A Framework for E-business Web Designing Based on Web Usage Mining: A Case Study

Babak Sohrabi, Babak Abedin
Department of management
University of Tehran
University of Tehran, Jalal-Ale-Ahmad High way, Tehran
Iran
Tel: +98(21) 2549950 , Fax: +98(21) 2555850

Abstract

Website plays a significant role in success of an e-business. It is the main start point of any organization and corporation for its customers, so it's important to customize and design it according to the online behavior of web site visitors. In this paper, we will introduce web mining, as a new field of research in data mining and knowledge discovery, and will focus on web usage mining to extract useful knowledge about website usage pattern. Then, we will use web usage mining and its techniques for an Iranian governmental website and explain how it's used to decrease the cost of web site development costs and enhance customization and ease of use according to online users' behavior.

Keywords: web usage mining, website design, knowledge discovery, customization

1 Introduction

With the proliferation of the WWW, providing more intelligent Websites has become a major concern in the e-business industry. Recently, this trend has been even more accelerated by the success of Customer Relationship Management (CRM) in terms of product recommendation and self after service, etc. Giving more intelligence to e-commerce sites is popularly recognized as one of the effective strategies that increase customer satisfaction because they react intelligently and can give a personalized response to each customer [1].

The World Wide Web is an immense source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world [4].

This bunch of data has hidden various knowledge about different aspects of organization's website. Mining these data and discovering the hidden knowledge will show to an organization how its website is performing and how is used by its customers and visitors. The results will extract new opportunities and also show weak points of the e-business to decision makers and will lead them to make better decisions.

Here, data mining is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from data in large databases. Data mining is the principle core of knowledge discovery process, which also includes data integration, data cleaning, relevant data selection, pattern evaluation and knowledge visualization. Traditionally, data mining has been applied to databases. The wide spread of the World-Wide Web technology has made the large document collection in the World-Wide web a new ground for knowledge discovery. [3,4].

Simply, mining data in this new ground is called web mining. It refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from Web documents and services [5].

Web-mining is an increasingly important and very active research field which adapts advanced machine learning techniques for understanding the complex information flow of the World Wide Web [6]. Also , due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce, the area of research is so huge today [7].

In this paper, an introduction to web mining and its techniques is provided. Then, defining real case in Iran, we will use web mining to discover online behavior of web users and the use the results for future development plans of the website.

2 Web mining review

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web [8].

Fig.1 shows taxonomy of web mining [5, 2]. It can be broadly categorized as:

• Web Content mining of multimedia documents, involving text, hypertext, images, audio and video Information. • Web Structure Mining of inter-document links, provided as a graph of links in a site or between sites. • Web Usage Mining of the data generated by the users' interactions with the Web, typically represented as Web server access logs, user pro8les, user queries and mouse-clicks. This includes trend analysis (of the Web dynamics information), and Web access association/sequential pattern analysis ; source data mainly consist of the (textual) logs that are collected when users access Web servers,; typical applications are those based on user modeling techniques, such as Web personalization, adaptive Web sites, and user modeling.

Among this taxonomy, we will concentrate on web usage mining and will explain it's capabilities in web site design and personalization.

3 Web usage mining applications

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web [14].

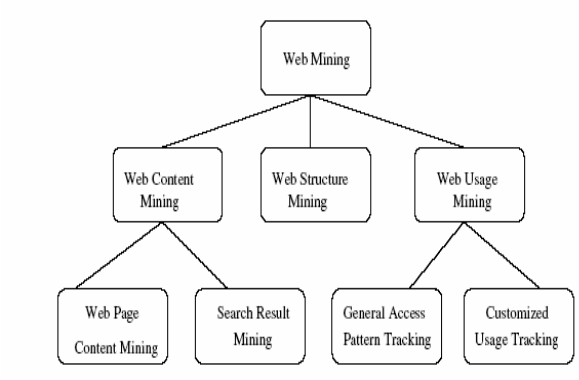

Fig.1: a taxonomy of web mining [4]

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize Web users). This information can be exploited later to improve the Web site from the users' viewpoint. The results produced by the mining of Web logs can used for various purposes: (i) to personalize the delivery of Web content; (ii) to improve user navigation through pre-fetching and caching; (iii) to improve Web design; or in e-commerce sites (iv) to improve the customer satisfaction [8,9].

4 Web usage mining data source and process

A web site consists of a hyperlinked set of pages. Fig.2 represents such a web site where the nodes are web pages and lines are hyperlinks. Organizations collect large volumes of data and analyze it to determine the life time value of customers, cross marketing strategies across products and effectiveness of promotional campaigns.

In the Web, such information is generally gathered automatically by Web servers and collected in server or access logs. Analysis of server access data can provide information on how to restructure a Web site for increased effectiveness, better management of workgroup communication, and analyzing user access patterns to target ads to specific groups of users [10]. The order in which visitors choose to view pages indicates their steps through the buying process. The similarities and differences in navigational behavior of various classes of visitors, such as new visitors vs. repeat visitors,

purchasers vs. non-purchasers, 1st time purchasers vs. repeat purchasers, could hold clues towards improving the web site design, offer personalization opportunities, and help streamline the e-commerce environment [11].

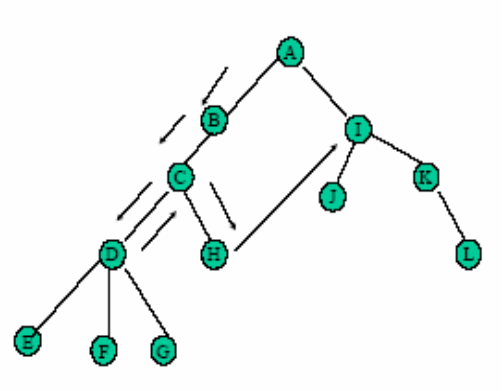

Fig.2 : A sample path of a website [11]

The overall process of web usage mining is generally divided into two main tasks; data preparation and pattern discovery. The data preparation tasks build a server session file where each session is a sequence of requests of different types made by single user during a single visit to a site. The pattern discovery tasks involve the discovery of association rules, sequential patterns, usage clusters, page clusters, user classifications or any other pattern discovery method [12]. Fig.3 shows the overall process of web usage mining.

Web Usage Mining applications are based on data collected from three main sources: (i) Web servers, (ii) proxy servers, (iii) Web clients [2] and (iv) user profiles.

The server side: 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 [7]:

remotehost rfc931 authuser date "request" status bytes

The proxy side: Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that a proxy server collects data of groups of users accessing huge groups of web servers.

The client side: Usage data can be tracked also on the client side by using Javascript, Java applets, or even modified browsers. These techniques avoid the problems of users_ sessions identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviors

User profiles: however not all the users' have profile in a website, but this is another useful source for registered users to be recognized and be tracked.

5 Web usage mining techniques

Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst, such as path analysis, discovery of association rules and sequential patterns, and clustering and classification.

Discovering Association Rules: In of Web mining, an example of an association rule is the Correlation among accesses to various files on a server by a given client, For example, using association rule discovery techniques we can find the following correlations 60% of clients who accessed the page with URL/company/products/ also accessed the page company/products/product1.html. [10].

Discovery of Sequential Patterns: Given a database of time stamped transactions, the problem of discovering sequential patterns is to find inter-transaction patterns, i.e. the presence of a set of items followed by another item, in the time-stamp ordered transaction set. In Web server transaction logs, a visit by a client is recorded over a period of time.

By analyzing this information, we can determine temporal relationships among data items such as: 30% of clients who visited /company/products/product1.html/ had done a search in Yahoo, within the past week on keywords w1 and w2 [10].

Discovering classification rules: allows one to develop a profile of items belonging to a particular group according to their common attributes. This profile can then be used to classify new data items that are added to the database. In Web usage mining, classification

techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns. For example, classification on WWW access logs may lead to the discovery of relationships such as: 50% of clients who placed an online order in /company/product1 were in the 20-25 age groups and lived on the same city [8].

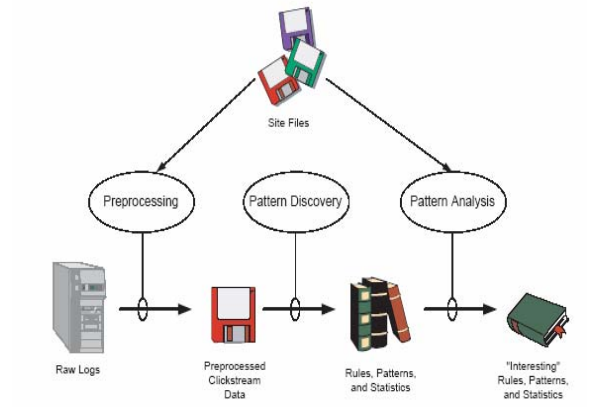

Fig.3 :high level web usage mining process[13]

Path analysis: analyzing path, usually demonstrating by graphs, which a user has passed. The most obvious is a graph representing the physical layout of a Web site_ with Web pages as nodes and hypertext links between pages as directed edges. Most of the work to date involves determining frequent traversal patterns or large reference sequences from the physical layout type of graph_ Path analysis could be used to determine most frequently visited paths in a Web site. Other example of information that can be discovered through path analysis is : 70% of clients who accessed /company/product2 did so by starting at /company and proceeding through /company/new-company-products and /company/product1[8].

6 The real-world case

In this section, we describe a real case of Civil Aviation Organization (CAO) of Iran. Among different applications of web usage mining, we concentrate on web site design issues and will try to improve the usability of website, according to real data.

Developing ICTs in recent years, most of government organizations in Iran has focused mostly on their websites. As websites are the window to the external world of organizations, the content and it's layout are so important for top managers and also CIOs. In some cases, website design, content and online traffic works as a performance measure which shows the level of development of using ICTs in an organization.

To make better decisions for web contents, CAO arranged a committee to decide about the appearance and structure of website, but after four months, there was not a common agreement about the layout of website. Because of high financial and time cost of changing the layout of the website, a list of contents for website is proposed and they are prioritized according to their usage pattern.

7 Analyzing the situation

To analyze the usability and performance of the website, we use data from Server Logs. To obtain the analysis, we use LiveStat 5.03, a product of Media House Software Inc. LiveStat is log analyzer software which can track online behavior of users in a website and generate various statistical reports. Obtaining the analysis by the LiveStat, the results were not clean. So unnecessary data was eliminated and then the addresses were grouped.

Table 1: Most accessed pages from www.cao.ir, Jul04 to Dec04 – numbers in percentage

Address	Dec04	Nov04	Oct04	Sep04	Aug04	Jul04
/index	26.7	27.9	28.5	18.9	19.5	20
/delay-value	10.4	10.6	8.5	*	*	*
/FIDS	11.1	7.54	9.8	13.2	14.6	15.3
/news	6.23	6	4.7	13.4	15.3	2.9
/timetable	4.2	4.3	2.8	11.2	9.3	10.4
/ticket	1.7	1.4	<1	*	*	*
/airport	<1	1.6	<1	2.1	<1	1.4
Total	<61.33	59.34	<56.3	<58.1	<59.7	<50.7

* Service was not available in this date

Listing the addresses, the ones which were visited more that 1 percent per month was added to the table. Table 1 shows addresses, from Jul04 to Dec04, which are visited more than %1 per months.

During last six month, the home page of CAO was frequently and dramatically changed and services and head topics were arranged in various styles. But, according to Table 1, despite the design of website in this period of time has frequently changed, but the most accessed pages are almost alike in different months.
Other services and topics in the website are rarely requested and visited.

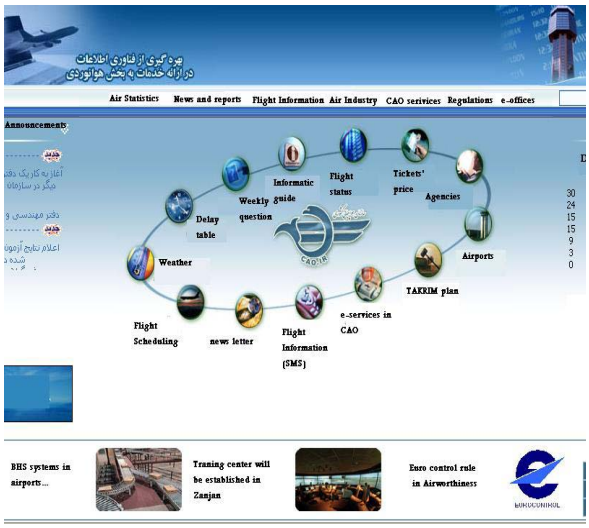

Fig. 4: home pages of www.cao.ir (Jan2005)

Table 2 shows the percentage of accessed page for Jan-2005. It makes it clear that there are some topics that have very low level access level and are not interested for most of users. So, we group them in a "detail list". As a second source, we use "CEO & CIO feedback". Due to the data, there are some topics that must be highlighted in the home page, which are: *news, announcements and annual civil aviation statistics.* So, regarding to online behavior of users and the feedbacks, we offer a framework for website layout in the next section.

8 Proposed website framework

From the previous section, the most frequent visited links were identified. From customer point of view, these links should be more highlighted in the website. We also consider the top manager's point of view. There are some topics, according to the previous section, which are critical for him and they must be highlighted in the website as well. So, we will have three sections in CAO's website layout. First, the top manager's critical topics, Second, most frequent visited links (main body) and finally less visited links.

Table 2: accessed pages in Jan05

Jan 2005	
Topic	*Percentage*
index	26.7
Delay rep.	10.4
FIDS	11.1
News	6.23
Timetable	4.2
Ticket price	1.7
Aviation stat.	<1
Air industry	<1
CAO services	<1
Regulations	<1
e-offices	<1
Weekly question	<1
Weather	<1
Newsletter	<1
TAKRIM	<1
Airports	<1
Agencies	<1

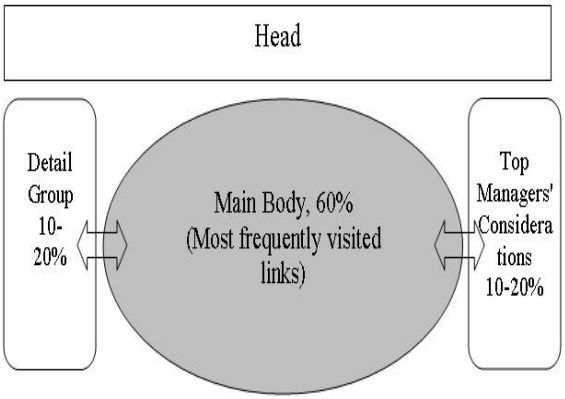

Fig.5 :A framework for website design and content arrangement

We call the third group as "detail group" in website. So, by making the design simpler to navigate, it will be easier for visitors to find required services sooner.
Fig.5 depicts the proposed framework of website design according to web usage mining and considerations of top manager.

9 Conclusion

Rapid growth of internet has made websites as huge bunch of raw data. Web mining makes it

possible to discover useful knowledge from this web data for organizations.

In this paper, web mining was used to offer an efficient and effective web design framework to attract more visitors, make web navigation easier and decrease web site design and development costs. The result took us to a framework for web designing in which, contents are arranged due to their visiting pattern and top manager's priorities.

References

[1] Kim,Wooju. Song,Yong U. Hong ,June S. (2004) Web enabled expert systems using hyperlink-based inference. *Expert Systems with Applications.* pp:1-13

[2] Michele Facca, Federico. Luca Lanzi, Pier (2004). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*. (ARTICLE IN PRESS)

[3] Hsu, Jeffrey. DATA MINING TRENDS AND DEVELOPMENTS: The Key Data Mining Technologies and Applications for the 21st Century. *Proc. of ISECON*. 2002

[4] Zaiane, Osmar R (2001). Building virtual web views. *Data & knowledge engineering*. Vol. 39. pp:143-163

[5] Arotaritei, Dragos. Mitra, Sushmita(2004). Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems*. Vol. 148. PP: 5–19

[6] Larsen, Jan. Lars Hansen, Kai. Szymkowiak Have, Anna. Christiansen, Torben. Kolenda, Thomas (2002). Webmining: learning from the World Wide Web. Computational *Statistics & Data Analysis.* 38 .517–532

[7] Eirinaki, Magdalini. Vazirgiannis, Michalis (2003). Web Mining for Web Personalization. *ACM Transactions on Internet Technology*, Vol. 3, No. 1, Pages 1–27.

[8] Cooley ,R. Mobasher , B. and Srivastava, J. (1997) Web Mining: Information and Pattern Discovery on the World Wide Web. *Proc. of the 9th IEEE Int'l Conf. on Tools with Artificial Intelligence (ICTAI'97).*

[9] DeYoung, Colin G. Spence, Ian (2004). Profiling information technology users: en route to dynamic personalization. *Computers in Human Behavior*. Vol. 20. pp: 55–65

[10] Mobasher, B. Jain, N. Han, Eui Hong (Sam). Srivastava, J (1997) .Web Mining: Pattern Discovery from World Wide Web Transactions. *Technical Report TR96-050*, Department of Computer Science, University of Minnesota.

[11] Theusinger, Christiane . Huber , Klaus-Peter (2000). Analyzing the footsteps of your customers. *Proc. of the Sixth ACM SIGKDD Internat. Conf. on Web KDD*.

[12] Ho Cho, Yoon. Kyeong Kim, Jae. Hie Kim, Soung. A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications*. 23 (2002) 329–342

[13] Srivastava, Jaideep. Cooley, Robert. Deshpande, Mukund. Tan, Pang-Ning (2000). Web Usage Mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations. ACM SIGKDD*. Vol 1. Issue 2. pp: 12-23

[14] R. Kosala, H. Blockeel, Web mining research: a survey (2000), *SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM* 2 (1)-pp:1–15