CHAPTER 3

PERSONALIZED WEB SEARCH ARCHITECTURE

3.1 REQUIREMENTS OF AN EFFECTIVE PERSONALIZED WEB SEARCH

3.1.1 Tracking of User Data and Actions

The automatic categorization of user’s web activity could serve useful in developing more intuitive and effective information filtering systems. Many information filtering systems rely on implicit measures to infer user interest, such as dwell time, mouse and keyboard activity, and document interactions (e.g., copy, cut, save, print) (Oard and Kim 2001). The ability to automatically infer a user’s task would allow any information filtering system to apply the most appropriate measures of user interest, thereby improving the effectiveness of the information filtering system.

3.1.2 Maintaining Privacy in spite of Tracking the User Visited Pages

A problem in the web domain is the inherent conflict between the analysis needs and privacy needs. The analysts want more detailed usage data collected and the privacy needs of users who want as little data exposed as possible. Richard Mason developed the PAPA (Privacy, Accuracy, Property, and Accessibility) model, which is rooted in the deontological tradition (Mason 1986). Mason postulates the universal maxim that information technology and the information it handles must be used to enhance the dignity of mankind. Mason further holds that the main questions concerning privacy
are to what extent a person must reveal personal information to others, under what circumstances, and how preventive measures can be taken to guarantee privacy. For the purpose of personalization the personal information of a user is not required and the user’s identity may be kept anonymous.

3.1.3 Searchable Index for Collected Data

User data collected by any personalized search system must be indexed based on the following factors: content, context and user. The challenge is to develop methods for exploiting all evidence, including the web structure, meta-data, and user context information, to find high quality information for users (Allan and Bruce 2002).

3.1.4 Context-based Search

Contextual information retrieval is all about combining search technologies and knowledge about query and user context into a single framework in order to provide the most ‘appropriate’ answer for a user’s information needs (Frakes and Baeza-Yates 1992). In general, interactions with web search systems could be characterized as ‘one size fits all’. This means that all queries are treated as simple web queries where the aim is to locate useful home pages and the burden is placed on the user to scan and navigate the retrieved material to find the answers. There is no representation of user preferences, search context, or the task context. Despite some recent attention to this problem, the difficulty of capturing and representing knowledge about users, context, and tasks in a general web search environment makes this area more challenging.
3.1.5 Identifying and Filtering Irrelevant Pages

Identifying irrelevant pages in web search is very important as it helps to reduce the amount of work done during web search. A number of web agents use various information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter and categorize them (Frakes and Baeza-Yates 1992; Morita and Shinoda 1994). Many works (Malone et al 1987; Foltz and Dumais 1992; Alton-Scheidl et al 1997; Avery and Zeckhauser 1997) in the literature exists on how to identify relevant and user interested pages by using explicit user ratings. The Web4Groups Voting Service (Alton-Scheidl et al 1997) encompasses the similarities of voting and rating at the level of social action. Various examples of rating activities in everyday life illustrate the importance of evaluating objects according to the knowledge, interests, or opinions of individuals, experts, or groups of people.

The results of surveys conducted on voting and rating were the basis for the design of a Web4Groups voting tool for group decision-making, as well as a Web4Groups rating tool for enhancing context information in discussion forums. The information Lens (Malone et al 1987) is a message filtering system which emphasizes the cognitive approach, to information filtering. It exploits, in a simple way, the techniques from artificial intelligence like frames, production rules, and inheritance. Foltz and Dumais (Foltz and Dumais 1992) used keyword match-word profile method to compared words in the employee's profile to words in the abstracts of new Technical Memos.

3.1.6 Prediction and Usage of Implicit Browsing Behavior Patterns

Implicit browsing behavior patterns might prove very essential for identification of irrelevant/non–interesting pages while performing web
search. Work done by Oard and Marchionini utilize implicit information (Oard and Marchionini 1996) and rates are automatically inferred from user's behaviour. User browsing patterns also aid in identifying the following: i) relevant pages that are not in the results returned by existing search engine, ii) relevant pages (from the results returned by existing search engine) that are missed by the user and iii) page relevancy based on user actions.

3.1.7 Predicting User Interests

Tracking user preferences and interests is a learning problem that has been investigated by researchers in information retrieval (Allan 1996) and intelligent agents (Sheth 1993; Lang 1995; Moukas and Zacharia 1997; Balabanovic 1997 and 1998; Chen and Sycara 1998; Billsus and Pazzani 1999a; Billsus and Pazzani 1999b). Rocchio’s relevance feedback is an algorithm for learning user interests that has been well studied in information retrieval (Rocchio 1971; Salton and McGill 1983; Jurafsky and Martin 2000). Learning the dynamics of user interests is closely related to learning long-term and short-term interest models (Widyantoro 1999; Widyantoro et al 1999). In order to customize search through the WWW, any personalized search system must maintain the users' information as precisely as possible. In addition automatic prediction of user interests with least effort from the users becomes a critical issue for these systems.

3.1.8 Personalized Re-Ranking Based on User Interests

Many of today’s search engines produce thousands of retrieved pages related to the query and sort this list by some ranking criterion. There is also a substantial amount of literature on various theoretical aspects of learning a ranking function (Cohen et al 1999; Díez et al 2002). One successful and well-publicized ranking system is PageRank (Brin and
Lawrence 1998), the ranking system used by Google. Actually, for pages related to a query, an IR score is combined with a PR (PageRank) score to determine an overall score, which is then used to rank the retrieved pages (Schneider et al 2003). Topic-Sensitive PageRank (Haveliwala 2002), instead of computing a single global PageRank value for every page, the search engine computes multiple Topic-Sensitive PageRank values, one for each topic listed in the Open Directory (ODP 2009). A more recent investigation (Jeh and Widom 2003), uses a different approach: it focuses on user profiles built from user page-views and pages are ranked based on the personalized page preference scores. But ranking pages based on user interests is what needed for effective personalization.

Any personalized search system that does the above mentioned functionalities will deliver an effective web search. This work utilizes the search results returned by any existing search engine like Google or Yahoo!. The results are then analyzed for relevancy, re-ranked and recommended to the user. Thus the overall objective of this work is to provide an architecture that binds the various factors of personalization to provide an effective and efficient web search.

3.2 INTEGRATED PERSONALISED WEB SEARCH ARCHITECTURE

The proposed personalized search system (Figure 3.1) is made up of the following six layers: Presentation, Pre-processing, Data Storage, Data Extraction, Knowledge and Analysis layers.
Figure 3.1 Personalized Web Search Architecture
3.2.1 Presentation/User Layer

Personalization relies heavily on the user data. Though personalization seems a simple concept, it is difficult to implement. By its very nature, personalization means different things to different people. Data at the user side is valuable (Konstan et al. 1997; Mobasher et al. 1999) in the context of personalization.

The presentation layer is the place where the user interacts with the personalized search system and the data about the user’s search is collected. There are many problems associated with user data collection. The usage data are usually obtained from either web server log, at the server side, or web browser, at the client side. Data acquisition via server log involves preprocessing which is an overhead and hence it practically renders on-line mining of user behaviors impossible (Brij et al. 2002). Unfortunately, cache hits are missing from the server log, rendering it as an incomplete source of information to acquire spatial features of user interactions such as hit-count (Konstan et al. 1997; Mobasher et al. 1999).

Temporal features such as page view-time of pages are considered highly informative in deducing user preferences (Konstan et al. 1997). However, timestamps recorded for each server log entry includes the network transfer time. Hence page-view time extracted from the server side is not accurate. User session identification and data cleansing are the most difficult and time-consuming tasks performed during preprocessing of the server log. Since there is a many-to-many relationship between users and IP addresses which are recorded in the log entry for each HTTP request, it becomes an unreliable source of information for user session identification. Due to proxy servers and/or IP masquerading, and missing cache hits in the server log, the user identification process becomes even harder. Hence server log is not reliable as a source of usage data for web usage findings.
Researchers have proposed various methods to resolve this problem, but none of these methods are without serious drawbacks. One such method is exploiting the cookies, which allow inter-session tracking of users, but this method violates users privacy; web-sites requiring user registration are often neglected by anonymous users; dynamic URLs embedding session ID restricts intermediate caching and does not correctly handle the exchange of URLs between people (Pitkow 1997); cache-busting defeats the speed up advantage gained by caching. To avoid all these problems the architecture proposed in this thesis suggests a client-side web usage data acquisition.

The presentation layer keeps track of all the user data collected from the machines, from where the user performs the search. Users submit their search queries through a specially designed browser and view the result pages. The user queries are submitted to an existing search engine like Google or Yahoo!. Recent studies showed that users often do not examine lower-ranked results (Joachims et al 2005). Hence the presentation layer collects top 30 results given out by the existing search engine and the pages visited. These data are analyzed, re-ranked based on page relevancy with respect to the user interests and current context of search. The pages are then recommended to the users through the browser. The user interface is designed in such a way that it keeps track of user data like the search queries given, pages visited, time spent on a page, and actions performed (save, copy, print, bookmark) on a page.

The tracked user information is deposited into the data layer after pre-processing. Thus the user data is collected implicitly from the users without any user intervention. For experimental study, we involved 15 users to work on our personalised search system. Among the 15, eight of them were post graduate students, five were undergraduate students in computer science and 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 15 users performed regular searches using the intelligent browser (Figure 7.1) developed for the proposed personalised search architecture. Their behaviour on every page they visited was captured by the browser. Table 3.1 highlights the list of user’s behaviours (implicit indicators) gathered by the browser for each user on each access to a page.

Table 3.1  List of User’s Behaviors (Implicit Indicators) Gathered by the Browser

<table>
<thead>
<tr>
<th>User Behaviour</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymous User Names</td>
<td>Essential for any user to use the browser &amp; to perform search through the WWW</td>
</tr>
<tr>
<td>Information Viewed</td>
<td>Textual Contents of a visited page that is viewed</td>
</tr>
<tr>
<td>URL of the accessed page</td>
<td>Collected and displayed in graphs</td>
</tr>
<tr>
<td>Parent URL of the accessed page</td>
<td>Collected and used for graph construction</td>
</tr>
<tr>
<td>Page-view time</td>
<td>Measured in seconds normalized by page size</td>
</tr>
<tr>
<td>Print</td>
<td>0/1, indication of any print action performed on a page</td>
</tr>
<tr>
<td>Bookmark</td>
<td>0/1, indication of adding a page to bookmark list</td>
</tr>
<tr>
<td>Save</td>
<td>0/1, indication of saving a page to hard disk</td>
</tr>
<tr>
<td>Copy/paste</td>
<td>0/1, indication of any copy/paste action performed on the page</td>
</tr>
<tr>
<td>Search Queries</td>
<td>Essential for any search &amp; direct indicators of user’s information need. Utilized for user and page categorization</td>
</tr>
<tr>
<td>Search Query usage time</td>
<td>Measured in seconds &amp; Indicative of how long a search was performed</td>
</tr>
</tbody>
</table>

In many cases a potential roadblock is users' unwillingness to reveal personal information to fine-tune personalization features. Privacy policy statements are often confusing and arouse suspicion in users' minds,
leading to reluctance to share personal information. To overcome this, the users of personalized search system are provided with anonymous user name as login. Hence the problem of privacy is also eliminated. The user data thus collected are pre-processed before storing them into the data storage layer.

3.2.2 Pre-Processing Layer

The user data tracked by the browser is raw data like the search query string, URLs of web pages, time in seconds, etc. The web pages collected might be large sized HTML documents. Storage and maintenance of such large unstructured/semi-structured data is very difficult when collected over a large period of time. In addition mining on such data is time consuming. Hence pre-processing is done to extract important concepts that represent the content of such unstructured/semi-structured documents. The visited pages are maintained by indexing their URLs with the concepts extracted from the pages. Such keywords that are used for indexing the URLs are called as Index Word (IW).

The users are allowed to issue natural language queries (e.g. ‘What is Personalization?’) and the concept(s)/keyword(s) in the search queries (e.g. personalization) are the best indicators of user’s interests. Pre-processing also involves extraction of such concepts from the user queries. The pre-processing layer performs the following activities: HTML to text conversion, POS tagging, noun extraction, computing term-weight based on Term-Frequency (TF) and Inverse Document Frequency (IDF) and feature selection.

All the top 30 ranked pages and visited pages are parsed. The parser used for this purpose is Monte Lingua parser (Liu 2004) an end-to-end natural language processor for English, for the Python/Java platform. The parser eliminates the HTML tags and keeps the textual content of a page. The textual
content of a page are then tagged for its parts-of-speech. Unigrams like Nouns and Adjectives as well as bi-grams like Noun-Noun and Adjective-Noun represent the various concepts that a page speaks about. Such unigrams and bi-grams are extracted and their TFs and IDFs are computed. Then the TF * IDF weight is computed for all the keywords present in a page.

Experiments conducted by Salton and Buckley (1988) indicate that top 3% of the unigrams/bi-grams in a document highlight the concepts spoken in a page. Hence we have chosen the top 3% unigrams and bi-grams are selected as the features of each page. Such keywords are used to index the page URLs in the Data storage layer. The complete pre-processing process explained above is shown in Figure 3.2. At the end of pre-processing every page visited by the user are semantically tagged with the concepts that represent the content of a page.

![Figure 3.2 Steps in Data Pre-processing](image)

3.2.3 Data Layer

The data collected by the browser from the user end are populated into tables present in the data layer. The data layer acts as a central warehouse of all the data needed for achieving an effective personalization. Every web page visited by the user is represented by the page’s URL and indexed by the
features extracted during the pre-processing stage. The user data is organized in special structures called the transactions and search sessions.

**Definition 3.1 (Transaction):** Given a search query $SQ_i$ and the initial set of pages retrieved by existing search engine, a transaction $T$ is a triple $<Session_{id}, SQ_i, P_i>$ where $Session_{id}$ is a unique number that represents every search session performed by the user, $SQ_i$ is the search query issued by the user and the page-view $P_i = \{p_1, p_2,\ldots, p_n\}$ is a set of pages visited by the user for a search query $SQ_i$.

**Definition 3.2 (Search Session):** Given a set of search queries $Q$, where $SQ \in Q$, a session is defined as a set of transactions that occurred until the user closes the browser and is represented as $S = \{T_1, T_2,\ldots, T_m\}$, where every $T_i$ represents one unique $SQ_i$. Thus a search session represents all the transactions that take place in between the opening and closing of the browser.

Transactions that house the page-views are semantically meaningful entities which can be used for extracting useful information like users' interested web pages, relation between various pages, user interests, etc. The data layer is the source for data extraction, knowledge and analysis layer. The data storage layer houses the initial user profiles which are constructed automatically in the data extraction layer. In general the data layer comprises of the entire user browsing data structured as mentioned above.

The preliminary user profiles store approximations of the interests of a given user. User's interests are identified automatically from the
collection of various search queries and the relevant pages visited by the user. The terms in the user’s search queries, the Index Word (IW) of the visited pages, the frequency of usage of similar search terms indirectly highlights the user’s interest in a particular concept. In addition the user profile contains the user’s favourite pages which are identified by frequency of visits to a particular page.

3.2.4 Data Extraction Layer

The data extraction layer is essential for deriving the conceptual relations between the search queries and their relevant pages, structuring the personalized information needed for final re-ranking process. The data extraction layer also structures the users search behaviour as a graph, identifies relevant and irrelevant pages and hence this layer aids filtering of irrelevant pages according to the current context of search.

User profiles store the approximation of user’s interests and disinterests. After identifying the relevant and irrelevant information for individual users they are then populated into the user profile, which is yet another function of the data extraction layer. In this work the methodology used for the generation of user profiles differs from the majority of other approaches in that the profiles are: generated automatically, without explicit user feedback, and dynamic, i.e., not based on a fixed period of time, but it evolves over time.

In a nutshell, the data extraction layer does the following functions:
1) Offline computation of the User Conceptual Index (UCI) using the factors like page-hit, page-view time, query hit and query usage time, 2) Construction of Transaction Feature Matrix (TFM) where each element of the matrix holds the conceptual relation (quantified as UCI value) between the search query
and its relevant pages, 3) User Search Behavior (USB) Graph construction, 4) Automatic construction of user profiles called Personalized Page-View (PPV) graph using conceptual graphs and 5) Identification of user interest categories using taxonomical data and categorising user interests as long and short-term interests.

The data extraction layer extracts the spatial data like the pages visited and their hits, search queries used and their hits, the temporal data like page-view time and query usage time and the implicit indicators of user interests like user actions on pages. Such information is used to compute the UCI. A detailed discussion on UCI computation is postponed until chapter 4.

3.2.5  Knowledge Layer

The knowledge layer is the place where reference knowledge and personalized knowledge are being stored. The personalized knowledge is generated and represented as conceptual graphs called Personalized Page-View (PPV) graph which is user specific.

Another important factor which when utilized leads to an effective context-based web search is the content relations between the pages visited by the user. Deeper semantics on the page content can be inferred with the help of ontological representation of the visited pages. With the aim of utilizing the semantically related pages for page recommendation and to refine the proposed UCI with page link factors, the data from the data layer are used in the knowledge layer for the construction of the PPV graphs. The PPV graph is constructed automatically from the set of pages visited by the user. More elaborate discussion on PPV graph construction is postponed until chapter 5.
The conceptual PPV graphs and USB graphs thus constructed are then analyzed for further re-ranking and recommendation. This analysis is carried out in the analysis layer and the details of all the analysis carried out are discussed in the following section.

3.2.6 Analysis Layer

The PPV graphs constructed in the knowledge layer exhibits various semantic paths that exist between the concepts of various pages. Such semantic paths are assigned with weights based on user actions performed and the shortest paths are ranked based on the weights computed. The ranked shortest search paths and the relevant pages in those search paths are then updated into the user profiles for aiding page recommendations. Individual user interest analysis is essential for updating the user profiles with their interests and disinterests.

In a multi user search environment semantic search paths of individual users are compared for identifying similar search behaviours which proves very important for page recommendations for a new user. Analysis of user behaviours, interests, search paths are thus very essential for final page re-ranking. Also this layer utilizes ODP taxonomy (ODP 2009) for labelling user interests and categorizing such interest labels into long and short-term interests. More discussion on user interest categorization is postponed until chapter 6.

The results of all the above mentioned analysis are combined and utilized for final page re-ranking and recommendation. The various analyses that are performed for page re-ranking in the analysis layer are as follows: Semantic analysis and ranking of search paths, user interest analysis and evolving interest based user groups for recommending pages to new users.
Finally the re-ranked pages are displayed to the user in the presentation layer through the browser.

The integrated architecture for personalized web search, proposed in this chapter, provides the following benefits which is crucial for any personalized search system. They are: 1) enabling context and content-oriented search, 2) effective methodology for eliminating the most appropriate irrelevant pages according to the current context of search and 3) recommending pages that are left unvisited. Personalization using such conceptual graphs can produce better results as compared to keyword-based searching by providing conceptual link between visited and unvisited pages and thus pages that are unvisited but relevant can also be recommended to the users. Any such personalized search architecture can produce an effective small scale multi-user personalized search environment.

The UCI based web search narrated in this thesis involves the following processes, the corresponding layers where these processes occur and the chapter in which they are discussed as listed in Table 3.2.

The processes mentioned in Table 3.2 aids to build an effective personalized web search system. In addition, the proposed personalized search system depends on the results returned by existing search engines. The next four chapters are not going to discuss about the various layers of personalized search architecture one after the other. Instead the next four chapters illustrate the various processes used in this work to achieve an effective personalized search. However, the processes involved are part of the various layers of the personalized search architecture mentioned in this chapter. Therefore the explanation of various processes involved in our system will be accompanied with citations to the corresponding layers of our architecture.
Table 3.2  Processes involved in UCI based Search and Chapters discussed

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Process</th>
<th>Layer involved</th>
<th>Chapter No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User data collection</td>
<td>Presentation Layer</td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>Feature (Index Word ) extraction for various pages visited by the user</td>
<td>Pre-processing Layer</td>
<td>3</td>
</tr>
<tr>
<td>3.</td>
<td>Search Query and Page weight computation</td>
<td>Pre-processing Layer</td>
<td>4</td>
</tr>
<tr>
<td>4.</td>
<td>Construction and storage of Transaction-Feature Matrix (TFM)</td>
<td>Data and Data Extraction Layer</td>
<td>4</td>
</tr>
<tr>
<td>5.</td>
<td>Computation of Basic and Modified UCI and updating into TFM</td>
<td>Data Extraction Layer</td>
<td>4</td>
</tr>
<tr>
<td>6.</td>
<td>Constructing User Search Behaviour (USB) Graph</td>
<td>Data Extraction Layer</td>
<td>4</td>
</tr>
<tr>
<td>7.</td>
<td>Analysis of USB graph for identifying relevant and irrelevant pages</td>
<td>Analysis Layer</td>
<td>4</td>
</tr>
<tr>
<td>9.</td>
<td>Construction and Updating User Profiles</td>
<td>Knowledge and Data Layer</td>
<td>5</td>
</tr>
<tr>
<td>10.</td>
<td>User’s search path analysis and ranking of search paths</td>
<td>Analysis Layer</td>
<td>5</td>
</tr>
<tr>
<td>11.</td>
<td>Evaluations of PPV graph</td>
<td>Analysis Layer</td>
<td>5</td>
</tr>
<tr>
<td>12.</td>
<td>Assigning user’s search with interest Labels</td>
<td>Analysis Layer</td>
<td>6</td>
</tr>
<tr>
<td>13.</td>
<td>User interest Analysis</td>
<td>Analysis Layer</td>
<td>6</td>
</tr>
<tr>
<td>14.</td>
<td>Categorization of user interests into long and short-term</td>
<td>Analysis Layer</td>
<td>6</td>
</tr>
<tr>
<td>15.</td>
<td>Personalized page recommendations</td>
<td>Presentation Layer</td>
<td>7</td>
</tr>
<tr>
<td>16.</td>
<td>Evaluation</td>
<td>Analysis Layer</td>
<td>7</td>
</tr>
</tbody>
</table>

The following chapter introduces User Conceptual Index (UCI) and its associated data structure called Transaction Feature Matrix (TFM) which is the major contribution of our research work.