A novel Web usage mining approach for search engines

Dell Zhang [a,b,*], Yisheng Dong [c]

[a] *Department of Computer Science, National University of Singapore, S15-05-24, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *Singapore-MIT Alliance, E4-04-10, 4 Engineering Drive 3, Singapore 117576, Singapore*
[c] *Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China*

Abstract

Web usage mining can be very useful to search engines. This paper proposes a novel effective approach to exploit the relationships among users, queries and resources based on the search engine's log. How this method can be applied is illustrated by a Chinese image search engine. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Web information retrieval; Multimedia retrieval; Data mining

1. Introduction

The World Wide Web contains an enormous amount of information, and it is becoming more and more complex while its size continues to grow at a remarkable rate. So it can be exceedingly difficult for users to locate resources that are both relevant to their information needs and high in quality.

Today's search engines usually help users locate information based on the textual similarity of a query and potential documents. Experienced users can make effective use of these search engines by searching for tightly constrained keywords and phrases. However, such search engines are far from satisfactory to common users. In particular, a search engine will typically return thousands of Web resource pointers for a general query, while a user will only be willing to look at an extremely small part of the search results. Moreover, keyword matching techniques may not work for various types of multimedia Web resources such as images, MP3 songs, video clips, etc., Here our goal is to explore other ways to decide if resources are "of value" to the user.

This paper is organized as follows. In Section 2, we will provide an overview of our dominant idea. In Section 3, we will present the details of our work. Section 4 will describe the *eeFind* Chinese image search engine, where the proposed approach can be applied. And Section 5 will discuss the relationship between our method and other related works. Our concluding remarks will be given in Section 6.

2. Model

In order to distill a large search topic on the Web down to a size that will make sense to a

---

* Corresponding author. Address: Department of Computer Science, National University of Singapore, S15-05-24, 3 Science Drive 2, Singapore 117543, Singapore.

E-mail addresses: dell.z@ieee.org (D. Zhang), ysdong@seu.edu.cn (Y. Dong).

human user, we wish not only to locate a set of relevant resources, but rather the relevant resources of the highest quality. The concept "quality" here means both "authority" and "freshness" on the search topic. This notion of quality adds a crucial second dimension to the notion of relevance.

As we think about the types of resources we hope to discover, and the fact that we wish to do so automatically, we are quickly led to some difficult problems. It is not sufficient to just apply purely text-based methods to rank the quality of resources, because high quality resources are often not self-descriptive.

On the other hand, the log of a search engine contains plenty of latent human annotation that can be extremely valuable for automatically inferring notions of quality. Just like the case in a traditional library, the documents being frequently accessed recently should have high quality. In fact, accessing a Web resource can represent an implicit type of "recommendation" of the resource. By mining the collective judgement contained in the set of such recommendations, we can obtain a richer understanding of both relevance and quality of the resources on the Web.

If the high quality resources on the Web do not explicitly describe themselves as such, and they do not link to one another (e.g., multimedia resources), how can we determine that they are indeed the high quality resources for this topic? We could say that they are valuable to this topic because a large number of experienced users have frequently accessed them in recent time. For example, if a paper about "data mining" has been frequently accessed by many database experts (such as Jim Gray, Jeff Ullman, etc.), it should be a high quality paper on this topic. Furthermore, if we take query terms into account, the recommendation can work in a recurring way: a user is "good" if he/she issues many "good" queries, while a query is "good" if it can retrieve many "good" resources, while a resource is "good" if it is accessed by many "good" users. This circular reinforcing relationship among users, queries and resources (Fig. 1) will serve as a central theme in our exploration of Web usage mining.
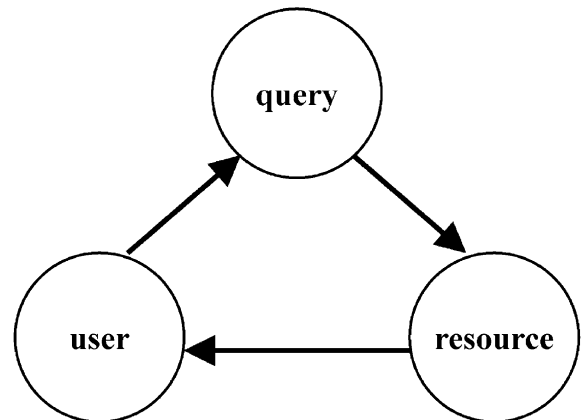

Fig. 1. The circular reinforcing relationship among users, queries and resources.

3. Method

We now describe the concrete algorithm, which can generate a list of relevant resources for a search topic. Beginning with a initial query q^*, specified by one or more keywords, our algorithm applies four main steps. First, the algorithm looks for the set of all users who have issued the query q^* recently based on the search engine's log. Second, the set of all queries these users have issued recently is constructed, also based on the search engine log. These queries are believed to be relevant to the initial query q^* in different degree so they are worthy of further evaluation. Silverstein et al. claimed that queries for a single information need come clustered in time, and there is a gap before the user returns to the search engine [20]. Then the set of all resources relevant to these queries can be constructed through traditional keyword-based IR techniques. The last step is to compute the numerical quality estimates of the found resources by an iterative procedure. The Web resources with the highest quality weights are returned in order for the search topic.

Assume that the initial query term is q^*. The good resources can be extracted from the base set by giving a concrete numerical interpretation to the intuitive notions developed in the previous section. Define T as the time-window's width, e.g., a week, then events occurring after the time point

$(t_{now} - T)$ are taken as "recent" events, where t_{now} denotes the current time. Let U represent the set of all users who have issued the query q^* recently, $|U| = m$. Let Q represent the set of all queries the users in U have issued recently, $|Q| = n$. We obtain the expanded set of relevant resource candidates R, $|R| = s$. We associate a non-negative experience weight u_i with each user u_i in U, a non-negative utility weight q_j with each query q_j in Q, and a non-negative quality weight r_k with each resource r_k in R. That is to say, a user with a larger weight u_i will be viewed a "better" (more experienced) user, a query with a larger weight q_j will be viewed a "better" (more effective) query, and a resource with a larger weight r_k will be viewed as a "better" (more valuable) resource. We will only be interested in the relative values of these weights, not their actual magnitudes; so they are normalized to satisfy

$$\sum_U u_i = 1, \quad \sum_Q q_j = 1, \text{ and } \sum_R r_k = 1.$$

The actual choice of normalization does not affect the results; we maintain the invariant that all weights sum to 1. We will see later, however, that the final results are essentially unaffected by this initialization. So we do not impose any a priori estimates, and set all u, q, r values to a uniform constant initially.

The weights of users, queries and resources can be calculated as follows:

A user is "good" if he/she issues many "good" queries, i.e.,

$$u_i \leftarrow \sum_{j=1}^{n} a_{ij} \cdot q_j, \quad a_{ij} = \text{num}(u_i, q_j).$$

Here $\text{num}(u_i, q_j)$ represents how many times the user u_i issued the query q_j recently.

A query is "good" if it can retrieve many "good" resources, i.e.,

$$q_j \leftarrow \sum_{k=1}^{s} b_{jk} \cdot r_k. \ b_{jk} = \text{sim}(q_j, r_k).$$

Here $\text{sim}(q_j, r_k)$ represents the similarity between the query q_j and the resource r_k.

A resource is "good" if it is accessed by many "good" users, i.e.,

$$r_k \leftarrow \sum_{i=1}^{m} c_{ki} \cdot u_i, \ c_{ki} = \text{hit}_{q^*}(r_k, u_i) + \alpha \cdot \text{hit}_{q'}(r_k, u_i).$$

Here $\text{hit}_{q^*}(r_k, u_i)$ represents how many times the user u_i accessed the pointer of r_k in the search results of the query q^*, $\text{hit}_{q'}(r_k, u_i)$ represents how many times the user u_i accessed the pointer of r_k in the search results of the queries other than q^*, and α $(0 < \alpha < 1)$ is a constant decreasing factor for non-initial queries. Thus we get three equations in a strictly recurring fashion.

There is a more compact way to write these updates, and it turns out to shed more light on what is going on mathematically. Let us write the set of all u_i-values as a m-dimensional vector $\mathbf{u} = (u_1, u_2, \ldots, u_m)^T$, and similarly define $\mathbf{q} = (q_1, q_2, \ldots, q_n)^T$, $\mathbf{r} = (r_1, r_2, \ldots, r_n)^T$. Let us also define the weight matrix $\mathbf{A} = (a_{ij})_{m \times n}$, $\mathbf{B} = (b_{jk})_{n \times s}$, and $\mathbf{C} = (c_{ki})_{s \times m}$. Then our update rules can be rewritten as

$$\mathbf{u} \leftarrow \mathbf{Aq},$$

$$\mathbf{q} \leftarrow \mathbf{Br},$$

$$\mathbf{r} \leftarrow \mathbf{Cu}.$$

Unfolding these three rules further, we have

$$\mathbf{u} \leftarrow \mathbf{Aq} = \mathbf{A}(\mathbf{Br}) = \mathbf{A}(\mathbf{B}(\mathbf{Cu})) = (\mathbf{ABC})\mathbf{u},$$

$$\mathbf{q} \leftarrow \mathbf{Br} = \mathbf{B}(\mathbf{Cu}) = \mathbf{B}(\mathbf{C}(\mathbf{Aq})) = (\mathbf{BCA})\mathbf{q},$$

$$\mathbf{r} \leftarrow \mathbf{Cu} = \mathbf{C}(\mathbf{Ar}) = \mathbf{C}(\mathbf{A}(\mathbf{Br})) = (\mathbf{CAB})\mathbf{r}.$$

Thus we get three iterative equations.

To get the stable vector $\mathbf{u}$, we multiply the initial vector by $\mathbf{ABC}$ repeatedly. The computation of $\mathbf{q}$ and $\mathbf{r}$ is similar.

Theorem 3.1 (Perron–Frobenius Theorem). *If a n-dimensional matrix* $\mathbf{M}$ *is positive or non-negative irreducible, then*

- *The spectrum radius of* $\mathbf{M}$, ρ, *is also a latent root of* $\mathbf{M}$.
- *There is a positive eigenvector of* $\mathbf{M}$ *corresponding to* ρ.
- *The eigenfunction of* $\mathbf{M}$ *has a single root* ρ, *i.e.,* $\text{mult}\rho(\mathbf{M}) = 1$.

The above well-known *Perron–Frobenius Theorem* in linear algebra [11], tells us that this sequence of iterates, when normalized, converges to the principal eigenvector of **M**. That means,

- **u** converges to the principal eigenvector of the $m \times m$ matrix **ABC**,
- **q** converges to the principal eigenvector of the $n \times n$ matrix **BCA**,
- **r** converges to the principal eigenvector of the $s \times s$ matrix **CAB**,

To satisfy the condition of Theorem 3.1, all the zero elements in matrix **A**, **B** and **C** can be replaced by ε which is a very tiny real number, then **A**, **B** and **C** are all positive matrixes, consequently their product **ABC**, **BCA** and **CAB** are also positive matrixes.

In fact, such power iteration will converge to the principal eigenvector for any "non-degenerate" choice of initial vector. In our case, for example, for any vector all of whose entries are positive. This says that the quality (experience) weights we compute are truly an intrinsic feature of the resources (users), not an artifact of our choice of initial weights or the tuning of arbitrary parameters. Intuitively, the resources with large weights are those being frequently accessed recently. In our experience, the relative values of the large components in these vectors typically resolve themselves within a few power iterations, obviating the need for more sophisticated eigenvector computation methods. Finally, the output of the algorithm for the given initial query is a short list consisting of the resources with the highest quality weights.

This method is denoted as MASEL (matrix analysis on search engine log). Our preliminary experiments suggest that MASEL provide surprisingly good search results for a wide range of queries.

4. Application

In the *eeFind* Chinese image search engine (http://www.eefind.com/), a simplified version of the MASEL algorithm can be used to achieve better ranking and query expansion effects.

A fragment from the log of *eeFind* Chinese image search engine is illustrated in Fig. 2.

There are two kinds of entries in the image search engine log:

```
202.108.50.66     973233934 国旗
202.108.50.66     973233967 国旗    http://www.qingdao.org.cn/chinese/lishi/157.jpg
202.108.50.66     973237215 模特
202.108.125.66    973500340 微软
202.108.125.66    973500621 微软    http://go.163.com/~happy100/cartoon/politic/zzmh19.jpg
202.178.245.230   973745516 孔子
202.178.245.230   973753709 孔子    http://www.njnet.edu.cn/www/eastnor/shandong/kz4.gif
210.74.165.87     973760844 猫
210.74.165.87     973760853 猫      http://www.hello.com.tw/~w372/cat-10.jpg
210.74.165.87     973760932 猫      http://lynx.uio.no/jon/gif/animals/cats/serval1.jpg
166.111.214.28    973761603 国旗
166.111.214.28    973761750 国旗    http://www.crwflags.com/fotw/images/cn.gif
166.111.214.28    973762725 国旗    http://sq.k12.com.cn/~xld/scjp/sc/flag/020.GIF
210.131.20.2      973820859 长城
......
```

Fig. 2. A sample snippet from the log of *eeFind*. Chinese image search engine.

1. *host timestamp query*
2. *host timestamp query url*

For example:

210.74.165.87   973760844 •

means the remote computer 210.74.165.87 issued a query "•" in the timestamp 973760844.

210.74.165.87   973760853 •
http://www.hello.com.tw/~w372/cat-10.jpg

means the remote computer 210.74.165.87 accessed the image http://www.hello.com.tw/~w372/cat-10.jpg while searching for "•" in the timestamp 973760844.

Given an initial query, the matrixes **A**, **B**, and **C** in the MASEL algorithm can be constructed based on the above style log. Then the algorithm can generate a list of pointers to relevant image resources, ranked by their weights. Fig. 3 contains two screen snapshots while looking for "car" in our search engine: the left one is the original result page, the right one is the result page after several days and utilizing MASEL algorithm to adjust the rank of each image resource.

Other than better ranking, MASEL has a side effect that some implicit query expansion occurs. For example, the query "car" can return some images labeled with "BMW", "Porsche" or "Rolls Royse" because they are often queried by the uses with similar interests recently. This feature is especially helpful to improve recall in multimedia retrieval, since the labels automatically extracted from Web pages is usually very short.

To provide a better insight how MASEL works in practice, here we present a small sample. Given the query "car", three matrices **A**, **B** and **C** will be constructed (as shown in Fig. 4). The parameter α is set to $1/10$. After several iterations, the resource weight vector **r** comes to be $\mathbf{r} = (0.96, 0.28, 0.03, 0.01)^{\mathrm{T}}$, it is just the principal eigenvector of the matrix $\mathbf{M} = \mathbf{CBA}$.

Fig. 3. Two screen snapshots while looking for "car" in our search engine, with or without MASEL.

	car	Auto	bus
Tom	1	1	1
Jack	1	0	1
Rose	2	1	0

matrix **A**

	img1	img2	img3	img4
car	1	1	0	0
auto	0	0	1	0
bus	0	0	0	1

matrix **B**

	Tom	Jack	Rose
img1	2	1	2
img2	1	1	0
img3	0	0	0.1
img4	0	0.1	0

matrix **C**

Fig. 4. A illustration of the constructed matrices for MASEL.

5. Related work

Today, many new style search engines have emerged. Experienced researchers may discover that our approach brings ideas from two significant search engines: DirectHit (http://www.direct-hit.com/) and Google (http://www.google.com/). DirectHit estimates the popularity of a Web page by counting how many times its pointer is clicked. But DirectHit does not distinguish between different user groups, i.e., "click" of an experienced user is equally weighted as a beginner. Google estimates the authority of a Web page through linkage analysis, the hyperlinks from more authoritative Web pages are considered more important. But in some situations, the linkage information cannot represent the authority accurately, e.g., two websites reference each other frequently will both get high scores in Google's algorithm. Moreover, for multimedia Web resources, there is no linkage information directly available. It's obvious that the relevance and quality of a Web page cannot represent the relevance and quality of all images inside it.

The MASEL algorithm has intellectual antecedents in the HITS algorithm, which is proposed in the CLEVER search engine [5,6] and pioneers in Web structure mining. Such techniques have been utilized for the discovery of interesting images on the Web [17]. There are some similarities between MASEL and HITS, but MASEL is mining on a search engine's log but not Web structure, and MASEL takes queries into consideration. The name HITS seems to be more appropriate for our algorithm since we take advantage of users' "hits". It is worth drawing some contrasts between MASEL and HITS, and we believe that MASEL and HITS can complement each other.

There are many other mechanisms to do query expansion, e.g., using meta-data and conceptual hierarchies. However, such approaches are usually labor-intensive while MASEL is totally automatic.

The scientific literature contains few analyses of data collected from large-scale commercial search engines. Jansen et al. reported their analysis on 51,000 query logs [14]. Silverstein et al. reported statistics accumulated from a billion-entry set of user queries to Altavista [20]. Beeferman proposed an agglomerative clustering method on "click-through" records in Lycos's log [1].

Researchers in the multimedia area have built some content-based image retrieval system using "relevance feedback" techniques to improve the quality of search results [8,19]. These systems usually need more than five feedback iterations to reach a satisfactory accuracy. One inherent deficiency with these systems is that they do not take advantage of the knowledge from previous users' relevance feedback. However, our approach can make use of all users' feedback implicitly through log analysis.

Collaborative filtering is a process by which information on the preference and actions of a group of users is tracked by a system which then, based on the patterns it observes, tries to make useful recommendations to individual users [12,13,18,22]. Based on this idea, we would like to capture access information that may tell us within limits of privacy concerns which user groups are accessing what data, and how frequently. MASEL has

learned from collaborative filtering to share searching expertise among users.

For further reading, see [2–4,7,9,10,15,16,21,23].

6. Conclusion

Inspired by the apparent nature of browsing activities in traditional libraries, we try to improve search engines through Web usage mining. This paper proposed the MASEL algorithm, which can exploit the relationships among users, queries and resources based on the search engine's log. Illustrated by a Chinese image search engine, the proposed approach reveals its power to achieve better ranking and query expansion effects.

Certainly, more experiments need to be conducted to verify the effectiveness and evaluate the performance of our method. Also, we are considering how to make this approach more scalable, and more robust to malicious attacks.

References

[1] D. Beeferman, A. Berger, Agglomerative clustering of a search engine query log, in: Proceedings of ACM KDD 2000, Boston, MA, USA.

[2] S. Brin, L. Page, The anatomy of a large scale hypertextual Web search engine, in: Proceedings of WWW7, Brisbane, Australia, April, 1998.

[3] S. Chakrabarti, B. Dom, M. van den Berg, Focused crawling: a new approach for topic-specific resource discovery, Proceedings of the 8th World Wide Web Conference, 1999.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, Automatic resource compilation by analyzing hyperlink structure and associated text, in: Proceedings of the 7th World Wide Web Conference, Elsevier, Amsterdam, 1998.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Mining the link structure of the World Wide Web, IEEE Computer 32 (8) (1999) 60–66.

[6] S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Experiments in topic distillation, in: ACM SIGIR Workshop on Hypertext Information Retrieval, Melbourne, Australia, 1998.

[7] S. Chakrabarti, B. Dom, P. Indyk, Enhanced hypertext classification using hyperlinks, ACM SIGMOD Conference on Management of Data, 1998.

[8] I.J. Cox, M.L. Miller, T.P. Minka, T.V. Papathornas, P.N. Yianilos, The Bayesian image retrieval system, Pic Hunter: theory, implementation, and psychophysical experiments, IEEE Trans. Image Processing 9 (1) (2000) 20–37.

[9] L. Egghe, R. Rousseau, Introduction to Informetrics, Elsevier, Amsterdam, 1990.

[10] D. Florescu, A. Levy, A. Mendelzon, Database techniques for the World Wide Web: a survey, SIGMOD Record 27 (3) (1998) 59–74.

[11] G. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, 1989.

[12] D. Goldberg, D. Nichols, B.M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, Communications of the ACM 35 (12) (1992) 51–60.

[13] W. Hill, L. Stead, M. Rosenstein, G. Furnas, Re-commending and evaluating choices in a virtual community of use, Proceedings of ACM CHI, 1995, pp. 194–201.

[14] B. Jansen, A. Spink, J. Bateman, T. Saracevic, Real-life information retrieval: a study of user queries on the Web, SIGIR Forum 32 (1) (1998) 5–17.

[15] J.M. Kleinberg, Authoritative sources in a hyperlinked environment. in: Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998 and IBM Research Report RJ 10076, May 1997.

[16] S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling emerging cyber-communities automatically, in: Proceedings of the World Wide Web Conference, 1999.

[17] R. Lempel, A. Soffer, PicASHOW: Pictorial AUthority Search by Hyperlinks On the Web, in: Proc. 10th World Wide Web Conference, Hong Kong, May 2001.

[18] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Center for Coordination Science, MIT Sloan School of Management Report WP #3666-94, 1994.

[19] Y. Rui, T.S. Huang, S. Mehrotra, Content-based image retrieval with relevance feedback in MARS, in: Proceedings of IEEE International Conference on Image Processing, 1997.

[20] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a very large AltaVista query log, DEC SRC Technical Note 1998-014, 1998.

[21] G. Salton, M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[22] H.R. Varian, P. Resnick (Eds.), CACM Special issue on recommender systems, Comm. ACM 40(3) (1997).

[23] D. Zhang, Y. Dong, An efficient algorithm to rank Web resources, Proceedings of the 9th World Wide Web Conference, 1999.

Dell Zhang is a Research Fellow in the National University of Singapore under the Singapore-MIT Alliance (SMA). He has received his B.Eng. and Ph.D. in Computer Science from the Southeast University, Nanjing, China. He is currently focusing on data mining and information retrieval.

Yisheng Dong is a Professor and also the Head of the Department of Computer Science and Engineering in Southeast University, Nanjing, China. He graduated in 1965, from Nanjing Institute of Technology, Nanjing, China. His research interests include database, software engineering and information system.