WELCOME TO THE GATHERING PLACE

Computer Networks and ISDN Systems 30 (1998) 621–623

COMPUTER NETWORKS and ISDN SYSTEMS

Short Paper

Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval

Chia-Hui Chang *,[1], Ching-Chi Hsu [1]

Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan

Abstract

Keyword based querying has been an immediate and efficient way to specify and retrieve related information that the user inquired. However, conventional document ranking based on an automatic assessment of document relevance to the query may not be the best approach when little information is given. In this poster, we propose an idea to integrate two existing techniques: query expansion and relevance feedback to achieve a concept-based information search for the Web.

Keywords: Query expansion; Relevance feedback; Document clustering

1. Introduction

The main issues of keyword-based query model lie in the difficulty of query formulation and the inherent word ambiguity in natural language. The problem is best illustrated through the scenario of information search on the Web, where the queries are usually of two words long and a large number of "hit" documents are returned to the user. Part of the reason comes from the inherent ambiguity of word in natural language. Another part is the difference of interpretation for a query. That is, given the same query expression by different users, the information inquired could range from various perspectives.

To overcome these problems, researchers have focused on automatic query expansion to help the user formulate what information is really needed. Another research topic is on relevance feedback from the user which gives the relevance of documents to clarify the ambiguity. In fact, these two techniques complement each other. However, the mechanisms of relevance feedback based on words or documents in the past research both have their own deficiencies. Word feedback has its upper bound performance in lexical-semantic expansion [3] and document feedback is sometimes too tiring for the users.

In this paper, we propose the conceptual feedback together with a joined mechanisms for query expansion. This is continuing research from our previous work based on clustering [2]. The idea is to organize the initial documents retrieved by the original query into conceptual groups such that the user could get a quick overview of what the query actually retrieves. Under this designing philosophy, we choose document clustering as our first step toward conceptual feedback. The hypothesis is that similar documents are more related to the same topic than documents that are less similar to each other.

Indeed, providing a concept-based information results as well as an interactive feedback has attracted many researchers in these two years. The dynamic browsing paradigm of Scatter/Gather that clusters documents into topical-coherent groups is applied in conventional similarity search to navigate the retrieved documents by Hearst and Pedersen [4]. On the other hand, static clustering of the database contents has also been exploited by An-

* Corresponding author.
[1] E-mail: {f2506005.CCHsu}@csie.ntu.edu.tw

0169-7552/98/$19.00
PII S0169-7552(98)00076-2

ick and Vaithyanathan [1]. They discuss the cognitive load required to assess the content of the clusters from the key terms and introduce natural language processing techniques to extract noun phrases for describing cluster contents.

In this paper, the target is how concept-based feedback can be achieved in a personalized Web information search assistant by integrating existing search engines and techniques of query expansion and relevance feedback. We focus on the mechanisms of keyword extraction for both cluster digesting and query expansion. Furthermore, the personalized Web search assistant can be enhanced by automatic discovery agents to search for more information based on the recorded query history.

2. Query expansion with relevance feedback

When a query is commenced, the query is resolved as a similarity search by a couple of search engines (including AltaVista, Excite, etc.). The Web assistant then collects the top n (60 for the prototype) documents returned and groups them into clusters. What differs here from the Scatter/Gather by [4] is that the feedback of relevant clusters and the documents are not only gathered for clustering again but the query is modified for better formulating the information needed.

As we mentioned earlier, relevance feedback has long been suggested as a solution for query modification. Rocchio describes an elegant approach and shows how the optimal vector space query can be derived using vector addition and subtraction given the relevant and non-relevant documents [6]. The probabilistic model proposed by Robertson and Sparck Jones shows how to adjust the individual term weight based on the distribution of the terms in relevant and non-relevant document set [5].

Now, given the cluster or concept as feedback unit, we would expect an approach to join these two models. Once each cluster has been digested as a document vector, the query can be modified by Rocchio's algorithm [6]. Thus, the problem becomes how keywords can be extracted as clusters digest and the weighting of terms in probabilistic model can be adjusted for this purpose.

The basic idea of feature selection for a concept is to highlight those words that have high frequency with respect to some contrast concept. Past research has applied Robertson and Sparck Jones' term weighting to query expansion [3], given the top 10–30 documents as relevant and all other documents in the corpus as non-relevant. For keyword extraction from a cluster, a simple application is to divide the initial documents into "belonging" and "not belonging" with respect to the cluster.

However, we find the direct application of the probabilistic weighing has some problems. Since the number of documents in a cluster is not large (about 10), the weighting is useless for words that appear only in the cluster. Hence, modification of the weighting is needed in this scenario. At the time of writing, the best performance is a rewrite form of "cue validity" joined with the majority principle for keyword selection.

The application of the Web search assistant is not only for online search help but also for constructing automatic discovery agents. Given all the queries the user has submitted, they can viewed as an interest profile for the user. Thus, a background domain manager is responsible for organizing the queries into interest groups. Based on the categorization, the Web assistant suggests related queries that the user has submitted and gives hints for query formulation whenever a query is commenced. On the other hand, we could apply the technique of genetic algorithm to spawn new queries. Taking each query as a chromosome with each word viewed as a gene, the words from the queries in the same interest domain are remixed to derive new queries and search results thereafter.

3. Summary

In this paper, we focus on integration of query expansion and relevance feedback. The employment of conceptual feedback with query expansion based on the two models is a new approach in information retrieval. By expanding a query, we could not only increase the number of relevant documents retrieved but also rank better the candidate documents. In the same time, the constructed summaries of the queries serve as the description profile of the user's information need. Thus, automatic information discovery can be carried out by generating new queries and filtering through the existed evidence.

There are a number of advantages in this Web assistant. First, it accelerates the browsing speed by dividing the initial results into similar document groups. Second, relevance feedback is constructed by dichotomy of documents and simply excludes the ones we do not like. Finally, the creation of a search agenda through genetic algorithm grants the property of autonomy and helps to execute the discovery automatically.

References

[1] Anick, P.G. and S. Vaithyanathan, Exploiting clustering and phrases for context-based information retrieval, in: *Proceedings of ACM SIGIR '97 International Conference on Re-*

search and Development in Information Retrieval, 1997, pp. 314–323.

[2] Chang, C.H. and Hsu, C.C., Multi-engine search tool with clustering, in: *Proc. of the 6th International WWW Conference*, Santa Clara, CA, USA, Apr. 7–11, 1997.

[3] Harman, D., Relevance feedback revisited, in: *Proceedings of ACM SIGIR 1992, International Conference on Research and Development in Information Retrieval*, 1992, pp. 1–10.

[4] Hearst, M.A. and J.O. Pedersen, Reexamining the cluster hypothesis: scatter/gather on retrieval results, in: *Proceedings of ACM SIGIR '96 International Conference on Research and Development in Information Retrieval*, 1996.

[5] Robertson, S.E. and Sparck Jones, K., Relevance weighting of search terms, *Journal of the American Society for Information Science*, 27(3): 129–146.

[6] Rocchio, J.J., Relevance feedback in information retrieval, in: G. Salton (Ed.), *The SMART Retrieval System*, Prentice-Hall, Inc., Englewood Cliffs, NJ, pp. 313–323.