Sentiment Classification Algorithm of Web Data

Mr. Sanjay Singh Bhadoria **Dr. Dhanraj Verma**

1. *Research Scholar, Dr. A.P.J. Abdul Kalam University, Indore*
Email Id.: sanjay.bhadoria@gmail.com
2. *Professor, Dr.A.P.J. Abdul Kalam University, Indore*
Email Id.: dhanrajmtech@gmail.com

Abstract:-

Semantic could be considered as a magical term to bridge the gap in the diversity of data. Semantics can be used in meaningful data integration, decision system which makes it possible to detect inconsistency of data, discovers new (hidden) knowledge, etc. Semantic analysis presents the data in a more efficient manner and makes it useful as a source for knowledge discovery and comprehension. Se- mantic analysis is method to better understand the implied or practical meaning of the input dataset. It is mostly applied with ontology to analyze content mainly in web resources. This field of research combines text analysis and Semantic Web technologies. The use semantic knowledge is to aid sentiment analysis of queries like emotion mining, popularity analysis, recommendation systems, user profiling, etc. Different sources with information can be linked through semantic annotations. Some of the application areas of sentiment analysis are product/services research for marketing purposes, better search engines, trend monitoring (e.g. so- cial, cultural, political etc.), recommendation systems etc.

Key words: *Educational Datamining (EDM), Learning Analytics (LA), Technology, Enhanced learning (TEL), Inquiry based learning (IBL)*

Introduction

The number of data produced across the world is increasing and could keep on growing at an accelerating rate for years to come. At organizations total businesses, servers have been overflowing with usage logs, transaction records, message flows and business operations records, detector data along with cell device data. Efficiently analyzing these massive collections of data, big data since it's often known, can make considerable value-creating trong interest in pros having the capability to perform such analysis. Effective big data analysis requires skills in a range of computer science areas such as data processing and storage, statistical data analysis and computational linguistics, and also the skill to combine this specific knowledge in novel ways.

Given such large sizes of text data sets, mining programs, that arrange the writing data sets into structured knowledge, will enhance efficient document access. This eases semantic search and at precisely the same period, provides an efficient platform for classification as data becomes extremely huge. Semantic analysis can be a process to better understand the implied or practical significance of the

input data set. It is largely applied with ontology to analyze content chiefly from web resources. Ontology fitting is a remedy to the semantic heterogeneity problem. It finds correspondences between semantically related entities of ontology. All these correspondences can be employed for various tasks, such as for example ontology blending, query answering, or statistics translation. More conventional machine learning algorithms aren't efficient enough to extract the semantic information (hidden) generally exhibited in big data. By extracting such attributes, semantic analysis empowers using classification, forecast, which is essential when developing models to manage the scale of data that is big. The telltale knowledge is used to aid sentiment investigation of queries like emotion mining, popularity analysis, recommendation systems, user profiling, etc.. Separate sources together with advice could be linked through semantic annotations. Some of these application areas of sentiment analysis are product/services research for promotion purposes, better searchengines, trend tracking (e.g. societal, cultural, political, etc.), recommendation systems, etc..)

Objective

The objective of this research work is semantic analysis of big data with an emphasis on sentiment analysis from web-based data. We have divided the major tasks for this dissertation as follows:

a) To study critically on semantic analysis of big data and its recent development.

b) To propose a new hybrid distributed rule mining method for semantic analysis of big data.

c) To propose a new distributed semantic text summarization method for big data analysis.

d) To propose a novel sentiment analysis approach of web-based big data.

e) To analysis the proposed methods.

Problem Formulation

The main focus on big data analysis is to efficiently process a very large dataset. There is an equally hard problem of how to analyze, integrate, and transform the data from many sources. Knowledge can be represented from big data using large-scale analysis and visualization tools. Our research philosophy is to build a model designed for large scale operations that expose extraction of semantic information for big data. Using this philosophy as the cornerstone, our research work has spanned the broad area of semantic analysis of big data of both theoretical and practical system-oriented research. In this work, we have considered two characteristics of big data i.e. volume and velocity. We have considered big data analysis in terms of large in size and dynamic in nature. Main goal of this work is to design a new method to analyze big data as it can not be processed with traditional methods. Semantic representation of an input data gives inner knowledge which is very useful in different applications of the data. Here, we have tried to handle big data with semantic analysis. Also, we have considered another task for this research as sentiment analysis from web-based data.

Contribution in the Research

In this research work, we have provided a systematically study and solution of the problems that traditional methods face when these are applied for big data analysis. We have used association rule mining, information clustering and domain ontology to represent semantic analysis of big data. This thesis offers an innovative analysis and proposed methodological approaches for semantic analysis and an efficient sentiment analysis from web-based data using semi-supervised classification approach.
A brief about contributions of the thesis are as follow.

a) Different research issues and challenges in semantic analysis of big data as well as sentiment analysis.

b) Implementation of a smart web crawler to collect domain specific web con- tents.

c) Design of a new distributed evolutionary rule mining method for semantic analysis of big data.

d) Design of a new semantic text summarization technique for dynamic large data.

e) Design of a novel semi-supervised approach for sentiment analysis of large web-based data.

Data Collection and Processing

The first step of the proposed method is to make collection of different web reviews datasets. Reviews datasets which are already available in some data repository, can be downloaded. Also, these can be collected directly from different web sources. Here, a focus web crawler can be designed to collect related reviews datasets which are available in web. Next, these raw datasets should be preprocessed as these have some irrelevant data. Data preprocessing is an important process as it reduces the number of words to be processed by eliminating unnecessary words.
HTML Preprocessing: For each web reviews article, irrelevant HTML tags are removed to get the exact actual information since these articles are downloaded from web in HTML format.
Stopwords Removal : Stopwords should be removed from reviews articles before sentiment analysis.

Creation of Sentiment Terms Matrix

A sentiment term matrix or term-sentiment matrix of a reviews datasets is a mathematical matrix that describes the frequency of sentiment terms that occur in the reviews datasets. In a sentiment term matrix, rows correspond to reviews documents in the datasets and columns correspond to sentiment terms and each element of the matrix contains the numerical value (strength) of the sentiment terms in the review document. Here, lists of sentiment words are considered from *http://ptrckprry.com*. Two lists are considered for positive sentiment word list and negative sentiment word list.

Number of sentiment terms (columns) in the sentiment terms matrix is considered as 2. First column is considered for positive sentiment terms and next column is considered for negative sentiment terms. When a term in the dataset is present in the positive sentiment word list then, value 1 is added for the

element corresponding with the dataset and positive sentiment of the matrix. Similarly, when a term in the dataset is present in the negative sentiment word list then, value 1 is added for the element corresponding to the dataset and negative sentiment of the matrix. This technique is followed for all sentiment terms in the processed dataset.

When *not* is present before a negative sentiment term in the dataset, value 1 is added for the element corresponding to the dataset and positive sentiment of the matrix. Similarly, when *not* is present before a positive sentiment term in the dataset, value 1 is added for the element corresponding to the dataset and negative sentiment of the matrix. Intensity of sentiment terms plays a vital role in sentiment analysis. Extra 1 value is added for sentiment terms where an intensifier is present before the sentiment term in the processed dataset.
For example:

D1 = *The camera is awesome and sound quality is too good.*
D2 = *It is a bad product.*
D3 = *The picture quality is good but battery performance is very poor.*

Table 5.1 represents the sentiment term matrix for documents D1, D2 and D3.

Dataset	Positive Sentiment	Negative Sentiment
D1	3	0
D2	0	1
D3	1	2

Table 5.1: An Example of Sentiment Term Matrix for Datasets

Sentiment Classification

Here, we have considered a large amount of web reviews datasets. Similarly, the size of sentiment terms matrix which is represented the large datasets is also large. Traditional classification techniques fail to handle these large datasets. In this technique, firstly the large input large dataset is divided into several clusters using fuzzy c-means clustering algorithm which is discussed in chapter 2. This clustering technique is implemented in MapReduce framework which provide dis- tributed parallel processing environment to efficiently handle large datasets. Then, each cluster is assigned into any of three classes (positive, negative and neutral) by applying Gaussian naive Bayes classifier which is discussed in chapter 2.

A. ***Distributed Clustering Method for Large Datasets***: The sentiment terms matrix is divided into several clusters using fuzzy c-means clustering (FCM) algorithm. Fuzzy c-means clustering is suitable for reviews datasets as it gives the best result for overlapped datasets and comparatively bet- ter then the k-means algorithm. To handle these large datasets, we adopt Hadoop MapReduce programming model. Since, datasets are converted into terms matrix with numerical values, MapReduce performs efficiently. The

FCM iteration is decomposed into two stages of *Map* and *Reduce*. On Map nodes, a function can effect different datasets in the different data node

B. ***Classification Technique for Numerical Dataset*** : Finally, we have to assign proper sentiment to each dataset. After applying the proposed clustering technique, a set of final cluster centers is generated along with clusters of datasets. Basically, cluster center represents all elements in a cluster. In this work, cluster centers are assigned classes using Gaussian naive Bayes classifier which is discussed in Algorithm 2.1. Before applying the classifier, training dataset is preprocessed and converted into a sentiment terms matrix (training matrix). Finally, datasets are classified with classes which are assigned to corresponding cluster centers.

Algorithm 5.1: Proposed Algorithm of FCM in MapReduce for Sentiment

---

Analysis

Input :

(a)Sentiment Terms Matrix. (b)Number of clusters (*n*).
(c)Threshold value, *e*, for termination criterion.

Output:

(a) Set of clusters, C = $\{c_1,...,c_n\}$

(b) Membership Matrix.

1 Copy the term matrix data to the distributed file system (HDFS).

2 Select randomly *n* numbers of clustering centers.

3 do

4 Generate fuzzy membership matrix.

5 Produce the key-value pairs of (document, sentiment), in which the document is the matrix's row number and sentiment contains the value of the document.

6 Divide all the key-value pairs into several data nodes, and transfer to the Map function to calculate membership degree according to the Equation 2.7.

7 Store the result in the intermediate key-value pair of (k, μ_i) where, k is clustering number, and μ_i is membership degree to the i^{th} clustering.

8 Merge the intermediate key-value according to the clustering number in Reduce function using calculation result of the Map function and obtain a new clustering center by calculating according Equation 2.8.

9 Transfer these data to the data nodes participating in the MapReduce according to the clustering center.

10 while *membership matrix does not satisfy the termination criterion.*

11 Export clustering results including the clustering number, the clustering centers and the final membership matrix.

LTS and Hadoop 2.2.0. The experiment is conducted using on one name node as master node and 10 data nodes (Map) as slave nodes. The input database is distributed to each Map node from HDFS. Before evaluation of the tweets, we need to pre-process in order to remove the noise from the data and create sentiment terms matrix. We have conducted two experiments.

In the first experiment, training dataset and test dataset are used which is freely available in *http://help.sentiment140.com/for-students*. In the second experiment, we have collected web reviews dataset from Twitter directly.

Experimental Results

We have tested our classifier on a set of real Twitter posts. For the first experiment, we have used the first 500 records of the training set [79] as actual training dataset and next 500 records are considered for test dataset. The characteristics of the training dataset are presented.

Sentiment	No. of Records in Training Set
Positive	219
Negative	198
Neutral	83
Total	500

Table 5.2: Characteristic of the Considered Training Datasets

We compute accuracy of the classifier on the whole evaluation dataset where, M is the number of correct classifications, and N is the number of all classifications.

Table 5.3 shows the performance of our method which has better efficiency than other methods.

Sentiment Classification Method	Accuracy
FLSPR [68]	83.62%
UFCTS [69]	76.52%
STCSVM [73]	80.47%
Proposed Method	88.18%

Table 5.3: Results with other methods using Sentiment140 Twitter dataset

Conclusion

In this, we proposed a novel sentiment classification method for large web data. Traditional classification methods are not suitable for large datasets. In this method, large datasets are divided into several clusters. After that, datasets are assigned classes by assigning class for each corresponding cluster centers. Here, we have applied MapReduce framework to divide large datasets into clusters efficiently. That is why, the proposed method becomes more effective and efficient in term of processing time. Results are provided to establish the proposed efficient sentiment analysis method.

Nowadays big data analysis plays a vital role in different applications. Big data can not be analyzed with traditional database system due its characteristics like volume, velocity and variety. The purpose of this research work is to establish an efficient mechanism to extract data from the web and discover rich information for a decision system through semantic analysis of big data. This thesis offers an innovative analysis and methodological approaches for semantic analysis of big data as well as for sentiment analysis from web-based big data. Here, we have considered large volume and dynamic data for semantic analysis. In this thesis, we have proposed two methods for semantic analysis of big data and a method for sentiment analysis from web-based data. Hadoop MapReduce framework provides distributed parallel processing environment where big data can be processed efficiently. This framework is suitable for handling with large structured and homogeneous datasets. In the first proposed method, semantic rules are identified from large structure dataset using evolutionary rule mining in MapReduce platform. But it is not suitable for unstructured dataset like text dataset. In the second proposed method, we have developed a semantic model for large and dynamic text data analysis using knowledge representation. Finally, we have developed a method on sentiment analysis from web-based data as it can be an excellent source of information and can provide insights. Thesis is con- cluded by comparing possible parameters for combining different methods as well as implementation issues and limitations in the context of big data.

Future Work

The research work discussed in this thesis can be extended in many directions. Here, we provide some directions which can be considered as future work.

- This work can be extended to provide suitable data fusion method to handle big data where different heterogeneous sources are considered. Input dataset with different structures can be mapped with semantics into a single format and then desired knowledge can be retrieved easily with some query management system.

- Information about anything is available in the web. Nowadays, one of the major challenges is to collect (semantically) relevant information from huge complex data sources efficiently. So, issues associated with big web data analysis in semantic clustering, semantic search, etc. can be a future direction in this research area.

- In sentiment analysis method, natural language processing (NLP) can be utilized to tackle natural language terms.

Reference

[1] M. El-Hajj and O. R. Za¨ıane, "Parallel bifold: Large-scale parallel pattern mining with constraints," Distributed and Parallel Databases, vol. 20, no. 3, pp. 225–243, 2006.

[2] C. Pitangui and G. Zaverucha, "Genetic based machine learning: merging pittsburgh and michigan, an implicit feature selection mechanism and a new crossover operator," in Hybrid Intelligent Systems, 2006. HIS'06. Sixth Inter- national Conference on, pp. 58–58, IEEE, 2006.

[3] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic accep- tance," Physica A: Statistical Mechanics and its Applications, vol. 391, no. 6, pp. 2193–2196, 2012.

[4] C. Saranyamol and L. Sindhu, "A survey on automatic text summarization," Int. J. Comput. Sci. Inf. Technol, vol. 5, no. 6, pp. 7889–7893, 2014.

[5] A. R. Pal and D. Saha, "An approach to automatic text summarization using wordnet," in Advance Computing Conference (IACC), 2014 IEEE Interna- tional, pp. 1169–1173, IEEE, 2014.

[6] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "Knn based machine learning approach for text and document mining," International Journal of Database Theory and Application, vol. 7, no. 1, pp. 61–70, 2014.

[7] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de França Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment," Expert Systems with Applications, vol. 41, no. 13, pp. 5780–5787, 2014.

[8] S. T. Babekr, K. M. Fouad, and N. Arshad, "Personalized semantic retrieval and summarization of web based documents," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no. 1, pp. 177–86, 2013.

[9] A. S. Corrêa, C. Borba, D. L. da Silva, and P. Corrêa, "A fuzzy ontology- driven approach to semantic interoperability in e-government big data," In- ternational Journal of Social Science and Humanity, vol. 5, no. 2, p. 178, 2015.

[10] R. Ragunath and N. Sivaranjani, "Ontology based text document summariza- tion system using concept terms," ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 6, p. 2638, 2015.