Abstract- Web mining is the most important application of data mining techniques to extract knowledge from web data including the documents, hyperlinks between documents, usage logs of web sites etc. Web mining has been explored to a vast degree and different techniques have been proposed for a huge variety of applications that includes search engine enhancement, optimization of web services, Business Intelligence, B2B and B2C business etc. Most research on web mining has been from a ‘process-centric’ point of view which defined web mining as a sequence of tasks. In this paper, we highlight the significance of studying the evolving nature of the web pattern analysis (WPA). 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. A Web usage mining system performs five major tasks: i) data collection ii) information filtering iii) pattern discovery iv) pattern analysis and visualization techniques, and v) Knowledge Query Mechanism (KQM). Each task is explained in detail and its related technologies are introduced. The web mining research is a converging research area from several research communities, such as database system, information retrieval, information extraction and artificial intelligence. In this paper we implement how the usage mining techniques can be applied for the customization i.e. web visualization.

Keywords: Usage Mining, Pattern Analysis, Information Filtering, KQM, Visualization Techniques.

I. INTRODUCTION

With the significantly quick and volatile growth of information available over the Internet, World Wide Web has become a very powerful platform to store, propagate 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 usage mining research has encountered a lot of challenges: web security, privacy of user, user authentication issues etc. As a result, web users are always drowning in an “ocean” of information and facing the several 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].

Figure 1: Taxonomy of Web Mining

II. WEB MINING TECHNIQUES

The web mining techniques can be divided into a major three areas as follows; 1. web content mining 2. web structure mining 3. web usage mining.

2.1 Web Content Mining: The web content mining focuses on the information discovery/retrieval of the useful information from the Web documents. It may consist of audio, text, images or structured records such as tables and lists. The research activities in the field also involve using techniques used are from other disciplines like, Information Retrieval (IR), Information Extraction (IE) or Natural Language Processing (NLP) and Knowledge Extraction from Image can be a challenging area of web content mining.

2.2 Web Structure Mining: The web structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. The web structure mining as a process of discovering structure information from the Web. This type of mining can be performed either hyperlinks nor document structure.(i) Hyperlinks: A hyperlink is a structural unit that connects a web page to different location either within the same web page or to a different web page.(ii)Document structure: 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.

2.3. Web Usage Mining: The web usage mining is relative independent but is not isolated category, which primarily describes the web usage mining techniques that to discover the users interesting usage patterns from Web data and try to predict the users’ behaviors, in order to understand and better serve the needs of Web based applications. Usage data captures the
III. PATTERN ANALYSIS CLASSIFICATIONS

3.1. Information Retrieval: Information Retrieval (IR), which is a pull technique, rests on need expression of an individual through a query formulated in a more or less structured free language. However, in Information Retrieval, the real intention of the user is not always obvious in his manner of formulating his query and that can generate ambiguities on the sense of words that it contains. Many solutions exist in order to precise the sense of a query: the method of relevance feedback that uses the user relevance judgments on information to reformulate his query and thus to refine his research; the use of long term profile concept and short term profile (or context) concept of a user, to interpret his queries in order to re-evaluate and re-order the search results; the use of contextualization and individualization concepts for a personalized search.

3.2 Information Filtering (IF), which is a push technique, is a relatively passive task because the user does not explicitly formulate his needs through a query, as it is the case in IR. In Information Filtering, we rather use a representation of the user called user profile to send information to him. There are several methods of filtering (i) cognitive or content-based filtering which uses the description of information contents for determining to which users profiles it corresponds; (ii) social or collaborative filtering which uses the users judgments concerning a set of information to carry out recommendations; (iii) demographic filtering which uses users demographic data (age, profession, etc.) to make recommendations. Information Filtering systems learn user profiles from their interaction with systems and then use this information to analyse new documents.

3.3. Knowledge Query Mechanism (KQM): SQL-like query mechanism for querying the discovered knowledge from patterns. The Microsoft SQL based server is type of relational database system. The basic query language is the Transact-SQL, which is the implementation of the ISO / ANSI standard SQL (Structured Query Language) used by the Sybase and IMS.

3.4. Visualization Techniques: As the volume of data collected and stored in databases grows, there is a growing need to provide data summarization (e.g., through visualization [1,2]), identify important patterns and trends, and act upon the findings. The basic idea of our visual data mining techniques is to represent as many data items as possible on the screen at the same time by mapping each data value to a pixel of the screen and arranging the pixels adequately. The major goal is to evaluate our visual data mining techniques and to compare them to other well-known visualization techniques for multidimensional data. It is developed the WebViz system for visualizing WWW access patterns. A Web path paradigm is proposed in which sets of server log entries are used to extract subsequences of Web traversal patterns called Web paths. WebViz allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions. The web is visualized as a directed graph with cycles, where nodes are pages and edges are (inter-page) hyperlinks.

The visualization technique used may be classified into different categories
- Standard 2D/3D displays, such as bar charts and x-y plots as used in visualization system.
- Geometrically transformed displays, such as landscapes and parallel coordinates as used in Scalable Framework.
- Icon-based displays, such as needle icons and star icons as used in MGV.
- Dense pixel displays, such as the recursive pattern and circle segments techniques (see figures 2 and 3) [4] and the graph skecthes as used in MGV.
- Stacked displays, such as treemaps [5] [6] or dimensional stacking [7]

3.4.1. Standard 2D/3D displays: There is a large number of visualization techniques which can be used for visualizing the data. In addition to standard 2D/3D-techniques such as x-y (x-y-z) plots, bar charts, line graphs, etc., there are a number of more sophisticated visualization techniques. The classes correspond to basic visualization principles which may be combined in order to implement a specific visualization system.

3.4.2. Geometrically-Transformed Displays: geometrically transformed display techniques aim at finding “interesting” transformations of multidimensional data sets. The class of geometric display techniques includes techniques from
exploratory statistics such as scatterplot matrices [8] [9] and techniques which can be subsumed under the term "projection pursuit" [10]. Other geometric projection techniques include Projection Views, Hyperslice, and the well-known Parallel Coordinates visualization technique. The parallel coordinate technique maps the k-dimensional space onto the two display dimensions by using k equidistant axes which are parallel to one of the display axes. The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension. Each data item is presented as a polygonal line, intersecting each of the axes at that point which corresponds to the value of the considered dimensions (see figure 4).

3.4.3 Iconic Displays: Another class of visual data exploration techniques are the iconic display techniques. The idea is to map the attribute values of a multi-dimensional data item to the features of an icon. Icons can be arbitrarily defined: They may be little faces, needle icons as used in MGV (star icons, stick figure icons, color icons and TileBar.. The visualization is generated by mapping the attribute values of each data record to the features of the icons. In case of the stick figure technique, for example, two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the angles and/or limb length of the stick figure icon. If the data items are relatively dense with respect to the two display dimensions, the resulting visualization presents texture patterns that vary according to the characteristics of the data and are therefore detectable by preattentive perception.

3.4.4 Dense Pixel Displays: The basic idea of dense pixel techniques is to map each dimension value to a colored pixel and group the pixels belonging to each dimension into adjacent areas. Since in general dense pixel displays use one pixel per data value, the techniques allow the visualization of the largest amount of data possible on current displays (up to about 1,000,000 data values). If each data value is represented by one pixel, the main question is how to arrange the pixels on the screen. Dense pixel techniques use different arrangements for different purposes. By arranging the pixels in an appropriate way, the resulting visualization provides detailed information on local correlations, dependencies, and hot spots. Well-known examples are the recursive pattern technique and the circle segments technique. The recursive pattern technique is based on a generic recursive back-and forth arrangement of the pixels and is particular aimed at representing datasets with a natural order according to one attribute (e.g. time series data). The user may specify parameters for each recursion level, and thereby controls the arrangement of the pixels to form semantically meaningful substructures. The base element on each recursion level is a pattern of height hi und width wi as specified by the user. First, the elements correspond to single pixels which are arranged within a rectangle of height h1 and width w1 from left to right, then below backwards from right to left, then again forward from left to right, and so on. The same basic arrangement is done on all recursion levels with the only difference that the basic elements which are arranged on level i are the pattern resulting from the level (i − 1) arrangements. In figure 2, an example recursive pattern visualization of financial data is shown. The idea of the circle segments technique is to represent the data in a circle which is divided into segments, one for each attribute. Within the segments each attribute value is again visualized by a single colored pixel.

The arrangement of the pixels starts at the center of the circle and continues to the outside by plotting on a line orthogonal to the segment halving line in a back and forth manner. The rational of this approach is that close to the center all attributes are close to each other enhancing the visual comparison of their values. Figure 3 shows an example circle segment visualization of the same data (50 stocks) as shown in figure 3.

3.4.5 Stacked Display: Stacked display techniques are tailored to present data partitioned in a hierarchical fashion. In case of multidimensional data, the data dimensions to be used for partitioning the data and building the hierarchy have to be selected appropriately. An example of a stacked display technique is Dimensional Stacking. The basic idea is to embed one coordinate systems inside an other coordinate system, i.e. two attributes form the outer coordinate system, and two other attributes are embedded into the outer coordinate system, and so on. The display is generated by dividing the outmost level coordinate systems into rectangular cells and within the cells the next two attributes are used to span the second level coordinate system. This process may be repeated one more time. The usefulness of the resulting visualization largely depends on the data distribution of the outer coordinates and therefore the dimensions which are used for defining the outer coordinate system have to be selected carefully. A rule of thumb is to choose the most important dimensions first. Other examples of stacked display techniques include Worlds-within-Worlds, Treemap, and Cone Trees.

IV. CONCLUSION

As the popularity of the World Wide Web continues to increase, there is a growing need to develop tools and techniques that will help improve its overall usefulness. Since one of the primitive goals of the Web is to act as a world-wide distributed information resource, a number of efforts are underway to develop techniques that will make it more useful in this regard. The term Web mining has been used to refer to different kinds of techniques that encompass a broad range of issues. However, while meaningful and attractive, this very broadness has caused Web mining to mean different things to different people and...
there is a need to develop a common vocabulary for all these efforts. Next, we concentrated on the aspect of Web mining which focuses on issues related to understanding the behavior of Web users, called Web usage mining. We provided a detailed study of the efforts in this area, even though the study is short because of the area’s newness. We provided a general pattern analysis of a system to do Web usage mining, and we have identified some of the issues and problems in this area that may require further research and development. Future work will involve the tight integration of visualization techniques with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation and computer modelling. Integration of visualization techniques and these more established methods would combine fast automatic data mining algorithms with the intuitive power of the human mind, improving the quality and speed of the visual data mining process. Visual data mining techniques also need to be tightly integrated with the systems used to manage the vast amounts of relational and semistructured information, including database management systems, data mining and data warehouse systems. The ultimate goal is to bring the power of visualization technology to every desktop to allow a better, faster and more intuitive exploration of very large data resources.

REFERENCES