A Novel Approach of Web Search Based on Community Wisdom

Weiliang Zhao and Vijay Varadharajan
Department of Computing
Macquarie University
NSW 2109, Australia
Email: wzhao,vijay@ics.mq.edu.au

Abstract

In this paper, we propose a novel approach for Web search based on the statistical information of local setting data of web browsers in a community. The members of the community share their local setting data of browsers and this enables them to take advantage of the peer community members's opinions in their Web search. Then we develop a new scheme that combines PageRank's link-based ranking scores with our proposed community based popularity scores for web sites. This hybrid scheme provides a rank-ordering method for search query results that integrates the content consumers' opinions with the content producers' opinions in a balanced manner. The users' opinions of web sites provide a solid starting point of trust for combatting web spam and improving the quality of Web search.

1. Introduction

The size of the World Wide Web is huge. The current indexed Web contains more than 20 billion pages [1]. The size of Web is still growing every day. With the information explosion, information navigation and retrieval is becoming increasingly difficult. Multiple search engines are currently being used to search and retrieve information on the Web. The most popular search engines include Google, Windows Live Search, Yahoo and Ask. All search engines are struggling to address rapidly expanding size of documents, vague queries, heterogeneous documents and web spam pages.

For a specific search query, no match message, a single or multiple found items can be returned. Typically, boolean search engines return items which match the criteria exactly without regarding to the order. Probabilistic search engines search the database against the matching criteria; they use some algorithms to rank the matched items and return them in order. The ranking algorithms are designed based on certain specific measures such as similarity, popularity or authority.

Pure boolean search engines are seldom used for providing final searching results because ranking measures are normally desirable for users, particularly when a large amount of items are included for a broad query. Even when a search engine claims that there is no order for the search result of a query, the displayed order of a searching result cannot be avoided. For users, the order of items in the returned list of a query is important. The items on the top of the list will get more attention. The items are very far away from the top of the list are possibly ignored.

In order to achieve higher-than-deserved rankings in the query results, various techniques are used by web spam pages to mislead search engines. The vast size of the web and the diversity of web documents make the ranking of query results a complex task. Web spam makes the task even more complicated.

There have been many research about how to rank the items in the searching results. Quite a few algorithms have been designed and deployed for ranking query results for different search engines. These ranking algorithms are normally based on the following strategies [2]:

- Linkage Analysis [5]: uses link information among pages on the web to calculate the importance scores and rank query results.
- Collaborative Web Search (CWS) [3]: uses the search experience of community members to promote the search results.
- Human Evaluation [4]: uses human editors to make judgements manually.

In this paper, we propose a new approach to create a community that shares statistics on specific information such as trusted sites, restricted sites, favorites and search history in the the design of the ranking in the web search algorithm. That is, the proposed approach of *community based web search* (CMWS) takes the advantages of users' opinions in the community. Then we propose to combine the widely used link-based ranking with our usage-based

978-0-7695-3163-2/08 $25.00 © 2008 IEEE
DOI 10.1109/ICIW.2008.66

ranking. Hence the popularity of web pages is calculated based on both the views of "author-to-author" as well as "user-to-author".

The remainder of this paper is organized as follows: Section 2 provides an overview of related work including PageRank, TrustRank, and Collaborative Web Search. Section 3 describes our proposal of *community based web search*. Section 4 describes the method of combining PageRank with our *community based web search*. Section 5 concludes the paper and briefly describes possible future work.

2. Related Work

In this section, we will provide a brief review of PageRank, TrustRank, and Collaborative Web Search.

2.1. PageRank

PageRank [5] was developed by Larry Page and Sergey Brin. PageRank forms the basis for Google's web search tools. It is a link analysis algorithm assigning a score for each element of a hyperlinked set of documents based on its popularity.

Consider a Web with N pages and assume there are some directed links that connect these pages. Assume that a web page p in the Web has multiple hyperlinks to other pages. The page p has incoming links and outgoing links. Let the number of outgoing links be denoted as $LO(p)$. The philosophy of PageRank is that the importance of a web page influences and is being influenced by the importance of other pages. A web page is considered to be more important if there are more important web pages pointing to it. The equation for PageRank score $PR(p)$ has the form:

$$PR(p) = \alpha \cdot \sum_{q \in M(p)} \frac{PR(q)}{LO(q)} + (1-\alpha) \cdot \frac{1}{N} \tag{1}$$

The $M(p)$ includes all incoming links to p. The α is a damping factor. The PageRank scores can be expressed by eigenvector of modified adjacency matrix. The eigenvector is

$$\mathbf{PR} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

The equation for PageRank eigenvector $\mathbf{PR}$ has the form:

$$\mathbf{PR} = (1-\alpha) \cdot \begin{bmatrix} 1/N \\ 1/N \\ \vdots \\ 1/N \end{bmatrix} + \alpha \cdot \begin{bmatrix} L(p_1,p_1) & \dots & L(p_1,p_N) \\ L(p_2,p_1) & \dots & L(p_2,p_N) \\ \vdots & L(p_i,p_j) & \\ L(p_N,p_1) & \dots & L(p_N,p_N) \end{bmatrix} \cdot \mathbf{PR} \tag{2}$$

The $L(p_i,p_j)$ is a function with the value $\frac{1}{LO(p_j)}$ if p_j links to p_i; otherwise the value is 0.

PageRank scores are computed iteratively based on the above eigenvector equation. Each page has the same initial score $1/N$. 0.85 is a typical value selected for α. With the iterative computing process, the eigenvector approaches convergence. PageRank scores express the degree of popularity of pages on the Web based on the link analysis.

2.2. TrustRank

Web spam is the term for web pages which are created to mislead search engines for achieving higher-than-deserved rankings in query results. There are different kinds of spam techniques. One kind of spamming technique is to mislead search engines using a large number of bogus web pages to point to a single target page. In order to combat such web spam, TrustRank [4] was proposed. The TrustRank employs human experts to identify spam in selected seed pages. Then it uses the link structure of the Web to evaluate other pages.

A binary *oracle function* O formalizes the notion of checking a page by a human editor:

$$O(p) = \begin{cases} 0 & if\ p\ is\ bad, \\ 1 & if\ p\ is\ good. \end{cases}$$

A trust function T is defined as the probability that a page is good; it has the form:

$$T(p) = Pr[O(p) = 1]$$

Select a seed set S out of N pages and call the *oracle* on the elements of S. The subsets of good and bad pages are denoted by S^+ and S^- respectively. The trust score is $1/2$ for the pages that have not checked by the human expert. Hence we have

$$T(p) = \begin{cases} O(p) & if\ p\ \in S, \\ 1/2 & otherwise. \end{cases} \tag{3}$$

Trust propagation is achieved by assuming that good pages point to other good pages only; the trust score is 1 for all pages that are reachable from a page in S^+ in M or less steps. Then T has the form:

$$T(p) = \begin{cases} O(p) & if\ p\ \in S, \\ 1 & if\ p \notin S\ and\ \exists q \in S^+\ : q \rightsquigarrow_M p, \\ 1/2 & otherwise. \end{cases} \tag{4}$$

where $q \rightsquigarrow_M p$ denotes that there is a link path from q to p and the path has less than M steps. Trust attenuation can be obtained using the assumption that trust is reduced if a page is further away from good seed pages. Some trust scores between $1/2$ and 1 may be introduced for pages linked directly or indirectly from good seed pages.

The Equation 2 can be rewritten as:

$$\mathbf{TR} = \alpha \cdot \tilde{L} \cdot \mathbf{TR} + (1 - \alpha) \cdot \mathbf{d} \tag{5}$$

where $\mathbf{TR}$ is the eigenvector of TrustRank scores and $\tilde{L}$ is the same matrix in Equation 2. The $\mathbf{d}$ is a static distribution vector of non-negative entries summing up to one. In PageRank algorithm, each entry of $\mathbf{d}$ is assigned the same static score $1/N$. In TrustRank algorithm, the values of entries of $\mathbf{d}$ are biased assigned. In Ref. [4], the $d(p_i)$ is assigned to 1 only when p_i is good seed page; otherwise $d(p_i)$ is assigned to 0. Then $\mathbf{d}$ is normalized by $\mathbf{d}/|\mathbf{d}|$. In Ref. [4], trust scores other than 1 and 0 in Equation 3 or Equation 4 are not used in TrustRank algorithm for deciding the static score distribution vector even that they have been listed and discussed.

2.3. Collaborative Web Search

There have been several research works on Collaborative Web Search [3]. The CWS is based on the high degree of query repetition and selection regularity among communities of like-minded searchers. In CWS, the search histories including queries and selections are recorded. These search histories provide the basis of a preference model for a community. CWS can be viewed as a form of case-based reasoning. The new queries as search problems are solved by retrieving and adapting the results of previous search cases.

Smyth's Group at Dublin [3] has developed an CWS implementation I-SPY, which captures search histories and uses them in the ranking metrics to reflect user behavior. The histories of queries and selections are stored in the *Hit-Matrix* H. H_{ij} denotes the number of times that a page p_j has been selected for query q_i. The *Hit-Matrix* is used as the direct source of relevancy information. The relevance value of page p_j and query q_i has the form:

$$Relevance(p_j, q_i) = \frac{H_{ij}}{\sum_{\forall j} H_{ij}} \tag{6}$$

In order to take the advantages of multiple similar queries, the normalized weighted relevance metric combines the relevance values of similar queries. This has the following form:

$$\begin{aligned} &WRel(p_j, q_T, q_1, \ldots, q_N) \\ &= \tfrac{\sum_{i=1,\ldots,N} Relevence(p_j,q_i) \cdot Sim(q_T,q_i)}{\sum_{i=1,\ldots,N} Exists(p_j,q_i) \cdot Sim(q_T,q_i)} \end{aligned} \tag{7}$$

where $Sim(q_T, q_i)$ denotes the similarity of query q_T and query q_i. $Exists(p_j, q_i) = 1$ if $H_{ij} > 0$ and 0 otherwise. The above relevance metric is used as the basis for rank-ordering the query results. There are also some further research on collaborating search communities which is beyond our present concerns in this paper. For further details, refer to [3].

3. Community Based Web Search

In the previous section, we have reviewed the fundamentals of PageRank, TrustRank, and CWS. As a link analysis algorithm, PageRank assigns a numerical weighting to each element of a hyperlinked set of documents. PageRank exploits the macro-scale link structure to evaluate the popularity of web sites based on author-to-author opinions. TrustRank employs human experts to evaluate web sites as trust seeds. It is a biased PageRank version for combating web spam. CWS takes advantage of the high degree of query repetition and selection regularity in a community of like-minded searchers for rank-ordering of query results. These approaches are successful in some aspects and while they have their own limitations.

In this paper, we propose a novel approach of web search based on the community wisdom. We believe that the searchers in a community will be willing to share their personal opinions of web sites. The users' personal opinions of web sites provide the basis for the analysis of popularity and trustworthiness of web sites. Our approach helps to create an incentive community, which is composed of web searchers sharing their opinions of web sites which will be used to evaluate the popularity and trustworthiness of web sites. The members of the community can take advantage of statistics of peers' opinions in the community.

3.1. Community Creation

In our scheme, the searchers are represented by their browsers on the Web. Hence our community is composed of a group of web browsers which agree to share their local setting data. Let us assume that there are M browsers in the community. For a browser $B_i (i = 1, \ldots, M)$, it has local setting data which includes Trusted Sites, Restricted Sites, Favorites and History which are denoted as TS_i, RS_i, FS_i, and HS_i. For a search engine SE, the members of the community agree to share their local setting data of their browsers. When they search the Web, the rank-orderings of search results are calculated based on the statistics of local setting data of browsers who have joined the community. From the view of members in the community, the search engine becomes a virtual search engine for the community members. We denote the virtual search engine for the community as CSE. The CSE has functions to collect and record the local setting data of browsers in the community. We assume that the CSE uses relational database and it has two tables to record local setting data of browsers, namely BrowserSetting_Table (*BrowserID, WebSiteURL, SetType, SettingTime, CheckingTime*) and SurfHistory_Table(*BrowserID, WebSiteURL, SurfingTime*). *SetType* includes types, namely *TrustedSite*, *RestrictedSite*, and *FavoriteSite*. When a browser B_i joins the community, a client software is installed as a plug in tool to perform client side tasks for checking and collecting TS_i, RS_i, FS_i, and HS_i. The browser is assigned a unique *BrowserID* (here we denote it as B_i) which can be recognized by the CSE. The unique *BrowserID* guarantees that one browser provides only one set of local setting data. It also provides the possibility for the future analysis of the behavior of the browser in the community. We assume that each time the web site of CSE is browsed, the client software checks the status of Trusted Sites, Restricted Sites, Favorites, and History, and sends the updated information of these local setting data to the CSE. In this paper, we only provide an outline of the scheme to collect the data in local browser setting and update the database on CSE server. The implementation details of data collection, delivery, and table updating are beyond the scope of this paper. The CSE should have privacy policy to guarantee that the local setting data of browsers can only be used by the CSE for the analysis of popularity and trustworthiness of web sites.

3.2. Community Based Popularity Scores of Web Sites

Here we use a $CURank(p_i)$ to express the popularity and trustworthiness of a web site p_i based on the community members' opinions. At this stage, we consider the popularity and trustworthiness as a single score, without differentiating them. (In our future work, we will be separating these two aspects). The score is denoted as $CURank(p_i)$ and it is calculated based on the local setting data of browsers in the community. The $CURank(p_i)$ has a subjective nature and now we devise the formula for calculating the score of web site p_i :

$$\begin{aligned} CURank(p_i) &= \alpha \cdot NumT(p_i) + \beta \cdot NumF(p_i) \\ &+ \gamma \cdot NumH(p_i) - \delta \cdot NumR(p_i) \end{aligned} \tag{8}$$

The following notations have been used in the above equation:

- p_i is the web site with URL $WebSiteURL(p_i)$.
- $NumT(p_i)$ is the repetitive number of $WebSiteURL(p_i)$ in the data table BrowserSetting_Table where the *SetType* = *TrustedSite*.
- $NumF(p_i)$ is the repetitive number of $WebSiteURL(p_i)$ in the data table BrowserSetting_Table where the *SetType* = *FavoriteSite*.
- $NumH(p_i)$ is the repetitive number of $WebSiteURL(p_i)$ in the data table $SurfHistory\_Table$.
- $NumR(p_i)$ is the repetitive number of $WebSiteURL(p_i)$ in the data table BrowserSetting_Table where the *SetType* = *RestrictedSite*.
- α, β, γ, and δ are parameters to express the relative weight factors of data sources of the community for trusted sites, favorites, history, and restricted sites.

The parameters α, β, γ, and δ are set subjectively. Trusted Sites are normally believed to be more trustworthy and popular than Favorite Sites. Hence α is assigned to a value that is bigger than the value of β. Favorite Sites are normally believed to be much more trustworthy and popular than surfing sites. Hence β is assigned to a value that is much bigger than the value of γ. For example, we can assign the value set for (α, β, γ, δ) as $\{8, 6, 1, 8\}$. The initial setting up of the parameter set (α, β, γ, δ) is based on intuition. It will be judged and adjusted later based on whether relative $CURank$ scores of web sites are reasonable or not.

The $CURank(p_i)$ is not normalized. The relative values of $CURank$ are more meaningful than their individual values. The $CURank(p_i)$ can be negative. The absolute values of $CURank$ normally become bigger when the community grows with more members. The $CURank$ scores provides a standard measure for the relative popularity of different web sites. Let us consider the case where for a query q, a set of web sites $\{p_1, p_2, \ldots, p_{M_q}\}$ are the query results. These web sites in the query result are then reordered based on their $CURank$ scores. This rank-ordering reflects the users' opinions of these web sites.

3.3. Discussion

Our approach presented above suggests a novel way to employ community-based information in the design of Web search algorithm. The local setting data of browsers are collected by the search engine and the statistics of these data are used as the measure of popularity of web sites. A new score system has been developed to express the popularity and trustworthiness of web sites based on the collected data including trusted sites, restricted sites, favorite sites, and sites in the search history list. Our community based rank score system provides a new way of order-ranking of query result in the Web search. Our approach has the following novel features:

- The proposed approach of community based web search suggests a reputation system purely based on users' opinions of web sites. The reputation system collects local setting data of browsers which have joined the community. $CURank$ scores of web sites are calculated based on the statistics of the collected data.

- The search engine is the server with the power of central control. The unbiased role of search is achieved because the algorithm of $CURank$ scores only accepts the opinions on web sites from all the members in the community. These opinions are believed to be embedded in the local setting data of browsers in the community.

- Comparing our approach with CWS, the $CURank$ scores have obvious advantages over the relevance metric in CWS. The trusted sites, restricted sites, favorite sites, and sites in the search history list reflect more accurately the popularity and trustworthiness of web sites from the viewpoint of users. The query repetition and result selection regularity of search results can only provide relatively weaker evidences or clues of relevance and popularity of web sites.

- In the formula for $CURank$ score, the time factor of trusted sites, restricted sites, favorite sites and sites in the search history list is not considered. We can envisage extensions where recent searching history lists of web sites having greater weighting relationship with the current popularity of web sites than an older search history list of web sites. We can easily extend this basis scheme with such temporal characteristics, which we believe is important. However as this initial stage, in our approach of community based web search (CMWS), we keep the formula 8 to be as simple as possible and have not introduced this time factor.

- Many reputation systems only take positive opinions, we take into account of negative opinions. The $CURank$ scores calculated by the formula 8 can be negative. The restricted sites contribute negative opinions of web sites in the community. A negative $CURank$ score corresponds to the bad reputation of a web site.

- We assume that the community based web search provides the appropriate incentive for the members in the community. In general, the community members can be expected to share their local setting data of browsers, as they can take advantage of peer users' opinions (and statistics of these data) in the community when they search the Web. In the implementation, different mechanisms can be deployed to achieve this sharing. For example, instead of becoming a permanent member of the community, the search engine may allow a user to share his local setting data of his browser for one time and take advantage of community based web search only in one session or a period of time.

- The size of the community (the number of members) affects the accuracy of $CURank$ scores for expressing the popularity and trustworthiness of web sites. Statistically, the community with more members has more data sources of opinions of web sites; hence the $CURank$ scores should be more accurate. The CMWS will become better as the size of the community increases.

- If the community is composed of like-minded searchers, the $CURank$ scores will be more accurate in expressing the popularity of interested web sites in the community. For example, if a community consisting of only rally car racing fans agree to share their local setting data of their browsers, then the community based web search would provide more accurate judgement of rally car racing related web sites.

4. Community Based TrustRank (CB-TrustRank)

In this section we propose a scheme that uses $CURank$ and $PageRank$ to complement each other. In Equation 5, we will use $CURank$ as the basis to compute the static score distribution vector $\mathbf{d}$. We still assume that $\mathbf{d}$ has non-negative entries summing up to one. For a web page p_i, $CURank(p_i)$ can be negative. We define $MCURank(p_i)$ to adjust $CURank(p_i)$. $\mathbf{TR}$ in Equation 5 is eigenvector of CBTrustRank. The whole process of calculating the CB-TrustRank has the following steps:

1. $CURank(p_i)$ is calculated for each page p_i in the web $(i = 1, 2, \ldots, N)$.

2. If $CURank(p_i) \geq 0$, set $MCURank(p_i) = CURank(p_i) + D$; otherwise set $MCURank(p_i) = 0$. Similarly to α, β, γ, and δ, D is set subjectively.

3. Compute $TCURank = \sum_{k=1,\ldots,N} MCURank(p_k)$.

4. Compute $d(p_i) = \frac{MCURank(p_i)}{TCURank}$ for each $p_i, i = 1, \ldots, N$.

5. CBTrustRank scores are computed iteratively with Equation 5 until the required convergence condition is satisfied.

The above CBTrustRank follows the initial idea in TrustRank to bias PageRank. The above CBTrustRank score system combines the advantages of proposed CURank scores and PageRank scores. The CBTrustRank has multiple new features which are beyond the original TrustRank. In the following, we provide some discussion of CBTrustRank scores:

- TrustRank employs human editors to manually evaluate a set of seed pages as the starting point of trust evaluation of web pages. The involvement of human editors introduces several issues and is often considered to be undesirable (at least minimized). Our proposed CBTrustRank scores employ $CURank$ scores to compute the biased static score distribution vector. Hence the human editors can be removed.

- PageRank scores are based on the pure link analysis. A $PageRank$ score reflects other authors' opinions about the web site. The $CURank(p_i)$ score is calculated based on local setting data of browsers. It aggregates community members' opinions about the web site p_i. The authors' opinions of a web site is usually more static than the users' opinions of a web site. CBTrustRank scores take the opinions of both the users and authors of web sites.

- In TrustRank, the maximum value of the entries of static score distribution vector is limited. All the believed good sites have the same $d(p_i)$ (if p_i is a good seed site or a good site based on trust propagation). The original PageRank scores are lightly biased by the static score distribution vector. In CBTrustRank, the maximum value of the entries of static score distribution vector can be quite big. The $d(p_i)$ has a broad distribution for all the web site p_i on the Web. The original PageRank scores can be heavily biased by the static score distribution vector. For a web site p_i with a very high $CURank(p_i)$ score, it can have a high CBTrustRank score even it is unreferenced. In TrustRank, the web link structure is still the dominant factor in the calculation of ranking scores. In CBTrustRank, both the link structure and CURank scores from community opinions of web sites can be dominant.

- The parameter α in Equation 5 can be used to adjust the relative weights of the content consumers' opinions and the content producers' opinions in the calculation of CBTrustRank scores. When $\alpha = 0$, CBTrustRank scores become the CURank scores and the link structure analysis is not included in the CBTrustRank scores. When $\alpha = 1$, CBTrustRank scores will be only based on the link structure of web sites.

- In CBTrustRank calculation, D is introduced to differentiate web sites with negative reputation from web sites without any user's opinion.

5. Concluding Remarks

We have proposed a novel approach for Web search based on statistics of collected local setting data of browsers in a community. The members of the community share their local setting data of browsers and this enables them to take advantage of the peer community members's opinions in their Web search. We have provided initial implementation outline of how to go about creating the community and collecting the required data, which are used in the calculation of popularity scores of web sites. The popularity scores of web sites form the basis for the rank-ordering of query results. The incorporation of the popularity statistics of users' opinions in the community, we believe can greatly improve the quality of Web search.

Then we have proposed a scheme that combines the ranking scores of web sites based on link structure analysis with our proposed community based popularity scores of web sites. This hybrid scheme provides a rank-ordering method of Web search query results that integrates the content consumers' opinions with the content producers' opinions of web sites in a balanced manner. Comparing this new scheme with TrustRank, the community based popularity scores provide a solid starting point of trust for combatting web spam and reducing the involvement of human editors. The main objective of this paper has been to introduce this new approach of web search based on community wisdom. As the next step, we are currently in the process of implementing a series of experiments to evaluate the performance of the proposed schemes and refining the algorithms.

References

[1] *The Size of the World Wide Web.* http://www.worldwidewebsize.com/.

[2] K. Bharath and G. Mihaila. *Hilltop: A Search Engine based on Expert Documents.* http://ftp.cs.toronto.edu/pub/reports/csrg/405/hilltop.html.

[3] J. Freyne and B. Smyth. Cooperating search communities. *Lecturer Notes in Computer Science*, 4018:101–110, 2006.

[4] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. *Proceedings of the International Conference on Very Large Data Bases*, 30:576–587, 2004.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report, Stanford University, http://citeseer.ist.psu.edu/page98pagerank.html, 1998.