International Journal of Advanced and Innovative Research (2278-7844) /

Volume 7 Issue 4

A Discovery Knowledge Extraction Hybrid approach using the Synaptic data discovery Model

Surbhi Tiwari#1, Asst. Prof. Nitesh Gupta

#NRI College, CS Department, RGPV University
India

1surbhitiwari01@gmail.com

***Abstract*--** **Web Mining is an extraction of knowledge from the web data. A number of data get generate while working with web usage. An analysis of such data and finding the usable entity to provide a better user experience can be a advantage of algorithms. Thus an knowledge discovery and providing a quick solutions to the input query can be performed. Many approaches for the data analysis, weight analysis and data processing is performed by previous author. TF-IDF, Semantic, FP growth algorithm and such other techniques are used by previous research for knowledge analysis. In this paper, an advance synaptic data discovery model for the web data extraction and analysis is performed. The proposed algorithm work with the tree architecture based discovery and enable finding the relevant terminology. Thus finding a better solution for the prediction and finding a better knowledge query output is performed. The experiment result shows the effectiveness of proposed approach over the traditional algorithm.**

***Keywords*--** **Semantic knowledge, Data extraction, Data understanding, NLP, Data pruning.**

I. INTRODUCTION

In the recent era, a large amount of raw data is being gathering day by day and storing in databases anywhere across the world, which is mainly collecting from different industry and social media sites. There is a requirement to extract and determine useful data and knowledge from such a data that is being collected. Data mining is an interdisciplinary field of computer science. It is referred to as mining knowledgeable data from large databases. It is the process of performing automated extraction and generating the predictive information from a large database. It is the process of searching the hidden information from the repositories .The fields that use Data mining techniques include medical research, marketing, telecommunication, and stock markets, health care and so on. In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [1].

Data mining consists of the different technological methods including machine learning, statistics, database system etc. The aim of the data mining process is to discover knowledge from large databases and transform into a human understandable format. Data mining with knowledge discovery are important parts to the organization due to its decision making strategy. Classification, clustering and regression are three methods of data mining. In these methods instances are grouped into identified classes. Classification is a popular task in data mining especially in knowledge discovery. It gives an intelligent decision making. Classification is not only studies and examines the existing sample data but also predicts the future behaviour of that information. It maps the data into the predefined class and groups. It is used to predict group membership for data instances.

A semantic TF-IDF based weighting method is proposed in the current paper. The vector is used for redefining semantic weights and thus the similarity of tweets. For a given tweet, T, the tags of the Top N similar tweets are recommended. The classical metrics of data mining is used for evaluating current approach. Semantic similarity and relatedness algorithms are compared and results showed significant improvement than normal TF-IDF weighting schema.

Usually tags which are semantically related to the terms are used not semantically similar. Consider "plasticsurgery" 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 HirstStOnge (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.

II. RELATED WORK

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 I text pdf 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].

In the above section, multiple literature survey algorithm are discussed. This section contains multiple author approaches which participate in data processing [13, 14].

III. PROPOSED METHODOLOGY

Hierarchical clustering provides an easily visualized way to modelling the underlying relationships among the data objects. Hierarchical clustering can be considered an agglomerative approach, which suffers from the problem of one-way construction, that is, it cannot be undone during the hierarchical tree construction procedure.

Step 1: calculate the mutual distance of two data points (distance matrix) as the clustering criteria;

Step 2: decompose the dataset into a set of levels of the nested aggregations based on the distance matrix (i.e. the tree of clusters);

Step 3: cut the hierarchical tree at the desired level by selecting a predefined threshold, and then explicitly merge

TECHNIQUES	ACCURACY	PRECISION	RECALL	MAE
Old	80.04	85.35	75.35	65.35
New	82.89	88.2	78.2	68.2

all connected subjects below the cut level to create various clusters;

Step 4: output the dendrograms and the clusters.

IV. RESULT ANALYSIS

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

Computing Parameter:

There are mainly three parameter, which is taken for the comparison analysis is taken. Computing parameter such as computation time, computation cost and bandwidth consumption is observed.

Computation Time:

Computing time is the time difference which is observed by subtracting final executing time to initial loading time. A time difference between both the times is observed and call as computation time.

Computing time = final execution time – initial time;

Ct=fet-it;

Computation Cost:

Computing cost is the total cost which can be observed by monitoring different usage resources and aspects such as bandwidth, data consumption, resources etc.

Computing cost = bandwidth consumption cost + Resources consumption cost + cost per second;

Cc= bcc + rcc+cps;

Bandwidth consumption:

It is the total data consumption per unit of time which is taken by the token and complete access monitoring.

Bandwidth consumption = total data consumption/ unit time;

Bc = tdc/ut;

STATISTICAL ANALYSIS

In this section we will explain about the several calculations performed over different algorithms.

Table 1.1 Statistical Analysis.

In the above table we have explained about the accuracy, precision, recall, mae which are calculated over different techniques.

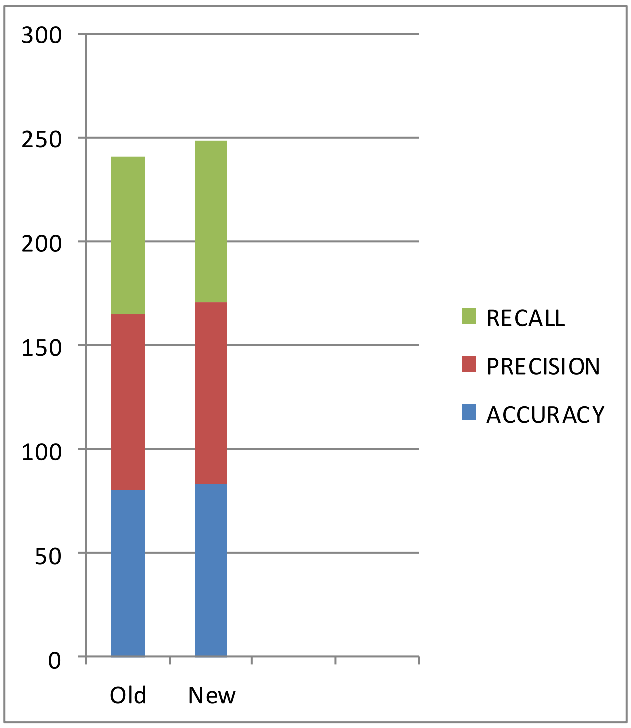

Figure 1.1 Graph for comparison in between techniques.

The above graph explains about the techniques difference on the basis of the different aspects.

V. CONCLUSION

Data discovery from the web data help in extracting the usable entity and further finding many prediction as per requirement. Web data generates many unit and sparse amount of data. In this paper an synaptic based hybrid data processing model is presented which help in web extraction and finding usable data points from variance. The algorithm proposed used a tree based architecture and use the web dataset for processing. The execution is performed on web service data and finally performing prediction on the data. Result analysis and implementation shows the effectiveness of proposed algorithm while comparing with traditional TF-IDF approach. A future work can be done in algorithm application such as health care and cloud implementation of proposed scenario.

REFERENCES

[1]. Dr. Pranav Patil,"Application for Data Mining and Web Data Mining Challenges",IJCSMC, Vol. 6, Issue. 3, March 2017, pg.39 – 44.

[2]. Manoj Kumar,Mrs Meenu,"A Survey on Pattern Discovery of Web Usage Mining", 2017, IJARIIT.

[3]. Sharma K., Shrivastava G. & Kumar V.,"Web Mining: Today and Tommorrow". In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[4]. I. Mele, "Web usage mining for enhancing search-result delivery and helping users to find interesting web content," in ACM 6th International conference on Web search and data mining, 2013, pp. 765–770.

[5]. Bhatia C.S. & Jain S.,"Semantic Web Mining: Using Ontology Learning and Grammatical Rule Interface Technique". In IEEE 2011.

[6]. R. Nayak and A. Bose, "A Data Mining Based Method for Discovery ofWeb Services and their Compositions," Real World Data Min. Appl., vol. 17, pp. 325342, 2015.

[7]. R. Geng and J. Tian, "Improving web navigation usability by comparing actual and anticipated usage," IEEE Trans. HumanMachine Syst., vol. 45, no. 1, pp. 84–94, 2015.

[8]. Ramakrishna, Gowdar ,"Web Mining: Key Accomplishments, Applications and Future Directions", in the International Conference on Data Storage and Data Engineering 2010.

[9]. Singh A., Juneja D. and Sharma A.K.,"Design of Ontology-Driven Agent based Focused Crawlers". In proceedings of 3rd International Conference on Intelligent Systems & Networks (IISN-2009),Organized by Institute of Science and Technology, Klawad, 14 -16 Feb 2009, pp. 178-181. Available online in ECONOMICS OF NETWORKS ABSTRACTS, Volume 2, No. 8: Jan 25, 2010.

[10]. Esfahani PM, Habibi J, Varaee T ,"Application of social harmony search algorithm on composite web service selection based on quality attributes". In: Sixth International Conference on Genetic and Evolutiona, 2012

[11]. Aarti Singh,"Agent Based Framework for Semantic Web Content Mining". Published in International Journal of Advancements in Technology,Vol. 3 No.2 (April 2012), ISSN 0976-4860.