Query and Content Suggestion Based On Latent Interest and Topic Class

Noriaki KAWAMAE
NTT Information Sharing Platform Laboratories
3-9-11, Midori-cho Musashino-shi
Tokyo 180-8585 JAPAN
kawamae. noriaki@lab.ntt.co.jp

Hideaki SUZUKI
NTT Information Sharing Platform Laboratories
3-9-11, Midori-cho Musashino-shi
Tokyo 180-8585 JAPAN
suzuki. hideaki@lab.ntt.co.jp

Osamu MIZUNO
NTT Information Sharing Platform Laboratories
3-9-11, Midori-cho Musashino-shi
Tokyo 180-8585 JAPAN
mizuno. osamu@lab.ntt.co.jp

ABSTRACT

To improve the process of user information retrieval, we propose the concept of a latent semantic map (LSM), along with a method of generating this map. The novel aspect of the LSM is that it can archive user models and latent semantic analysis on one map to support instantaneous information retrieval. With this characteristic, the LSM can improve search engines in terms of not only user support but also search results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Query formulation and Retrieval models

General Terms

Algorithms, Human factors

Keywords

Information retrieval, query suggestion, document suggestion, document categorization, latent semantic map

1. INTRODUCTION

Search engines are designed to support user information retrieval from Web pages. For this purpose, these engines provide maps from queries to contents. This type of mapping can be viewed as a directed divergent graph, where a set of queries maps to a set of contents. We can thus conclude that user information retrieval means the discovery of contents by following the divergent graph.

One goal of information retrieval is to enable users to efficiently find contents based on their interests beyond what they include in queries. The maps provided by existing search engines are often redundant or too complex. The corresponding directed divergent graphs are inappropriate for achieving the above goal, since they employ keyword matching technology to determine whether contents contain the query used by a user.

To improve user information retrieval, we instead deal with the relation between users and contents in terms of user interests and content topics. To relate user interests and content topics, we propose the latent semantic map (LSM), along with a method of generating this map. The novel aspect of the LSM is that it can archive user models and latent semantic analysis on one map to support instantaneous information retrieval.

By utilizing an LSM, a search engine can enable users to navigate according to the underlying user interests and content topics. This could be one approach to meet the ultimate goal of information retrieval, which is to enable a user to efficiently seek contents about topics of interest outside the scope of his or her current query.

2. PREVIOUS WORK

Among previous works related to our research, user query vectors have been used to measure query similarities [6] and applied to the Web [2] [3]. Another group [4] proposed a way to calculate the similarities of all pairs of objects. This technique has since been applied to find relevant website queries [1]. The problems with these works are their sensitivity with respect to observed data and lexical disagreement.

3. HOW TO MAKE AN LSM

3.1 Search Engine Based on LSM

The effect of utilizing an LSM is to enable a search engine to suggest better user queries and categorize search results as follows.

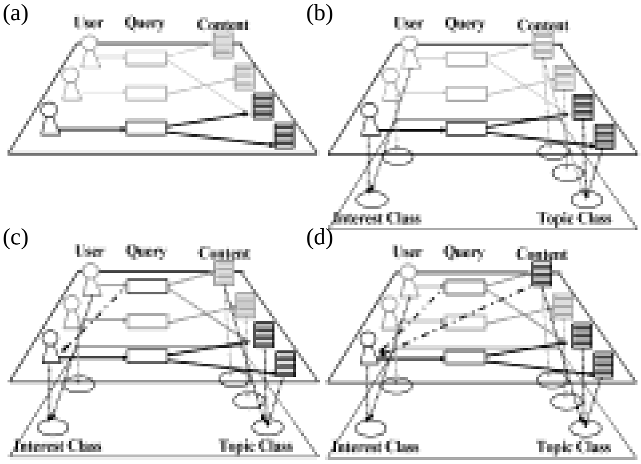

Figure1. Query and content suggestion based on LSM

In the LSM, we assume that all users, queries, and contents have probabilities of belonging to certain topic or interest classes. On this assumption, the LSM deals with user information retrieval in terms of a directed divergent graph, where the interest class maps to the content class. Figure 1(b) shows the discovery of other users belonging to interest classes similar to that of the active user and other contents belonging to topic classes similar to that of the

WWW 2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

content containing the query. Figure 1(c) shows how the search engine suggests queries to the active user based on similar users. Figure 1(d) shows content categorization based on similar topic classes. As for search results, the search engine can suggest contents belonging to the same topic classes as the content sought by a query.

3.2 Extraction of Interest and Topic Class

To generate the LSM, we must extract interest classes and topic classes. We assume that the relation between the content and the query used to seek it can be explained in terms of the relation between the topic classes to which the content belongs and the interest classes to which the query belongs. Based on this assumption, we can define the equation for the co-occurrence of content d_i sought by query q_h as follows.

$$P(q_h, d_j) = \sum_{k=1}^{O} P(q_h|t_k)P(t_k)P(d_j|t_k) \quad (1)$$

To calculate this equation and the membership probability of content d_j in each topic class, we take a Bayesian statistical approach[5]. From the calculated values of $P(d_j|t_k)$ and $P(t_k)$, we can define the membership probability of content d_j for each topic class by applying Bayes' rule as follows.

$$P(t_k|d_j) = \frac{P(d_j|t_k)P(t_k)}{\sum_{k=1}^{O} P(d_j|t_k)P(t_k)} \propto P(d_j|t_k)P(t_k) \quad (2)$$

To extract the interest class, we consider the relation between a user and interest class to be explainable in terms of the relation between the interest classes to which the user belongs and the interest classes to which a query belongs. In addition, we assume that each query can be explained in terms of a previously extracted topic classes. Given these assumptions, we can define the equation for the co-occurrence of user u_i selecting query q_g as follows.

$$\begin{aligned} P(q_g, u_i) &= \frac{1}{\left|d_{q_g}\right|} \sum_{t_k \in T} \left|d_{q_g \in t_k}\right| * P(t_k, u_i) \\ &= \frac{1}{\left|d_{q_g}\right|} \sum_{t_k \in T} \sum_{l=1}^{R} \left|d_{q_g \in t_k}\right| * P(t_k|c_l)P(c_l)P(u_i|c_l) \end{aligned} \quad (3)$$

where $\left|d_{q_g}\right|$ denotes the number of contents obtained by query q_g, and $\left|d_{q_g \in t_k}\right|$ denotes the number of d_{q_g} belonging to t_k. We calculate this equation in the same way as the above equation. After this calculation, we can define the membership probability of user u_i for each interest class c_l:

$$P(c_l|u_i) = \frac{P(u_i|c_l)P(c_l)}{\sum_{l=1}^{R} P(u_i|c_l)P(c_l)} \propto P(u_i|c_l)P(c_l) \quad (4)$$

3.3 Suggestion and Categorization Based on LSM

To measure the similarity of users or contents, we use an equation based on the relative entropy [5]. With this equation, we can calculate these similarities in terms of interest classes or topic classes, respectively, on the LSM. As a result, this similarity calculation enables an LSM-based search engine to suggest queries and contents according to the similarity of interests or topics.

4. Experiments

We have analyzed the feasibility of generating an LSM from a search log. The search log was generated by an Internet search engine over a 24-hour period on September 1, 1999. It consisted of 952,666 lines. We found 128,211 unique words and 120,677 unique URLs in the log. We show the results of analyzing the search log in Figures 2 and 3

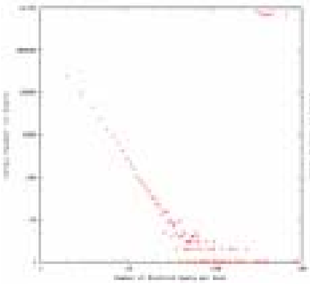

Figure 2. Log-log plot of the total number of users grouped by the number of distinct queries per user

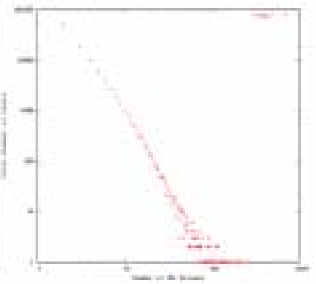

Figure 3. Log-log plot of the total number of users grouped by the number of URLs browsed per user

As for the relation between URLs and queries, we analyzed it and obtained similar results. These experiments results support the validity of generating an LSM from a search log.

5. Conclusion

The main contribution of this paper for information retrieval is to propose the LSM and a method of generating it. The effect of utilizing an LSM is to improve search engines in terms of not only user support but also search results. In our future work, we will extend the LSM by introducing link analysis.

6. REFERENCES

[1] **B. D. Davison, D. G. Deschenes and D. B. Lewanda. Finding Relevant Website Queries, In Proceeding of the 12th World Wide Web Conference (WWW12), pages 162-168, Budapest, Hungary, May 2003.**

[2] **L. Fitzpatrick and M. Dent. Automatic feedback using past queries: Social searching? In Proceedings of the 20th International ACM SIGIR, Philadelphia, PA, USA, July 1997.**

[3] **N. S. Glance. Community search assistant. In Artificial Intelligence for Web Search, pages 29--34. AAAI Press, July 2000.**

[4] **G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In Proceedings of the 8th ACM SIGKDD, Edmonton, Alberta, Canada, July 2002.**

[5] **N. Kawamae, H. Suzuki and O. Mizuno. Collaborative Filtering Via Bayesian Statistical Model, to appear.**

[6] **V. V. Raghavan and H. Sever. On the Reuse of Past Optimal Queries. In Proceedings of the 18th ACM I SIGIR'95, Seattle, WA, USA, July 1995.**