

Table of Content

Content	page no
Abstract - - - - -	2
Introduction - - - - -	2
1. What is the web mining? - - - - -	3
1.1 Web Content Mining - - - - -	3
1.2 Web Structure Mining - - - - -	3
1.3 Web Usage Mining - - - - -	4
2. Mining the World Wide Web in E- Commerce architecture- -	6
2.1 Internet basic services - - - - -	6
2.2 E- commerce business services - - - - -	7
2.3 Business enabling services - - - - -	8
3. Web development algorithms and application programmes - -	9
3.1 What sort of data we need? - - - - -	9
3.2 Algorithms and application programmes - - - - -	9
4. How do people view the method and what are the possible applications of this particular method	16
4.1 What are the advantages of web mining? - - - - -	16
4.2 How should I view the web mining? - - - - -	17
Conclusion - - - - -	18
References - - - - -	19

Abstract

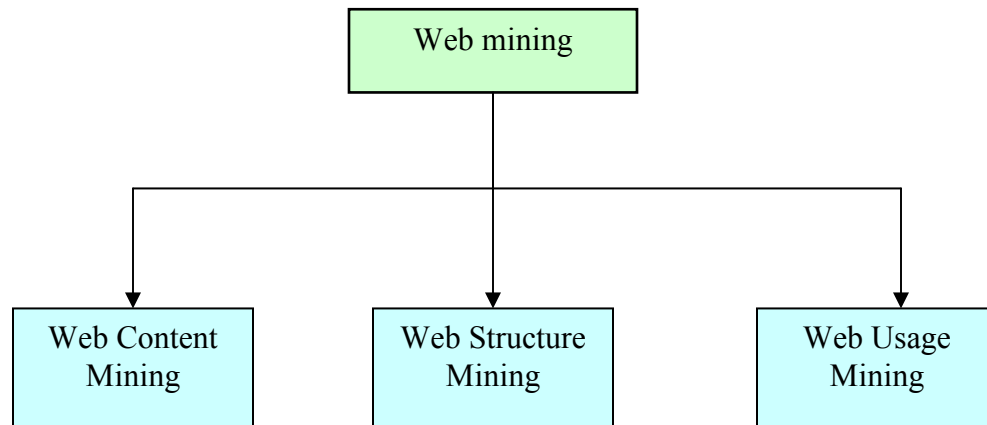
This report provides a detailed classification of the work in this area, including research effort. Also, in report described web mining taxonomy such as, web Content Mining, Web structure mining and the web usage mining. Also, more discusses the application of web usage mining and that is the process of applying data mining process of applying web mining technology for the discovery of usage Patten. In addition, in report explained a general architecture for Web Usage Mining. Moreover, explain the data cleaning, transactional identification, path Analysis, Pattern Discovery, Pattern analysis, association Rules, sequential Pattern, Clustering and Classification. Moreover, discuss the web mining in the E- commerce architecture. In that section mainly discuss three sections such as, Internet basic services, electronic commerce business services and Intra business. Also, explain the web development algorithms and application programmes. Firstly discuss sort of data we need such as weblog and CGI data and secondly disused algorithms and application programmes such as, XML, XHTML, LOGML, XGML and SQL. Furthermore, more discussed the how do people view the method and what are the possible applications of this particular method such as, electronic commerce

Introduction

In this report discuss mining the World Wide Web now undeniably the richest and one of the most dense source of information the world. Moreover, the WW W is the indubitably the richest and most numbers of source of information the world has ever seen, yet its structure makes it difficult to make use of that information in a systematic way. The methods and tools described in this article will enable developers familiar with the most common technologies of the Web to quickly and easily extract the Web-delivered information they need. In addition, the methods and tools enable developers familiar with the most common technologies of the web to quickly and easily exact the Web-delivered inferred information they need. Moreover, web pages great chalange grate challenge for effective resource and knowledge discovery such as, the web is highly dynamic information source, The web seems to be too huge for data warehousing and data mining, The web serves a broad diversity of user communities and Only the small person of the information of the Web is truly relevant or useful. Moreover, it is widely distributed financial managements, electronic commerce, government services.

1. What is web mining?

Web mining is the use of data mining techniques to automatically discover and mine information from Web documents and information services. Mining for World Wide Web can divide for three categories.



1.1 Web Content Mining

That is focuses on techniques for supporting a user in finding documents that meet a certain standard. The technique, involving mining web data contents, describes the automatic search of information resource available online such as, the HTML files, images, or E-mails and also, it already goes beyond some simple statistics of words and phrases in documents. Moreover, that is building a local knowledge base data model on the web. “Store locally abstracts characterisations of web pages. A query language enables to query the local repository at several levels of abstraction. As a result, the query the system may have to request pages from the web in more detail is needed” (Han, Zaiane, 2000). In addition, database sight of Web data is used to have the most useful information management and querying on the Web. Also, information retrieval techniques are taken to work with the unstructured data mining. Moreover, the web mining always tries to understand the structure of the Website to transform a Website to become a database.

1.2 Web Structure Mining

Web structure mining is a research field focused on using the analysis of the link structure of the web, and one of its purposes is to identify more preferable documents. In addition, that is exploiting hyperlink structure; as a result, its aims at developing techniques to take advantages of the collective judgement of web page quality.

1.3 Web Usage Mining

Web Usage mining focuses on techniques to study the user behaviour when navigating the web. According to user mining software developers develop the application of data mining and knowledge discovery techniques to discover patterns from e-commerce and clickstream data. Also, analysing and exploring regularities in Web log records can recognize potential customers for electronic commerce and as well as improve web server system, and enhance the quality and delivery of Internet information services to the end user. The web usage mining includes two data mining technique such as, WEBMINER and WebAnalyst.

Web Analyst is an intelligent e-Commerce application server. It increases the functionality of the existing web server by adding data and text mining capabilities.

WEBMINER is a Web usage mining system integrating techniques to discover association rules, sequential patterns, and classification rules from WWW transaction data (Mobasher, 1995-2003). The system includes a knowledge query mechanism and new algorithms for inferring and identifying single user sessions and transactions from missing data. In figure(1) displays general architecture of Web Usage Mining

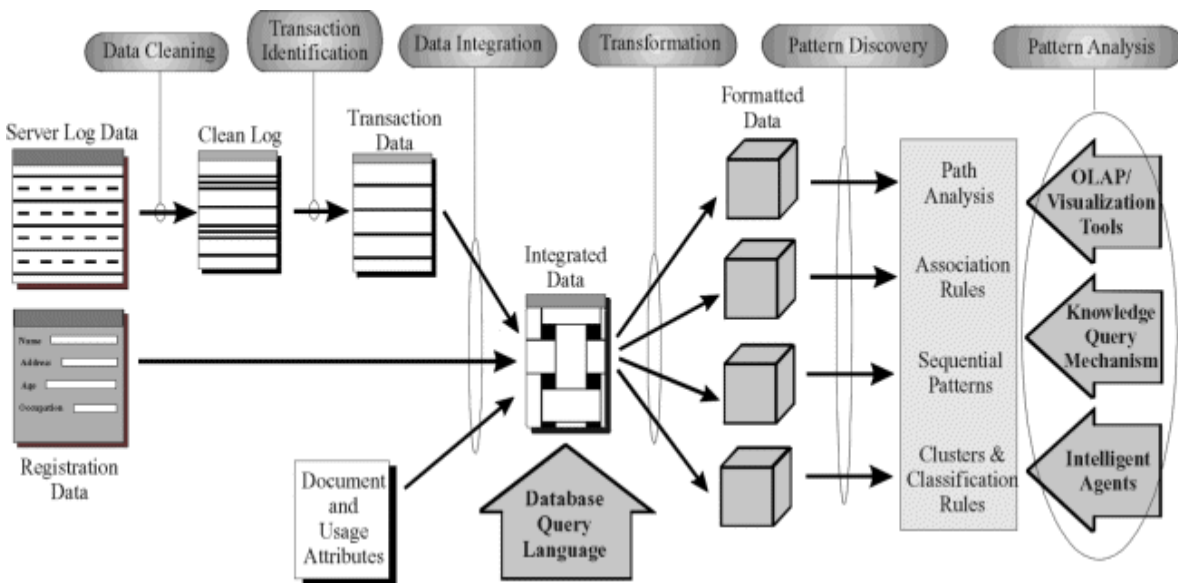


Figure (1.1)

A general architecture for Web Usage Mining (Mobasher, 1997)

In table (1.1) describe the components of web usage mining architecture

	Description
Data Cleaning	Data cleaning is the first step performed in the Web usage mining process and that is eliminating the impact of the irrelevant items to the analysis result.
Transactional identification	That is creating the meaningful clusters of references for each user.
Path Analysis	Most frequent paths traversed by users and it is distribute user session durations.
Pattern Discovery	That is the main part of the Web mining, which converge the algorithms and techniques from data mining, statistics, and machine learning categories.
Pattern Analysis	That is the final stage of the Web usage mining. The main purpose of this process is to eliminate the irrelative rules or patterns
Association Rules	This technique can be used to discover unordered connection between items found in a database of transactions.
Sequential Pattern	That is helping web market to predict the future trend and this technique intends to find the inter-session pattern
Clustering and Classification	Clustering is a technique to group together data items with the similar characteristics and the classification is a technique to map a data item into one of several predefined classes.
Server logs data	Server log data is an input of the mining process

Table (1.1)

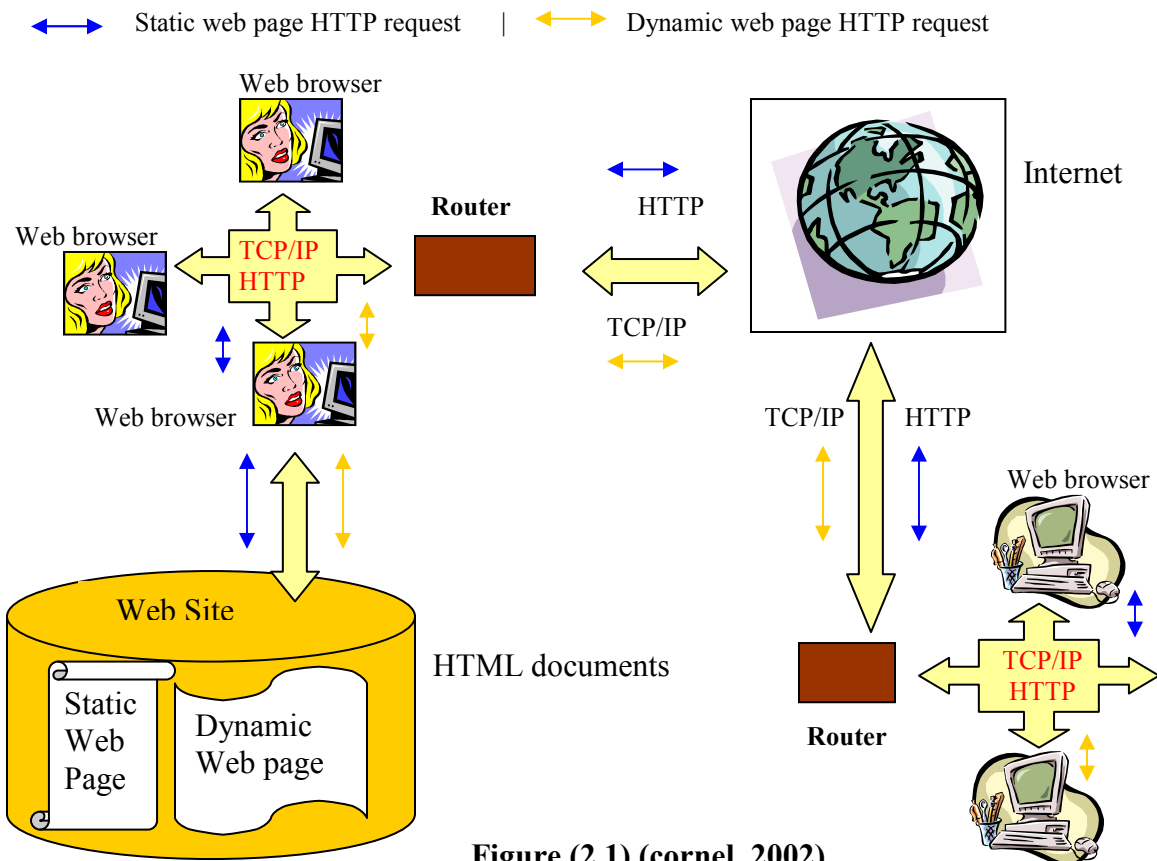
2. Mining the World Wide Web in E-commerce architecture

Web Analyst helps to improve knowledge of Electronic commerce visitor interests by collecting and analysing information generated by interactions with their website, such as clickstream data, search requests, and cookies. In addition, it can use the well-known knowledge to rank the resources by their relevance to the web user's interests. It is presenting a user request for information with the best matching issues in a higher visitor-to-customer conversion rate for your e-business.

Moreover, many companies deal with both technological and marginal issues. Technological issues include the hardware and software components that provide the better support for reliable and secure e-commerce transaction. Marginal issues range from establishing partnerships with suppliers, vendors, and distributors to the design and development well known business plans. E-commerce architecture can be dividing it into a three layers such as, Internet basic services, E-Commerce business services and business Enabling Services.

2.1 Internet basic services

The Internet provides the services that facilitate the transaction of data and information between computers. In figure (2.1) shows the relationships between internet basic services.



Web browser: - That is a end user application used to brows or navigate through the Internet. A client uses the web browser to request web pagers from a web server. For example; Microsoft Internet is explorer and Netscape Navigator.

Static web page: - Static Web pages are used to display information that does not change much over time. Example of Statistic is Standard price list.

Dynamic Web page: - Web pages are used to present information that does not change over time. Also, the dynamic page whose contents are modified to an end user's needs each time the end user request page; for example, an online ordering system.

HTTP (Hypertext Markup protocole):- That is a standard protocol used by the Web browser and Web server to communicate that is to send request and reply between server and browser.

TCP/IP:- The basic network protocol that determines the rules used to create and route the packet of data between computers in the same or different network.

Router: - That is a hardware equipment that connects multiple and diverse network.

2.2 E- commerce business services

E-Commerce business services include three main e-commerce styles such as, Business to business, Business to consumer and Intra-business.

Business-to-Business is electronic commerce between businesses. It covers the online ordering, order taking, product delivery, product support and online procurement.

Business to Consumer mean is electronic commerce between business and consumer. It is consist of personalization and real-time marketing applications. Personalization applications are doing dynamic update of customer profiles, On-line/off-line customer classification and building customer profile from usage history. In addition, real-time marketing applications are creating instant cross sales and up sales, instant advertisements and real-time recommendation.

Intra business is hold by internal electronic commerce activities, number of which includes interaction among employers and their employees.

2.3 Business enabling services

Business enabling services provide additional support for business transaction. Normally business-enabling services include web development, database integration, personalization, usability testing, transaction processing and search services

- Web development** : - That is providing the means by which to add business logic to web pages. Adding business logic into web pages using by web based programming environments such as VB script and JavaScript as well as these programs help create dynamic web pages in e-commerce web sites.
- Database integration** :- To integrate the database using SQL and Oracle. In addition, business transaction data is stored in database. Normally e-commerce business third party vendors provide solutions corporate databases into a company Web sites
- Usability testing** : - Usability testing is concerned with ensuring that the web site features and services are presented in a user-friendly manner. That is identify is the search functions hard to find? Are the options presented logically?
- Personalization** : - Normally personalization makes the site user friendly to attract more users coming back to the web sites. Also, its characteristics are concerned with customizing web pages for individual users.
- Search services** : - That is an enable web sites searches the contents such as, customer supports data and electronic commerce applications. Search services have all electronic commerce web sites.

3. Web development algorithms and application programmes

3.1 What sort of data we need?

Mainly web mining need hypertext data structured. Hypertext data containing text, markups and linkages. Also, the sources of data were of the old area of business such as, transaction records, point-of-sale terminals, inventory databases. In addition, web server normally registers a weblog entry for every access of a web page. Also, weblog data provide information about what kind of users will access what kind of web pages and weblog information can be integrated with web content. Moreover, sources of data automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies.

Also, web mining use CGI data. CGI stands for Common Gateway Interface. In addition CGI use script files that perform specific function based in client parameters that are passed to the web server. Also, the script file's contains can be used to connect to the database and retrieved data from it .In addition, the script files is a small program containing commands written in a programming language usually PERL, C++ and Visual basic (Coronel, 2002).

Normally discuss about the data field the available data fields are specified by the HTTP protocol. In addition, the server chooses the logged format of the fields. such as Netscape. These servers log requests in the Extended Log Format.

Example of data fields

```
Client IP : 128.101.228.20
Time/Date: [17/May/2003:15:20:30 -0600]
Request : "GET / HTTP/1.0"
Status : 300
Agent : "Mozilla/4.61 [en] (WinNT; I)"
```

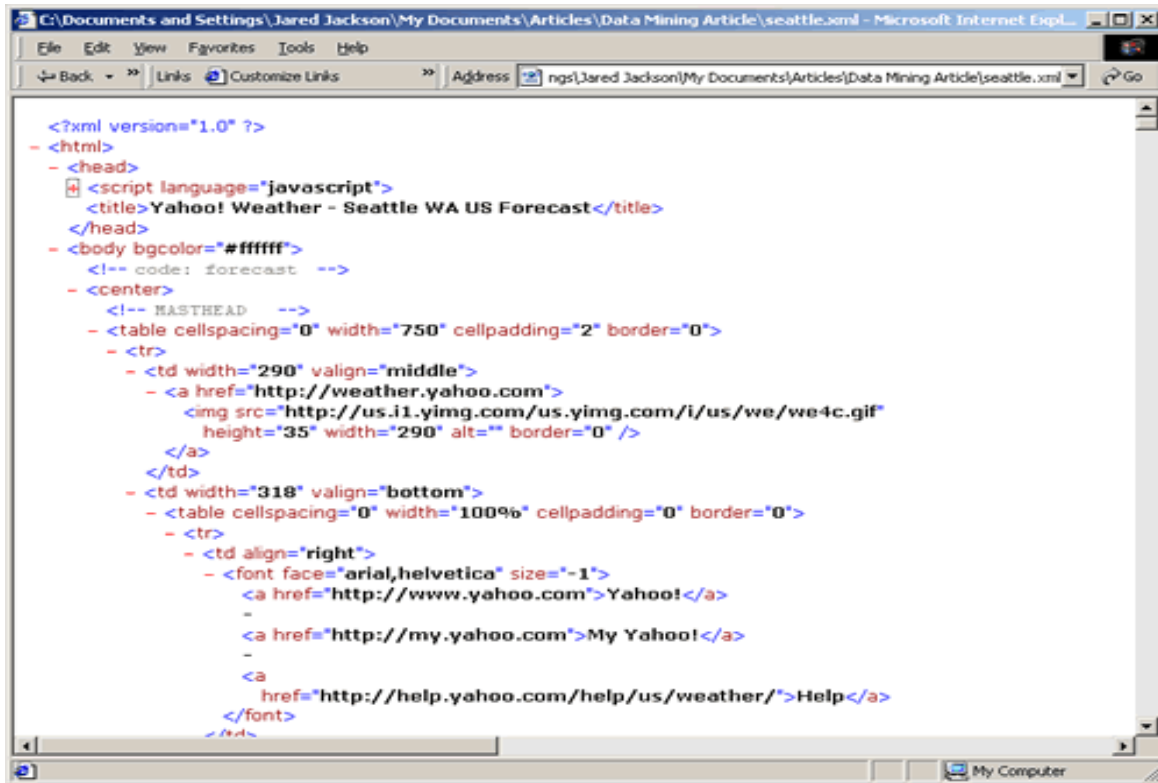
3.2 Algorithms and application programmes

The World Wide Web is driving the development of a new generation of information system providing web base applications like XML, XHTML, and web base data bases like SQL (structured Query language).

XHTML (Extensible Hypertext Markup Language) is a markup language that specifies the format of text that displayed in a web browser such as, Netscape's Communicator and

Microsoft's Internet explorer. Also, XHTML has a set of special codes that can be embedded text to add formatting and linking information. There are some advantageous aspects of XHTML for data miners such as XHTML is specified as TAGS in an XHTML document. Also, the body section of an XHTML document specifies the document content, which may include text and tags such as, (<table> and </table>) and (<p> and </p>). It is allowing the extraction process to work exclusively within a small portion of the document. In the deficiency of client-side-scripting, there is only one way to define drop-down menus. In addition, XHTML techniques are particularly useful for databases and CCS (Cascading Style Sheets).

Below In figure (3.1) displays the Yahoo Weather Web page converted to XHTML in this figure functionality provided by the Tidy library to do conversion in the method XMLHelper.tidyHTML(). This method takes in a URL as a parameter and returns an XML Document as a result.



Displays the Yahoo Weather Web page converted to XHTML

Figure (3.1) (Carickhoff Rich, January 1997)

In figure (3.2) displays the anchor is found by looking for a table containing the text appear.

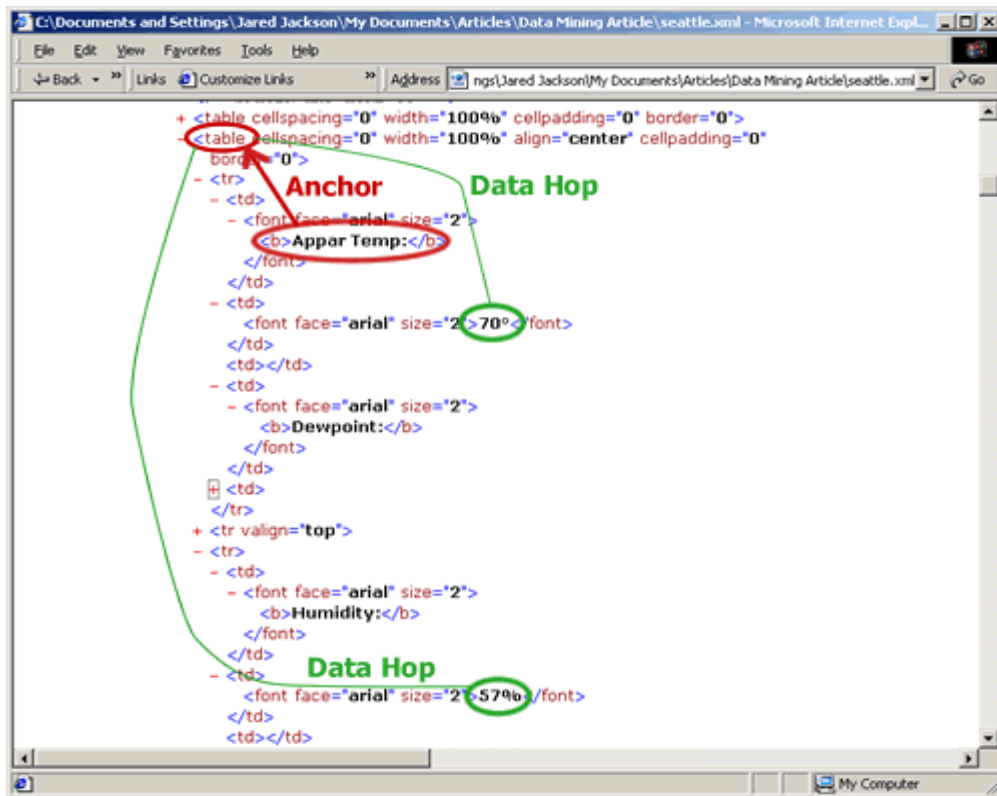


Figure (3.2) (Carickhoff Rich, January 1997)

XML (Extensible Markup Language) is a meta-language used to represents and manipulate data elements. Mainly XML has gained wider acceptance in both commercial and research establishments. In addition XML meta-language allows the definition of new tags to describe data elements such as, `<S_CODE>` and `<ProductId>`. Also, XML tags must be properly nested.

For example, `<Subject><SubjectCode>SCC337</SubjectCode></Subject>`

According to figure(3.1) and figure(3.2) all we need to do now is to repeat extraction process over and over again, merging the results into a single XML data file instead we will create one last method for merging XML files. In figure (3.3) displays the results of our Web extraction.

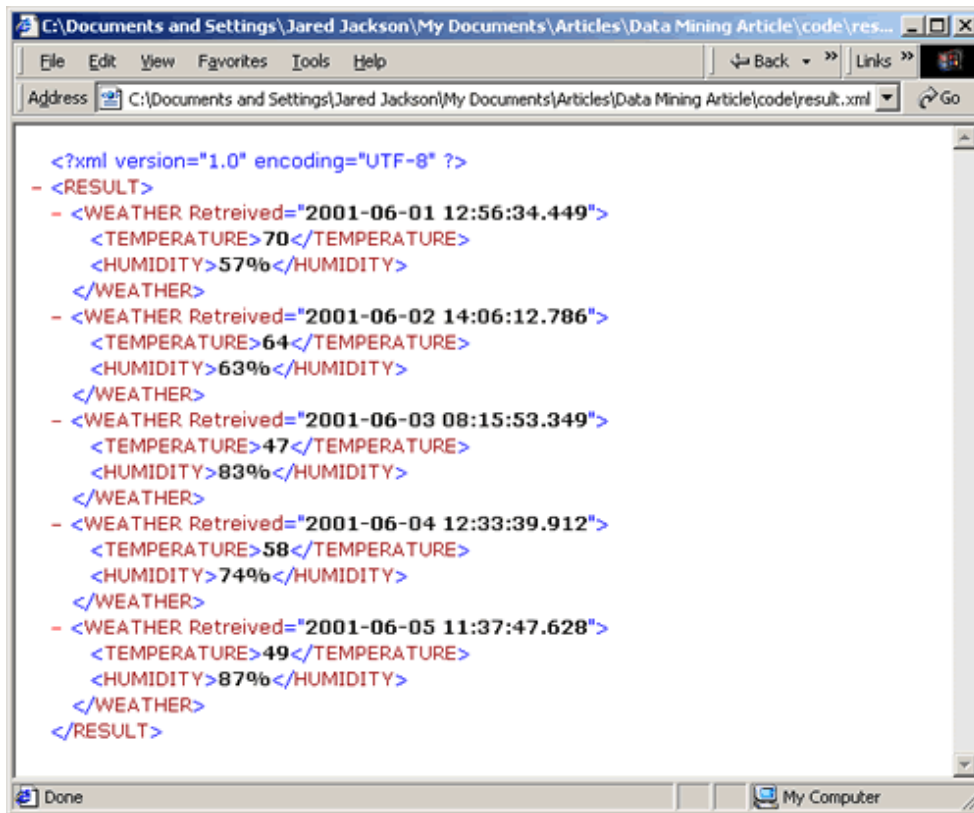
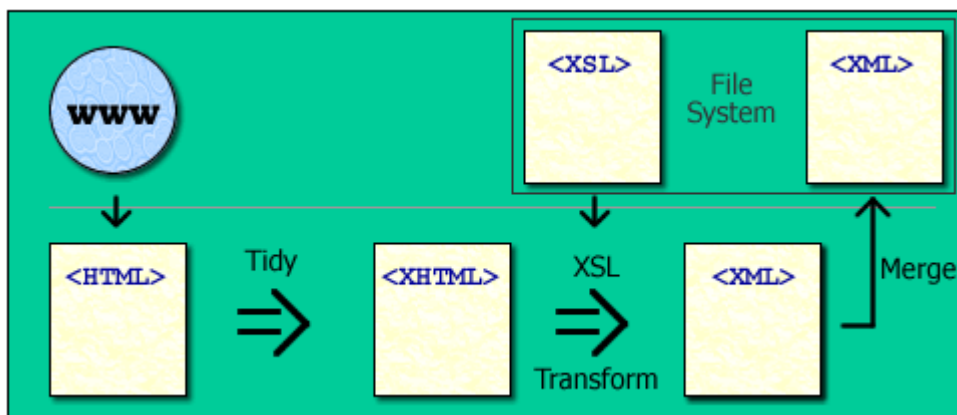


Figure (3.3) (Carickhoff Rich, January 1997)

Web pages are processed until a data set is created that can be incorporated into an existing data set. In figure (3.4) illustrates an overview of the extraction process.



Overview of the extraction process (Carickhoff Rich, January 1997)

Figure (3.4)

The database appearance is to web mining have normally focused on techniques for integrating and organizing the heterogeneous and semi-structured data in a web. Moreover, Web has a more structured and high-level collections of resources, such as in relational databases, and using standard database querying mechanisms and data mining techniques to access and analyze this information (Mobasher, 1997). There are two main types of databases in data mining such as, Multilevel Databases and the Web Query Systems.

Multilevel Databases have proposed a multilevel database approach to organizing web based information. Examples of multilevel functions are relational or object-oriented databases.

Web Query System database and language develop a standard database query languages such as SQL is a language it is use to make databases. A data is an interpreted collection of data. Database Management System (DBMS) provides mechanisms for storing and organizing data in a way consistent with a database format. Also, it allows users to access and store data without addressing the internal representation of database. However, relational databases composed of data that composed to one another. Many Relational database system use a language called SQL .

There are two XML languages such as LOGML and XGMML.

LOGML (Log Markup Language)

That is mining data that has been collected from web server logfiles. These files helped well-organized web pages. The root element of LOGML document is the logml element. LOGML elements can classified to the three sections

- 1) Report of the log graph
- 2) Report the general statistics of the web server such as top pages, top referer URLs, top visiting user agents.
- 3) Reports the user sessions.

Log elements used many global attribute

- Label - text representation of the LOGML element
- id - unique number to identify the elements of LOGML document
- html pages - number of HTML pages requested from the web server. For example, the number of html pages requested by a specific site.
- Path - the path element contains all hyperlinks that the user has traversed during the user session.

In below displays the algorithm is the report of one user session in a LOGML document:

```
<userSession name="proxy.artech.com.uy" ureferer="No referer"
entry page="http://www.cs.rpi.edu/
¢
puninj/XGMML/" start time="12/Oct/2000:12:50:11"
access count="4">
<path count="3">
<uedge source="3" target="10" utime="12/Oct/2000:12:50:12"/>
<uedge source="3" target="21" utime="12/Oct/2000:12:51:41"/>
<uedge source="21" target="22" utime="12/Oct/2000:12:52:02"/>
</path>
</userSession>
```

Figure (3.5) (John, Mukkai, Mohommed)

XGMML (Extensible Graph Markup and Modeling Language)

An XGMML document describes a graph structure. Also, the Graphs can be described as data objects whose elements are nodes and edges. Web pages are nodes and the hyperlinks are edges. Example of graph is structure of www. The node element describes a node of a graph and the edge element describes an edge of a graph. In example, represent a graph with one node.

```
<?xml version="1.0"?>
<! DOCTYPE graph PUBLIC "-//John Punin//DTD graph description//EN"
"http://www.cs.rpi.edu/
¢
puninj/XGMML/xgmml.dtd">
<graph directed="1" id="2">
<node id="1" label="Node 1"/>
</graph>
```

Figure (3.6) (John, Mukkai, Mohommed)

Web Site

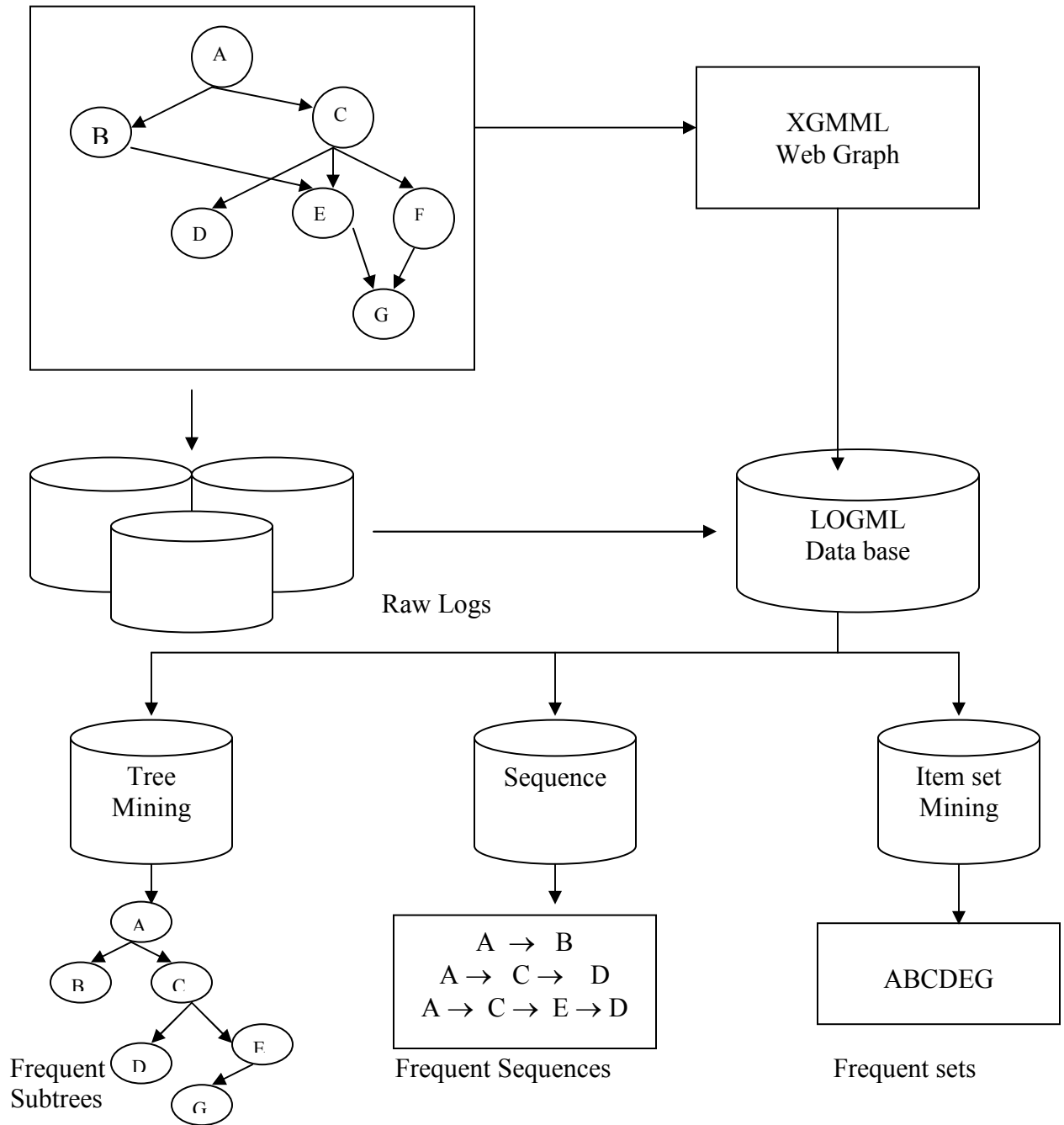


Figure (3.7)

Web Usage Mining architecture (John, Mukkai and Mohommed)

4. How do people view the method and what are the possible applications of this particular method

Most of people point of view is these applications; the use of Internet technologies facilitates the sharing of heterogeneous information in an environment that provides many benefits for software developers. In addition, For information system departments the new boundary is the use of internet technologies to provide services to customers, partners and the general public.

4.1 What are the advantages of web mining?

Advantages of web mining		
	Advantages	Description
1	Improve the visitor's experience at the website	That is mean the website to act interactively and proactively and deliver the most relevant customized resources to the visitor. In addition, combination of data and text mining techniques can help determine user interests; as a result, it is increasing the customer's satisfaction and reduce attrition.
2	Can collect information in new ways	This mean while for the number of electronic vendors the task of collecting data is just an intermediate step necessary for better targeting the vendors marketing. Also, can collect many advanced problems including low response rates, poor accuracy, and high cost
3	Match your available resources to visitor interests	That is explain meta data of e-mail Fragments from a mailing list, anything else you distribute online, information fragments you distribute online, banner ads from your client advertisers or products you sell are then stored in a database.
4	Can test the relevance of content and web site architecture	Normally, web miners, electronic commerce businessmen and internet users are looking for the characteristics of your website's content and architecture. That is help to you can increase the usability of internet architecture. Moreover, Log analyzers can help you visualize the most navigated paths trough their site, averaged over all visitors.
5	Increase the value of each visitor	This knowledge significantly increases the value of a customer for an e-business when used in individualized cross-selling and up-

		selling promotions, and thus increasing revenue.
--	--	--

Table(4.1)

Also, a number of business people and Internet users are thinking electronic commerce events are often only valid to specific domains, and the definition of certain events can vary from site to site. In addition, electronic commerce data is generally product oriented and also, there can be one, several, or no e-commerce events linked with a particular page-view. However, usage data is page view oriented. Furthermore, usage events are well defined and have consistent meaning across all Web sites.

Moreover, they believe web mining given benefit to the electronic finance applications According to the web mining electronic finance applications, personalize information aggregation on the Web and application of data mining to aggregated information. They argue data mining analysis of aggregation users such as, competitive analysis for credit card offering organizations, financial & lifestyle preferences of banking customers, segmentation analysis of brokerage customers and dynamic targeting for ads, e-mail campaigns (Cooly, Srivastava).

4.2 How should I view the web mining?

In my point of view web mining given benefit to the electronic commerce applications such as, supply chain optimization, Business to Business, Business to Consumer, customer relationship, finance and production chain optimization.

Advantage of electronic commerce is to provide quick and convenient comparison shopping, thereby increasing competition and so reducing cost. Also electronic commerce Can easier to access and potential to a global market. Also, it is improve the customer decision making.

However, electronic commerce have there are few disadvantages. In electronic commerce technology is not perfect; as a result, network unreliability is a continuing unease. According to the network failure business is become trouble. Moreover, lost privacy of the customers and vendors.

Moreover Web mining is given mining unstructured data such as understand competitive directions and combining with relational data to track new products.

Conclusion

During the report mainly discussed four main topics. Firstly, given a description about what is the data mining. Then compare and mining the World Wide Web in electronic commerce architecture. After that more detail describes about web development algorithms and applications programmes and finally, explain how people view the method and what the possible applications of this particular methods. If we go through the report we can identify Web mining is a key technology to help make sense of all of this data and better understand and facilitate electronic commerce. Also, that report mainly explains web based electronic commerce is providing a unique opportunity to collect detailed data about all aspects of the commerce activity.

References

1. (Megaputer Intelligent, 2000-2003) “Web Mining with WebAnalyst”
Available from: - <http://www.megaputer.com/products/wa/index.php3>
2. (Mousa Reshed, Sundaresan Neel, May 2-5, 2001) “Algorithms and Programming Models for Efficient Representation of XML for Internet Applications”
Available from: - <http://www10.org/cdrom/papers/542/>
3. (Greening Dan R.) “Data Mining on the Web There's Gold in that Mountain of Data”
Available from: - <http://www.webtechniques.com/archives/2000/01/greening/>
4. (John R. Punin, Mukkai S. Krishnamoorthy, Mohammed) “Web Usage Mining – Languages and Algorithms”
Available from: - <http://www.cs.rpi.edu/~puninj/LOGML/TR01-3.pdf>
5. (GRIP, 1995, 2003) “ ReseArch WEB Usage mining”
Available from: - <http://www.global-reach.com/research/usagemining/>
6. (Whang Yan) “Web mining and knowledge discovery of usage patterns - A survey”
Available from: <http://db.uwaterloo.ca/~tozsu/courses/cs748t/surveys/wang-slides.pdf>
7. (Mobasher Bamshad, Wed Jul 16) “Web Usage Mining Architecture”
Available from: - <http://maya.cs.depaul.edu/~mobasher/webminer/survey/node23.html>
8. (Robert Cooley, Bamshad Mobasher, Jaideep Srivastava) “Web Mining: Information and Pattern Discovery on the World Wide Web”
Available from: - <http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html>

Books

9. (Kamber Jiawei, 2001) “Data Mining concept and technique” printed in a USA
10. “Database System design, implementation & Management” by Rob Cornal