Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

Document Clustering with Feature Behavior based Distance Analysis

A. Kanimozhi., M.C.A.[1] , M. Subha., M.Sc., M.Phil., M.C.A., (Ph.D)., [2]

[1](M.Phil., Research Scholar), Department of Computer Science,
Kaamadhenu Arts and Science College, Sathyamangalam.

[2] Assistant Professor, Department of Computer Science,
Kaamadhenu Arts and Science College, Sathyamangalam.

Abstract—Machine learning and data mining methods are applied to perform large data analysis. Clustering methods are applied to group the related data values. Partitional clustering and hierarchical clustering methods are applied to handle the clustering operations. Tabular format data processing is carried out under the partitional clustering models. Tree based data clustering is adapted in the hierarchical clustering models. Clustering techniques are also applied to group the text documents. Distance measures are employed to estimate the document relationships in clustering process. Cosine and Euclidean distance measures are widely used in the clustering operations. Dimensionality is the key factor in the document clustering process. Document contents are parsed and represented as vector model. Features and associated weight values are assigned under the document vector model. Feature behavior distance model faces the High dimensionality and sparsity issues. Feature based similarity estimation is carried out using Similarity Measurement for Text Process (SMTP). Clustering and classification operations are performed with the SMTP distance measure. Text document clustering is performed using the Hybrid Similarity Measure for Text Process (HSMTP). Feature appearance and weight factors are integrated in the HSMTP scheme. The HSMTP scheme is integrated with the Spherical K-Means clustering algorithm to partition the documents. Feature reduction process is initiated to minimize the dimensionality of the document vector. Ontology is used to fetch the concept relationship values. Concept relationship based distance model is also supported by the HSMTP scheme.

I. INTRODUCTION

Texts are themselves linguistic expressions and if the desired information were contained in a single text it would be entirely reasonable to simply find that text. The requirement that the patterns be novel can be omitted from the definition of TDM for the same reason. The Data mining step includes model selection, transformation of the data into a format appropriate for the selected model and choice of a method for finding the appropriate parameters for that model. At a very general level, data mining models can be grouped by what they seek to produce. Functions traditionally associated with data mining include clustering models, classification models and dependency models.

Viewed abstractly, a model is simply a parameterized set of ways in which the data can be used to produce a result, a criterion that describes which results are preferred and an algorithm for searching the space of possible combinations. The output of the data-mining step is the model and the parameters that have been found. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait - often proximity according to some defined distance measure.

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Besides the term data clustering, there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis.

A TDM system is designed to identify topically related news stories. A clustering model would be well suited to this task, since clusters could be used to represent topical relationships. If the data storage system contains the frequency of each word appearing in the title, a simple data transformation could reduce this to a Boolean value reflecting the presence or absence of each word. The parameters of a clustering model are the number of categories and the category to which each text is assigned.

One simple preference criterion is to assign two texts to the same category if their titles share at least half the words in the shorter title. In this case, a greedy agglomerative clustering algorithm that sequentially makes pair-wise comparisons and either forms or joins clusters will quickly discover the optimal parameters for the model. Better clustering models can be designed for this task, but the key parts of the data-mining step are illustrated by this example.

1.1 Objective

To perform document clustering with Similarity Measurement for Text Processing (SMTP) and Hybrid SMTP schemes, estimate the term weights using the statistical relationships, use the Ontology for concept relation based weight estimation process and perform document clustering using C.

II. RELATED WORKS

There are many other graph partitioning methods with different cutting strategies and criterion functions, such as Average Weight and Normalized Cut, all of which have been successfully applied for document clustering using cosine as the pairwise similarity score [5]. An empirical study was conducted to compare a variety of criterion functions for document clustering.

Another popular graph-based clustering technique is implemented in a software package called CLUTO. This method first models the documents with a nearest neighbor graph, and then splits the graph into clusters using a min-cut algorithm. Besides cosine measure, the extended Jaccard coefficient can also be used in this method to represent similarity between nearest documents. Given non unit document vectors u_i, u_j (d_i = ui/||u_i||, d_j = uj/||u_j||), their extended Jaccard coefficient is

$$Sim_{\in Jacc}(u_i, u_j) = \frac{u_i^t, u_j}{\|u_i\|^2 + \|u_j\|^2 - u_i^t u_j} \quad (1)$$

Compared with euclidean distance and cosine similarity, the extended Jaccard coefficient takes into account both the magnitude and the direction of the document vectors. If the documents are instead represented by their corresponding unit vectors, this measure has the same effect as cosine similarity. Strehl et al. compared four measures: euclidean, cosine, Pearson correlation, and extended Jaccard, and concluded that cosine and extended Jaccard are the best ones on web documents.

In nearest neighbor graph clustering methods, such as the CLUTO's graph method above, the concept of similarity is somewhat different from the previously discussed methods. Two documents may have a certain value of cosine similarity, but if neither of them is in the other one's neighborhood, they have no connection between them. In such a case, some context-based knowledge or relativeness property is already taken into account when considering similarity. Ahmadand Dey [9] proposed a method to compute distance between two categorical values of an attribute based on their relationship with all other attributes.

Subsequently, Ienco et al. [10] introduced a similar context-based distance learning method for categorical data. For a given attribute, they only selected a relevant subset of attributes from the whole attribute set to use as the context for calculating distance between its two values. More related to text data, there are phrase-based and concept-based document similarities. Lakkaraju et al. [12] employed a conceptual tree-similarity measure to identify similar documents.

This method requires representing documents as concept trees with the help of a classifier. For clustering, Chim and Deng [1] proposed a phrase based document similarity by combining suffix tree model and vector space model. They then used Hierarchical Agglomerative Clustering algorithm to perform the clustering task.

There are also measures designed specifically for capturing structural similarity among XML documents. They are essentially different from the document-content measures that are discussed in this paper.

In general, cosine similarity still remains as the most popular measure because of its simple interpretation and easy computation, though its effectiveness is yet fairly limited. We propose a novel way to evaluate similarity between documents, and consequently formulate new criterion functions for document clustering.

III. DOCUMENT CLUSTERING PROCESS

Text processing plays an important role in information retrieval, data mining, and web search. In text processing, the bag-of-words model is commonly used. A document is usually represented as a vector in which each component indicates the value of the corresponding feature in the document.

The feature value can be term frequency, relative term frequency or term frequency and inverse document frequency (tf-idf) [5]. The dimensionality of a document is large and the resulting vector is sparse, i.e., most of the feature values in the vector are zero. Such high dimensionality and sparsity can be a severe challenge for similarity measure which is an important operation in text processing algorithms [1], [6], [7].

A lot of measures have been proposed for computing the similarity between two vectors. The Kullback-Leibler divergence is a non-symmetric measure of the difference between the probability distributions associated with the two vectors. Euclidean distance [3] is a well-known similarity metric taken from the Euclidean geometry field. Manhattan distance, similar to Euclidean distance and also known as the taxicab metric, is another similarity metric.

The Canberra distance metric is used in situa- tions where elements in a vector are always non-negative. Cosine similarity [5] is a measure taking the cosine of the angle between two vectors. The Bray-Curtis similarity measure is a city-block metric which is sensitive to outlying values.

The Jaccard coefficient is a statistic used for comparing the similarity of two sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. The Hamming distance between two vectors is the number of positions at which the corresponding symbols are different. The extended Jaccard coefficient and the Dice coefficient [11] retain the sparsity property of the cosine similarity measure while allowing discrimination of collinear vectors. An information-theoretic measure for document similarity, named IT-Sim. Chim et al. [1] proposed a phrase-based measure to compute the similarity based on the Suffix Tree Document (STD) model.

Similarity measures have been extensively used in text classification and clustering algorithms. The spherical k-means algorithm introduced by Dhillon and Modha adopted the cosine similarity measure for document clustering. Zhao and Karypis reported results of clustering experiments with 7 clustering algorithms and 12 different text data sets, and concluded 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".

D'hondt et al. [8] adopted a cosine-based pairwise adaptive similarity for document clustering. Zhang et al. [4] used cosine to calculate a correlation similarity between two projected documents in a low-dimensional semantic space and performed document clustering in the correlation similarity measure space. Kogan et al. proposed a two step clustering procedure in which the sPDDP is used to generate initial partitions in the first step and a k-means clustering algorithm using the Kullback-Leibler divergence is applied in the second step. Dhillon et al. proposed a divisive information-theoretic feature clustering algorithm for text classification using the Kullback-Leibler divergence.

Euclidean distance is usually the default choice of similarity-based methods, e.g. k-NN and k-means algorithms. Kogan et al. combined squared Euclidean distance with relative entropy in a k-means like clustering algorithm. Chim et al. [1] performed document clustering based on the proposed phrase-based similarity measure. The extended Jaccard coefficient can be used for document data and it reduces to the Jaccard cofficient in the case of binary attributes.

We propose a new measure for computing the similarity between two documents. Several characteristics are embedded in this measure. It is a symmetric measure. 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. An absent feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The measure is applied in several text applications, including single label classification, multi-label classification, k-means like clustering, and hierarchical agglomerative clustering, and the results obtained demonstrate the effectiveness of the proposed similarity measure.

IV. SIMILARITY MEASURE FOR TEXT PROCESS (SMTP)

Let a document d with m features $w_{1,}$ w_2, . . . , $w_{m,}$ be represented as an m-dimensional vector, i.e., d = < d_1, d_2, . . . , d_m >. If w_i, $1 \leq i \leq m$, is absent in the document, then d_i= 0. Otherwise, d_i> 0. The following properties, among other ones, are preferable for a similarity measure between two documents:

1) The presence or absence of a feature is more essential than the difference between the two values associated with a present feature. Consider two features w_i and w_j and two documents d_1 and d_2. Suppose w_i does not appear in d_1 but it appears in d_2. Then w_i is considered to have no relationship with d_1 while it has some relationship with d_2. In this case, d_1 and d_2 are dissimilar in terms of w_i. If w_j appears in both d_1 and $d_{2.}$ Then w_{jj} has some relationship with d_1 and $d_{2 \text{ simultaneously}}$. In this case, d_1 and d_2 are similar to some degree in terms of w_i.

For the above two cases, it is reasonable to say that w_i carries more weight than w_i in determining the similarity degree between d_1 and d_2. For example, assume that w_i is absent in d_1 i.e., $d_{1i} = 0$, but appears in d_2, e.g., $d_{2i} = 2$, and w_j appears both in d_1 and d_2, e.g., $d_{1j} = 3$ and $d_{2j} = 5$. Then w_i is considered to be more essential than w_j in determining the similarity between d_1 and d_2, although the differences of the feature values in both cases are the same.

2) The similarity degree should increase when the difference between two non-zero values of a specific feature decreases. The similarity involved with d_{13}= 2 and d_{23} = 20 should be smaller than that involved with d_{13}= 2 and d_{23} = 3.

3) The similarity degree should decrease when the number of presence-absence features increases. For a presence-absence feature of d_1 and d_2, d_1 and d_2 are dissimilar in terms of this feature as commented earlier.

4) Two documents are least similar to each other if none of the features have non-zero values in both documents.

5) The similarity measure should be symmetric. That is, the similarity degree between d_1 and d_2 should be the same as that between d_2 and d_1.

6) The value distribution of a feature is considered, i.e., the standard deviation of the feature is taken into account, for its contribution to the similarity between two documents. A feature with a larger spread offers more contribution to the similarity between d_1 and d_2.

Based on the preferable properties mentioned above, we propose a similarity measure, called SMTP (Similarity Measure for Text Processing), for two documents $d_1 = < d_{11}, d_{12}, . . . , d_{1m} >$ and $d_2 = < d_{21}, d_{22}, . . . , d_{2m} >$. Define a function F as follows:

$$F(d_1, d_2) = \frac{\sum_{j=1}^{m} N*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N \cup (d_{1j}, d_{2j})} \qquad (2)$$

where

The proposed measure takes into account the following three cases: a) The feature considered appears in both documents, b) the feature considered appears in only one document and c) the feature considered appears in none of the documents. For the first case, we set a lower bound 0.5 and decrease the similarity as the difference between the feature values of the two documents increases, scaled by a Gaussian function as where σ_j is the standard deviation of all non-zero values for feature w_j in the training data set. For the second case, we set a negative constant $-\lambda$ disregarding the magnitude of the non-zero feature value. For the last case, the feature has no contribution to the similarity.

V. PROBLEM STATEMENT

A document is represented as a vector. In document vector each component indicates the value of the corresponding feature in the document. The feature value can be term frequency, relative term frequency. High dimensionality and sparsity can be a severe challenge for similarity measure. Similarity Measurement for Text Process (SMTP) is used to compute the similarity between two documents with respect to a feature. Presents and options of the features in both documents are used to estimate the similarity values.

The SMTP is extended to estimate similarity between two set of documents. The SMTP scheme is used with text clustering and classification task. The following problems are identified in the text similarity analysis.

- Term relationships are not considered
- Dimensionality reduction is not performed
- Future weights are not used in the similarity estimation
- Limited clustering accuracy

VI. FEATURE BEHAVIOR BASED DISTANCE ANALYSIS FOR CLUSTERING

6.1 Spherical K-Means Clustering

The text document clustering process is performed with Spherical K means Clustering algorithm. The Spherical K means Clustering algorithm is composed by integrating the K-means clustering algorithm with Cosine similarity measure. Term weights are used in the similarity estimation process. Term weights are integrated with occurrence based similarity model. The system also uses the semantic relationship values. Relatively efficient: O(tkn), where n is # of objects, k is # of clusters, and t is # of iterations. Normally, k, t << n. Often terminates at a local optimum.

K-Means algorithm is applicable only when mean is defined. Need to specify k, the number of clusters, in advance. K-means algorithm is unable to handle noisy data and outliers. Reallocation methods - start with an initial assignment of items to clusters and then move items from cluster to cluster to obtain an improved partitioning. Single pass method - simple and efficient, but produces large clusters, and depends on order in which items are processed. The basic algorithm:

1. Select K data points as the initial representatives.
2. For i = 1 to N, assign item xi to the most similar centroid.
3. For j = 1 to K, recalculate the cluster centroid Cj.
4. Repeat steps 2 and 3 until there is no change in clusters.

Example: Clustering Terms

Initial assignment:
C1 = {T1, T2},
C2 = {T3, T4},
C3 = {T5, T6}
Cluster Centroids :
Example: K-Means

Now using simple similarity measure, compute the new cluster-term similarity matrix. Now compute new cluster centroids using the original document-term matrix. The process is repeated until no further changes are made to the clusters.

The most popular distance metrics for text clustering is the cosine of the angle between the two vectors given by the following formula:

$$\cos\Theta = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^{2}} * \sqrt{\sum_{i=1}^{N} y_i^{2}}}$$

where θ refers to the angle between two vectors. For example, the cosine between two vectors (1, 2, 3) and (3, 5, 7) is (1*3 +2*5+3*7)/sqrt(1+4+9)* sqrt (9+25+49). This ratio defines the cosine angle between the vectors, with values between 0 and 1. As the angle between vectors lessens the Cosine angle approaches to 1 i. e when angle becomes 0 it will be 1.

VII. SEMANTIC ANALYSIS BASED DOCUMENT CLUSTERING WITH HYBRID SIMILARITY MEASURE

The system is designed to perform document clustering using Similarity Measurement for Text Process (SMTP). Spherical K means algorithm is used for the clustering process. Concept relationships are identified with the support of ontology. Dimensionality reduction functions are applied to minimize the features. The system is designed to cluster the text documents with statistical and conceptual relationships. Weight values are integrated in the presence and absence based similarity analysis mechanism. Infrequent features are removed in dimensionality reduction process. The system is divided into six major modules. They are document preprocess, term analysis, semantic analysis, distance analysis with SMTP, distance analysis with HSMTP and clustering process.

Document parsing, stopword elimination and stemming process are carried out under document preprocess module. Term weights are estimated under term analysis module. Semantic analysis is performed to identify concept relationships. Appearance based relationship analysis is performed in Distance analysis with SMTP module. Distance analysis with HSMTP module is designed to estimate the similarity using appearance and influence values. The clustering process module is used to partition the document collection based on the similarity values.

Ontology is a specification of a conceptualization. It refers to the subject of existence. In the context of knowledge sharing, the term ontology is used to mean a specification of a conceptualization. That is, ontology is a description of the concepts and relationships that can exist for an agent or a community of agents.

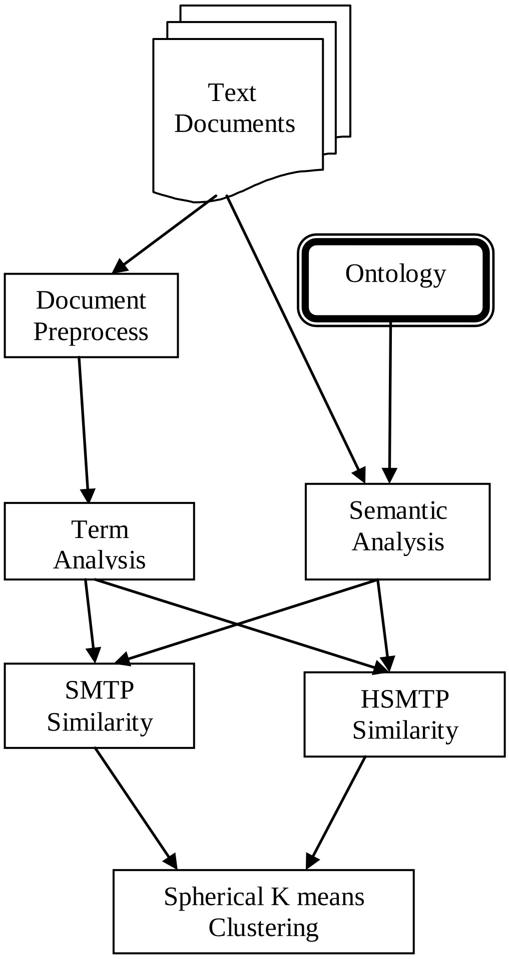

Fig.: 7.1. Semantic Analysis based Document Clustering

7.1. Document Preprocess

Document preprocess is performed to parse the text documents into words. Document cleaning is applied to remove stop words. Stemming process is applied to detect the base term. Terms are updated with their frequency values.

7.2. Term Analysis

The term analysis is performed to estimate the term weight values. Statistical method is used for the term weight estimation process. Term frequency (TF) and Inverse Document Frequency (IDF) are used for the term weight estimation process. Dimensionality reduction process is carried out to remove infrequent feature values.

7.3. Semantic Analysis

The semantic analysis is performed to identify the concept relationships. Ontology is constructed for the selected domains. Terms and associated concept relationships are identified using the Ontology. Semantic weights are assigned with reference to the concept relationship type.

7.4. Distance Analysis with SMTP

Documents relationship is estimated using Similarity Measurement for Text Process (SMTP) scheme. Feature appearance behavior is used in the SMTP scheme. Distance values are estimated with presence and absence of features. Term features and semantic features are used in the similarity estimation process.

7.5. Distance Analysis with HSMTP

Hybrid Similarity Measure for Text Process (HSMTP) mechanism is used for the distance analysis process. Feature appearance and feature influence details are integrated in the distance estimation process. Feature influence weight value is used in the distance estimation process. Distance analysis is carried out with the term and semantic features.

7.6. Clustering Process

The document clustering is performed using Spherical K means clustering algorithm. The clustering process is carried out with user specified cluster count values. Term weight and semantic weight values are used in the clustering process. SMTP and HSMTP measures are used for the distance analysis between the text documents.

VIII. CONCLUSION

Similarity measurement is used to estimate the relationship between the records or documents. Similarity Measurement for Text Process (SMTP) scheme is used to estimate the distance values. Spherical K means algorithm is used for text document clustering process. Statistical weight and concept weight models are used to improve the similarity measurement process.

Concept relationship based similarity analysis model is adopted for the clustering process. Efficient similarity computation mechanism produces document relationships. Clustering accuracy is improved in the system. The system reduces the process time and memory requirement. In Future the system can be enhanced with incremental mining model. The text document clustering scheme can be improved to cluster the XML documents and web documents.

REFERENCES

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

[2] C. G. González, Jr.and A. L. V. Rodrigues, "Density of Closed Balls In Real-Valued And Autometrized Boolean Spaces For Clustering Applications," in Proc. 19th Brazilian Symp. Artif. Intell., Savador, Brazil, 2008, pp. 8–22.

[3] T. W. Schoenharl and G. Madey, "Evaluation of Measurement Techniques For The Validation Of Agent-Based Simulations Against Streaming Data," in Proc. ICCS, Kraków, Poland, 2008, pp. 6-15.

[4] T. Zhang, Y. Y. Tang and Y. Xiang, "Document Clustering In Correlation Similarity Measure Space," IEEE Trans.Knowl.Data Eng., Jun.2012,pp. 1002-1013.

[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, Elsevier, 2006, pp. 614-624.

[6] K. M. Hammouda and M. S. Kamel, "Hierarchically Distributed Peer-To-Peer Document Clustering And Cluster Summarization," IEEE Trans. Knowl. Data Eng., vol. 21, no. 5, May 2009, pp. 681-698.

[7] C. Silva, U. Lotric, B. Ribeiro and A. Dobnikar, "Distributed Text Classification With An Ensemble Kernel-Based Learning Approach," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., May 2010, pp. 287-297.

[8] J. D'hondt, "Pairwise-Adaptive Dissimilarity Measure For Document Clustering," Inf. Sci., 2010.

[9] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," Pattern Recognit. Lett., vol. 28, no. 1, pp. 110 – 118, 2007.

[10] D. Ienco, R.G. Pensa, and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Proc.Eighth Int'l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.

[11] M. Steinbach and V. Kumar, Introduction to Data Mining. Boston, MA, USA: Addision-Wesley, 2006.

[12] P. Lakkaraju, S. Gauch, and M. Speretta, "Document Similarity Based on Concept Tree Distance," Proc. 19th ACM Conf. Hypertext andHypermedia, pp. 127-132, 2008.