Proceedings of KFIS Spring Conference 2006 Vol. 16, No. 1.

사용자 선호도 기반의 퍼지 랭킹모델에 관한 연구

A Study on Fuzzy Ranking Model based on User Preference

김대원

서울시 동작구 중앙대학교 컴퓨터공학부
E-mail: dwkim@cau.ac.kr

요 약

A great deal of research has been made to model the vagueness and uncertainty in information retrieval. One such research is fuzzy ranking models, which have been showing their superior performance in handling the uncertainty involved in the retrieval process. In this study we develop a new fuzzy ranking model based on the user preference. Through the experiments on the TREC-2 collection of Wall Street Journal documents, we show that the proposed method outperforms the conventional fuzzy ranking models.

Key Words : Fuzzy similarity measure, relevance ranking, information retrieval

1. Introduction

In recent years a great deal of research in information retrieval has aimed at modelling the vagueness and uncertainty which invariably characterize the management of information. Many approaches belonging to this class goes under the name of IR. The main levels of application of fuzzy set theory to IR have concerned the representation of documents and the query, the associative mechanism such as fuzzy thesauri, and the fuzzy ranking models.

Many fuzzy ranking models have been showing their superior performance in handling the uncertainty involved in the retrieval process. The ranking is achieved by calculating a similarity between two fuzzy sets, a document (D) and a query (Q). The best-known ranking models are the MMM, PAICE, and P-NORM. However, in spite that the user has an ability to reflect their preference for the information need in searching, these conventional fuzzy ranking models are limited to incorporate the user preference when calculating the rank of documents. Taking the problems of existing methods into account, in this study we develop a new fuzzy ranking model based on the user preference.

2. Fuzzy Ranking of Documents

Having established the index terms from given documents, a ranking model to calculate the similarity between a document and a query is required. We introduce a notion of user preference, which can provide a more clear ranking result. In this study, each document is represented and regarded as a fuzzy set:

$$D=(t_i, u_D(t_i))$$

where t_i is a term in the index set (I) and $u_D(t_i)$ represents a measure of degree to which the document D is characterized by each index term.

A variety of ranking measures between fuzzy sets have been proposed. However, most of these measures have no mechanism to reflect the user preference. Thus we propose a novel similarity measure incorporating the user preference or intention. Firstly, the similarity measure computes the degree of overlap between a document and a query. For each document

and query represented in fuzzy set, we obtain the overlap value between two fuzzy sets at each membership degree before computing the total overlap. The overlap function $f(u)$ at a membership degree (u) between D and Q is defined as:

$$f(u; D, Q) = \sum_{i=1}^{n} d(t_i, u; D, Q)$$

where d-func is 1.0 if $u_D(t_i), u_Q(t_i) > u$; 0.0 otherwise. It determines whether two sets are overlapped at the membership degree for index term. It returns an overlap value of 1.0 when the membership degrees of the two sets are both greater than u; otherwise, it returns 0.0. Based on this calculation, we derive the following definition of the similarity measure between a document and a query:

$$S(D, Q) = \sum_{u} f(u; D, Q) p(u)$$

where S-func is obtained by summing f-func over the whole range of membership degrees. A larger value of S means that two sets D and Q are more similar to each other, indicating that D is more relevant to Q.

Here, $p(u)$ is a preference function of membership, which is determined by the user. When two ranking results that have different fuzzy sets yield the same degree of similarity, the preference function is able to discern the two ranking results by focusing on the higher range of membership degrees. When users search the Web for information, they tend to focus on the document with the terms of highest matching. Thus the relevance of the highest-matched document plays an important role in user satisfaction. In such cases, $p(u)$ is given a value in the range [0.7, 1.0] when $u_D(t_i)$ is considered significant. Under this case, index terms with higher weights place greater emphasis on the calculation of the similarity between D and Q. Conversely, $p(u)$ is given a value in the range [0.0, 0.3] when $u_D(t_i)$ is considered insignificant.

3. Results

The search result obtained by the proposed method using the preference function with $u_p = 0.5$ was compared with the search result obtained using the PAICE, P-NORM and vector model.

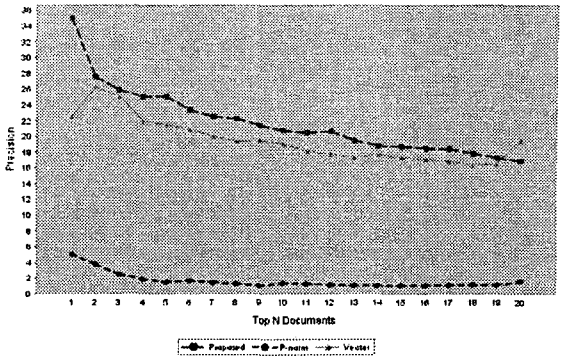

Figure 1 Comparison of Search Performance

The search performance of the PAICE model was much similar with that of P-NORM. The average precision of the PAICE and P-NORM models for 40 queries ranging from Top 1 to Top 20 are 1.48% and 1.62%, and the average recall of the PAICE and P-NORM are 0.47% and 0.59%, respectively. The P-NORM and vector models give average precision of 1.62% and 19.48% respectively. In contrast, the proposed model gives the higher average precision of 21.76%. Moreover, we see that the average precision of the proposed ranking model for the Top-ranked document (35.00%) is remarkably higher than those of the other two models.

References

[1] R.R. Yager and F.E.Petry, "A framework for linguistic relevance feedback in content-based image retrieval using fuzzy logic", Information Sciences (in press), April 2005.

[2] R. Baeza-Yates, et al., "Modern information retrieval", Addison-Wesley, 1999.

[3] J. Fan, W. Xie, "Some notes on similarity measure and proximity measure", Fuzzy Sets and Systems, vol. 101, pp. 403-412, 1999.

[4] W.J. Wang, "New similarity measures on fuzzy sets and on elements", Fuzzy Sets and Systems, vol. 85, pp. 305-309, 1997.