ISSN 2394-739X

IJRSET JUNE Volume 9 Issue 6

International Journal for Research in Science Engineering & Technology (IJRSET)

https://www.doi.org/10.5281/zenodo.6793309

SURVEY ON DATA MINING FOR WEB SEMANTIC ANALYSIS

[1] H. Ramprasanth, [2] Dr. M. Shanmugapriya
[1, 2] Associate Professor,
[1, 2] Department Of Computer Science,
[1, 2] Park's College, Chinnakarai, Tirupur, 641605
[1, 2] Tamilnadu, India.

ABSTRACT - Data mining plays an important role in various human activities because it extracts the unknown useful patterns (or knowledge). Due to its capabilities, data mining become an essential task in large number of application domains such as banking, retail, medical, insurance, bioinformatics, etc. To take a holistic view of the research trends in the area of data mining, a comprehensive survey is presented in this paper. This paper presents a systematic and comprehensive survey of various data mining tasks and techniques. Further, various real-life applications of data mining are presented in this paper. The challenges and issues in area of data mining research are also presented in this paper. The main objective of the paper is to analyze and compare different data mining techniques used in the medical applications. We present a summary of the results and provide comparison analysis of the data mining methods employed by the reviewed articles.

Keywords: [Data mining, Neural Network, Naive Bayes.]

1. INTRODUCTION

The World Wide Web is generally available to people while machines just have an extremely simple comprehension of its content. The original vision behind the Semantic Web (hence SW) is that computers ought to by one way or another has the option to comprehend and abuse information offered on the Web. The SW depends on two segments: (1) Formal ontology's give area explicit foundation knowledge as a jargon that is shared by a few gatherings and that depicts abstract item classes, predicate classes and their interdependencies, formalized in legitimate proclamations; (2) Annotations of web assets with explanations which can be perused and interpreted by machines through the basic ontological knowledge present started up certifiable perceptions.

Web-users exhibit various kinds of conduct contingent upon their information needs and their proposed errands. These undertakings are caught certainly by an assortment of actions taken by users during their visits to a site. For instance, in a dynamic application-based internet business Web site, user assignments might be reflected by successions of interactions with Web applications to look through a list or to make a buy. Then again, in an information-escalated site, for example, an entry or an online news source, user undertakings might be reflected in a progression of user taps on an assortment of Web pages with related content.

Semantic Web

The Semantic Web is certifiably not a different Web yet an extension of the current one, in which data is given all around characterized importance, better-enabling computers and individuals to work in cooperation. For novices to the Semantic Web, which is regularly taken as the beginning stage for the research zone, is as great a beginning stage as any? The objective of the Semantic Web is in some sense a contrast to the Web of 2001. That Web was designed as a worldwide report archive with simple routes to access, distribute, and connect records and Web records were made to be accessed and perused by people. The Semantic Web is a machine-readable Web. As suggested over, a machine-readable Web encourages human-PC cooperation. As suitable and required, certain classes of errands can be appointed to machines and consequently handled naturally. Obviously, the plan opportunities for a machine-readable Web are exceptionally large, and a more number of plan.

Data mining

Information mining predominantly manages organized information coordinated in a database while text mining primarily handles unstructured information/text. Web mining lies in the middle and adapts to semi-organized information or potentially unstructured information. Web digging calls for inventive utilization of information mining or potentially text mining methods and their particular methodologies. Mining the web information is quite possibly the most difficult errands for information mining and information the board researchers in light of the fact that there are enormous heterogeneous, less organized information accessible on the web and we can without much of a stretch get overwhelmed with information.

2. LITERATURE SURVEY

1. **Adamov, A. (2014)** et.al proposed "Data mining and analysis in depth. Case study of Qafqaz University HTTP server log analysis" The touchy development of data accessible online brings new challenges of mining and examination of the web to discover helpful information. Since Web mining is predominantly founded on text mining techniques, it works with unstructured or semi-

organized data. Web mining is the way toward investigating and changing over immense volumes of pointless data into significant information by determining intriguing patterns and relations among gatherings of data. Considering the fast development and solid rivalry in the market of internet business and online services, effectiveness and comfort are key variables for progress. Following user action on the website through investigation of the log documents is a fundamental advance in determining the best utilization patterns and adjusting the user interface to the individual user's necessities.

Merits

1. Web mining is the process of investigating and changing over gigantic volumes of futile data into significant information by determining fascinating patterns and relations among gatherings of data.
2. The assortment of data types and the unstructured nature of the substance utilizes unfeasible and hard. It is the reason Web mining strategies and application have gotten one of the key research zones of data mining.

Demerits

1. The dangerous development of data accessible online brings new difficulties of mining and analysis of the web to discover valuable information is somewhat difficult.

2. Singh, S. P., & Meenu. (2017) et.al proposed "Analysis of web site using web log expert tool based on web data mining" Web Usage Mining (WUM) is important for Web Mining. There is a data mining instrument to get and find information from web data. Web usage mining utilizes the data mining measure for the research of the usage pattern from data brought from the weblog files. The web is the assortment of scholarly educational institute marry server data was dissected to help the institute further improving the terms and policies of the service they give. It helps in the assessment of the viability of the web website, and supportive in acquiring achievement in a marketing effort. Web mining likewise considers extricating patterns in data with the assistance of construction mining, content mining, and usage mining. The web analyzers investigate them cut off log files for the assurance of the framework blunder. The scholastic educational institute marry server data was dissected to help the institute further improving the terms and policies of the service they give.

Merits

- Web Usage Mining is utilized for E-Learning, E-business, E-Commerce, E-Government, E-Newspapers, and Digital Libraries Making.
- Personalization for a client can be performed by monitoring beforehand access pages.

Demerits

- The Web mining process would be not effective if the samples are not a better representation of the bigger body of data.

3. Chen, J., Li, K., Liu, Z., Zhang, T., Wen, W., Song, Z., Huang, T. (2019) et. Al proposed "Data Analysis and Knowledge Discovery in Web Recruitment Based on Big Data Related Jobs" Presently, as per the operation model; the online recruitment industry can be separated into classified data recruitment websites, exhaustive recruitment websites, vertical recruitment websites, search recruitment websites, neighbourhood recruitment websites and social recruitment websites. Network data mining depends on insights, machine learning algorithm and data mining innovation to measure and dissect the huge network data, find the network data recruitment interest through modelling, and afterward foresee the future profession request pattern. Enormous data-related positions centre on back-end development, operations, and items, trailed by DBA, client, and promoting. It is clarified that the principle occupation of large data post is data-driven, giving vital choice help to key zones of big business like item system, operation strategies, client research, market pattern and client picture.

Merits

- TF-IDF strategy is picked to choose the accompanying stop words: I, have, is, and so forth the chose stop words are added to the stop words rundown, and afterward the stop words list is utilized to filter the stop words.
- LDA model is broadly utilized in text clustering, similitude calculation and different fields.

Demerits

- Network data mining depends on statistics, machine learning algorithm and data mining technology to process and examine the enormous network data, find the network data recruitment interest through displaying.

4. S. Zhou, X. Zhang, X. Li, G. Zheng and G. Zhang (2017) et. Al proposed "Design and Implementation of Data Mining and Analysis Platform Based on Web Service Techniques," Taking into account the straightforwardness of conventional data mining and analysis stage application situations, a high data mining and analysis stage dependent on B/S Architecture was planned, which was called HSP-DMA. The stage incorporated the calculation service with the application service and utilized the REST Web Service to distribute as a segment service to encourage the outsider system call. HSP-DMA embraces the intuitive modelling strategy, which is not difficult to work and decreases the trouble of modelling. It joined the business situation to build the application model and fulfilled the real business needs of the power grid. It incredibly improved the application degree and diminished advancement costs. The HSP-DMA stage is planned dependent on Web Service, and every module has been delivered by segment service, which is advantageous for the outsider system call. The stage incorporates an assortment of develop data mining items and supports an assortment of data files, standard database, with internet modelling, model management, and model publishing and other data mining coordinated management capacities.

Merits

- Web service is a platform-independent, low-coupled, self-contained, web-based application. It is regularly used to develop circulated, interoperable applications.
- Data decrease strategies can be utilized to acquire the data set protocol said it is a lot smaller yet at the same time close to keep up the uprightness of the original data.

Demerits

- A definitive users of the platform and they accomplish their every day work through the browser interface is hard for different users.

5. Hidayat, W., & Yaqin, A. (2019) et.al proposed "Business Trends Based on News Portal Websites for Analysis of Big Data Using K-Means Clustering" Business analysis is performed to decide the business that is well known, in Indonesia with text mining can take information from a few news portals in Indonesia. Text pre-processing is utilized to change the text and tags on the news to be changed over into loads. The heaviness of the information will be handled utilizing the K-Means algorithm to be assembled into clusters and each cluster will be pictured utilizing Word Cloud so that words that regularly show up as famous word recognizable proof are known. Testing utilizes the Silhouette Coefficient to calculate the quality of every part against the cluster. Moreover, every part will be deciphered by the test outcomes. The analysis is completed each month in 2018 with a sum of 995 information with a month to month normal of 6 clusters, in January were the most well known business as indicated by the quantity of individuals from 64 information shaped 6 clusters, the most part clusters were cluster 1 the Silhouette Coefficient test results are solid 0.00%, medium 65.22%, feeble 30.43%, not generous 4.35%, Word Cloud framed was a cowhide pack business. In the wake of realizing the number k moreover decides the centroid of the cluster haphazardly and calculates the distance utilizing the Euclidean equation. Besides, the assurance of the centroid of each gathering is taken from the normal (mean) of all information esteems in each cluster and recalculated the distance between information to centroid until the cluster part has no change.

Merits

- Testing utilizes the Silhouette Coefficient to calculate the quality of every part against the cluster.
- Text mining from title and tag in news entries at that point processed utilizing text pre-processing and standardized to defeat a lot of distance to the load between documents.

Demerits

- The more weight the relevant data builds the nature of clusters and the more similar words will explain, Word Cloud not to be changed after the analysis process.

6. M. Driss, A. Aljehani, W. Boulila, H. Ghandorh and M. Al-Sarem (2020) et.al proposed Servicing Your Requirements: An FCA and RCA-Driven Approach for Semantic Web Services Composition. The widespread use of the Web and the development of network technologies is the next step in the evolutionary implementation chain of distributed applications that led to the emergence of the Web service paradigm. Web services have emerged as a new technology that, through interoperability opportunities it offers, ranks now as a focal point of multiple technological actors from various fields such as e-commerce, e-learning, e-government, or other fields. In this paper, we presented a novel requirement-driven approach that ensures the discovery, the selection, and the execution of optimal semantic Web services satisfying user's functional and non-functional requirements specified in terms of QoS, QoE, and QoBiz properties. This proposed approach experimented using an extended version of the OWLS-TC dataset, which includes more than 10830 semantic Web services descriptions.

Merits

- The optimal composition satisfying the user's and organization's requirements with high accuracy and efficiency.
- The validation of the user's satisfaction is ensured by monitoring the obtained service-based applications.

Demerits

- Perform more experiments with large and complex datasets of micro services did not collect from different cloud.

7. S. Arora and N. Baliyan (2019) et.al proposed Extraction and Analysis of Information in News Domain Using Semantic Web. The present news search engines are created in such a way that they give the news to the users based on their rankings which is based on the relevance of the news. The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C).This paper focuses on the need of gathering news from different result sets of the query and integrating the relevant result obtained at one single place which reduces the human efforts and also the time that is spent in finding the relevant news. Later the words are grouped together so as to find the frequency of their occurrence and their relevance too. Further they are classified and drawn into RDF graphs. Finally, the data can be retrieved using various queries over the Ontologies created.

Merits

- The semantic analysis of the information we extract so that only positive and true news will be delivered to the user and the fake news is discarded and not delivered to the user.
- The semantic web techniques are better than those which were used earlier , as the Ontologies leverage the technique used.

Demerits

- No more Improvement in scalability and smooth deployment.
- Semantically if the content is enriched it did not increase the ability of the audience to discover, navigate and share the content and the information.

8. B. Vijaya and P. Gharpure (2019) et.al proposed Candidate Generation for Instance Matching on Semantic Web.The world is experiencing an exponential growth in data creation from various sources such as Knowledge graphs, social networks, ubiquitous computing devices such as smart phones, wearable devices, sensors etc. For large-scale multinational organizations storing, analysing, consuming and deriving knowledge from the data is primary concern. In this paper, we studied the working of inverted index structure that can support approximate search queries. Comparison of four algorithms that uses trigram-based index was carried out. It was seen that candidate generation and reduction using probabilistic approach yielded good result. One of the possible future directions is to extend these methods on data sets described using larger Ontologies with complex structure. The candidate reduction approaches using threshold and probabilistic model that uses linear interpolation improves the efficiency and effectiveness. Since heterogeneity is an inherent feature of data on the semantic web, hybrid approaches that combine different algorithms will be of great interest.

Merits

- The candidate reduction approaches using threshold and probabilistic model that uses linear interpolation improves the efficiency and effectiveness
- Generation and reduction using probabilistic approach yielded good result.

Demerits

- Redundancy is another serious issue as same resources or instances appear in multiple data sources however with different resource identifiers.

9. X. Chen, C. Tian and T. Wu (2018) et.al proposed The Semantic Web Approach for the Collaborative Analysis and Visualization of Ethnic Education and Vocation.economic growth period compared with developed eastern regions of China. Special policy support for ethic areas and population is long being adopted by the central government to foster educational, vocational and economic growth for minority nationalities. Recently released initiatives such as “the Belt and Road Initiative” (B&R), “Massive Entrepreneurship and Innovation” (E&I), etc., indeed cover different facets of the ethnic development and help with its growth. This paper brings together the meaningful expressiveness of Semantic Web, powerful ranking ability of Page Rank algorithm and rich features offered by Baidu Map to conduct the collaborative analysis and visualization of ethnic educational and vocational statistics. According to the demonstration and experiment results based on Southwest Minzu University published data, hidden relationships can be intuitively mined from raw datasets in a user-friendly manner powered by visual added SPARQL queries and Page Rank ranking. In our future work, data of various aspects of ethnic areas and population, such as public sentiment, healthcare, etc., are going to be collected to reveal a more comprehensive ethnic development status.

Merits

- Semantic Web approach with Baidu Map offers a more comprehensive analysis and visualization platform.

Demerits

- Data of various aspects of ethnic areas and population did not work, such as public sentiment, healthcare, etc.

10. J. Fabra, M. J. Ibáñez, P. álvarez and J. Ezpeleta (2018) et.al proposed Behavioral Analysis of Scientific Workflows with Semantic Information.In the last years; scientific computing workflows have gained a lot of interest in different areas related to science and human life. Scientific workflows are a special type of workflows which often underlies many large-scale complex e-science applications such as climate modelling, structural biology and chemistry, medical surgery or disaster recovery simulation, among others. Petri nets and model checking techniques are widely used in different application domains. This work has focused on their application to the area of scientific workflow analysis. However, the proposed method has some important limitations. For a given task specification there are different types of post conditions that could be defined. Many of them could be evaluated without executing the task invocation, and there are sets of post conditions that could only be evaluated by means of the task execution. Finally, the COMBAS framework allows the use of different RDF storages, so we are carrying out a study to analyse and improve the efficiency of the overall system as each RDF solution exposes different costs depending on the inference engine.

Merits

- Semantic aspects to workflow models allow a higher flexibility for analysis and improve resource usage when dealing with complex problems.
- Problems and challenges that were too heavy or time consuming solved in a more efficient manner.

Demerits

- A discovered and provided operation (service) fits our needs for a specific task for some data.

CONCLUSION

Today, most enterprises are actively collecting and storing large databases. Many of them have recognized the potential value of these data as an information source for making business decisions. The dramatically increasing demand for better decision support is answered by an extending availability of knowledge discovery and data mining products, in the form of research prototypes developed at various universities as well as software products from commercial vendors. Decision tree algorithm is used when the user doesn’t know the in depth knowledge of the domain. SVM is used in the context of minimum execution time. Naïve bayes is used if the application follows independent feature model. The goal of Classification algorithms is to produce precise and accurate results.

REFERENCES

[1]. A. Adamov, "Data mining and analysis in depth. case study of Qafqaz University HTTP server log analysis," 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 2014, pp. 1-4, doi: 10.1109/ICAICT.2014.7035947.

[2]. S. P. Singh and Meenu, "Analysis of web site using web log expert tool based on web data mining," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/ICIIECS.2017.8275961.

[3]. J. Chen et al., "Data Analysis and Knowledge Discovery in Web Recruitment—Based on Big Data Related Jobs," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2019, pp. 142-146, doi: 10.1109/MLBDBI48998.2019.00033.

[4]. S. Zhou, X. Zhang, X. Li, G. Zheng and G. Zhang, "Design and Implementation of Data Mining and Analysis Platform Based on Web Service Techniques," 2017

International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, 2017, pp. 545-549, doi: 10.1109/ICCTEC.2017.00124.

[5]. W. Hidayat and A. Yaqin, "Business Trends Based on News Portal Websites for Analysis of Big Data Using K-Means Clustering," 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 445-450, doi: 10.1109/ICOIACT46704.2019.8938413.

[6]. M. Driss, A. Aljehani, W. Boulila, H. Ghandorh and M. Al-Sarem, "Servicing Your Requirements: An FCA and RCA-Driven Approach for Semantic Web Services Composition," in IEEE Access, vol. 8, pp. 59326-59339, 2020, doi: 10.1109/ACCESS.2020.2982592.

[7]. S. Arora and N. Baliyan, "Extraction and Analysis of Information in News Domain Using Semantic Web," 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, 2019, pp. 1-6, doi: 10.1109/IoT-SIU.2019.8777502.

[8]. B. Vijaya and P. Gharpure, "Candidate Generation for Instance Matching on Semantic Web," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862131.

[9]. X. Chen, C. Tian and T. Wu, "The Semantic Web Approach for the Collaborative Analysis and Visualization of Ethnic Education and Vocation," 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), Nanjing, China, 2018, pp. 414-419, doi: 10.1109/CSCWD.2018.8465182.

[10]. J. Fabra, M. J. Ibáñez, P. álvarez and J. Ezpeleta, "Behavioral Analysis of Scientific Workflows with Semantic Information," in IEEE Access, vol. 6, pp. 66030-66046, 2018, doi: 10.1109/ACCESS.2018.2878043.