International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Deriving Concept Based User Profiles for Search Engine Personalization

**Supriya Vinod Koratkar[1], Sheetal A. Takale[2]**

[1]Computer Engineering Department, Vidya Pratishthan's College of Engineering, Baramati, Pune, India

[2]Information Technology Department, Vidya Pratishthan's College of Engineering, Baramati, Pune, India

**Abstract:** ***Search engine is the vital key role of today's life. Usually users use short and ambiguous terms for searching. Hence it's difficult to get the exact required result. Hence in this approach we have designed a technique that will help to find out the required result. In this method we will first collect the clickthrough data from user and then apply the calculated measures to generate the expected result. After collecting the clickthrough, user's profile is generated which shows the desired result.***

**Keywords:** clickthrough, personalization, profile.

## 1. Introduction

Search engine contains a large amount of miscellaneous data. Hence it is always difficult to extract the relevant information from this huge dataset. Mostly the single short query contains multiple meanings. Such as, a query 'Apple' may contains the information about apple as a computer or apple as an iPod or apple as a fruit or anything else like apple as a toy. Hence at this time to give the exact result to the user is a challenging task. Lets suppose a technician is given a query as 'Apple' then he must be interested to find out the information related with the apple computer or iPod or Apple Company. But if a farmer is giving the same query as 'Apple' then he must be interested in apple as a fruit or apple plant or fertilizers used for apple tree etc.

Hence to distinguish the result as per the interested area, we are fist calculating the user's clickthrough. User's clickthrough is nothing but the concept associated with the web snippet clicked by the user. We are considering the clickthrough because we believe that the user scans the document and clicked on the interested pages. Hence clickthrough is the best way o find out user's interest.

## 2. Related Work

Previous strategies can be distinguish into two approaches,

1) **Document-based approach:** In this approach, user's clicking and browsing behaviour is taken into consideration, it shows that user is interested more in some documents and less in others [1],[2].
2) **Concept-based approach:** this approach considers browsing behaviour and search histories. It tries to find out the concept or topics of user's interest[3],[10].

### 2.1 Click-Based Method ($P_{click}$)

When the use clicks on the web snippet, its degree of interest id calculated for the extracted concept. It find outs concepts which are having similar meaning and the interesting query with its neighbourhood it uses following formula

$$click(s_j) \Rightarrow \forall c_i \in s_j, w_{ci} = w_{ci} + 1$$

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

It calculates the concept space for a particular query and calculates the weighted concept vector to create user profile [4].

### 2.2Joachims-C Method ($P_{joachims\text{-}C}$) :

This method says that, user scans the document from top to bottom and then clicks on the interested document only. Lets suppose document $d_i$ Comes first than document $d_j$ and if user clicks on the document $d_j$, it means that user has gone though the document $d_i$ and decided not to click on it because he is interested in document $d_j$ i.e C($d_j$)<r C($d_i$) where r is the user's preference order.
It uses the feature vector defined by following formula

$$Feature\_c_k = \begin{cases} 1, & if\ k = i, \\ sim_R(c_i,c_j) & sim_R(c_i,c_k) > 0, \\ 0, & otherwise. \end{cases}$$

The target weight vector will be

$$\vec{w} = (w_{Feature\_c1}, w_{Feature\_c2}, ...., w_{Feature\_cn})$$

this is used to create the concept preference profile

$$P_{Joachims-C} = (w_{Feature\_c1}, w_{Feature\_c2}, ...., w_{Feature\_cn})$$

This simply indicates the user's document interest. It uses page ranking algorithm [5].

### 2.3 mJoachims–C Method ($P_{mJoachims\text{-}C}$)

It considers only the unclicked pages. It says that suppose use clicked on a document $d_i$ and then next clicked on the document $d_i$ . But in-between them a document $d_k$ is present (i.e. $i<k<j$). Then again it assumes that user scan it but haven't clicked on $d_k$ .It means this document $d_i$ is less relevant than $d_i$ and $d_j$. Hence these predictions should be combined with prediction of Joachim-c method.[5]

### 2.4 SPY NB-c method($P_{SpyNB\text{-}C}$)

This approach is somewhat different than previous approaches. Instead of considering totally relevant or irrelevant pages, this method considers clicked pages as

positive samples and unclicked pages as unlabeled samples, from which the relevance or irrelevance is find out and later defined as negative samples. This is helpful to find out the clear idea of users interest. It uses spy Naive Bayes technique. Naive Bayes classifier is built by estimating the prior probabilities (Pr(+) and Pr(-)) and likelihoods $Pr(w_j|+)$ and $Pr(w_j|-)$.

## 3. Proposed System

**Personalized concept based query clustering**

Underling idea is based on concepts and their relations extracted from the submitted user queries, web-snippets and extracted data. When a user submits a query, search engine will return the associated search result; important terms from the web-snippets and their relations are online mined to build a concept relationship graph. This graph is first derived without taking use's clickthrough. Then clickthroughs are collected to predict user's conceptual preferences. After then conceptually close queries are find out and return as query suggestion for use's query, for this algorithm, concept relationship graph along with use's conceptual preferences is used as an input [6].

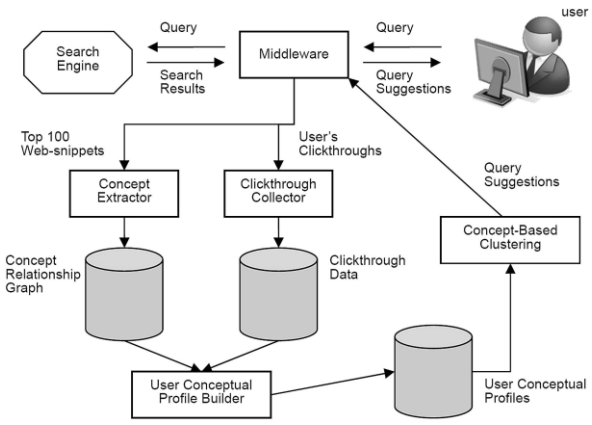


**Figure 1**

The working is shown in fig 1. This can be divided majorly in two parts.i.e Concept Extraction and Concept Based clustering as follow,

### 3.1 Concept Extraction

It is composed of three basic steps 1) extracting concepts using the web-snippets returned from the search engine, 2) mining concept relations, and 3) creating a user concept preference profile using the extracted concepts, concept relations, and user's clickthroughs

#### 3.1.1 Concept Extraction Using Web-Snippets

It is assume that if a keyword or a phrase appears frequently in the web-snippets of a particular query, it represents an important concept related to the query. We use the following support formula for measuring the interestingness of a particular keyword/phrase $t_i$ with respect to the returned web snippets arising from a query q:

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

where, $n$ is total no. Of web snippets returned $sf(t_i)$ is the snippet frequency of the keyword/phrase $t_i$

#### 3.1.2 Mining Concept Relations

To find relations between concepts, we apply a well-known signal-to-noise ratio formula from data mining [7] to establish similarity between terms $t_1$ and $t_2$.

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

where n is the number of documents in the corpus, df($t_1$ U $t_2$) is the joint document frequency of $t_1$ and $t_2$, and df($t$) is the document frequency of the term $t_1$.Therefore, we use the formula for the three different cases in our context as follows:

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

#### 3.1.3 Creating user concept preference profile

The concept relationship graph is first derived without taking user clickthroughs into account. Intuitively, the graph shows the possible concept space arising from user's queries. User's clickthroughs should gradually favor the concept "recipe" and its neighbourhood (by assigning higher weights to the nodes), but the weights of the unrelated concepts such as "iphone," "ipod," and their neighbourhood should remain zero. Therefore, we propose the following formulas to capture user's interestingness $w_{ti}$ on the extracted concepts $t_i$ when a clicked web-snippet $s_j$, denoted by click($s_j$), is found as follows

$$click(s_j) \Rightarrow \forall t_i \in s_j, w_{ti} = w_{ti} + 1$$

$$click(s_j) \Rightarrow \forall t_i \in s_j, w_{tj} = w_{tj} + sim_R(t_i,t_j) \quad if\ sim_R(t_i,t_j) > 0$$

where $s_j$ is a web-snippet, $w_{ti}$ is the interestingness weight of the concept $t_i$ , and $t_j$ is the neighbourhood concept of $t_i$.

### 3.2 Concept Based Clustering

It can be achieved through Personalized Agglomerative Clustering Algorithm [6],[8],[9] as given

**Algorithm:**

| |
|---|
| Input: A Query-Concept Bipartite Graph G |
| Output: A Personalized Clustered Query-Concept Bipartite Graph Gp |
| // Initial Clustering |
| Step 1: Obtain the similarity scores in G for all possible pairs of queries using the noise-tolerant similarity function given in (2). |
| Step 2: Merge the pair of most similar queries (qi, qj) that does no contain the same queries from different users. |
| Step 3: Obtain the similarity scores in G for all possible pairs of concepts using the noise-tolerant similarity function given in (2). |
| Step 4: Merge the pair of concepts (ci,cj)having highest similarity score. |
| Step 5. Unless termination is reached, repeat steps 1-4. |
| // Community Merging |
| Step 6. Obtain the similarity scores in G for all possible pairs of queries using the noise-tolerant similarity function given in (2). |
| Step 7. Merge the pair of most similar queries (qi, qj)that contains the same queries from different users. |
| Step 8. Unless termination is reached, repeat steps 6 and 7. |

## 4. Experimental Result

Here, user has given the query as "orange". So initially he has received all the links related with name orange. This Includes, orange as a colour, orange as a fruit, orange as an amplifier, orange as a company and much more data. Now when user clicks on link orange as a fruit, as shown in fig.(a) his profile is created, containing all the information related with orange as a fruit. Which is shown in fig.(b).

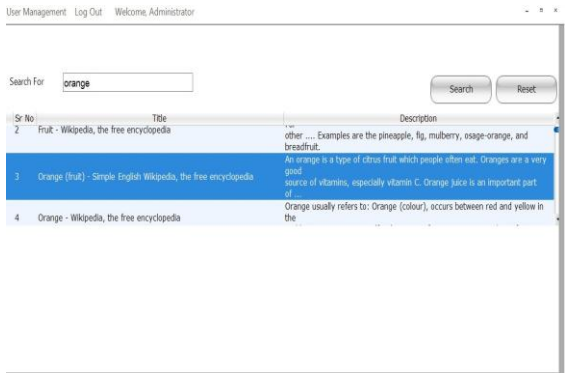


**Figure (a)**

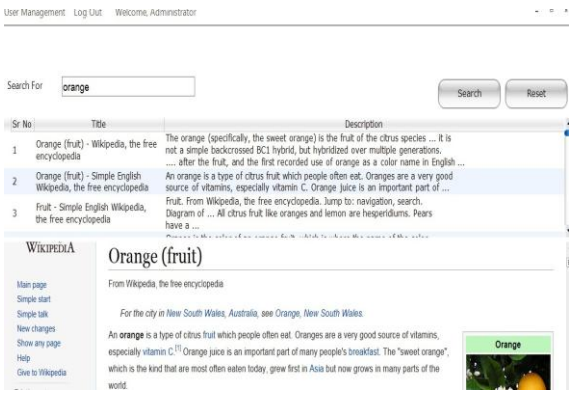


**Figure (b)**

When the user clicked in orange as a colour, his profile is generated containing all the web snippets associated with orange as a colour as shown in fig.(c)

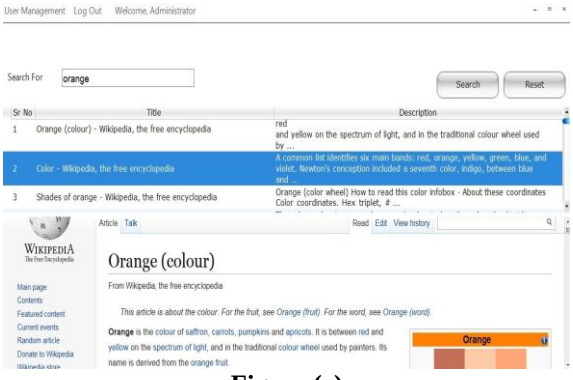


**Figure (c)**

When the user clicked on orange amplifier, his profile is generated containing all the information regarding orange amplifier. As shown in fig.(d).

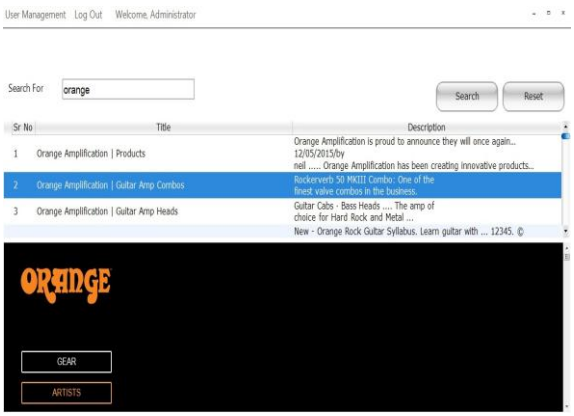


**Figure (d)**

In this way a profile is generated by considering user's interest.

## 5. Conclusion

Personalized profile creation method is more useful to find out the optimal result for the user's queries. It gives most relevant result of what the user is really want to search. It clearly differentiates personalization of search engine and employing the user profiling strategies on it to find an easy way to come up with the closer solution of the user's precise need.

**Future Scope:** we can add user's clickthrough preferences in positive and negative ways to achieve the clear differentiation between interested and not interested queries. i.e. to achieve more precise solution.

## 6. Acknowledgment

I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Prof. S. A. Takale madam for giving her valuable guidance, inspiration and encouragement to embark this paper. Without her guidance and reviewing, this task could not be completed alone.

## References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Rango, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. 29th Ann. Int"l ACM SIGIR Conf. (SIGIR), 2006
[2] E. Agichtein, E. Brill, S. Dumais, and R. Rango, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. 29th Ann. Int'l ACM SIGIR Conf. (SIGIR), 2006.
[3] 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.
[4] Kenneth Wai-Ting Leung and Dik Lun Lee, "Deriving Concept-Based User Profiles from Search Engine Logs" ,vol 22, No 7, july2010
[5] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD, 2002
[6] 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

[7] K.W. Church, W. Gale, P. Hanks, and D. Hindle, "Using Statistics in Lexical Analysis," Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, U. Zernik, ed., Lawrence Erlbaum,1991

[8] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. ACM SIGKDD, 2000

[9] 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.

[10] Y. Xu, B. Zhang, Z. Chen, and K. Wang, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int"l World Wide Web Conf. (WWW), 2007

[11] Personalized approach for user profiling Strategies: A Survey, ijcse, vol 9, 2011