International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Similarity Matching of Pairs of Text using CACT Algorithm

CH N Santhosh Kumar, V Pavan Kumar, K S Reddy

Abstract*: In data mining, shorter text analysis is performed more widely for many applications. Based on the syntax of the language, it is very difficult to analyze the short text with several traditional tools of natural language processing and this is not applied correctly either. In short text, it is known that there are rare and insufficient data available and further it is difficult to identify semantic knowledge with the great noise and ambiguity of short texts. In this paper, the authors proposed to replace the coefficient of similarity of Cosine with the measure of similarity of Jaro-Winkler to obtain the coincidence of similarity between pairs of text (source text and target text). Jaro-Winkler does a better job of determining the similarity of the strings because it takes an order into account when using the positional indices to estimate relevance. It is presumed that the performance of CACT driven by Jaro-Wrinkler with respect to one-to-many data links offers optimized performance when compared to the operation of CACT driven by cosine. In this paper, the ensemble algorithm CACTS and SAE is adopted with Jaro-Winkler similarity approach. The new algorithm is employed for short text analysis and better results. An evaluation of our proposed concept is sufficient as validation.

Keywords*: text mining, Cosine's similarity coefficient, Jar -Winkler similarity.

I. INTRODUCTION

In data mining, text mining is the domain most commonly used to extract knowledge from semantic data. Text mining is the method of extracting data from similar varied information and has the necessary relationship between the entities extracted. Within the text mining space, the classification of texts is one of each technique. It is one of many difficult data mining problems, since it manages high-dimensional data indexes with subjective samples of missing information. The address of the problem in this document was to classify text documents without a label. The problem would be described by taking a large group of tagged text documents and designing the data mining classifier.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Ch. N. Santhosh Kumar*, Dept. of CSE, Anurag Engineering College, Kodada, India.
V Pavan Kumar**, Dept. of CSE, Anurag Engineering College, Kodada, India.
Dr.K.S.Reddy***, Researcher, Anurag Group of Institutions, Hyderabad, India.

The wide accessibility of Internet records in electronic structures requires a programmed strategy to mark documents with a predefined set of topics, which is known as automatic short-text categorization (ASTC). Since the last few years, a large number of machine learning algorithms have been observed to handle this test task. In designing the TC journey as a meeting problem, several current learning methodologies are connected, but its unit of constraint area was discovered once the basic content is small.

There is an extensive type of use to handle short messages. Short messages present new problems with connected content assignments as well as knowledge retrieval (IR), order, and clustering. Unlike long files, two short messages that have a comparable meaning do not share very different words. For example, the implications of "uploading and returning Macintosh items" and the "new iPhone and iPad" area unit are firmly connected, however, they do not share traditional words. The absence of adequate factual knowledge causes successful challenges in the estimation of similarity, and therefore, the calculations of varied current content examinations do not specifically distinguish short messages.

Stanchion to short messages specifically. All the a lot of considerably, the absence of factual knowledge likewise implies problems which will be firmly forgotten after the researchers handle long reports find yourself basic for brief messages. Take lexical ambiguity for example. "Apple" offers ascend to varied implications in "apple item" and "apple tree". Due to the shortage of relevant knowledge, these ambiguous words build short messages arduous to understand by machines. In this paper, the proposed system replace Cosine's similarity coefficient with Jaro-Winkler similarity measure to obtain the similarity matching of text pairs(source text and destination text). Jaro-Winkler does a much better job at determining the similarity of strings because it takes order into account using positional indexes to estimate relevancy. It is presumed that Jaro-Wrinkler driven CACT's performance with respect to one-to-many data linkages offers an optimized performance compared Cosine driven CACT's workings. An evaluation of our proposed concept suffices as validation.

II. REVIEW OF LITERATURE

In this section, the authors present and appreciate the work done by several researchers in the area of text mining in general and coefficient part in specific.

The researcher A. McCallum and W. Li [2] from the tagged data, it is known that this method obtains the seeds for the lexicons and this is called Web List, according to the HTML rules and the service as the search.

For reference, based on the appearance of Arnold Palmer in the tagged data, the researchers have collected the data from the Web that is an extensive list of other golf players, including Tiger Woods (a phrase that is difficult to detect as a name without a good lexicon).

The researchers G. Zhou, et.al have presented the "Entity recognition named using a fragment tagger based on hmm". This article proposes the Hidden Markov Model (HMM) [3] and a fragment tagger based on HMM, from that entity recognition (NE) system (NER) is built to recognize and classify the names, times and quantities numerical.

The researchers D. M. Blei, et.al, proposes "Assignment of Latent Dirichlet". This article describes the allocation of latent Dirichlet [4] (LDA), a generative probabilistic paradigm for discrete data groups, such as text bodies. The LDA is a three-level, hierarchical, Bayesian model, in that each element of a meeting is recorded as a measurable mix over an underlying set of topics. Each theme, in turn, has modeled as an infinite mix over an underlying set of subject probabilities.

III. METHODOLOGY

A. Ensemble Semantic Knowledge Approach

The challenging problem of inducing taxonomy from a set of keyword phrases instead of large text corpus is the current project context. Conceptualization is the most widely used multiple inferencing mechanisms to work under various contexts [5]. All these contain a huge number of probabilistic concepts, interconnected and fine-grained and it is also called as Probase API. To express the semantics explicitly the concept based on the data is most powerful in comprehending the understanding of the short text [6]. For comparing the two short texts and also for classification and it is also alone to not compatible for the tasks. Consider the same two short texts: "upcoming apple products" and "new iphone and ipad".

In this system, for every short text, it is known that there are no common terms. To improve the performance such as similarity of their semantics by using the inferencing mechanism on Pro-base to retrieve the most popular similar terms such as noun and these are checked for new contexts for that short text. The first process is clustering the Pro-base terms based on their similar relationship, and with this, it is known that this belongs to the same cluster [7]. For instance, If there is a keyword "dogs", our Pro-base driven extracts other polysemy words like "mutt", "canines", "mongrel" etc which definitely forms a cluster group and the researchers shall repeat the process for other nouns in the short text and then from the obtained results the researchers shall identify the most commonest matching entities using a 3-layer stacked auto-encoders for hashing terms to reduce processing complexity.

Cosine similarity coefficient, a measure that is commonly used in semantic text classifications which measures the similarity between two texts and determines the probable measure.

CACT's approach to use Cosine's similarity co-efficient increases time complexity exponentially [8].

So the authors propose to replace Cosine's similarity coefficient with Jaro-Winkler similarity measure to obtain the similarity matching of text pairs (source text and destination text).

Jaro-Winkler does a much better job at determining the similarity of strings because it takes order into account using positional indexes to estimate relevancy [9].

It is presumed that Jaro-Wrinkler driven CACT's performance with respect to one-to-many data linkages offers an optimized performance compared Cosine driven CACT's workings [10].

An evaluation of our proposed concept suffices as validation.

B. Algorithm:

Step 1: Start
Step 2: Begin ETL (Extract, Transform and Load) process
Step 3: Load datasets
Step 4: Pre-processing of data
Step 4.1: Cleansing
Step 5: Implement CACTS
Step 6: Implement SAE
Step 7: Implement Cosine Similarity
Step 8: Short Texts
Step 9: Calculate time
Step 10: Display time
Step 11: Results
Step 12: Analysis of Results with the existing approaches

IV. EVOLUTION RESULTS

The evolution of the algorithm was done by using set of Software applications includes, Java Development Kit 1.8, Apache Tomcat Webserver on a higher end hardware machine. The experiment was deployed on the NetBeans 8.x, that is an Integrated Development Environment on the machine with the 8 GB RAM (Primary Memory) that loads huge datasets in faster pace and Intel Core i3 processor with 3.6 GHz clock speed. This set of sophisticated technology used as the experiment needs to process the huge datasets and load the articles by browsing from datasets. The analysis of the results is done in four different phases. In the first phase, the long texts are analyzed. TF-IDF + LongTexts where only LONG TEXTS that are analyzed as it was on an existing system. The articles are be categorized and article Classes also generated by running the proposed algorithm. The time complexity for TF-IDF Classification Build completed in 20.214587612 seconds for 15 files.

In the second phase, CACTS + SAE + Cosine Similarity + ShortTexts where only SHOTTEXT can be analyzed mostly using the algorithm. When the algorithm is applied then the articles are categorized as each and individual articles classified and given the result. The time complexity of TF-IDF Classification build was completed in 20.214587612 seconds for 15 files and for CACTS + SAE + Cosine Similarity + ShortTexts algorithm the time complexity is 13.500744525 seconds for 15 files.

In the third phase, CACTS + SAE + Jaro-Winkler Similarity + ShortTexts where only SHOTTEXT can be analyzed, mostly used algorithm, when the algorithm is applied then the articles is categorized as each and individual articles classified and given the result.

The time complexity of CACTS + SAE + Cosine Similarity + ShortTexts Classification Build completed in 13.500744525 seconds for 15 files and for CACTS + SAE + JaroWinkler Similarity + ShortTexts algorithm the time complexity is 8.576099063 seconds for 15 files.

Fig: 1 Semantic knowledge Results 1

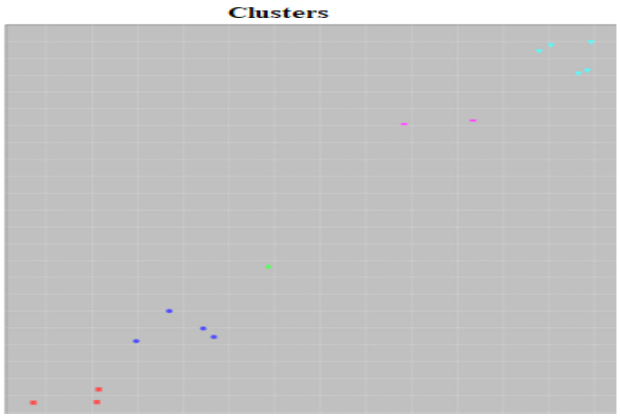

Fig: 2 Semantic knowledge Representation - Results 2

V. CONCLUSION

In this paper, the ensemble algorithm CACTS and SAE is adopted with Jaro-Winkler similarity approach. The new algorithm is employed for short text analysis and better results. Many existing systems performed semantic knowledge to understand the shortest text but those existing systems does not work properly. Finally, the researchers conclude that the proposed algorithm is an efficient, accurate and optimal in terms of time complexity & space complexity. The algorithm is proved that it works better and reduce the computation time to produce the better results within optimal time.

ACKNOWLEDGMENT

The researchers would express the deep sense of gratitude and thank the Management of Anurag Institutions for their unconditional support and extended cooperation. The researchers also express the heartfelt thanks to the colleagues at the Dept. of Computer Science and Information Technology for their support and encouragement throughout.

REFERENCES

1. Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, Senior “Understand Short Texts by Harvesting and Analyzing Semantic Knowledge”, IEEE Trans on Knowledge and Data Engineering. vol. 29, no. 3, Mar 2017.
2. A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction, and web enhanced lexicons", Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Volume 4, ser. CONLL '03, Stroudsburg, PA, USA, 2003, pp. 188–191.
3. G. Zhou and J. Su, “Named entity recognition using an hmm-based chunk tagger,” 40th Annual Meeting on Association for Computational Linguistics, ser. ACL ’02, Stroudsburg, PA, USA, 2002, pp. 473–480.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
5. M. Rosen-Zvi, "The author-topic model for authors and documents," 20th Conference on Uncertainty in Artificial Intelligence, ser. UAI '04, Arlington, Virginia, United States, 2004, pp. 487–494.
6. R. Mihalcea and A. Csomai, "Wikify! Linking documents to encyclopedic knowledge," 16th ACM conference on information and knowledge management, ser. CIKM '07, New York, NY, USA, 2007, pp. 233–242.
7. D. Milne and I. H. Witten, "Learning to link with Wikipedia," 17th ACM conference on Information and knowledge management, ser. CIKM '08, New York, NY, USA, 2008, pp. 509–518.
8. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of Wikipedia entities in web text", 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD 2009, New York, NY, USA, 2009, pp. 457–466.
9. X. Han and J. Zhao, "Named entity disambiguation by leveraging Wikipedia semantic knowledge," 18th ACM conference on Information and knowledge management, ser. CIKM '09, New York, NY, USA, 2009, pp. 215–224.
10. “Structural semantic relatedness: A knowledge-basedmethod to named entity disambiguation," 48th Annual Meeting of the Association for Computational Linguistics, ser. ACL '10, Stroudsburg, PA, USA, 2010, pp. 50–59.
11. KS Reddy, GPS Varma, MK Reddy, An Effective Preprocessing Method for Web Usage Mining, International Journal of Computer Theory and Engineering 6 (5), 2014, pp. 412-415.
12. CNS Kumar, VS Ramulu, KS Reddy, S Kotha, CM Kumar, Spatial data mining using cluster analysis, International Journal of Computer Science & Information Technology 4 (4), pp. 71-75.

AUTHORS PROFILE

Dr. CH. Naga Santhosh Kumar (CSI Membership No: 2010000525) is a professor at ANURAG Engineering College, KODADA, Telangana, India. He has received his Ph.D in Computer Science and Engineering from JNUTH, Telangana. He has 20 years of experience in teaching. He has authored 22 publications in premier indexed journals. He is serving as Reviewer and Editorial Board Member of reputed journals. His research interests are in the field of Data Mining, Big data, Software Engineering, Machine Learning and Artificial Intelligence.

Surya Pavan Kumar is currently working as Assistant Professor in CSE Department at ANURAG Engineering College, KODADA, Telangana, India. He has 9 years of teaching experience. His area of Expertise is Networking and Information security.

Author-3 Photo

Dr. K.S.Reddy has about 16 years of experience. He is currently working as a Professor in Information Technology Dept., at Anurag Group of Institutions, Hyderabad.