ISSN:0975-9646

New Approaches to Web Personalization Using Web Mining Techniques.

Beatric , Ryan Fernandes, Leo. J. Peo, Nikhila Kamat, Sergius Miranda

Department of Computer Engineering,
Xavier Institute of Engineering, Mahim, Mumbai, India.

Abstract--**With millions of pages available on web, it has become difficult to access relevant information. One possible approach to solve this problem is web personalization. Web personalization is defined as any action that customizes the information or services provided by a web site to an individual. It includes Web mining, as application of data mining techniques to extract knowledge from Web. Web mining has been explored to a vast degree and different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. Web mining has been a form of 'data-centric' view. In this paper, the significance of the evolving nature of the Web personalization have described. Web usage mining is used to discover interesting user navigation patterns and can be applied to many real-world problems, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, etc. A Web usage mining system performs five major tasks: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. The Web mining research is a converging research area from several research communities, such as Databases, Information Retrieval and Artificial Intelligence. In this paper, the web mining techniques also described. It also highlights the applications and tools of web personalization.**

Keywords-- **Usage Mining, Navigation Patterns, Pattern Analysis, Content Mining, Structure Mining**

I. INTRODUCTION

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web. A user interacts with the web, there is a wide diversity of user's navigational preference, which results in needing different contents and presentations of information. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized web pages to the user by learning user navigational pattern knowledge [1,2,3].

Interest in the analysis of user behavior on the web has been increasing rapidly This increase stems from the realization that added value for web site visitors is not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form. To improve the Internet service quality and increase the user click rate on a specific website.

Thus, it is necessary for a web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and presents the customized web pages to the user by learning user navigational pattern knowledge.

Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling, social networks, web data mining, human-machine interactions to Web usage mining.

The process of website personalization starts with the collection of user data, which is then integrated with user needs, demands and the result is a specifically designed website. Some basic elements of website personalization include content, imagery and even the theme of the website .Benefits offered by Website Personalization include a customer- focused and a much comfortable website, Information catered and presented in a much better manner resulting in saving of time, productive results, and higher call-to-action

The tremendous growth in the number and the complexity of information resources and services on the Web has made Web personalization an indispensable tool for both Web-based organizations and for the end users. The ability of a site to engage visitors at a deeper level, and to successfully guide them to useful and pertinent information, is now viewed as one of the key factors in the site's ultimate success.

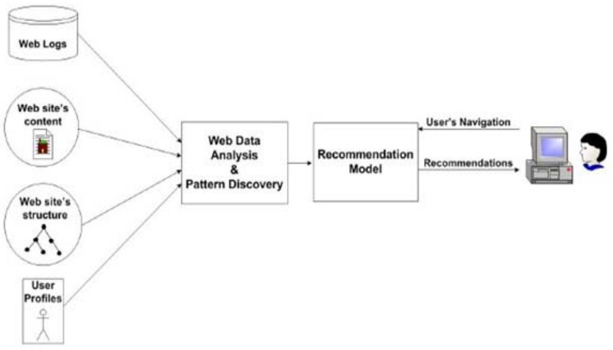

Figure 1: The web personalization process

II. WEB MINING

Web mining is the application of data mining techniques to extract knowledge from web data.

A. *Types of web mining*

1)Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

2) Web Structure Mining: The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Thus, Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the(intra-page) document level or at the (inter-page)hyperlink level (Figure 1).

3) Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Some of the typical usage data collected at a Web site include IP addresses, page references, and access time of the users.

a. The Different steps in web usage mining are

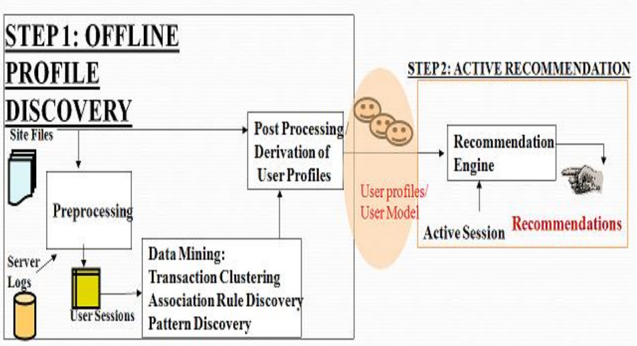
STEP 1: OFFLINE PROFILE DISCOVERY
Site Files
Preprocessing
Server Logs
User Sessions
Data Mining:
Transaction Clustering
Association Rule Discovery
Pattern Discovery
Post Processing/ Derivation of User Profiles
User profiles/ User Model
STEP 2: ACTIVE RECOMMENDATION
Recommendation Engine
Active Session
Recommendations

Figure 2: Steps in web personalization system

B. ***Knowledge discovery:***

Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

III. WEB DATA

Web data are those that can be collected and used in the context of Web personalization. These data are classified in four categories according to [6]:

A. Categories of web data

1) Content data: Content data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.

2) Structure data: Structure data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Website together, such as hyperlinks connecting one page to another.

3) Usage data: Usage data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.

4) User profile data: User profile data provide information about the users of a Web site. A user profile contains demographic information for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

IV.PERSONALIZATON ON THE WEB

Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. Correctly executed, personalization of the visitor's experience makes his time on your site, or in your application, more productive and engaging. Personalization can also be valuable to you and your organization, because it drives desired business results such as increasing visitor response or promoting customer retention. Unfortunately, personalization for its own sake has the potential to increase the complexity of your site interface and drive inefficiency into your architecture. It might even compromise the effectiveness of your marketing message or, worse, impair the user's experience. Few businesses are willing to sacrifice their core message for the sake of a few trick web pages. Contrary to popular belief, personalization doesn't have to take the form of customized content portals, popularized in the mid-to-late 90s by snap.com and My Yahoo!. Nor does personalization require expensive applications or live-in consultants. Personalization can be as blatant or as understated as you want it to be. It's a tired old yarn, but if you hope to implement a web personalization strategy, the first and most important step is to develop and mature your business goals and requirements. It is important to detail what it is you hope to do and, from that knowledge, develop an understanding of how you get from an idea to implementation. You might be surprised to discover that it won't require most of next year's budget to achieve worthwhile results.

Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling, social networks, web data mining, human-machine interactions to Web usage mining; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories selection. Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also include, in addition to user contextual information, contextual information of real-time interactions with the web proposed a multi-agent system based on three layers: a user layer containing users profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents propose reformulation query by adding implicit user information. This helps to remove any ambiguity that may exist in query: when a user asks for the term "conception", the query should be different if he is an architect or a computer science designer. Requests can also be enriched with predefined terms derived from user's profile develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level. Other approaches also consider social-based filtering and collaborative filtering. These techniques are based on relationships inferred from users' profile. Implicit filtering is a method that observes user's behavior and activities in order to categorize classes of profile [3,8,11,12].

V.PERSONALIZATION STRATEGIES

The web personalization strategies are as follows in the increasing order of importance and complexity.

A. Memorization

In this simplest and most widespread form of personalization, user information such as name and browsing history is stored (e.g. using cookies), to be later used to recognize and greet the returning user. It is usually implemented on the Web server. This mode depends more on Web technology than on any kind of adaptive or intelligent learning. It can also pose a threat to user privacy.

B. Customization

This form of personalization takes as input a user's preferences from registration forms in order to customize the content and structure of a web page. This process tends to be static and manual or at best semi-automatic. It is usually implemented on the Web server. Typical examples include personalized web portals such as My Yahoo and Google.

C. Guidance or recommendation system

A guidance based system tries to automatically recommend hyperlinks that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website. It is usually implemented on the Web server, and relies on data that reflects the user's interest implicitly or explicitly. This approach forms the focus of Web personalization [8,13].

D. Task performance support

In these client side personalization systems, a personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information. This approach requires heavy involvement on the part of the user, including access, installation, and maintenance of the personal assistant software. It also has very limited scope in the sense that it cannot use information about other users with similar interests.

The Web personalization process can be divided into four distinct phases [13]:

1) Collection of Web data: Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from registration forms and rating questionnaires. Additional data such as demographic and application data (for example, e-commerce transactions) can also be used. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.

2) Preprocessing of Web data: Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis (example: automatically generated requests to embedded graphics will be recorded in web server logs, even though they add little information about user interests), and completing the missing links (due to caching) in incomplete click through paths. Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.

3) Analysis of Web data: Also known as Web Usage Mining, this step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.

*4) Decision making / Final Recommendation Phase :*The last phase in personalization makes use of the results of the previous analysis step to deliver recommendations to the user. The recommendation process typically involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming.

VI.PHASES OF WEB PERSONALIZATION

A. Collection of web data

Data can be collected either in an implicit or an explicit manner.

1) Implicit data: It includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. This helps to study user's behavior at the website.

2) Explicit data: It comes from registration forms which the user Fill's while signing up with the website. The user rating questionnaires. Additional data such as demographic (i.e. study of the characteristics of visitors) and application data (example: e- commerce transactions) can also be used. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.

B. Pre-processing of web data

Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include the following steps:

- Cleaning data of inconsistencies
- Filtering out irrelevant information,
- Completing the missing links in incomplete click through paths.

Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.

C. Analysis of web data

Also known as Web Usage Mining, this step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.

D. Decision making and recommendation stage

The last phase in personalization makes use of the results of the previous analysis step to deliver recommendations to the user. The recommendation process typically involves generating dynamic Web content on the fly, such as adding hyperlinks to the last webpage requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming. "CGI" stands for "Common Gateway Interface." CGI is one method by which a web server can obtain data from (or send data to) databases, documents, and other the pre-processing phase is data preparation about the visitor's identification is stored, along with password information. Additional information such as credit card details, if one is used during a transaction, as well as details concerning the visitor's activities at the Web site, for example which pages where .

VII.REQUIREMENTS OF WEB USAGEMINING

It is necessary to examine what kind of features a Web usage mining system is expected to have in order to conduct effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to:

Gather useful usage data thoroughly,

- Filter out irrelevant usage data,
- Establish the actual usage data,
- Discover interesting navigation patterns,
- Display the navigation patterns clearly,
- Analyze & interpret the navigation patterns correctly and apply the mining results effectively.

VIII. COMMONLY USED IMPORTANT TECHNIQUES

A. User Profiling

In order to personalize a Website, the system should be able to distinguish between different users or groups of users. This process is called user profiling and its objective is the creation of an information base that contains the preferences, characteristics and activities of the users. A user profile can be either static, when the information it contains is never or rarely altered (e.g. demographic information), or dynamic when the user profile's data change frequently. Such information is obtained either explicitly, using on line registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behavior and/or the preferences of each user, resulting in dynamic user profiles.

1) Data Collection: A way of uniquely identifying a visitor through a session is by using cookies. W3C [WCA] defines cookie as "the data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent requests". In other words, a cookie is simply an HTTP header that consists of a text-only string, which is inserted into the memory of a browser. It is used to uniquely identify a user during Web interactions within a site and contains data parameters that allow the remote,HTML server to keep a record of the user identity, and what actions she/he takes at the remote Web site. The contents of a cookie file depend on the Website that is being visited.

A user can be identified making the assumption that each IP corresponds to one user. In some cases, IP addresses are resolved into domain names that are registered to a person or a company, thus more specific information is gathered.

2) Issues

a) Overdependence on browser cookies: First of all, in-case a system depends on cookies for gathering user information, there exists the possibility of the user having turned off cookie support on their browser. Other problems that may occur when using cookies technology are the fact that since a cookie file is stored locally in the user's computer, the user might delete it and when she/he revisits a website will be regarded as a new visitor. Furthermore, if no additional information is provided (for example some logon id), there occurs an identification problem if more than one user browses the Web using the same computer.

B. Log Analysis

By applying statistical and data mining methods to the Web log data, interesting patterns concerning the users navigational behavior can be identified, such as user and page clusters, as well as possible correlations between Web pages and user groups.

1) Web Log: Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format. The fields of the common log format are: *remotehost rfc931 authuser date "request" status bytes* where *remotehost* is the remote hostname or IP number if DNS hostname is not available, *rfc931* is the remote log name of the user, *authuser* the username as which the user has authenticated himself, available when using password protected WWW pages, date the date and time of the request, "request" the request line exactly as it came from the client (the file, the name and the method used to retrieve it),status the HTTP status code returned to the client, indicating whether or not the file was successfully retrieved and if not, what error message was returned, and bytes the content-length of the documents transferred. If any of the fields cannot be determined a minus sign (-) is placed in this field.

2) Data Pre-processing: There are some important technical issues that must be taken into consideration during this phase in the context of the Web personalization process, since it is necessary for Web log data to be prepared and pre-processed in order to use it in the consequent phases of the process. The first issue in the pre-processing phase is data preparation. Depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file.

Most important of all is the user identification issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate since for example, a visitor may access the Web from different computers, or many users may use the same IP address (if a proxy is used). A further assumption can then Clustering is used to group together items that have similar characteristics. After discovering patterns from usage data, a further analysis has to be conducted. Additionally, visualization techniques are used for an easier interpretation of the results. Using these results in association with content and structure information concerning the Web site there can be extracted useful knowledge for modifying the site according to the correlation between user and content groups be made, that consecutive accesses from the same host during a certain time interval come from the same user. More accurate approaches for Apriori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration.

3) Log analysis

Log analysis tools (also called traffic analysis tools), take as input, raw web data and process them in order to extract statistical information. Such information includes statistics for the site activity(such as total number of visits, average number of hits, successful / failed /redirected/cached hits, average view time, average length of a path through a site), diagnostic statistics (such as server errors, page not found errors), server statistics (such as top pages visited, entry / exit pages, single access pages), referrers statistics (such as top referring sites, search engines, keywords), user demographics, client statistics etc. Some tools also perform click-stream analysis, which refers to identifying paths through the site followed by individual visitors by grouping together consecutive hits from the same IP, or include limited low-level error analysis, such as detect in unauthorized entry points or finding the most common invalid URL. These statistics are usually outputted to reports and can also be displayed as diagrams. This information is used by administrators for improving the system performance, facilitating the site modification task and providing support for marketing decisions. However, most advanced Web mining systems further process this information to extract more complex observations that convey knowledge, utilizing data mining techniques such as association rule and sequential pattern discovery, clustering and classification.

4) Web usage mining in log analysis

Log analysis is regarded as the simplest method used in the Web usage mining process. The purpose of Web usage mining is to apply statistical and data mining techniques to the pre-processed Web log data, in order to discover useful patterns.

Association rule mining is a technique for finding frequent patterns, associations and correlations among sets of items. Association rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Aside from being exploited for business applications, such observations also can be used as a guide for Web site restructuring, for example by adding links that interconnect pages often viewed together, or as a way to improve the system's performance through pre-fetching Web data. Sequential pattern discovery is an extension of association rules mining in that it reveals patterns of co-occurrence incorporating the notion of time sequence. In the Web domain such a pattern might be a Web page or a set of pages accessed immediately after another set of pages. Using this approach, useful users trends can be discovered, and predictions concerning visit patterns can be made.

Clustering is used to group together items that have similar characteristics. After discovering patterns from usage data, a further analysis has to be conducted. Additionally, visualization techniques are used for an easier interpretation of the results. Using these results in association with content and structure information concerning the Web site there can be extracted useful knowledge for modifying the site according to the correlation between user and content groups.

IX.NEW APPROACHES TO WEB PERSONALIZATION

A. Filter Bubbles

A filter bubble is the intellectual isolation which assumes the information a user would want to see and make use of algorithms selectively through websites. The information given to the user, such as former click behavior, browsing history, search history and location, is based on this assumption. Due to this, only the information that will abide by the users' past activity will be presented by the websites. Filter bubble causes users to contact less with contradicting viewpoints ad become intellectually isolated. Examples, Personalized search result from Google and personalized news stream from Facebook.

B. Crowd turfing

There is a growing underground market on the Web for malicious crowd-sourcing. For just a few cents, you can buy Facebook likes, Twitter followers, bulk social networking accounts, and fake reviews on Yelp. These types of social spam are extremely difficult for existing security systems to stop because the damage is caused by real people, not automated programs. In our work, we have measured malicious crowd turfing systems, and we are actively engaged in devising new solutions to stop these insidious threats.

C. Social Sybils

Fake accounts, otherwise known as Sybils, are a pervasive threat on the social web. Sybils generate a large portion of the spam on social networks, and steal personal information that is used to power targeted phishing attacks. Examples: Spam on social networking websites like Facebook and Twitter.

X. WEB 3.O TECHNOLOGY

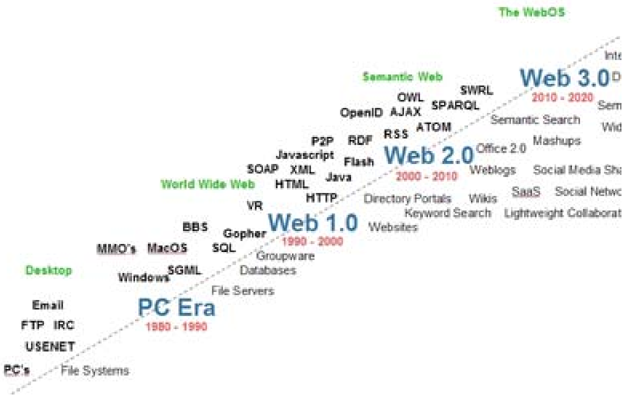

Figure 3: The web 3.0 Technology.

The web 3.0 Technology is the next paradigm shift of the internet taking the best of web 2.0,including rich internet applications and social media.

Information is searched for, filtered, personalized and delivered to end users based on preferences, biofeedback and location.

Semantic Web is an evolving extension of the web 3.0 where information is tagged in relation to use and context so that similar information can be delivered more and effectively to humans and machines.

A prime example of a Web 3.0 technology is "natural-language search", which refers to the ability of search engines to answer full questions such as 'which US Presidents died of disease?'.In some cases, the sites that appear in the results do not reference the original search terms, reflecting the fact that the web knows, for instance, that Reagen was a US President, and that Alzheimer's is a disease.

XI. CONCLUSION

Web personalization is the process of customizing the content and the structure of a Web site to the specific and individual needs of each user, without requiring from them to ask for it explicitly.

Enterprises expect that by exploiting the information hidden in their Web server logs they could is cover the interactions between their Web site visitors and the products offered through their Web site. Using such information, they can optimize their site in order to increase sales and ensure customer retention .Apart fromWeb usage mining, user profiling techniques are also employed in order to form a complete customer profile. Lately, there is an effort to incorporate web content in the recommendation process, in order to enhance the effectiveness of personalization. Thus web personalization has made web activities user centric and has made the users an integrated part of the web environment.

REFERENCES

[1] Agrawal R. and Srikant R., “Privacy preserving data mining, In Proc. of the ACMSIGMOD Conference on Management of Data”, 2000.

[2] Berners - Lee J, Hendler J, Lassila. O., “The Semantic Web. Scientific American”, vol.184, pp34-43, 2001.

[3] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J., “Measuring the accuracy of sessionizers for web usage analysis, In Workshop on Web Mining, at the First SIAM International Conference on Data Mining”, 7-14, 2001.

[4] Berendt B., HothoA. And Stumme G., “Towards semantic webmining. In Proc. International Semantic WebConference (ISWC02)”, 2001.

[5] Cecconi A, Galanda M, “Adaptive Zooming in Web Cartography. In Proceedings of SVG Open 2002 (Zurich, Switzerland), pp787-799, 2001.

[6] Chen L, Sycara K. “A Personal Agent for Browsing and Searching. In Proceedings of the 2nd International Conference on Autonomous Agents”, Minneapolis/St. Paul, May 9-13, pp132-139, 1998.

[7] Desikan P. and Srivastava J., “Mining Temporally Evolving Graphs. In Proceedings of Web KDD- 2004 workshop on Web Mining and Web Usage Analysis”, 2004.

[8] EirinakiVazirgiannis M., “Web mining for web personalization. ACM Transactions on Internet Technology (TOIT)”, 3(1), 1-27, 2003.

[9] Ghani, R. and A. Fano. “Building Recommender Systems using a Knowledge Base of Product Semantics in Proceedings of the Workshop on Recommendation and Personalization in E-Commerce”, 2ndInternational Conference on Adaptive Hypermedia

and Adaptive Web Based Systems., p. 11-19, Malaga,Spain, 2002.

[10] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations”, Vol. 1, Issue 2 pp. 12-23, January 2000.

[11] Kargupta H., Datta S.,WangQ., and Sivakumar K., “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, In Proc. of the 3rd ICDM IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL, 2003.

[12] Mobasher, B., Web Usage Mining and Personalization, in Practical Handbook ofInternet Computing, M.P. Singh, Editor. 2004,CRC Press. p. 15.1-37.

[13] Maier T., “A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. In Proc. of “WebKDD-2004 workshop on Web Mining and Web Usage Analysis”, part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA, 2004.