Collaborative Web Search Using Likewise Users' Recommendations

Nakulraj K R
Student
Department of Computer Science,
Vidya Academy of Science and Technology, Thrissur, India

Abstract **- Many tasks in both professional and casual settings can benefit from the ability to jointly search the Web with others. The intuition was that such situations might be commonplace. Most users seek the help of their friends or knowledge domain experts to find the relevant information from the web. Web search thus mostly become a group activity rather than a solo one. Collaborative Web Search exploits repetition and regularity within the query-space of a community of like-minded individuals in order to improve the quality of search results. In short, search results that have been judged to be relevant for past queries are promoted in response to similar queries that occur in the future. The main aim of Collaborative Web Search is to provide users a platform to work cooperatively by sharing their web search experiences and thus improving search quality. Collaborative methods are less supported by current web browsers and search engines. We know that experts often find it easy to get better results through search engines due to their domain knowledge and so on. The sharing of these experts' search experiences will be always helpful for other users to get better search results. A web browser toolbar or add-on can be used for sharing search experiences between collaborators. This toolbar catches the search histories and is uploaded into a recommendation server. These collected search histories are converted into hierarchical user profile according to some rules at the recommendation server. Then these experts' profiles are used wisely to provide valuable recommendations in the search activities of collaborators. The key advantage of collaborative web search is the information gain with the fulfillment of user attributes, along with the acceptance of user community and with the saving of time that might be spent in the irrelevant documents.**

Index Terms **- Collaborative Web search, shared Web search experiences, user profile, personalization.**

I. INTRODUCTION

Collection size, document diversity, and limited searcher expertise all combine to make the Web a very challenging information retrieval environment. The entire World Wide Web consists of billions of information and it grows every single day. Some of the data are useful while some are not. Web search results too many pages that might not be of the user's interest. User spends lots of time in searching data and information on the rich web. Usage of online material is playing an important role for the learners in self-discovery learning process but it is difficult to find relevant pages from the web because of gigantic web pages.

For accessing information online, Web search engines have become the dominant tool over the past few years. However, the vast information still results in such problems and phenomena as "getting lost in information" and "information overloading" because information requirements of end users are so limited. Therefore, many research projects are targeted on how to make it easier for end users to find the information they want efficiently and accurately. In a typical interaction with a Web search engine, end users enter a specific information need, expressed as a query, and obtain a great many search results. Among these results, some are relevant to the query, but some are not. Typically, users expect to find relevant information in the top-ranked results. However, relevant results are always mixed with, and even presented after irrelevant ones. It indicates that ranking schemes should take into account not only the overall page quality and relevance to the query, but also the match with the users' real search intents when they formulate the query. Nevertheless, a typical Web query contains only two or three terms. The short Web search queries are so vague that it is difficult to distinguish the searcher's true information needs.

A survey revealed that a large proportion of users engage in searches that include collaborative activities. A key problem here is the misinterpretation of collaborators' written messages to each other. Browsing computerized information resources has a social and collaborative dimension which will be increasingly remote and asynchronous. The storage and re-use of the search process provides a mechanism to support a variety of activities which users may wish to undertake. The visualization of a search process can be a useful means to abstract information and aid collaboration between information workers. Sharing the search knowledge within the community will be of great benefit to the individual users. Search communities are interesting because of the high likelihood that similarities will exist among the members of the community.

The need and advantage of collaborative web search has been identified and various studies, surveys and researches have been conducted in this field and are still going on. The studies says that even though all the information are available on the web and are easily accessible with today's high speed internet and developed search engines, people still seek the co-operation from other in their search activity. The main reason for this are:

- Lack of confidence (mainly to students and beginners due to their lack in experience) in their search activity.
- The unawareness of the correct resource location.
- To get the best result out of many.
- To save the time spending in searching irrelevant documents.

The survey paper by M R Morris et al. shows out various situations in which people engage in co-operative search. Students or children engage in web search to gain knowledge about their learning subjects. But the often have to spend too much of time to find the relevant documents. So they mostly seek the help of their friends, elders or teachers during the search process. Teachers or friends help them suggesting the correct query keywords or the links to follow. Business person engage in cooperative web search to find right market places, current economic scenarios, purchasing online products etc. Common people cooperates for travel planning (Researching travel info for a group trip, to match budgets & personal tastes), general shopping tasks, job related tasks and fact finding. So the fact is that knowingly or unknowingly people get into the joint search process. This paper presents an approach for collaboration in web search and thereby improves the search quality.

II. PROPOSED METHOD

This paper presents a convenient way for users to share and utilize experiences of collaborators through a Web browser toolbar for collaborative Web search. This toolbar is built using plug-in or add-on or extension facility of web browsers and these facilities are supported by almost all modern web browsers. This special toolbar can control the document model of a web page and capture her click actions and extract the title, url, and others of the page. In addition, all data extracted can be uploaded to a recommendation engine server for processing with the help of some software development technologies and recommendations also be downloaded and merged with the return-list by a search engine. The architecture of the system is as follows:

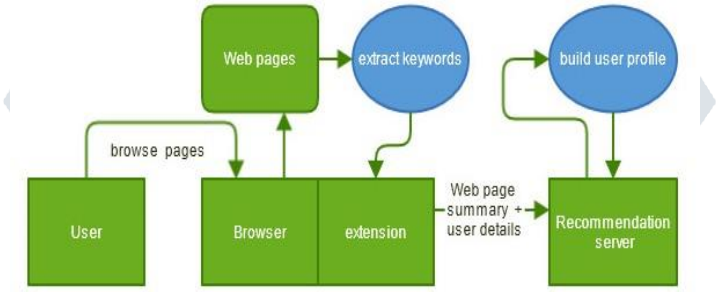

Figure 1: Architecture of system for building user profile

One important issue here is, browsing history is mostly unstructured. In addition, it is also difficult to incorporate unstructured data with search engines without summarization. So, for the purpose of both web personalization, it is necessary for an algorithm to collect, summarize, and organize a user's personal information into a structured user profile. To solve this issue it offers a scalable way to automatically build a hierarchical user profile on the client side. It's not realistic to require that every user to specify their personal interests explicitly and clearly. Thus, an algorithm is implemented to automatically collect personal information that indicates an implicit goal or intent. The user profile is built hierarchically so that the higher-level interests are more general, and the lower-level interests are more specific.

Constructing a Hierarchical User Profile

Any personal documents such as browsing history on a user's computer could be the data source for user profiles. Our hypothesis is that terms that frequently appear in such documents represent topics that interest users. This focus on frequent terms limits the dimensionality of the document set, which further provides a clear description of users' interest. This approach proposes to build a hierarchical user profile based on frequent terms. In the hierarchy, general terms with higher frequency are placed at higher levels, and specific terms with lower frequency are placed at lower levels.

D represents the collection of all personal documents and each document is treated as a list of terms. D(*t*) denotes all documents *covered* by term *t*, i.e., all documents in which *t* appears, and |D(*t*)| represents the number of documents covered by *t*. A term *t* is *frequent* if |D(*t*)| ≥ *minsup*, where *minsup* is a user-specified threshold, which represents the minimum number of documents in which a frequent term is required to occur. Each frequent term indicates a possible user interest. In order to organize all the frequent terms into a hierarchical structure, relationships between the frequent terms are defined below. Assuming two terms t_A and t_B, the two heuristic rules used in our approach are summarized as follows:

1. **Similar terms**: Two terms that cover the document sets with heavy overlaps might indicate the same interest. Here we use the Jaccard function to calculate the similarity between two terms: $\text{Sim}(t_A, t_B) = | D(t_A) \cap D(t_B) | / | D(t_A) \cup D(t_B) |$. If $\text{Sim}(t_A, t_B) > \delta$, where δ is another user-specified threshold, we take t_A and t_B as similar terms representing the same interest.
2. **Parent-Child terms**: Specific terms often appear together with general terms, but the reverse is not true. For example, "badminton" tends to occur together with "sports", but "sports" might occur with "basketball" or "soccer", not necessarily "badminton". Thus, t_B is taken as a child term of t_A if the condition probability $P(t_A | t_B) > \delta$, where δ is the same threshold in Rule 1.

Rule-1 combines similar terms on the same interest and Rule-2 describes the parent-child relationship between terms. Since $\text{Sim}(t_A, t_B) \leq P(t_A | t_B)$, Rule-1 has to be enforced earlier than Rule 2 to prevent similar terms to be misclassified as parent-child relationship. For a term t_A, any document covered by t_A is viewed as a natural evidence of users' interests on t_A. In addition, documents covered by term t_B that either represents the same interest as t_A or a child interest of t_A can also be regarded as

supporting documents of t_A. Hence *supporting documents* on term t_A, denoted as $S(t_A)$, are defined as the union of $D(t_A)$ and all $D(t_B)$, where either $Sim(t_A, t_B) > \delta$ or $P(t_A \mid t_B) > \delta$ is satisfied.

Utilizing Expert's Experiences

Based on the above rules, a hierarchical user profile can be automatically built in a top-down fashion. The profile is represented by a tree structure, where each node is labeled a term *t*, and associated with a set of supporting documents *S*(*t*), except that the root node is created without a label and attached with a user's name of *D*, which represent all personal cases. Starting from the root, nodes are recursively split until no frequent terms exist on any leave nodes.

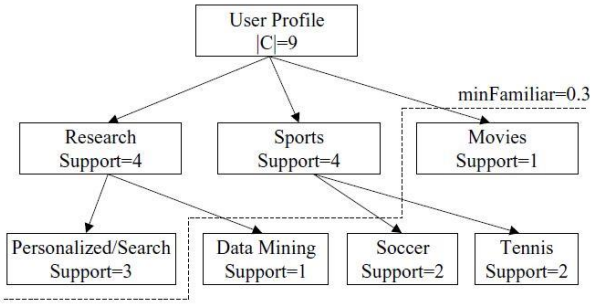

Figure 2: A hierarchical User Profile

With the hierarchical user profile constructed above, every term with supporting search cases can be detected. The support of an topic of a term *t* is *Sup*(*t*), and *S*(*t*) represents all the supporting cases for term *t*. $\sum Sup(t) = |D|$ is for all terms *t* on the leave node, where |*D*| represents the total number of supports received from a user's search cases. In addition, our hypothesis is that a term *t* with larger *Sup(t)* represents a user's familiar topic and partial search cases in *S(t)* are her valuable experiences.

The user profile is established as an indicator of the user's possible individual interests. According to probability theories, the possibility of one interest (or a term) can be calculated as P(*t*)=Sup(*t*)/|*D*|. Within the context of information theory, the amount of information about a certain interest of the user is measured by its *self-information*:

$$\mathbf{I(t) = \log(1/P(t)) = \log(|D|/\,Sup(t))}, \text{ for any term } t. \qquad (1)$$

This measure has also been called *surprisal* by Myron Tribus, as it represents the degree to which people are surprised to see a result. More specifically, the smaller *Sup(t)* is, the larger the self-information associated with the term *t* is, and more surprise occurs if the term *t* is exposed. More specifically, the smaller *Sup*(*t*) is, the larger the self-information associated with the term *t* is, and the search case including term *t* is more valuable as it is a special search case for a user. This leads to a parameter for specifying the requirement of recommendation.

minFamiliar: The user profile above is organized from high-level to low-level. Terms associated with each node become increasingly specific as the list progresses, and same level terms are sorted from left to right in descending order of their supports. A threshold of *minFamiliar* is defined to measure users' familiar topics on both vertical and horizontal dimensions. With a specified *minFamiliar*, any term *t* in the user profile with $P(t) = Sup(t)/|D| \geq minFamiliar$ will be taken as a user's familiar topic.

Firstly, the possibility of every topic of a user is calculated. Then, for every user a subtree of his hierarchical user profile, *U[Fam]* is constructed such that *U[Fam]* consists only those nodes which have possibility of term in node, $P(t) \geq minFamilier$. For conventional, *U*[*Fam*] is transformed into a list of weighted terms and the weight of each term in *U*[*Fam*] is estimated by applying the concept of IDF (Inverse Document Frequency). Given a term *t*, the weight of *t*, denoted by w_t, is calculated as:

$$\boldsymbol{w_t = log(|D|/Sup(t)),} \qquad (2)$$

where |*D*| represents the total number of search cases of *U*[*Fam*], and *Sup*(*t*) is the support of this term on the node in *U*[*Fam*]. The user profile is expressed by a list $< t, w_t >$, where *t* is a term in *U*[*Fam*] and w_t is the weight.

In order to incorporate the user profile with results returned by a search engine, *U*[*Fam*] is transformed into a list of weighted terms where a search wrapper calculates a score for each of the returned search results. The final ranking of the search results is decided by the search engine and *U*[*Fam*].

Next is to choose the right experts' experiences to recommend. The question here is whether a solution can be found where users' experiences can be effectively filtered to improve the search quality. As a hierarchical user profile can summarize user's experiences into different levels with different supports, general topics with more supports can be taken as familiar topics and experiences under such topics can be taken as experts' experiences.

When a user inputs a query, a set of terms, in a search engines, the toolbar would capture the snippets of the results provided by the search engine and upload the terms and other details of each snippets to the recommendation server. The server would find valuable experts' experiences through travelling expert profile (say E). But which experiences are most valuable for the searcher? In our opinions, search cases, which don't appear in her profile but include terms with larger self information according to E, are more valuable. So we firstly choose such search cases according to E: their support topics include terms appeared in the query.

Recommendations are built when a query is submitted to the recommendation server in five steps:

1. The expert profile of every user is built and represented by a set of $< t,w_t >$ pairs in the recommendation engine server.
2. When a user makes a query the toolbar captures a query and the search results returned by a search engine and they are uploaded to the recommendation engine server. Each result comprises of a set of links related to the query, where each link is given a rank from the search engine, called DefaultRank.
3. For each of the returned link *l*, a score called *EPScore (Expert Profile Score)* is calculated by the expert profile as follows:

$$EPScore(l) = \sum_t w_t \times f_t \qquad (3)$$

where *t* is any term in the expert profile and f_t is the frequency of the term *t* in the snippet of the link *l*. An *EPRank* is assigned to each link according to its *EPScore*, and the link with the highest *EPScore* will be ranked first.
4. Re-ranking results by combining ranks from both DefaultRank and EPRank. The final rank, ECRank (Enhanced Collaborative Rank), is calculated as:

$$ECRank = \alpha * EPRank + (1 - \alpha) * DefaultRank, \qquad (4)$$

where the parameter $\alpha \in [0, 1]$ indicates the weight assigned to the rank from the expert profile. If $\alpha = 0$, the expert profile is ignored, and the final rank is decided by the expert profile instead of the search engine when $\alpha = 1$.
5. The toolbar downloads the final ranking of the search results and recommends them to the user.

III. Experiments and Evaluations

All experiments are conducted with the following objectives: to verify the effectiveness of the clustered user profile to help search quality improvement, and to explore the relationship between search quality and expert's experiences.

In the performance evaluation of Collaborative Web Search Using Likewise Users' Recommendations, there are mainly two things to be evaluated. The prior one is to verify the effectiveness of system to improve search quality (ie, to check whether users find it more useful to get the relevant results better than normal search engine results). At the same time it is to be ensured that the recommendations (ie, re-ranking of the search engine results) are provided by the CWS system (Collaborative Web Search System) according to the expert's user profile.

Evaluating the effect of expert's profile in re-ranking search engine results:

In this section, tests are conducted with objective to ensure that the experts' profile is effectively used to provide recommendations and to check the effect of the user profile in re-ranking the results. All the test are conducted with following values *minsup = 2, α= 1, δ = 0.6, minfamiliar = 0.3.*

First, two user are made to register into the system (say User A and User B). The User B is made to follow user A. Then a search case which occur in various domain was randomly selected (here, say term 'Rose'). Here, by various domain we mean that when we search the word 'Rose' in Google, we can see that the each link of results are from different domain. In the next step, we copied 3 random snippets out of the result links of term 'Rose'. Then these 3 snippets are used to create 10 different html pages. The first copied snippet is placed in 6 out of 10 html pages. The second snippet is placed in 3 html pages and third in 2 html pages.

In the next stage, these 10 html pages are run in a local server and User A was made to visit these pages and user profile was created for user A. Then User B was made to login to system with User A as his expert to provide recommendation and the same word 'Rose' was searched by him in the Google. The results were obtained and evaluated. The snippet which occurred in most html pages of User A are ranked 1 and snippet which occurred in 3 html pages is ranked second. All other snippets are ranked with rest of the ranks. No special rank was there for the third copied snippet and it may be because it occurred in few html pages that User A visited. The figure below (figure 3) shows the screenshot of re-ranked search results.

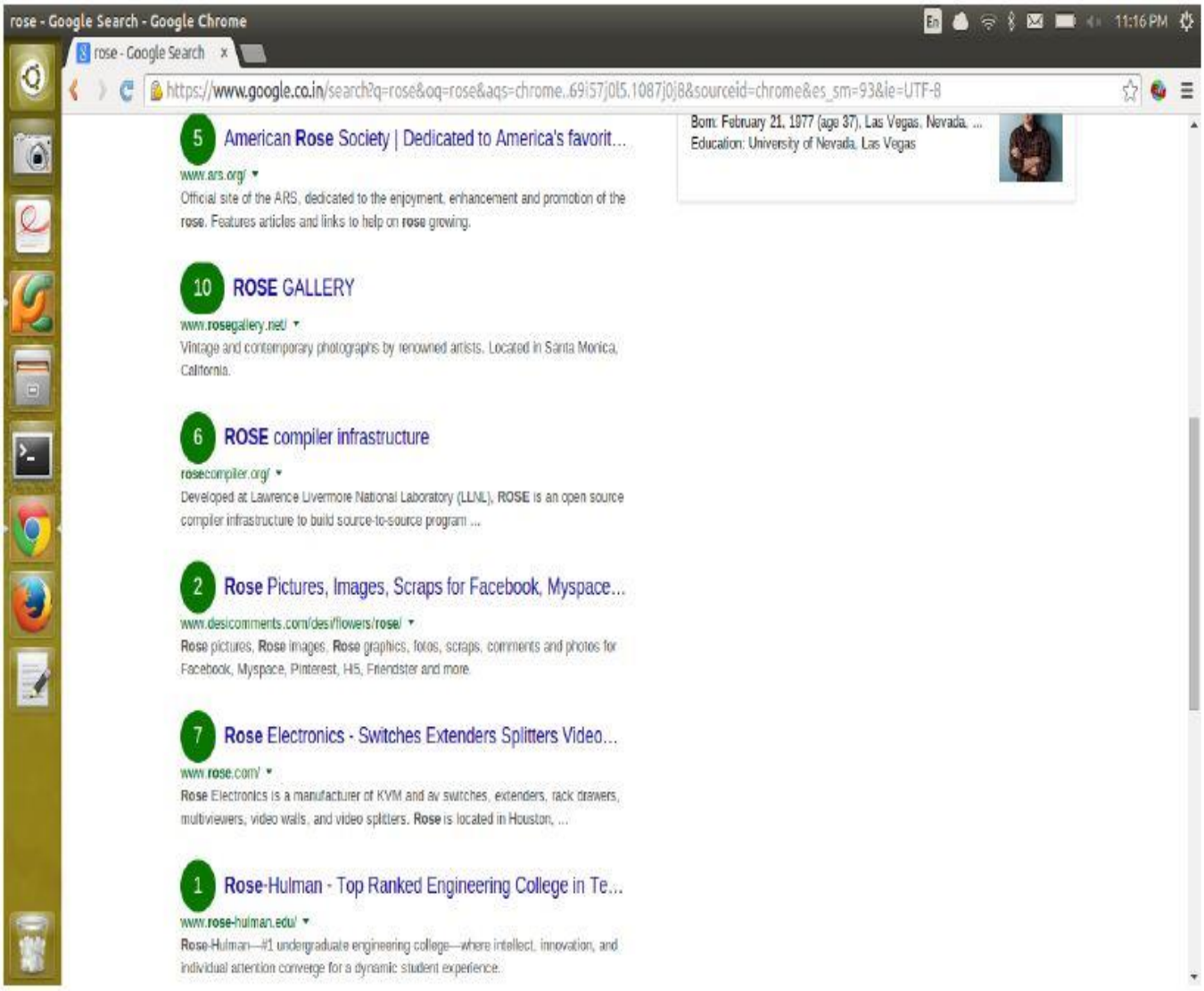

Figure 3: Re-ranked search results

Performance evaluation by users:

In this approach of collaboration, one user's profile which is built according to his search behavior is used to make recommendation for another user helping him to get better results. This scenario works ideally for users with same domain of interest, especially when one user with less domain knowledge uses recommendation from another user with good search experience in the same domain. And these kinds of users are supposed to be the best beneficiaries, at the same time the best evaluators the system. In order to get the users with same domain of interest, 5 groups of students doing their degree project were selected. Each group consisted of 4 members. The system was then clearly explained to these 20 students along with some examples.

Then each group is divided into two (say Team-A and Team-B) consisting of two members from the same group. The Team-A of each group were requested to search terms related to their field of project through the search engines. And they were also requested to copy down the links of 30 to 50 webpages that they find relevant during their searching. Then an account was created for Team-A of each group in our collaborative search system. Then each group's Team-A members were allowed to login into the system and were requested to browse their related webpages whose links are previously copied down during the search. Then the user profile are generated for Team-A of each group.

Later, account was created for each Team-B. Each Team-B account was made to follow their corresponding Team-A account as their expert to provide recommendation during their search. Then each Team-B was requested to make 10 search queries each in their field of their project to Google search engine. For each query applied, the results were obtained which was associated with two ranks ie, the normal Google rank and rank given by the collaborative search system. Along with these ranks, each corresponding Team-B was requested to give their own rank to each link of Google search results accordingly as the relevance they felt.

The irrelevant links are ignored. The participants were requested to note down these three ranks in a spreadsheet document for each relevant links identified by them for each query. The following chart (figure 4) shows deviation of ranks made by CWS system and Google search from that of ranks given to the relevant links for the query made by Team-B of Group-1.

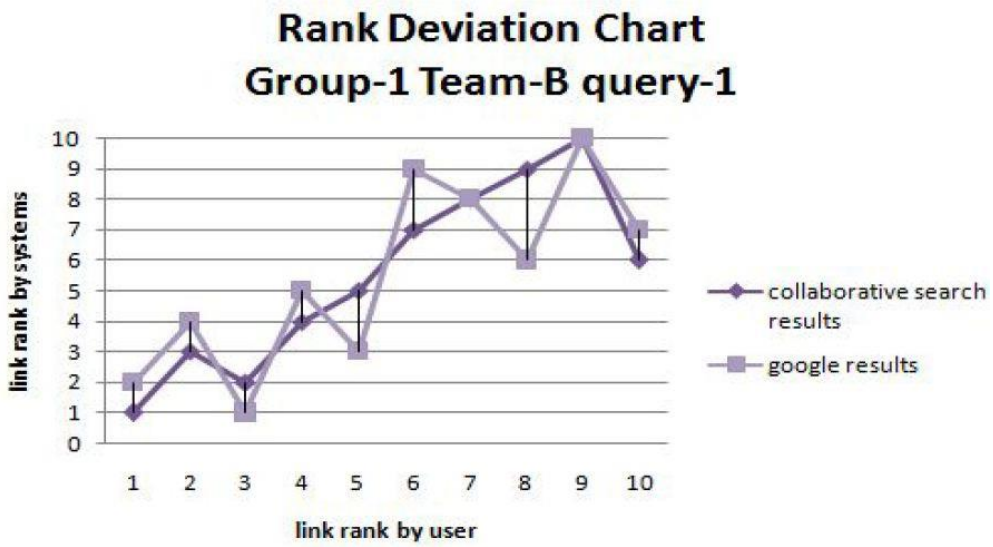

Figure 4: Graph showing deviation of ranks made by CWS system and Google search

Later the following measure, 'average deviation' was calculated for each query which denotes how close the system is able to provide the results in the interest of the user.

$$\textit{average deviation} = (_{i=1}\sum^{n}|(1-l_i.rank)|)/n \qquad (5)$$

where l_i denotes the i^{th} relevant link identified for a query, and n is the number of relevant links. Each relevant link l_i identified by participants will be associated with two ranks: EERank which represents the final rank that is determined by the collaborative search system, and DefaultRank, which is the original Google ranking. Average deviation was calculated for both different rankings for each query. Intuitively, a lower average deviation indicates a higher search quality. That is, if ranking made by the participant and the system are same, the average deviation becomes zero, which denotes the best search quality. The table 1 shows the *average deviation* values of CWS system and default Google search for the 10 queries made by Team-B of Group-1.

Table 1: Table showing *average deviation* values of CWS system and default Google search

Group-1 Team-B		
Query no:	Average deviation (CWS ranking)	Average deviation (Google ranking)
1	1	1.8
2	0.87	1.56
3	1.13	1.95
4	0.73	1.5
5	0.77	1.35
6	1.2	1.63
7	1.13	1.52
8	0.6	1.1
9	0.63	0.9
10	0.96	1.42
Average	0.902	1.473

The same procedure was done for each group. Finally, a measure called 'final deviation' was calculated for Team-B of each group based on the all 10 queries made by them using the following formula.

$$\textit{final deviation} = (_{i=1}\sum^{N}(query_i.averagedeviation))/N \qquad (6)$$

where $query_i.averagedeviation$ is the *average deviation* calculated for $query_i$ of that user and N denotes number of queries performed (here it is 10). Here also the final deviation was calculated for collaborative search results and for default Google results.

Table 2: Table showing final deviation of CWS system and default Google search for all 5 groups

Group no:	Final deviation	
	CWS system	Google
1	0.902	1.473
2	0.782	1.257
3	0.642	1.013
4	0.81	1.12
5	0.603	0.97
Average	0.7478	1.1666

The table 2 shows the final deviation for both systems for all 5 groups. The values were analyzed to calculate the performance improvement of the collaborative search system with that default search engine. By taking the average of final deviation produced by both systems, it can be noted that the deviation is reduced by a value of 0:4188 ie, by a percentage of 35.89. It means that the collaborative web search system has given an improvement of about 36% than the default search using search engine. With even better user profile, the performance is expected to be even more.

IV. CONCLUSION

Collaborative Web search is a promising way to improve search quality by users working in cooperation. However, this approach requires a convenient way for users to work together. But current Web browsers and search engines provide limited support for this. A feasible solution through a browser toolbar to combine a Web browser and major search engines like Google, Yahoo is introduced. An approach utilizing users' experiences was proposed for this goal based on a hierarchical user profile. The methodology can be summarized as follows. First, a method is provided to the user for collecting, summarizing, and organizing her search cases into a hierarchical user profile, where general terms are placed to higher levels than specific terms. Through this profile, any user can be taken as an expert for a given topics and search cases under general terms are taken as experts' experiences. In addition, users' experiences were organized into a hierarchical expert profile and recommendation rules were proposed in order to utilize them for Collaborative Web Search with the higher search quality. This project is an exploratory work on the following aspects: First, it explores a way to combine current Web browsers and search engines for collaborative Web Search. Secondly, it deals with unstructured data of various web documents and organize them it into a structured hierarchical format and thirdly it try to define experts' experiences and utilize them to improve the search quality. There are a few of promising directions for future work. In particular, ways of finding right experts and their valuable experiences for a given query from expert-finding system are being considered. Also, it is suspected that personalized collaborative Web search can be achieved if difference of the hierarchical expert profile and the user profile is measured for a specific query.

REFERENCES

[1] Meredith Ringel Morris; "A Survey of Collaborative Web Search Practices "; CHI 2008, April 5-10, 2008, Florence, Italy.
[2] Jingyu Sun, Xueli Yu and Ning Zhong; " Collaborative Web Search Utilizing Experts' experiences"; 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology, pp. 120-127.
[3] Michael B. Twidale and David M. Nichols; "Collaborative Browsing and Visualization of the Search Process".
[4] Meredith Ringel Morris, Saleema Amershi; "Shared Sensemaking: Enhancing the Value of Collaborative Web Search Tools."
[5] Barry Smyth, Evelyn Balfe, Oisin Boydell, Keith Bradley, Peter Briggs, Maurice Coyle, Jill Freyne; "A Live-User Evaluation of Collaborative Web Search". Enterprise Ireland Informatics Initiative.
[6] Anand.S. S. and B. Mobasher, "Intelligent techniques for web personalization," Lecture Notes in Computer Science 3169pp.1-36.
[7] Cheqian Chen, K. L," Personalized search based on learning user click history". Cognitive Informatics (ICCI), 2010 9th IEEE International Conference pp. 490-495.
[8] Spink, A. and B. J. Jansen, "A study of web search trends." Webology 1(2):4.
[9] Gauch, S., M. Speretta, "User profiles for personalized information access.", Lecture Notes in Computer Science 4321: 54
[10] Namita Mittal, Richi Nayak, MC Govil, KC Jain, "A Hybrid Approach of Personalized Web Information Retrieval",2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 105-116, Aug. 2010.