VOL. 13, NO. 3, FEBRUARY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ANALYSING EFFECTIVE METHODOLOGIES USED FOR TEXT CLUSTERING USING WEIGHTED ALGORITHMS

S. Sree Dharinya
School of Information Technology and Engineering, VIT University, India
E-Mail: dharinya_sk@yahoo.com

ABSTRACT

Clustering of text documents is an important technique for enhancing automated learning. Matching is the technique used in order to relate or match the various set of related documents. Clustering groups a set of documents which are similar and dissimilar for unsupervised learning where the user has learning materials which are from raw data which requires further classification. Established feature extraction strategies intend to change over the representation of the major dimensional data set into a lower-dimensional informational collection by anticipating process through mathematical changes. The concept of feature clustering is to aggregate the features into clusters with a high level of pair wise semantic relations. Each cluster is dealt as a single new feature, and, hence, feature dimensionality can be radically lessened. HAC, K-Means, TF/IDF-weighted vectors and cosine similarities is used for the various vectors of data and is applied to text in a direct way to optimize the vectors.

Keywords: NMI, term frequency, cluster similarity, weighted algorithm, hyponyms, hyponyms.

INTRODUCTION

The text document clustering can incredibly simplify browsing huge collections of documents by redesigning them into more number of reasonable clusters. Existing algorithms have the constraints like grouping but cannot overcome the ambiguity as well as the exact synonyms and also existing text clustering algorithms uses the frequent and relevant word sets to cluster the data sets. A common and important task that finds many applications in IR and other places There are various matching techniques present but they are retrieving data in time consuming for large number of data [1]. There are various steps used in order to indicate the matching techniques, the graph is being extracted into data models like trees and they are being used in order to represent the various set of data, they are being separated into words and that words are related or matched with the Word Net database and the resultant graph is being produced. Figure-1 represents the sample image of data clustered.

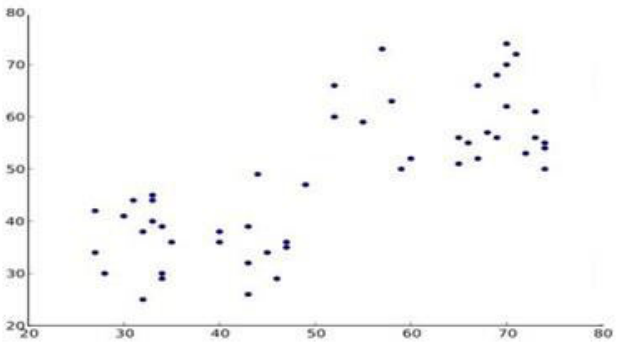

Figure-1. Sample data cluster.

The text classification of records outlines the dimensionality of the element vector which is normally extremely very high for example, 25 Newsgroups and Reuters21578 top-10, which are two certifiable datasets; both have more than 15,000 components. This high dimensionality can be a severe obstacle for classifying the data. To alleviate this difficulty, feature reduction methodologies are applied before document classification assignments are performed [2-5].The major two approaches, which are feature selection along with feature extraction, have been proposed for feature reduction. The feature extraction methodologies are more reliable than feature selection techniques, yet are all the more computationally costly. Therefore, creating scalable and proficient element extraction calculations is exceedingly requested for managing high-dimensional archive informational collections. The classical feature extraction strategies intend to change over the portrayal of the first high-dimensional informational collection into a lower-dimensional informational collection by an anticipating procedure through mathematical changes [6-7]. Each cluster is dealt with as a single new component, and, in this manner, the feature dimensionality can be radically diminished to identify the relevant terms. Matching is the technique used in order to relate or match the various set of related documents [8]. There are various matching techniques present but they are retrieving data which is time consuming for a large number of data. There are various steps used in order to indicate the matching techniques, the graph is being extracted into data models like trees and they are being used in order to represent the various set of data, they are being separated into words and that words are related or matched with the Word Net database and the resultant graph is being produced [9-11].

The weighted clustering algorithm is numerical, unsupervised, time-series data stream clustering and iterative clustering algorithm which uses distance method for clustering The distance is the ordinary distance between two points that one would measure with a ruler which in this algorithm is used for calculating distance between the centres and the

documents using K-means method[12]. Clustering contrasts from multidimensional scaling, thereby plans to delineate all the assessed objects in a way that limits the topographical identification by utilizing a few measurements as possible [13-15]. Clustering algorithms segment information into a specific number of clusters like subsets, groups and classifications.

RELATED WORK

Matching time series based data through KNIP Algorithm

To improve the searching problem, the KNIP algorithm is used to carry through rough grouping and matching. As KNIP is an algorithm for string coordinating, it is basic to transform time series into 0-1 string and rapidly look rough through similar subsequence from significant sequence and finally to reduce the measurement of raw time series information [16]. Harr wavelet change is utilized to represent the sequence to be compared and Wavelet Transformation coefficients are utilized to figure the closeness of two arrangements.

Documents clustering based on weighted cosine measures

XML data is subject to change in practical application and the algorithm used is the weighted cosine measure (WCM) which is improved from the existing algorithms, which is used to calculate the similarity between two clusters [17]. XML documents are clustered and the results of using the WCM are better than using the existing cosine measure. Whereas the XML documents in each cluster, have similar dimensions and are not frequently changed.

Clustering algorithms for an changing data sets

A three-step clustering algorithm is proposed to enhance the accuracy of clustering which presents the idea of anomaly hindsight. In rDenStream grouping, dropped small clusters are shown on outside memory which will be given new opportunity to have clustering to improve the accuracy. Tests displayed at the entering of data stream in Poisson method[18], and the outcomes over standard informational collection demonstrated its leverage over different techniques in the early period of new identification of pattern.

METHODOLOGY

Text clustering is done by only relating documents that use identical terminology. The existing and most popular bag of words representation is used for the various clustering methods [19][20]. Text document clustering assumes a vital role in providing new navigation and browsing mechanisms by sorting out a lot of data into few important clusters by organizing large amounts of information into a small number of chunks of clusters. Preferred outcomes for an assortment of temporal information clustering data are weighted and grouped and techniques can combine any information groups to produce a clustering outfit [21-23]. Currently existing text clustering arrangements just relate data and document that have distinguished terminology and the terms are defined with resources like “Word Net”. For the pre-processing the documents in text clustering we integrate the conceptual account of terms found in Word Net which is nothing but a lexical database for grouping English words into a set of synonyms called syn sets. The weighted algorithms have a frequent concept to cluster the text documents. It also uses the algorithm to utilize the semantic relationship between words to create concepts. The set of techniques used for text pre-processing are discussed as follows. The following techniques are used in the proposed work:, Load Documents (Datasets),Document Pre-processing, Text Classification ,Cluster Weighted Value, Feature Document Cluster which is represented in Figure-2.

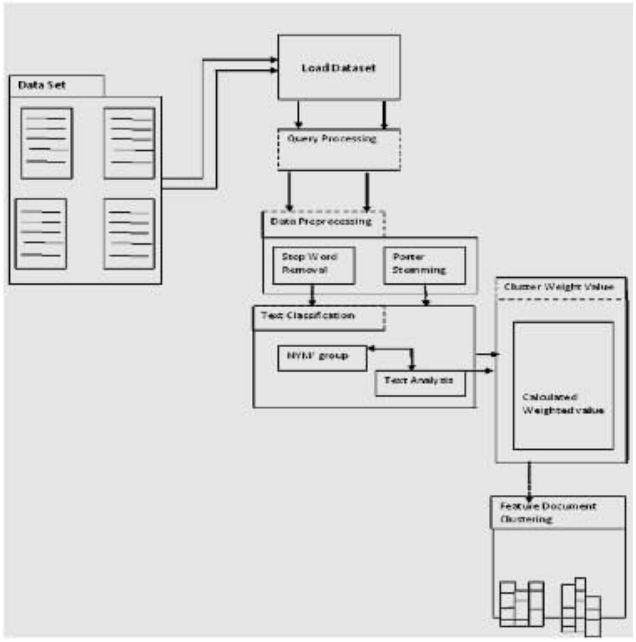

Figure-2. Overview of proposed work.

Load documents (Datasets)

An object (data record) typically has dozens of attributes and the domain for each attribute can be large data records which can be accessed by the user. The dimensionality of the feature vector is usually huge. For example, 20 Newsgroups and Reuters21578 top-10, which are two real-world data sets, both have more than 15,000 features. These standard data sets are used for the weighted cluster analysis.

Document pre-processing

Stop word removal: Stop words are words which are removed earlier to, or after, processing of the natural language data which can be represented in the text format. It totally relies on user input and is not automated. These are some of the most common, short function words, such as *the*, *is*, *at*, *which* and *on*.

Porter stemming: This algorithm uses a lookup table which has relations between different forms of data. In order to stem a word, the table is queried to find an appropriate match. When a match is found, the associated root form is returned. By incorporating the stemming algorithm in the example the words "matching", "matched", "match", and "matcher" to the root word, "match" are reduced. The similarity of the data values using cosine similarity is calculated using the following concept which involves the vector values of the data obtained in the form of terms.

$$sim(c_j,c_k)=\frac{(\vec{s}(c_j)+\vec{s}(c_k))\bullet(\vec{s}(c_j)+\vec{s}(c_k))-(|c_j|+|c_k|)}{(|c_i|+|c_k|)(|c_i|+|c_k|-1)}$$

Text classification

Nym word calculation: Words ending in nym's are often used to describe different classes of words, and the relationships between words like hypernym (a word with a generalized meaning), hyponym (a word with a specific meaning) and synonym (words with the same meaning). The term frequency retrieval based on text classification for the nym, word calculation for the data sets is computed. The text analysis is carried out using the ontology using the word net The featured results produced by the sentence-based, document-based, corpus-based, and the combined approach concept analysis have higher quality than those produced by a single-term analysis similarity. Weight clustering (WC) algorithm, which is an incremental feature clustering approach to reduce the number of features for the text classification accuracy. The following indexing techniques are used for the text clustering. The similarity of the nym words is calculated using the following method which involves the similarity values to obtain the TF/IDF which are appropriate in a specific domain. $sim(c_i,c_j)=\max_{x\in c_i,y\in c_j} sim(x,y)$ The K means approach is used to compute the distance based on the data centroids and centre of gravity.

$$\vec{\mu}(c)=\frac{1}{|c|}\sum_{x\in c}\vec{x}$$

Cluster weighted value

The cluster identifier is executed by detecting the combination of distinct weights of document terms automatically. The extracted feature corresponding to a cluster is a weighted combination of the words contained in the cluster. There are three ways of weighting, hard, soft, and mixed. Feature clustering is an efficient approach for feature reduction, which groups all features into some clusters, where features in a cluster are similar to each other. Modified Huber's Index (MHI): The proximity matrix of objects and "N" is a cluster distance matrix derived from the partition. Validity Index (DVI) is used to determine the dissimilarity metric between clusters $c_1,c_2,c_3\ldots c_n$. Normalized Mutual Information (NMI), where the proposed method is to measure the consistency between two groups of data and the amount of information shared between two groups.

The outline of the normalized mutual indexing in weighted clustering algorithm is given as follows:

Normalized Mutual Indexing in weighted clustering algorithm.
Input : Key words
Output : Indexed key word based on content, style, presentation and adjacency.
Step 1: The entered input belongs to the data clustering the clustering distribution P_c in c where c is c= {c1,c2,..cn}.
Step 2: Input the number of points n_i into ith cluster Ci and N is the summative number of points in the ith cluster and $P_c(i)= n_i N$.
Step 3: While NMI for any distribution P_{c1}=(p1,...,pn)and P_{c2}=(q1,....qn), the mutual index is $I(p,q)=\Sigma\ i,jR(i,j)\log R(i,j)\ PC1(i)PC2(j)$ *where* $R(i,j)$ *is the probability distribution*
Step 4: If indexed element in data content is equal to same word, presentation style, adjacency and data content in the cluster where threshold=0.5.
Step 5: Then retrieve key word indexed based weighted clustering.

Feature document cluster

In feature document cluster the pattern of the cluster is considered consecutively. The user is a novice and need not have any idea related to clusters. Clusters do not exist initially, and clusters can be created according to the need if necessary. For the identified word pattern, the similarity of the identified word pattern to each existing cluster is calculated to finalize if it is combined into an existing cluster else a new cluster is formed. When a new cluster is created, the associated membership function should be started. The normalized mutual indexing algorithm is used to retrieve data units which are clustered into different groups with similar semantics. Based on data content, presentation style based on font weight, style, color and text decorator, data style, and adjacency based on preceding and succeeding data the key word containing the data units is retrieved from the text node.

RESULTS AND DISCUSSIONS

In the datasets utilized for this work the Reuters dataset has a substantial number of reports per classification (around 4000) when contrasted with the other datasets. The medium size data is provided in the

Reuters dataset and lesser in the 20 newsgroup dataset. Records of same classification in Reuters dataset are in particular comparative when contrasted with reports of various classifications. In 20 newsgroups dataset numerous classes are connected thus the archives of various classifications are additionally comparative. The data sets of 20 news group have applied the techniques for the set 19,282 unique documents and 20 categories. The number of unique removing of word after text classification for nym word is computed as 51288.Similarly the data sets for Reuters 21578 has around 20,012 unique documents and 6 categories and stopping words of 40,176.The broader categories have been used for the discussion of the results based on text clustering.

The Figure-3 shows the results of using the proposed algorithm for asset of data clusters from the data set of Reuters and 20 news group. The comparative analysis shows accurate clustering of data performed using the proposed work. The entropy of the documents retrieved when the threshold value for similarity mapping is 0.5 for a set of clusters is shown for the different data sets used.

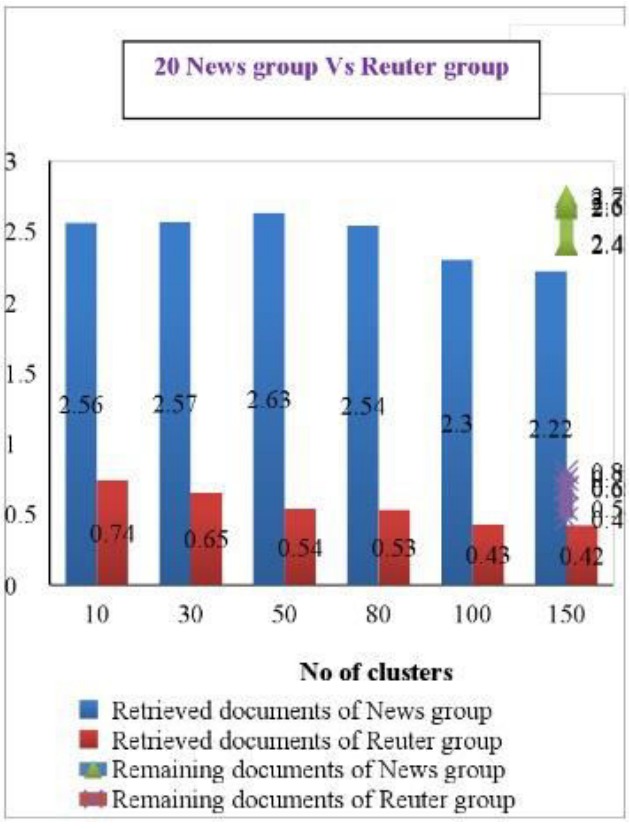

Figure-3. Results of proposed work for the data sets.

CONCLUSIONS

In this work many existing approaches and new techniques for text clustering has been discussed. Though the best in clustering algorithms is yet to be achieved the proposed word has shown better retrieval of data clusters for the given data sets. The NMI algorithm for the data sets with the given threshold has proved better results for specific domains. Finally it is inferred that however numerous calculations have been proposed for clustering however it is as yet an open issue and considering the rate at which the web is developing, for any application utilizing web reports, clustering will turn into a basic technique. In the feature based clustering, the similarity measure used is simple, and these techniques can be changed according to the applications for better and accurate results.

REFERENCES

[1] SL. Bing, W. Lam, T.L. Wong, and S. Jameel, "Web query reformulation via joint modeling of latent topic dependency and term context", ACM Transactions on Information Systems. Vol.33, No.6, 2015.

[2] L. Li, L. Zhong, Z. Yang, and M. Kitsuregawa, "Qubic: An adaptive approach to query-based recommendation", Journal of Intelligent Information Systems, Vol.40, No.3, pp. 555- 587, 2013.

[3] A. D'Ulizia, F. Ferri, A. Formica, and P. Grifoni, "Approximating geographical queries", J. Comput. Sci. Technol., Vol. 24, No. 6, pp. 1109-1124, 2009.

[4] W. Toyohide and K. Kei "Computer-supported interaction for composing documents logically". International Journal of Knowledge and Learning, Vol. 4, No. 6, pp. 509-526, 2008.

[5] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science". Parts II,III, J. Amer. Soc. Inform. Sci. Technol., Vol. 58, No. 13, pp. 2126-2144, 2007.

[6] L. T. Su, "The relevance of recall and precision in user evaluation", J. Amer. Soc. Inform. Sci., Vol. 45, No. 3, pp. 207-217, 1994.

[7] V. M. Megler and D. Maier, "Finding haystacks with needles: Ranked search for data using geospatial and temporal characteristics", in: Proc. 23rd Int. Conf. Sci. Statist. Database Management., 2011, pp. 55-72.

[8] D. Maier, V. M. Megler, A. Baptista, A. Jaramillo, C. Seaton, and P. Turner, "Navigating oceans of data",in Proc. 24th Int. Conf. Sci. Statist. Database Manage, 2012, Vol. 7338, pp. 1-19.

[9] D. R. Montello, "The measurement of cognitive distance: Methods and construct validity", J. Environ. Psychol., Vol. 11, No. 2, pp. 101-122, 1991.

[10] V. Markl, M. Kutsch, T. Tran, P. Haas, and N. Megiddo, "MAXENT: Consistent cardinality

estimation in action", in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2006, pp. 775-777.

[11] S. P. Harter, "Variations in relevance assessments and the measurement of retrieval effectiveness", J. Amer. Soc. Inform. Sci. Vol. 47, No. 1, pp. 37-49, 1996.

[12] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas, "Do user preferences and evaluation measures line up?", in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2010, pp. 555-562.

[13] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents", In: Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2002, pp. 324–330.

[14] Passier and J.T. Jeuring, "Ontology Based Feedback Generation in Design-Orientated E-Learning Systems," Proc. IADIS Int'l Conf. E-Soc., P. Isaias, P. Kommers, and M. McPherson, eds., pp. 992-996, 2004

[15] E. Voorhees and D. M. Tice, "The TREC-8 question answering track evaluation, In: Proc. 8th Text Retrieval Conf., 1999, vol. 8, pp. 83-105.

[16] G. Demartini, T. Iofciu, and A. de Vries, "Overview of the INEX 2009 entity ranking track", In: Proc. 8th Focused Retrieval Eval. 8th Int. Conf. Initiative Eval. XML Retrieval, 2010, pp. 254-264.

[17] J. G.-P. A. El Semaray, J. Edmonds, and M. Papa, "Applying data mining of fuzzy association rules to network intrusion detection," presented at the IEEE Workshop Inf., pp 122-210 United States Military Academy, West Point, NY, 2006.

[18] Li Wei, College of Computer Science and Technology, Jilin University, China autumnal_mood@163.com "XML Documents Clustering Research Based on Weighted Cosine Measure", pp 221-234, 2010

[19] J. Luo, "Integrating fuzzy logic with data mining methods for intrusion detection," Master's thesis, Dept. Comput. Sci., Mississippi State Univ., Starkville, MS, pp 123-155, 1999.

[20] M. .E. Voorhees, " On test collections for adaptive information retrieval, Information Processing and Management ,Vol. 44 , No. 4, pp. 1879-1885, 2008.

[21] S. Tsumoto, "Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model", Information Sciences, Vol. 162 , No. 2 , pp. 65- 80. 2004.

[22] M. Ricardo, V. Joao and D. Dulce, "Modelling and learning controlled flexibility in software processes", International Journal of Knowledge and Learning, Vol. 5, No. 5, pp. 423-444, 2009.

[23] J.Mostow, J. and J. Beck. Some Useful Tactics To Modify, Map and Mine from Data from Intelligent Tutors, Natural Language Engineering, Vol. 12, No. 2, pp. 195-208, 2006.