Clustering Sentence-Level Text Using a Novel Fuzzy
Relational Clustering Algorithm

ABSTRACT:

In comparison with hard clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. However, because most sentence similarity measures do not represent sentences in a common metric space, conventional fuzzy clustering approaches based on prototypes or mixtures of Gaussians are generally not applicable to sentence clustering. This paper presents a novel 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. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a
variety of text mining tasks. We also include results of applying the algorithm to benchmark data sets in several other domains.

**EXISTING SYSTEM:**

Clustering text at the document level is well established in the Information Retrieval (IR) literature, where 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 (e.g., tf-idf values of the keywords).

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.

**DISADVANTAGES OF EXISTING SYSTEM:**

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.

**PROPOSED SYSTEM:**

This paper presents a novel 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. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks.
ADVANTAGES OF PROPOSED SYSTEM:

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 FRECCA is capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated by external measures.

ALGORITHM USED:

This section presents the proposed clustering algorithm. We first describe the use of PageRank as a general graph centrality measure, and review the Gaussian mixture model approach. We then describe how PageRank can be used within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm. The final section discusses issues relating to convergence, duplicate clusters, and various other implementation issues. Since PageRank centrality can be viewed as a special case of eigenvector centrality, we name the algorithm Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA).
MODULES:

(1) User Module
(2) Input Dataset
(3) Fuzzy clustering
(4) Page Rank

MODULES DESCRIPTION:

(1) User Module
The user login and register for the specific query search, NLP Request and to cluster sentence level text using FRECCA 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 pairwise similarities.
Text Rank and Lexmark apply a single instance of Page Rank to the collection of sentences.

**SYSTEM CONFIGURATION:**

**HARDWARE CONFIGURATION:**

- ✓ Processor - Pentium –IV
- ✓ Speed - 1.1 Ghz
- ✓ RAM - 256 MB(min)
- ✓ Hard Disk - 20 GB
- ✓ Key Board - Standard Windows Keyboard
- ✓ Mouse - Two or Three Button Mouse
- ✓ Monitor - SVGA

**SOFTWARE CONFIGURATION:**

- ✓ Operating System : Windows XP
- ✓ Programming Language : JAVA/J2EE
- ✓ Java Version : JDK 1.6 & above.