according to (6), (7); If all the documents has been processed, go to step 5;

Step 5. Re-ranked documents based on their *QDUScore*(*q*,*d*), a document with the highest score ranks top of the document set;

Step 6. Return the re-ranked documents.

IV. EXPERIMENTS AND EVALUATION

In order to evaluate the effectiveness of the proposed method, we conducted our experiments using the NTCIR-5 information retrieval test collections. We used TITLE field in the query set, and chose 20 out of 50 titles in the query set to conduct our experiments, and top 1000 documents of initial information retrieval results conducted by Lucene 3.0. TREC evaluation tool trec eval [12] was used to evaluate the experimental results.

A. Parameter Setting

First of all, this paper will confirm three parameters in (5): α, β and λ by experiments. Since *Ex* normally has greater importance than *En* and *He*, and *En* has greater importance than *He*, we will analyze how the accuracy of the information retrieval was affected by the change of α first. The experiment conducted used two methods: cloud method as cloud which applied (6), using the score of a document gets to express its relevance to the query to re-rank the documents; LC method as the method we proposed in this paper which applied (7), using the results of (6) that the score of a document gets to express its relevance to the query as one of the two factors to re-rank the documents. Fig. 2 shows the changes of the two method's accuracy when α changes.

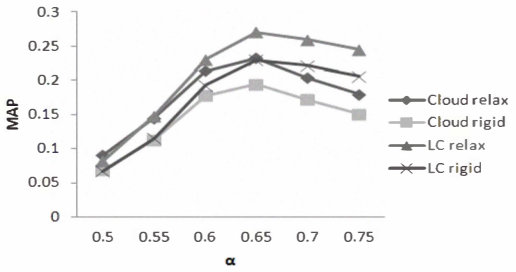

Figure 2. Accuracy of cloud method and LC method with relax and rigid assessments when α changes.

It shows that when α is between 0.6 and 0.7 the two methods will both get the highest accuracy rate. In order to get a more accurate result, a detailed experiment was performed as following in Fig. 3:

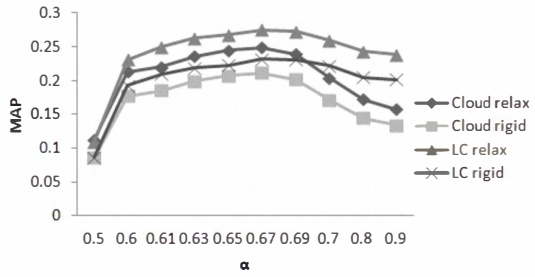

Figure 3. Detailed figures of Fig. 2.

Therefore, we set α=0.67 to obtain the best result according to Fig. 3. And also, we need to figure out how the two method's accuracies change when α or β changes. Since $\alpha+\beta+\lambda=1$ and $\lambda<\beta$, we adjust parameter λ from 0 to 0.16 when α is set as 0.67. The experiment result is shown in Fig. 4.

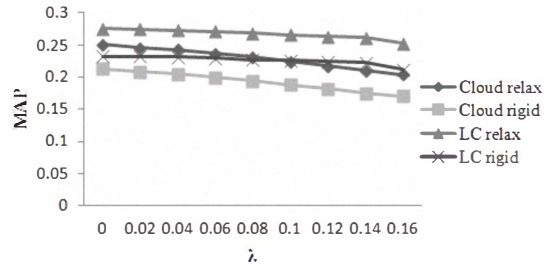

Figure 4. Accuracy of cloud method and LC method with relax and rigid assessments when λ changes.

As shown in Fig. 4, the accuracies of the two methods decrease when λ increases and is bigger than 0.02. The result is sound because *He* is the second-order entropy of the entropy. Hence, the value of parameter λ should be set between 0 and 0.02 to gain the best result. Moreover, the effect of *He* on the uncertainty degree cannot be ignored. Therefore we set λ=0.01 in this paper, and α=0.67, β= 0.32.

B. Comparison of Experimental Results

1) Comparison of the Mean Average Precision

This paper compares the final results' MAP(Mean Average Precision) in Table Ⅰ. The table shows the precision for each case. For each method, we give the percentage of improvement over the initiative information retrieval as base in parentheses. We demonstrate that the performance of our proposed method is better than cloud method which applied (6), using the score of a document gets to express its relevance to the query to re-rank the documents. Cloud method and LC method both get more accurate retrieval result than the initiative information retrieval.

TABLE I. COMPARISON RESULTS ON NTCIR-5 COLLECTION

Standard	Base	Cloud Method		LC Method		
	MAP	*MAP*	*Change over base (%)*	*MAP*	*Change over base (%)*	*Change over cloud method (%)*
Rigid description	0.1834	0.2112	+15.16%	0.2320	+26.50%	+9.85%
Relax description	0.2328	0.2487	+6.83%	0.2749	+18.08%	+10.53%

The table shows the cloud model theory will improve the performance of the information retrieval. Experiments show that cloud method achieves a 6.83% and 15.16% improvements comparing to the initial retrieval without any re-ranking while our method achieves an 18.08% and 26.50% improvements with relax and rigid assessments. These results could be explained. It is practical to research information retrieval in a statistical way. However, it cannot be ignored that information retrieval is natural language related. Hence, there are not only statistical characteristics which natural language have, but also randomness and fuzziness. Cloud model theory made a good combination by characterizing the randomness and fuzziness in natural language by statistical characteristics. Therefore it improves the information retrieval accuracy by re-ranking the documents based on the cloud model theory of uncertainty.

2) Comparison of the precision at R documents

Fig. 6 and Fig. 7 below show the precision at R documents after re-ranking. We can see that the number of the related documents in the top position increases after re-ranking using the two methods. Precision at 5 documents with relax assessment is improved from 0.37 of the initiative information retrieval to 0.45 and 0.42 respectively by using cloud method and LC method, while it is improved from 0.26 to 0.35 and 0.32 with the rigid assessment. The query- related documents are more intensive at the top rank.

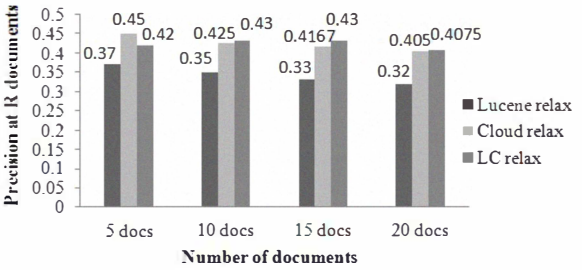

Figure 5. Precision at R documents of the three methods with relax assessment.

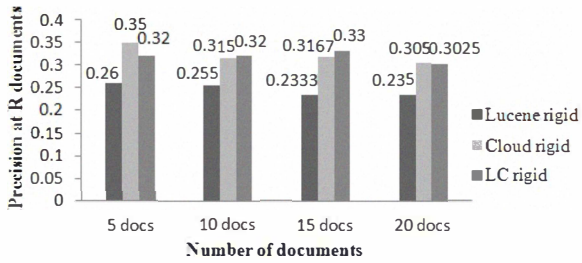

Figure 6. Precision at R documents of the three methods with rigid assessment.

V. CONCLUSION

This paper proposes a novel method for re-ranking to improve the performance of Chinese information retrieval systems by introducing the cloud model theory. That method introduced three numerical characteristics of cloud model, and dug out the uncertainty inside the relevance between query and documents. And then, the method took the uncertainty degree into consideration for re-ranking of information retrieval. Thus information retrieval was expressed from the uncertainty perspective by using the method. In addition, the improvement of the accuracy rate by re-ranking introducing the cloud model theory proved that the method is practical and effective to express the relevance by uncertainty.

ACKNOWLEDGMENT

This work was supported by the Major Research Plan of National Natural Science Foundation of China (No. 90920005), the National Natural Science Foundation of China (No. 61003192), the Program of Introducing Talents of Discipline to Universities (No. B07042), Chenguang Program of Wuhan Municipality (No. 201050231067), and the self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU10A02009, No. CCNU10C01005).

REFERENCES

[1] Kyung-Soon Lee, Young-Chan Park, Key-Sun Choi, "Re-ranking model based on document clusters," Information Processing and Management, Oxford: Pergamon, vol. 37, pp. 1-14, 2001.

[2] Jaroslaw Balinski, Czeslaow Danilowicz, "Re-ranking method based on inter-document distances," Information Processing and Management, Oxford: Pergamon, vol. 41, pp. 759-775, 2005.

[3] HE Ting-ting, XU Ting, QU Guo-zhong, TU Xin-hui, "Re-ranking based on topic word pairs," Computer Engineering and Applications, Beijing, vol, 43(11), pp. 161-163, 2007.

[4] Dong Zhou, Vincent Wade, "Latent Document Re-Ranking," Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1571–1580, 2009.

[5] ZHANG Min, SONG Rui-Hua, MA Shao-Ping, "Document Refinement Based on Semantic Query Expansion," Chinese Journal of Computers, Beijing:Science, vol. 27(10), pp. 1395-4001, 2004.

[6] Hongbo Deng, Michael R. Lyu, and Irwin King, "Effective Latent Space Graph-based Re-ranking Model with Global Consistency," In WSDM '09, Spain, pp. 212-221, 2009.

[7] Zhou Bo, Cen Rongwei, Liu Yiqun, Zhang Min, Jin Yijiang, Ma Shaoping, "A Document Relevance Based Search Result Re-Ranking," Journal of Chinese Information Processing, Beijing, vol. 24(3), pp. 19-36, 2010.

[8] C. J. van Rijsbergen, "A new theoretical framework for information retrieval," In Proceedings of the 1986 International Conference on Research and Development in Information Retrieval (SIGIR '86), Italy, pp. 194-200, 1986,

[9] D.Y. Li and Y. Du, "Artificial Intelligence with Uncertainty," Beijing: National Defense Industry, 2005.

[10] D.Y. Li, X. Shi, and M.M. Gupta, "Soft Inference Mechanism Based on Cloud Models," LPSC 1996, Germany, pp. 38-62, 1996.

[11] Michael McCandless, Erik Hatcher, and Otis Gospodnetić, "Lucene in Action, Second Edition," New York:Manning Publications Co., 2010.

[12] Voorhees, E., Harman, D., eds., "TREC - Experiment and Evaluation in Information Retrieval," Masseachusettes: MIT, 2005.