HFRECCA : A New Sentence Clustering Technique For Text Mining

1, P.Siva Rama krishna 2. P.Srikanth
1, 2 Dept., of CSE, Sri Aditya Engineering College, Surampalem, EG.dist, AP, India.

Abstract: Clustering is the process of grouping or aggregating of data items. Sentence clustering mainly used in variety of applications such as classify and categorization of documents, automatic summary generation, organizing the documents, etc. In text processing, sentence clustering plays a vital role this is used in text mining activities. Size of the clusters may change from one cluster to another. The traditional clustering algorithms have some problems in clustering the input dataset. The problems such as, instability of clusters, complexity and sensitivity. To overcome the drawbacks of these clustering algorithms, this paper proposes a algorithm called Hierarchical Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (HFRECCA) is extension of FRECCA which is used for the clustering of sentences. Contents present in text documents contain hierarchical structure and there are many terms present in the documents which are related to more than one theme hence HFRECCA will be useful algorithm for natural language documents.

1. Introduction
Sentence clustering plays an important role in many text processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi-document summarization helps avoid problems of content overlap, leading to better coverage. However, sentence clustering can also be used within more general text mining tasks. For example, consider web mining, where the specific objective might be to discover some novel information from a set of documents initially retrieved in response to some query. By clustering the sentences of those documents we would intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; however, other clusters may contain information pertaining to the query in some way hitherto unknown to us, and in such a case we would have successfully mined new information. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement. Automatic text summarization researchers since Luhn work, they are trying to solve or at least relieve that problem by proposing techniques for generating summaries. Rath et al. in 1961 proposed empirical evidences for difficulties inherent in the notion of ideal summary. Both studies used thematic features such as term frequency, thus they are characterized by surface-level approaches. In the early 1960s, new approaches called entity-level approaches appeared; the first approach of this kind used syntactic analysis. The location features were used in, where key phrases are used dealt with three additional components: pragmatic words (cue words, i.e., words would have positive or negative effect on the respective sentence weight like significant, key idea, or hardly).

Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Over the past decades, many clustering algorithms have been proposed, including k means clustering, mixture models, spectral clustering, and maximum margin clustering. Most of these approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap. Thus, for items that belong to two or more clusters, it may be more appropriate to assign them with gradual memberships to avoid coarse-grained assignments of data. This class of clustering methods is called soft- or fuzzy-clustering.

Text mining mainly depends on geometric examination of a phrase, word or term. Sentence level clustering is an application of text classification. The most common objectives in text classification are to classify texts into fairly objective categories such as topics, but in sentiment mining the core objective is to identify the polarity of opinions, emotions, and evaluations.

Clustering has become an increasingly important topic with the explosion of information available via the Internet. It is an important tool in text mining and knowledge discovery. Its ability to automatically group similar textual objects together enables one to discover hidden similarity and key concepts, as well as to summarize a large amount of text into a small number of groups.

Methods used for text clustering include decision trees, conceptual clustering, clustering based on data summarization, statistical analysis, neural nets, inductive logic programming, and rule-based systems among others.
1. Related Work

The vector space model has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence. However, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at the document level, the assumption does not hold for small-sized text fragments such as sentences, since two sentences may be semantically related despite having few, if any, words in common. To solve this problem, a number of sentence similarity measures have recently been proposed. Rather than representing sentences in a common vector space, these measures define sentence similarity as some function of inter-sentence word-to-word similarities, where these similarities are in turn usually derived either from distributional information from some corpora (corpus-based measures), or semantic information represented in external sources such as Word Net [20] (knowledge-based measures).

2. Literature Survey

We present a statistical similarity measuring and clustering tool, SIMFINDER, that organizes small pieces of text from one or multiple documents into tight clusters. By placing highly related text units in the same cluster, SIMFINDER enables a subsequent content selection/generation component to reduce each cluster to a single sentence, either by extraction or by reformulation. We report on improvements in the similarity and clustering components of SIMFINDER, including a quantitative evaluation, and establish the generality of the approach by interfacing SIMFINDER to two very different summarization systems.

A novel method for simultaneous keyphrase extraction and generic text summarization is proposed by modeling text documents as weighted undirected and weighted bipartite graphs. Spectral graph clustering algorithms are used for partitioning sentences of the documents into topical groups with sentence link priors being exploited to enhance clustering quality. Within each topical group, saliency scores for keyphrases and sentences are generated based on a mutual reinforcement principle.

3. Existing System

The vector space model has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence.

Disadvantages

- The results often suffered from instability in the optimization algorithms that were used.
- A limitation of existing approach is the high dimensionality introduced by representing objects in terms of their similarity with all other objects.

4. Proposed System

This paper presents a novel Hierarchical fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as a likelihood.

Advantages

- It able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set of famous quotations, and applying the algorithm to a recent news article has demonstrated that the algorithm is capable of identifying overlapping clusters of semantically related sentences.
- Comparisons with the ARCA algorithm on each of these data sets suggest that HFRECCA is capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated by external measures.

5. System Architecture

```
Text data (XML Files) -> Similarity Measure (Page Rank) -> HFRECCA -> Clusters with some hierarchical order with sentences
```
6. ALGORITHM

step1: let each sentence be a cluster and let the membership degree of each clustering belonging to another sentence level text cluster be equal to 1

step2: find the proximity matrix of distances (similarity value)

step3: assign sequence numbers to clustering and then perform hierarchal fuzzy relational clustering algorithm for sentence clustering

step4: after performing hierarchal fuzzy relational algorithm check whether the membership value is smaller than the particular threshold value or not using the value obtained from FRC if yes then compute the new membership value and update the mixing coefficient otherwise go to next step.

step5: combine the first sentence cluster with second cluster into a new sentence cluster. Suppose first cluster contains n sentences and second cluster contains m sentences.

step6: then again find the degree of similarity between each pair of sentence clusters.

step7: if the degree of similarity between any two sentence clusters is smaller than the threshold value then stop.

step8: otherwise go to step 5 and repeat process.

7. Modules

MODULES DESCRIPTION:

(1) User Module
The user login and register for the specific query search, NLP Request and to cluster sentence level text using HFRECCA algorithm.

(2) Input Dataset
- The input dataset is taken from the already extracted information that is presented in the paper itself.
- The dataset is the collection of data.
- Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question.

(3) Fuzzy Clustering
- Clustering text at the document level is well established in the Information Retrieval (IR) literature.
- Here documents are typically represented as data points in a high-dimensional vector space in which each dimension corresponds to a unique keyword, leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents (values of the keywords).
- This type of data, which we refer to as “attribute data,” is amenable to clustering by a large range of algorithms.
- And we propose a Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) for clustering datasets.

(4) Page Rank
- By applying the Page Rank algorithm to each cluster, and interpreting the Page-Rank score of an object within some cluster as a likelihood, we can then use the Expectation-Maximization (EM) framework to determine the model parameters (i.e., cluster membership values and mixing coefficients).
- The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities.
- Text Rank and Lexmark apply a single instance of Page Rank to the collection of sentences.

8. Conclusion

Clustering, one of the conventional data mining strategies is an unsubstantiated knowledge pattern. Here clustering methods endeavor to recognize intrinsic alignments of the text documents, so that a set of clusters is formed in which clusters display high intra-cluster likeness and low inter-cluster likeness. Normally, text document clustering endeavors to separate out the documents into groups where every group characterizes some subject that is different from the topics characterized by the other groups. In this paper, a survey of sentence level clustering algorithms for text data is presented. A good clustering of text requires effective feature selection and a proper choice of the algorithm for the task at hand. Many algorithms are used to find the solutions to the above problems are discussed in detailed manner.

9. References


Mr.P.Siva Rama Krishna is a student of Sai Aditya College of Engineering & Technology, Surempalem. Presently he is pursuing his M.Tech [Computer Science Engineering] from this college and he received his B.Tech from Sai Aditya Institute of Science and Technology, affiliated to JNT University, Kakinada in the year 2012. His area of interest includes Computer Networks and Object oriented Programming languages, all current trends and techniques in Computer Science.

Mr.P.Srikanth, well known Author and excellent teacher Received Sri Aditya engineering college M.Tech from is working as Associate Professor, Department of M.Tech, He is completing M.Tech in Aditya college of Engineering and Technology 2013.He has teaching experience in engineering college. To his credit couple of publications both national and international conferences /journals . His area of Interest includes Data Warehouse and Data Mining, information security, flavors of Unix Operating systems and other advances in computer Applications.