Collaborative User Building Concept based Profile

R.Murugeswari
PG Student
Department of Computer Science &Engineering (PG)
National Engineering College

D.Vijayakumar, B.E., M.S,
Asst. Professor
Department of Computer Science &Engineering (PG)
National Engineering College

ABSTRACT

One of the most promising and potent remedies against information overload comes in the form of personalization. It aims to customize the interactions on a website depending on the user's explicit and /or implicit interests and desires. User profiling is a fundamental component of any personalization applications. In this paper, the focus is on search engine personalization and to develop concept-based user profiling methods. The research results show that the profile which capture and utilize both of the users' positive and negative preferences perform the best by means of p-Click and SpyNB-c method. To improve the quality of information access and infer users' intentions for personalization using concept based user profile, collaborative filtering will be used. Finally, the concept-based user profiles can be integrated into the ranking algorithms of search engine.

Keywords

Positive preference; Negative preferences; clickthrough data; collaborative filtering

1. INTRODUCTION

With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Coping with ambiguous queries has long been an important part of the research on Information Retrieval, but still remains a challenging task. Personalized search has recently got significant attention in addressing this challenge in the web search community, based on the premise that a user's general preference may help the search engine disambiguate the true intention of a query. However, studies have shown that users are reluctant to provide any explicit input on their personal preference. For example, a farmer may use the query "apple" to find information about growing delicious apples, while graphic designers may use the same query to find information about Apple Computer. Personalized search is an important research area that aims to resolve the ambiguity of query terms.

To increase the relevance of search results, personalized search engines create user profiles to capture the users' personal preferences and as such identify the actual goal of the input query and the learned user preferences. Most personalization methods focused on the creation of one single profile for a user and applied the same profile to all of the user's queries. For example, a user who prefers information about fruit on the query "orange" may prefer the information about Apple Computer for the query "apple." Personalization strategies employed a single large user profile for each user in the personalization process. Existing click through-based user profiling strategies can be categorized into document-based and concept based approaches. They both assume that user clicks can be used to infer users' interests, although their inference methods and the outcomes of the inference are different.

On the concept based profiling methods aim to derive topics or concepts that users are highly interested. These two approaches will be reviewed in Section 3. While there are document-based methods that consider both users positive and negative preferences, to the best of our knowledge, there are no concept-based methods that considered both positive and negative preferences in deriving users' topical interests. Most existing user profiling strategies only consider documents that users are interested in (i.e., users' positive preferences) but ignore documents that user's dislike (i.e., users' negative preferences).

In reality, positive preferences are not enough to capture the fine grain interests of a user. Profiles built on both positive and negative user preferences can represent user interests at iner details. Personalization strategies such as [3, 11] include negative preferences in the personalization process, but they all are document-based, and thus, cannot reflect users' general topical interests.

2. MOTIVATION

Most existing user profiling strategies considers only document based methods. The relevant search result is not accurate to infer the users' implicit and explicit interests but in the concept based user methods gives better performance than the document based methods. In case of single user personalization, relevance of search results is not effective results in obtaining the user's explicit and implicit interests. Community based user interest may increase the relevance of search. In item-based method, by identifying similarities between different items, recommendations for users will be computed. The item-based query expansion method provides better performance than the user-based method. Item-based method recommended better expansion terms than the user based method, which is important in helping web users to easily access information needs by formulating qualified queries.

3. RELATED WORKS

Users' browsed documents and search histories are automatically mapped into a set of topical categories. User profiles are created based on the users' preferences on the extracted topical categories. Joachim's [10] proposed a method which employs preference mining and machine learning to model users' clicking and browsing behavior. Joachim's' method assumes that a user would scan the search result list from top to bottom. If a user has skipped a document d_i at rank i before clicking on document d_j at rank j, it is assumed that he/she must have scan the document d_i and decided to skip it. Thus, we can conclude that the user prefers document d_j more than document d_i. Using Joachim's' proposition and the example click through data in Table 1.

Table 1.An Example of Clickthrough for the Query "apple

Links	**Action**	**Search Results**	**Extracted Concepts**
http://www.apple.com/	clicked	Apple computer	Macintosh
http://www.hort.purdue.fruits/		Apple corps	Apple Fruit
http://www.macworld.com/	clicked	Macintosh Products	Macintosh, catalog
http://www.apples.htm/		Apple hill growers	Fruit, apple hill
http://www.info.apple.com/		Apple support	product
http://www.appleinsider.com/	clicked	Apple store	Apple store, Macintosh

More recently, Agichtein et al. [1] suggested that explicit feedback (i.e., individual user behavior, click through data, etc.) from search engine users is noisy. One major observation is the bias of user click distribution toward top ranked results. To resolve the bias, Agichtein suggested cleaning up the click through data with the aggregated "background" distribution.

.Liu et al. [13] proposed a user profiling method based on users' search history and the Open Directory Project (ODP) [16]. The user profile is represented as a set of categories, and for each category, a set of keywords with weights. The categories stored in the user profiles serve as a context to disambiguate user queries. If a profile shows that a user is interested in certain Categories, the search can be narrowed down by providing suggested results according to the user's preferred categories.

Xu et al. [20] proposed a scalable method which automatically builds user profiles based on users' personal documents (e.g., browsing histories and e-mails). The user profiles summarize users' interests into hierarchical structures. The method assumes that terms that exist frequently in user's browsed documents represent topics that the user is interested in. Frequent terms are extracted from users' browsed documents to build hierarchical user profiles representing users' topical interests.

4. GENERATION OF CONCEPT BASED USER PROFILE

A. concept extraction method

After a query is submitted to a search engine, a list of Web snippets is returned to the user.

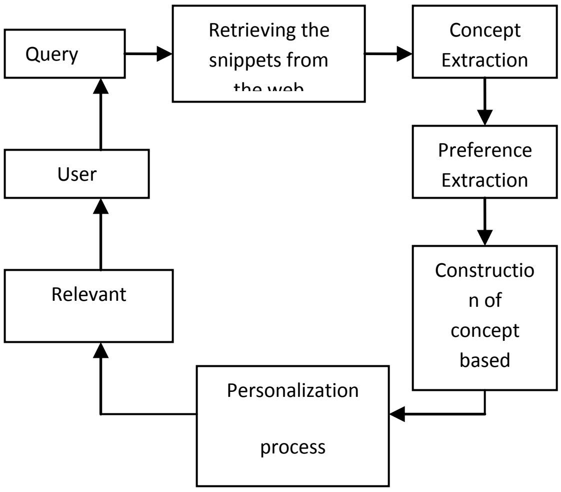

Fig 1: Generation of concept based user profile

Assume that if a keyword/phrase exists frequently in the Web-snippets of a particular query concept related to the query because it coexists in close proximity with the query in the top documents., which is inspired by the well-known problem of finding frequent item sets in data mining to measure the interestingness of a particular keyword/phrase ci extracted from the Web-snippets arising from q:

$$\text{Support}(c_i) = \frac{sf(c_i)}{n} \cdot |c_i|$$

Where $sf(c_i)$ is the snippet frequency of the keyword/phrase ci (i.e., the number of Web-snippets containing ci), n is the number of Web-snippets returned, and $|c_i|$ is the number of terms in the keyword/phrase ci. If the support of a keyword/phrase ci is greater than the threshold s (s = 0.03 in our experiments), we treat ci as a concept for the query q.

Table2. Example Concepts Extracted for the Query "apple"

Concept c_i	**Support(c_i)**
mac	0.1
iPod	0.1
iPhone	0.1
hardware	0.09

Before concepts are extracted, stop words, such as "the," "of," "we," etc., are first removed from the snippets. The maximum length of a concept is limited to seven words. These not only reduce the computational time, but also avoid extracting meaningless concepts.

B. click-based method (pclick)

The concepts extracted for a query q using the concept extraction method discussed in Section 3 describe the possible concept space arising from the query q. The concept space may cover more than what the user actually wants. For example, when the user searches for the query "apple," the concept space derived from our concept extraction method contains the concepts "macintosh," "ipod," and "fruit." If the user is indeed interested in "apple" as a fruit and clicks on pages containing the concept "fruit," the user profile represented as a weighted concept vector should record the user interest on the concept "apple" and its neighborhood (i.e., concepts which having similar meaning as "fruit"), while downgrading unrelated concepts such as "macintosh," "ipod," and their neighborhood. Therefore, the following formulas to capture a user's degree of interest W_{ci} on the extracted concepts W_{ci}, when a Web-snippet S_j is clicked by the user (denoted by click (S_j)):

$$click(s_j) \Rightarrow \forall_{ci} \in s_{j,} w_{ci} = w_{ci} + 1$$

$$click(s_j) \Rightarrow \forall_{ci} \in s_j, w_{cj} = w_{cj} + sim_R(c_i, c_j) if$$
$$sim_R(c_i, c_j) > 0,$$

Where S_j is a Web-snippet, W_{ci} represents the user's degree of interest on the concept C_i and C_j is the neighborhood concept of C_i .When a Web-snippet Sj has been clicked by a user, the weight w_{Ci} of concepts C_j appearing in is incremented by 1. For other concepts C_j that are related to on the concept C_j relationship graph, they are incremented according to the similarity score click-based profile PClick in which the user is interested in information about "macintosh." Hence, the concept "macintosh" receives the highest weight among all of the concepts extracted for the query "apple." The weights W_{ti} of the concepts "mach os," "software," "apple store," "iPod," "iPhone," and "hardware" are increased because they are related to the concept "macintosh." The weights W_{ci} for concepts "fruit," "apple farm," "juice," and "apple grower" remains zero, showing that the user is not interested in information about "apple fruit."

Training the Naive Bayes Algorithm

Input:

$L = \{ I_1, I_2, ...I_N \}$ /∗ a set of links ∗/

Output:

Prior probabilities: Pr (+) and Pr (−);

Likelihoods: $P_r(w_j|-)\forall_j \in \{1,.....,M\}$

Procedure:

1: $$P_r(+) = \frac{\sum_{i=1}^{N} \delta(+|l_i);}{N}$$

2: $$P_r(-) = \frac{\sum_{i=1}^{N} \delta(-|l_i);}{N}$$

3: for each attribute $w_j \in W$ do

4. $$\Pr(w_j|+) = \frac{\lambda + \sum_{i=1}^{N} Num(w_j, l_i)\delta(+|l_i)}{\lambda M + \sum_{k=1}^{M}\sum_{i=1}^{N} Num(w_k, l_i)\delta(+|l_i);}$$

5. $$P_r(w_j|-) = \frac{\lambda + \sum_{i=1}^{N} Num(w_j, l_i)\delta(-|li)}{\lambda M + \sum_{k=1}^{M}\sum_{i=1}^{N} Num(w_k, l_i)\delta(-|l_i)}$$

6: end for

C. click + spyNB-c method

Similar to Click+Joachims-C and Click+mJoachims-C methods, the following formula is used to create a hybrid profile PClick+SpyNB-C that combines PClick and PSpyNB-C:

$w(C+sNB)_{ci} = w(C)_{ci} + w(sNB)_{ci}$,

if $w(sNB)_{ci} < 0, w(C+sNB)ci = w(C)_{ci}$, otherwise ,

$w(C+sNB)_{ci} \in$ Pclick + w(spyNB-C),$w(C)_{ci} \in$ PClick, and $w(sNB)_{ci} \in$ $_{PspyNB-c}$. If a concept ci has a negative weight in PspyNB-C,the negative weight will be added to $w(C)_{ci}$ in PClick forming the weighted concept vector for the hybrid profile $P_{Click+SpyNB-C}$

5. COLLABORATIVE FILTERING

Collaborative filtering (CF) is a very popular technique, especially in commercial applications, for recommending products of some kind to clients. It requires a large database of user data to work properly, when such data exists, it is not difficult to implement. Collaborative filtering can be done in a user-based or item-based form. The user-based form matches the description above: users rate every product, and the filtering process identifies users who have made similar ratings to the user requiring a recommendation. The idea it to combine the ratings of similar users to predict how any given user would rate products he or she has not seen. By giving users access to others' prior experience with an information source, collaborative information filter is created.

Item Based Method

Item-Based method is based on query similarity, not on user similarity. The idea is to recognize relations between items by analyzing the user-item matrix and for a given pair predicate related items based on these relations. In other words, this method first computes similarity between items and then selects the most similar ones. In determination of similarities, Log-likelihood ratio will be used.

Query q;

For each query {

Compute similarity between each q and query

}

6. EXPERIMENTAL RESULTS

In this section, concept based user profiling strategies are evaluated. The click through data together with the extracted concepts is used to create the concept-based user profiles. Joachim's-C, PmJoachims-C, and PSpyNB-C are able to capture users' negative preferences, yield worse precision and recall ratings comparing to PClick. This is attributed to the fact that PJoachims-C, PmJoachims-C, and PSpyNB-C share a common deficiency in capturing users' positive preferences as shown in the fig: 3. A few wrong positive predictions would significantly lower the weight of a positive concept. Although PJoachims-C, PmJoachims-C, and PSpyNB-C are not ideal for capturing user's positive preferences, they can capture negative preferences from users' clickthroughs very well. SpyNB-C produces a more reliable set of negative concepts compared to the others. With a more accurate set of negative preferences, PClick+SpyNB-C achieves better precision and recall results comparing to PClick+Joachims-C and PClick+mJoachims-C.

” Table 3.Feature weights obtained for the query “apple”

Feature	Weight(Joachim’s-C)	Weight(Joachim’s-C)	Weight(spyNB-C)
Entertainment	-0.369	-0.275	-0.029
Traveller	-0.092	-0.030	-0.022
Receipe	-0.333	-0.272	-0.435
Fruit	1.941	1.871	1.765
Farm	2.048	2.629	1.497

PClick achieves a high average similarity value (0.3217) for similar queries, showing that the positive preferences alone from PClick are good for identifying similar queries. PJoachims_C, PmJoachims_C, and PSpyNB_C achieve negative average similarity values (-0.0154, -0.0032, and -0.0059) for dissimilar queries. These methods are good in predicting negative preferences to distinguish dissimilar queries. The wrong positive predictions significantly lower the correct positive preferences in the user profiles, and thus, lowering the average similarities (0.1056, 0.1143, and 0.1044) for similar queries. PClick+Joachims_C, PClick+mJoachims_C, and PClick+SpyNB_C achieve high average similarity values (0.2546, 0.2487, and0.2673) for similar queries, but low average similarities (0.0094, 0.0087, and 0.0091) for dissimilar queries. Both the accurate positive preferences of PClick and the correctly predicted negative preferences from PJoachims-C; PmJoachims-C; and PSpyNB-C:

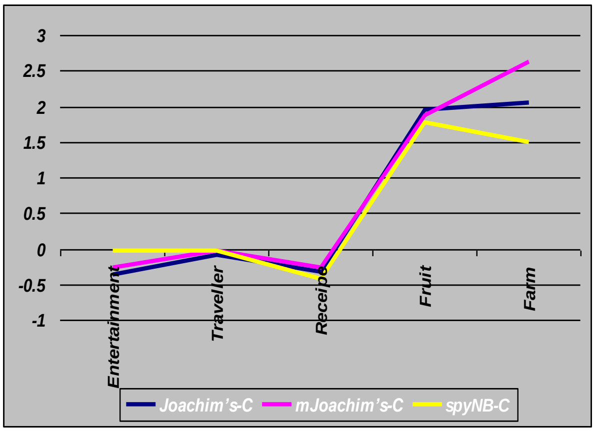

Figure 2: Comparison of clickthrough weights using joachims-C, mJoachims-C, spyNB-C

Thus, PClick+Joachims-C,PClick+mJoachims-C, and PClick+SpyNB-C perform the best among all the proposed user profiling strategies. Accurate positive preferences of PClick and the correctly predicted negative preferences from Joachim’s-C; PmJoachims-C; and PSpyNB-C:

SpyNB-C performs better mainly because it is able to discover more accurate negative samples (i.e., results that do not contain topics interesting to the user). With more accurate negative samples, a more reliable set of negative concepts can be determined. Since the sets of positive samples (i.e., the clicked results) are the same for all of the three methods, the method (i.e., SpyNB-C) with a more reliable set of negative samples/concepts would outperform the others. Thus, PClick + Joachim’s-C, PClick+mJoachims-C, and PClick+SpyNB-C perform the best among all the proposed user profiling strategies as shown in the fig:3. The user profiles are employed to group similar queries together according to users’ needs by the item based collaborative filtering method.

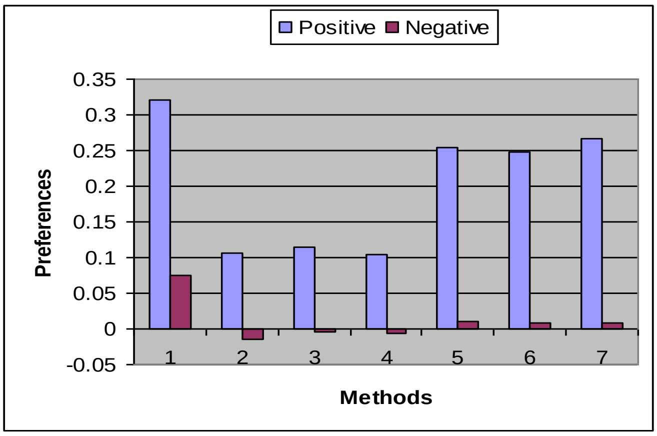

Figure 3: Average Similarity Values for Similar/Dissimilar Queries Computed Using Pclick, Joachim’s-C, PmJoachims-C, PspyNB_C, Pclick+Joachims-C, PClick+mJoachims-C, and PClick+SpyNB-C

1 $\rightarrow$ Pclick

2 $\rightarrow$ $P_{Joachims-C}$

3 $\rightarrow$ $p_{mJoachims-C}$

4 $\rightarrow$ $P_{spyNB-C}$

5 $\rightarrow$ $p_{click+Joachims-C}$

6 $\rightarrow$ $P_{click+mJoachims-C}$

7 $\rightarrow$ $P_{click+spyNB-C}$

7. CONCLUSION

Several user profiling strategies have been discussed. These strategies consider only the positive preferences. It is not enough to capture the fine grain interests of the user for personalization. The above problems by experimental results show that user profiles which capture both positive and negative preferences perform the best profiling strategies studied for single user personalization. Here, relevance of search results is not effective in obtaining the user’s explicit and implicit interests. Community based user interest may increase the relevance of search results. To improve the quality of information access and infer users’ intentions for personalization using concept based user profile, collaborative filtering will be used which allows the users with the similar interests to share their concept based user profiles.

8. REFERENCES

[1] Burcu et.al, “Guided Navigation Using Query Log Mining through Query Expansion”, Anadolu University, Turkey.

[2] Pin-Yu Pan et.al, “Wireless and Mobile Communication Laboratory,” The Development of An Ontology-based Adaptive Personalized Recommender System”, Department of Computer Science and Information Engineering National Cheng Kung University, Taiwan.

[3] Jingqiu et.al, “Personalized Web Search Using User Profile”, College of Computer Science, Chongqing University.

[4] Christos Bouras, “Personalized News Search in WWW: Adapting on user’s behavior”, Professor Research Academic Computer Technology Institute, MsC Research Academic Computer Technology Institute, Greece.

[5] Fang Liu et.al, “Personalized Web Search by Mapping User Queries to Categories”, Department of Computer Science, University of Illinois at Chicago, IL 60607,(312) 996-4881,fliu1@cs.uic.edu.

[6] Thorsten, Cornell University, “Optimizing Search Engines using Click through Data”, Department of Computer Science, Ithaca, NY 14853,USA,tj@cs.cornell.edu.

[7] Mirco Speretta, “Personalizing Search Based on User Search Histories” Electrical Engineering and Computer Science, University of Kansas Lawrence.

[8] Magdalini Teriyaki et.al, “Web Mining for Web Personalization”, Department of Informatics, Athens University of Economics and Business Patision76, Athens.

[9] H.R. Kim, P.K. Chan , “Learning implicit user interest hierarchy for context in personalization”, the 8th international conference on Intelligent user interfaces, Miami, Florida, USA, 2003, pp. 101 - 108.

[10] Chau D. Zeng, H. Chen. “Personalized spiders for web search and analysis”, In Proceedings of the 1st ACM-IEEE Joint Conference on Digital Libraries, pp 79 - 87, 2001.

[11] Ng, L. Deng, and D.L. Lee, “Mining User Preference Using Spy Voting for Search Engine Personalization”, ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.

[12] “Proc. Int’Workshop Current Trends in Database Technology, ”Query Recommendation using Query Logs in Search Engines”, pp. 588-596, 2004.

[13] “Proc. Acm E. Agichtein, E. Brill, And S. Dumais,”Improving Web Search Ranking by Incorporating User Behavior Information”, SIGIR, 2006.

[14] Indu Chawla, “An overview of personalization in web search”, Computer Science”, JIITNoida.

[15] Demetrius Pierrakos et.al, “Personalizing Web Directories with the Aid of Web Usage Data”, IEEE Transactions on Knowledge and Data Engineering, vol.22, No.9, Sep 2010.

[16] Nicola's Mathis, “Personalizing web search using Long Term Browsing History”, feb9-12, 2011, copyright 2011 ACM.

[17] Kenneth Wai -Ting Leung et.al, “Personalized Concept-Based Clustering of Search Engine Queries”, IEEE Transactions on Knowledge and Data Engineering vol.20, No.11, Nov 2008.

[18] Daniela Godoy and Anglia, “User profiling for Web Page Filtering”, IEEE Transactions on Knowledge and Data Engineering vol.20, No.11, August 2005.

[19] Fang Liu et.al, "Personalized Web search for improving Retrieval Effectiveness”, IEEE Transactions on Knowledge and Data Engineering, vol no.16, No.1 Jan 2004.

[20] Kenneth Wai-Ting Leung and Dik Lun Lee,” Deriving Concept-Based User Profiles from Search Engine Logs”, IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 7, July 2010.