ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2016; 2(5): 563-566
www.allresearchjournal.com
Received: 05-04-2016
Accepted: 06-05-2016

Dr. A Pappu Rajan
Assistant Professor,
Department of Management
Studies St. Joseph's Institute
of Management, St. Joseph's
College (Autonomous),
Tiruchirappalli, Tamil Nadu,
India.

Correspondence
Dr. A Pappu Rajan
Assistant Professor,
Department of Management
Studies St. Joseph's Institute
of Management, St. Joseph's
College (Autonomous),
Tiruchirappalli, Tamil Nadu,
India.

Web sentiment analysis

Dr. A Pappu Rajan

Abstract

In the real world, data are represented in the form of facts, numbers and text. The data are accumulated in vast quantity and have also grown in different formats and databases. They are categorized as operational, transactional, nonoperational and metadata. These data are analyzed and the knowledge is derived from the original data using data mining. Data Mining is a process of analyzing the data in different perspectives such as association, clustering, classification and regression, prediction. Web data mining can be defined as the discovery and analysis of data from all over the world. World Wide Web has huge volume of data, which may be very useful or sometimes may be useless Meta data. Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage. The information is collected from the analysis in the form of patterns, association or relationships among the data. The collected information is converted into knowledge, which is gathered from the historical terms and is applied as future trends. Data can be collected from the different repositories. They may contain noisy data, redundant, irrelevant and insignificant features. In this scenario, web sentiment prediction plays a vital role of identifying the relevant predictions and data from the dataset. In this article deals concept of data mining, web mining, Information Retrieval, web log mining and opinion mining.

Keywords: Web Mining, Information Retrieval, Opinion mining

1. Introduction

Data mining refers to extracting or mining knowledge from large amounts of data. The data mining should have been more appropriately named as knowledge mining. The knowledge mining as a shorter term may not reflect the emphasis of mining form large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both data and mining became a popular choice. Many other terms carry a similar or slightly different meaning of data mining, such as knowledge mining from data, knowledge extraction, data pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for other popularly used terms. The steps in the process of knowledge discovery in data mining.

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation. In the broad view of data mining functionality data mining is the process of discovering interesting knowledge form large amounts of data stored in databases, data warehouses, or other information repositories

Knowledge Discovery Process in Data mining involves the following steps:

1. Data Cleaning – this step is to remove noise and inconsistencies in the data.
2. Data Integration - multiple data sources may be combined in this step.
3. Data Selection – the data relevant to the analysis are retrieved from the database.
4. Data Transformation where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining – This is an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern Evaluation – This stage is to identify the truly interesting patterns representing knowledge based systems.

7. Knowledge Presentation – This is the final step where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2. Web Data Mining Systems

Web data mining can be defined as the discovery and analysis of useful and relevant information from the World Wide Web data. WWW has a lot of useful or useless Meta data, web log data of users who retrieved multiple web pages and the structured and unstructured data. Over the past Fifteen years, we have already faced lot of explosion type of information resources available over the web. Web mining can be applied to all the fields of artificial intelligence system, human interaction, cloud computing, neural data mining, geographical data mining, and information retrieval etc. Many Web applications focused on extraction of knowledge from the web, extraction of knowledge from the user's behavior, getting information from the web, providing information to the web, downloading and uploading data over the web were developed.

The main objective of the web mining is to provide data mining algorithms which can improve the content, structure, usage, performance, and categorization of web documents, snippets and user sessions. The Web data mining can be classified into three categories namely Web Structure Mining, Web Content mining and Web usage mining. All these three categories focus on the process of discovering unknown data and potentially very useful information from the web. Though each of them focuses on the same attribute each may be using it with different mining objectives on the web.

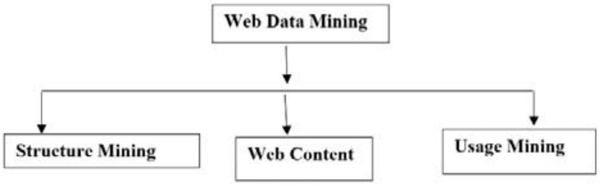

Fig 1: Web Mining Systems

2.1 Web Structure Mining

Web Structure Mining involves mining the structure of web document's and links. Useful insights can be given by mining the structural information on the web. WSM is very useful in generating information such as visible web documents, luminous web documents and luminous paths, a path common to most of the results returned, linkage information useful to improve search engine results, hyperlink structure analysis, link analysis, graph, categorization and mining the document structure.

2.2 Web Content Mining

Web Content mining examines the contents of web pages as well as the results of web searches. WCM is described as the automatic search of information resources available on-line. It represents structured, unstructured, semi structured documents and builds model for interactive retrieval view and Data Base View. It is all about extracting and integrating of useful data with the objective of information and knowledge discovery from Web page contents. WCM has two different major approaches. One is Agent based approach and another one is Data base based approach. First approach is on improving the information finding and filtering which are carried out using intelligent search agents, or personalized web agents. Second approach is on modeling the data on the web into a more structured form by connecting with multilevel data bases and web query systems

2.3 Web Usage Mining

Web Usage Mining focuses on several techniques that could help in learning or predicting user behavior and navigation pattern of users using the web round the clock. It includes the data from server access logs, user registration or profiles, user sessions or transactions etc., It also depends on the collaboration of the user to allow the access of the web log records.

3. Information Retrieval Systems

IR involves retrieving desired information from textual data. The historical development of IR was based on effective use of libraries. Many universities and public libraries use IR systems to provide access to books, journals and other documents. IR effectiveness can be measured by using a technique called test collection. The test collection consists of document collection, a test suite of information needs express as queries and a set of relevant judgments, standard binary assessment of either relevant or non-relevant response for each query-document pair. Automated IR systems are used to reduce information overload. Web search engines are the most visible IR applications. In the work of searching, retrieving data from web, we are using several other words such as data retrieval, document retrieval, information retrieval, text retrieval. There a number of words overlapping but with similar meaning and tasking almost similar. For the information retrieval to be efficient, the documents are typically transformed into a suitable representation. There are several representations available like theoretical, probabilistic and future based retrieval models. As a result, some traditional data mining methods are not applicable to web mining data retrieval. The key problem of information retrieval in web mining is how to improve comprehensive and correlated information accessed from the web data base and make efficient information retrieval for classification of different web pages for retrieving relevant information. The challenges for IR is to deal with the structure of the hyperlinks within the web itself. IR Link analysis is an old area of research. However with the growing interest in web mining, the research on structure analysis has increased and this had resulted in a new emerging research called link mining which will help in retrieving accurate information without spreading spam, even from irrelevant or unwanted web pages.

4. Web Log Data Preprocessing

Web Usage Mining is one of the applications of data mining techniques to discover usage patterns from global Web data. WUM includes data preprocessing, pattern discovery and pattern analysis. In the preprocessing phase raw Web logs are cleaned, analyzed and then converted in to pattern mining process. Data pre-processing includes the following process: cleaning, normalization, transformation, feature extraction and selection etc., of the data recorded in the server logs system. The logs are used to identify users and sessions. Sometimes server logs analysis is not accurate and reliable. So the system also considers cookies and sessions.

The server logs authentication of server logs must be formalized with standard set of format and it should be updated to capture user access data. Most of the preprocessing techniques suffer from low quality. So the system should improve the quality of preprocessed data and their algorithms. A new technique is essential to analyze the log file. The Basic Process of web Log Mining should concentrate in Data Preparation, Data Mining, and Pattern Analysis. Using these techniques, the problem can be solved and ultimately data can be converted into knowledge. By doing a survey of literature on web log preprocessing, it was found that web log systems have these essential techniques to get exact pattern. The steps are data cleaning, data filtering, path completion, user identification, and session identification, cluster of web session and data visualization. WLM is used to enhance server performance, improve web site navigation, improve the design of web applications, and improve the multidimensional web log analysis, identifying web access association or pattern analysis. The pattern analysis is used to analyze web cashing, pre-fetching, swapping and frequently used predefined reports. The report should include the following information: HITS information, list of top requested URLs, referred, list of common browsers, HIT per time and error report.

5. Web Sentiment Analysis

Sentiment mining or Opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes. In conversation mining the additional task is to determine the nature of opinion whether it is positive or neutral in general. Opinion mining is a type of Natural language processing for tracking the attitudes, feelings or appraisals of the public about particular topic, product or services. Sentiment analysis can be useful in several ways. For example, it tracks and judges the success rate of an advertisement campaign or launch of new product, determines popularity of products and services with its versions and also tells us about demographics which like or dislike particular features. The analyst company gets a much clearer picture of public opinion than surveys or focus groups, if this kind of information is identified in a systematic way.

5.1 Level of Sentiment Analysis

The sentiment analysis tasks can be accomplished at the following levels of granularity, namely word, sentence, document, and feature level.

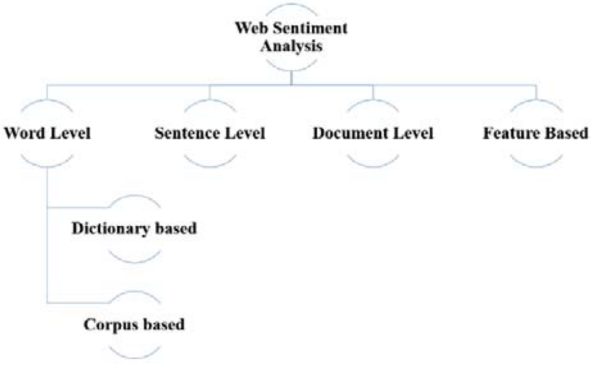

Fig 2: Level of Sentiment Analysis

5.2 Word Level Sentiment Analysis

The word level sentiment analysis is used to find semantic orientation at phrase level. Most previous works use the prior polarity of words and phrases for sentiment classification at sentence and document levels. There are two methods of automatically annotating sentiment at the word level such as dictionary-based and corpus-based ones. The Dictionary based sentiment analysis means a small seed list of words with known prior polarity is created. This seed list is then extended by extracting synonyms or antonyms iteratively from online dictionary sources like Word Net. Corpus based methods rely on syntactic or statistical techniques like co-occurrence of word with another word whose polarity is known.

5.3 Sentence Level Sentiment analysis

At sentence level sentiment analysis is used to detect subjective sentences in a document from a mixture of objective and subjective sentences and also the sentiment orientation of these subjective sentences is determined.

5.4 Document Level Sentiment Analysis

Document-level sentiment analysis is considering the whole document as the basic unit whose sentiment orientation is to be determined. To simplify the task, it is presumed that each text‘s overall opinion of each text is completely held by a single opinion holder and is about a single object.

5.5 Features Based Sentiment Analysis

This level of sentiment analysis is used to extract system feature and the corresponding opinion about it. The opinion may be positive or negative of a particular system. The person may like some features and dislike some, even though the general opinion of the system may be positive or negative.

6. Conclusion

In this article the basic concepts of data mining, web mining systems, IR, Web mining, Sentiment analysis and the relevant terminologies are discussed. In real world, web analysis plays an important role in understanding the data and knowledge discovery from the real input data. In web sentiment analysis, understanding the data and knowledge discovery are the crucial parts of data mining process. In analyzing the data, the relevant data are extracted and used for prediction in data mining. In relevant data extraction, sentiment prediction plays the role of identifying the highly relevant features from the original data. Web sentiment mining is a new and promising research area which will to help users in gaining insight into overwhelming information on the web social media.

7. References

1. Abdulrahman Al-Senaidy M, Tauseef Ahmad, Mohd Mudasir Shafi. Privacy and Security Concerns in SNS: A Saudi Arabian Users Point of View, International journal of Computer Applications. 2012; 49(14):1-5.
2. Akshi Kumar, Teeja Mary Sebastian. Sentiment Analysis on Twitter, International Journal of Computer Science. 2012; 9(4):372-378.
3. Alexander Liu Y, Bin Gu, Prabhudev Konana, Joydeep Ghosh. Predicting Stock Price from Financial Message Boards with a Mixture of Experts Framework

Intelligent Data Exploration & Analysis Laboratory, 2006, 1-14.

4. Andreas Mastel, Jurgen Jacobs. Mining User-Generated Financial Content to Predict Stock Price Movements, Technical Reports and Working Papers, Leuphana Universitat Luneburg, 2012, 1-30.
5. Anitha Anitha B, Pradeepa S. Sentiment Classification Approaches – A Review, International Journal of Innovations in Engineering and Technologies. 2013; 3(1):22-31.
6. Arti Buche, Chanak MB, Akshay Zodgaonkar. Opinion mining and analysis: a survey, International Journal on Natural Language Computing, 2013; 2(3):39-48.
7. Ashok Srivastava N, Mehran Shani. Text Mining, Classification, Clustering and Applications, CRC Press, 2009.
8. Bernard Jansen J, Mimi Zhang, Kate Sobel, Abdur Chowdury. Twitter power: Tweets as Electronic Word of Mouth, Journal of the American Society for Information Science and Technology. 2008; 6(11):2169-2188.
9. Blessy Selvam, Abirami S. A Survey on Opinion Mining Framework, International Journal of Advanced Research in Computer and Communication Engineering, 2(9), 3544-3549.
10. Bo Pang, Lillian Lee. Opinion mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, 2012; 2(1):1-135.
11. Brain Clifton. Advanced Web Metrics with Google Analytics, 3rd Edition, John Wiley & Sons, 2012.
12. Carvalho AX, Tanner MA. Mixtures-of-Experts of Autoregressive Time Series: Asymptotic Normality and Model Specification, IEEE. 2005; 16(1):39-56.
13. Carvalho AX, Tanner MA. Mixtures-of-Experts of Autoregressive Time Series: Asymptotic Normality and Model Specification, IEEE. 2005; 16(1):39-56.
14. Chhavi Rana. Trends in Web Mining for Personalization, International Journal of Computer Science and Telecommunications. 2012; 3(1):260-265.
15. Christian Bissattini, Kostis Christodoulou. Web Sentiment Analysis for Revealing public Opinions, Trends and Making Good Financial Decisions, Journal of Advanced Research in Computer Science and Software Engineering, 2013, 1-9.
16. Devmane, Rana. Privacy Issues in Online Social Networks, International Journal of Computer Applications, 2012, 41(13).
17. Han Jiawei, Kamber Micheline, Pei Jian. Data Mining: Concepts and Techniques, 2nd Revised Edition, Morgan Kaufmann Publishers, 2006.
18. Jagtap VS, Karishma Pawar. Analysis of different approaches to sentence-level sentiment classification, International Journal of Scientific Engineering and Technology. 2013; 2(3):164-170.
19. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations 2000; 2:12-23.
20. Joshila Grace LK, Maheshwari V, Dhinaharan Nagamalai. Analysis of Web Logs and Web User in Web Mining, International Journal of Network Security & Its Applications. 2011; 3(1):99-110.
21. Avinash Kaushik. Web Analytics 2. 0 - The Art of Online Accountability and Science of Customer Centricity, Sybex, Wiley, 2009.
22. Liu Hongyan, Lin Yuan, Han Jiawei, Liu Hongyan. Journal Knowledge and Information Systems archive. 2011; 26(1):1-30.
23. Prabhakar Raghavan, Christopher Manning D, Hinrich Schutze. An Introduction to Information Retrieval, Online Edition, Cambridge University Press, 2009.
24. Padmaja S, Sameen Fathima. Opinion mining and sentiment analysis – An assessment of peoples' belief: A survey, International Journal of Ad hoc, Sensor & Ubiquitous Computing. 2013; 4(1):21-33.
25. Ralf Mikut, Markus Reischl. Data Mining Tool, Wiley Inter disciplinary Review: Data Mining and Knowledge Discovery 2011; 1(5):431-443.
26. Songwattana. Mining Web logs for Prediction in Perfecting and Caching, IEEE. 2008; 2:1006-1011.