A COMPUTATION OF SIMILARITY MEASURE IN TEXT MINING BY INCORPORATING MULTI-PASS CLUSTERING AND CLASSIFICATION METHODOLOGY

Vinothkumar, M., Kannan, S and P.Solainayagi

Dept. of Computer Science and Engineering, Madha Engineering College, Kundrathur, Chennai-69.

ABSTRACT: Computation of a Text (word, term, label & phrase) statistically with predefined feature/measure is the traditional methodology used in the area of Text Mining. A model which is based on conceptual text mining with new and enhanced features has been introduced in this paper. The concept based model will minimize the demerits of mathematical approach and increase the performance of the text mining system. It computes the text at corpus, document and sentence level. Each term is syntactically analysed and measures are defined based on their conclusions. It computes the frequency of a term conceptually (Cft) and also analysis the Document term frequency (Dft) and Corpus level document frequency (Cdf) using conceptual mining model. Finally, the similarity of a feature/measure is computed by incorporating Multi-pass clustering and classification methodology. The term that identifies the semantics of the text has to be given importance more than the frequency of the term in the corpus/document. The proposed conceptual text mining model has been used on various datasets for analysis and their results are grouped into the text clusters. The experimental outcomes shows significant improvement in quality of clusters formed using the concept based computation on sentence and document level. The comparisons between the conceptual text mining and traditional analysis approach have been done and a result shows the significance of the proposed model.

Keywords: Pre-processing of the text, Concept based Analysis; Concept based Document Similarity, Classification methodology, Clustering Techniques.

I.Introduction

In a traditional statistical approach, representation of document is in vector format to indicate the value of respective features associated with it. Each feature value can be frequency of the term in the document, correlation between the frequency of the term (Cft) and gross occurrences of all terms in the document set and composition of inverse document frequency (Cft_Dif) and the term frequency. The vector representation of document resulted in one-dimensional sparse array. Computations of similarity for dataset are used in text classification and clustering algorithms [1]. There are many features proposed for analysis of similarity between two vectors. A non-symmetric measure proposed in Kullback-Leibler divergence finds the variance between two vectors. Euclidean distance is a commonly used method in statistical approach to determine the similarity measure. If the feature in a vector is always non-negative, then the Canberra distance metrics is used. Cosine similarity is an approach which calculates the measure by taking the angle between two vectors. Dhillon and Modha have adopted the similarity measure for document clustering used in cosine similarity to propose spherical k-means algorithm [2].

To discover the new and undetermined information, Text mining will be incorporating methodologies like machine learning and NLP. Text Clustering is a traditional approach where it comprises unsupervised learning and identifies the clusters which has less similarity with the document outside the cluster and high similarity for document within the clusters. Choosing the feature is most significant in text mining which has an impact in cluster formations. Then the frequency of the tern is analysed to find the significance of the same in the document. But, document can contain the term which have same or different frequencies [3]. In this case, one term may contribute more to the description of its sentences than the other and obtaining the correlation between arguments and verbs in same sentence has much impact in analysing the same.

Classification is a methodology for analysing the document and describes the semantics of the same. It has much application some of them are email document classification to filter spam, classifying the document to identify the line of interest and different feeds. Finally to get the customer feedback and reports. We have various approaches in text mining, some of them are rough set approach and fuzzy approaches. Clustering is grouping the document set that pertains to similar set/class. it is also helps in classifying the document and used in distribution of data to understand the characteristics of the cluster.

II.System Model

In the proposed model, the term that identifies the semantics of the text has to be given importance more than the frequency of the term in the corpus/document. The Fig 2.1 describes the overall system design of the proposed model. The measure used for determining the similarity is based on concepts that matching document and sentence at term and label level. The computed labelled terms capture the each sentence structure and its frequencies to measure the description of each sentence to the complete document clusters. The correlation between document and semantic structure to sentences in text are only classified with statistical approach whereas in the proposed model, both of them are analysed by understanding the semantics of the text to classify and cluster using the multi-pass clustering and classification algorithms [4].

The proposed model has different segments, an input block used to get the dataset to the model then the dataset is split into different labels according to the requirement. Pre-processing of the label information is carried out to remove terms which does not have any impact on the document set. Finally the dataset is computed at each level (Sentence, corpus and document).

III.Previous Work

Calculating the similarity of document set is a trivial activity in the text mining area. A new concept based similarity measure for a feature is introduced in this paper. To analyse the similarity between documents based on a feature, the suggested measure looks into the following metrics: 1) The feature that available in both sets of document, 2) the feature available in any one of the document set and 3) the feature that is not available in any of the document set. The efficiency of the similarity is directly proportional to the feature availability in the document set. Hence the first metrics has higher similarity measure since the difference between the documents set decreases. For the second metrics, the similarity will be nominal and scalable, since the measure is predetermined. Finally, for the last case the there is no significant feature matching between the document, hence very low similarity measure. The prospective of our measure is examined on various data sets and the outcomes show the significance of the proposed model.

The most commonly adopted measures for analysing the similarity between documents sets are listed below. Let the m_1 and m_2 be the two document set represented in the vector forms. The measure defined by Euclidean distance method [1] is represented as the square root of difference between the specific co-ordinates of m_1 and m_2.

$$EucD\ (m_1, m_2) = [(m_1 - m_2) . (m_1 - m_2)]^{1/2}, \ldots\ldots\ldots\ldots \quad (1)$$

where X*Y represents the inner product of the two vectors X and Y.

The Cosine similarity method [1] defines a measure which is represented as the cosine of the angle between doc_1 and doc_2:

$$Cos\_S\ (doc_1, doc_2) = [doc_1 . doc_2\ (doc_1 . doc_1)]^{1/2} / [(doc_2 . doc_2)]^{1/2} \ldots\ldots. \quad (2)$$

The measure defined by Pairwise-adaptive [1] will dynamically select a number of features out of m1 and m2:

$$PairD(m_1,m_2)=m_1,K.m_2,K.[(m_1,K\cdot m_1,K)]^{1/2}(m_2,K\cdot m_2,K) \qquad \ldots (3)$$

where mj, K is a subset of mj, j = 1, 2, which contains the values of the features that are the union of the K features appearing in m1 and m2, respectively.

The measure defined by Extended Jaccard coefficient [1] is an extended version of the Jaccard coefficient for similarity calculation

$$SEJ\ (m_1, m_2) =$$

$$m_1 . m_2\ m_1 . m_1 + m_2 . m_2 - m_1 . m_2 \ldots\ldots\ldots\ldots \quad (4)$$

and finally dice coefficient appears similar to above and it is represented as

$$SDic\ (doc_1, doc_2) = 2doc_1 . doc_2\ m_1 . doc_1 + doc_2 . doc_2 \ldots\ldots. \quad (5)$$

Hofmann`s Cluster Abstraction Model [5] employs word existence statistics on specific contents and its completely data driven. This model extracts correlation between document sets hierarchically. A specific algorithm (Expectation-Maximization) used to obtain the evaluation of parameters used in the same. The merits of this model are fast cluster summarization and data retrieval.

Jurafsky describes the need of analyzing the structure in form of semantic for unstructured text [6]. The technique of semantic parsing which defines the sentences structure in text. This technique helps in understanding and solving the problem of classification using VSM. It constitutes the multi classification problem in which the supervised data is used.

Feldman and Dagan, [7] work integrates the text categorization and Knowledge discovery to provide advanced solution on both of the areas. An enhanced framework is defined with various components which include concept hierarchy definition, text categorization by concepts and comparing the same with hierarchy.

Rinck classified the text into three formats [8] as words, predicate and position in text. The data is extracted to categorize the same. The paper compared this approach with other approach to classify the text. To achieve the same, relations between attribute in text are represented with combination of word context relations.

In all of the above described papers, the correlation between document and semantic structure to sentences in text are only classified with statistical approach whereas in the proposed model, both of them are analysed by understanding the semantics of the text to classify and cluster using the multi-pass clustering and classification algorithms. For attribute extract from the document set, the bag-of-words representation of sentence is used and it is integrated with the simple word context to make their relation adequate. By this, the similarity measure has been increased which is useful for grouping and further analysis in text mining.

IV.PROPOSED METHODOLOGY

In a document m with n features $q_1, q_2. . . q_n$ it can be represented in n-dimensions of vector, where, $m = \{ m_1, m_2, . . . , m_n \}$. If q_i, $1 \leq i \leq n$, is absent in the document, then mi = 0. Otherwise, $m_i > 0$. The document will be pre-processed at the minimum level comprising word, character and stem to evaluate the same. It is the imperative part of any Machine learning and NLP system to split that into the fundamental units to process at different stages to compute the components. The Analysis of the sentence, term and label is carried out by conceptually computing the same. The Document and lowest level in the document is analysed and similarity is measured by incorporating conceptual mining model with Classification and Clustering Techniques. The paragraph in the document is identified by the REGEX technique, it separates the sentence in the paragraph and label the term and also remove the stop words in the sentence. Finally in the stem words, the suffix and prefix are completely removed and outputs only the required word from the document.

The Concept based mining model analyse the document by obtaining the concept hierarchy with their object collection and properties. Each concept and its sub-concept in hierarchy are represented by its objects to analyse the document effectively. It uses different technique to analyse, some of them are frequency of term is analysed, document frequency is analysed and finally both of them are analysed conceptually. The below derived properties are to be considered for computing the similarity measure of the document set.

Property 1: The existence and non-existence of a feature to compute a measure is important than the difference between the values associated with the existing feature. Let the document m_1 and m_2 have a two features q_i and q_j. If q_i not in m_1 and exist in m_2, then qi have no correlation with m_1 and some relation with m_2. Here, m_1 and m_2 are variant in terms of qi. If q_j exist in both m_1 and m_2, then q_j has correlation with m_1 and m_2 concurrently. Hence m_1 and m_2 are similar with q_j.

Property 2: The similarity ratio should increase when the difference between two non-zero values of a feature decreases. Hence the similarity measure proposed will have higher efficiency than the other measures defined.

Property 3: The similarity ratio should decrease when the number of existence-non-existence features increases. For an existence-non-existence feature of m_1 and m_2, m_1 and m_2 is variant w.r.t feature. Therefore, as the number of existence and non-existence features increases, the variant between m_1 and m_2 increases and hence the similarity decreases.

Property 4: The document sets are least significant to each other if any of the features does not have non-zero values in all documents set.

Property 5: The similarity measure should be symmetric. The similarity ratio between m_1 and m_2 should be the same as of that between the document m_2 and m_1.

Property 6: The distribution of a value to the feature need to be considered, i.e., the feature`s standard deviation to be taken into account, for measuring the similarity between document sets.

Computation of Similarity between Document sets: By incorporating the properties mentioned above, a similarity measure is proposed, called Similarity Measure for Text Mining(SMTM), for document sets $m_1 =< m_{11}, m_{12}, \ldots, m_{1n} >$ and $m_2 = < m_{21}, m_{22}, \ldots, m_{2n} >$

$$G(m1,m2) = \sum_{j=1}^{n} N \star (d1j, d2j) / \sum_{j=1}^{m} N \cup (d1j, d2j)$$

Then the similarity measure proposed, SMTM, for m_1 and m_2 is

$$S_{SMTM}(m1,m2) = G(m1,m2) + \lambda/1 + \lambda$$

The proposed measure has considered the following:

1) The feature will be available in all document set,

2) Feature is available in only one document set and

3) Finally, Feature does not available in any of the document set. We set a non-positive constant -λ for the non-zero feature value.

The concept-based similarity measure proposed calculates the importance of each context at the sentence part by the Cft measure, term frequency at document level and finally at the corpus level. The similarity between the document set is assessed by incorporating the concept based analysis algorithm. The similarity measure defined in this model is the function of the below elements:

In each document, the no. of similar concepts, x, in argument structures.

In each document m, the total no. of sentences, sx that has similar concepts cp_i.

In each sentence s, the total no.of.labeled argument structure defined.

The Ctf_i of each concept cp_i in s for each document m, where i=1, 2, n

The Tf_i of each concept cp_i in each document m, where i= 1, 2, n.

The Df_i of each concept c_i, where i =1, 2, n.

The length, r, of all concepts in the argument structure of each document m,

In each corpus, the total number of documents is represented as M.

The similarity between concepts can exist or does not exist. If exist then concepts have similar words (wr), else no similar words with it. In the following concepts, cp_1=”$wr_1wr_2wr_3$” and cp_2=”wr_1wr_2” where cp_1 and cp_2 are concepts and wr_1, wr_2, wr_3 are the separate words.

Once the stop words are extracting and if $cp_{1>}$ cp_2 it specifies that cp_1 has more conceptual statistics than cp_2. In such scenarios, to measure the similarity between cp_1 and $cp_{2,}$ the length of cp_1 is used.

The objective of the concept based analysis is to distinguish the concepts in the document which correlate to the semantics of the document. The concept can be a least part of the document either a term or label which denotes the context of the text information.

These concepts are computed based on the semantic structure which tags to the each label or term in the sentence of the document. Each sentence will have one or more argument structures. The following algorithm describes the proposed model.

mdoc$_i$ is a new Document Q is a matched concept list (Initially empty) Stdoc$_i$ is a new sentence in mdoci Build concepts list Cpdoc$_i$ from Stdoc$_i$ for each concept cp$_i$ € Cp$_i$ do compute Ctf$_i$ of cp$_i$ in ddoc$_i$

compute Tf$_i$ of cp$_i$ in ddoc$_i$ compute Df$_i$ of cp$_i$ in ddoc$_i$ D$_k$ is seen document, where k = {0; 1; . . . ; doci – 1}

S$_k$ is a sentence in D$_k$ Build concepts list Cp$_k$ from sk for each concept Cj € Cp$_k$ do if (cp$_i$ == cp$_j$) then update df$_i$ of c$_i$ compute ctf weight = avg (ctf$_i$, ctf$_j$) add new concept matches to list Q end if end for end for output the matched concepts list Q

The defined algorithm explains the process of calculating the term frequency (Ctf), document frequency (dF) derived by conceptually analysing the same. The algorithm starts with reading the document set and labelling them semantically. The concepts and their argument structures which are similar are filtered and their length is calculated. Each concept in the structure denotes the sentence semantics and processed sequentially. To find the similarity between the document set, all present and previous patterns in the document set are analysed and computed accordingly. Once the dataset is processed, the Q contains the concepts that are matching between each previous and current document (m).

For Concept (cp) in Sentence (st), Measuring conceptual term frequency (Ctf):

The Ctf is the total no. of occurrences of the concept (cp) in the argument structure of sentence st. The concept (cp), which has the different structure of sentence (sp) to provide the meaning of (sp) at sentence level.

In Document (m), Measuring Ctf of Concept (cp)

A concept (cp) can have many Ctf values in different sentences in the same document m. Thus, the Ctf value of concept (cp) in document m is calculated by:

$$Ctf = \sum_{n=i}^{spn} Ctf/sp$$

where spn is the total number of sentences that contain concept (cp) in document m. The average of the Ctf values of concept (cp)in its sentences (st) of document m calculates the overall significance of concept (cp) to the context of its sentences in document m. A concept (cp), which has Ctf values in most of the sentences in a document m, has a major contribution to the context of its sentences that leads to determining the subject of the document.

Calculating the Similarity Measure using Document-Based Concept Analysis:

At the document level, to compute each concept (cp) the conceptual term frequency Ctf, the number of occurrences of a concept (cp) in the document set, is measured. The tf is a measure on the document level only.

Calculating the Similarity Measure using Concept-Based Analysis:

Input: Test document M and a label profile L then for the Output: Similarity measure value, do similarity Measure between the M & L by:

$$\text{cos\_sim}(M, L) = \frac{\text{dot}(M, L)}{|M|\,|L|}$$

$$\text{dot}(M, L) = \sum_{i=0}^{v} mi.\, li$$

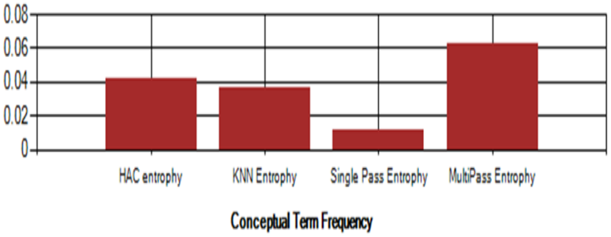
$$|M| = \sqrt{}\sum_{i=1}^{v} mi$$

$$|L| = \sqrt{}\sum_{i=1}^{v} li$$

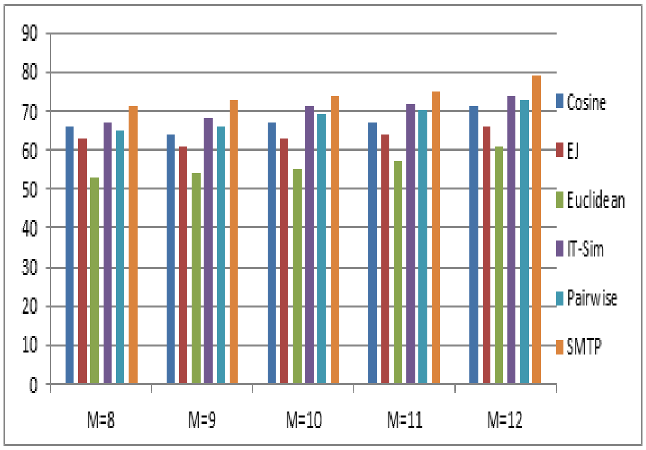
V.Simulation/Experimental Results

In this section, we investigate the effectiveness of our proposed similarity measure. The investigation is done by applying our measure in several text applications, including k-NN based single-label classification (SL-kNN) , k-NN based multi-label classification (MLkNN) , k-means clustering (k-means) , and hierarchical agglomerative clustering (HAC) and single pass classification. We also compare the performance of measure with that of other five measures, Euclidean, Cosine, Extended Jaccard (EJ), Pairwise-adaptive (Pairwise), and IT-Sim described. Note that the percentage of features taken into account for the proposed measure is set to be 90%. For the Pairwise-adaptive measure, K is determined by the product of the minimum number of non-zero features in the two documents and the percentage of features taken into account.

To evaluate the effectiveness of concepts that are matching in determining a measure of the similarity between document sets, large sets of experiments are conducted using the concept-based term analysis and the similarity measure is calculated. The similarity measure calculated using the Multi-pass clustering and classification algorithms shows significant improvement in quality of clusters formed using the sentence, document and combined approach on concept based computation.



5.1 Results of Ctf for the defined feature measurement



5.2 Comparison of Ctf Results for a feature measurement with Different Clustering Algorithms

VI.CONCLUSION

The proposed conceptual text mining model with new and enhanced features bridges the gap between machine learning and text mining principles. This model composed of different components which are proposed to improve the similarity measure computed. The sentence-based concept analysis which computes the semantic structure of each sentence in the document set to calculate the term frequency Ctf measure. The document-based concept analysis, measures each concept at the document level using the concept-based term frequency tf. Then the document frequency is captured at corpus level and finally calculating the similarity measure of each concepts w.r.t semantics of the sentence in document set.

VII.FUTURE SCOPES

There are various scopes to extend this paper, some of them are to extend the work on the web document clustering and another area of implementation would be text classification schemes. Similarity measures defined using feature sets for Retrial of images can be applied in the field of bio-medical. Furthermore, scientists in this field are working on the organizing and retrieval of data from various dataset on multiple signal data.

REFERENCES

1. Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee. 2014. “A Similarity Measure for Text Classification and Clustering” IEEE Transactions on Knowledge and Data Engineering,. 26 (7): 1-5.
2. Meenambigai Krishnamoorthy, Menaka Mani. 204. “ A Brief Survey on Text Mining and its Applications” Int.J.Computer Technology & Applications, 5 (5),1637-1640.
3. Jin, M.-L. Wong, and K.S. Leung, 2005. “Scalable Model-Based Clustering for Large Databases Based on Data Summarization,” IEEE Trans.Pattern Analysis and Machine Intelligence, 27 (11): 1710-1719.
4. Sebastiani, F. 2002. “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, 34 (1): 1-47.
5. Hofmann, T. 1999. “The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data,” Proc. 16th Int’l Joint Conf. Artificial Intelligence (IJCAI ’99), pp. 682-687, 1999.
6. Pradhan, K. Hacioglu, W. Ward, J.H. Martin, and D. Jurafsky, “Semantic Role Parsing: Adding Semantic Structure to Unstructured Text,” Proc. Third IEEE Int’l Conf. Data Mining (ICDM), pp. 629-632, 2003.
7. Feldman and I. Dagan, “Knowledge Discovery in Textual Databases (KDT),” Proc. First Int’l Conf. Knowledge Discovery and Data Mining, pp. 112-117, 1995.