IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Value Model For Text Mining

K.Thyagarajan [1*], R.Nanthini [2]

[1*] Dept of Computer Science, AVC College, Mayiladuthurai, India
[2] Dept of Computer Science, AVC College, Mayiladuthurai, India

[*]*Corresponding Author: nanthini.11d@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 17/May/2018, Published: 31/May/2018

Abstract— Data Retrieval can be characterized as the movement of acquiring data or information significant to data needs from a gathering of data assets. In like manner, Resources are accessible in colossal, huge and unstructured. From the accessible Resources, just a section or some is needed by the clients. Hence some algorithm has to be derived to carry out this retrieving process. There are many previously proposed algorithms. Almost they are unique in nature of implementing the retrieving process. Many algorithms retrieve information based on term similarity where the result is not quite accurate to the user need. Here in this newly proposed algorithm called "Value Model for Text Mining (VM)" the Information retrieving process is performed based on semantic similarity. Hence this algorithm results in better performance in retrieving more related information by ranking the terms in the Fuzzy set.

Keywords— Data mining, Value Model, Ranking Tool, Fuzzy Set

I. INTRODUCTION

In general, knowledge can be defined as the Information. The information is stored in the database or repositories by an individual which can be fetched or retrieved by another. In order to have the ultimate use of the stored information it is necessary to retrieve the stored information more efficiently. The concept of Information Retrieval (IR) deals with storing, maintaining and searching of information. The information that is retrieved should be quiet enough for the user. It should be efficient and accurate. Information available on the Internet is retrieved by different users at different time with different speed. Increasing the speed of information retrieval activity can satisfy the user quickly. It should also be relevant to the user query. Such information retrieval activity is not an easy task. Information retrieval is a major process which needs more attention to undertake all the above said. Data Retrieval can be characterized as the movement of acquiring data or information important to a data needs from a gathering of data assets [1a].

IR provides the required information according to the user query within a collection of data. The Corpora that are available on the Internet and information stored in the databases of libraries and companies are huge in recent years. This leads to the difficulty in searching of information needed from the vast data that exists. The need to efficiently organize, search and manage textual corpora for knowledge has brought a new interest in the process of information retrieval (IR) and data mining strategies. In IR system the two important factors that directly affect the efficiency of the retrieval results is the approach to represent text and the measure to evaluate the similarity between query and documents [6]. Information Retrieval consists of three different types of classical models such as Set Theoretic, Algebraic and Probabilistic models. Set Theoretic model further has sub models such as Standard Boolean, Extended Boolean and Fuzzy models. In this paper Fuzzy model is used to retrieve information from the static collection of data. Fuzzy logic is the logic of imprecision and approximate reasoning. More specifically, fuzzy logic may be seemed as an attempt at mechanization of two remarkable capabilities [7a]. Fuzzy logic is a form of many. It deals with reasoning that is approximate rather than fixed and exact or accurate. In traditional binary sets the variables take two values such as values. In fluffy rationale, factors have a reality esteem that extents in degree in the vicinity of 0 and 1. Fluffy rationale has been stretched out to deal with the idea of fractional truth. In this reality esteem may go between totally evident and totally false [8a]. Further, when semantic factors are utilized these degrees might be overseen by particular capacities [9a]. In this proposed algorithm, the semantic similarity terms are found with the help of the WordNet. The documents are retrieved with semantically similar terms. In order to avoid duplication of terms, this algorithm uses the Porter stemmer algorithm and Fuzzy concept is used for further demonstration.

The rest of this paper is organized as follows: related works in Section 2; mining techniques in Section 3; methodology in

Section 4; experimental results and discussion in Section 5; conclusion in Section 6.

II. RELATED WORK

Here some previous works done by others based on Information retrieval are discussed.

Min Lu et. al. [11] proposed an algorithm with the similarity search at large scale. In order to perform the similarity search effectively hash functions are used. It referred to top k Rank Hash Similarity, in which a ranking loss function is designed for learning a hash function. Compared with the existing approaches, the proposed method has the same order of computational complexity. The target of fast similarity search is to find the similar data for given query efficiently and effectively from a large dataset. They have used experiment results on three text datasets demonstrate that the proposed method achieves high accuracy than the state-of-the-art techniques.

Wei Song et. al. [12] proposed a fuzzy control genetic algorithm (GA) in conjunction with a novel hybrid semantic similarity measure for document clustering. In order to evaluate the performance of the algorithm, two standard data sets such as Reuter (21578 version) corpus and 20-newsgroup (18828 version) corpus, are used for test. Thesaurus-based and corpus-based semantic methods are used to solve the complicated term indexing method. WordNet is used as the thesaurus-based ontology. It is concluded that Fuzzy control GA performed better than conventional GA with the same similarity measures.

Philip resnik et. al. [10] in "The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues" annotate the Biblical text to create the aligned corpus such as Bible for linguistic research which also includes the automatic creation and evaluation of translation lexicons and similar tagged text. It has the feature of parallel translations over huge number of languages. It also represents the comparison with dictionary and corpus resources for modern English. By this it makes the Bible a multilingual corpus which considered be a unique resource for linguistic research.

R. ThamaraiSelvi et. al. [13] proposed an algorithm based onBoolean Information Retrieval (BIR) that has the significance of its simplicity. Boolean model is used to predict whether each document is relevant or not.They used an online lexical reference called WordNet to find the semantically similar terms. Stemming is the process used for reducing inflected words to their stem, base or root form.

Wen-der Yuet. al. [15] proposed system for the retrieval of CAD documents. They designed their computational algorithm based on contents. The system has been tested with a sample CAD database comprising 2094 CAD documents from three sources. Two performance indexes namely Recall and Precision were measured for each sample for verification. They too have done the verification with the selected Chinese annotated CAD documents.

S.Niveditha et. al. [16] proposed an algorithm that works with various user's search goals for a query and reflecting each idea with some keywords. They have done the above process by clustering the proposed feedbacks. They have determined number of user search goals for a query and clustered the Pseudo Documents using Fuzzy Self Constructing Algorithm to get the final restructured search result. The final results on user click from a commercial search engine demonstrated the effectiveness of their proposed methods. It has been finalized that in future when a user searches for a same topic for many time that topic will be sent to the user.

Earlier research on sentiment analysis primarily focused on polarity classification, i.e., determining the sentiment orientation of a sentence or a document [19]. However, these tasks are all coarse-grained and cannot provide more detailed information. Recently, there has been a shift towards fine-grained aspect-based tasks that can identify both the text expressing the opinion and the aspect of the opinion as well as analysing its polarity (e.g., positive, neutral or negative) [17], [18]. The Aspect Polarity collocation extraction is the basic task of aspect-based sentiment analysis.

III. METHODOLOGY

In this proposed algorithm called "Value Model for Text Mining" information retrieval is done in an effective way by reducing the time taken for retrieving information most accurately. It considers the semantic similar terms for retrieving the documents. Fig. 1 shows the complete architecture of the proposed algorithm.

This algorithm is based on stemming process and ranking technique. Today search engine is very helpful for retrieving information for any sort of queries. It may be easy for the user to get large amount of information with a single word or text as the query to the search engine. But search engine has to find the related information from all the online repositories or databases which contains huge amount of information. This proposed algorithm performs the information retrieval by using terms in the user query. The information given by the user in the form of text, words or sentences is considered as the query. Each query is composed of terms. This algorithm performs the information retrieval with the help of terms in the query.

This algorithm considers each term as search wordR_w. The maximum possibilities of semantic similar terms are found with help of the dictionary or from the Online Dictionaries. Those similar terms are considered as the synset n. Each term in the synset is verified whether the

term is 'NOT' prefixed or not. If the terms in the synset is 'NOT' prefixed, semantic dissimilar terms are found for those 'NOT' prefixed terms from the WordNet and stored in set A. If the terms in the synset are with no 'NOT' prefixed, semantic dissimilar terms are found for terms without 'NOT' prefixed from the WordNet and stored in set B. Now the set A or Set B contains the collection of terms called as Acquired Text Collection (ATC). The processed terms from the ATC undergoes the stemming process to make the terms more accurate and for finding the root words.

These stemmed words or root words help the Information Retrieval (IR) process to precisely find the relevant documents D. This reduces the time consumption for IR process. After stemming process each $termt_i$ is given a value from the range of 0 to 1 called Ranking. By giving values to the terms or ranking the terms fuzzy set has been created. Therefore fuzzy set contains the valued terms.

Fluffy sets will be sets whose components in the sets have degrees of enrollment. Fluffy set hypothesis permits the continuous evaluation of the enrollment of components in a set with the assistance of a participation work esteemed in the genuine unit interim [1] [14].

By using the elements in the fuzzy set the related documents D is retrieved from the database or repository. After retrieving, each document d_i is weighted. Weighting the document is as follows: Each element in the fuzzy set is checked with the document d_i for its existence. If $elementt_j$ exists, the value of that element in ranking set R(t) is multiplied with the frequency of t_i in d_i which is denoted as $f(t_j, d_i)$ and the result is summed up with the document's weight. Similarly all the documents d_i € D is weighted. Further the weighted documents are ordered from maximum to minimum weights. The accuracy of the above process is experimented with ranking shown below:

ALGORITHM 1
VALUE MODEL FOR TEXT MINING

1. Search word r_w
2. SR<=SYN(r_w)
3. For all sr_i in SR ~~v~~i= 0 to count (SR)
   If(sr_iprefixedwith "NOT")
   ATC<= ANT(sr_i)
   ATC<=sr_i
   Else
   ATC<="NOT"+ANT(sr_i)
   ATC<=sr_i
   If (sris EMO)
   Search(EMO DB)
   RANK(sr)
4. RS<=STEMMING(ATC)
5. For all fs_iwherei= 0 to count (RS)
   RANK(fs_i) from 0 to1
6. For all d_i in DB
   D<=d_i where di € anyof(fs_i)
7. $w(D) = \sum_{i=0}^{n} w(tj, di)$
8. Order(D)
9. Disp(D)

Where SR means search results, ATC means Acquired Text Collection, ANT means function to find the Antonyms of search results, SYN means function to find the Synonyms of

Terms and conditions used in this algorithm is as follows: r_w as user search word, SYN (r_w) is the function which finds the synonyms for the search words and assigned in SR. For each term sr_i for i=0 to count of SR in SR is checked for NOT prefix and assigned in ATC (Acquired Text Collection).If the term is NOT prefixed, antonym is found else antonym is found and concatenated with NOT.After the completion of the term processing, stemming process is carryout as shown below:

FS<=STEMMING (ATC)

STEMMING (ATC) is a function which finds the root word for the terms and assigned in RS. RANK(rs_i) evaluate the rank for each term, fs_i. Ranking process takes place as shown below:

Rank of term t € RS is assumed as 1 if it is a root word which is given by user. The Rank of term t € RS is set to 0.9 if the term is the first level synonym of root word

RANK (t) =1 if t is root word
RANK (SYN(t))=0.9
Likewise RANK (SYN(SYN(t))) =0.8
Similarly
RANK(ANT(t)) =0.9 and RANK(SYN(ANT(t)))=0.8

From the existing documents, the documents that are related to the user needs is represented as D. In the related documents, the weight(w) of each document d_i is found as given:

$$w(D) = \sum_{i=0}^{n} w(tj, di)$$

$w(t_j,d_i)$ is the weight of term t_i in the document d_i
and it is calculated by
$w(t_i,d_i)=RANK(t_i)*tf(t_i,d_i)$
where tf is represented as term frequency which is calculated as

$$tf(tj, di) = \frac{f(tj, di)}{\max\{f(w, di) : w \in di\}}$$

The weighted documents are then ordered from largest weight to the smallest weight. The ordered documents are then displayed.

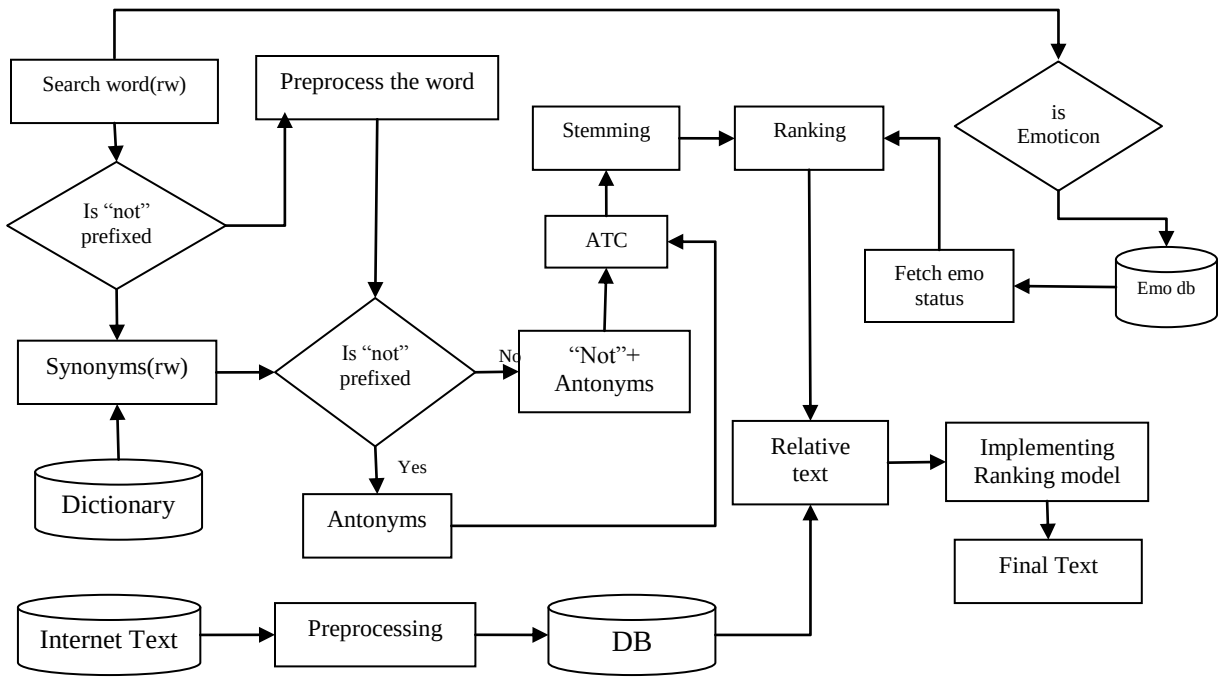

Fig .1 Architecture of proposed Algorithm

IV. EXPRIMENTS AND RESULTS

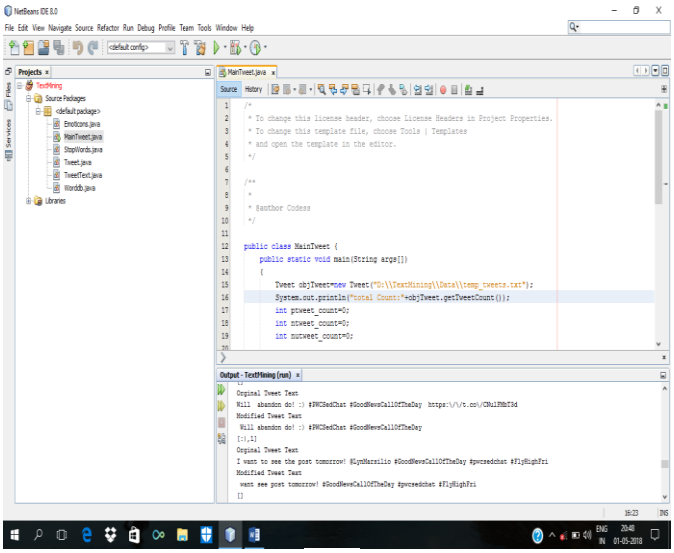

Fig.2

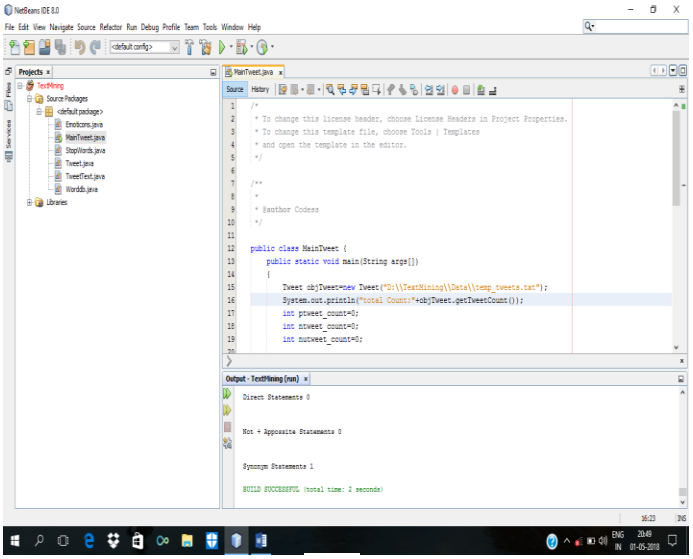

Fig.3

V. CONCLUSION AND FUTURE SCOPE

In the proposed algorithm, the Information retrieval is done based on semantic similarity for the user query or term with value model. This shows a good start by searching semantically similar results and it further undergoes stemming process. Ranking model used in this algorithm

improves the performance of the retrieving process such a way that it displays the most related many documents to the user by ranking the terms in the retrieved documents and manipulating the total weight of each retrieved document. Thus this algorithm performs well by retrieving more related information up to its maximum level to the user.

REFERENCES

[1] http://en.wikipedia.org/wiki/Information_retrieval

[2] Fayyad, Usama, Piatetsky-Shapiro, Gregory, Smyth, Padhraic (1996), "From Data Mining to Knowledge Discovery in Databases", Retrieved 17 December 2008

[3] "Data Mining Curriculum",ACM SIGKDD. 2006-04-30, Retrieved 2011-10-28.

[4] Clifton, Christopher (2010), "Encyclopædia Britannica: Definition of Data Mining", Retrieved 2010-12-09.

[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[6] Wei Song · Soon Cheol Park. "Latent semantic analysis for vector space expansionand fuzzy logic-based genetic clustering". © Springer-Verlag London Limited 2009.

[7] Lotfi A. Zadeh *,1." Is there a need for fuzzy logic?". _ 2008 Elsevier Inc

[8] Novák, V., Perfilieva, I. and Močkoř, J. (1999)Mathematical principles of fuzzy logic Dodrecht: Kluwer Academic. ISBN 0-7923-8595-0

[9] Ahlawat, Nishant, Ashu Gautam, and Nidhi Sharma (International Research Publications House 2014) "Use of Logic Gates to Make Edge Avoider Robot."International Journal of Information & Computation Technology (Volume 4, Issue 6; page 630) ISSN 0974-2239 (Retrieved 27 April 2014)

[10] Philip Resnik, Mari Broman Olsen and MONA DIAB The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues" Computers and the Humanities 33: 129–153, 1999.© 1999 Kluwer Academic Publishers. Printed in the Netherlands. 129

[11] Min Lu a, YaLou Huang a,b, MaoQiangXieb,c,, Jie Liu a "Rank hash similarity for fast similarity search ",Information Processing and Management 49 (2013) 158–168

[12] Wei Song a,b, Jiu Zhen Liang a, Soon Cheol Park b "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering", Information Sciences 273 (2014) 156–170

[13] R. ThamaraiSelvi, Dr. E. George Dharma Prakash Raj. "UFSBIR:A Semantic based Boolean Information Retrieval Algorithm with User Feedback ".

[14] http://en.wikipedia.org/wiki/Fuzzy_set.

[15] Wen-der Yu, Jia-yang Hsu "Content-based text mining technique for retrieval of CAD documents" © 2012 Elsevier B.V. Automation in Construction 31 (2013) 65–74

[16] S.Niveditha,T.Malathi,S.R.Sivaranjhani,"Efficient Information Retrieval using Fuzzy Self Construction Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 104 – No.1, October 2014

[17] J. Kaur, S.S. Sehra, S.K. Sehra, "A Systematic Literature Review of Sentiment Analysis Techniques", International Journal of Computer Sciences and Engineering, Vol.5, Issue.4, pp.22-28, 2017.

[18] Naveen Kumar Laskari, Suresh Kumar Sanampudi , "Aspect Based Sentiment analysis",IOSR Journal of Computer Engineering, Vol.18, Issue.2, pp.24-28, 2016.

[19] C. Nanda, M. Dua, "A Survey on Sentiment Analysis", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.67-70, 2017.

Authors Profile

Dr.K. Thyagarajan received his M.Sc (Computer Science) degree from Vinayagamission University, M.Sc(maths) degree from Madurai Kamaraj University, M.Phil (Computer Science) ,M.Phil(Maths) and Ph.D (Computer Science) degrees from Bharathidasan University. He is currently working as Hod and Associate Professor in the Department of Computer Science at AVC College (Autonomous), Mayiladuthurai. He has published serveral research papers in international jounals. His research area is Data Mining.

Miss.R. Nanthini has completed her M.Sc (Computer Science) degree from Bharathidasan University. Currently she is doing M.Phil (Computer Science) at A.V.C College(Auto), Mayiladuthurai. She is doing research in the area of Data mining.

