A PROPORTIONAL ANALYSIS ON THE ILLUSTRIOUS PRACTICES FOR THE EXTRACTION AND DISCOVERY OF HIDDEN PATTERNS - DATA AND WEB MINING

SHEILINI JINDAL

Lecturer

Department of Computer Science & Engineering

Chitkara University

Rajpura, Punjab, India

E-mail : er.sheilinijindal@gmail.com

GAURAV KUMAR

Asstt. Professor

Department of Computer Applications

Chitkara University

Rajpura, Punjab, India

E-mail : kumargaurav.in@gmail.com

ABSTRACT

In the era of information and knowledge management, Data Mining and Web Mining are the major and inclusive areas of research in the stream of computer science and information technology. This blend is sometimes referred to as Web mining that has been the focus of several recent research projects and papers throughout the globe. The term Web mining has been used in two distinct ways. The first, referred to as Web content, describes the process of

information or resource discovery from millions of sources across the World Wide Web. The second, known as Web Usage Mining, is the process of mining Web access logs or other user information user browsing and access patterns on one or more Web localities. The World Wide Web is the major source of data and information for all domains. It is not only an accessible and searchable information source but also one of the most important communication channels, almost a virtual society. Web mining is an important and challenging activity that aims to discover new, relevant and reliable information and knowledge by investigating the web structure, its content and its usage. This paper is the comparative analysis of the illustrious information extraction techniques Data Mining and Web Mining. In this paper we present various software products and suites available for information extraction using data mining and web mining.

Keywords : Data Mining, Commercial Applications, Web Mining, Web Usage Mining, Web Mining Techniques, Web Mining Software

INTRODUCTION

Web mining is the discovery and analysis of useful information extracted from the Global Network or World Wide Web. This expansive definition describes the automatic search and retrieval of information and resources available from millions of sites and on-line databases, Web content mining is also referred to as the discovery and analysis of user access patterns from one or more web servers or on-line services, web usage mining.

The growing area of research, the Web mining is divided into three different types
1. Web Content Mining
2. Web Usage Mining
3. Web Structure Mining

WEB CONTENT MINING

Web content mining is an automatic process that is far beyond simply keyword extraction. As the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in a representation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. Techniques making use of lexicons for content interpretation is still the research issue.

WEB STRUCTURE MINING

World Wide Web divulges more information rather than the information contained in documents. For example, links pointing to a specific document show the popularity of the document, whereas links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. Web structure mining, one of three categories of web mining for data, is a tool or technique that is used to identify the relationship between Web pages that are linked by information or direct link connection. The main purpose for web structure mining is to extract earlier unknown relationships between various Webpages or applications. The structure data mining provides the use for a business to link the information of its own Web site to enable navigation and cluster information into sitemaps or site hierarchy.

WEB USAGE MINING

Web usage mining is the process to extract useful information from Server Log Files or Cyber Forensic Databases. Web Usage Mining extracts out what different categories of users are looking or searching for on the Internet or Global Network. Some clients might be looking for textual data, whereas others might be searching and download multimedia objects. Web

Servers or Forensic Databases record and accumulate data about user interactions when requests for resources are received.

DATA MINING

In contrast, the process or technique called Data mining is the process of extracting and digging patterns from huge data sets by combining methods from statistics and artificial intelligence or heuristics with database management.

Data mining comprise four classes of tasks performed therein:

- Association rule learning – It is the searching for relationships between number of variables. It may include the analysis of market survey or customer purchasing pattern or behavior. It is also referred to as market basket analysis.
- Clustering – It is the task of finding and extracting groups and structures in the data in some way without making use of structures in the data.
- Classification – It is the activity of generalizing known structure to apply to the new data set. It may include the analysis of E-mail whether it is valid or a Spam.
- Regression – It tries to extract a function that models the data with the least error.

METHODS AND ALGORITHMS IN DATA MINING

- Association rule learning
- Cluster analysis
- Constructive induction
- Data analysis

- Decision trees
- Factor analysis
- Knowledge discovery
- Neural nets
- Predictive analytics
- Reactive business intelligence
- Regression
- Structured data analysis (statistics)
- Text mining

APPLICATIONS OF DATA MINING

- Customer analytics or Market Basket Analysis
- Cyber Forensics and Investigation
- Data Mining in Agriculture
- Law Enforcement Agencies
- Ocean Analysis and Satellite Predictions
- Meteorology
- Surveillance

COMMERCIAL DATA MINING SOFTWARE

- FastStats
- DB Miner
- Speed Tracer
- Commerce Trends
- Clementine

- Estard
- WUM
- Sawmill 5
- Funnel Web Pro
- Knowledge Studio
- Net Analysis
- DBMyne
- SAS Enterprise Miner
- SPSS Modeler
- STATISTICA Data Miner

FREE AND OPEN SOURCE DATA-MINING SOFTWARE AND APPLICATIONS

1. Carrot
2. ELKI
3. GATE
4. JHepWork
5. KNIME
6. NLTK or Natural Language Toolkit
7. Orange
8. R
9. RapidMiner
10. Tanagra
11. UIMA
12. Weka

DIVERGENCE BETWEEN DATA MINING AND WEB MINING

	Data Mining	**Web Mining**
Definition	Data mining makes use of the techniques for structure and relationships in large amount of data.	Web mining is the analysis of Web server logs of a Website or Web Application.
Description	Data mining is having five categories: • Neural Networks • Knowledge Discovery • Data Visualization • Fuzzy Query Analysis • Case-Based Reasoning	The Web server logs contain the entire collection of requests made by a potential or current customer through their browser and responses by the Web server. The information in the logs varies depending on the log file format and option selected on the Web server.
Application	The applications of Data Mining include Mailing, Telemarketing, Medical Diagnosis, Clinical Help, Credit Card Fraud, Cyber Crime and Intrusion Detection.	The applications of Web Mining includes include usability, Purchase Pattern, content marketing, user profiling.

Table-1: Difference between Data mining and Web mining

CORPORATE WORLD USING WEB MINING

- Johnson & Johnson
- GE Capital
- Fingerhut
- Procter & Gamble
- Harrah's Casino

- Advent Software, Inc
- Dell, Inc. and Intel
- Denodo Technologies
- MicroStrategy Incorporated
- Oracle Corporation

WEB CONTENT EXTRACTION / WEB MINING SOFTWARE

Web Sundew	Twin Soft
Protitero	Web Extractor
Addtoit	Automation Anywhwere
Xtractly	Connotate
Irobot Soft	Mozenda
Djunggler	30 Digits Web Extractor
QL2	Megaputer
ZOLOPAGES	Text Mining Tool
iMacros Web Automation and Web Testing	Web Icons
TS WEB.	Knowlesys Web Crawler 1.0 Custom Web Data
Knowlesys Web Data Extractor 1.0 Custom Web Data Extractor	Web Data Extractor
WebExtractor System	Concordance
R.J.C. Watt	Deep Log Analyzer
SmarterStats	WebGrab
gINET Software	NetTools Spider
QuesTronix Software	NetTools
Sobolsoft	TextPipe Pro
DataMystic	WebIESoft

Web Power Data Mining	Bitware Specialists
Web Scraper Lite	Velocityscape
Bget Software Studio	Data Record Extractor

WEB MINING COMMERCIAL SOFTWARE

11Ants Model Builder	AlterWind Log Analyzer Professional
Amadea Web Mining	ANGOSS KnowledgeWebMiner
Affinium NetInsight	ClickTracks
ConversionTrack from Antssoft	Download Analyzer
Megaputer WebAnalyst	Nihuo Web Log Analyzer
SAS Webhound, analyzes	SPSS Web Mining
Surf Pattern Visual Analyzer	The Data Miner
WebLog Expert	XAffinity(TM)
123LogAnalyzer	

FREE AND OPEN-SOURCE WEB MINING SOFTWARE

- AlterWind Log Analyzer Lite
- Analog (from Dr. Stephen Turner)
- jwanalytics
- htminer
- Visitator
- WUM : Web Utilization Miner

APPLICATIONS AND FUTURE OF DATA AND WEB MINING

Generate User Profiles - improving customization and provide users with pages, advertisements of interest

Targeted Advertising - Ads are a major source of revenue for Web portals (e.g., Yahoo, Lycos) and E-commerce sites. Internet advertising is probably the "hottest" web mining application today.

Fraud - Maintain a signature for each user based on buying patterns on the Web (e.g., amount spent, categories of items bought). If buying pattern changes significantly, then signal fraud.

Performance Management - Annual bandwidth demand is increasing ten   fold on average, annual bandwidth supply is rising only by a factor of three. Result is frequent congestion. During a major event (World Cup), an overwhelming number of user requests can result in millions of redundant copies of data flowing back and forth across the world.

Fault management - Analyze alarm and traffic data to carry out root cause analysis of faults.

Information Retrieval (Search) on the Web - Web Mining tools analyze web logs for useful customer-related information that can help personalize web sites according to user behavior. Web mining tools are also used to search the web for key words, phrases, or other content.

CONCLUSION

The World Wide Web today is the major source of data and information for all domains. The scientists and research scholars executes millions of queries to fetch the information from the Global Network. Data and Web mining are challenging activities with the aims to discover new, relevant and reliable information and knowledge by investigating the structure of data, its content and its usage. In this paper we have presented a proportional analysis on data and web mining as major techniques to dig the useful information.

REFERENCES

1. URL : http://www.galeas.de/webmining.html. Last Accessed : 22 November 2010

2. URL : http://en.wikipedia.org/wiki/web-mining. Last Accessed : 22 November2010

3. Web Mining: Information and Pattern Discovery on the World Wide Web

Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, Department of Computer Science University of Minnesota, Minneapolis, MN USA, 2010

4. Design and Implementation of A Web Mining Research Support System, Research Proposal, University of Notre Dame, by Jin Xu, M.S.,25th November 2010

5. URL : http://paginas.fe.up.pt/~ec/filesfi0405/slides/06%20WebMining.pdf Last Accessed: 30 Nov. 2010

6. Web Mining: Accomplishments & Future Directions, Jaideep Srivastava

University of Minnesota, USA URL : http://www.cs.umn.edu/faculty/srivasta.html

7. Advanced AI Techniques for Web Mining, Ioan Dzitac. IT Department Agora University, Piata Tineretului, Oradea, Romania Ioana Moisil, Hermann Oberth, Faculty of Engineering - Computer Science and Automatic Control, Lucian Blaga, University of Sibiu

8. URL : http://searchoracle.bitpipe.com/olist/Data-Mining.html Last Accessed : December 2, 2010

9. URL : http://findarticles.com/p/articles/mi_m0OOL/is_2_4/ai_99824637/ Last Accessed : December 2, 2010

10. URL : http://maya.cs.depaul.edu/~mobasher/webminer/survey/node1.html Last Accessed : December 2, 2010

11. URL : http:// www.mineit.com Last Accessed : December 2, 2010

12. URL : http://www.information-management.com/news/5458-1.html Last Accessed : December 2, 2010

13. URL : http://www.galeas.de/webmining.html PatricioGaleas.org

14. URL : http://www.sigkdd.org/curriculum.php

15. Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining"

16. URL : http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

17. Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases".