Similarity Spreading: A Unified Framework for Similarity Calculation of Interrelated Objects

Gui-Rong Xue[1] Hua-Jun Zeng[2] Zheng Chen[2] Wei-Ying Ma[2] Yong Yu[1]

[1]Computer Science and Engineering
Shanghai Jiao-Tong University
Shanghai 200030, P. R. China

grxue@sjtu.edu.cn, yyu@cs.sjtu.edu.cn

[2]Microsoft Research Asia
5F, Sigma Center, 49 Zhichun Road
Beijing 100080, P. R. China

{hjzeng, zhengc, wyma}@microsoft.com

ABSTRACT

In many Web search applications, similarities between objects of one type (say, queries) can be affected by the similarities between their interrelated objects of another type (say, Web pages), and vice versa. We propose a novel framework called similarity spreading to take account of the interrelationship and improve the similarity calculation. Experiment results show that the proposed framework can significantly improve the accuracy of the similarity measurement of the objects in a search engine.

Categories & Subject Descriptors:

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Search process, Clustering*; H.2.8 [**Database Management**]: Database Applications - *Data mining*; I.5.3 [**Pattern Recognition**]: Clustering - *Similarity measures*

General Terms: Algorithms, Experimentation.

Keywords: Similarity Spreading, Mutual Reinforcement, Interrelated

1. INTRODUCTION

Various applications in a search engine require a measurement of similarities between objects. One obvious example is to suggest related terms in interactive query expansion, which require the similarity calculation of query terms to other terms. Existing algorithms measure the similarity of objects based on the content features [1][2][4][5] or object interrelationship [3][6]. However, the influence of similarities between objects of one type on the similarities between objects of another type has not been taken into account.

As shown in Figure 1, the two types of objects are queries and Web pages. They are interrelated by click-through relationships. It is obvious that, when we compute the similarity of the any two queries, the similarity of the corresponding Web pages should be considered. Meanwhile, when computing the similarity of the Web page, the similarity of the corresponding queries should be considered.

The similarity of two objects of one type can propagate similarity of their respective interrelated objects of the other type through the mutual effect above. The propagated similarity also propagates similarity of the two original objects conversely.

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

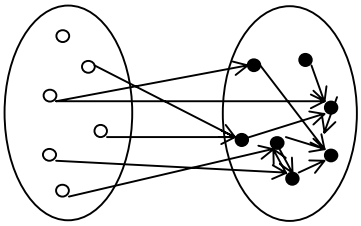

Figure 1 Objects of a search engine

In this paper, we propose a unified framework to calculate the similarities over the interrelated objects in a search engine. Under this framework, the intra-type feature, the inter-type feature and mutual effect of two types of similarity are considered. Besides, such process is iterative until the similarity converges to a stable state.

2. ALGORITHM

We model objects on the Web and their relationships as a directed graph $G=(V, E)$, where nodes in V represent Web objects and edges E represent relationships between Web objects. In this paper, we consider V as consisting of two subsets $Q=\{q^1, q^2, \ldots, q^m\}$ and $P=\{p^1, p^2, \ldots, p^n\}$, where Q represents the query objects and P represents the Web page objects.

Three kinds of relationships among the Web objects are studied in this paper, including Web page in-link relationship (*IL*), Web page out-link relationship (*OL*), and query-page click-through relationship (*CT*). For any Web object v in G, the set of the adjacent objects which have any of the three relationships with the object v, is denoted as $M_R(v)$ (R represents the corresponding relationship). For example, $M_{IL}(v)$ denotes the set of Web pages that contains hyperlinks leading to Web page v. Individual objects in the set $M_R(v)$ are denoted as $M_R^i(v)$, where $1 \le i \le |M_R(v)|$.

Furthermore, we use S to denote a similarity matrix for objects; thus, $S[a,b]$ is the similarity between objects a and b.

In this section, we propose the a unified framework of similarity spreading algorithm, to calculate the similarity of different Web objects by utilizing different kinds of object relationships in a mutually reinforcing manner.

The query similarities can be calculated as:

$$S_{QC}[q^s, q^t] = \frac{Keyword(q^s) \cap Keyword(q^t)}{Keyword(q^s) \cup Keyword(q^t)}$$

$$S_{QCT}[q^s, q^t] = \frac{C}{|M_{CT}(q^s)||M_{CT}(q^t)|} \sum_{i=1}^{|M_{CT}(q^s)||M_{CT}(q^t)|} \sum_{j=1} S_P[M_{CT}^i(q^s), M_{CT}^j(q^t)] \quad (1)$$

$$S_Q[q^s, q^t] = \alpha S_{QC}[q^s, q^t] + \beta S_{QCT}[q^s, q^t]$$

where q^s and q^t are two queries; $\alpha+\beta=1$; S_Q and S_P are the similarity matrices of the queries and the Web pages, respectively. As shown in Eq. 1, the inter-type similarity of the queries is affected by the similarity of Web pages via the click-through relationship (M_{CT}). The similarity matrix (S_Q), for the current iterative result, is a linear combination of intra- and inter-type query similarity.

Similarity of Web pages can be defined in a very similar fashion as:

$$S_{PC}[p^s,p^t]=\frac{Keyword(p^s)\cap Keyword(p^t)}{Keyword(p^s)\cup Keyword(p^t)}$$

$$S_{OL}[p^s,p^t]=\frac{C_{PC}}{|M_{OL}(p^s)||M_{OL}(p^t)|}\sum_{i=1}^{|M_{OL}(p^s)||M_{OL}(p^t)|}\sum_{j=1}^{}S_P[M_{OL}^i(p^s),M_{OL}^j(p^t)] \quad (2)$$

$$S_{IL}[p^s,p^t]=\frac{C_{PR}}{|M_{IL}(p^s)||M_{IL}(p^t)|}\sum_{i=1}^{|M_{IL}(p^s)||M_{IL}(p^t)|}\sum_{j=1}^{}S_P[M_{IL}^i(p^s),M_{IL}^j(p^t)]$$

$$S_{PCT}[p^s,p^t]=\frac{C_{CT}}{|M_{CT}(p^s)||M_{CT}(p^t)|}\sum_{i=1}^{|M_{CT}(p^s)||M_{CT}(p^t)|}\sum_{j=1}^{}S_Q[M_{CT}^i(p^s),M_{CT}^j(p^t)]$$

$$S_P[p^s,p^t]=\alpha'(\omega^1 S_{OL}[p^s,p^t]+\omega^2 S_{IL}[p^s,p^t])+\beta' S_{PCT}[p^s,p^t]+\gamma' S_{PC}[p^s,p^t]$$

where p^s and p^t are two Web pages, and $\alpha'+\beta'+\gamma'=1$. The similarity of the queries is spread to the similarity of Web pages through the click-through relationship (M_{CT}), and the similarity matrix (S_P), for the current iterative result, is a linear combination of intra- and inter-type Web page similarity. The similarity calculation can be continued iteratively until values converge.

3. EXPERIMENTS

We used a trace of query sessions from the MSN search engine that was collected in August, 2003. It contains almost 1.2 million requests recorded over a period of three hours. We define a quantitative measure Precision to evaluate the performance of the algorithm. Given an object, the top 10 of similar objects as M is present to the user. Then ten volunteers were asked to identify which object is similar to the given object and the set N is the voting results we collected. The precision of the algorithm is defined as $|N|/|M|$. We do two types of experiments: one is to measure the performance of finding the similar queries and the other is to measure the performance of finding the related Web pages.

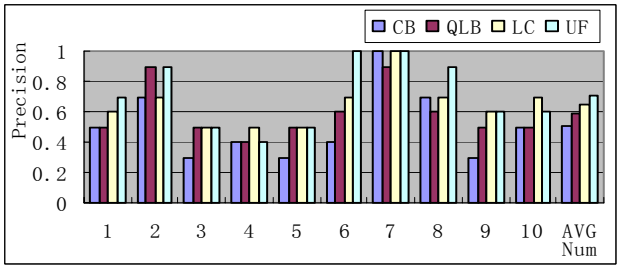

Figure 2 Precision of the similarity between the queries

We compare several methods: our unified framework (UF), content based method (CB) which is only based on the keywords, hyperlink based method (HB) which is only based on the hyperlink relationship, and query log based method (QLB) which is only based on the clickthrough data. For query similarity calculation, linear combination method (LC) is linear combination of CB and QLB, while for Web page similarity calculation, linear combination method (LC) is linear combination of HB and QLB.

The comparison of 4 algorithms is shown in Figure 2. The right-most label "AVG" stands for the average value for the 10 queries. Our algorithm can improve the accuracy of similarity among the queries.

The ten volunteers were also asked to evaluate the precision of the similarity calculation for the random selected 10 Web pages. The comparison of 4 algorithms is shown in Figure 3.

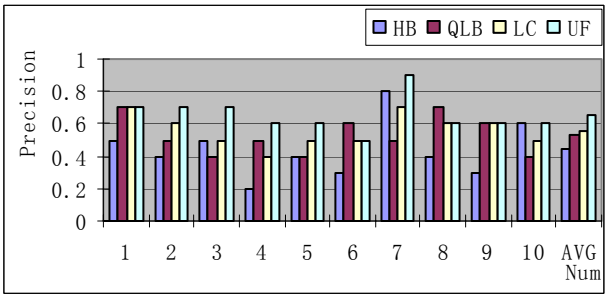

Figure 3 Precision of the similarity between the Web pages

4. REFERENCES

[1] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," In In Proceedings of the AAAI 2000 Workshop on Articial Intelligence for Web Search, pp. 58--64, Austin, Texas, July 2000.

[2] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G.Karypis, V. Kumar, B. Mobasher, and J. Moore, "Partitioning-based clustering for web document categorization," Decision Support Systems 27:329-341, 1999.

[3] J. D. Wang, H. J. Zeng, Z. Chen, H. J. Lu, L. Tao, and W.-Y Ma. ReCoM: reinforcement clustering of multi-type interrelated data objects. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 274-281, Toronto, CA, July 2003.

[4] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration," In Proceedings of SIGIR '98, pp. 46--53, 1998.

[5] R. Baeza-Yates and B.Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999.

[6] S. Chakrabarti, B.E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. Computer, 32(8), 1999.