ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Web Warehouse: Issues and Challenges for Web Data Mining

Jai Prakash Verma
Assistant Professor
Institute of Technology, Nirma University
Ahmedabad, India

Dr. Atul Patel
Professor & dean,
CMPICA, CHARUSAT University
Changa, Nadiad, Gujarat, India

Abstract: Web Mining defines as extracting knowledge from web. Its types are web uses mining, web content mining, and web structured mining. Each category requires handling the issues of heterogeneous behaviour of web data. This paper will focus on the issues of web data, web mining and its applications. The detailed architecture of web warehouse is also proposed. Each component of architecture requires more exploration and research. As well as paper describe Knowledge as a Service (KAAS) and Web Warehousing as a Service (WWAAS) as web mining applications with cloud computing. The implementation of web warehouse as KAAS and WWAAS with cloud computing is proposed as future extension of this work.

Keywords: Web Data, Web Data Mining, Web Warehouse, Web Uses Mining, Web Content Mining, Web Structure Mining

I. INTRODUCTION

Our day to day life is impacted by World Wide Web (in short WWW) in the form of information seeking, purchasing, entertainment, communication, businesses, decision making, marketing, and prediction etc. All web based applications require mine web pages to extracting knowledge and information. Web Mining uses data mining techniques to extract patterns and information from web data and web pages. There are basically three classes of knowledge that can be discovered from web are web uses mining, web content mining, and web structure mining. Web uses mining applications includes server logs, web browser activity tracking, and user behavior on web as data resources. Web structure mining applications includes links and hyperlinks between web pages, people as data resources. Web content mining applications includes data found on web pages and inside documents as data resources[1, 2].

When comparing web mining with traditional data mining, there are three main differences to consider. Scale- In traditional data mining, processing 1 million records from a database but in web mining data records are large even 10 million pages. Access- When doing data mining information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. Structure – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup [11, 12].

The main objective of this paper is to discuss the major issues in web data mining applications. First is heterogeneous behavior of web data. Drastically increasing the uses of web and mobile applications generates huge amount of data and information. These are available in the form of interconnected web pages. Web pages provide data in three form structured, semi-structured and unstructured. Second is dynamic nature of web data and web applications. Third is searching and finding patterns form web in the form of web uses mining, web content mining, and web structure mining [4]. This paper also proposed a detailed architecture of web warehouse. At the end the research scope of web mining in the area of cloud computing is also discussed. Paper proposed the two concepts, Knowledge as a Service (KAAS) and Web Warehousing as a Service (WWAAS) as an extension of cloud computing.

II. WEB MINING AS EXTENDED VERSION OF DATA MINING

Web Mining: extracting knowledge from web. This definition generates many questions in our mind. Why web mining? Which types of information or knowledge are mined? How does it perform? How are heterogeneous web data transformed in furnished data? All these questions generate a lot of issues in our mind that justify the requirement of web mining. Data mining commonly defined the processes to find pattern and knowledge form different data sources like databases, text files, images, videos, Web etc [17, 18]. But in all cases data are predefined and static and in case of web data mining, data are dynamic in nature. Due to dynamic nature of web data, extracting knowledge from web data is different from traditional data mining. That why we can say that web mining is an extended version of data mining. Figure -1 shows steps for Web Mining: that shows the issues to be handle at each level.

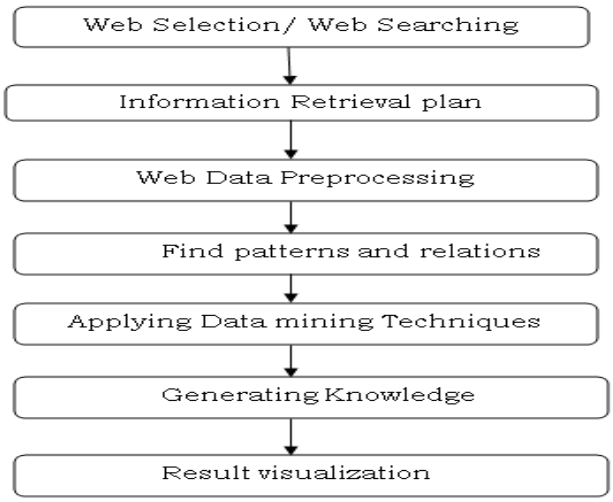

Figure 1: Steps of Web Mining

Web mining task can be categories into three types: web content mining, web uses mining, and web structure mining. Web Content Mining: extract useful information and knowledge form web page contents. Application areas for web content mining are Web search engine, Web personalization, Recommendation system, and Web site adaptive interfaces etc. [1]. In it we can automatically cluster and classify web pages content wise as per user requirements. We can also mine customer reviews and remarks on different forum to discover trends and patterns. These are not traditional data mining tasks [11, 18].

Web Uses mining: covers extracting useful information and knowledge form user logs and clicks. In this types of mining technique system require to maintain all information of user traveling on web through different web pages. The key issues for web uses mining are Data preprocessing, Data Integration, Pattern Discovery, and Pattern Analysis etc. of clicking stream data in usage logs in order to produce the right data for mining [11, 18].

Web Structure Mining: extracts knowledge and information regarding structure of web-sites using different hyperlinks or relations between different web pages. These types of information and knowledge can help to maintain and modify web site navigation for making it more interactive and interesting. Traditional data mining does not perform such task because there is usually no link structure in a relational table [11, 18].

III. HETEROGENEOUS BEHAVIOR OF WEB DATA

Uses of internet are increasing drastically. The way in which we are communicating, gathering information, conducting businesses and making purchases with the help of internet are generating huge amount of information and web data which are heterogeneous in nature. These heterogeneous web data [4] can be categories in three classes: structured, semi-structure, and un-structure. Databases and knowledge bases are typical examples of structured data and information. The sources of semi-structured data are HTML pages, Web forms and languages of text description. And the source of unstructured does not have any form of standard organization and there are no precise relations among the data. It is very challenging task to develop a system to manage and maintain these heterogeneous web data for retrieval. Following are relevant work done by different researchers for handling the issues of heterogeneous behavior of web data.

Daniel M. Herzig and Thanh Tran [5] construct an Entity Relevance Model (ERM) that can be seen as a compact representation of relevant results mirroring the underlying information need for structured data. They assume that today rapid amount of structured data are available on the web. Based on large-scale experiments using real-world datasets, they observe that the data integration approach consistently provides better results than keyword search.

Abdolreza Hajmoosaei, Sameem Abdul Kareem [6] proposed an approach to resolves semantic heterogeneities between web data source and user query through semantic mapping between the domain ontology and related local ontology. Because of syntax and semantic heterogeneity of web data sources fails to answer user queries.

Alexandra Cernian [7] performed clustering on unstructured data using on the concept of clustering by compression that was proposed by Rudi Cilibrasi and Paul Vitanyi [8]. The results of these tests encourage to including the clustering by compression procedure into a framework for clustering heterogeneous web data.

Abdolreza Hajmoosaei [9] recommends system architecture for web data integration focusing on resolving the problems of semantic schema heterogeneity between web data sources. Paper proposed an ontology-based approach as a solution for the reconciliation of semantic conflicts between web data at the schema level.

Yongtao Ma [10] proposed an unsupervised approach for learning subtypes and the subtype-specific blocking keys and key values for heterogeneous web data. Paper provides a solution to solve the subtasks of selecting discriminative attributes and representing their values. It shows how this approach can be used for blocking and instance matching. Compared to state-of-the-art instance matching approaches, blocking and instance matching approach greatly improves result quality.

IV. RELATED WORK AND RESEARCH ISSUES FOR WEB DATA MINING

Increasing the uses of internet in the area of entertainment, business, research, and security generates huge amount of web data and information. This web data and information can be used for finding knowledge and pattern for decision making and prediction in several application areas. Following are some related work in the area web data mining.

Georges Dupret [11] proposed user visit and absence time for a website to analyzing user engagement with the website. Also proposed various engagement matrices include click-through rates, page views, time spent on a site ("dwell time"), loyalty metrics, or more generally "activity" metrics which relate to the user behavior during an online

session. Based on it, paper proposes if users find the website is interesting then they will return sooner. Author also discuss the issues of ranking function, effects of visitors sessions, click-through rates, query reformulation, abandonment rates, survival analysis, a case study. This requires future research for analyzing relationship between the user behavior during a session and user decision to return to the site and their long term engagement.

Jai Prakash Verma [3], proposed Smart Inbox, to categorize the incoming mails to a specific mail recipient using one of the classification methods, Bayesian Learning. It uses the available knowledge to classify the training data and then predicts the classes of the new incoming mail. This is my first work for web mining that motivates me to work in this direction. We have implemented this concept at a primitive level using mathematical approach. It can be extended to achieve better efficiency using more efficient classifiers. Further, different attributes may be assigned different weights to determine the priorities dynamically. A comparison based analysis of several classifiers on a relatively large dataset can give significantly reliable output.

Milad Eftekhar [13], proposed two models (intrinsic burst model social burst model) for finding knowledge and information from different user behavior on social network sites. They are using two graphs: action graph and holistic graph. Action graph emphasis on user actions, timings, and links of actions triggers other actions; holistic graph emphasizes on contents generated or shred by individual users. These models characterize and identify information bursts in social networks and presented algorithms to identify bursty subgraphs. Paper proposed several research issues for future work like dynamically generation of subgraph, time based evaluation, assigning reputation value or weightage to different users etc.

Nicola Barbieri [14] proposed the Community – Cascade (CCN) Model that uses to guide user behavior within the network. They also apply CCN model to real-world social networks and information cascades: the results witness the validity of the proposed CCN model, providing useful insights on its significance for analyzing social behavior. This paper proposes a graph and a set of cascades, that tackle the community detection task by fitting a unique stochastic generative model to the observed social graph and cascades.

Takeshi Kurashima [15] proposes a model called Geo Topic Model that analysis the user's interest and location log data to recommend locations to be visited. This model is experimented on real location logs of landmarks and places visited to evaluate the recommendation accuracy and performance. This paper also shows that the model can estimate latent features of locations such as art, nature and atmosphere as latent topics, and describe each user's preference based on them.

Jing Liu [16] considers the problems of linking users of across multiple online communities. This user linking tasks uses the concept of alias-disambiguation, which is meant to differentiate users with the same username. The alias-disambiguation step is casted as a pair-wise classification problem and proposes a novel unsupervised approach. That classifies usernames in the classes of rareness and commonness using n-gram probabilities. Linking users across community websites can be used for user bonding with websites, aggregating public profile user data, protecting users from privacy risk, enables users to keep updated with their online friends and in many more applications.

V. PROPOSED WEB WAREHOUSE ARCHITECTURE

Based on the above survey and study the architecture of web warehousing is proposed in figure -2. It shows different components of web warehousing. Basically there are three main components one is extracting, transforming, and loading information from World Wide Web (WWW) to web warehouse and handling the issues of heterogeneous behavior of Web Data. Second component includes integration and aggregation of web data into web warehouse implementing different data mining techniques. Third component encapsulate and handling the issues of query generation, query rewriting, query mapping and query redirection. All three components raises research issues in the area of web data mining.

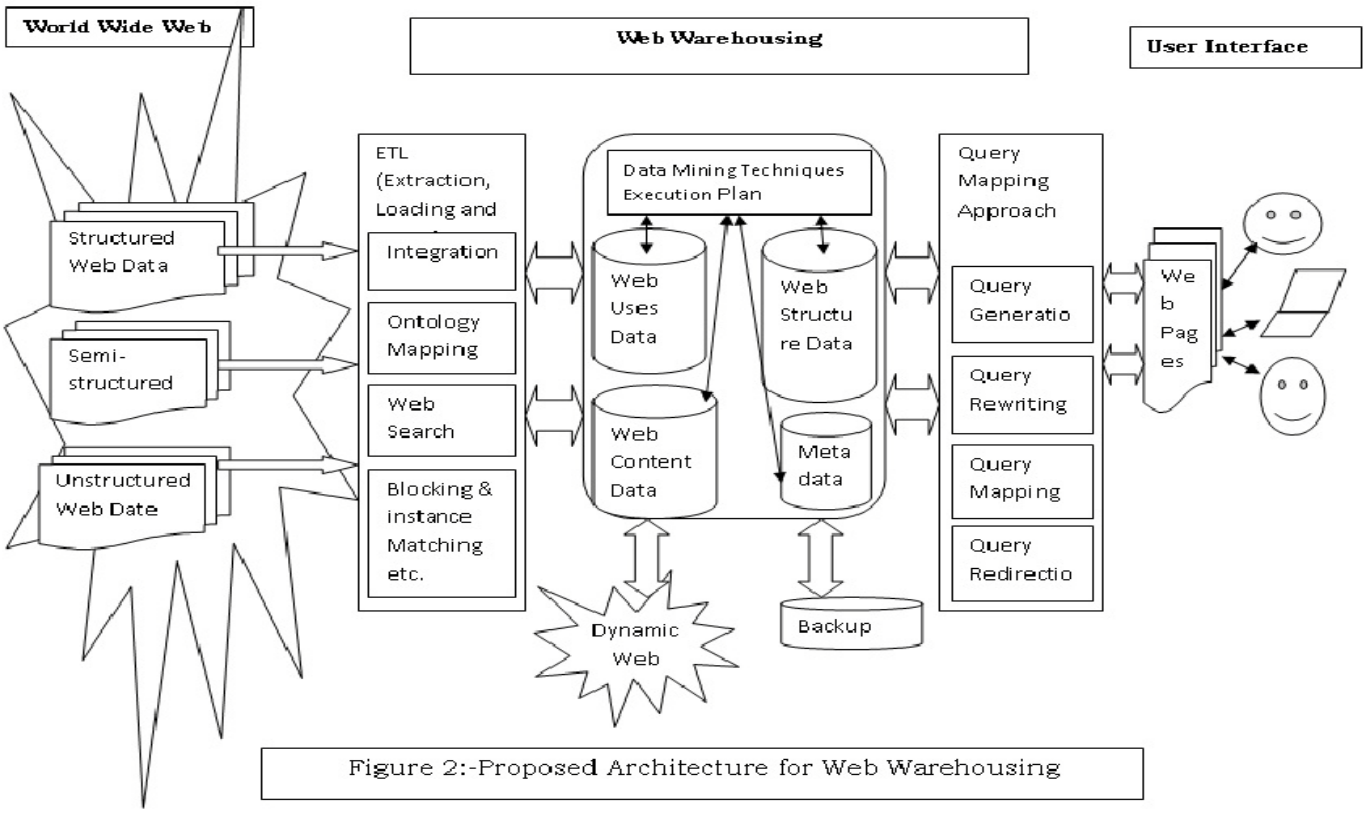

Figure 2:-Proposed Architecture for Web Warehousing

VI. CLOUD MINING

We also propose two concepts Knowledge as a Service (KAAS) and Web Warehousing as a Service (WWAAS) as extension of cloud computing. Cloud computing refer as delivering computing resources like hardware, software etc. through internet. Cloud computing provide three services Software as a Service (SAAS), Infrastructure as a Service (IAAS), and Platform as a Service (PAAS). Today uses of internet are increasing drastically in the area of entertainment, business, research, and security that generating huge amount of web data and information. This web data and information can be used for knowledge and pattern discovery for decision making and prediction. In extension of cloud computing web mining emerges the concept of KAAS and WWAAS. Knowledge as a Service (KAAS) is the future of web mining. In this service provider establish a knowledge cloud and knowledge seeker consume their services as per requirement and need. Web Warehouse as a Service (WWAAS) provides a complete infrastructure of web warehouse as a service to client.

VII. CONCLUSION AND FUTURE WORK

This paper suggests complete design architecture for an expert system that uses to solve all types web mining queries based on different business and scientific applications. Design an intelligent and predictive model for huge amount of web data. This paper suggest architecture of an intelligent warehouse for structured, unstructured and semi-structured web data and an expert system that use to solve all web mining queries based on different business and scientific applications using different types of data mining techniques. We also propose two concepts Knowledge as a Service (KAAS) and Web Warehousing as a Service (WWAAS) as extension of cloud computing. These require more work and research in the direction of cloud mining.

VIII. REFERENCES

[1] Ujwala Manoj Patil, J.B. Patil, 2012 Web Data Mining Trends and Techniques,International Conference on Advances in Computing, Communications and Informatics (ICACCI-2012) 962-965

[2] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, 2011, Web Mining: Today and Tomorrow, 978-1-4244-8679-3/11 ©2011 IEEE

[3] Jai Prakash Verma and Sapan Mankad, 2011, Smart Inbox:A comparison based approach to classify the incoming mails International Journal of Artificial Intelligence and Knowledge Discovery Vol.1, Issue 1, Jan, 2011

[4] Hicham Snoussi, Laurent Magnin, and Jian-Yun Nie, Heterogeneous Web Data Extraction using Ontology

[5] Daniel M. Herzig, Thanh Tran, 2012, Heterogeneous Web Data Search Using Relevance-based On The Fly Data Integration, WWW 2012 – Session: Semantic Web Approaches in Search April 16–20, 2012, Lyon, France

[6] Abdolreza Hajmoosaei, Sameem Abdul Kareem, 2008, An Approach for Semantic Query Mapping on the Heterogeneous Web Data Sources, 978-1-4244-2624-9/08 ©2008 IEEE

[7] Alexandra Cernian, Dorin Carstoiu and Adriana Olteanu, 2008, Clustering Heterogeneous Web Data Using Clustering by Compression, Cluster Validity, 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing

[8] Rudi Cilibrasi, Paul M.B. Vitanyi, Clustering by compression. IEEE Transactions on Information Theory, Vol. 51, No. 4, pp 1523–1545, 2005

[9] Abdolreza Hajmoosaei, Sameem Abdul-Kareem, 2007, An ontology-based approach for resolving semantic schema conflicts in the extraction and integration of

query-based information from heterogeneous web data sources, Proc. 3rd Australian Ontologies Workshop (AOW 2007), Gold Coast, Australia

[10] Yongtao Ma, Thanh Tran, 2013, TYPiMatch: Type-specific Unsupervised Learning of Keys and Key Values for Heterogeneous Web Data Integration, WSDM'13, February 4–8, 2013, Rome, Italy. Copyright 2013 ACM 978-1-4503-1869-3/13/02.

[11] Web Data Mining Exploring Hyperlinks, Contents, and Usages Data By Bing Liu Published by Springer

[12] Georges Dupret, Mounia Lalmas, 2013, Absence Time and User Engagement: Evaluating Ranking Functions, WSDM'13, February 4–8, 2013, Rome, Italy

[13] Milad Eftekhar, Nick Koudas, and Yashar Ganjali, 2013, Bursty Subgraphs in Social Networks, WSDM'13, February 4–8, 2013, Rome, Italy.

[14] Nicola Barbieri, Francesco Bonchi, Giuseppe Manco , 2013 , Cascade-based Community Detection, WSDM'13, February 4–8, 2012, Rome, Italy.

[15] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya and Ko Fujimura, 2013, Geo Topic Model: Joint Modeling of User's Activity Area and Interests for Location Recommendation, *WSDM'13,* February 4–8, 2013, Rome,

[16] Italy. Jing Liu, Fan Zhang, and Xinying Song, 2013, What's in a Name? An Unsupervised Approach to Link Users across Communities, *WSDM'13,* February 4–8, 2013, Rome, Italy.

[17] Jiawei Han and Micheline Kamber. Data Mining : concept and Techniques.Elseveir Pubilcation.

[18] Jai Prakash, Bankim Patel, Atul Patel,"Web Mining: Opinion and Feedback Analysis for Educational Institutions", 2013, IJCA, Volume 84 – No 6, December, 2013.