International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 6 (2013), pp. 603-608

http://www. irphouse.com /ijict.htm

An Analysis of Data Mining, Web Image Mining and their Applications

[1]Preetika Bhateja[*], [2]Pinki Sehrawat and [3]Avdesh Bhardawaj

[1,2]Department of Computer Science Engineering and Information Technology, ITM University.
[3]Department of Applied Sciences and Humanities, ITM University
Sector 23 (A), Gurgaon, Haryana, India.

Abstract

Opening up of new sources of data has left no realm possible untouched in the modern world. This has led to an unprecedented proliferation of data. The average daily intake of technology was 10 hours 45 minutes in 2009. Everyday addition of fresh content is leading to more such incredible numbers. Data Mining involves extracting information from data and using it in the future through clustering and anomaly detection. After manual extraction of the data, Bayes' theorem and regression analysis came into being as data expanded by leaps and bounds. Neural networks and Genetic algorithms were used for analysis. Data Mining when constrained to multimedia is referred to as Web Image Mining. Emergence of social media like Flickr, Facebook, LinkedIn, Plurk, etc. has augmented computing rich amount of multimedia data. Image Mining methods include classification and clustering. Applications of Data Mining include games, business, science and engineering, medical, surveillance (PRISM programme of USA). Image Mining plays a crucial role in are audio-visual applications. The data currently existing in the world is expected to see exponential growth thus giving a boom to the job sector of data analysis in the near future. This paper analyses the current applications of Data Mining and Web Image Mining. It also suggests some possible near future applications specially for sustainable and green computing.

Keywords: Data mining, multimedia, social media, web image mining, clustering.

1. Introduction

It is estimated that everyday 2.5 quintillion bytes of data is created. In 2013, the amount of stored information in the world is being estimated to be around 1200 exabytes of which less than 2% is non-digital. Earlier the problem was the storage but now with advanced storage technologies the techies are combating to handle this data. Data is no longer considered waste instead is reused to create new economic value in every domain be it medical , business, policy making, weather forecasting ,social media networking et cetera. With the rapid increase of WWW, websites are an abundant source of information and hence their usage patterns are to be brought to the book. The task of mining is difficult considering the fact that Web traffic volume is enormous. Data collected is both structured and unstructured hence requiring different analytical techniques. IBM gauged that 80% of data captured today is unstructured-social networking sites (videos, audios, images, web text, and geographical information), sensors, and GPS signals being some of the sources.

2. Literature Review and Discussion

2.1 Bayes' Theorem and Regression Analysis

The basic foundation of data mining came from Bayes theorem which was proposed by Thomas Bayes in 1700s.It is used in conditional probabilities and statistical inference. Bayesian interpretation is used measure degree of belief. Regression theorem is used in data mining to define a pattern of the given data set by fitting an equation to it. It is used for predictions and forecasting. The pitfall of regression theorem is that they require scaling predictors before fitting the model and problem arises when predictors are non-linear functions because scaling in that case is not sensible.

2.2 Neural Networks

Neural networks is a must have component in any data mining tool. The ANN (Artificial Neural Networks) was pioneered by McCulloch and Pitts in 1943. It has been inspired from biological neural system. It is a non-linear, statistical computational tool with the ability to learn from the existing data by analyzing its patterns. Connection between two neurons has an associated weight. Each neuron adds this weight to the incoming neuron value and passes the result to the subsequent neuron as an input. It has the capability to detect non-linear relationships and its major disadvantage is its 'Black Box' nature i.e. it's internal working is not transparent.

2.3 Genetic Algorithm

John Holland is regarded as the father of the Genetic Algorithm, who invented it in early 1970's with inspiration from natural evolution process .GA follows Charles Darwin's principle "survival of the fittest ". It is a heuristic search algorithm meant to solve the optimization problems. Generation consists of a population of individual (character strings). Individual solutions compete with each other for resources and

mates, the successful individual is allowed to crossover and mate to produce a new generation of solutions by exchanging chromosomes, hence heading towards a solution better than the either parent. GA cannot be used in real time applications because response time is greater than the conventional methods.

2.4 Hadoop

Hadoop is open source software that manages the data based distributed systems. It provides scalability, reliability, authentication and authorization. It started with batch based MapReduce engines in 2006, inspired by Google paper on GFS and Map Reduce published in 2004. Distributed parallel computing requires reliability and this feature was added to Hadoop with the advent of HBase in 2008. HBase is a first non-batch component in Hadoop world that can retrieve data in real time, it also supports batch processing and is integrated with past computing engines, thus when used with MapReduce, transfer of data is not required,hence providing authentication,authorization and reliability. Lately Impala is added to Hadoop that serves online queries.

2.5 Clustering

In this method firstly the clusters of images are made. The RGB values of the image chosen are relatively close, so that the images are not identifiable. A set of steps which involves, clustering of images, analysis of the patterns obtained by clustering algorithm, to see the changes in the cluster once the input is changed and the patterns are worked out. Also a minimum spanning tree based clustering algorithm came into being in 2010. A minimum spanning tree is obtained on the basis of the clustering algorithm used. This approach helps in case of huge and unstructured clusters.

3. Applications of Data Mining and Web Image Mining

3.1 Telecommunications

Data mining is extensively used in this industry due to adequate amount of data available in structured form, which helps in network analysis, tracking the customers and post services.

3.2 Marketing and Retailing

Marketers can be made aware of the products their customers are interested in buying taking into consideration the patterns obtained. Data Mining can help decide some new marketing strategies, stock their products, discount schemes can be arranged and help in identifying profitable customers.

3.3 Banking Sector and Finances

Data mining can help in credit reporting, credit rating, loan and credit approval by predicting good customers, help in making the policies (insurance) that might attract a good numbers of customers.

3.4 Medicare
Data mining helps in identification of relationships among diseases, keeping track of new drugs and of effective medicines. Information regarding local health care systems and tracking the high-value physicians is made easy. In 2009, with the outbreak of H1N1 virus most parts of the world feared a pandemic, United States' Centers for Disease Control and Prevention was unable to trace the realms it had already affected. Here Google stepped in with its publication of a paper in the journal Nature, which played a huge role in tracking the areas already affected and the ones about to be affected by the virus with its data, processing power, and statistical know how.

3.5 Criminal Investigation
Tracking of fraud relating to funds or chit scams, suspecting criminals by keeping a track record, following the trends of crime type, location or habitat are some of the latest applications of data mining.

3.6 DNA Analysis
One of the most important usages of data mining has been in identifying gene sequence patterns that plays role in different diseases, studying a collection of genes and also their linkages.

3.7 Surveillance
With several countries of the world brining up their surveillance system, it is today one of the most sorted uses of Data Mining. A track record of phone calls, mails, social networking sites, and blogs of the civilians is maintained. Often corporate are too spied on due to government interventions and to keep their monetary activities in check.

3.8 Image Mining
E-commerce has been one of the biggest success stories in the era of digitalization. Image mining can aid in understanding auction behavior in sites like eBay, Personalized Portal for the Web in sites like MyYahoo and Personalized customer experience in B2C E-commerce like Amazon.com.

4. Mining Technology: Future Perspectives
5.1 Surveillance Systems
With the coming up of U.S.A's surveillance program PRISM, it has been made clear that with the data growing by leaps and bounds every day, it must be tracked and analyzed expeditiously. India too is all set with its central Monitoring System (CMS) that is apparently being considered more lethal than PRISM. All this tracking of data has also led to a serious issue that is the violation of the privacy of the citizens and the sovereignty of a nation. With US tracking Petrobras, Brazil's oil producing giant, the question of nation using their new found Data Mining missions against each other in the coming future is a big threat.

5.2 Green Computing
The focus of computation has shifted to efficient power consumption; less computation by improving Algorithmic efficiency and alternative energy sources, product can remain green in all of its four stages: designing, manufacturing and implementation, utilizing and disposal.

5.3Disaster Management
The frequency of disasters has been increasing in the recent past. Data and web image mining can be utilized for early warning information dissemination and thus saving precious lives.

5. Conclusion
Data can be utilized to create wonders which can be further used for Artificial Intelligence, surveillance systems, better tracking systems, designing of government policies, retailing, medicare, etc.It can lead to efficient management of world as a whole if used positively. It can be concluded that data mining and web image mining will play a pivotal role in not only scientific advancements but also sustainable development in the future.

6. Acknowledgements
The authors acknowledge the support and guidance of Mr. RajatBhateja of Mu Sigma during the preparation of this Paper.

References
[1] AleksanderPivk, OlegasVasilecas, Diana KalibatieneandRokRupnik (2013), On approach for the implementation of datamining to business process optimisation in commercial companies, *Technological and Economic Development of Economy*, Volume 19, Issue 2, pp 237-256

[2] Brijesh Kumar Bhardwaj, Saurabh Pal (2012), Mining Educational Data to Analyze Students' Performance, ***International Journal of Advanced Computer Science and Applications,*** Vol. 2, No. 6, pp 63-69

[3] E.W.T. Ngai, Yong Hu,Y.H. Wong, Yijun Chen, Xin Sun (2011), The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision Support Systems*, Volume 50, Issue 3, pp 559–569

[4] Leah P. Macfadyen, Shane Dawson (2010), Mining LMS data to develop an "early warning system" for educators: A proof of concept, *Science Direct:*Computers and Education, Volume 54, Issue 2, pp 588–599

[5] PaulrajPonniah (2010), *Data Warehousing Fundamentals: a Comprehensive guide for IT Professionals*, Wiley-Interscience Publication, India

[6] Siddhartha Bhattacharyya, SanjeevJha, KurianTharakunnel, J. Christopher Westland(2011), Data mining for credit card fraud: A comparative study, *Science Direct:Decision Support Systems*, Volume 50, Issue 3, pp602–613

[7] Viktor Mayer Schonberger and Kenneth Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt Publishing Company, New York.