ISSN:2320-0790

An International Journal of Advanced Computer Technology

Web Mining: Penning an Era of Information Age

Anshika Goel[1], Dinesh Sahu[2], Manish Kumar Singh[3]

[1]Assistant Professor, Jagannath International Management School, GGS IP University, Delhi,
[2]M. Tech, Ph. D (Resource Allocation and Scheduling Models in Mobile Grid Computing)
[3]M.C.A (Master in Computer Applications)

Abstract: Today's age is rightly pronounced as "Information Age" which stands on the edifice of Information Technology and is operated by the Internet through the concept of web mining and is maintained & evolved through the high-speed technology of cloud computing. In short, if we try to summarize the situation, we would find that web mining concept has fuelled the entire process. This paper is an attempt to put light on the aspect of how web mining has penned the information age by contributing in large to the information technology.

Keywords: Information Age; Evolution; Challenges; Web Spider; Web Bag; Virtual Society; Semantic Web; Cloud Computing

1. INTRODUCTION

We are living in the age that is primarily characterized by the essence of information that guides the lives of billion people belonging to this planet. The world which was once termed as "global village" has now transformed into the "information age" where the wide range of information & communication technologies have bound the whole society in a common thick line of internet & have compelled it's residents to live under the showered information of cloud computing. The entire mankind has been benefited from it and so has been loomed by its negative impacts too. But whatever be the context, we can't escape from the fact that the huge development witnessed by the world in the past decade could primarily be attributed to the boom- boom progress in information technology, which is mainly brought by the development in web mining strategies over internet and its associated field of research, including web personalization, semantic web, artificial intelligence, neural networks, database management and the highly configured cloud computing.

Information has now become the "lifeline" for today's generation which stands on the edifice of information technology. Further, this edifice of information technology relies heavily on the integrated framework of digital networks, comprising data management institutions, advanced computing technologies, computer networks and a regulatory system [1]. The advent of internet and the concept of web mining played the vital role, similar to veins in human body, of controlling the infrastructure and functionality of information technology hugely, thus bringing the revolution to the today's information age. There are two aspects to understand this:

- One aspect is to understand the basic formulation of information technology that is comprised of internet as a mechanism to interconnect computer and users which includes the transmission media, including the telephone lines, satellites and antennas, and so the routers [2].

- Another aspect is to understand that the infrastructure of information technology can be broadly viewed as everything that supports the flow and processing of information.

This survey paper uses the second aspect of understanding the information technology in the face of web mining and how it has revolutionized the information science. The section 2 of the paper discusses how the information technology has evolved so far, section 3 comprises of the brief discussion about how web mining contributes to information technology, and finally we wind up the paper with a conclusion in section 4.

2. EVOLUTION OF INFORMATION TECHNOLOGY

Humans have a long history of storing, retrieving, manipulating and communicating information. It is believed that the dissemination of information is done since the Sumerians (in Mesopotamia) when they had developed writing during 3000 BC [3]. But the term "information technology" in modern generation was first used in 1958 in an article published in Harvard Business Review by authors H.J. Levitt and T.L. Whisler. They commented in the article about “the new technology”

which did not have any name of recognition at that time which led them to propose the name of "Information Technology (IT)" for it [1]. This short but a significant view defined the information technology into three manifold ways:

1. The techniques for processing;
2. The application of statistical and mathematical methods to decision-making; and
3. The simulation of higher-order thinking through computer programs [4].

In short, the term information technology was used as a synonym for computers and computer networks including several other information distribution technologies such as television and telephone. Moreover, Information Technology (IT) has myriad industries associated with it including computer hardware & software, electronics & semi-conductors, internet & e-commerce, telecom equipment & computer services [5][6].

IT can be distinguishably divided into four phases on the basis of it's evolution as per storage and processing technologies employed [1]:

- *Pre-mechanical* (3000 BC to 1450 AD);
- *Mechanical* (1450 to 1840);
- *Electro-mechanical* (1840 to 1940); and
- *Electronic* (1940 to present).

The last phase can be further divided into six phases of development [7][8]:

i. The first phase involved people employing terminals to connect to powerful mainframes that were shared by many users;
ii. The second phase involved stand-alone PCs becoming powerful and reliable enough to satisfy user's daily work;
iii. The third phase involved the computer networks allowing multiple computers to connect to each other;
iv. The fourth phase involved the connection of one local area network to another local area network to build a much fascinated global network;
v. The fifth phase let the world to witness the electronic grid, facilitating the shared computing power and storage resources; and
vi. The sixth phase involves the latest innovative cloud computing allowing the exploration of all available resources on the internet in a scalable and simple way.

The sixth phase incorporates information/data analytics techniques, noticeably working on the principle of web mining (an extension of data mining) to harness web data, plus big data to explore the path of healthy progress of several businesses and industries today.

Data mining is the principle concept of extracting data from large data repositories to represent knowledge/information that indulges the concrete convergence of statistics, machine learning, AI (Artificial Intelligence), computational techniques and pattern recognition techniques. All these methodologies follow the following key principle strategies to mine data to constitute and represent knowledge:

- *Clustering* to explore data and finding natural groupings.
- *Classification* to predict a specific outcome such as buy/not buys, yes/no, etc.
- *Association* to search rules associated with frequently occurring items needed for selling product, its in-store storage and analyzing it for the defects.
- *Regression* to predict an outcome whose numerical value must be continuous which is generally achieved through machine learning technique in order to fit an equation to a dataset [9].

Since the advent of Internet, plus the development seen in web languages, the data mining strategies have evolved into web mining strategies over the decade that has facilitated the significant representation of knowledge and has further nourished the methods to provide services in quick span of time via Internet by extracting web data as well as big data. The importance of web mining has been felt in almost all the umpteen businesses, industries and markets across the globe. Besides the myriad benefits and services exploited through the significant strategies of web mining, various issues and norms have also grown over the years ranging from encountering malicious activities to intrusion & terror practices, security and intrusion detection have become major challenges for the webmasters and web organizations. To counter them and making web surf efficient & safe, several new and effective strategies have been inducted into web mining techniques including [10]:

- *Attribute importance*: method to rank attributes as per the degree of relationship with the searched/target attribute.
- *Anomaly detection*: method to detect and recognize the unusual, many times suspicious activities/pattern based on the deviation from the actual norm. Example: Frauds associated with banking, taxation, online dealing and transaction, healthcare, etc.
- *Feature extraction*: method to generate attributes that are new, using linear combination of existing attributes. This method is applicable for generating text data, latent semantic analysis, data decomposition & projection, data compression and pattern recognition.

3. HOW WEB MINING CONTRIBUTES TO INFORMATION TECHNOLOGY- A DISCUSSION

Revolutions in human life such as agricultural, industrial and now information, they occur so quickly that no one can predict whether the change they will have on our

make-up, or even whether there will be a change. But we can't escape from the fact of swashbuckling development that has impacted the life of billions across the globe that couldn't have been possible in the absence of techniques, technologies and sophisticated infrastructure contributing to the fast access of information in fractions of seconds from any corner of the world. The internet/web has created a kind of "superhuman" intelligence where the information can be exchanged within fractions of seconds by anyone residing at any part of the world, just like the human brain which exchanges information from different sensory organs of the human body [11]. Computer and Internet are playing significant role in every genre of human life. It is very hard to find a person without computer and internet knowledge and usage. Whether at home, school or workplace, computer plus internet is finding place in everyone's life including naive users, students, employee, employer, business men, scientist, medical professions and so on. The mass growth of internet has increased one's knowledge in various fields as one can easily get information about anything with ease and comfort, thus bringing a great revolution in one's daily life [12].

The internet has made the world "smaller" which helps to bring the world's information to one's fingertips and him/her to the world if he/she so chooses [13].

- One can easily communicate through e-mail, streaming video, chat and instant messaging.
- One can easily get information out through Blogs, MySpace, etc.
- One can easily be informed since today almost all newspapers are now available online and numerous reliable sites are there like Wikipedia, Google, buying websites, opinion websites, etc. that facilitate the fast online browsing, access and availability of information.

Besides rapid and steady advancement in semi-conductor technology, information strategy, networking, and applications, the interaction of Information Technology with Internet has improved the Knowledge Acquisition considerably at the exponential rate. The "impact" or "effect" of Internet and Web Mining on the development of Information Technology has proved to be multi-dimensional as well as multi-directional. Over the past few decades, many studies have been done and several research papers published in this regard. Large business houses have invested heavily on the development of Internet, Web Mining and in the expansion of e-commerce through the building of powerful data repositories, databases and informational retrieval support management, customer services, logistics, online project design, marketing and competitive analysis [14] that has made the edifice of Knowledge Acquisition a rock-strong and has provided a wind to the noticeable revolution of development in almost every field across the planet today, ranging from nuclear science to agriculture, nanotechnology to networking, astronomy to meteorology, energy exploitation to archaeological findings, education to healthcare improvements, security and defense research, etc.

The growth in e-commerce over the past decade due to the development in Web Mining strategies and technologies has changed the perception of the economists and strategists who earlier used to work on the strategy of how IT applications within organizations would improve internal operations, are now working on how businesses use IT to communicate with Customers and Suppliers in order to develop new distribution chains and new methods of marketing and selling [14]. If we talk about customer relationship management (CRM), Web mining is the process of integrating information gathered through traditional data mining methodologies and techniques such as Clustering, Classification, Association rules. Web mining is an extension of Data Mining concept. Web mining is used to understand the behavior of customer, derive the capability of a Web site, and help quantify the effect of a campaign made for a product's marketing [15].

Web mining concept allows one to obtain patterns out of data collected through various mining techniques-content, structure, and usage. The first kind of web mining, i.e. content mining, is the process that is frequently employed to examine data that are collected through search engines and Web spiders. Web spider is a program that visits websites and other information on Web in order to create entries for a search engine index [16] .Structure mining, on other hand, is another web mining technique that is used to examine data related to the structure of a particular Web site, and finally there is usage mining which is the famous web mining technique to examine data related to a particular user's browser as well as data gathered through online registration or signup forms available at various websites.

3.1 Challenges before Web Mining

Web mining, unlike to other sources to Information Technology, is more complex as well a challenging task which requires a sheer planning, an excellent management and a talented workforce. There are myriad challenges to web mining. A Web is characterized by a huge amount of data/information [17] that keeps growing day by day at every passing minutes and hours. The coverage of web information is quite wide as well as diverse where anyone can search for any kind of information as per his/her area of interest since all kinds of data exist on the web including text, images, audio, video, structured tables, etc. In other words, web consists of heterogeneous type of data/information where similar or different formats and syntax based myriad web pages exist whose information integration is a daunting and cumbersome task. Further, due to the semi-structured nature of the most of the web/HTML document, the need to present information in a simplified and readable form to achieve human viewing and browsing puts extra burden on the web mining. Due to the links among web pages within a site and across

different sites that serve as an information organization tool and as an indicator of trust/authority in the linked pages and sites, much of the information on web is redundant [17].Though this property of redundancy of web data is helpful in various web data mining tasks, this also leads web data to be noisy. One of the important feature of web data is that they can be either surface web data (consisting of pages that could be processed through normal web browser) or deep web data [18][19][20][21] (comprising of database that could be searchable only through popular search engines which employs parameterized queries of query form for accessing web data) which takes the task of web mining to certain level of complexity. Today, web also plays a significant role in providing myriad services but in some cases certain loopholes occur in the method adopted to provide services online such as the induction of intrusion and malicious code to the web query by hackers and crackers to gain access of transactions made for such services. To counter such problems, there are numerous web mining techniques available that help to build secure websites and web servers. Since the web is dynamic, the information on the web keeps on changing constantly. For the purpose of keeping up the changes and monitoring them on the web, web mining provides several sophisticated tools and reliable techniques.

3.2 Search Engine and Web Spider

A successful web mining demands an efficient, reliable and a powerful search engine working on a compatible web browser. All major search engines on the web consist of a program known as web spider or "crawler" or "bot" which typically visits sites that have been submitted by their owners as new or updated [16]. For example [22]:

a. Googlebot is the search bot used by the Google.com;
b. Yahoo Slurp is the search bot used by the Yahoo.com;
c. MSNbot is the search bot used by the MSN for its Search.MSN.com;
d. Ask Jeeves/Teoma is the search bot used by the Ask.com;
e. Architext Spider is the search bot used by the Excite.com; and
f. ia_archiver is the search bot of Alexa.com.

But it is noteworthy that most of the search engines fail to handle the following knowledge discovery goals [23]:

A. From the query's result returned by search engine, a user may wish to locate the most visible websites [24] or documents for future reference i.e. many paths (high fan in) can reach that sites or documents, as compared to the way where at present he/she can do so only by manually visiting the documents and then download the visible documents as files on user's hard disk for future reference which is quite a cumbersome task.

B. If we understand this concept in the context of visibility, it could be possible that a user may try to determine the most luminous websites/ documents [23] that could be referred in future i.e. web sites/documents having the greatest amount of outgoing links, as compared to the way where at present he/she may determine this information only by manually visiting each web documents which is again a quite daunting task.

C. Furthermore, a user may wish to find out the most traversed path for a particular query result which is important by the matter of fact that it helps the user to identify the set of most popular interlinked web documents that have been traversed frequently to obtain the query result. This if compared to the way where at present he/she may do so only by visiting each document in the search result and compare their link information, is quite time consuming.

3.3 Web Bag

Sourav Madaria and his colleagues [25] has discussed about the concept of Web Bag to deal the problems mentioned in above section while working with search engines to acquire knowledge. Basically, a Web Bag is a web table containing multiple occurrences of similar web tuples. In general, a web tuple comprises of a set of inter-linked documents that are usually accessed from WWW which helps in satisfying a query graph. The creation of a Web Bag requires the projection of some of the web tuples' nodes contained in the web table by making the usage of web project operator (it is used to isolate the data of interest) which allows the considerable queries to run over a smaller, perhaps more structured web data. In contrast to this, a relational table never allows a web project operator to remove similar web tuples in automatic way, thus leading to further anomalies. Whereas Web Bags simply discover web documents that are visible by searching luminous paths [25].

3.4 Virtual Society

Over the past decade, the web has evolved into a virtual society which involves interactions among people, organizations and automatic system. When the first social networking site, SixDegrees.com was launched in year 1997 after the earlier famous chat community sites like Theglobe.com (1995), Geocities (1994) [26] and Tripod.com (1995), the world had witnessed an awesome virtual society prevailed over the web which brought people together online to interact with each other through chat rooms where users can share ideas, personal experiences and updated information over the personalized web pages that led them to have an easy access to publishing tools and web space that were basically free or inexpensive [27]. With the further advancements and development in web mining tools and techniques, such virtual society has been transformed into a platform over the years where the users can socialize themselves over the web through social networking sites. Here they can share

information of their interest, can perform numerous activities, and can reveal background details or real-life instances & experiences. Moreover, in order to recognize each user online, today's social networking sites like Facebook, YouTube, Google+, LinkedIn have included the features to create profile, his/her social links, and a variety of additional services.

The latest development in the field of above discussed virtual society is the incorporation of new age information as well as communication tools, noticeably the mobile connectivity facilities, sharing of images, audios, videos and posting blogs [28]. This has contributed the Information Technology a lot by providing an individual-centered service rather than online community services which are group-centered that allow users to share audio, video, ideas & information, real life experiences, posts and to perform activities, participate in events within their network only. All this has become possible only because of the advancement brought in web mining concept like web personalization and semantic web mining.

Today various organizations, news agencies, e-commerce industries and other business houses have also become the part of these virtual societies and sites to advertise & market their products and to build a reliable customer relationship. This is one of the major developments that have taken place in Information Technology over the last decade that is mainly attributed to astounding breakthroughs achieved in the development of web mining strategies.

3.5 Semantic Web Mining

The Semantic Web Mining is a visionary idea of the inventor of WWW, Tim Berners-Lee, which focuses on the improvements that could be made at the levels of Web service, plus it addresses problems related to the existing Web services. It uses Web Ontology Language (OWL) [29] to mark up hypertext pages by allowing more complex assertions about a page (for instance, its methods of proof, access rules, and links to other pages) as compared to the Web-as-database approach which is restricted to just simple metadata. Further, such assertions are provided by the language (OWL) with explicit semantics which makes it machine interpretable [30]. The knowledge of Semantic Web not only helps to achieve Web mining easily, but also helps in improving the capability of Web mining tools which simply helps to contribute Information Technology at significant level to a larger extent in knowledge acquisition and intelligence building [31].

3.6 Cloud Computing

The greatest discovery in the Information Technology field that has been made through the web mining concept in the last decade is Cloud computing. The term cloud, or cloud computing can be used as a synonym for the internet that revolves around the building of infrastructure and business model whereby the data, news, entertainment and other products/services are delivered to one's device (including PCs, Laptops, Mainframes, and Mobiles) in real time from the internet, rather than storing such information resources on the device. The creation of the cloud has been helpful not only to hosting companies but also to consumers who just need any of the above discussed device and internet access to fulfill most of their computing needs. Google is perhaps the most significant cloud-based company that has shown the potential of a cloud platform over the years to derive a hugely successful business model including Gmail, Google documents, Google map and the latest development in building the digital library. There are basically two kinds of clouds:

A. *Sector storage cloud*: It provides block or file based storage service which is basically a distributed storage that can be deployed over a wide area network (WAN) and allows users to consume & download large dataset from any location with a high speed network connection to the system. This kind of cloud replicates files automatically in order to provide better access, reliability and availability.

B. *Sphere compute cloud*: It provides the computational services and is built on the top of the sector storage kind of cloud which facilitates developers to build certain distributed, parallel and especially data intensive applications with numerous simple APIs. The key factor of this kind of cloud is data locality that enhances the performance of cloud computing.

So, cloud computing is the most promising application of web mining concept that contributes a huge to the expansion of Information Technology. The following issues are catered while developing such a sophisticated computing technology:

- since the development of cloud demands an appreciable consumption of energy, it is important to pre-determine how big is the cloud in electricity consumption and Green House Gas emissions and how big will it become;
- which is the most suitable location for building the cloud at the expense of what kind of sources of energy for powering it;
- how many large data centers would impact the surrounding load center's demand for fossil fuels; and
- What would be the extent of efficiency as well as design improvements in order to minimize the rate of growth?

There are major efforts that have been carried out by cloud computing honchos like Google and Facebook to resolve above issues and the various plans & initiatives are being under proposal that could effectively enhance Green Computing in the coming years. So we can simply hope for having a better, greener and safer way to boost Information Technology in future.

4. CONCLUSION

This paper is an honest effort to let the readers know about the impact brought by the concept of web mining to "Information Age". Various papers and books have been thoroughly studied, and several web pages have been surf by us to present this paper where we have put our ideas and thoughts to deliver how the advancement made in web mining tools and technology to benefit the Information Technology. We have tried to show that if Information Technology is the fabric of daily life, supporting daily activities at home, work and school, Web mining in today's era is its sewing tool. The past few decades have seen myriad improvements and breakthroughs in the field of web mining strategies that have made the task of accessing information a mere child's play and, has consequently widened the horizon of Information society to a larger extent. A major factor contributing to this is the fast pace development in the Internet languages, technologies and web mining tools and methodologies such as search engines, web spider, web bags, semantic web and cloud computing.

Even today with the uneven growth in malicious code and intrusion activities across the web, the fascination for web mining and it's usage among masses haven't felt any setback, rather this fascination & usage have kept growing more than which hasn't been experienced before. It's all because of new web mining techniques, tools and methodologies applied to build ever evolving, reliable and secure web applications to contribute massively to today's information age.

5. ACKNOWLEDGEMENTS

We are very thankful to each other who have maintained a thorough understanding among ourselves throughout the development of this survey paper and co-ordinate each other in getting out fantastic ideas and words to describe the myriad issues and concepts. Moreover, we are thankful to our family and friends who have also participated in our discussion while preparing this paper. At last but not the least, we are extremely filled with gratitude for all those authors and researchers whose research papers and their work have provided us good anecdote to carry out our work

6. REFERENCES

[1] http://www.en.m.wikipedia.org/ wiki/ Information_ Technology.

[2] http://www.computerhope.com/ jargon/r/router.htm.

[3] Butler, J. G. "A history of information technology and system".

[4] Levitt, H. J. and Whisler, T. L. 2011. "Management in the 1980s". Harvard Business Review.

[5] Chandler, Daniel, Mundrey and Rod. "Information Technology".

[6] Dictionary of Media and Communication (first edition). Oxford University Press. 1 August 2012.

[7] Voas, J. and Zhang, Z. 2009. "Cloud Computing: New Wine or Just a New Bottle". IEEE Internet Computing Magazine.

[8] http://www.cmlab.csie.ntu.edu.tw/~jimmychad/cmCN2011/Readings/CloudComputingNewWine.pdf.

[9] Petre, R.S. 2012. "Data Mining in Cloud Computing". Database Systems Journal. Volume-III, Issue-3

[10] Radhy, N., Mishra, P. and Panigrahi, R. June 2012. "The survey of Data Mining Applications & Future scope". International Journal Conference on Software Engineering and Information Technology (IJCSEIT). Volume-II, Issue-3, pp. 13-59.

[11] http://www.amiquote.tublr.com/post/10199018126/what-is-the-signficance-of-the-internet-and-today's-communication-revolution-for-the-evolution-for-the-mind?

[12] http://www.vidyaprakash.expertscolumn.com/article/signinficance-internet.

[13] http://www.answers.yahoo.com/question/index?qid=20061221005220AACS45b.

[14] http://www.nsf.gov/statistics/seind02/c8/c832.htm.

[15] http://searchcrm.techtarget.com/definition/Web-mining.

[16] http://Whatis.techtarget.com/definition/spi.

[17] Ling, B. and Chen, Chuan, Chang, K. "Editorial: Special issue on web content mining".

[18] He, B. and Chen, Chuan, Chang, K. 2003. "Statistical Schema Matching across Web Query Interfaces". ACM SIGMOD.

[19] He, B. and Chen, Chuan, Chang, K. 2004. "Discovering Complex Matching across Web Query Interfaces". KDD.

[20] He, H., Meng, W., Yu, C. and Wu, Z. 2003. "WISE-Integrator: An Automatic Integrator of Web Search Interfaces for E-Commerce". VLDB.

[21] Zhang, Z., He, B. and Chen, Chuan, Chang, K. 2004. "Understanding Web Query Interfaces: Best Effort Parsing With Hidden Syntax". ACM SIGMOD.

[22] http://www.searchguru.com/blog/search_engine-they-belong-to/.

[23] Madaria, S., Bhowmick, S.S., Nag, W.K. and Lim, E.P. "Research Issues in Web Data mining".

[24] Bray, T., 1996. "Measuring the Web". In Proceedings of the 5th International WWW Conference, Paris, France.

[25] Madaria, S., Bhowmick, S.S., Nag, W.K. and Lim, E.P. 1998. "Web Bags: Are they useful in Web Warehouse?". In Proceedings of the 5th International Conference on Foundation of Data Organization, Japan.

[26] Cotriss and David. "Where are they now: TheGlobe.com". May 29, 1998.

[27] http://en.wikipedia.org/wiki/Social_networking_service.

[28] Journal of Computer-Mediated Communication. October 2007. Volume-13, Issue-1, pp. 210-230.

[29] http://www.w3.org/TR/owl-features.

[30] "Web Mining Research and Practice". Web Engineering article. July/August 2004, pp. 49-53.

[31] Berendt, B., Hotho, A. and Stumme, G. 2006. "Semantic Web Mining: State of the art and future directions". Web Semantics Science, Services and Agents on the World Wide Web, pp. 124-143.