ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

Web Mining Over Cloud – A Prerequisite for Today's Enterprises

Jitendra Singh Tomar

Assistant Professor, Amity University, Gautam Buddha Nagar (UP), India

Abstract: This paper surveys web mining techniques to mine information on cloud computing platform. WM is extraction of relevant information from the relics and mannerism recorded by the stakeholders on WWW for knowledge management purpose. The online information resource is very large and organizations are utilizing web content mining and access through the web servers. Web mining techniques and applications are much needed in cloud computing. The effective management of information kept over cloud for knowledge and business management could be achieved if the web mining techniques over virtually integrated warehouse are efficiently practiced by an end users.

Keywords: Web Mining, Data Mining, Knowledge Management, Cloud Mining, Web Usage Mining, Knowledge Discovery.

I. INTRODUCTION

With growth and affordability of technology in recent years has lead to enormous increase in its usage. The numbers of users are exponentially increasing with the convenience of Internet too which is affecting day to day life of the people, both personally and professionally. The modern businesses are certainly supported by the internet architecture and of late business are moving to cloud computing because of enormous advantages which it has in offering.

Since most people are connected to the WWW, a sea of information is floating on the medium. This electronic information available on the internet could be accessed and managed for new strategy formulation by the organizations as people on the internet are also the stakeholders of one or the other enterprise. The prominent observation is that the information on the internet is majorly readily and freely available, and could be mined for developing business strategies.

The organizations are focussing on efficient usage of mining techniques to extract information from the web and ascertain hidden and unknown patterns in the information. There are various mining techniques that the organizations could use but Web Mining not only is limited to applying mining techniques to the data available on the Web, but technically it makes sure that the algorithms are modified to suit better demands of the Web[1] to meet business requirement of the current times. In recent years, data mining techniques have progress to great extent since discovering new patterns in information, and ascertaining knowledge in databases is proving imperative in various business domains and functional areas be it medicine, business, science and technology, spatial data [2] [3].

The Cloud is an IT / IS service over the public network of Internet which uses IT resources like hardware, software, and data repositories. Many companies, especially the start-ups, are plumping for this service provided by various service providers as an alternative to have their own IT infrastructure for which high investment is essential. The organization opting for the cloud service would be using the servers, software, and databases over the internet, assigned to them by the third party cloud service providers. In addition to the low cost, the facility of mobility and availability of the cloud helps it making it preferred by the organizations. But the pros of the service brings in cons too, the biggest of which is the threat to the organization's information which is the most important resource.

A. Web Mining in Yester times.

It had been a great time for development of cloud computing and tremendous R&D backed its potential along with the ease and scalability it has. The culmination started when the internet based data repositories were used to store business data of the organizations and gradually it appealed with all the advantages leading to improvements in data access and the navigation through the data which it offers. With the improvement ranging across the activities of storage to retrieval of business information, the medium saw tremendous growth. The ability to respond with key business analysis promptly and accurately out of the large databases is the unique feature of web mining.

Web mining could be seen as an update over simple data mining with its versatility on online medium where information is stored on the web servers and web log, compared to data mining which is associated with offline medium and data been accessed from databases and data warehouse. The web mining involves newer trends in technology such as artificial intelligence, NLP, and neural network which are incorporated with techniques amalgamated with high performance DB engines and data integration make these technologies practical of current data warehousing environment [4].

II. CHARACTERISTICS OF CLOUD COMPUTING

Cloud computing could be considered as a facility of hardware, software, and data repository offered by third

DOI 10.17148/IJARCCE.2015.4969

party purveyor over the internet. Last decades have seen a paradigm shift in computing technology to lately see the cloud computing concept grow. The gradual growth has seen various phases as given:

- Phase 1: Resourced on powerful mainframe computers were shared by multiple end users on their terminals.
- Phase 2: Advancement in PC computing served the need of end users for day to day working.
- Phase 3: Growth of network technologies allowed multiple computers to connect to each other for sharing of resources including hardware, software, and information.
- Phase 4: Various independent networks could connect to each other to form a more wide and global network.
- Phase 5: The rise of electronic grid eased up sharing of computing power and storage resources across varied networks and platforms.
- Phase 6: The capability to use all the available resources on the public network as shared resources in a scalable and easy way gave rise to cloud computing.

A computing model that enables omnipresent, convenient, on-demand network access to a shared pool of configurable computing resources, including networks, servers, storage, applications, and services, that can be rapidly provisioned and released with minimal management effort or service provider interaction could be called as cloud computing as characterized by National Institute of Standards and Technology. The cloud computing model is an amalgamation of three service models and four deployment models [5].

A. Essential features of Cloud Computing [6]

- On demand self services: Computer services such as email, applications, network or server service can be provided without requiring human interaction with each service provider. Cloud service providers providing on demand self services include Amazon Web Services (AWS), Microsoft, Google, IBM and Salesforce.com. New York Times and NASDAQ are examples of companies using AWS (NIST). Gartner describes this characteristic as service based.
- Broad network access: Cloud Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms such as mobile phones, laptops and PDAs.
- Resource pooling: The provider's computing resources are pooled together to serve multiple consumers using multiple-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The resources include among others storage, processing, memory, network bandwidth, virtual machines and email services. The pooling together of the resource builds economies of scale (Gartner).
- Rapid elasticity: Cloud services can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
- Measured service: Cloud computing resource usage can be measured, controlled, and reported providing transparency for both the provider and consumer of the utilised service. Cloud computing services use a metering capability which enables to control and optimise resource use. Just as utility companies sell power to subscribers, and telephone companies sell voice and data services, IT services such as network security management, data centre hosting or even departmental billing can now be easily delivered as a contractual service.
- Multi Tenacity: It is the 6th characteristics of cloud computing advocated by the Cloud Security Alliance. It refers to the need for policy-driven enforcement, segmentation, isolation, governance, service levels, and chargeback/billing models for different consumer constituencies. Consumers might utilize a public cloud provider's service offerings or actually be from the same organization, such as different business units rather than distinct organizational entities, but would still share infrastructure.

III. CHARACTERISTIC OF DATA MINING

The technique of viewing information from data repository with a new perspective to find hidden patterns in the information is termed as Data Mining. It could be defined as "Type of database analysis that attempts to discover useful patterns or relationships in a group of data using advance statistical tools through artificial intelligence, NLP, and neutral network techniques to discover previously unknown relationships among the data, especially when the data come from different databases."[7]

In current business structure, the data mining techniques could be used in various business domains and functional areas to offer advantage to the business, few are given below:

- Clustering – Clustering is the process of making a group of abstract objects into classes of similar objects. Cluster analysis is used for number of purposes such as market research, pattern recognition, data analysis, and image processing. Clustering could help in recognition of distinct group in the customer base. It helps in classifying the documents on the web for information discovery, and could serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster [8].
- Classification – It is a data mining function that assigns items in a collection to target categories or classes. It is used to predict the aimed class for each case in the data. A classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification model could be used by the financial institution to predict credit risk based on the observed data for many loan applications over a period of time. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer

DOI 10.17148/IJARCCE.2015.4969

would constitute a case. The technique could also be used for tracking employment history, home ownership or rental, years of residence, number and type of investments, and so on [9].

- Association – Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository [10]. It is used to find rules connected with frequently co-occurring items and is used for market basket analysis, cross-selling, up selling, and to do root cause analysis. This analysis is of great use in product bundling, in-store placements, and defect analysis.
- Regression – It is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. A regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors [11].
- Attribute Importance – Attribute importance is a supervised function that identifies and ranks the attributes that are most important in predicting a target attribute [12]. It ranks the attributes according to strength of relationship with target attribute. This may include finding most important factors associated with customers who respond to an offer.
- Fault & Anomaly Detection – The goal of anomaly detection is to identify cases that are unusual within data that is seemingly homogeneous. Anomaly detection is an important tool for detecting fraud, network intrusion, and other rare events that may have great significance but are hard to find [13].
- Feature Extraction – Feature extraction projects a data set with higher dimensionality onto a smaller number of dimensions. As such it is useful for data visualization, since a complex data set can be effectively visualized when it is reduced to two or three dimensions. Some applications of feature extraction are latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. Feature extraction can also be used to enhance the speed and effectiveness of supervised learning [14].

IV. CHARACTERISTICS OF WEB MINING

Web Mining is extraction of relevant information from the relics and mannerism recorded by the stakeholders on WWW. WM works along with the intelligent agents to certain targets, like competitors sites' [15]. These agents collect information from the host web server and collect as much information from analysing the web page itself by visiting the hyperlinks, cookies, and recording traffic patterns. WM could be used for tracking customers' online behaviour and the organizations could use the collected knowledge to establish better customer relationships, offers, and target potential buyers with exclusive deals. The dynamism of WWW could be dealt with ease through effective web crawling that could yield effective results.

In WM operations, Internet and agent technologies are the basis behind WM that are based on fuzzy logic techniques.

The IR tools that are supported by Semantic Web Technologies provide the course for the intelligent agents to explore and search WWW [16]. For easing out the retrieval process of information from WWW, the software are developed with semantics over web published content on the WWW and are continuously growing to enhance the searching process.

The intelligent agents are programmed to perform mining techniques by analysing the HTML document, parsing, and extracting all information from various hyperlinks, multimedia, and other content on the web. The tracking of online account and user preferences is done to identify the mannerism of the users to provide them with the most relevant business content. The sessions and transactions of the users are logged, data traffic along with various activities is logged and analysed to give the stakeholder best and non-tedious experience to work on web as per their preferences and habits.

WM is divided into three main categories as given [17].

- Web content mining WCM – Locate and retrieve text, multimedia, and hyperlinks in the context of the search.
- Web structure mining WSM – Analyze the traffic flow and site maps of certain sites to explore the movement of the users.
- Web usage mining WUM - collects browser history, bookmarks, site logs, cookies, and metadata of the users. It may also be used to mine social network mannerism of users [18].

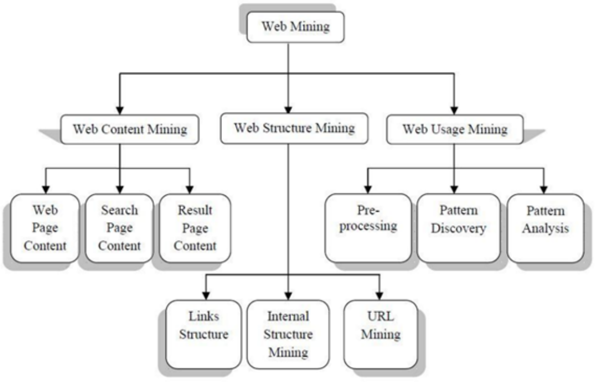

Fig. 1 Classification of Web Mining

All three mentioned tasks are web oriented and play a major role in investigation of network, intranets, and VPNs. It is growingly used in multimedia mining to mine pictures, graphics, movies, audio, and its applications.

Web Mining is foresaid to have four tasks of Information retrieval i.e. resource discovery, Information Extraction i.e. Selection or Pre-processing, Generalization i.e. Pattern recognition, Analysis Validation i.e. Interpretation.

All these tasks help in effective knowledge management and information processing. Also, Web Mining is broadly classified into three phases which are Pre-processing of information, Pattern Discovery in the information, and Pattern Analysis for better decision making.

A. Bottlenecks in efficient Web Mining

Since the usage of technology is growing and web has been a convenient medium to store and access the information, whether personal or professional, the web repositories are over burdened with the information. With increase in size of the medium, the response time to the user may be superficial and it may take too long to avail a service online. The exponential augmentation of Web has imposed a heavy demand on networking resources and Web servers. Hence, an apparent resolution of this workload is to increase the bandwidth which will involve huge investments.

Web caching scheme seems to provide bit of breathing but it too has three significant limitations. If the updation of proxy is not efficiently update, the user may retrieve inappropriate data leading to poor business implementation. Also, the cache servers too have limited system resources like memory space, disk storage, I/O bandwidth, processing power, and networking resources. However, even if the cache space is unlimited, the difficulty of updating such a enormous anthology of Web entities will be unworkable. Also, there may be the web objects that are not requested by the users but due to the pre-fetching policies, the network traffic and the web servers will suffer the overload [19].

V. WEB MINING THROUGH CLOUD COMPUTING

The evolution of cloud computing over the past few years is potentially one of the major advances in the history of computing. It's a new style of computing in which dynamically scalable and virtualized resource are provided as a services over the internet or web. The cloud computing is now amalgamated with Cloud Mining or cloud hashing which allows users to have mining power of the information that is kept of the cloud by sharing resources like hardware, software, and data repositories placed in remote data centres. This could be done without any offline hassle, such as electricity, hosting issues, or installation and upkeep trouble.

The term cloud is a symbol for the Internet, an abstraction of the Internet‛s underlying infrastructure, used to mark the point at which responsibility moves from the user to an external provider. The cloud mining techniques are introduced and invariably used by various organizations and individuals to extract relevant information from the cloud based service on relics and mannerism recorded by the stakeholders on the cloud. It is a new approach to enhance search interface of the data on the cloud. The technology is framed with new dimensions to support cloud computing and modifications are brought in web mining to counter the demands of cloud computing. Utilities like SaaS (Software-as-a-Service) is used for reducing the cost of web mining and try to provide information management with cloud mining and computing technique. The powerful frameworks are developed for doing predictive analytics over the cloud as compared to complex distributed information sources. However, despite increased activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model [20]. But with focus on the cloud computing in recent times, the emergence of the phenomenon of cloud computing represents a fundamental change in the way information technology (IT) services are invented, developed, deployed, scaled, updated, maintained and paid for [21].

VI. CONCLUSION

The future of computing is very strong with cloud computing with economies of scale strongly justifying the concept. The computing resource and applications are generic in nature and if consolidated, could offer tremendous economies of scale and business organizations are certainly taking advantage of cloud computing and its mining features.

Cloud computing is the pulse of today's computing environment and is rich in features providing advantages such as on demand self services, broad network access, resource pooling, rapid elasticity, measure services, and multi tenacity. This adds on the dynamism to computing requirements of today.

Mining too is a great application required in analysis of information and hence knowledge building. The data mining tools are enhance to support cloud mining with the basic operations of supporting information management through techniques like clustering, classification, association, regression, attribute importance, fault & anomaly detection, and feature extraction to help in personal as well in professional arena. Content mining, structure mining, and usage mining are helping the business organization to strategise accordingly. Also, the good of cloud computing and mining is undergoing a tremendous pressure as the upcoming demands in the field are exponentially growing and is adding on to the pressure due to limited resource availability, be it hardware, networking services, processing power, or memory management.

Inspite of the fluid and uncertain environment that may surround cloud computing, it is going to grow as technology and its adoption by the industry is growing. Its evolution over the past few years is potentially one of the major advances in the history of computing and is here to grow. The major reason behind this growth is a new style of computing where the resourced are dynamically scalable and virtualized as a service over the internet or web. With the new applications and tools for mining over the cloud, the cloud computing is along with hashing tools for mining are here to stay and support the business organizations in effective policy making.

REFERENCES

[1] Voas J and Zhang J, ―Cloud Computing: New Wine or Just a New Bottle. IEEE Internet Computing Magazine, 2009. http://www.cmlab.csie.ntu.edu.tw/~jimmychad/CN2011/Readings/ CloudComputing New Wine.pdf.

DOI 10.17148/IJARCCE.2015.4969

[2] Almeida V, Bestavros A, Crovella M, and Oliveira A, ―Characterizing reference locality in the WWW, In IEEE International Conference in Parallel and Distributed Information Systems, Miami Beach, Florida, USA, December 1996. http://www.cs.bu.edu/groups/oceans/papers/ Home.html

[3] Chen, M. S, Han, J. and Yu, P. S. ―Data Mining: An overview from a database perspective, IEEE transaction on knowledge and data engineering, Vol. 08, No. 6, pp: 866-883, 1996.

[4] Kosala R, Blockeel H, Web Mining Research: A Survey, In ACM SIGKDD, July 2000.

[5] Mell P, and Grance T, ―The NIST Definition of Cloud Computing, The National Institute of Standards and Technology, USA, 2011, Link: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[6] http://www.isaca.org/Groups/Professional-English/cloud-computing/GroupDocuments/Essential%20characteristics%20of%20Cloud%20Computing.pdf, accessed on 8th Sep 2015, 15:19 Hrs.

[7] Merriam-Webster Dictionary, ―Definition of data mining, Link: http://www.merriamwebster.com/dictionary/datamining

[8] http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm, accessed on 8th Sep 2015, 14:34 Hrs

[9] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm, accessed on 8th Sep 2015 14:43 Hrs

[10] http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining, accessed on 8th Sep 2015, 14:48 Hrs

[11] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm, accessed on 8th Sep 2015, 14:53 Hrs

[12] http://gerardnico.com/wiki/data_mining/attribute_importance, accessed on 8th Sep 2015, 14:58 Hrs

[13] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/anomalies.htm, accessed on 8th Sep 2015, 15:04 Hrs

[14] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/feature_extr.htm#i100652, accessed on 8th Sep 2015, 15:08 Hrs

[15] Wasilewska A, (2011) "Web Mining Presentation 1" CSE 590 Data Mining, Stony Brook.

[16] Berlanga R, Romero O, Simitsis A, Nebot V, Pedersen T, Abelló A, Aramburu M (2012) "Semantic Web Technologies for Business Intelligence" IGI.

[17] Pal S, Talwar V, Mitra P, (2002) "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions" IEEE Transactions on Neural Networks, Vol. 13, No. 5.

[18] Facca F, Lanzi P, (2005) "Mining interesting knowledge from weblogs: a survey" Data & Knowledge Engineering, 53, Elsevier.

[19] Khalil F, Li J, and Wang H ―A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses, Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184.

[20] Li D, Di K, Li D, and Shi X, ―Mining association rules with linguistic cloud models‖, Research and Development in Knowledge Discovery and Data Mining ,Lecture Notes in Computer Science, 1998.

BIOGRAPHY

Jitendra Singh Tomar is Mathematics Graduate and received his MCA degree in 2001. He is a Microsoft Certified Systems Engineer since 2002 and is working as an IS Consultant and an Academician. He has worked upon various software and network security projects in the industry. As a trainer and academician, he has conducted various MDPs and training programs for the professionals and has been an active associate for various academic activities including curriculum design. He is currently working with Amity University, UP, India, since 2006 and held imperative positions over a period of time.

DOI 10.17148/IJARCCE.2015.4969