IJCSN International Journal of Computer Science and Network, Volume 4, Issue 1, February 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

Next Generation Business Intelligence Techniques in the Concept of Web Engineering of Data Mining

[1] M Vijaya Kamal, [2] P Srikanth, [3] Dr. D Vasumathi

[1] Asst. Professor, University of Petroleum & Energy Studies, Research Scholar in JNTUH Dehradun, Uttarakhand, India

[2] Asst. Professor, University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India

[3] Professor, Jawaharlal Nehru Technological University , Dehradun, Uttarakhand Hyderabad

Abstract - Web mining plays vital role in day-to-day applications to improve intelligence of web in the context of business must be able to identify useful business intelligence. To achieve our model in web engineering, we are using mining techniques for next generation business intelligence development. In this research our approach identifies the weblogs error reports using comprehensive algorithms, applies the mining techniques to detect noisy and integrates the different models, finally our information patterns satisfies the need of client inputs. For web engineering retrieval system, list of web log bugs and web architecture, the system uses mining techniques to explore valuable web data patterns in order to meet better projects inputs and higher quality web systems that delivered on time. Our research uses association and machine learning applied to web architecture model pertaining to source code mining implementation tools improves software debugging business rules for novel projects and also presents strategies for efficient study text, graph mining. Presents the Geo Tracking system to identify messages from terrorist or threat persons and also from hackers detects the negative rates and improves the high positive which increases the quality of Government Private and Public sectors.

Keyword - *Business Intelligence, Web mining, Geo-Tracking, Text Mining, Pattern Analysis*

1. Introduction

Data mining is the nontrivial process of identifying valid novel, potentially useful, and ultimately understandable patterns in data Fayyad. The most commonly used techniques in data mining are artificial neural networks, decision trees, genetic algorithm, nearest_neighbour method, and rule induction. Data mining research has drawn on a number of other fields such as inductive learning, machine learning and statistics etc. Machine learning is the automation of a learning process and learning is based on observations of environmental statistics and transitions. Machine learning examines previous examples and their outcomes and learns how to reproduce these make generalizations about new uses. Inductive learning means inference of information from data and Inductive learning is a model building process where the database is analyzed to find patterns. Main strategies are supervised learning and unsupervised learning. Statistics used to detect unusual patterns and explain patterns using statistical models such as linear models. Data mining models can be a discovery model – it is the system automatically discovering important information hidden in the data or verification model – takes an hypothesis from the user and tests the validity of it against the data.

The web contains collection of pages that includes countless hyperlinks and huge volumes of access and usage information. Because of the ever-increasing amount of information in cyberspace, knowledge discovery and web mining are becoming critical for successfully conducting business in the cyber world. Web mining is the discovery and analysis of useful information from the web. Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services (content, structure, and usage).

Two different approaches were taken in initially defining web mining: i. Process centric view web mining as a sequence of tasks ii. Data centric view web mining as a web data that was being used in the mining process. The important data mining techniques applied in the web domain include Association Rule, Sequential pattern discovery, clustering, path analysis, classification and outlier discovery. Text mining is concerned with the task of extracting relevant information from natural language text and to search for interesting relationships between the extracted entities. Text classification is one of the basic techniques in the area of text mining. It is one of the more difficult data-mining problems, since it deals

with very high-dimensional data sets with arbitrary patterns of missing data.

2. Related Work

Geo graphical systems captures manipulate analyze manage and present all types of geographical data used for geographical information science or geographical to extract large data. It digitally creates and manipulates spatial terms that may be jurisdictional. Generally geographical custom designed for an organization. One of the first applications of spatial analysis in epidemiology is the Rapport sur la marche et les effects du cholera dans paris modern geographical technology use digital information for which various digitized data creation methods are used where a hard copy map or survey plan is transferred into a digital capabilities. Data representation can be classified into discrete and continuous fields.

Web has tremendous success in building of users to identify such communities is useful for many purposes. Gibson identified web communities as core of central authorized pages linked together by hub pages. Web crawling applies the maximum flow minimum cut model to the web graph for identifying web communities flow approaches and strengths bipartite graph method followed to find a related concept of friends and neighbors was introduced who are in the cyber world would form a community. Analyzing web log data with visualization tools has evoked a lot of interest developed a web ecology visualization is to understand the relationship between web content, structure, usage over a period of time. Interactive web log site at a global level display each browsing path on the pattern displayed in an incremental manner. Google is popular search engine provides access to information from over billion web pages that it has indexed on its server. Previous search engines only concern on web content to return pages to a particular query but Google is quality quick search engine makes it most successful search engine and Google tool bar is service provider that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages.

Mining techniques like Association Rule Mining predict the association and correlation among set of items "where the presence of one set of items in a transaction implies with a certain degree of confidence. 1) Discovers the correlations between pages that are most often referenced together in a single server session/user session. 2) Provide the information: i. what are the set of pages frequently accessed together by web users? ii. What page will be fetched next? iii. What are paths frequently accessed by web users. 3) Associations and correlations: i. Page association from usage data – user sessions, user transactions ii. Page associations from content data – similarity based on content analysis iii. Page associations based on structure – link connectivity between pages. Advantages: a) Guide for web site restructuring – by adding links that interconnect pages often viewed together. B) Improve the system performance by prefetching web data. Sequential pattern discovery is applied to web access server transaction logs to discover sequential patterns that indicate user visit patterns over a certain period. That is, the order in which URLs tend to be accessed benefits are a) useful user trends can be discovered b) predictions concerning visit pattern can be made c) to improve website navigation d) personalize advertisements e) dynamically reorganize link structure and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles.

Clustering is a group together items that have similar characteristics. a) Page clusters groups of pages that seem to be conceptually related according to users' perception. b) User Cluster groups or users that seem to be behave similarly when navigating through a web site. Classification maps a data item into one of several predetermined classes for example describing each users category using profiles. Classification algorithms are decision tree, naïve Bayesian classifier, neural networks. Path Analysis is a technique that involves the generation of some form of graph that "represents relations defined on web pages. This can be the physical layout of a web site in which the web pages are nodes and links between these pages are directed edges. Most graphs are involved in determining frequent traversal patterns/ more frequently visited paths in a web site for example what paths do users traversal before they go to a particular URL.

3. Web Mining Techniques

Web mining can be categorized into three areas of interest based on which part of the web to mine

3.1. Web Content Mining

Useful information from the web contents/documents is the application of data mining techniques to content published on the Internet. The web contains types of data. Basically, the web content consists of several types of data such as plain text (unstructured), image, audio, video, meta data as well as HTML (semi Structured), or XML (structured documents), dynamic documents, multimedia documents. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts/ text data mining/ text mining. Hence consider text mining as an instance as an instance of web content mining.

Issues in Web content Mining:
Developing intelligent tools for information retrieval
Finding keywords and key phases

- Discovering grammatical rules collections
- Hypertext classification/categorization
- Extracting key phrases from text documents
- Learning extraction rules
- Hierarchical clustering
- Predicting relationships

Web content mining involves in artificial intelligence systems that can "act autonomously or semi autonomously on behalf of a particular user, to discover and organize web based information". Agent Based approaches focus on intelligent and autonomous web mining tools based on agent technology. i. Some intelligent web agents can use a user profile to search for relevant information, then organize and interpret the discovered information. Example Harvest. ii) Some use various information retrieval techniques and the characteristics of open hypertext documents to organize and filter retrieved information. iii) Learn user preferences and use those preferences to discover information sources for those particular users. Example: Xpert Rule Rminer.

Data base approach: focuses on "integrating and organizing the heterogeneous and semi-structured data on the web into more structured and high level collections of resources". These organized resources can then be accessed and analyzed. These "metadata or generalization are then organized into structured collections and can be analyzed.

3.2. Web Structure Mining

Operates on the web hyperlink structure can provide information about page ranking or authoritativeness and enhance search results through filtering i.e., tries to discover the model underlying the link structures of the web. This model is used to analyze the similarity and relationship between different web sites. This type of mining can be based on the kind of structural data used. a) A hyperlink is a structural unit that connects a web page to different location, either within the same web page (intra_document hyperlink) or to a different web page (inter_document) hyperlink. b) Document structure, the content within a web page can also be organized in a tree structured format, based on various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

Web structure analysis used for:

- Ordering documents matching a user query
- Deciding what pages to add to a collection
- Page categorization
- Finding related pages
- Finding duplicated web sites
- Find out similarity between them.

3.3. Web Usage Mining

To discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. It tries to make sense of the data generated by the web surfer's sessions/behaviors. While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The web usage data includes the data from web server logs, proxy server logs, browser logs, and user profiles. (The usage data can also be classified into three different kinds on the basis of the source of its collection, on the server side (there is an aggregate picture of the usage of a service by all users), the client side (while on the client side there is complete picture of usage of all services by a particular client), and the proxy side (with the proxy side being somewhere in the middle). Web usage mining analyzes results of user interactions with a web server, including web logs, click streams, and database transactions at a web site of a group of related sites.

Web usage mining process can be regarded as a three-phase process consisting:

a) Preprocessing/ data preparation - web log data are preprocessed in order to clean the data – removes log entries that are not needed for the mining process, data integration, identify users, sessions, and so on.

b) Pattern discovery - statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns.

c) Pattern analysis phase - discovered patterns are analyzed here using OLAP tools, knowledge query management mechanism and Intelligent agent to filter out the uninteresting rules/patterns.

After discovering patterns from usage data, analysis has to be conducted. The most common ways of analyzing such patterns are either by using query or by loading the results into a data cube and then performing OLAP operations. Then, visualization techniques are used for a results interpretation. The discovered rules and patterns can then be used for improving the system performance / for making modifications to the web site. The purpose of web usage mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns. Usage mining tools discover and predict user behavior in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

4. Business Intelligence

Business intelligence is software represents the tools and systems that play a role in the strategic planning processes of the corporation. System allows an organization to gather store access and analyze corporate data in decision making. Generally systems illustrate business intelligence in the areas of customer profiling, support, market research, segmentation, product profitability, statistical analysis inventory. Most of the companies collect large data from their business operations. A business would need to use a wide range of software programs, such as Excel, Access, databases applications. Business intelligence software for web mining help to gather information in a timelier, this will search for the trade magazines and newspapers relevant to business to provide the growth information. Entering a new revenue market is always frightening but diversification is key factor to surviving difficult timelines. Business intelligence software for mining provides predictive analysis of various growth potentials according to the search criteria determine. With assistance of a business intelligence service can we face the most difficult of financial times with more confidence and need to diversify because will have intelligence smart choice.

Web mining can help in improving the business decision is challenging task to engineer, implement the search engine. This specifies that indexing of web pages involves a huge task as per tens of millions of queries are given to search engine. The problems of scaling traditional search techniques to magnitude new technical challenges are involved in using the additional information present in hypertext to produce better search results. Hypertext information can be answered by using web mining techniques and improving the capabilities of the search engines by giving better results to clients. Web mining applications have been used by these web sites such as web search for example Google Yahoo web vertical search Amazon, applications of web mining such as ERP CRM , E-Business.

Business Intelligence Architecture: web content is responsible for fetching content from diverse sources into the web business intelligence system.

As a component of the SAP Netweaver platform business intelligence brings together a powerful business intelligence infrastructure and numerous tools and functions for planning for data warehousing. We can integrate internal and external data and convert it into valuable information.

Data warehousing extraction transformation and importing of data will allow setting up data warehouses to model the information architecture on structure and to manage data from various sources.

Online analytical processing data mining alerts data to be accessed and represented to be searched for patterns. Queries reports and analysis as well as the development of web applications will allow creating analysis reports to support decision making all levels of business solutions available on the internet. Business content and metadata as well as collaborative business intelligence can track progress report templates ensure data consistency and promote cooperation between decision make. Functionally business intelligence improves the efficiency of queries simplify administrative tasks and speed up background processes.

5. Problem Definition

In previous research we found some text mining, sequence mining. Research in web mining is ultimate goal of developing computational approaches for monitoring public opinion in regions of conflict, indicators, and social media correlating these risk signals with commonly accepted quantitative assessments. Serious concern to the international community country risks has traditionally been assessed by monitoring economic indicators, to view the conversation of terrorist to attack our country or in the world we can identify appropriate data sources in our analysis. Our aim to track all the messages and call conversations of mobile networks and conduct the analysis using mining algorithms, this system consists of tracking of message and also phone calls, analytic approach, visualization technique. Geo-tracking web system includes blogs, social media forum data from several countries. Our system supports automated social media collection updates from multiple search, visualization and machine translation through web interface application only.

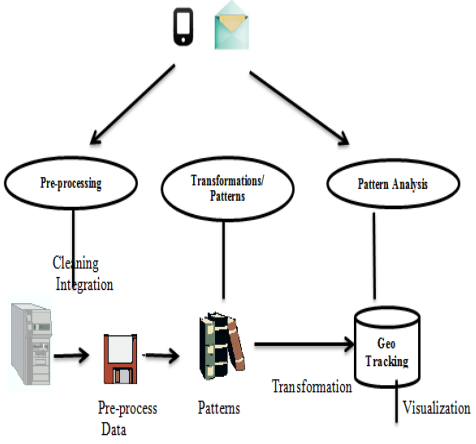

Figure 2: Geo Tracking system Mining

6. Early Research on Web Mining

Web mining research on Google apps, text mining, semantic webs, information retrieval, graph mining, and museum mining, we briefly describe on web mining applications.

6.1. Text Mining

Text mining is to process of extracting relevant patterns (text, numeric indices) from unstructured information. It filter out automatically most undesirable junk emails based on certain terms or words that are not likely to appear in legitimate message but instead identify undesirable electronic mail. Automatically process the contents of web pages in a particular domain for example open a webpage and begin crawling the links to find there process all web pages that are referenced. Text mining can be summarized as a process of numericizing text at the simple level all words found in the input documents will be indexed and counted in order to compute a table of documents words that a matrix frequencies that enumerate that number of times that each word occurs in each document. The automatic search of large numbers of documents based on key words is the domain for example the popular internet search engines that have been developed over the last decade to provide efficient access to web pages with certain content.

6.2. Sequence Analysis

A set of sequence is to find the complete set of frequent subsequences. An element may contain a set of items within an element are unordered and we list them alphabetically. Huge number of possible sequential patterns are hidden in databases mining algorithm should find the complete set of patterns when possible satisfying the minimum support frequency threshold. Highly efficient scalable involving only a small number of databases scans and incorporates various kinds of user-specific constraints. A sequence of two events is generated from frequent sequences consisting of one event and so on after generating a new sequence is checked in a database of customer histories.

6.3. Graph Mining

Graph mining has become an important research because of numerous applications to a wide variety of data mining problems in computational biology chemical data analysis, drug discovery and communication networking. Traditional data mining algorithms such as clustering classification frequent pattern and indexing have now been extended to the graph scenario. A graph is a set of nodes pairs of which might be connected by edges in a wide array of disciplines data can be intuitively cast into this format. Computer networks consists of routers computers and links between them. Social networks consists of individuals and interconnections, aims to design a graph mining tool that provides facilities for input data preprocessing for upload of source data into graph representations, frequent substructure discovery dense substructures especially the interactions within themselves. Mining the knowledge from graph data has become a major research topic in recent data mining, graph matching is that of finding either an approximate or a one-to-one correspondence among the nodes of the two graphs.

6.4. Museums Mining

Museums development is a text mining to improve image access the goal is to collect digital image art historians and personalize retrieval. Text combines social tagging and trust inferencing to enrich metadata retrieval. By processing related text through the CLiMB toolkit evidence for evaluating the role of trust and for assessing the relationship between tags and text terms.

The fundamental and driving research issue in this project concerns the relationship between the language of image description and an image itself. The University of Maryland's Institute for Advanced Computer Studies and College of Information Studies, the Indianapolis Museum of Art, and fourteen other museums have joined to conduct research on new methods to improve user access to digital image collections in museums and libraries. Studies on image searching indicate that current subject description and cataloging practices in museums, libraries and other art historical collections are inadequate for many end user needs. Trant, et al. 2007, as part of the steve.museum project, report that search behaviors for users of the Guggenheim collection do not match the descriptive practices of museum personnel. This disconnect results in unsatisfactory and unsuccessful image access for users.

7. Evaluation

In our work web mining is a most innovative role to identify messages from hackers and terrorist which focus to capture mobile information data. To report all the information it uses business intelligence technique at the same time it also identify the hackers, anyone going to attack our Geo Tracking.

Step 1: Capture mobile information from all the networks
Step 2: Preprocessing the data (relevant information to be mined from large database)
Step 3: Transforms in Pattern analysis.
Step 4: Apply the Mining Technique.
Step 5: Track the message into Geo Tracking System.

Evaluates the Geo Tracking system using information object that are available in particular dataset results will be shown either in percentage or notification alerts.

8. Conclusion

Aassociation and machine learning applied to web architecture model pertaining to source code mining implementation tools improves software debugging business rules for novel projects and also presents

strategies for efficient study text, graph mining. Capture the information available in mobile calls internet and call conversations from all the networks availability apply the data mining technique to track the alerts or attacks. We implements the system ”Geo Tracking” to identify messages from terrorist or threat persons which only finds the alert messages or misuse conversations at time Geo Tracking not expose to capture the confidential information is advantage and also from hackers detects the negative rates and improves the high positive which increases the quality of Government Private and Public sectors.

References

[1] K. Bollacker, S. Lawrence, and C.L. Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Katia P. Sycara and Michael Wooldridge, editors, Proceedings of the Second International Conference on Autonomous Agents, pages 116–123, New York, 1998. ACM Press.

[2] J. Borges and M. Levene. Mining Association Rules in Hypertext Databases. In Knowledge Discovery and Data Mining, pages 149–153, 1998.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107–117, 1998.

[4] Ananiadou, S. and McNaught, J. (Editors) (2006). *Text Mining for Biology and Biomedicine*. Artech House Books. ISBN 978-1-58053-984-5

[5] Bilisoly, R. (2008). *Practical Text Mining with Perl*. New York: John Wiley & Sons.

[6] Feldman, R., and Sanger, J. (2006). *The Text Mining Handbook*. New York: Cambridge University Press. ISBN 9780521836579