Personalized Web Page Ranking Using Trust and Similarity

Lara Srour         Ayman Kayssi         Ali Chehab
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
{les03,ayman,chehab}@aub.edu.lb

Abstract

Search engines, like Google, use link structure to rank web pages. Although this approach provides an objective global estimate of the web page importance, it is not targeted to the specific user preferences. This paper presents a novel approach for the personalization of the results of a search engine based on the user’s taste and preferences. The concepts of trust and similarity, captured from explicit user input and implicit user behavioral patterns, are used to compute personalized page rankings.

1. Introduction

Web personalization is any action that makes the Web experience of a user personalized to the user’s taste and preferences [1]. Recommender Systems use similarity computation based on the Collaborative Filtering (CF) approach to provide people with recommendations for products they may appreciate, taking into account their past rating profiles and history of purchases. However, Massa [2] revealed that similarity computations tend to scale poorly for very large product sets (such as web pages). Moreover, CF techniques require high computational cost and thus lack scalability. There is a need to substitute CF with trust-aware techniques especially in a system with a very large population and a large product set. However, this substitution only makes sense if one can prove that trust reflects similarity. Ziegler and Lausen [3] revealed that trust and interest similarity tend to positively correlate but they argue that trust should supplement rather than replace existing filtering techniques. Thus, pre-filtering of like-minded peers based on social networks ensures the credibility of recommendations and reduces the computational cost. Applying supplementary filtering based on profile similarity guarantees that the resulting set of trusted peers resembles the active user [4]. In this paper, we propose a novel approach to personalize the results of search engines to the opinion of the specific user using trust and similarity. The system computes the similarity and trust scores between users and combines this information in order to come up with a local metric that is used to produce the weights assigned to users’ ratings, which are then aggregated into a personalized ranking. Since the incentive to provide explicit ratings is low, the system is capable of rating pages implicitly based on the length of time spent on the page.

2. Related Work

There are many attempts to rank web sites and web pages and recommend them to end users. Services like Scandoo.com, SiteAdvisor.com and TrustWatch.com provide users with recommendations about web sites to help keep them safe from spyware, spam, viruses and online scams. Some of these services monitor users’ activities in order to weigh the importance of the web site. However, none of these sites computes personalized ranking tailored to the user’s opinion and interests. The PageRank algorithm [5] is the most popular algorithm for ranking the results of a web search engine. It attempts to provide an objective global estimate of web page importance by using the link structure of the web to determine whether web pages are authoritative sources of information. This algorithm considers all links into a page to be a vote for that page. Massa et al. [6] suggest extending the web language so that semantic links can be expressed in order to discriminate between sites that are highly linked and sites that are highly trusted. Although these algorithms provide an objective global estimate of the web page importance, they might not necessarily capture the importance of a page for a given individual user. Eirinaki et al. [7] use a PageRank-like algorithm to bias the ranking based on the usage of a web site. The algorithm aims to propose “next pages” to a user based on his/her current visit and past users’ navigational pattern. However, this algorithm cannot be used by search engines, and is subject to cold start problem. Aktas et al. [8] introduce a methodology to personalize PageRank scores based on URL features such as Internet domains. A user is expected to input his/her interests as a set of domain features before query time. The algorithm is simple and can be better adopted as a filtering technique rather than a personalization engine. Mabasher et al. [1] present a web personalization system based on web usage mining which provides effective personalization using anonymous and implicit user behavioral pattern. The system, however, has a high real-time computation overhead.
3. The Personalized Page Ranking Approach

3.1. System Architecture

The overall system is divided into two components. The offline component includes the trust and similarity computation engines that result in the derivation of the local metrics. The other component is the online component that computes in real-time the implicit and explicit page ratings based on the user query and on the local metrics derived from the first component. The result obtained is a personalized ranked list of web pages.

3.2. User Profile

Each user completes a profile containing a username and password, and information related to his/her profession and interests. Also, each user can indicate the users s/he trusts in his/her “web of trust”. These are users that s/he either knows and trusts, or whose comments and reports are deemed by the user as useful and accurate. Users are also requested to download a browser plug-in that allows the search engine to monitor the user activities, such as page visits, time spent on each page, and user ratings, in order to use this information to compute personalized ranking.

3.3. Trust and Similarity Computation

Due to the excessive computational cost of determining the similarity between any two users in real-time for a large number of pages and users, we restrict similarity computation to only neighborhoods of trustworthy peers. Trust scores are computed by propagating trust through the webs of trust. On the other hand, similarity values are computed from three parameters: page ratings, pages visited in common and users’ profession and interests. The local weight assigned to each user is a function of the predicted trust score and similarity value.

3.3.1. Computing Trust. A user is much more likely to believe statements from a trusted peer than from a stranger. And recursively, since a trusted friend will also trust the thinking of his/her friends, trust may propagate (with appropriate decay) through the trust network. Propagating trust permits us to predict trust between any two users from a small number of expressed trusts per user. The system thus provides personalized rankings quickly to new users; there are no “cold start users” as in computing similarity. Thus, additional filtering based on profile similarity needs to be applied to the neighborhood of trustworthy peers to guarantee that the resulting set resembles the active user. Accordingly, the system computes the similarity of the active user with the peers that are reachable within the propagation horizon. First, the system computes Sim1, the similarity between users based on the pages visited in the past. The larger the overlap between the pages visited by both users, the more similar the two users are:

\[
Sim_1 = \frac{\text{PagesInCommon}}{\text{PagesVisitedByX} + \text{PagesVisitedByY}}
\]

Second, the system computes Sim2, the similarity between two users based on their profession and interests, as entered in the user profiles. Users with same profession and similar interests are more likely to have similar views and opinions:

\[
Sim_2 = x \ast \text{ProfSim} + y \ast \text{IntSim}; \ x + y = 1
\]

- ProfSim is the similarity between the two users’ professions ranging between 0 and 1.
- IntSim is the similarity between the interests of the two users and is calculated as follows:

\[
\text{IntSim} = \frac{\text{NbrOfInterestsInCommon}}{\text{NbrOfInterestsOfX} \times \text{NbrOfInterestsOfY}}
\]

Third, the system computes Sim3, the similarity of the two users based on the ratings entered by both users for the same pages. This is done by calculating the Pearson correlation coefficient and normalizing it to the interval [0-1]. The Pearson correlation coefficient is used in Collaborative Filtering to assess user similarity in Recommender Systems. The coefficient allows for detecting negative correlations (the values obtained range from –1 to 1) which occur when users have completely diverging interests [3]. In our case, if users have diverging interests, they will have zero similarity. However, this coefficient cannot be computed if the users have rated less than 5 items in common. Pearson correlation coefficient calculated between users a and u is given by:

\[
w_{a,u} = \frac{1}{\sqrt{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_{a})^2 \sum_{i=1}^{m} (r_{u,i} - \bar{r}_{u})^2}}
\]

Where m is the number of items rated by both users; \(r_{a,i}\) and \(r_{u,i}\) are the ratings given by users a and u, respectively to page i; \(\bar{r}_{a}\) and \(\bar{r}_{u}\) are the mean of ratings provided by a and u, respectively.

Finally, the similarity coefficient between two users X and Y is computed as:

\[
\text{Sim}(X, Y) = A \ast \text{Sim1} + B \ast \text{Sim2} + C \ast \text{Sim3}; \ A + B + C = 1
\]
since all users must have a profile and must have visited at least a few pages. When it is difficult to calculate similarity values for new users who have not visited or rated any pages, trust scores alone are used as an alternative weight to rank pages for these new users.

3.3.3. Combining Trust and Similarity. People usually ask for recommendations from trusted friends with similar tastes and interests. For this reason, we select for each user the group of similar and trusted users who will help him/her get personalized results. After calculating the similarity and trust scores that each user has with his/her neighbors, we combine this information in order to compute a local metric that will be used to produce the weights assigned to users’ ratings and aggregate them into a personalized ranking. The local metric assigned to user \( X \) by user \( Y \) is given by:
\[
LM(X, Y) = a \times TV(X, Y) + b \times Sim(X, Y);
\]
\( a + b = 1 \)  

Users whose local metric is below a certain threshold (\( \theta \)) will be neglected and thus will not contribute to the personalized recommendations. New users, for whom the similarity cannot be computed with anyone in their trust group, will have the local metric assigned to other peers equal to their trust score. Finally, we note that similarity is symmetric but trust is not.

3.4. Personalized Ranking

3.4.1. Implicit and Explicit Rating. Since users’ incentive to provide explicit ratings tends to be low [3], we should provide a way that allows the system to rate pages implicitly. Implicit ratings are captured as a result of the user interactivity, using for example, web page visit rates, length of time spent on the web page, etc. It is also possible to combine both implicit and explicit ratings to get a final rating value for a page.

3.4.2. Decay of Ratings with Time. A rating that was made a long time ago may not actually reflect the importance of the page now. For this reason, more recent ratings should be given more importance. Thus, ratings should decay with time. The change with time is made exponential according to:
\[
e^{-\frac{t-t_0}{\tau}}
\]
where \( t_0 \) is the time when the rating was computed and \( \tau \) is an empirical constant that determines how quickly or slowly the rating becomes invalid [10].

3.4.3. Computing a Personalized Page Rank. Among the selected peers for whom an above-average local metric was calculated (i.e. whose \( LM > \theta \)), some may have rated at least one of the pages of interest, while others have just visited some of these pages. From the first group, an explicit rating for the page is calculated. From the second group, an implicit rating is calculated based on the length of time spent on the page. The aggregated rating is a weighted function of user ratings with weights assigned to each user equal to his/her local metric. Finally, the assigned rating to a page will be a weighted function of both implicit and explicit ratings:

\[
LocalRating(P) = \alpha \times ExplicitRating(P) + \beta \times ImplicitRating(P)
\]

4. Simulation

4.1. Simulation Setup

In order to evaluate our proposed model, we developed a simulation prototype using Microsoft Visual C++. We simulated a network consisting of twelve users and twelve pages (see Tables 2 and 3). The constant values used in the simulation are shown in Table 1. The value of \( \theta \) is taken as the average of all calculated local metrics and its value is determined from the simulation.

Table 1: Simulation Constants

<table>
<thead>
<tr>
<th>Parameter</th>
<th>d</th>
<th>x</th>
<th>y</th>
<th>( \theta )</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>Tmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>3</td>
<td>0.75</td>
<td>0.25</td>
<td>0.45</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 2: User Profiles

<table>
<thead>
<tr>
<th>User</th>
<th>Profession</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0-1]</td>
<td>engineering</td>
<td>business</td>
</tr>
<tr>
<td>[2-3]</td>
<td>engineering</td>
<td>art</td>
</tr>
<tr>
<td>[4-5]</td>
<td>business</td>
<td>art</td>
</tr>
<tr>
<td>[6-7]</td>
<td>business</td>
<td>engineering</td>
</tr>
<tr>
<td>[8-9]</td>
<td>art</td>
<td>engineering</td>
</tr>
<tr>
<td>[10-11]</td>
<td>art</td>
<td>business</td>
</tr>
</tbody>
</table>

Table 3: Page Profiles

<table>
<thead>
<tr>
<th>Page</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1-4]</td>
<td>engineering</td>
</tr>
<tr>
<td>[5-8]</td>
<td>business</td>
</tr>
<tr>
<td>[9-12]</td>
<td>art</td>
</tr>
</tbody>
</table>

We assumed that each user has three other users in his/her web of trust, two of them having the same user’s profession and the third one having the same profession as the user’s interest. Also, each user belongs to the web of trust of three other users. Initially each user visits three pages, two of which have the same topic as the user’s profession and the third one has the same topic as the user’s interest. We assume that all pages are equally visited, thus each page is visited by three users. We also assume a 50% probability that a user rates the page that s/he visits. If no explicit rating is provided,
the system records the time spent by the user on the page and uses this time to calculate implicit rating. Ratings and time spent are randomly generated in the ranges [1-5] and [5min-60min], respectively. Since pages considered in the simulation are very few, we assumed that Pearson correlation coefficient can be calculated if the users have at least one rated page in common.

4.2. Simulation Results

Since all users are assumed to have similar characteristics, we will observe the results obtained for user ‘0’. The trust, similarity, and local metric assigned by user ‘0’ to other users in the system are shown in Figure 1. Since $\theta = 0.45$, users that are locally trusted by user ‘0’ and will contribute to his/her recommendations are those users whose $LM > 0.45$. These are users 1, 2, 3, 4 and 6. If only trust is considered with threshold 0.5, three additional users (5, 8 and 10) will contribute in making the recommendations. On the other hand, if similarity alone is to be considered with a threshold equal to the average similarity (found to be 0.25), the selected users would be 1, 2, 3 and 6. In this case, user 4 is eliminated although s/he belongs to the user’s web of trust. We can conclude that using trust alone as local metric increases coverage; however, it may result in receiving recommendations from users with diverging interests. Also, using similarity alone may result in ignoring recommendations given by directly-trusted peers.

![Figure 1. Trust, Similarity and LM Calculated by User 0](image1.png)

The personalized ratings generated to user ‘0’ for each page in the above mentioned three cases are shown in Figure 2. Negative ratings mean that a local rating could not be calculated. This is due to the fact that none of the locally trusted users has visited or explicitly rated the page. We notice that three pages could not be rated in case 2 (Sim in Fig. 1) and two pages in case 3 (LM in Fig. 1), while all pages were successfully rated in case 1 (TV in Fig. 1). This is expected since using trust alone increases coverage as stated previously. If we want to rank the pages, we notice that different ranked lists are obtained in each case. We expect the third case to give the most accurate result.

![Figure 2. Personalized Page Ratings for User 0](image2.png)

5. Conclusion

In this paper, we proposed a novel approach to personalize the results of search engines to the opinion and interests of the specific user based on his/her behavior in the past, and similarity and trust towards other users. Due to the excessive computational cost of determining the similarity between any two users in real-time, we restrict similarity computation to neighborhoods of trusted peers. Since users’ incentive to provide explicit ratings tends to be low, we provide a way that allows the system to rate pages implicitly based on the length of time spent on a page. Simulation results show that combining trust and similarity is the best approach to get accurate personalized recommendations for web pages.

References