Web Usage Mining: Personalization based on User Positive and Negative Preferences

Ravi.D
Research Scholar
Periyar University
Salem

G.M.Nasira, Ph.D
Assistant Professor
Chikkanna Govt. Arts College
Tirupur

ABSTRACT

Most commercial search engines give the same results for the same query, not considering the user's interest. User profiling is a fundamental component of any personalization application. Most existing user profiling strategies are based on object that users are interested in (positive preferences), but not the objects that users dislike (negative preferences). This paper focuses on search engine personalization and develop several concept-based user profiling methods that are based on both positive and negative preferences

Keywords

User Profiling, Personalization, Support, Positive preference, Negative preference

1. INTRODUCTION

Most existing search engine retrieval system could be characterized as "one size fits all" . This means that the Information Retrieval(IR) decision is based solely on the query document matching. The query is the only evidence that specifies the user need and information about the user. Since queries submitted to search engines tend to be short and ambiguous, they are unable to draw the user's precise needs. For example, a user may use ''mouse'' to find information about rodents when the user is viewing television news about a plague, but would want to find information about computer mouse products when purchasing a new computer. Generic search engines are unable to distinguish between such cases

Personalized is the process of presenting the right information to the right user at the right moment. System can learn about user's interests collecting personal information , analysing the information , and storing the results in user profiles. Information can be captured from users in two ways that go explicitly and implicitly, asking for feedback such as preferences or ratings, observing user behaviors such as the time spent reading on online document. Explicit construction of user profiles has several drawbacks.

The user provides inconsistent or incorrect information, the profile created is static whereas the user's interests may change over time, and the construction of the search engine can provide the users with the query results that accord with their interests and backgrounds . In order to achieve this goal, it is first needs to recognize the user's personalized behaviour , then analyses their behaviour and finds their patterns, finally uses the existing resource to match their pattern and puts the personalized information to them. The general architecture of a personalized search system is depicted in Fig.1.

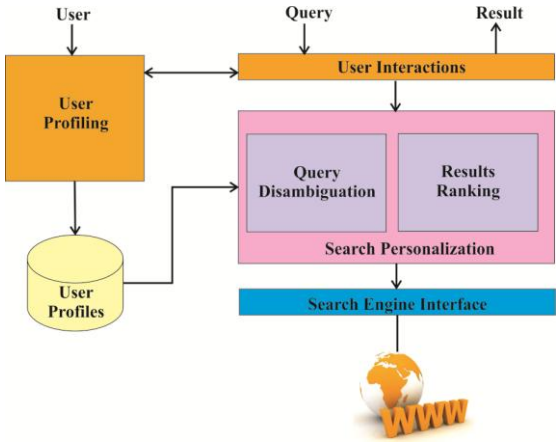

Fig 1: Architecture of a Personalized Search System

A user profile is a collection of personal data associated to a specific user. A profile can be used to store the description of the characteristics of person. This information can be exploited by systems taking into account the persons characteristics and preferences. For instance profiles can be used by adaptive hypermedia systems that personalize the human computer interaction.

A good user profiling strategy is an essential and fundamental component in search engine personalization. Most existing user profiling strategies only consider documents that users are interested in (i.e., user positive preferences) but ignored documents that users dislike(i.e., users negative preferences) In reality ,positive preferences are not enough to capture the fine grain interests of a user.

This paper propose and studies seven concept-based user profiling strategies that are capable of deriving both of the user's positive and negative preferences. The negative preferences improve the separation of similar and dissimilar queries.

2. RELATED WORK

User profiling strategies can be broadly classified into two main approaches that is document-based and concept-based approaches. Document-based user profiling methods aim at capturing users' clicking and browsing behaviours. Whereas concept-based user profiling methods aim at capturing user's conceptual needs. User's browsed documents and search histories are automatically browsed documents and mapped

into a set of topical categories. User profiles are created based on the user's preferences on the extracted topical categories.

2.1 Concept Extraction

The Concept extraction method contains there basic steps : 1) extracting concepts using the web-snippets or query log 2) mining concept relations, and 3) creating a user concept preference profile using the extracted concepts, concept relation, and user's click throughs.

Our Concept extraction method is inspired by the well known problem of finding frequent item sets in data mining. If a keyword or a phrase appears frequently in the web-snippets or query log of a particular query, it represents an important concept related to the query because it coexists in close proximity with the query in the top documents. The support formula for measuring the interestingness of a particular keyword/phrase *ti* with respect of a query *q*:

$$\mathbf{support}(\mathbf{t_i}) = \frac{\mathbf{sf}(\mathbf{t_i})}{\mathbf{n}} \cdot |\mathbf{t_i}|$$

Where n is the total number result returned. s f (ti) is the snippet frequency of the keyword/pharse ti (ie. The number of web-snippets containing ti) and |ti| is the number of terms in the keyword/phrase ti. If the support of a keyword/phrase ti is bigger than the threshold s (support(ti) > s). Then treat ti as a concept for the query q.

2.2 Mining concept Relations

If two concepts from a query q are similar if they coexist frequently in the Web-snippets arising from the query q. Then apply the following well-known signal-to-noise formula from data mining to establish the similarity between term t1 and t2

$$sim(t_1, t_2) = \frac{n.df(t_1 \mathbf{U} t_2)}{df(t_1).df(t_2)} / \log n,$$

Where n is the number of documents in the corpus, df(t) is the document frequency of the term t and df(t1ut2) is the joint document frequency of t1 and t2. The similarity sim(t1,t2) obtained using the above formula always lies between [0,1].

In the search engine context, two concepts tit j could coexist in the following situations 1) ti and tj coexist in the title 2) ti and tj coexist in the summery or 3) ti exists in the title, while tj exists in the summary(or vice versa). Similarities for the three different cases are computed using the following formula

$$sim_R(t_i, t_j) = sim_{R,title}(t_i, t_j) + sim_{R,summary}(t_i, t_j) + sim_{R,other}(t_i, t_j)$$

$$sim_{R,title}(t_i, t_j) = \propto .\log\left(\frac{n.sf_{title}(t_i \mathbf{U} t_j)}{sf_{title}(t_i).sf_{title}(t_j)}\right) \Big/ \log n$$

$$sim_{R,summary}(t_i, t_j) = \propto .\log\left(\frac{n.sf_{summary}(t_i \mathbf{U} t_j)}{sf_{summary}(t_i).sf_{summary}(t_j)}\right) \Big/ \log n$$

$$sim_{R,others}(t_i, t_j) = \propto .\log\left(\frac{n.sf_{other}(t_i \mathbf{U} t_j)}{sf_{other}(t_i).sf_{others}(t_j)}\right) \Big/ \log n$$

3. USER PROFILING STRATEGIES

In this section, three user profiling strategies whish are both concept-based and utilize user's positive and negative preferences . They are

Table 1. User Profile strategies

User Profile	Description
Click-Based	Which capture only Positive preference
Joachims-c	Which capture only negative preference and consider only un clicked page above clicked page
mJoachims-c	Which capture only negative preference and consider only un clicked page both above and below clicked page

TABLE 2.An Example of Click through for the Query "apple"

Doc	Clicked	Search Results	Extracted Concepts
d 1	√	**Apple computer**	**Macintosh**
d 2		Apple Support	Product
d 3		Apple Inc. Official Downloads	Mac os
d 4		Apple Store	Apple store, iPod
d 5	√	**The Apple Store.**	Apple store, Macintosh
d 6		Apple Hill Growers	Fruit, apple hill
d 7		Apple Corps	Fruit
d 8	√	Macintosh Products Guide	Macintosh, catalog

3.1 Click-Based Methods (PClick)

PClick is good in capturing user's positive preferences. 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 click 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 neighbourhood (i.e., concepts which having similar meaning as "fruit"), while downgrading unrelated concepts such as "Macintosh," "iPod," and the neighborhood

3.2 Joachims Methods

Definition 1. Given two retrieved links, di and dj , for a given query q, the pair wise, *di <q dj*, means that the user prefers dj to di with respect to the query q.

Interpretation 1: When a user scans the ranked list of the search results with respect to the query q, if he or she does not click on a link di,but clicks on a lower link dj, where j>i then

this indicates that the user prefers link dj to di. In this case, the reference is identified by the partial order <q, and is denoted as ***di <q dj*** The rationale is that when the user scans the search results from top to bottom, he or she must have observed di and decided to skip it, before he or she clicks on dj.

To exemplify Joachims' algorithm, consider the click through example in Table 2. According to Interpretation all the preference indentified by Joachims algorithm are shown in Table 3

Table 3. Pair wise preferences identified by Joachims' algorithm from the click through data shown in Table 2.

Preferences containing d 1	Preference Pairs containing d5	Preference Pairs containing d8
Empty Set	**d 2 < q d5**	**d 28 <q d8**
	d 3 < q d5	**d 3 < q d8**
	d 4 < q d5	**d 4 < q d8**
		d 6 < q d8
		d 7 < q d8

3.3 mJoachims Methods

Interpretation 2 : Supports di is a clicked link dj is the next clicked link right after di (no other clicked links between di and dj) and dk is any unclicked link between di and dj (i<k<j). When the user clicks on dj, he or she must have observed link dk(k<j) and decided not to click on it. Therefore besides Interpretation 1, the click through also indicates that the user prefers link di and dk. Thus, the additional preferences ***dk <q di*** can be identified

Table 4. Pair wise preferences identified by mJoachims' algorithm from the click through data shown in Table 2**.**

Preferences containing d 1	Preference Pairs containing d5	Preference Pairs containing d8
d 2 < q d1	**d 2 < q d5**	**d 2 <q d8**
d 3 < q d1	**d 3 < q d5**	**d 3 < q d8**
d 4 < q d1	**d 4 < q d5**	**d 4 < q d8**
	d 6 < q d5	**d 6 < q d8**
	d 7 < q d5	**d 7 < q d8**

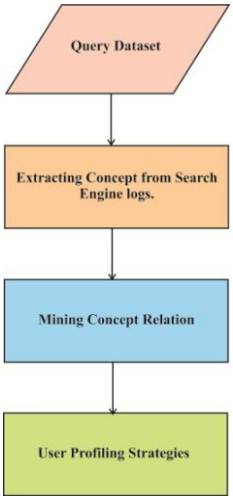

Figure 2 : Process Diagram

4. CONCLUSION

Search Engine personalization and develop several concept based user profiling methods that are based on both positive and negative preferences. The proposed system focus on relationships between users can be mined from the concept based user profiles to perform collaborative filtering. This allows users with the same interests to share their profiles

5. REFERENCES

[1] Di JIANG, Kenneth LEUNG, Wilfred NG and Hao LI. Beyond Click Graph: Topic Modeling for Search Engine Query Log Analysis. International Conference on Database Systems for Advanced Applications DASFAA 2013.

[2] Da YAN, Zhou ZHAO and Wilfred NG. Monochromatic and Bichromatic Reverse Nearest Neighbor Queries on Land Surfaces. ACM Conference on Information and Knowledge Management. ACM CIKM 2012.

[3] Di JIANG, Kenneth LEUNG and Wilfred NG. Context-Aware Search Personalization with Concept Preference. ACM Conference on Information and Knowledge Management. ACM CIKM 2011.

[4] Leung, K.W.-T., Lee, D.L., and Lee, W.-C., Personalized Web Search with Location Preferences, Proc. of IEEE ICDE Conference, Long Beach, USA, 2010.

[5] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.

[6] Open Directory Project, http://www.dmoz.org/, 2009.

[7] W. 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

[8] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. ACM SIGIR, 2006.

[9] M. Speretta and S. Gauch, "Personalized Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, 2005

[10] R. Baeza-yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Workshop Current Trends in Database Technology, pp. 588-596, 2004.

[11] Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-training to Click through Data for Search Engine Adaptation," Proc. Database Systems for Advanced Applications (DASFAA) Conf., 2004

[12] T. Joachims, "Optimizing Search Engines Using Click through Data," Proc. ACM SIGKDD, 2002.

[13] F. Liu, C. Yu, and W. Meng, "Personalized Web Search by Mapping User Queries to Categories," Proc. Int'l Conf. Information and Knowledge Management (CIKM), 2002.

Dr. G. M. Nasira received M.C.A. and M.Phil degree in the year 1995 and 2002 respectively and the Doctorate degree from Mother Teresa Women's University, Kodaikanal in the year 2008. She is having around 19 years of teaching experience in College. Her area of interest includes Artificial Neural Networks, Fuzzy Logic, Genetic Algorithm, Simulation and modelling. She has presented 38 technical papers in various Seminars / Conferences. She is a member of Indian Society for Technical Education (ISTE).

Ravi.D received the MCA degree a from the IGNOU, New Delhi in the year 2003, the M.Phil from the Bhrathidasan University, Trichy in the 2005. the M.E (CSE) in Anna University of Technology, Coimbatore in the year 2011. He having 10 years of Teaching Experience. His area of Interest in Data Mining and SEO. He published 04 National Conference and 03 International Conference. He is a member of ISTE, IACSIT , UACEE and IAENG