ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Document Classification and Clustering using Feature Extraction for Similarity Measure

Somnath Melasagare[1], Vikas Thombre[2]

P.G. Student, Department of Computer Engineering, SKNSIT's College, Lonavala, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, SKNSIT's College, Lonavala, Maharashtra, India[2]

ABSTRACT: The similarity between documents are the new innovative concept now days in data mining and information retrieval. Standard text similarity measures perform poorly because of data sparseness and the lack of context. In text processing, bag of words model is used. Measuring the similarity between documents is a main task in the document processing and text classification. To compute the similarity between two documents with respect to a feature, the proposed measure takes the following cases into account a) The feature present in both documents, b) the feature present in only one document, and c) the feature that appears in none of the documents. For first one, similarity increases as the difference between the two involved feature values decreases. For second, a fixed value is contributes to the similarity. For last, the feature has no contribution to similarity. The effectiveness of measure will be evaluated on several real world data sets for document classification and clustering. In context of clustering and retrieval of text data, system provides a fine classification of data set. This system performs creation of flat clusters, pre-processing (text feature extraction) and data point representation (normalization, filtering, tokenization etc.) This system achieves reduction in noisy data that helps for more accurate classification of data sets as flat clusters. Use of hierarchical clustering in this system gives high level accuracy and classification of data.

KEYWORDS: Document Classification, Feature Extraction, Document Clustering, Clustering Algorithm.

I. INTRODUCTION

Today in many fields, the amount of available information is growing continuously. This growth is not limited just as an increase of the available knowledge. For usability of such large bulk of information, different methods and tools are used to filter and process the data. Now days, growth of such information primarily hails from internet, which is a main source of information retrieval. Obviously, controlling or filtering such is a big issue. Current information retrieval method is based over similarity of input query, which gives every possible output which matches with that query. Another issue in this case, can be integrity of searched information.

Document or Text clustering is an automatic document organization, topic extraction or filtering[2]. It is closely related to data clustering. Clustering methods can be used to group the retrieved documents into a list of meaningful categories. clustering involves use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.

The feature value can be term frequency (the number of occurrences of a term appearing in the document), relative term frequency (the ratio between the term frequency and the total number of occurrences of all the terms in the document set), or tf-idf (a combination of term frequency and inverse document frequency). Usually, 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[9].

Similarity measures extensively used in text classification and clustering. The spherical k means algorithm introduced by Dhillon and Modha [10] adopted the cosine similarity measure for document clustering. D'hondt et al. [3] adopted a cosine-based pairwise adaptive similarity for document clustering. 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

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

relative entropy in a kmeans like clustering algorithm. Chim et al. [5] 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 coefficient in the case of binary attributes.

One approach to differentiate this bulk of information based on their domain is document similarity. This approach is being used since last 20 years in the field of computer science and artificial intelligence. Similarity is a formal description of concepts in a domain of classes, properties of each content describing various features and attributes of concepts and restrictions on slots. Classes describe contents in the domain and slots describe properties of classes.

Clustering is task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. It is a common technique of statistical data analysis. There are several algorithms used for clustering techniques according to type of cluster, also includes no. of algorithm categories. E.g. Affinity propagation, fuzzy clustering, means shift, k means++, hierarchical clustering etc.

Paper is organized as follows. Section II describes the related work.After related work, how system works is explained using system architecture technique that is given in Section III where the number of phases represents the flow of the system. Section IV and V presents algorithm implementation and experimental results. Finally, Section VI presents conclusion.

II. RELATED WORK

Various measures which have been popularly known for computing the similarity between documents are briefly presented here. Let d1 and d2 be two documents represented as vectors.

1. The Euclidean distance or Euclidean metric is the ordinary (i.e. straight-line) distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.
   The Euclidean distance measure is defined as the root of square differences between the respective coordinates of d1 and d2, i.e.,
   $$dEuc(d1, d2) = [(d1 - d2).(d1 - d2)]^{1/2} \quad (1)$$

where A.B denotes the inner product of the two vectors A and B.

2. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle.
   In Cosine similarity [4] measures the cosine of the angle between d1 and d2 as follows:
   $$Scos(d1, d2) = \frac{d1.d2}{(d1.d1)^{1/2}(d2.d2)^{1/2}} \quad (2)$$
   As a prominent example, cosine similarity is often used in information retrieval, within the Vector Space Model, in which a document (i.e. a piece of text) is represented as a vector of all its terms (words), by using term frequency (or tf-idf, or variants thereof). In that model, it is beneficial to abstract out the magnitude of the term vector because it takes out the influence of the document length: only the relative frequencies between words in the document, and across documents, are of importance, and not how big the document is. This enables that e.g. in computation of document similarity big documents do not always come on top.
3. In Pairwise-adaptive similarity [3] dynamically selects a number of features out of d1 and d2 and is defined to be
   $$dpair(d1, d2) = \frac{d_{1,k}.d_{2,k}}{(d_{1,k}.\ d_{1,k})^{1/2}(d_{2,k}.\ d_{2,k})^{1/2}} \quad (3)$$
   wheredi,K is a subset of di, i = 1, 2, containing the values of the features which are the union of the K largest features appearing in d1 and d2, respectively.
4. Hotho, Staab, Stumme [11] proposed an integration of core ontologies as background knowledge into the process of clustering text documents. As text document clustering plays key role in providing intuitive navigation and browsing mechanisms by organizing large sets of documents into a small number of meaningful clusters. In this paper they propose that ontologies improve text document clustering.
5. The use of hierarchical clustering for learning the ontologies in recommendation systems has proposed by Vincent Schickel Zuber and BoiFaltings [12]. Since ontologies are being successfully used to overcome semantic heterogeneity, and are becoming fundamental elements of the Semantic Web, This paper show that it is possible to learn ontologies autonomously by using clustering algorithms.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

III. SYSTEM ARCHITECTURE

Proposed system consists of eight such phases which define system overall architecture, these are as follows,
Phase 1: Data set collection
In this stage, the collection of available data sets can be obtained via internet or if required, data sets for proposed field's i.e. books and news can be created. The data sets should be in standard formats which are widely available. For e.g. books are available in word, pdf, eBook format.
Phase 2: Preprocessing
This stage will consist cleaning of collected data. This includes removing of noisy data. Noise is an error or variance in measured variable such as unnecessary information, for e.g. missing values, inconsistent data etc. There are many procedures which cleans the noisy data. Some of them are Binning method, clustering, combined human and computer inspection, regression etc.
Phase 3: Feature extraction
The process of creating features for a given learning or classification instance is called feature extraction. We are going to use text features for better data representation. Some of the text features are chart/table, appendix, fonts, diagram, graphs, and maps and so on. In this context, approach to be followed can be like: selecting features, loading features and then extracting them. There are various text feature extraction systems already available in text mining.Respective algorithms are meant to be used according to java environment.
Phase 4: Data point representation
In this stage, extracted features should be converted in previous stage into input from suitable for clustering, since clustering algorithm needs information in numeric form. So the extracted information will be converted to numeric form. K means clustering or any similar method will be very useful here. Because k means clustering is one of unsupervised learning technique used to understand the underlying structure of datasets.
Phase 5: clustering
Clustering is the most common form of unsupervised learning. This stage will do actual clustering over the extracted information. This will create flat clusters. Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other.

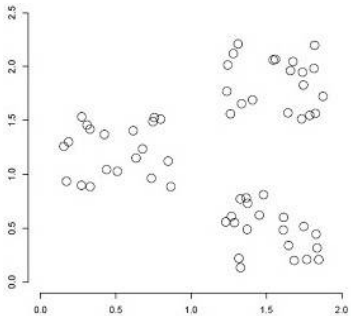

Figure 1.1 An example of data set with a clear cluster structure

Flat clustering is efficient and conceptually simple for primary clustering. Figure 1.1 shows an example of cluster data sets.

Phase 6: Hierarchical clustering
Hierarchical clustering creates a hierarchy of clusters. So in this stage cluster tree will be built for each cluster (flat). For example, one flat cluster includes a cluster named documents, then its sample cluster tree will be somewhat like figure1.2.Hierarchy is a structure that is more informative than the unstructured set of clusters returned by flat clustering. Some popular algorithms in this category are BIRCH, CURE, ROCK, HFRECCA, Chameleon etc.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

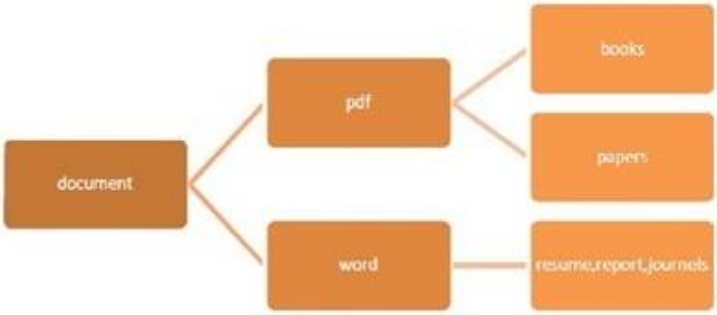

Figure 1.2 Sample Cluster tree for flat cluster documents

Phase 7: Input Query
In this system, a query will be provided to obtain required information retrieval. User can upload customized data set or any standard data set as an input.
Phase 8: Data retrieval
This is a final stage where representation of system is delivered to the user. In this stage where user will get result in form of flat clusters. Data set itself will contain folder of newly created flat clusters. Flat clusters are classified into hierarchy of themselves that will be contained in same folder.

IV. ALGORITHM AND IMPLEMENTATION

In the proposed system, a basic feature extraction algorithm is used to extract features of the data. Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers that, unlike NB, are expensive to train. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature is one that, when added to the document representation, increases the classification error on new data.

Select features (D, c, k)
1. V ← Extract Vocabulary (D)
2. L ← []
3. **for each** t ϵ V
4. **Do** A (t, c) ← ComputeFeatureUtility (D, t, c)
5. **Append** (L, <A (t, c), t>)
6. **Return** FeaturesWithLargestValues (L, k)

Algorithm 2.1 Basic Feature Selection Algorithm

For a given class c, we compute a utility measure A (t, c) for each term of the vocabulary and select the k terms that have the highest values of A (t, c). All other terms are discarded and not used in classification. We will introduce three different utility measures in this section: mutual information, A (t, c)= I (Ut, Cc) ; the X^2 Test, A(t, c) = X^2 (t, c) ; and Frequency , A(t, c)=N(t, c).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Feature selection for multiple classifiers

In an operational system with a large number of classifiers, it is desirable to select a single set of features instead of a different one for each classifier. One way of doing this is to compute the X^2 statistic for an n×2 table where the columns are occurrence and nonoccurrence of the term and each row corresponds to one of the classes. We can then select the k terms with the highest X^2 statistic as before.

More commonly, feature selection statistics are first computed separately for each class on the two-class classification task c versus ĉ and then combined. One combination method computes a single figure of merit for each feature, for example, by averaging the values of A (t, c) for feature t, and then selects the k features with highest figures of merit. Another frequently used combination method selects the top k/n features for each of n classifiers and then combines these n sets into one global feature set.

Classification accuracy often decreases when selecting k common features for a system with n classifiers as opposed to n different sets of size k. But even if it does, the gain in efficiency owing to a common document representation may be worth the loss in accuracy.

V. EXPERIMENTAL RESULTS

In this system, we examine how the different strategies solve the Problem, which is the core focus of our work.

Cluster accuracy of implemented system is compared with existing (old) system. The graph on Figure no.1.3 shows implemented system is more accurate than existing system though we increase no. of files in dataset.

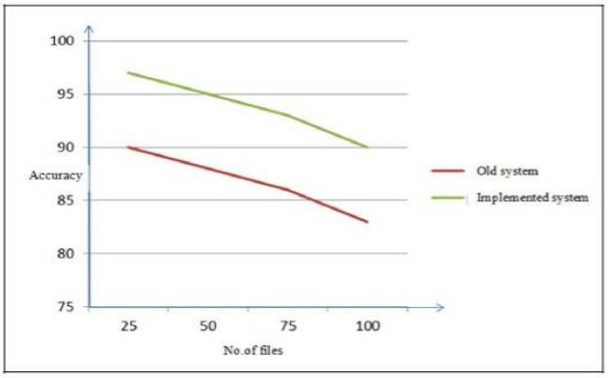

Figure 1.3 Accuracy Graph

Various data classification techniques are compared with this system. This comparison is made on the basis of data type compatibility. Table no. 3.1 shows comparison of NEWSD, Uclassify and BDS (business document system) with new system. NEWSD is the news classification system. BDS classifies business related documents such as pdf and word. Uclassify is the system that classifies text data.

type	NEWSD	BDS	Uclassify	Implemented system
Webpages	✓	X	X	✓
Text	X	X	✓	✓
PDF	X	X	X	✓
Word	X	✓	✓	✓

Table 3.1 Compatibility Table

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Time required for retrieval of each type of document differs from each other. Files in documents have different sizes, though the files have same no. of pages. Figure no. 1.4 shows graph with the variations in time required for data retrieval. For this analysis, 50 files are taken from each document types viz. pdf, word, text and webpages. as a different data sets.

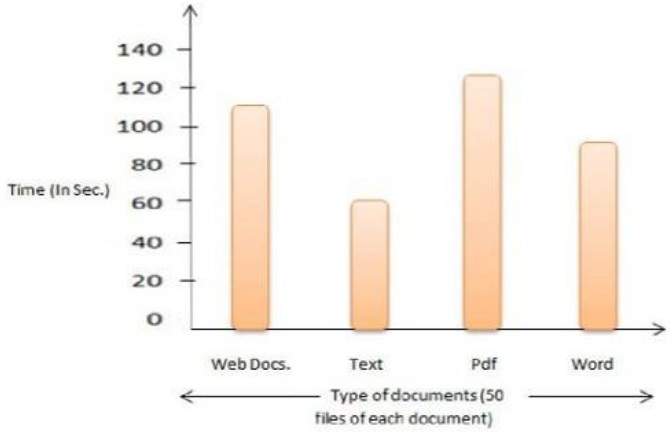

Figure 1.4 Time requirement for various document types

VI. CONCLUSION

We have proposed system which is helpful to summarize and classify the given data set.In context of clustering and retrieval of text data, system provides a fine classification of data set.This system performs creation of flat clusters, pre-processing (text feature extraction) and data point representation (normalization, filtering, tokenization etc.) This system achieves reduction in noisy data that helps for more accurate classification of data sets as flat clusters.Use of hierarchical clustering in this system gives high level accuracy and classification of data.The advantages of document similarity and hierarchical clustering are helpful for increasing processing speed and retrieval of information.

VII. ACKNOWLEDGMENT

We would like to thank IJIRSET for giving such wonderful and useful platform for the PG students to publish their research work. Also would like to thanks to our guide & respected teachers for their constant support and motivation for us. Our sincere thanks to SKN Sinhgad Institute of Technology and Science for providing a strong platform to develop our skill and capabilities.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Trans. Knowl. Data Eng., VOL. 26, no. 7, July 2014.
[2] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing", IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.
[3] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering",. Sci., vol. 180, no. 12, pp. 2341–2358, 2010.
[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.
[5] 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.
[6] A. Strehl and J. Ghosh, "Value-based customer grouping from large retail data-sets", in Proc. SPIE, vol. 4057. Orlando, FL, USA, Apr. 2000, pp. 33–42.
[7] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, "Document clustering in correlation similarity measure space", IEEE Trans. Knowl. Data Eng., vol. 24, no. 6, pp. 1002–1013, Jun. 2012.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

[8] D. Renukadevi , S. Sumathi , "TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING USING FUZZY C-MEANS ALGORITHM", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 4, April 2014.
[9] P. K. Agarwal and C. M. Procopiuc, "Exact and approximation algorithms for clustering", in Proc. 9th Annu. SODA, Philadelphia, PA, USA, 1998, pp. 658–667.
[10] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering", Mach. Learn., vol. 42, no. 1, pp. 143–175, 2001.
[11] Andreas Hotho, Steffen Staab, GerdStumme, "Text Clustering Based on Background Knowledge",Institute of Applied Informatics and Formal Description Methods AIFB, University of Karlsruhe, D–76128 Karlsruhe, Germany.
[12] Vincent Schickel-Zuber and BoiFaltings, "OSS: A Semantic Similarity Function based on Hierarchical Ontologies", Swiss Federal Institute of Technology - EPFL Artificial Intelligence Laboratory.