International Journal of Research and Development in Applied Science and Engineering (IJRDASE)

An Approach for Personalization on Web Based Systems

Anshu Srivastava

anshu_qrat@yahoo.co.in

Raghawendra Sharma

robie2007@gmail.com

***Abstract*- Web search engines are one of the most useful and popular online services used today. There are two types of search engines: directory-based and query-based search engines. Normally, directory-based search engines have more complicated interfaces than query based search engines. Moreover, directory-based search engines are challenged by how to automatically classify Web contents efficiently and reduce human effort. Contrarily, query-based search engines have compact interfaces and a structure that is easy to maintain, making them easy to implement and use. However, a traditional search engine faces the difficulty of distinguishing intents of users. The personalization on Web systems is a method to relieve this difficulty. A search engine can be integrated with a user modeling system, by which the likely interests and preferences of users can be utilized for re-ranking search results. This will help users gain easy access to specific site.**

Keywords- CBC, Conventional Search, Query, Search Engine, Web Page.

1 Introduction

The user modeling approach consisting of a user's interest and preference models is based on the premise that frequently visiting certain types of content indicates that the user is interested in that content. The proposed approach can be divided into three steps: monitoring the user's navigational data; using the Web page classification method developed to determine a Web page's content; and employing the Naïve Bayes theory for updating the user's interest model. The same three steps can also be utilized to determine a user's Web preference model. In this work, the Web preference model assigns a value to a Web site (information source) based on the degree to which a user prefers to retrieve information from that Web site.

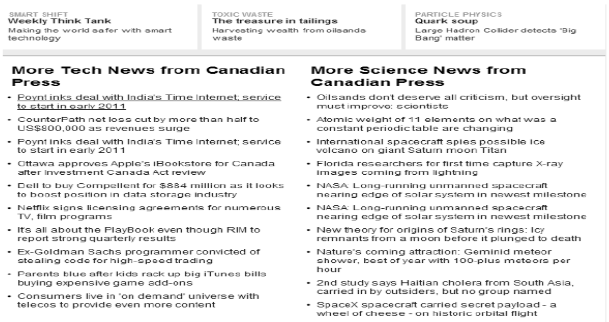
A method of auto-classifying Web pages is employed as a part of the proposed user modeling process. The work employs a well-designed ontology database, WordNet [1],[2], to represent valid terms for the automatic classification method. The *tf–idf* (term frequency–inverse document frequency) method is adopted to determine how important a word is to a document (category), e.g., the weight of the term to the category. First, three components of a Web page are classified separately: classification of meta information, classification of effective content area, and classification by the Web directory service[9] [13]. The classifications of meta information and effective content area are based on the cumulative weight of the terms, while a Web directory service can directly classify a small number of Web pages on the Internet. The classification results from the three components are fused to classify the Web pages into categories. Tracking and recording a user's selections from the search results built up the user's search context model. In each Web search session, the Web pages viewed by a user in sequence represent the user's judgment on the search result of the query. The set of a user's selections in search sessions forms the user's search context model.

2. Conventional Web Search Engines

It is a well known fact that the Web is a huge repertory, consisting of various types of Web sites and online documents. The basic data structure of the Web is the connection between a URL address and an online target, such as a Web site, a Web page, or a digital document [7]. Therefore, the straightforward strategy of retrieving online information is direct navigation that requires a user to input a link address into the browser. This method is often used while a user logs into the frequently visited Web sites. However, this strategy is not efficient for Web sites with complex structure, and not applicable for the Web sites that a user cannot correctly input the URL addresses. Another strategy of retrieving Web information is to click a link to open a Web page. In a Web site where the contents are well-organized and the links to the contents are provided to users, this method is effective and adopted by users. Figure 1 shows an example of retrieving Web information by links in the CBC Web site. However, without the organization of links, it is difficult for users to access Web sites of interest by clicking links in between them. Therefore, an information retrieval tool, which is known as a search engine, is used for users to efficiently obtain the Web information of interest.

A directory-based search engine has three processing steps: collecting information of Web sites, classifying Web sites, and presenting links to Web sites hierarchically. Figure 2 shows the interface of a directory-based search engine. In this type of search engine, Web links are categorized into different topics. Therefore, a user can reach the links of interest by selecting the relevant topics. Along with the rapidly increasing number of Web sites, directory-based search engines have gradually shown their limitations. Firstly, to classify Web sites either increases machine computational burden or requires a lot of manual works. Secondly, it is difficult to present the relevant Web sites to users, because the number of Web sites belonging to a category topic is very large. Thirdly, a directory-based search engine only classifies Web sites and provides links of Web sites, which cannot guide a user directly to the Web pages of interest.

Query-based search engines are more popular than directory-based search engines because they are intuitive, simple to use, and easy to maintain. A query-based search engine can be considered as a tool for retrieving information from a database. The input of a search activity is a set of query terms, while the output is a set of Web links whose targeting Web pages contain the query terms. In the search engine, the search results are presented right after the query terms are submitted



SMART SHIFT
Weekly Think Tank
Making the world safer with smart technology

TOXIC WASTE
The treasure in tailings
Harvesting wealth from oilsands waste

PARTICLE PHYSICS
Quark soup
Large Hadron Collider detects 'Big Bang' matter

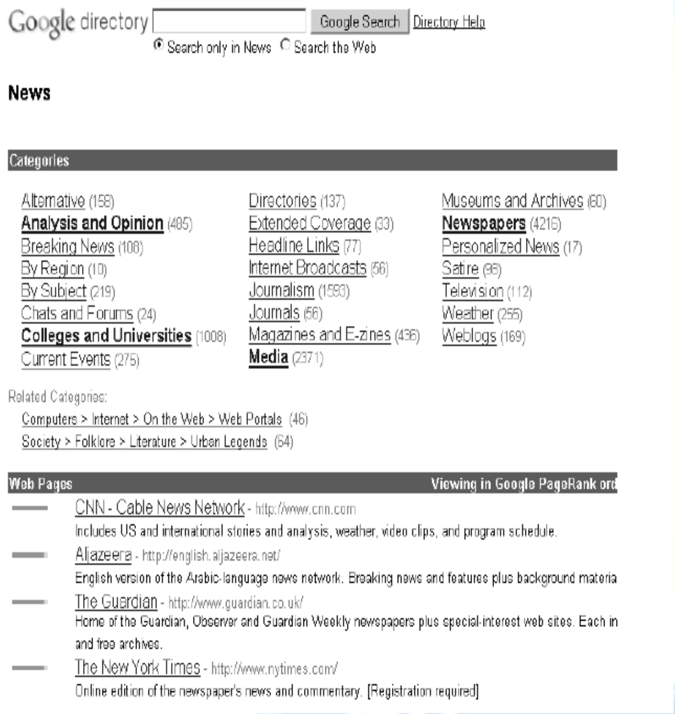
More Tech News from Canadian Press

- Poynt inks deal with India's Time Internet; service to start in early 2011
- CounterPath net loss cut by more than half to US$800,000 as revenues surge
- Poynt inks deal with India's Time Internet; service to start in early 2011
- Ottawa approves Apple's iBookstore for Canada after Investment Canada Act review
- Dell to buy Compellent for $884 million as it looks to boost position in data storage industry
- Netflix signs licensing agreements for numerous TV, film programs
- It's all about the PlayBook even though RIM to report strong quarterly results
- Ex-Goldman Sachs programmer convicted of stealing code for high-speed trading
- Parents blue after kids rack up big iTunes bills buying expensive game add-ons
- Consumers live in 'on demand' universe with telecos to provide even more content

More Science News from Canadian Press

- Oilsands don't deserve all criticism, but oversight must improve: scientists
- Atomic weight of 11 elements on what was a constant periodic table are changing
- International spacecraft spies possible ice volcano on giant Saturn moon Titan
- Florida researchers for first time capture X-ray images coming from lightning
- NASA: Long-running unmanned spacecraft nearing edge of solar system in newest milestone
- NASA: Long-running unmanned spacecraft nearing edge of solar system in newest milestone
- New theory for origins of Saturn's rings: Icy remnants from a moon before it plunged to death
- Nature's coming attraction: Geminid meteor shower, best of year with 100-plus meteors per hour
- 2nd study says Haitian cholera from South Asia, carried in by outsiders, but no group named
- SpaceX spacecraft carried secret payload - a wheel of cheese - on historic orbital flight

Figure 1 An example of retrieving Web information by links in the CBC Web site



● Search only in News ○ Search the Web

News

Categories

Alternative (158)
Analysis and Opinion (485)
Breaking News (108)
By Region (10)
By Subject (219)
Chats and Forums (24)
Colleges and Universities (1008)
Current Events (275)

Directories (137)
Extended Coverage (33)
Headline Links (77)
Internet Broadcasts (56)
Journalism (1593)
Journals (56)
Magazines and E-zines (436)
Media (2371)

Museums and Archives (60)
Newspapers (4216)
Personalized News (17)
Satire (96)
Television (112)
Weather (255)
Weblogs (169)

Related Categories:

Computers > Internet > On the Web > Web Portals (46)
Society > Folklore > Literature > Urban Legends (64)

Web Pages — Viewing in Google PageRank ord

CNN - Cable News Network - http://www.cnn.com
Includes US and international stories and analysis, weather, video clips, and program schedule.

Aljazeera - http://english.aljazeera.net/
English version of the Arabic-language news network. Breaking news and features plus background materia

The Guardian - http://www.guardian.co.uk/
Home of the Guardian, Observer and Guardian Weekly newspapers plus special-interest web sites. Each in and free archives.

The New York Times - http://www.nytimes.com/
Online edition of the newspaper's news and commentary. [Registration required]

Figure 2 The interface of the Google directory-based search engine

3. Personalized Web Search Engine

This research attempts to identify the relationship between a user's Web search behaviors and his or her interest and preference models, so that a personalized Web search method can be developed based on this relationship. In other words, based on analyzing a user's search behaviors, the system will determine how the interests and preferences of the user influence the search results. The system will then provide a personalized search solution for the user based on that influence.

4. Proposed Personalized Search Ranking

A user's interest and preference models are built and updated based on his or her long-term navigational data. For each user's very first search query, the results presented to the user are the exact duplicates of one generated by the conventional search engine. Subsequently, the vector of Web pages viewed by a user is compared with the vector of Web pages resulting from the Web search engine. The user's search behavior data is utilized to determine how the user's interest and preference models impact his or her personal judgment on the given Web pages. The re-ranked results will then be presented to the user and the user's click-through is tracked. Any ranking change from the given search results to the user's context model indicates the dissimilarity between the search engine's ranking algorithm and the user's judgment. For example, if the vector of Web pages mr (=($mr(1)$, $mr(2)$, $mr(3)$, …, $mr(lw)$)) is the search result of a user's rth search session, and the user's search context vector in this session is ($mr(3)$, $mr(1)$), e.g., the user

viewed Web pages $mr(3)$ and $mr(1)$ successively, then there are two instances of ranking jump-over observed, which are $mr(3)$ over $mr(2)$, and $mr(3)$ over $mr(1)$. The pseudo code of determining the total number of instances of ranking jump-over from the search result to the user's context model is given in the next page:

```
read user's context vector $v=(v(1), v(2), \dots v(lo))$,
read search result $m$,
for $i=1, lo, i++$
    find $n$, that $m(n)=v(i)$,
    for $j=1, n-1, j++$
        if $m(j)$ is prior than $v(i)$ in the user's context model then
            no ranking jump-over,
        else
            ranking jump-over is determined,
            $Sumjumpover= Sumjumpover+1$,
        end if
    end for
end for
```

The next step is to utilize the ranking jump-over to determine the influence of factors in a user's interest model and preference model. Using the same example discussed above, the Web pages involved in ranking jump-over can be mapped to categories: $Cat(mr(1))$, $Cat(mr(2))$, and $Cat(mr(3))$, by using the classification method. If the user's interest model shows that the user's degree of interest in $Cat(mr(3))$ is higher than the degree of interest in $Cat(mr(1))$, then the ranking jump-over of $mr(3)$ versus $mr(1)$ is considered as the case that the user's interest model influences his or her search behavior. After taking into account the ranking jump-overs in all search sessions, the influence factor of a user's interest model can be determined statistically:

$$\eta 4=(\Sigma \textit{jump-overs influenced by a user's interest})/(\Sigma \textit{jump-overs}) \quad (1)$$

The same method can also be applied to determine the influence factor of a user's preference model:

$$\eta 5=(\Sigma \textit{jump-overs influenced by a user's preference})/(\Sigma \textit{jump-overs}) \quad (2)$$

Since the conventional search engines only provide ranking results without the appearance rates of Web pages, an inverse distance measure is used to estimate the rates in a search session (Teevan *et al*., 2010), which is formulated as:

$$ER(m(k))=1/(\alpha+\beta\log(k)) \quad (3)$$

where $ER(m(k))$ is the estimated rate of the *kth* Web page in a search result; α is the initial coefficient which is set to 2 in this work; and β is the attenuation coefficient which is set to 1. The choice of the coefficients is taken into account for the relatively smooth decreasing and dominance of the top ten ranks, which is shown in Figure 3.

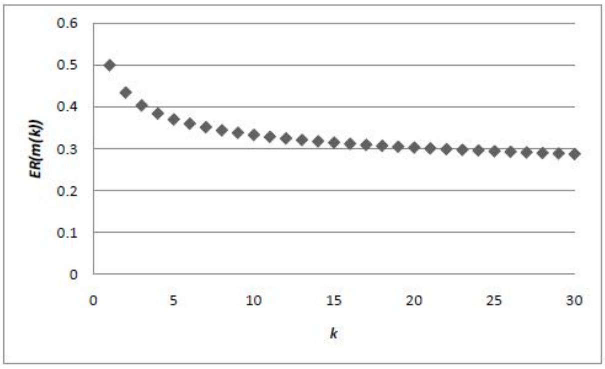


Figure 3 The rate estimation of the Web pages in a search result

Once a user's interest model, preference model, and influence factors have been determined, a personalized search rank algorithm can be applied, which is

$$RR(m(k)) = ER(m(k))(1+\eta 4\,UserInterest(Cat(m(k))))(1+\eta 5\,UserPreference(IS(m(k)))) \quad (4)$$

where *RR* represents the re-rating function; *UserInterest* calls equation to retrieve the user's degree of interest in category $Cat(m(k))$; $IS(m(k))$ is the function of identifying the information source (Web site) of Web page $m(k)$; and *UserPreference* calls equation to retrieve the user's degree of preference on Web site $IS(m(k))$. Ultimately, the web pages are arranged by their rerating value $RR(m(k))$, e.g., the Web page with the maximum re-rating value is listed at the top. Therefore, the personalization of a search engine is implemented by re-ranking the Web pages based on user models.

5. Experimentation and Evaluation

In this work, a user's click-through in a search session is considered as the user's search context model. In order to clarify the construction of a user's search context model. The purpose of developing a personalized search system is to change the search ranking results based on a user's models, so that the user can retrieve information of interest more effectively. From the results of re-ranking, it is shown that Web pages that are relevant to a user's topics of interest like "Computers", and a user's preferred Web site like "Google", are rearranged to the top of the list.

In order to evaluate the proposed personalized Web search method, 12 participants were asked to surf the Internet and conduct Web searches for their daily requirements under the system's supervision. All participants have at least a Bachelor's degree. The participants' interest, preference, and context models were constructed during their Web navigation. In order to alleviate the computational burden, the top 30 Web pages are reranked by the personalized search system. Even though the explicit measurements of relevance can clarify the precision rate and recall rate judged by the users, an implicit measurement is applied to evaluate the performance of the proposed personalized Web search method. Due to the fact that a user's experience of a practical search is the greatest concern, the implicit measurement approach is chosen. Table 5.3 shows the comparison of the users' average first-click on the personalized search results and the conventional search results tracked in 100 practical search sessions. The average first-click of the participants on the personalized re-ranking Web pages is 3.09, while it is 3.80 when the clicked Web pages are mapped to the conventional search results. Therefore, the average improvement of the first-click provided by the personalized re-ranking method is 0.71.

In order to check the statistical significance of the results shown in Table 1, a paired *t*-test (Box *et al*., 1978) is conducted based on the 12 participants' average rates of first-click. Denote participant *i*'s average rates of first-click on the personalized and conventional search results by $x1,i$ and $x2,i$, respectively. There are three steps for carrying out the paired *t*-test:

Table 1 Users' average first-click on the personalized search results and the conventional search results

User ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
Convention search	3.41	4.13	2.98	3.31	5.05	3.54	4.91	2.89	4.21	4.07	3.32	3.79	3.80
Personalized Search	2.56	3.62	2.60	2.71	3.88	2.85	4.21	2.70	2.96	3.25	2.89	2.94	3.09
Improvement	.85	0.51	0.38	.60	1.17	0.69	0.70	0.19	1.25	0.82	0.43	0.85	0.71

By conducting the paired *t*-test, the observed difference of the mean value between the personalized and conventional search models in Table 5.3 is found to be statistically significant having t=7.83 and p<0.01.

In order to evaluate the performance of the proposed system based on a user's overall click-through, the method of Discounted Cumulative Gain (DCG) is employed in this work (Jarvelin and Kekalainen, 2000). Considering that the higher ranked Web pages have higher scores in the evaluation, the *ith* Web page in a search result set can be scored by:

$$Score(i) = 1/\log2(1+i) \quad (5)$$

During a search session, a user will click the Web pages if they think it is relevant to the query. Therefore, the ranking-efficiency can be obtained by weighing the user's click-through in a search session. Equation (5.9) presents a measurement of ranking efficiency for a user's search session:

$$E = (\Sigma(Score(i)Click(i)))/\Sigma Click(i) \quad (6)$$

where *Score*(*i*) is the score of Web page *i* in a user's search session as defined in Equation (5); and *Click*(*i*) indicates if the user opens the *i*th Web page. *Click*(*i*) is 1 when the user opens Web page *i*, otherwise *Click*(*i*) is 0.

Table 2 shows the comparison of the average ranking-efficiency of both the personalized and conventional searches through the investigation of the participants' 100 rounds of search sessions. The average ranking-efficiency of the conventional search engine based on the participants' search behaviors is 0.44. The personalized search method improves the ranking-efficiency to 0.52. Therefore the ranking-efficiency of the personalized search is 18% higher than the conventional search. Moreover, there is no obvious regression of the ranking-efficiency observed on any participant's search data, which implies the generality of the proposed re-ranking method.

Table 2 Average ranking-efficiency of the personalized search and the conventional search

User ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Average**
Convention search	0.51	0.42	0.52	0.44	0.36	0.43	0.31	0.53	0.36	0.43	0.45	0.46	0.44
Personalized Search	0.58	0.51	0.63	0.49	0.49	0.51	0.41	0.57	0.46	0.49	0.51	0.50	0.52
Improvement	0.07	0.09	0.11	0.05	0.13	0.08	0.10	0.04	0.07	0.06	0.06	0.04	0.08

6. Conclusion:

A novel method for a personalized Web search system has been proposed in this paper. A Web page auto-classification method is utilized to analyze a user's navigational data, which helps to build and update the user's interest model. The user's preference model is constructed by analyzing the information sources (Web sites) of the user's navigational data. The user's click-through during a search session is interpreted as the user's search context model which is used to determine how the user's interest and preference models influence his or her search behavior. Finally, the personalized re-ranking algorithm is implemented by recalculating the score of a Web page based on the user's interest and preference models. The experimental results show an obvious improvement of the search ranking provided by the proposed personalized search method. The proposed system utilizes both content-based and behavior-based modeling approaches for the Web search personalization.

Reference:

[1] Aggarwal, C.C., S.C. Gates, and P.S. Yu (2004). On using partial supervision for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 245–255.

[2] Miller, G.A. (2009). WordNet – about us. *WordNet*, Princeton University, http://wordnet.princeton.edu

[3] Bernstein, M.S., D. Tan, G. Smith, M. Czerwinski, and E. Horvitz (2010). Personalization via friendsourcing. *ACM Transactions on Computer-Human Interaction*, vol. 17, no. 2, 6:1-6:28.

[4] Bobadilla, J., F. Serradilla, and J. Bernal (2010). A new collaborative filtering metric that improves the behavior of recommender systems. *Knowledge-Based Systems*, vol. 23, no. 6, pp. 520-528.

[5] Bomhardt, C. (2004). NewsRec, A SVM-driven personal recommendation system for news Websites. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Beijing, China, pp. 545-548.

[6] Choi, S.H., Y.-S Jeong, and M.K. Jeong (2010). A hybrid recommendation method with reduced data for large-scale application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 5, pp. 557-566.

[7] Godoy, D., S. Schiaffino, and A. Amandi (2010). Integrating user modeling approaches into a framework for recommender agents. *Internet Research*, vol. 20, no. 1, pp. 29- 54.

[8] Laroche, M. (2011). New developments in modeling Internet consumer behavior: Introduction to the special issue. *Journal of Business Research*, vol. 63, no. 9-10, pp. 915-918.

[9] Li, L., W. Chu, J. Langford, and R.E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, USA, pp. 661-670.

[10] Miller, G.A. (2012). WordNet – about us. *WordNet*, Princeton University, http://wordnet.princeton.edu. (Accessed in March, 2015.)

[11] Statistics Canada (2010). Canadian Internet use survey. Available online: http://www.statcan.gc.ca/daily-quotidien/100510/dq100510a-eng.htm. (Accessed in March, 2014.)

[12] Teevan, J., S.T. Dumais, and E. Horvitz (2010). Potential for personalization. *ACM Transactions on Computer-Human Interaction*, vol. 17, no. 1, 4:1-4:31.

[13]. Anshu Srivastva, et. al., "KDD based System Making Information Assessment Sustainable for Web Application" International Journal of Research and Development in Applied Science and Engineering, Volume 4, Issue 1, November 2013.