Web Mining and Knowledge Detection of usage Patterns

L . P . Sai Dhatri[1], N . Supriya[2], P . Nageswara Rao[3]
Department of Cse,Swetha Institute of Technology and Science::Tirupathi

***Abstract*—Web mining is a very hot explore issue which combine two of the start investigate region: Data Mining and World Wide Web. The Web mining explore transmit to more than a few investigate society such as Database, in sequence recovery and Artificial Intelligence. even though present exist fairly some puzzlement about the Web mining, the the majority documented move toward is to classify Web withdrawal into three areas: Web substance mining, Web formation mining, and Web tradition mining. Web substance mining focuses on the finding/repossession of the useful in sequence from the Web essence/data/papers, while the Web formation mining accentuate to the finding of how to model the underlying link structures of the Web. The difference between these two grouping isn't a very patent now and again. Web tradition mining is relation sovereign, but not inaccessible, group, which mainly portray the procedure that determine the user's institution model and endeavor to expect the user's behaviors.**

This paper is a converse support on the web mining. Besides given that an generally view of Web mining, this paper will focus on Web tradition mining. Normally speaking, Web tradition mining consists of three phases: Pre-processing, model innovation and Pattern psychoanalysis. A comprehensive report will be given for each part of them, nevertheless, extraordinary notice will be compensated to the user routing model detection and investigation. The client isolation is a new essential topic in this paper. An model of a classical Web ritual mining structure, Web SIFT, will be begin to make it easier to recognize the slant of how to apply data removal method to large Web data repositories in arrange to extract tradition patterns. Finally, along with some other interested explore problem; a short indication of the present explore work in the area of Web tradition mining is built-in

1.INTRODUCTION

It is not exaggerated to say the Web World Web is the most excited impacts to the human society in the last 10 years. It changes the ways of doing business, providing and receiving education, managing the organization etc. The most direct effect is the completed change of information collection, conveying, and exchange. Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information,and also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers, business analysts. The users want tohave the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the
problems encountered on the Web. Therefore, Web mining becomes an active and popular research field.

Web mining is the term of applying data mining techniques to automatically discover andextract useful information from the World Wide Web documents and services . AlthoughWeb mining puts down the roots deeply in data mining, it is not equivalent to data mining.The unstructured feature of Web data triggers more complexity of Web mining. Web miningresearch is actually a converging area from several research communities, such as Database,Information Retrieval, Artificial Intelligence , and also psychology and statistics as well.

As many believe, it is Oren Etzioni first proposed the term of Web mining in his paper 1996. In this paper, he claimed the Web mining is the use of data mining techniques toautomatically discover and extract information from World Wide Web documents andservices. Many of the following researchers cited this explanation in their works. In the samepaper, Etzioni came up with the question: Whether effective Web mining is feasible in practice Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has becomethe focus of quite a few research projects and papers. Some of the commercial considerationhas presented on the schedule.

a. Resource Discovery: the task of retrieving the intended information from Web.

b. Information Extraction: automatically selecting and pre-processing specific informationfrom the retrieved Web resources.

c. Generalization: automatically discovers general patters at the both individual Web sitesand across multiple sites.

d. Analysis: analyzing the mined pattern.

In brief, Web mining is a technique to discover and analyze the useful information from the Web data. The authors of claims the Web involves three types of data: data on the Web(content), Web log data (usage) and Web structure data. The authors classified the datatype as content data, structure data, usage data, and user profile data. M. Spiliopouloucategorized the Web mining into Web usage mining, Web text mining and user modelingmining; while

today the most recognized categories of the Web data mining are Web content mining, Web structure mining, and Web usage mining. It is clear that theclassification is based on what type of Web data to mine

1.1Web Content Mining

Web content mining describes the automatic search of information resource available online and involves mining web data contents. In the Web mining domain, Web content miningessentially is an analog of data mining techniques for relational databases, since it is possibleto find similar types of knowledge from the unstructured data residing in Web documents.The Web document usually contains several types of data, such as text, image, audio, video,metadata and hyperlinks. Some of them are semi-structured such as HTML documents or amore structured data like the data in the tables or database generated HTML pages, but mostof the data is unstructured text data. The unstructured characteristic of Web data forces theWeb content mining towards a more complicated approach.

The Web content mining is differentiated from two different points of view: InformationRetrieval View and Database View. R. Kosala et al. summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows thatmost of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus asfeatures. For the semi-structured data, all the works utilize the HTML structures inside thedocuments and some utilized the hyperlink structure between the documents for documentrepresentation. As for the database view, in order to have the better information managementand querying on the Web, the mining always tries to infer the structure of the Web site of to transform a Web site to become a database

1.2 Web Structure Mining

Most of the Web information retrieval tools only use the textual information, while ignore thelink information that could be very valuable. The goal of Web structure mining is to generatestructural summary about the Web site and Web page. Technically, Web content miningmainly focuses on the structure of inner-document, while Web structure mining tries todiscover the link structure of the hyperlinks at the inter-document level. Based on thetopology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.Web structure mining can also have another direction – discovering the structure of Webdocument itself. This type of structure mining can be used to reveal the structure (schema) ofWeb pages, this would be good for navigation purpose and make it possible tocompare/integrate Web page shemes. This type of structure mining will facilitate introducingdatabase techniques for accessing information in Web pages by providing a reference schema.

The detailed works on it can be referred toinformation in a third party's cloud scheme cause grave anxiety on information privacy. In arrange to offer strong the information measuring the frequency of the local links in the Web tuples in a Web table; the information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document; the information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites; the information measuring the frequency of identicalWeb tuples that appear in a Web table or among the Web tables. In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relationsmaybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlinks in the Web sites of a particular domain. This may help to generalize theflow of information in Web sites that may represent some particular domain, therefore the query processing will be easier and more efficient.

Web structure mining has a nature relation with the Web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the Web. It's quite often to combine these two mining tasks in an application

1.3Web Usage Mining

Web usage mining tries to discovery the useful information from the secondary data derivedfrom the interactions of the users while surfing on the Web. It focuses on the techniques thatcould predict user behavior while the user interacts with Web. M. Spiliopoulou [14] abstract the potential strategic aims in each domain into mining goal as: prediction of the user'sbehavior within the site, comparison between expected and actual Web site usage, adjustmentof the Web site to the interests of its users. There are no definite distinctions between theWeb usage mining and other two categories. In the process of data preparation of Web usagemining, the Web content and Web site topology will be used as the information sources,which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.

There are lots of works have been done in the IR, Database, Intelligent Agents and Topology, which provide a sound foundation for the Web content mining, Web structure mining. Web usage mining is a relative new research area, and gains more and more attentions in recent years. I will have a detailed introduction in the next section about usage mining, based on some up-to-date research works.

1.4The Usage Mining on the Web

Web usage mining is the application of data mining techniques to discover usage patternsfrom Web data, in order to understand and better serve the needs of Web-based applications. In the same paper, the Web usage mining is parsed into three distinctive phases:preprocessing, pattern discovery, and pattern analysis. I think it is an excellent approach todefine the usage mining procedure. It also clarified the research sub direction of the Webusage mining, which facilitates the researchers to focus on each individual process withdifferent applications and techniques. With the assistance of the diagram of the high-level.Web usage mining process shown in Figure 1, which is presented in , reader may

understand the architecture of the Web Usage Mining easily. I will give a detailed introduction as follows, encompassing these three-phase processing.

1.5Data Pre-processing for Mining

From the technique point of view, Web usage mining is the application of data miningtechniques to usage logs (secondary Web data) of large Web data repositories. The purpose ofit is to produce results that can be used in the design tasks such as Web site design, Webserver design and of navigating through a Web site . However, before applying the datamining algorithm, we must perform a data preparation to convert the raw data into the data

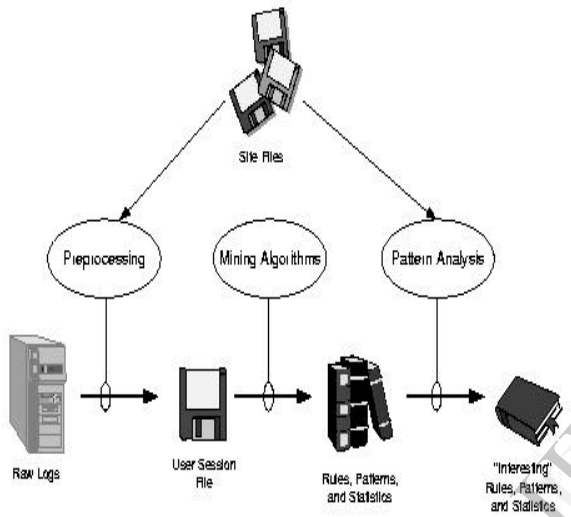

Figure 1: High Level *Web Usage Mining* Process

abstraction necessary for the further process. The data can be collected at the server-side,client-side, proxy servers, or obtained from database. For each type of data collection, the difference is not only the location, but also the available data type, the segment of population from which the data was collected and the method of implementation . The information sources available to mine include Web usage logs, Web page descriptions, Web site topology,user registries, and questionnaire . It's natural to think that the preprocess has three

different conversions: Usage converting, Content converting, and Structure converting

Since the data abstraction is very important in the data preprocess, it's necessary to clarify the definitions of the related data abstractions before the description of the different type of the data converting. The following definitions are from the Web characterization terminology & definition sheets drafty published by the World Wide Web Committee Web usage characterization activity.

User –The principal using a client to interactively retrieve and render resources or resource manifestations.

Page view – Visual rendering of a Web page in a specific client environment at a specific point in time.

Click stream – A sequential series of page view request.

User session – A delimited set of user clicks (click stream) across one or more Web servers.

Server session (visit) – A collection of user clicks to a single Web server during a user session.Also called a visit.

Episode - A subset of related user clicks that occur within a user session.

2.RELATED WORKS

As many researchers believe, it was Etzioni who first came up with the term of Web mining. He brought out a question: is it practical to mine Web data? He alsosuggested dividing the Web mining to three processes. The paper opened up a new activeresearch field. There are increasing number of researcher working on this field and do somesurveys around the data mining on the Web. The Web mining was clearly categorized as Webcontent mining, Web structure mining and Web usage mining in till 1999. Theresearch works have been well classified since then. There have been some works aroundcontent mining, and structure mining, based on the research of Data mining and InformationRetrieval, Information Extraction, and Artificial Intelligence. In the usage mining researcharea, several group s did distinguished work. R. Cooley et al. in University of Minnesota did in-depth research to all the procedure of usage mining. They proposed a miningprototype Web Miner and derived a system Web SIFT to perform the usage mining, which isrelatively practical. O. Zaiane et al. proposed the idea of how to implement the OLAPtechnique on the Web mining. Their works on the multimedia data also provided a valuablesolution for content mining. M. Spiliopoulou et al. focused on the applications ofthe usage mining. His works on the navigation pattern discovery and web site personalizationhas special meaning for the e-commerce society and the Web marketplace allocation, and willbe very helpful for both Web user and administrator. The Web Utilization Miner system is aninnovative sequential mining system. J. Borges et al. has explored some algorithms to minethe user navigation pattern in and his other papers. He proposed a data mining model toachieve an efficient mining, which captures the user navigation behavior pattern by using Ngrammarapproach.

3.DEVELOPMENT

WebSIFT: The Web Site Information Filter System

The Web Site Information Filter System is a Web usage mining framework, that uses thecontent and structure information from a Web site, and finally identify the interesting resultsfrom mining usage data [6]. The WebSIFT system is designed to perform usage mining fromthe server logs in the extended NSCA format. The preprocessing algorithms includeidentifying users, server sessions, and inferring cached page references through the use of the referrer field. Besides creating the server session, WebSIFT system performs content andstructure preprocessing, and provides the option to convert server sessions into episodes. Theserver session or episode files can be run through

sequential pattern analysis, association ruledискovery, clustering or general statistics algorithms

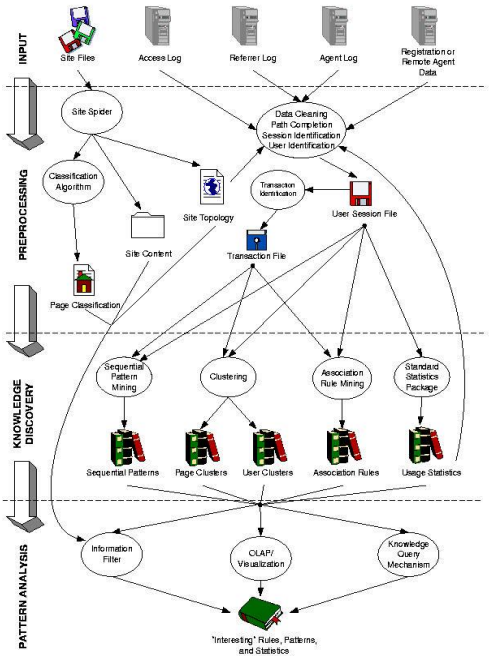

Figure 2: WebSIFT Architecture
– 3 –

The WebSIFT system is based on the WEBMINER prototypeand divides the Web usage mining process into three principal parts that are corresponding to the three phases of usage mining I described in Section 3. Figure 1 is also the high level architecture of the WebSIFT. provides a more details to show how to do usage mining in a particular Web site.

In input of the mining process includes three server logs – access, referrer, and agent; the HTML files that make up the site; and the optional data such as registration files, remote agent logs. In the preprocessing process, the input data is used to construct a user session file, to derive a site topology and to classify the pages of a site. The user session file will be converted to the transaction file and output to next phase – Pattern Discovery. Both the site topology and page classifications are fed into the information filter, which belongs to the Pattern Analysis process and makes use of the preprocessed content and structure information to automatically filter the results of the knowledge discovery algorithms for patterns that are potentially interesting . The pattern discovery phase uses the existing data mining techniques as mentioned in Section 3 (statistics, association rules, clustering, sequential) to generate rules and patterns. The discovered information is then fed into various pattern analysis tools, which includes the information filtering, OLAP, and knowledge query mechanism like SQL, to generate the final mining results.

The WebSIFT system has been implemented using a relation database, procedural SQL, and the Java programming language. Java Database Connectivity (JDBC) drivers are used to interface with the database. To the reader who is interested to know the experimental evaluation, please refer.

Personalization vs. User navigation pattern

The applications of Web usage mining can be classified into two main streams: personalizedvs. impersonalized. Personalized means learning a user profile of user modeling in adaptiveinterfaces, while impersonalized means learning user navigation pattern . With thetechnique of personalization, the Web user would prefer an intelligent Web server whichcapable to learn their information needs and preferences. On the other hand, with thetechnique of learning user navigation patterns, the information providers would be glad toview the improvement of the effectiveness on their Web sites, which results in adapting theWeb site design or by biasing the user's behavior towards satisfying the goals of the site.

Personalization

The Web provides a direct communication medium between the vendors of products andservices, and their customer with very low cost. There come tremendous opportunities for ecommercedevelopment. The Web personalization is a very important, if not necessary, partof the e-commerce. Even outside of the e-commerce, Web personalization has manyapplications.

In the context of Web mining, personalization is the provision to the individual of tailoredproducts, services, information or information relating to products or service. The goal ofpersonalization systems is to provide users with what they need or want without explicitindication . B. Mabasherbroadened the definition as the Web personalization can bedefined as any action that tailors the Web experience to a particular user, or set of users. Today, three of the major categories of existing personalization systems are manual decision rule systems, collaborative filtering system, and content-based filtering system. Mabasher compared these three kinds of system, and claimed that the new generation of Web personalization tools is attempting to incorporate techniques for pattern discovery from Web usage data.

Mabasher et al. also provided a system model for mining Web log files to discover profile for the provision of recommendations to current users based on their browsing similarities with previous users. There are several principal elements consisting of Web personalization in their framework. They are the modeling of Web objects (products, service, pages etc) and subjects (users), categorization of objects and subjects, mapping between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. The overall process of usage-based personalization is divided into two components: offline component vs. online component. The offline component is consisted of the data preparation and specific usage mining tasks that have been introduced in the previous sections. Online component uses the discovered patterns to provide personalized content to users, based on their current navigational activity. The authors introduced a personalization system based on the architecture they propose in the same paper – WebPersonalizer System. Currently, the system relies on only anonymous usage data provided by

Web server logs and hypertext structure of a site, and provides a list of recommended hypertext links to a user while browsing through a Web site. Please refer for the further details.

Some current open issues in this area are mentioned in such as the problems of the profiledata being subjective, as well getting out of date as user preferences change over time

User Navigation Pattern

The research of user navigation pattern focuses on the techniques to study the user behavior when navigating within a web site. While the World Wide Web turns to be the largest information resource available online, awareness of the user navigation preferences becomes an essential step. It is not only in the process of customizing and adapting the site's interface for individuals, but also in improving the site's static structure of the underlying hypertext system as well. Good knowledge on the way of visitors navigate in a web site could prevent disorientation and help the provider to place the information properly.

Analysis of user behavior has two aspects, one concerning the interests of the users and theaccessed information, the other concerning the way of accessing the information. The firstaspect is solved by techniques for the construct of user profiles and is not specific to the Webusage, while the second one is address by analyzing Web server logs, which falls in the field

of the Web usage mining [12]. In the paper, M. Spiliopoulou et al. proposed the exploitationof mining technology to discover access patterns with "interesting" statistical properties andpresented Web Utilization Miner (WUM) – a tool designed for the purpose. The miningmodel of WUM is in two aspects. First, it predicts that the "importance" indicators in userbehavior go far beyond than frequent access to some pages, such that the pattern discoverycan be done in the statistical domain, but also supports the subjective specification. Second,by processing aggregated sequences and applying optimization steps during the miningprocess, the high performance can be achieved.

Privacy on the Web

Due to the massive growth of the e-commerce, privacy becomes a sensitive topic and attracts more and more attention recently. The basic goal of Web mining is to extract information from data set for business needs, which determines its application is highly customer-related. As I mentioned in the above section, there exists unavoidable conflict between the Web user and the administrator in the view of privacy.

From the administrators point of view, many of the uses of data mining are innocuous, such as the data analysis to detect hidden behavioral patterns to allow supermarkets to arrange items in ways that will encourage customers to buy more of certain products or to look for seasonal buying variations. However, from individual point of view, many users believe that some applications of Web mining, may raise privacy concern, such as junk mails stuck mail account or personal information divulged during online shopping. The privacy concern has become the most critical concern for the Web user, and e-commerce developer.

The lack of regulations in the use and deployment of Web mining systems and the widely spread privacy abuses reports related to data mining has made privacy a hot iron like never before. Privacy touches a central nerve with people and there are no easy solutions. To solvethe problem, the privacy legislation is as important as the technique efforts.

Legislation efforts

In 1995, the European Union passed its Directive on Data Protection that introduces privacyprotection applying to the private sector. The Directive required member countries to adoptnational data protection laws that meet the standards of the Directive within three years . The European Union's European Data Protection Directives limits access to Internetbasedcustomer information. European companies can use data about customers to profile,but those profiles are encrypted to block out customers' names. Meanwhile, the Directiveprohibits member countries from transferring personal information to a non-member countryor to a business located in a non-member country, if the non-member country's laws do notprovide adequate protection for personal information. (European commission.The directiveon the protection of individuals with regard of the processing of personal data and on the freemovement of these data. Unfortunately in U.S. there is no unifying framework in place, although a Congress developlegislation has been recommended by U.S. Federal Trade Commission to regulate thepersonal information being collected at Web sited .

Technology development

While there are great efforts to address privacy issues by the legislative and regulation bodies,many researchers are working on new technologies to better protect consumers' privacy.Researchers at Xerox Corp.'s Palo Alto Research Centre have created an algorithm thatdesigned to keep the behavior of online shoppers hidden from Web site operator. Encirp, a vendor of marketing software designed to work in the electronicbilling environment, uses an engine that function on the *consumers' desktop* to sidestep theprivacy trap. It protects consumers' privacy by avoiding centralized data storage by theservice provider and provides personalized interface through the engine and data stored at theconsumers'desktops .

As J. Srivastava pointed in the main challenge is to come up with guidelines and rules.With the rules and guidelines, site administrator may perform various analyses on the usagedata without compromising the identity of an individual user. W3C has initiated a projectcalled Platform for Privacy Preferences (P3P), which provides a protocol try to solve theconflict between Web users and the site administrators. P3P is also in proceeding to provideguidelines for independent organization which can ensure that sites comply with the policystatement they are publishing . Please go to http://www.w3.org/P3P/

for the details.

It is expected a complete solution for the privacy issues around Web mining will not be easilyfound for many years to come. However, the process is sure accelerating with the publicattention, the efforts of the companies, the breakthrough technologies and the regulation of thegovernment agencies. The key issue for all sides is to maintain a balance in privacy concernand the use of data mining including both the results implementation and the data

collection.Only by maintaining a careful balance can the beauty of Web mining be fully explored

.

5.CONCULSION

In this paper, we survey the researches in the area of Web mining with the focus on the Web Usage Mining. Three recognized types of web data mining are introduced generally. Aroundthe key topic of this paper - usage mining, we provide detailed description of the three phasesof the process. An example of usage mining system is given to illustrate the overall usage23mining process. Moreover, the research of major applications of usage mining personalization and navigation pattern discovery are discussed. Finally, we wrap up thispaper with the most controversial topic - the user privacy.Besides the generalization of the current research work, we also try to clarify some confusionand reveal the up-to-data research issues.

REFERENCES

1. B. Berendt. Web usage mining, site semantics, and the support of navigation
2. J. Borges and M. Levene. Data mining of user navigation patterns. In Proceedings of theWEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, SanDiego, CA, USA, pages 31-39, 1999
3. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and patterndiscovery on the world wide Web. In Proceedings of the 9th IEEE International Conferenceon Tools with Artificial Intelligence (ICTAI'97), 1997
4. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide Webbrowsing patterns. Knowledge and Information Systems, 1(1), 1999
5. R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns fromWeb data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000
6. R. Cooley. WebSIFT: The Web Site Information Filter System.
7. Oren Etzioni. The world wide Web: Quagmire or gold mine. Communications of theACM, 39(11):65-68, 1996
8. R. Kosala, H. Blockeel. Web mining Research: A Survey
9. B. Mobasher, R. Cooley, J. Srivastava. Automatic Personalization Based on Web UsageMining. Communications of the ACM, Volume 43, Number 8 (2000)
10. S.K.Madria, S.S.Bhowmick, W.K.Ng, and E.P.Lim. Research issues in Web datamining. In Proceedings of Data Warehousing and Knowledge Discovery, First InternationalConference, DaWaK '99, pages 303-312, 1999
11. M.D.Mulvenna, S.S.Anand, A.G.Buchner. Personlization on the Net using Web MiningIntroduction. Communicaitons of the ACM, Volume 43, Number 8 (2000)
12. M. Spiliopoulou, L.C.Faulstich, K. Winkler. A Data Miner analyzing the NavigationalBehaviour of Web Users
13. M. Spiliopoulou. Web Usage Mining for Web site evaluation
14. M. Spiliopoulou. Data mining for the Web. InProceedings of Principles of Data Miningand Knowledge Discovery, Third European conference, PKDD'99, P588-589