International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 7: July 2015, pp 7-10. www.aetsjournal.com ISSN (Online) : 2394-6237

Text Optimizer Using Classification and Clustering

K. Prabhushankar , J. Ramkumar , U. Sugumar , N. Vasudevan

Abstract— A concept-based mining model analyzes the terms on sentences, documents and corpus levels. This model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the meaning sentence. Based on a new concept-based similarity measure similarity between the documents is calculated. A system is developed to determine the statistical value for the document by computing term frequency, conceptual term frequency and document frequency. The proposed mining model consists of sentence-based on document-based concept analysis, concept analysis, concept-based concept-analysis, and corpus-based similarity measure. Recapitulation reduces each document based on a set of rules. Similarity between documents is done in both the methods. And are clustered using Multi pass Clustering.

Keywords — Document classification, document clustering, entropy, accuracy, classifiers, clustering algorithms

I. INTRODUCTION

Many automated prediction methods exist for extracting patterns from various sample cases. In text mining, specifically text categorization, the raw cases are individual documents. We can transform these cases into a standard model of features and classes. For each case, we take a uniform set of measurements on the features. A dictionary is compiled from the collection of training documents. We measure the frequencies of occurrence of dictionary words in each document. Prediction methods look at samples of documents with known topics and attempt to find patterns for generalized rules that can be applied to new unclassified documents. We can describe sample cases in terms of dictionary words or phrases found in the documents. We also label each case to indicate the classification of the article it represents. Our objective is to compute decision criteria that distinguish between text categories. Given data classified using a standard numerical encoding; we can apply dozens of different data-mining methods. We're interested in the numerous secondary characteristics of collecting and applying the dictionary words, including stemmed or unstemmed words and binary (true or false) occurrence or word counts. Divisive information-theoretic feature clustering algorithm for text classification using the Kullback-Leibler divergence. We propose a new measure for computing the similarity between two documents. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with a present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. The measure is applied in several text applications, including singlelabel classification, multi-label classification, k-means like hierarchical agglomerative clustering, and clustering and the results obtained demonstrate the effectiveness of the proposed similarity measure. A lot of automated prophecy methods having for extracting patterns from sample cases, the raw cases are individual documents. We can transform these cases into a standard model of features and classes. For each case, we take a uniform set of measurements on the features. A dictionary is compiled from the collection of training documents. We measure the frequencies of occurrence of dictionary words in each document. Prediction methods look at samples of documents with known topics and attempt to find patterns for generalized rules that can be applied to new unclassified documents. We can describe sample cases in terms of dictionary words or phrases found in the documents. Each case consists of the values of a single article's features; these values could either be Boolean or numerical. We also label each case to indicate the classification of the article it represents. Our objective is to compute decision criteria that distinguish between text categories. Given data classified using a standard numerical encoding; we can apply dozens of different data-mining methods.

K. Prabhushankar, J. Ramkumar, U. Sugumar , UG Student, Department Of Computer Science And Engineering, Anand Institute Of Higher Technology, Chennai, India.

N. Vasudevan , Assistant Professor, Department Of Computer Science And Engineering, Anand Institute Of Higher Technology, Chennai, India.

II. RELATED WORKS

Similarity measures have been extensively used in text classification and clustering algorithms. In this method the unlabeled document collections are becoming increasingly common and also available. The text documentsare often represented as highdimensional and sparse vectors using words as features. The algorithm outputs k disjoint clusterseach with a concept vector that is the centroid of the cluster normalized to have unit Euclidean norm. Pair wise-adaptive similarity measure for large high dimensional document datasets improves the unsupervised clustering quality and speed compared to the original cosine similarity measure. Zoulficar younes reported results of clustering experiments with clustering algorithms and concluded that the objective functionbased on cosine similarity will leads to the best solutions irrespective of the total number of clusters. Daphe Koller and Mehran Sahami introduced a divisive

information-theoretic feature clustering algorithm for text classification using the Kullback-Leibler divergence. High dimensionality of text can be a discouraging process in applying complex learners such as Support Vector Machines to the task of text classification Kullback and Leibler combined squared Euclidean distance with relative entropy in a k-means algorithm. This algorithm is introduced recently as specifically designed to handle unit length document vectors. Zoulficar younes, conclude that the objective function based on cosine similarity leads to the best solutions irrespective of the number of clusters for most of the data sets. Chim and Deng performed document clustering based on the proposed phrase based similarity measure.

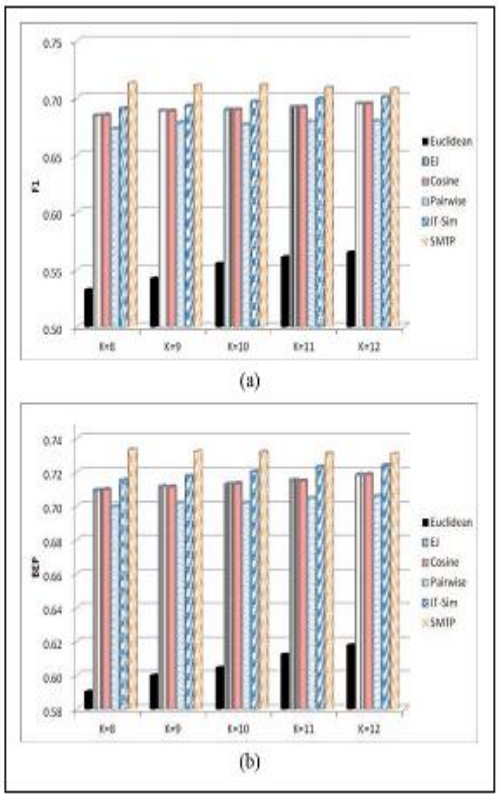

Fig 1. Classification Of Performance Compared By ML-kNN With Different Measures On Testing Data Of RCV1 In tf-idf. (a) F1. (b) BEP.

III. EXISTING SYSTEM

Assessing the main topics of texts and organizing them into meaningful structures is a time consuming and labour intensive process, which has become infeasible with the perpendicularvolume of electronically available information in today's world. Prediction methods look at samples of documents with known topics and attempt to find patterns for generalized rules that can be applied to new unclassified documents. The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful for only probably correct information. Usually, in text mining techniques, the term frequency of a term phrase or word is computed to explore the importance of the term in the document. However, two terms in the documents can have the same frequency, but one term contributes more meaning to the sentences than the other term. In this case, the mining model can capture terms that present the concepts of the sentence, which leads to the discovery of the topic of the particular document. A new concept-based mining model will analyzes the terms on the document,corpus levels and sentence is introduced. Text mining attempts to discover the new document, and the previously unknown information by applying techniques from natural language processing and data mining. This similarity measure outperforms other similarity measures that are based on term analysis models of the document only. The difference between documents is based on a combination of document-based, corpus-based, and sentence-based concept analysis.

IV. PROPOSED SYSTEM

A concept-based mining model has been proposed to overcome the disadvantages. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within the particular documents only. In the proposed model, there are three measures for analyzing concepts on the, document, corpusand sentence levels are computed.

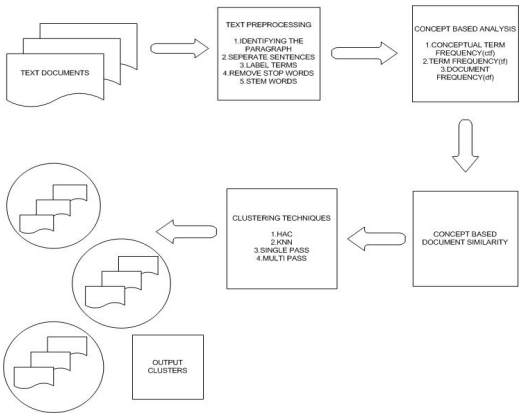

Fig 2. Implementation Of Proposed Architecture

Each of the sentence is labeled by a semantic role labeler which determines the terms used to contribute the sentence. The term which has a semantic role in the particular sentence, then that is called as a concept. Concept can be either phrases or words and are totally dependent on the semantic structure of the sentence. When a new document is introduced, the proposed mining model can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts. A new concept-based similarity measure which makes use of the concept analysis on the sentence, document and corpus levels which is proposed.

V. IMPLEMENTATION

A.Text Mining

Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the sremoval of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \quad (1)$$

Therefore, the formula of the surface subsidence due to underground mining is:

$$S(x,s) = \frac{ma}{r} \int_{x_1}^{x_2} \exp\left[-\pi \frac{(x-s)^2}{r^2} \right] dx \quad --(2)$$

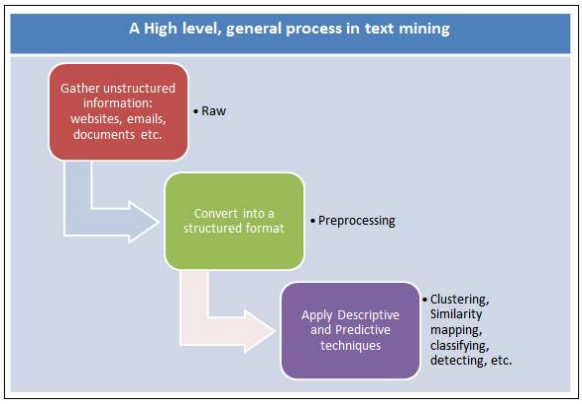

Fig 3. General Process Flow Of Text Mining.

B.Concept Based Mining Model

The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure. A raw text document is the input to the proposed model. Each document has well defined sentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based model on the sentence and document levels. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence.

C.Text Clustering

Clustering is an unsupervised classification process; differently from supervised classification no a priori information about classes is required. Document clustering is an optimization process which attempts to determine a partition of the document collection so that documents within the same cluster are as similar as possible (cluster compactness) and the discovered clusters as separate as possible (cluster distinctness).

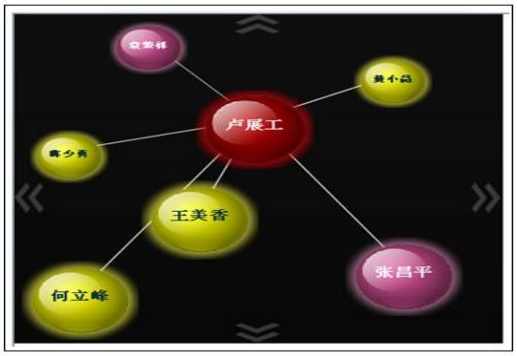

Fig 4. Result Of Clustering Information Of Chinese Persons

D. Concept Based Statistical Analyzer

The objective of this task is to achieve a concept-based statistical term analysis (word or phrase) on the sentence and document levels rather than a single-term analysis in the document set only. The ctfis the number of occurrences of concept c in verb argument structures (of sentence a. The concept c, appears frequently in different verb argument structures of the (same sentence a, has the main role of contributing to the meaning of a. The major challenge of clustering is to efficiently identify meaningful groups that are concisely annotated.

Calculation of the Odds Ratio:

The calculation of the odds ratio is very simple and the formula is as follows:

$$\text{Odds ratio} = \frac{PG_1 / (1 - PG_1)}{PG_2 / (1 - PG_2)} \quad --(3)$$

Where "PG_1" represents the odds of the event of interest for Group 1, and "PG_2" represents the odds of the event of interest for Group 2.

E. Conceptual Ontological Graph (COG)

The COG representation is a conceptual graph in which G=(C,R) where the concepts of the sentence, are represented as vertices (C). The relations among the concepts such as objects, actions and agents are represented as (R). C is a set of nodes (c1, c2,…,cn), where each node c represents a concept in the sentence or a nested conceptual graph G; and R is a set of edges (r1; r2,….,rm), such that each edge r is the relation between an ordered pair of nodes (ci,.., cj). The output of the role labeling task, hitch are verbs and their arguments are presented as concepts with relations in the COG representation as a graph. This relation allows the use of more informative concept matching at the sentence-level and the document-level rather than individual word matching. The concept-based model proposes new weight to each position in the COG representation to achieve more accurate analysis with respect

to the sentence semantics. Thus, each concept in the COG representation is assigned a proposed weight, which is weight COG, based on its position in the representation.

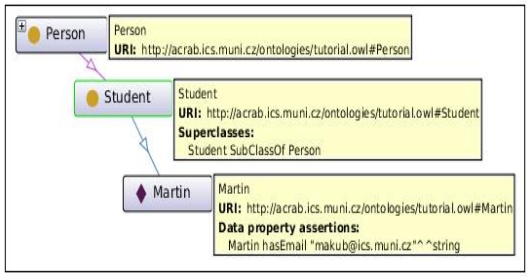

Fig 5. OWL Antology For A Set Of Axioms.

F. Concept-Based Extractor Algorithm

The concept extractor algorithm describes the process of combining the weight static(computed by the concept-based statistical analyzer) and the weightCOG(computed by the COG representation) into one new combined weight called weight comb. The concept extractor selects the top concepts that have the maximum weight combvalue. The proposed weight comb is calculated by:

weight combi= weight statici * weightCOGi(5) ---- (4)

The procedure begins with processing a new document which has well defined sentence boundaries. Each sentence is semantically labeled. For each labeled sentence, concepts of the verb argument structures which represent the semantic structures of the sentence are extracted to construct the COG representation. The concepts list L is sorted descendingly based on the weight combvalues. The maximum weighted concepts are chosen as top concepts from the concepts list L.

VI. CONCLUSION AND FUTURE ENHANCEMENT

In the proposed system we have presented a novel similarity measure between two documents. Several desirable properties are embedded in this measure. For example, the similarity measure is symmetric. The presence or absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity degree increases when the number of presence-absence feature pairs decreases. Two documents are least similar to each other if none of the features have non-zero values in both documents. Besides, it is desirable to consider the value distribution of a feature for its contribution to the similarity between two documents. The proposed scheme has also been extended to measure the similarity between two sets of documents. To improve the efficiency, we have provided an approximation to reduce the complexity involved in the computation. We have investigated the effectiveness of our proposed measure by applying it in *k*-NN based single-label classification, *k*-NN based multi-label classification, *k*-means clustering, and hierarchical agglomerative clustering (HAC) on several real-world data sets. The results have shown that the performance obtained by the proposed measure is better than that achieved by other measures.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, *Member, IEEE TRANS "* A Similarity Measure for Text Classification and Clustering," Jul 2003.

[2] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering," *Inf. Sci.*, vol. 180, no. 12, pp. 2341–2358, 2010.

[3] R. O. Duda, P. E. Hart, and D. J. Stork, *Pattern Recognition*. NewYork, NY, USA: Wiley, 2001.

[4]] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Clusteranalysis and display of genome-wide expression patterns," *Sci.*,vol. 95, no. 25, pp. 14863–14868, 1998.

[5] H. Fang, T. Tao, and C. Zhai, "A formal study of heuristic retrievalconstraints," in *Proc. 27th SIGIR*, Sheffield, South Yorkshire, U.K.,2004, pp. 49–56.

[6] P. K. Agarwal and C. M. Procopiuc, "Exact and approximationalgorithms for clustering," in *Proc. 9th Annu. SODA*, Philadelphia,PA, USA, 1998, pp. 658–667.

[7] D. W. Aha, "Lazy learning: Special issue editorial," *Artif. Intell.Rev.*, vol. 11, no. 1–5, pp. 7–10, 1997.

[8] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of informationretrieval based on measuring the divergence from randomness,"*ACM Trans. Inform. Syst.*, vol. 20, no. 4, pp. 357–389,2002.

[9] J. A. Aslam and M. Frost, "An information-theoretic measure fordocument similarity," in *Proc. 26th SIGIR*, Toronto, ON, Canada,2003, pp. 449–450.

[10] G. H. Ball and D. J. Hall, "A clustering technique for summarizingmultivariate data," *Behav. Sci.*, vol. 12, no. 2, pp. 153–155,1967.

[11] D. Cai, X. He, and J. Han, "Document clustering using localitypreserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12,pp. 1624–1637, Dec. 2005.

[12] H. Chim and X. Deng, "Efficient phrase-based document similarityfor clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9,pp. 1217–1229, Sept. 2008.

[13] S. Clinchant and E. Gaussier, "Information-based models forad hoc IR," in *Proc. 33rd SIGIR*, Geneva, Switzerland, 2010,pp. 234–241.

[14] M. Craven *et al.*, "Learning to extract symbolic knowledge formthe world wide web," in *Proc. 15th Nat. Conf. Artif. Intell.*, MenloPark, CA, USA, 1998.

[15] I. S. Dhillon, J. Kogan, and C. Nicholas, "Feature selection anddocument clustering," in *A Comprehensive Survey of Text Mining*,M. W. Berry, Ed. Heidelberg, Germany: Springer, 2003.

[16] I. S. Dhillon and D. S. Modha, "Concept decompositions for largesparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1,pp. 143–175, 2001.