International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : 10.32628/CSEIT11952247

Web Page Classification Using Sentiment Analysis and Natural Language Processing

Dr. Madhusudanan J[1], Shivcharan B[2], Jagan R[2], Gladson Solomon S[2]

[1]Professor, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry, Tamil Nadu, India

[2]B. Tech, Scholar, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry, Tamil Nadu, India

ABSTRACT

During past 10 years, the technology has improved really great in the field of web Technology and integration of a web pages with each other to form a websites has improved far way when compared to 19th century. Plagiarism plays a major role in contents of web page recently. Illegal and addictive content contaminate the quality of websites. This is becoming a major issue and must be solved is an efficient manner. So we came up with an idea of introducing classifications in web pages based on credit points that specifies the quality of website based on its CONTENT and QUALITY. Web mining under Data mining with help of Search Engine Optimization provides solution to this problem. Web content mining allows us to use the techniques of information extraction, integration, schema matching, knowledge synthesis and segmentation of web pages which in turn allows the content of web pages to be processed and provide a result based analysis.

Keywords : Classification, Web Content Mining, Quality, Schema matching, Segmentation, Knowledge synthesis.

I. INTRODUCTION

The ultimate goal of this paper is to identify the optimal way to classify the web pages based on the content and quality provided by them and to address the drawback of existing approaches and determining ways that can be implemented now practically by applying new methodologies. This idea comes under web personalization which can be defined as an action that obtains the information or services provided by a website to need of particular set of users that helps a user of website to determine the quality of content of a web page which is to be viewed or processed.

II. LITERATURE SURVEY

[1] With large amount of raw data and information available online the World Wide Web is a correct place for application of mining process with help of data to extract and obtain the information required in a efficient and optimal way. Each and every day the data in web in form of documents, texts and videos are increasing in count and information storage and retrieval becomes a huge task to be performed as the exact data to be found lies between several databases and time requirement for extraction is huge and is not precise. To overcome the difficulties in processing data from a webpage especially within the area of artificial intelligence, sub-areas of Natural Language Processing can be used to perform web content

mining for further determination of content quality through pattern evaluation.NLP is used in wide range of Speech recognition and Text Processing. Different text processing algorithms are discussed in paper along with the development in past decade. NLP is a subset of Machine Learning which comes under AI and moreover so important in future as it helps to process information in form of voice or text and manipulate them as per the algorithm in computer.

[2] This paper helps to analyze the personalization factor that helps the classification of web pages much easier rather than moving through complex algorithms. Personalization is similar to an information filtering system that allows to predict the mindset and behaviour of user through a model built through characteristics of search. It automatically extracts useful information from web and creates personalized results through content mining. Data Pre processing further helps in segregation of web page content and allows the usage of a Language Processor to evaluate using text processing and determine the nature of text.

[3] In web data mining process the hyperlink structure of web or web log data can be efficiently used in mining process. Since web data is huge, diverse and dynamic users usually encounter the problems in finding relevant information, creating new knowledge out of information available, information personalization but these issues can be handled using structured mining and usage mining. The subtasks in mining categorized according to the paper is categorized into a 5 step process of Resource discovery, Information selection and Pre-Processing, Generalization, Analysis and Visualization. Methods including Visual content description, Relevance feedback algorithm, Graph based overlapping, Cluster algorithm and SPARSE technique can be efficiently implemented in content mining process.

[4] Web page classification is essential in many web information retrieval tasks for maintaining directories in web and focused crawling. When compared to traditional text classification uncontrolled nature of web content provides additional challenges to webpage classification .Web specific features and algorithms can be applied to perform text classification and is stored in containers for further processing. Traditional text classification is usually performed on Structured corpora with controlled authoring styles whereas web classification do not have any particular style and property. Web page content can be observed and analyzed easily with the help of META tags of that particular web page which in turn allows to store the contents in a specific memory variable or location. Efforts are made to tweak traditional algorithms like k-Nearest Neighbor and Support Vector Machine in web classification context. Web page content processing and Dataset Selection and Generation are main issues in classification. It can be handled efficiently by preserving tags in case of preprocessing and usage of Labeled Training Instances in case of dataset selection and generation.

[5] Web page classification is much more difficult than pure text classification due to large variety of noisy information embedded in webpage. Based on web summarization, a new web page classification algorithm in this paper to improve the accuracy rate. In traditional methods web page classification can be directly done with help of machine learning literature for text classification. Using web-page summarization techniques for preprocessing in web page classification is a viable and effective technique when compared to pure text classification. In adapted Luhn summarization method every sentence is assigned with a significance factor and sentence with highest significance factor is selected for further summarization. To compute significance factor of a a sentence, we need to build a significant words pool. In order to customize the Procedure modification is done in Luhn's algorithm. Latent Semantic Analysis and Content Body Identification by Page Layout Analysis are also effective in web classification for text classification.

[6] As the data size grows day by day, internet grows at a very huge rate and it requires an effective data analysis method for web page classification. This paper utilizes the need for more effective method of Ant Colony Algorithm in field of web content classification and shows it's effectiveness when compared to C5.0 algorithm for decision trees. Ant miner algorithm is a powerful classification tool and produces accurate results when compared to c5.0 algorithm. As the content in web is dynamic in nature it is possible to adapt ant miner algorithm for continuous learning applications. It is capable of tolerating large amount of attributes in web mining field. If Ant miner algorithm copes with the attributes it can tolerate it would be more applicable to real world web mining applications.

III. METHODS AND MATERIAL

Arduino

IV. RESULTS AND DISCUSSION

COMPARISION TABLE

Author	Year	Approach	Description
Aditya Jain, et. al.,	2018	Text classificati on using Natural Language Processing.	Helps to build models and process information by text or voice and manipulate with system.
S.Jagan et. al.,	2015	Use of Personaliza tion techniques along with data processing for	Web personalization for analyzing behaviour of user in web along with processing of
Brijendra singh et. al.,	2011	Information	Implementation of Cluster
Xiaoguang Qi et. al.,	2007	Web specific features and	preserve tags in case of
Duo Shen et. al.,	2004	Web page Summariza tion Algorithm implement	Significance factor is assigned and summarized.
Nicholas Holden et. al.,	2004	Ant Colony Algorithm	Implementation of this

V. CONCLUSION

From this paper we infer that Web content mining along with Sentiment Analysis allows us to identify the context of the stored text, determine its relationship with other texts with help of Natural Language Processing and categorize it. This will further helps to identify the quality of content in webpage which helps in providing a good point of view. Behaviour of user can also be analysed and used for Personalization purpose and credit point ratings for determination of quality of a website will be implemented in future for ensuring great user experience.

VI. REFERENCES

[1]. Aditya Jain,Gandhar Kulkarni,Vraj Shah,"Natural Language Processing" in International Journal of Computer Science and Engineering, Volume 6 Issue I, January 2018.

[2]. Jagan.S,DR.S.P.RajaGopalan ,"A survey On Web Personalization using Web Usage Mining"in International Research Journal of Engineering,Volume 2 Issue I, March 2015.

[3]. Brijendra singh,Hemanth Kumar Singh,"Web data mining research" in International Conference

[4]. Petar Ristoski, Heiko Paulheim ,"Semantic Web in Data Mining and Knowledge Discovery" in Journal Of Web Semantics,November 2015.

[5]. Tarannum Shaila Zaman,Navisah Islam,Choudhry Farman Ahmed,"Single Pass

Approach For Web Content Sequential Pattern Mining" in GSTF Journal On Computing, January 2018.

[6]. Swapnil S. Patil, Hridaynath P. Khandagale,"Enhancing Web Navigation Usability Using Web Usage Mining Techniques" in International Research Journal Of Engineering And Technology(IRJET),June 2016.

[7]. V. A. Chakkarwar,Amruta A.Joshi,"Semantic Web Mining Using RDF data" in International Journal Of Computer Applications,Volume 133, No 10 ,January 2016.

[8]. N. A. Mohoto,A.Memon,A.Teevno,"Extraction Of Web Navigation Patterns by Sequential Pattern Mining" in Sindh University Research Journal,Volume 48 January 2016.

[9]. Dirk Burghardt, Moritz Neun, and Robert Weibel,"Generalization Services on the Web Classification and an Initial Prototype Implementation" in Cartography and Geographic Information Science, Vol. 32, No. 4, 2005, pp.257-268.

Cite this article as :

Dr. Madhusudanan J, Shivcharan B, Jagan R, Gladson Solomon S, "Web Page Classification Using Sentiment Analysis and Natural Language Processing ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1064-1067, March-April 2019.
Journal URL : http://ijsrcseit.com/CSEIT11952247