Abstract — In today’s world of internet, with whole lot of e-documents such as html pages, digital libraries etc. Occupying considerable amount of cyber space, organizing these documents has become a practical need. Clustering is an important technique that organizes large number of objects into smaller coherent groups. This helps in efficient and effective use of these documents for information retrieval and other NLP tasks. Clustering algorithm requires a metric to quantify how similar or different two given documents are. This difference is often measured by some distance measures such as Euclidian distance, Cosine Similarity, Jaccard Co-efficient to name a few. These measures are called as Similarity Measures. Similarity measure plays an important role in text classification and clustering. In this work, I will use a similarity measure called as SMTP (Similarity Measure for Text Processing) for processing e-mail data using Fuzzy c-means clustering algorithm.

I. INTRODUCTION

Text processing plays an important role in information retrieval, data mining and web search. As the amount of digital documents has been increasing dramatically over the years as the internet grows, information management, search and retrieval has become an important problem. Clustering is a technique that organizes large number of objects into smaller coherent groups. Clustering aims at grouping similar documents get converted into a matrix where each word add a new dimension being the row of matrix and each document is represented as a column vector. This implies that each entry in the matrix gives the frequency of world occurring in a particular document. It is easy to see that the matrix would be sparse. Higher the frequency of each word, is more descriptive off the document

II. INFORMATION RETRIVAL MODELS

An information retrieval system is responsible for storing of data , organization of data , presentation of data , and easy access to desired information. The ultimate goal of the defined similarity techniques is to facilitate an information retrieval process. Similarity measures in IR processes are applied on a set of data objects, select relevant objects, produce a ranked set of relevant objects, and identify best matches. Several IR models exist that make use of various similarity measuring techniques along with appropriate data processing methodologies, for example :-

Set-theoretic Models: - It is based on set theory. The data object forms sets and set theoretic operations are used to derive similarities.

Algebraic Models:- In algebraic models the information is represented as vectors/matrices and similarity between information has a scalar value.
Probabilistic Models:- These IR models are based on probabilistic inferrences. Information is retrieved based on the probability of relevance between data.

Knowledge based Models:- The knowledge based IR models use structural and domain knowledge and formalized linguistic information to discover semantic relevance between objects.

Structure based Models:- The structural information retrieval models combine the content and structural characteristics to achieve greater retrieval efficiencies in many applications.

One of the widely used IR model is algebraic models. In algebraic models information is represented as vectors/matrices and similarity between information has a scalar value. Among various algebraic models such as Latent Semantic Analysis based model, Neural Networks and Vector space model(VSM), VSM is the most popular one. In vector space model Information is represented as vectors in multidimensional space. Each dimension corresponds to a possible feature of the information for example term in document. A Distance function is applied to the information vectors, provide the match and rank Information. VSM based information retrieval is a good mathematical implementation for processing large amount of information . It provides possibilities of partial matching and ranked result output. But this approach lacks the control of Boolean model, and has no means to handle semantic or syntactical information.

III. DOCUMENT CLUSTERING

Given a set of documents, we want to partition them into a pre-determined number of k subsets, such that the documents assigned to each subset are more similar to each other than the documents assigned to different subsets. Document clustering techniques mostly rely on single term analysis of the document data set, such as the Vector Space Model. To achieve more accurate document clustering, more informative features including phrases and their weights are very important in such scenarios. Document clustering is particularly useful in many applications such as grouping search engine results, automatic categorization of documents, building taxonomy of documents, and others. Each document in a corpus corresponds to an n-dimensional vector d, where m is the total number of terms. Document vectors are often subjected to some weighting schemes, such as the standard Term Frequency-Inverse Document Frequency (TF-IDF), and normalized to have unit length.

IV. DOCUMENT PROCESSING STEPS

Tokenization:- It is a process where document is treated as a string (or bag of words), and then partitioned into a list of tokens.

Removing stop words:- Stop words are frequently occurring, insignificant words. This step eliminates the stop words. Example of stop words are articles appearing in the sentences such as a, an, the etc

Stemming: - This step is the process of converging tokens to their root form. For example a word ‘buying’ is converted into word ‘buy’.

Document representation:- Generating N-distinct words from the corpora and call them as index terms (or the vocabulary). The document collection is then represented as a N-dimensional vector in term space.

TFIDF Analysis:- By taking into account two factors : term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. A search result’s score ranking is the product of TF and IDF:

\[ \text{TF-IDF} = \text{TF} \times \text{IDF} \]

IV. DOCUMENT PROCESSING STEPS

Tokenization:- It is a process where document is treated as a string (or bag of words), and then partitioned into a list of tokens.

Removing stop words:- Stop words are frequently occurring, insignificant words. This step eliminates the stop words. Example of stop words are articles appearing in the sentences such as a, an, the etc

Stemming: - This step is the process of converging tokens to their root form. For example a word ‘buying’ is converted into word ‘buy’.

Document representation:- Generating N-distinct words from the corpora and call them as index terms (or the vocabulary). The document collection is then represented as a N-dimensional vector in term space.

TFIDF Analysis:- By taking into account two factors : term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. A search result’s score ranking is the product of TF and IDF:

\[ \text{TF-IDF} = \text{TF} \times \text{IDF} \]

IV. DOCUMENT PROCESSING STEPS

Tokenization:- It is a process where document is treated as a string (or bag of words), and then partitioned into a list of tokens.

Removing stop words:- Stop words are frequently occurring, insignificant words. This step eliminates the stop words. Example of stop words are articles appearing in the sentences such as a, an, the etc

Stemming: - This step is the process of converging tokens to their root form. For example a word ‘buying’ is converted into word ‘buy’.

Document representation:- Generating N-distinct words from the corpora and call them as index terms (or the vocabulary). The document collection is then represented as a N-dimensional vector in term space.

TFIDF Analysis:- By taking into account two factors : term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. A search result’s score ranking is the product of TF and IDF:

\[ \text{TF-IDF} = \text{TF} \times \text{IDF} \]

IV. DOCUMENT PROCESSING STEPS

Tokenization:- It is a process where document is treated as a string (or bag of words), and then partitioned into a list of tokens.

Removing stop words:- Stop words are frequently occurring, insignificant words. This step eliminates the stop words. Example of stop words are articles appearing in the sentences such as a, an, the etc

Stemming: - This step is the process of converging tokens to their root form. For example a word ‘buying’ is converted into word ‘buy’.

Document representation:- Generating N-distinct words from the corpora and call them as index terms (or the vocabulary). The document collection is then represented as a N-dimensional vector in term space.

TFIDF Analysis:- By taking into account two factors : term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. A search result’s score ranking is the product of TF and IDF:

\[ \text{TF-IDF} = \text{TF} \times \text{IDF} \]
a. Maximum likelihood estimation (MLE)
b. Maximum a posteriori (MAP) estimation
4. Extended/Additional Measures
   a. Similarity measures based on fuzzy set theory
   b. Similarity measures based on graph theory

Distance-Based Similarity Measures :-

One of the most popular similarity measures is distance-based similarity measures. According to a very popular theoretical assumption, the perceived similarity can be inversely associated with the distance in some suitable feature space. In most of the cases it is considered to be a metric space. Many psychological theories assume that closer objects are more similar than the objects that are far apart. Similarity measure

\( s = d - 1 \)

The most common form of dissimilarity calculation refers to distance calculation in metric space. Some measures which have been popularly adopted for computing the similarity between two documents are briefly presented here.

Euclidean distance:-

Euclidean distance is also called as the Minkowski distance. It is the most commonly used measure to determine distance between two points. It is described as,

\[
D_2 = \left( \sum_{i=1}^{n} |x_{i2} - x_{i1}|^{1/2} \right)^2
\]

(1)

Jaccard distance:-

The Jaccard distance measures, dissimilarity between sample sets. It is complementary to the Jaccard coefficient (size of the intersection divided by the size of the union of the sample sets) and is obtained by subtracting the Jaccard coefficient from 1.

\[
J_d(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}
\]

(2)

Dice’s Coefficient:-

Dice coefficient similarity measure is defined as twice the number of terms common to compared entities/strings (nt) divided by the total number of terms in both tested strings.

\[
s = \frac{2n_t}{n_x + n_y}
\]

(3)

Cosine similarity:-

Cosine similarity is a vector based similarity measure used in text mining and information retrieval. In this approach compared strings are transformed into vector space and the Euclidean cosine rule applied to calculate similarity. This approach is paired with other approaches to limit the dimensionality of the vector space.

\[
similarity = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||}
\]

(4)

Hamming distance:-

It is considered to be the most popular measure for binary attributes. It is defined as the number of bits which differ between two binary strings i.e. the number of bits which need to be changed to turn one string into the other. For example the bit strings 1011101 and 1001001 has a hamming distance of 2bits, (two bits are dissimilar). This approach is used for exact length comparisons.

VI. CONCLUSION

It can be concluded that not all similarity measures are suitable in all scenarios. Some similarity measure work best for one scenario while others work best for other scenario. While doing survey of various distance measures in the category of distance-based measures it is understood that except for the Euclidean distance measure, the other measures have comparable effectiveness for the partial text document clustering task. The Jaccard and Pearson coefficient measures can form more coherent clusters. Despite of the above differences, these measures overall performance is similar.

REFERENCES


