Using Web Archive for Improving Search Engine Results

Adam Jatowt[1], Yukiko Kawai[1] and Katsumi Tanaka[1,2]

[1] National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, 619-0289
Kyoto, Japan
{adam, yukiko}@nict.go.jp

[2] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, 606-8501
Kyoto, Japan
ktanaka@i.kyoto-u.ac.jp

Abstract. Search engines affect page popularity by making it difficult for currently unpopular pages to reach the top ranks in the search results. This is because people tend to visit and create links to the top-ranked pages. We have addressed this problem by analyzing the previous content of web pages. Our approach is based on the observation that the quality of this content greatly affects link accumulation and hence the final rank of the page. We propose detecting the content that has the greatest impact on the link accumulation process of top-ranked pages and using it for detecting high quality but unpopular web pages. Such pages would have higher ranks assigned.

1 Introduction

Search engines are the main gateway to the Web for many users seeking information. Most search engines use link structure analysis to evaluate the quality and ranks of web pages. The most popular ranking approach is derived from the computation of PageRank metric [5], which, in iterative process, defines the quality of pages based on the macro-scale link structure of the Web. The basic PageRank formula is shown in Equation 1; $PR[p_j]$ is the PageRank of page p_j, d is a damping factor, c_i is the number of links from some page p_i to the target page p_j, and m is the number of all pages containing links to page p_j. Basically, the larger the number of inbound links to the page and the higher their Pagerank values, the higher is the assigned PageRank value of the page.

$$PR[p_j] = d + (1-d) * \sum_{i=1}^{m} PR[p_i] / c_i \quad (1)$$

However, the search engines can affect user behavior and page popularity [3,4]. Since there are huge amounts of resources on the Web, relatively few of them are

likely to be found and visited by users seeking information. Pages must therefore compete for user attention. Successful pages are the ones that are frequently visited by many users. Unfortunately, the competition between web pages for user attention is not completely fair. Since users tend to click only on the top-ranked pages, lower ranked pages have difficulty gaining popularity even if they have high-quality content. Conversely, the top-ranked pages maintain their high ranks since they are frequently accessed by many users and consequently receive more new inbound links than pages with lower rankings. The accelerating growth of the Web is resulting in even more pages competing for user attention, making the negative-bias problem more serious. Consequently, we are witnessing "rich-get-richer phenomenon" [3,4].

The huge popularity of search engines among Web users and the impact of these engines on page popularity demands countermeasures that will enable pages to compete more fairly for user attention. One proposed solution is temporal link analysis. Temporal link analysis [1,2,4] focuses on analyzing link evolution and on using time-related information regarding links for estimating page quality or improving page rankings. Links are associated with time, and this factor is considered when analyzing page quality. We also use the notion of previous page quality, however unlike in previous approaches, we analyze past page content for use in evaluating the usefulness of those pages. Our assumption is that, for a certain query, there is "topic representation knowledge" that describes the topic of the query in the best way or is the most related to this query. If we can identify this knowledge, we can rank pages based on how well they cover it. However, this knowledge is distributed not only spatially among many existing web pages but also temporally among previous versions of pages. We use historical data for top-ranked pages to identify this "topic representation knowledge" by detecting the content that contributed most heavily to link accumulation of web pages. The previous content of a page and the quality of that content have a major effect on the current number of inbound links to that page. We demonstrate it by content-based link accumulation model of pages.

In the next section we introduce content-based link accumulation model of a web page. Section 3 discusses our approach to improving the search results by using past versions of pages extracted from web archives. It also shows some experimental results. Finally, we conclude with a brief summary.

2 Content-based Link Accumulation Model

Consider a web page that is ranked highly by a popular search engine for a certain query. Since it has a high ranking, it should have relatively many in-bound links, probably coming from high quality pages. Let t_0 denote the time at which the page was created, meaning that its age is t_n - t_0, where t_n is now. Let us imagine that at t_0, the page had few inbound links, denoted as $L(t_0)$, and a small stream of visitors, $U(t_0)$. Since its inception, the page has been accumulating links and increasing in popularity, reflected by a growing number of visitors. It has now reached the top rank, gained many links, $L(t_n)$, and enlarged its user stream $U(t_n)$. We assume that the query popularity, hence the number of web pages found in response to the query, stays the same. The number of visitors at some point of time t_i, that is, the size of the user stream, is

the number of visitors due to browsing plus the number due to finding the page by using a search engine:

$$U(t_i) = \eta * U^B(t_i) + \mu * U^S(t_i). \quad (2)$$

Thus, it is the number of users coming from other pages by following links, $U^B(t_i)$, plus the number coming from search engine results, $U^S(t_i)$. We can represent the first number as the sum of the PageRank values of pages linking to the target page, hence in approximation, the actual PageRank value of the page. Similarly, we can approximate the number of users visiting the page from search engines as its PageRank value. Hence, the user stream at a particular time depends on the PageRank value of the page at that time, $PR(t_i)$:

$$U(t_i) \approx \lambda * PR(t_i) \quad (3)$$

Some fraction of the visitors will place links pointing to the page on their pages or on those that they administer. This fraction greatly depends on the quality of the target page. Thus, similar to the approach of Cho et al. [4], we consider page quality as the conditional probability that a visitor will subsequently link to the page. Consequently, the increase in the number of inbound links can be represented as the product of the quality of the page and the size of the user stream at a certain point in time. However, the quality of a page at time t_i, $Q(t_i)$, is affected by the page content at that time. The quality can be in approximation represented as the sum of the qualities of particular elements on the page, such as for example the sentences and paragraphs. We should also evaluate the qualities of the inter-element relationships, however since evaluating relationships between elements is a difficult task, we consider now only independent qualities of elements. In this approach, each k element of the total K elements that ever stayed on the page during its lifetime has quality $q_k(t_i)$ at t_i contributing to the overall quality of the page $Q(t_i)$ at this time. These qualities are assumed to have zero values or fixed, positive values depending on whether they did not or did stay on the page at t_i. Thus, the overall quality of the page changes only when elements are added or deleted:

$$Q(t_i) = \sum_{k=1}^{K} q_k(t_i) \quad where: \quad \begin{array}{c} q_k(t_i) > 0 \quad if\ k\ exists\ at\ t_i \\ q_k(t_i) = 0 \quad otherwise \end{array} \quad (4)$$

Consequently, each element on the page has a direct impact on page quality and an indirect impact on the page's popularity and rank. If we assume that a page's inbound links are not deleted and omit λ in Equation 3, the number of inbound links at t_i is

$$L(t_i) = L(t_0) + \int_{t_0}^{t_i} Q(t) * PR(t)\, dt \quad (5)$$

where $Q(t)$ is a function describing the change in page quality and $PR(t)$ is a function describing the change in the PageRank value or in the user stream over time.

If the page had content changes at time points belonging to the sequence $T=t_1, t_2,\ldots,t_n$, the total number of inbound links is given by

$$L(t_n) = L(t_0) + \int_{t_0}^{t_1} Q(t_1) * PR(t)dt + \int_{t_1}^{t_2} Q(t_2) * PR(t)dt + ... + \int_{t_{n-1}}^{t_n} Q(t_n) * PR(t)dt \quad (6)$$

This equation shows that the current number of inbound links and hence indirectly the current PageRank depend on the page's history, i.e., its previous qualities and PageRanks. However, a page's PageRank depends to a large extent on the number of inbound links, hence also on the previous page qualities. Since the previous content of a page strongly affects the number of inbound links it now has and indirectly its rank, it can be considered the major factor in page popularity and can be used for detecting high-quality but low-ranked pages.

If we represent the quality of a page at certain point in time as the sum of the qualities of its parts at that time, the total number of inbound links at that time is

$$L(t_n) = L(t_0) + \int_{t_0}^{t_1} \sum_{k=1}^{K} q_k(t_i) * PR(t)dt + \int_{t_1}^{t_2} \sum_{k=1}^{K} q_k(t_i) * PR(t)dt + ... + \int_{t_{n-1}}^{t_n} \sum_{k=1}^{K} q_k(t_i) * PR(t)dt \quad (7)$$

The degree of impact of an element on the current link number of a page depends on its quality, PageRank value and on how long it is on the page. The longer the period, the stronger is the impact. Thus, the content that is relatively static has a higher impact on the page's rank than the content that is on the page for only a short while. This seems reasonable as the content that is on the page longer will be seen by more visitors. In this model we do not consider the qualities of the pages from where the inbound links originate. The focus is thus only on the changes in the number of inbound links.

3 System Implementation and Experiments

Since we do not know which pages could be used as topic representation knowledge for any arbitrary topic, we use some number of top-ranked pages from search engine results and retrieve their past versions from web archive. If lower ranked pages have historical content similar to that of higher ranked pages, their qualities should also be similar. In this case, one can ask why their ranks are still low. If the previous content of low-ranked pages demonstrated, on average, equally high quality, these pages should receive the same or a similar level of attention and trust from users as the top ranked pages. However this is not the case. For these pages to be treated fairly, their ranks should be increased. However, at the same time, we should check to see if their content is as relevant to the query as that of the top-ranked pages at the present time since their topics might have changed. Thus, additionally, in our approach, the present content of pages is compared with the present content of top-ranked pages.

To facilitate comparison of historical data, we combine the contents of previous versions of a page during certain time period to create a "super page" that presents the long-term history of the page. Each super page has a vector representation calculated using normalized *TF*IDF* weighting of the collection of super-pages of some number of top-ranked pages. Terms in the vector are stemmed and filtered using a stop list.

Because the more static content is contained in many consecutive page versions, it is repeated more often on the super page. Its importance is thus increased, and it has more impact on the super-page vectors. In this preliminary implementation we do not consider the *PR(t)* factor that is present in the link accumulation model (Equation 7). To achieve more accuracy we should combine historical content and link data. Thus the importance of the element of the page would also be dependent on the changes in page's PageRank values during the time periods when the element stayed on the page.

The new rank of a page is the result of re-ranking the search engine results based on the similarities between super pages and on the similarities between current page contents. The user issues a query to the search engine and specifies the number of top pages, *N*, to be re-ranked. The system then compares their historical and current contents with those of *W* user-defined top-ranked pages, where *W*<*N*. The first part of the new ranking computation, i.e., historical similarity, is calculated using

$$HR_j = \sum_{i=1}^{W} \left[\cos(VS_i, VS_j) * \left(1 + \varepsilon * \frac{W - R_i^{SE}}{W}\right) * \left(1 + \kappa * \frac{1}{Age_i}\right) \right], \tag{8}$$

where HR_j is the historical similarity value of page *j*, *co*s means cosine similarity, VS_i is the vector of the super page of page *i*, VS_j is the vector of the super page of page *j*, RS_i^{SE} is the search engine ranking of page *i*, and ε and κ are parameters. In our implementation the quality of the page depends on the query. It is represented as the average similarity of the page to the top-ranked pages for the certain query. It depends also on the ranks and ages of the top-ranked pages. The higher the ranks of the considered pages, the higher their weights in Equation 8. Thus, the *i-th* page from the *W* top-ranked pages is weighted according to its original ranking. Additionally, the younger the page, the more is important its content from the viewpoint of query topic representation. This means that a page had exceptionally important and/or high-quality content so it could achieve a high ranking in a relatively short time. Thus, HR_j also depends on the age of each top-ranked page. Consequently, sorting by HR_j produces page rankings based on historical similarity.

The second part of the new ranking computation, i.e., current similarity, is

$$CR_j = \sum_{i=1}^{W} \left[\cos(V_i, V_j) * \left(1 + \varepsilon * \frac{W - R_i^{SE}}{W}\right) \right], \tag{9}$$

where CR_j is the current similarity value of page *j*, V_i is the vector of page *i*, and V_j is the vector of page *j*. Unlike in Equation 8, the weight of page *i* depends only on its original ranking. To estimate the similarities of the current contents, we first calculate their *TF*IDF* vectors and then compute the cosine similarities between vector pairs.

The combined, new ranking of the page is the weighted average of its historical and current similarity rankings:

$$R_j^{new} = \frac{\alpha * HR_j^{rank} + \beta * CR_j^{rank}}{\alpha + \beta}. \tag{10}$$

We have conducted preliminary experiments to evaluate our approach by testing the system for several queries. We used the 30 top-ranked pages (*N=30*) for re-

ranking and the 10 top-ranked pages (*W=10*) from Google search engine for comparison. Super pages were created for the period from January 1 to December 31, 2004. We asked 3 subjects to assign ranks to pages. For example for query "Athens Olympics", the ranking computed by our system had Spearman rank correlation coefficient value [6] with relevance to user ranking equal to 0.31, whereas the search engine ranking had 0.24. However, for the query "Michael Jackson" the values were 0.25 and 0.26, respectively. The worse results for this query were probably due to large number of pictures and animations and few textual content that was observed in the top-pages of web sites devoted to the singer. We observed also that for many queries pages containing extensive descriptions of the query topic such as Wikipedia pages [7] were often assigned higher ranks by our system then the ranks computed by a search engine.

4 Conclusions

We have described a method for improving search engine results in order to alleviate rich-get-richer phenomenon by utilizing the previous content of web pages. The main assumption behind our approach is that the historical content of a web page greatly affects the page's current popularity and ranking. We have demonstrated it using content-based link accumulation model. By analyzing the previous versions of top-ranked pages, we can detect the core content for a certain query that has the main impact on the success of these pages. Then, low-ranked pages can be assigned higher ranks if their previous versions have high similarity to the core content for the query topic.

References

1. Amitay, E., Carmel, D., Herscovici, M., Lempel, R., and Soffer A.: Trend Detection Through Temporal Link Analysis. Journal of The American Society for Information Science and Technology, 55: 2004, 1–12
2. Baeza-Yates, R., Saint-Jean, F., and Castillo C.: Web Structure, Age and Page Quality. "String Processing and Information Retrieval", Springer, Lecture Notes in Computer Science (LNCS 2476) 117–130
3. Cho, J. and Roy, S.: Impact of search engines on page popularity. Proceedings of the 13th International World Wide Web Conference, New York, USA, 2004
4. Cho, J., Roy, S. and Adams, R.: Page quality: In search of an unbiased web ranking. In Proceedings of SIGMOD 2005, Baltimore, Maryland, USA, 2005
5. Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998
6. Spearman rank correlation coefficient: http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html
7. Wikipedia: http://www.wikipedia.org