International Journal of Advanced and Innovative Research (2278-844)/

Volume 7 Issue 11

Understand Short Texts by Harvesting and Analysis Semantic Knowledge

[1]Raheema Sheik, [2]K. Laxmaiah

[1]Research Scholar, [2]Assistant Professor, Department of Computer Science & Engineering

KODADA Institute of Technology & Science For Women, Kodad

Abstract: Semantic web mining and Synaptic web entropy mining is an important and recent research area today where number of technique presented in order to mine the web crawl efficiently and to find the web page rank of various data available in the web, in the present paper which is taken by us for the further research is the hybrid approach where the entropy is calculated based on the semantic-synaptic based approach and the important role of the entropy required today to monitor the todays web fluctuation and various stages of multiple portals and web data available today , here as the authors of the paper mentioned about the scope of the semantic monitoring ,we assume it to be a great way to make it better to experience and query search for the extracted data and opinion monitoring, we have monitored a paper which works with the short text analysis, challenges associated with it and finding data relevance in available knowledge base. Using the spatial data and related information with the input dataset. It can able to find more specific data about the user input. We here mentioning and further trying to make user experience better to utilize the data extracted from the user and visualize them efficiently as per the user requirement. ANN approach can further be utilized for efficient parameter computation.

Keywords:-Semantic Analysis; Segmentation; Datasets; Ontology; Hybrid Algorithm.

1. INTRODUCTION

In marketing and advertising domains Opinion Mining is being larger domain. Advertiser needs to analyze performance/ popularity of ads that he/she posted on site. Star rating based mechanism may go fraud, because of robots or automatic responders. So, current system needs to be analyzed using comments & natural language processing. Fraud comments could be removed by using irrelevant comment removal mechanism suggested in paper. In this paper the role and importance of social networks as preferred environments for opinion mining and sentiment analysis are discussed especially. In this paper, selected properties of social networks that are relevant with respect to opinion mining are briefly described and outline the general relationships between the two disciplines. It presents the related work and provide basic definitions used in opinion mining area. Then, the original method of opinion classification is introduce and test the presented algorithm on real world datasets acquired from popular Polish social networks, reporting on the results. The results are outperform and soundly support the main issue of the paper, that social networks exhibit properties that make them very suitable for opinion mining activities.

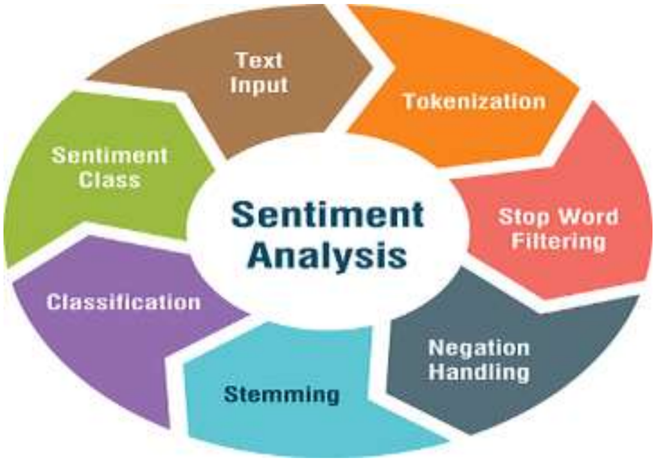

Figure 1.1: Semantic Knowledge Extraction And Learning Process.

In the figure 1.1 above, a knowledge extraction process from the large dataset by using semantic rules and semantic data annotation is performed.

Usually tags which are semantically related to the terms are used not semantically similar. Consider “plastic surgery” tag as an example, some of the terms with this tag are: surgery, body, arm, health, beauty. Where they are not semantically similar but are related. Semantic similarity algorithms usually takes a shortest path method on a IS A like graph, in order to calculate semantic similarity while semantic relatedness algorithms uses a graph with Has Part, Kind of, and Opposite edges. This is why First Stage (as semantic relatedness algorithm) has better results than other semantic similarity algorithms.

Twitter as a micro blogging system, allows users to share posts each containing maximum of 140 characters, known as tweets. Each tweet is enriched with content-based and context based tags.

2. LITERATURE REVIEW

1. Wen Hua, Zhongyuan Wang, Haixun Wang

In this paper an algorithm to determine short text using semantic knowledge is discussed. Here two modes of detection which is offline and online mode is provided by the author. The given processes first take the input from the user and then process it first by text segmentation process. The segmentation process creates the different segment of values. Further term building using the segmented value and tag generation from the value is performed. Then based on term understanding maximum clique is determined. Single chain and pair is detected so that data strength can be taken for processing. Weight detection is performed over the large data understanding and thus value output is generated. MaxCMC and CMaxC both the algorithms were used for computation. Twitter dataset is used for processing and further computation cost, precision is computed for the analysis purpose. A high precision is shown for the computation with data analysis pair wise and chain model [1].

2. Mir Saman Tajbakhsh, Jamshid Bagherzadeh, 2016

This paper work towards the TF-IDF approach which work with similarity measure algorithm with dataset. This approach work with similarity recommendation approach. Data text determination, computation of relation in between the algorithm given words such as #frd and #friend can be computed is solved in this paper. The algorithm computes with high accuracy, precision and better recall over previous IDF approach. A similarity measure score is computed and weight determination to solve the given issue. This paper lacks in processing with large number of data and noise removal entity [2, 3].

3. Godoy, D., Rodriguez, G., And Scavuzzo, F., 2014

In this work, Case-Based Reasoning (CBR) techniques for the data analysis is performed by author. In this research article author describes how jcolibri can serve to that goal. jcolibri is an object-oriented framework in Java for building CBR systems that greatly benefits from the reuse of previously developed CBR systems. The program analysis is given which work towards the user profile and processing. A Tag based processing, annotation data created and processing is performed for the input document. A similar resource finding technique based on the tag history, tag understanding is driven in paper. Semantic similarity pair score is generated which helps in computing [4, 5].

4. Bart P. Knijnenburg, Martijn C. Willemsen, Alfred Kobsa 2011

In this work, author works towards the data interaction and its behavior. Various component such as subjective system aspects, user experience, interaction and data detail consumption is performed by the author. Objective system aspects module process the algorithm which generate the proper recommendation for process. Feedback generation and its understanding using the text is performed by the system. It understands the meaning behind the provided feedback and overall rating over it. A local data generation and entity analysis performance over it driven. The limitation of their work is they performed observation over limited data and working with large data is left for the future processing [6,7].

5. Rishabh Upadhyay, Akihiro Fujii

In this paper [8] approach is performed with semantic algorithm and natural language processing hybrid approach is applied. Knowledge extraction from the various pdf file is extracted by them using itextpdf API. Further data extraction and word extraction from the pre-processed pdf text data is performed by them. A triple score is applied on the mining data obtained. A line triple score and its architecture generation is the main key concept of finding data statistics. Further an inference rule and public data optimization is used for any of the obtained data. A structure mining and semantic usage of the data mining is taken from the used dataset. A row of discourse element and data example keywords are extracted from the available dataset row [10].

6. *Gautam R. Raithatha*

In this paper [11] ontology and web ontology relation generation concept is taken. An ontology concept is the representation of entity in any of the semantic data, also it represent the relation between any of the data presented. It is the concept of specialization where the large data unit and processing row is presented. Ontology can get understand by the machine and human as well. There is a process which is extraction as syntactic extraction, further a semantic extraction and finally ontological operation extraction. Further an output as in the form of xml is extracted from the ontology processing result set [12].

3. PROBLEM IDENTIFICATION

Today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

Working with the short text and finding its proper meaning and usage is one of the important task objectives for the work [9].

Finding Relevant Information

People either browse or use the search service when they want to find specific information on the web. Today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

Creating New Knowledge Out Of The Information Available On The Web

This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already has collection of web data and extract potentially useful knowledge out of it [10].

Personalization Of Information

When people interact with the web they differ in the contents and presentations they prefer.

Learning About Consumers Or Individual Users

This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing.

Finding Or Analysing The Large Data

Large Amount of the data is unable to monitor and optimize according to the user requirement, so here the requirement is to find the best way to analyse it efficiently.

4. PROPOSED METHODOLOGY

Here we briefly describe a technique to discovered frequent item pattern and hybrid approach with make use of Line-Up Approach in proposed system.

A. Symantec Mining [6]

A Web mining from the crawl is done first,we are extracting the information from the web based on the similar type of object and their availability in semantic manner, the data is been extracted and use to create Entropy.

B. Synaptic Mining

In this algorithm, the patterns are categorized according to the length executed on lattice model. Patterns will form a lattice based on the pattern-length and pattern-frequency. And using this lattice, frequent patterns are searched depth first.

Lattice Construction: The basic element of the lattice is an atom i.e. single page. Each atom or page stands for length-1 prefix equivalence class. Beginning from bottom elements the frequency of upper elements with length n can be calculated by using two n-1 length patterns belonging to the same class.

C. Applying Proper Vocabulary Mined Data

We are applying now the visualize and ANN technique where the things can be substitute in various dataset and the result observed from the various semantic data and user can optimize according to the Visualize and observation required.

5. RESULT ANALYSIS

In this section, different observed result which is performed is presented. A statically analysis and graphical analysis using the existing as well as proposed technique is presented.

Experimental Setup

All experiment execution is performed on i3 Machine, Windows 10 Operating System with 750 GB HDD and 4 GB of RAM. Apache server foundation is used for running the application and WAMP server is used for data storage. A processing is performed with two different web service dataset.

The screen demonstrates an experimental framework derived for the existing Location based approach and proposed BDMFA (Big data matrix factorization ANN approach Algorithm) for web service recommendation generation from the large dataset of web service recommendation.

Table 1.1: Comparative Analysis Between Existing And Proposed Algorithm.

DATASET	ALGORITHM SYSTEM	ACCURACY %	PRECISION %	RECALL%	MAE
WSDREAM-Dataset 1	Location & Region based approach	78.4	87.5	81.10	63.49
	BDMFA	82.90	88.43	83	67.23
WSDREAM-Dataset 2	Location & Region based approach	84.3	78.23	81.68	65.78
	BDMFA	86.70	82	84.90	67.0

As per the statistical result in table 5.1, further a comparison is made individually using graph by which a proper monitoring and observation can be made.

Table 1.2: Comparative Analysis Computation Parameter.

DATASET	ALGORITHM SYSTEM	COMPUTATION TIME	COMPUTATION COST
WSDREAM- Dataset 1	Location & Region based approach	227	44.492
	BDMFA	174	34.104
WSDREAM- Dataset 2	Location & Region based approach	298	58.016
	BDMFA	213	41.748

In the table 5.2 above, other computation related to the computation time and computation costs were given.

Graphical Result Analysis

An analysis of result graphically is discussed which help in understanding the observe parameter and their graphical monitoring.

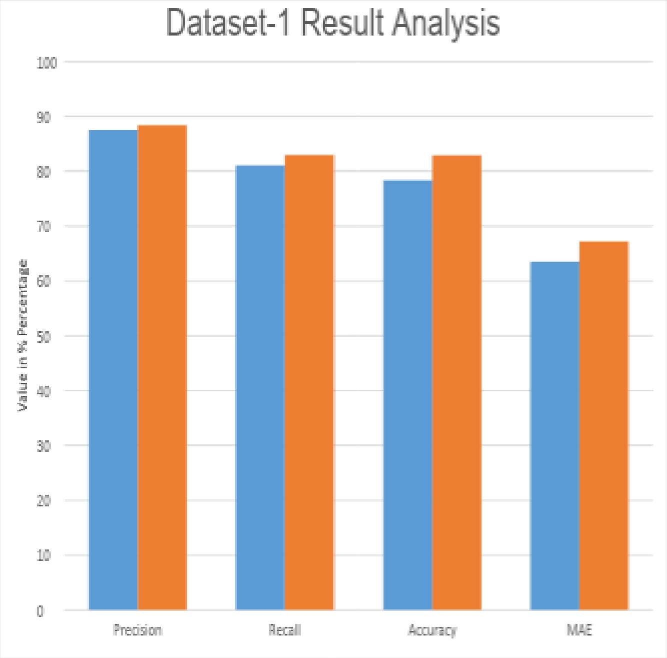

Figure 1.2: Result Analysis With The Dataset-1.

In the figure 1.2 above a graphical analysis of computational parameter, with the dataset-1 is computed.

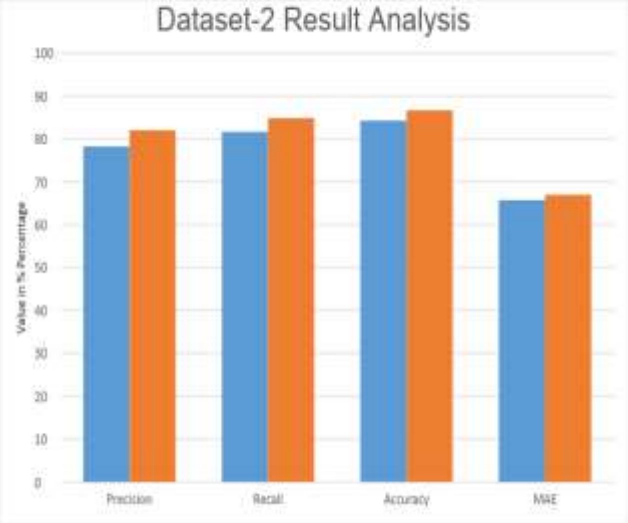

Figure 1.3: Dataset 2 Result Analysis.

In the figure 1.3 above a graphical analysis of computational parameter, with the dataset-2 is computed.

5. CONCLUSION & FUTURE WORK

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. One of the algorithms which is very simple to use and easy to implement is the Hybrid algorithm. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved Hybrid algorithm. The effective algorithm will be proposed with the improvements as well as the implementation of Hybrid Algorithm.

The forthcoming step in the research work shall be to design the improved version of the Hybrid Algorithm that shall be implemented on the synaptic mining and spatial data information.

REFERENCES

[1]. Zebang Chen, Takehiro Yamamoto, Katsumi Tanaka," Query Suggestion for Struggling Search by Struggling Flow Graph", IEEE 2016.

[2]. Huiping Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining" 2010 International Conference.

[3]. Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei "Web usage mining based on WAN users' behaviours" 2010 International Conference.

[4]. A. B. M. Rezbaul Islam; Tae-Sun Chung," An Improved Frequent Pattern Tree Based Association Rule Mining Technique",IEEE,2011.

[5]. C.P. Sumathi, r. padmajavalli,"An overview of pre-processing of web log files for web usage mining" 2011.

[6]. Hiteshwar Kumar Azad, Kumar Abhishek, "Entropy Measurement and Algorithm for Semantic-Synaptic Web Mining", IEEE 2014.

[7]. Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. Struggling or exploring?: disambiguating long search sessions. In WSDM'14, pages 53–62, 2014.

[8]. Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. Struggling and success in web search. In CIKM'15, pages 1551–1560, 2015.

[9]. Wen Hua, Zhongyuan Wang, Haixun Wang, Member, IEEE, Kai Zheng," Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE transaction, 2016.

[10]. D. Deng, G. Li, and J. Feng, "An efficient trie-based method for approximate entity extraction with edit-distance constraints," in Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, ser. ICDE '12, Washington, DC, USA, 2012, pp. 762–773.

.