A survey on: multiple approaches for knowledge extraction over semantic data

Surbhi Tiwari,Asst. Prof. Nitesh Gupta
Department of Computer Science
RGPV NRI Bhopal (M.P.),India

Abstract : Web mining and its usage over the different result, analysis over the large input is required. Web mining gets multiple techniques such as semantic, synaptic and other relevant approach. Different resource platform often generate the short text which is unable to understand by the software process and thus data usage cannot be apply over it. There are various ways of understanding such text which leads to noise removal, replication removal and processing large data over segmentation and other process. Semantic knowledge and matching content also help in understanding various given text. Different challenges such as typing New York City as “nyc” is not recognized by the standard system. In this paper our survey is performed over different approach which helps in understanding short text and processing its knowledge extraction over the large data.The paper discuss about the segmentation and other previous technique to solve in understanding short text and problem associate with the previous approach over it. A further analysis need to be done is to determine refined approach over it.

***IndexTerms* - 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 processes 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 technologicalmethods 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 importantparts 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 givesan intelligent decision making. Classification is not only studies and examines the existing sample data but also predicts the future behavior 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 “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 HirstStOnge (as semantic relatedness algorithm) has better results than other semantic similarity algorithms.

In the figure 1 above, a knowledge extraction process from the large dataset by using semantic rules and semantic data annotation is performed.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 contextbased tags.

II. LITERATURE REVIEW

2.1 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].

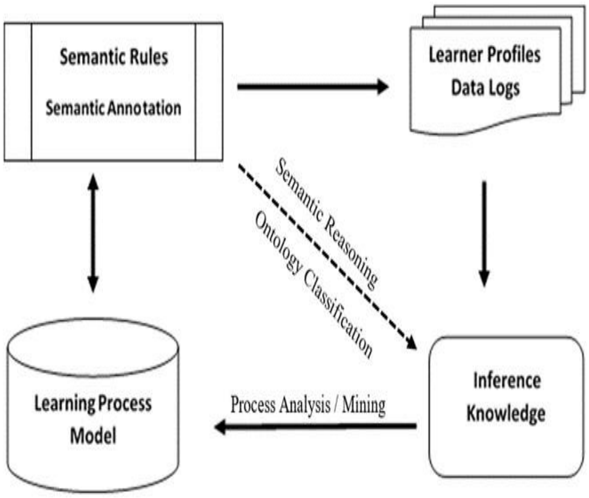

Figure 1: Semantic knowledge extraction and learning process

2.2 Mir SamanTajbakhsh, 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].

2.3 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].

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

In this work, author works towards the data interaction and its behavior. Various components such as subjective system aspects, user experience, and interaction and data detail consumption are 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].

2.5 RishabhUpadhyay, 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 itextpdfAPI. 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].

2.6 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].

Table 1: Comparison analysis between the previously defined approaches

Sr. No.	Author	Algorithm	Advantage	Remarks
1.	Wen Hua, Zhongyuan Wang, Haixun Wang	Short text analysis segmentation approach	Segmentation help in understanding short text over document.	High accuracy with large number of terminology.
2.	Mir SamanTajbakhsh, JamshidBagherzadeh	TF-IDF algorithm	Short text analysis with high accuracy	High end recommendation approach.
3.	Bart P. Knijnenburg, Martijn C. Willemsen, Alfred Kobsa	Behavior based recommendation	Limited detection but proper accuracy	Less number of detection is observed.
4.	RishabhUpadhyay, Akihiro Fujii	Hybrid approach with semantic and nlp process.	High accuracy and maximum detection is performed.	Can work with large amount of data.

In the table 1 above, multiple approaches for the semantic knowledge extraction and mining is performed.

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

III. CONCLUSION

Twitter and other social media platform where a rapid data generation is performed by the different platform unit. Social tagging and data generation with short text is often seen in the tags and reply by the audience. Hence understanding dealing with the large data is problem which can predict proper recommendation based on data utilization. In this paper tagging, segmentation, similarity function and similarity measure algorithm is studied which is done by previous authors. This paper shows the study done by previous authors which says that the different format of similarity measure algorithm with similarity function is computed. Short text understanding can be useful in better prediction and thus to provide better user support for a large number of tweets.

REFERENCES

[1]. Wen Hua, Zhongyuan Wang, Haixun Wang, “Understand Short Texts by Harvesting and Analyzing Semantic Knowledge”, IEEE transaction 2016.

[2]. Mir SamanTajbakhsh, JamshidBagherzadeh, “Microblogging Hash Tag Recommendation System Based on Semantic TF-IDF”, IEEE 2016 4th International Conference on Future Internet of Things and Cloud Workshops.

[3]. Otsuka, E., Wallace, S.A., and Chiu, D.: ‘Design and evaluation of a Twitter hashtag recommendation system’. Proc. Proceedings of the 18th International Database Engineering & Applications Symposium, Porto, Portugal2014 pp. Pages.

[4]. Givon, S., and MLavrenko, V.: ‘Predicting social-tags for cold start book recommendations’. Proc. Proceedings of the third ACM conference on Recommender systems, New York, New York, USA2009 pp. Pages.

[5]. Godoy, D.,Rodriguez, G., and Scavuzzo, F.: ‘Leveraging Semantic Similarity for Folks nomy-Based Recommendation’, IEEE Internet Computing, 2014, 18, (1), pp. 48-55.

[6]. Sigurbjornsson, B., and Zwol, R.v.: ‘Flickr tag recommendation based on collective knowledge’. Proc. Proceedings of the 17th international conference on World Wide Web, Beijing, China2008 pp. Page.

[7]. Bart P. Knijnenburg, Martijn C. Willemsen, Alfred Kobsa" A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems", RecSys’11, October 23–27, 2011, Chicago, Illinois, USA. ACM 978-1-4503-0683-6/11/10 (pp 321-324).

[8]. RishabhUpadhyay, Akihiro Fujii, “Semantic Knowledge Extraction from Research Documents”, Proceedings of the Federated Conference on Computer Science and Information Systems pp. 439–445 DOI: 10.15439/2016F221 ACSIS, Vol. 8. ISSN 2300-5963.

[9]. W. Shen, J. Wang, P. Luo, and M. Wang, “Linden: Linking named entities with knowledge base via semantic knowledge,” in Proceedings of the 21st International Conference on World Wide Web, ser. WWW ’12, New York, NY, USA, 2012, pp. 449–458.

[10]. Gautam R. Raithatha, “Knowledge Extraction for Semantic Web”, Knowledge Extraction for Semantic Web| ISSN: 2321-9939.

[11]. P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, “Computing term similarity by large probabilistic isa knowledge,” in Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management, ser. CIKM ’13, New York, NY, USA, 2013, pp. 1401–1410.

[12]. Erich Christian Teppan," Implications of Psychological Phenomenons for Recommender Systems", RecSys’08, October 23–25, 2008, Lausanne, Switzerland. ACM 978-1-60558-093-7/08/10 (pp 323-326).

[13]. A. S., and Yang, Y. (2008). Personalized active learning for Collaborative filtering. In SIGIR ’08: Proceedings of the 31st annual international ACM SIGIR conference on the research and development in the information retrieval, Singapore (pp. 91–98). New York: ACM.

[14]. G. Adomavicius and A. Tuzhilin.Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transaction Knowledge Data Eng., 17(6):734–749, 2005.

[15]. RishabhUpadhyay, Akihiro Fujii, “Semantic Knowledge Extraction from Research Documents”, Proceedings of the Federated Conference on Computer Science and Information Systems pp. 439–445 DOI: 10.15439/2016F221 ACSIS, Vol. 8. ISSN 2300-5963.

[16]. Gautam R. Raithatha, “Knowledge Extraction for Semantic Web”, Knowledge Extraction for Semantic Web| ISSN: 2321-9939.