Implicit User's Browsing Behavior Patterns for Filtering Irrelevant Pages in Web Search

Saravanakumar C, Sendhilkumar S, Geetha T.V

Department of Computer Science and Engineering
Anna University, Chennai- 600 025, India
saravanacsk@gmail.com, {thamaraikumar, tvgeedir}@cs.annauniv.edu

Abstract. The World Wide Web is a dynamic environment. It becomes very difficult to identify the users' interests while searching for some information need through the World Wide Web. Automatic inference of user's interests and the context of search are very important for recommending pages that are the direct answers to the user's information need. Graph-based modeling has emerged as a powerful abstraction capable of capturing many of the spatial and temporal characteristics of users during web search. This paper models every search conducted by multiple users through the web as a User's Search Behavior (USB) graph and predicts various graph patterns that exist in the USB graph. The meaning conveyed by such patterns with implicit feedback like page view time is utilized for predicting the user's interest on a page. The semantics of such patterns are used to identify both the relevant and irrelevant pages for a search query. The identified irrelevant pages can then be filtered while recommending pages to the users.

Key words. Web Search, Personalization, User Search Behavior Graph.

1. Introduction

World Wide Web (WWW) contains huge data that are both structured and unstructured and it is a dynamic environment as the data and the user changes frequently. In such a dynamic environment, the task of finding desired information quickly and exactly becomes crucial for the web surfers and tracking user's search behavior is also difficult for the search engines. Personalized web search is a popular remedy to this problem.

There are various factors that affect personalization, among them, page hit-count, which indicates frequency of page visits during a session, has been traditionally considered as an informative indicator of the user preferences. Temporal features such as page view time are also of significant concern, especially in the context of web personalization applications. The context of a search can be derived from the terms used in a search query. Individual search behaviors like, pages visited, order of visit and the actions performed on a visited page are the implicit indicators of the context of the search. In addition, order or sequence of page accesses is another important piece of information.

Advances in Computer Science and Engineering.
Research in Computing Science 34, 2008, pp. 249-260

Received 23/03/08
Accepted 26/04/08
Final version 04/05/08

All the above mentioned factors are content based and hence help to improve search through the WWW. Utilization of content factors of personalization will produce an effective personalization and this is the main motivation of the work explained in this paper.

Graphs are the mostly popularly used representation of the web data. To analyze graph structured data and to uncover important properties/patterns in graphs play an important role in the field of web mining. Hence, this paper models every user's search through the WWW as a graph by using a specially designed browser that captures all the user's actions and saves it in a database. The captured user data is analyzed for implicit information like queries used by the users, page view time, page hit counts, actions performed on pages visited and search paths leading to relevant/ irrelevant information. Such implicit information reflects the interests and search behavior of the user. The user's search behavior graph is then scanned for the existence of various sub graphs like linear paths and closed walks that convey the browsing behaviors of the user as well as the importance of the search paths (thereby the user's interests on the pages that constitute the search paths) traversed by users.

The rest of the paper is organized as follows: Section 2 describes the various works related with the one described in this paper. Section 3 is the core work of this paper that explains the USB graph construction and pattern identification. The experiments conducted, results and their analysis are explained in section 4. Finally section 5 gives the conclusion and future works.

2. Related Works

Oard and Kim [1] classification of observable implicit indicators as shown in table 1 that help to identify the user's interests when observed from the user's browsing behaviour were utilized in this work.

Table 1. Classification of behaviours for implicit indicators

Category	Observable behaviour
Examination	Page-view time, Scrolling Time, Selection (Content of a page)
Retention	Save, Print, Copy & Paste, Bookmark
Reference	Clicking the available links in a page

Examination refers to behaviors indicating that the user is examining an object (such as reading time), retention refers to behaviours indicating intention of future use of object and reference is a type of behaviour indented to link the relevant object to other objects. Scrolling time alone is not used in this work.

Menczer [2] utilized the relationships between content, link, and semantic topology in the Web to rank hits in response to user queries. However, this combination resulted in both false positives and false negatives because of many local optima. The connection between Web lexical and link cues, and between either of these and semantic characterizations of pages, has been previously studied in the context of hypertext document classification, topic distillation and navigation [3,4,5].

Web page ranking algorithms is based on text content analysis and/or link analysis. But most of the ranking algorithms are link-based, among which are HITS and Page-Rank. HITS algorithm, developed by Kleinberg [6], constructs a graph G-containing a set of Web pages relevant to the query, then expands this set with its in-links and out-links. Analyze the link structure of G to find the set of hub pages and the set of authority pages in an iterative fashion. Because the computation of HITS algorithm is carried out at query time, efficiency is a practical concern. Larry Page and Sergey Brin propose the best-known ranking algorithm PageRank [7, 8], which is adopted in the most successful search engine Google. Computation of PageRank is off-line, at background, and the computation involves the whole portion of the Internet (indexed by Google).

The Query Sensitive Self Adaptable Web Page ranking algorithm [9] a voting mechanism is involved. For each list of inverse table, the top 'n' Web pages are selected to construct a voting set. The construction of voting set is static one. The query sensitive algorithm is uncertain that all existing web pages remain unchanged throughout the time.

The UCAIR search toolbar developed by Shen et.al., [10] is a web browser plug-in that functions as a client-side personalized search agent. The UCAIR can collect implicit feedback information from a user and exploit such information to improve retrieval accuracy for this user without requiring any additional effort from the user. But the disadvantage in all the above mentioned works is that it does not provide any relation between the search queries and the content that is viewed. Such conceptual relations provide the context of search and can be used to recommend pages that are the direct answers to the user's information need.

Personalized Web Search for Individual User, by Sendhilkumar S and Geetha T.V [11] supports user searches by using a new search index called the User Conceptual Index (UCI) that highlights the conceptual relation between the search queries and the pages visited by the user. The UCI is computed based on the temporal features like page view time and query usage time and spatial features like frequency of page visit and query usage. However the results of the experiments conducted by Sendhilkumar S and Geetha T.V were affected by false positives and false negatives due to high page view time and greater hit-count for some irrelevant pages that were visited unknowingly by the user. Hence the UCI based re-ranking must be supported by other factors of personalization like the path through the web that the user used to reach his/her information need, the pages involved in such paths, the links that were made available to the user, and the various actions performed.

The aim of this work is not to develop a new search engine, instead utilize the factors of personalization like the path through the web that the user used to reach his/her information need, the pages involved in such paths, the links that were made available to the user, and the various actions performed to identify the importance of a page. This work also visualizes user's search behaviour as a graph and scans through the graph for various graph patterns which when analysed provides information about the pages that are irrelevant to the user's search context.

3. User Search Behavior Graph

A specially designed browser was used to search through the web. The user issues the search query in the browser which redirects the query to any user specified, existing search engine like Google. The search results returned by the search engine are provided to the user through the browser.

Some of the interesting features of the browser are as follows:

1. Privacy is preserved by providing anonymous User names
2. The browser keeps track (implicit data collection) of the various user data like the search queries issued, the pages visited and the page-view time.
3. The user data are automatically updated into a user database.
4. User actions on a page like SAVE, COPY, PRINT AND E-MAIL are tracked by the user.

Every search query is represented using specialized structure called transactions and every transaction is visualized as a graph.

Definition 1 (Transaction): A transaction $T_i = (SQ_i , P_i)$ is a set of pages $P_i = \{p_1, p_2, ...\}$ visited for a search query SQ_i.

Definition 2 (User Search Behavior Graph): A User Search Behavior graph *USB = (P, L) is a mathematical structure* consisting two sets P and L. The elements of P are called vertices/pages visited by the user for a search query (SQ_i) and represented as $P = \{p_1, p_2, ...\}$ such that every $P \varepsilon T$. The elements of L are called the labeled links/edges represented as $L = \{l_1, l_2, ...\}$, such that each link L_k is identified with a pair (p_i, p_j) of pages and the labels indicate the order of visit.

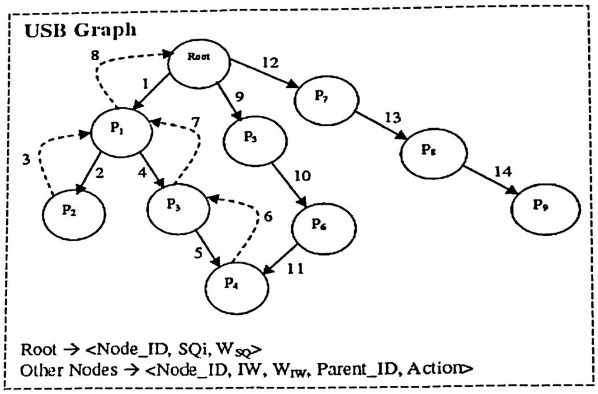

Figure 1. User's Search Behavior (USB) graph

The geometric representation of the USB graph is given in figure 1 and the 'Root' node represents the search query and the initial search result page. The descendents of the 'Root' represents the pages viewed by the user. The dashed arrows indicate the user's movement from the current page to the previously viewed page (backtracking using the 'back' button). The 'Root' node in the USB is represented using a **triple** <*Node ID, SQi,* W_{SQ}>, where SQ_i is the Search Query, W_{SQ} is the average SQ_i weight , *Node_ID* is represented as a combination of <*Page_ID, Depth, Offset*>, where *Page_ID* is a unique ID given to each page that is visited by the user, *Depth* represents the level and *Offset* is the node's position in the i^{th} level and it is {0,0} for the root node. The remaining nodes are represented using a **quintuple** <*Node_ID, IW,* W_{IW}*, Parent_ID, Action*>, where *IW* is the index word(s) for a page, W_{IW} is the average weight for the index words, *Parent_ID* is the Page_ID of the parent node and *Action* indicates the action performed on that page.

The following two hypotheses were proposed based on the USB graphs:

Hypothesis 1: *A linear path in the USB graph indicates that the page(s) that participate in the linear path is relevant to the search context provided atleast one user-action must have taken place through that path.*

Hypothesis 2: *Closed walks like dipoles and cycles in a USB graph indicate that the page(s) that participate in the closed walk are irrelevant to the search context provided no user action is performed on them.*

To prove the above two hypotheses the USB graphs were analyzed to spot the various graph patterns. For predicting the patterns hidden in the USB graphs a matrix representation of a graph and graph search algorithm was required. An adjacency matrix was used for representing the USB graph and a Depth First Search (DFS) was implemented to detect the various graph patterns.

3.1. Matrix Representation of USB

A *USB* graph with *n* pages/vertices and *e* links/edges is represented using an adjacency matrix $USB_{incidence} = [a_{ij}]$, which is a n x n square matrix where n correspond to the n-pages visited for s search query SQ_i, such that $a_{ij} = 1$, if the user has visited the j^{th} from i^{th} page and $a_{ij} = 0$, otherwise. An adjacency matrix $\boldsymbol{USB_{Adj}}$ of the USB graph given in figure 1 is shown in figure 2.

A Depth First Search (*DFS*), given in figure 3, was performed on the *USB* graphs to identify the various patterns that exist in them. DFS algorithm can be used to solve following problems:

1. Testing whether graph is connected.
2. Computing a spanning forest of graph.
3. Computing a path between two vertices of graph or equivalently reporting that no such path exists.
4. Computing a cycle in graph or equivalently reporting that no such cycle exists.

Given the adjacency matrix $\boldsymbol{USB_{Adj}}$ of a *USB* graph, the DFS algorithm will identify the various patterns (linear paths and cycles) that exist in the *USB* graph.

Later such patterns are analyzed using various factors of personalization inferring the semantics they convey.

USB_{Adj} =

	Root	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Root	0	1	0	0	0	1	0	1	0	0
P_1	1	0	1	1	0	0	0	0	0	0
P_2	0	1	0	0	0	0	0	0	0	0
P_3	0	1	0	0	1	0	0	0	0	0
P_4	0	0	0	1	0	0	0	0	0	0
P_5	0	0	0	0	0	0	1	0	0	0
P_6	0	0	0	0	1	0	0	0	0	0
P_7	0	0	0	0	0	0	0	0	1	0
P_8	0	0	0	0	0	0	0	0	0	1
P_9	0	0	0	0	0	0	0	0	0	0

Figure 2. Adjacency Matrix

DEPTH FIRST SEARCH

Input: n x n Adjacency matrix USB_{Adj}
Output: Pages that constitute *LINEAR PATH* or *CYLCE*

```
Step 1:   List_unvisited = {P1,..., Pn}; List_visited = Ø; List_Adjacency = Ø
Step 2:   For every page Pi in List_unvisited
Step 3:       If Pi != visited then
Step 4:           List_visited ← Pi;
Step 5:           For j = 1 to n Do Step 6 to 8
Step 6:               If (USB_Adj (i,j)==1 )then
Step 7:                   List_Adjacency ← Pj
Step 8:                   j = j + 1
Step 9:                   List_unvisited ← List_Adjacency
Step 10:          For all pages P in List_unvisited do Steps 3 to 9
Step 11:              Print All Pages in List_visited ← 'LINEAR PATH'
Step 12:          i = i + 1
Step 13:      Else
Step 14:          Print All Pages in List_visited ← 'CYLCE'
Step 15:  End
```

Figure 3. Depth First Search (DFS) Algorithm

The various patterns (linear paths and cycles) that were identified are listed and explain the next section.

3.2. Patterns Identified in USB Graphs

The patterns that were identified can be broadly classified into two categories: 1) open walk and 2) closed walk. Linear paths constitute the open walk and patterns like dipoles, dipole series and cycles constitute the closed walk.

Definition 3 (*Path Sub Graphs*): A sub graph USB_{PG} of a *USB* graph is called as a *Path Graph* with $|P_p|=|L_p|+1$ if it does not contain cycles and can be drawn so that all of its vertices, P_p and edges, L_p lie on a single straight line. A path sub graph USB_{PG} of the *USB* graph shown in figure 1 is given in figure 4 (a).

Definition 4 (*Dipoles*): A sub graph consisting of two pages (vertices) and n links joining them is called a dipole and is denoted D_n. A dipole with 2 links/edges D_2 is shown in figure 4 (b).

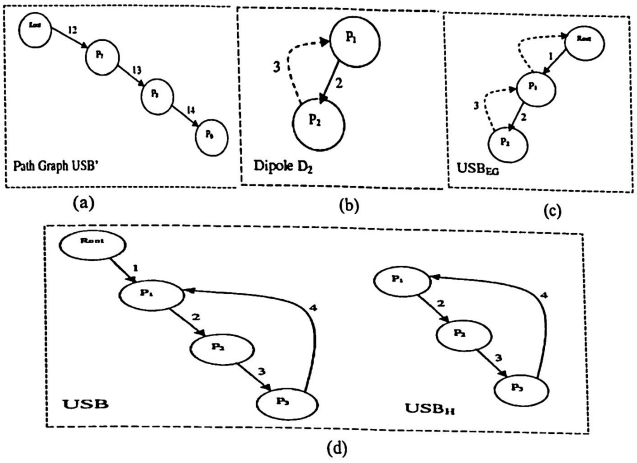

Figure 4. Graph Patterns - (a) Path Sub Graph, (b) Dipole, (c) Dipole Series, (d) Cycle

Definition 6 (*Dipoles Series*): A sub graph of *USB* graph that contains a series of dipoles D_n such that *n=even* and the *indegree* is equal to the *outdegree* of D_n, constitutes a dipole series USB_{EG}. Figure 4 (c) highlights one such dipole series that exists in figure 1.

Definition 5 (*Cycle*): A *cycle* in a *USB/sub graph of USB* graph is defined as a closed walk USB_H (*sub graph of USB*) shown in figure 4 (d), such that every page in that path is visited exactly once except the starting page at which the walk terminates.

Studies made by Mortia and Shinoda [13], Konstan et. al., [14], Zhang and Seo [15], Rafter and Smyth [16] confirm that page-view time and actions performed on page are good indicators of user's interest on that page. Hence these factors were utilized in this work for confirming the importance of a page in a USB graph with respect to the user's interests. The existence of the above mentioned graph patterns

were studied and the pages that were part of such patterns were analyzed for its relevancy by considering the explicit feedbacks given by the user. The experimental results and the analysis are explained in the next section

4. Experiment Study and Results

The users have submitted their search queries through a specially designed browser (Refer to Appendix). The browser is designed in such a way that it keeps track of all the user data. Table 2 highlights the list of user's behaviors (implicit indicators) gathered by the browser for each user on each access to a page. Thus the user data is collected implicitly from the users without any user intervention. The user queries are submitted to an existing search engine like Google or Yahoo. The results given out by the existing search engine are given to the user in the browser.

The number of users involved was six. Among the six four of them were post graduate students in computer science & engineering and the other two were research scholars in the same. Hence they all had at least five years experience of working with computers. All the six users performed regular searches using the new browser developed for this work. Their behavior on every page they accessed was recorded by the browser.

Table 2. List of User's Behaviors (Implicit Indicators) Gathered by the Browser

User Behavior	Remarks
Anonymous User Names	Essential for any user to use the browser & to perform search through the WWW
URL of the accessed page	Collected and displayed in graphs
Parent URL of the accessed page	Collected and used for graph construction
Page-view time	Measured in seconds normalized by page size
Print	0/1, indication of any print action performed on a page
Bookmark	0/1, indication of adding a page to bookmark list
Save	0/1, indication of saving a page to hard disk
Copy/paste	0/1, indication of any copy/paste action performed on the page
Search Queries	Essential for any search & direct indicators of user's information need. Utilized for user and page categorization
Search Query usage time	Measured in seconds & Indicative of how long a search was performed

The users were also asked to explicitly rate the relevancy of each page which was minimal disruption to their regular work, but necessary for the experiment. The users were asked to select one among the following six options: 0 – No Idea, 1- Not Relevant, 2 – Leads to Useful Link, 3 - Partially Relevant, and 4 – Exactly Relevant. When a user issues a new search query or modifies the previous search query he/she

is asked to rate the search as a whole for the given search query using one among the following four options: 0 – No Idea, 1 – Not Useful, 2 – Partially Useful and 3 – Very Useful.

The page-view time was measured in seconds. Catledge et al. [12] first measured a period of 25.5 minutes as the optimum timeout for a session and hence the proposed system also makes use of the same timeout value. 1716 pages were collected from which 28 pages were removed since the users had no opinion about their relevancy (rated as 0). Hence the remaining 1688 pages were used for the remaining processes and analyzed.

The graph in figure 5 presents the rating distribution where it is notable that users mostly rated pages as partially relevant (3) or exactly relevant (4).

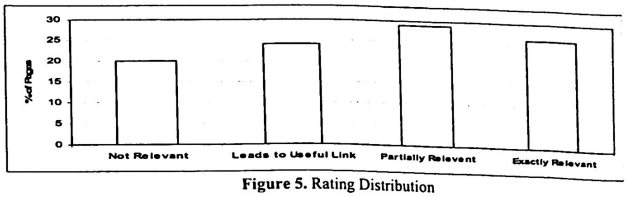

Figure 5. Rating Distribution

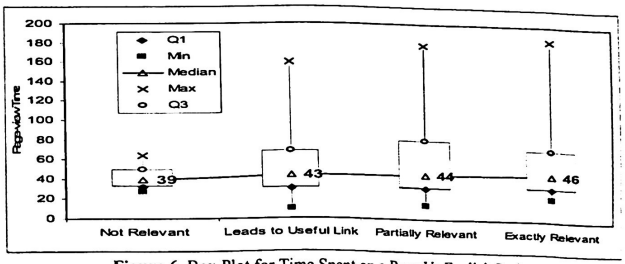

Figure 6. Box Plot for Time Spent on a Page Vs Explicit Ratings

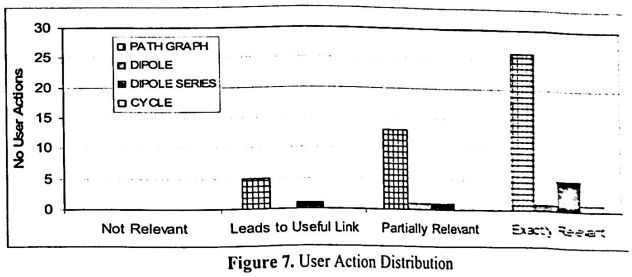

Figure 7. User Action Distribution

The graph in figure 6 is the box-plot of the time spent on a page versus explicit ratings. In the box plot the boxes show the range of values from the bottom quartile (25%) to the top quartile (75%) and the lines in the boxes show the medians of the page-view time. The box plot shows that there is a substantial increase in median (indicated by their values) for the reading time towards the explicit rating groups, which indicates that page-view time is a good indicator of user's interest on a page.

Proof of Hypothesis 1: *A linear path in the USB graph indicates that the page(s) that participate in the linear path is relevant to the search context provided atleast one user-action must have taken place through that path.*

The USB graphs were analyzed for the distribution of the various patterns according to the user feedback. Graph in figure 7 provides the distribution of the identified patterns in each user feedback category. The various observations that can be made from the pattern distribution graph in favor of hypothesis 1 are:

1. The number of path graphs USBPG is greater in the "Exactly Relevant" category when compared with the other feedback categories. Means area
2. Linear paths in the "Not Relevant" category is due to the fact that the user unknowingly had traced the web pages linearly and this is evident from the graph in figure 9 that the page-view times of the pages that constitute such linear paths in the "Not Relevant" category are characterized by very small page-view times. It also shows that the page-view time median is very high in "Exactly Relevant" when compared with the "Not Relevant" category.

Proof of Hypothesis 2: *Closed walks like dipoles and cycles in a USB graph indicate that the page(s) that participate in the closed walk are irrelevant to the search context provided no user action is performed on them.*

The observation|ns that can be made from the pattern distribution graph (figure 8) in favor of hypothesis 2 are:

1. The number of cycles like dipole, Euler and Hamiltonian cycles are large in number for the feedback category **"Not Relevant"** than any other feedback category. This clearly shows that cycles in the USB graph and the pages that constitute such patterns are not relevant to the user's context of search. Which is further confirmed by the fact that **"Not Relevant"** category pages do not contain any user actions is evident from the graph in figure 7.
2. Also it can be seen from the same graph that the **"Exactly Relevant"** category pages have more number of actions when compared with the other categories.

Evaluation: ANOVA is a common statistical test that examines whether there exists a significant difference between al least one of the group means. Since the group means are affected by outlier in the experiments conducted, a parametric test called Kruskal-Wallis test was performed on the patterns identified and the explicit rating groups, instead of ANOVA test. The Kruskal-Wallis test rejected the null hypotheses ($p<0.001$) in both the cases (hypotheses 1 and 2) signifying that median pattern values for at least one of the explicit feedback rating group was different.

Thus the pages that constitute the cyclic patterns can be deemed to be irrelevant and hence can be removed from the final results of a web search. The problem of false positives and false negatives due to high page view time and greater hit-count for some irrelevant pages that were visited unknowingly by the user can thus be eliminated and hence the search results can be improved.

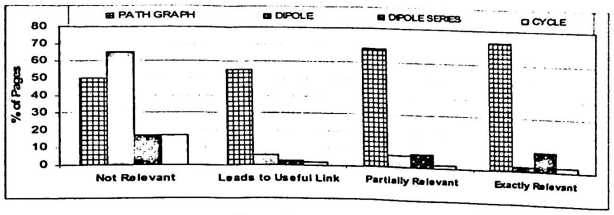

Figure 8. Pattern Distribution

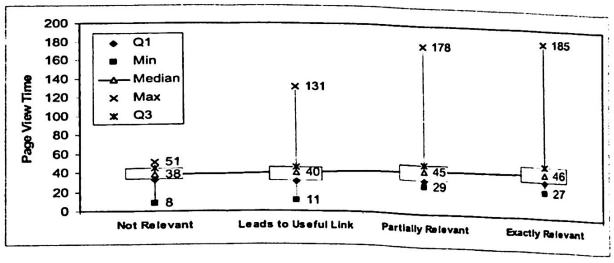

Figure 9. Box Plot for Page-view Time of Pages that constitute Path Graphs

5. Conclusion and Future Work

It can be concluded from this work that the existence of open and closed walks in the user's browsing behavior provide useful information of how a page is relevant or irrelevant to a search context. Also it has been proved that any linear path in the USB graph indicates that the page(s) that participate in the linear path are relevant to the search context provided atleast one user action must have taken place through that path and closed walks (Dipole, Dipole Series, Cycle) in a USB graph indicate that the page(s) that participate in the cycle are irrelevant to the search context provided no user action is performed on them.

Further improvement would be carried out in the following aspects: one is to provide personalizing features for many groups of users so that even for a new user, the personalized search system can recommend pages based on the commonalities in the search behavior. The other is to support for visual exploration of web search results. These additional features are be added to the system in near future.

References

1. D. Oard and J. Kim: Implicit Feedback for Recommender Systems. Proceedings of the AAAI Workshop on Recommender Systems, (1998) 81-83
2. Filippo Menczer: Combining Link and Content Analysis to Estimate Semantic Similarity. Proceedings of *IWIW2004*, (2004) 452 - 453
3. S. Chakrabarti, B. Dom, P. Raghavan, *et al.*: Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks*, 30(1–7): (1998) 65–74
4. K. Bharat and M. Henzinger: Improved algorithms for topic distillation in hyperlinked environments. In *Proc. 21st ACM SIGIR Conference*, (1998) 104–111
5. F. Menczer: Lexical and semantic clustering by Web links. *Journal of the Am. Soc. for Information Science and Technology*, 55(14), (2004) 1261-1269
6. Jon K.: Authoritative sources in a hyper-linked environment. Proceedings of the . ACM-SIAM symposium on Discrete Algorithms, New York: ACM Press, (1998) 668-677
7. Sergey B, Larry P.: The anatomy of large scale hyper textual web search engine. Computer Networks and ISDN Systems, 30(1-7): (1998) 107-117
8. Larry P. PageRank: Bringing order to the web. Stanford digital libraries working paper, (1997) 1-17
9. Wenxue Tao, Wanli Zuo: QuerySensitive SelfAdaptable Web Page Ranking Algorithm, *Second International Conference on Machine Learning and Cybernetics, Xi,* (2003) 413-418
10. X. Shen, B. Tan, and C. Zhai: Context-sensitive information retrieval with implicit feedback. *Proceedings of SIGIR 2005* (2005) 43 - 50
11. S. Sendhilkumar and T.V. Geetha: Personalized Web Search Using Enhanced Probabilistic User Conceptual Index. *International Journal of Intelligent Systems*, Vol. 17, No. 1-4, (2007) 199-213
12. S. Sendhilkumar and T.V. Geetha: User's Search Behavior Graph for aiding Personalized Web Search. 2nd International Conference on Pattern Recognition and Machine Intelligence (PReMI '07), Lecture Notes in Computer Science, Vol. 4815. Springer-Verlag, Berlin Heidelberg New York (2007) 415–438
13. Morita M, Shinoda Y: Information filtering based on user behavior analysis and best match text retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval (1994) 272-281.
14. Konstan J, Miller B. N., Gordon L.R., Reidl J , Grouplens: Applying collaborative filtering to Usenet news, communications of the ACM, 40(7), (1997) 77-87
15. Zhang B.T., Seo Y. W.: Personalized Web Document filtering using reinforced learning. Applied Artificial Intelligence 15 (7), (2001) 665-685
16. Rafter, R., and Smyth, B.: Passive Profiling from Server Logs in an Online Recruitment Environment. Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (2001) 35-41