Research Article Impact Factor: 4.226 ISSN: 2319-507X
Rakhi S. Waghmare, IJPRET, 2015; Volume 3 (9): 76-83 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

STUDY ON DIFFERENT SENTENCE LEVEL CLUSTERING ALGORITHMS FOR TEXT MINING

RAKHI S.WAGHMARE, PROF. RAM MANGRULKAR, PROF.VAISHALI BHUJADE

Computer Science & Engg., BDCOE, Sewagram, Wardha, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Clustering is the process of grouping of data items. The sentence clustering is used in variety of applications i.e. classify and categorization of documents, automatic summary generation, etc. In text mining, the sentence clustering plays a vital role this is used in text activities. Size of clusters can change from one cluster to another. The existing system many clustering methods and algorithms are used for clustering the documents at sentence level. In this paper, we study the different sentence clustering algorithm as a study. The main aim of this study is to present an overview of the sentence level clustering techniques are to find the drawback of the exiting work and how could overcome the all this drawback for clustering algorithm. And we can obtain the more efficient technique or we may propose the new method to overcome the problems in existing methods like time redundancy and data aqurency.

Keywords: Text mining, FRECCA, Hierarchical Structure, sentence level clustering, and Median Fuzzy CMeans Clustering.

Corresponding Author: MS. RAKHI S.WAGHMARE

PAPER-QR CODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Rakhi S. Waghmare, IJPRET, 2015; Volume 3 (9): 76-83

INTRODUCTION

Data mining is a process of extraction of useful information from the huge amount of data. The development of IT in the last two decades paved the way for a world full of data. But much of these data are of potentially not useful. In order to make it useful one, we need to expand the large amount of information or knowledge underlying the data. Clustering techniques can help in this data discovery and data analysis. Clustering sentences is mainly useful in Information Retrieval (IR) Process. Clustering text at the sentence level and document level has many differences. Document clustering partitions the documents into several parts and cluster those parts based on the overall theme. It does not give much importance to the semantics of each sentence in the document. So there may be content overlap or bad coverage of theme will happen in the case of multi document summarization. So there may be content overlap or bad coverage of theme will happen in the case of multi document summarization.

Sentence clustering is an important role in many text processing. For ex, the various authors have argued that incorporating sentence clustering into extractive multi document summarization helps avoid problems of content overlap and leading to better coverage. However, the sentence clustering can also used within more general text mining tasks. For example, consider web some novel information from a set of documents initially retrieved in response to some query and by clustering the sentences of those documents we could intuitively expect at one of the clusters to be closely related to the concepts described by the query; however, other clusters may contain information pertaining to the query in some way hitherto unknown and in such a case we would have successfully mined new information

The goal of text summarization is present the most important information in a shorter version of the original text while keeping its main content and helps the user to immediately understand large volumes of information. Text summarization addresses both the problem of selecting the most important sections of text and the problem of generating coherent summary. This way is significantly different from that of human based text summarization that why 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 researcher's work, they are trying to solve or at least relieve that problem by proposing techniques for generating summaries.

A new fuzzy relational algorithm that based on the popular fuzzy C-means a the application of the algorithm to four real and four synthetic data sets, and prove that this algorithm performs better than well-known fuzzy relational clustering algorithms on all these sets[2]. The novel

fuzzy relational algorithm was proposed on the data in the form of a square matrix of pair wise similarities between data objects and used a graph representation of the data, and operates in an EM framework in which the graph centrality of an object in the graph[1]. Richard Khoury had proposed the new technique to automatically cluster together similar sentences based on the sentences' part-of-speech syntax[3].

Median Fuzzy C-Means Clustering algorithm was combine the median C means algorithm with the fuzzy c-means approach which is only applicable for vectorial (metric) data in its original variant [7].The sentence level classification had used the text classification problem and compare the performance x to this task: namely, a Support Vector Machine (SVM) classifier versus a Language Modeling approach [5]. Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang was proposed to determine an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches.

I. II. CLUSTERING TECHNIQUES

A. Fuzzy relational clustering algorithm based on the fuzzy C-means algorithm

In this work, showed how one can take advantage of the stability and effectiveness of object data clustering algorithms when the data to be clustered are available in the form of mutual numerical relationships between pairs of objects [2]. More specifically, here propose a new fuzzy relational algorithm that based on the popular fuzzy C-means algorithm. This algorithm does not require any particular restriction on the relation matrix. Here discuss the application of the algorithm to four real and four synthetic data sets, and prove that this algorithm performs better than well-known fuzzy relational clustering algorithms on all these sets.

B. Novel Fuzzy Relational Clustering Algorithm

In this algorithm to association with hard clustering methods, in this work a pattern belongs to a single cluster, fuzzy clustering algorithms permit patterns to belong all clusters with differing degrees of membership. That important in domains such as sentence clustering, a sentence is to be related to more than one topic present within a document or set of documents. Because number of sentence similarity measures do not represent sentences in a common metric space, fuzzy clustering approaches used on prototypes or mixtures of Gaussians are generally not applicable to sentence clustering. Andrew Skabar and Khaled Abdalgader [1] worked a novel fuzzy clustering algorithm called FRECCA that operates on the data in the form of a square matrix of pair wise similarities between data objects. This algorithm used a graph

representation of the data, and operates in an EM framework in which the graph centrality of an object in the graph is interpreted as likelihood [1].

C. Clustering Using Parts-of-Speech

Clustering algorithms are used in many Natural Language Processing tasks. They have proved to be popular and effective tools to use to discover groups of similar linguistic items [3]. In this exploratory paper, propose a new clustering algorithm to automatically cluster together similar sentences based on the sentences' part-of-speech syntax. The algorithm generates and joins together the clusters using a syntactic similarity metric based on a hierarchical organization of the parts-of-speech. Here explain the features of this algorithm by implementing it in a question type classification system, in order to calculate the positive or negative impact of different changes to the algorithm.

D. Fuzzy-based sentence-level document clustering

Analysis is one of the popular text mining operations in which a document whose content is contradictory to the theme of a set of documents is identified [4]. It is a means to identifying Outlier documents that do not confirm to the overall sense conveyed by other documents. Most of the techniques work document-level comparisons, neglecting the sentence-level semantics, and often leading to vanish of vital information. Applications in domains like Defense and Healthcare require high levels of accuracy and identification of micro-level contradictions are vital. In this paper, propose an algorithm for identifying contradictory documents using sentence-level clustering technique along with an optimization feature. A visualization scheme is give the results to an end-user.

E. Sentence-level event classification

The ability to correctly classify sentences that describe events is an important task for many natural language applications such as Question Answering and Text Summarization. In paper, treat event detection as a sentence level text classification problem [5]. Overall, here compare the performance x to this task: namely, a Support Vector Machine (SVM) classifier versus a Language Modeling approach and also investigate a rule-based method that uses handcrafted lists of 'trigger' terms derived from WordNet. Two datasets used in the experiments to test each approach on six different event types, like Attack, Die, Meet, Transport and Charge-Indict.

F. Sentence Similarity Based on Semantic Nets

Y Li [6] Sentence similarity measures play an increasingly important role in text-related research and applications in areas likewise Web page retrieval, text mining, and dialogue systems. These methods for computing sentence similarity have been adopted from approaches used for long text documents. These methods process sentences very high-dimensional space and are consequently inefficient, required human input, and are not adaptable to some application domains. We focuses directly on computing the similarity between very short texts of sentence length and it gives an algorithm takes account of semantic information and word order information implied in the sentences. The semantic of two sentences is calculated using information from a structured lexical database and gives corpus statistics. The use of a lexical database enables this method to model human common sense knowledge and the incorporation of corpus statistics allows our method to be adaptable to various domains.

We proposed method can also be used in a variety of applications that involve text knowledge representation and discovery. semantic and syntactic information contained in the compared texts. A text is considered to be a sequence of words each of which carries meaningful information. The words, along with their combination structure, make a text convey a specific meaning. Texts considered this paper are assumed to be of sentence length. Unlike the methods that use a fixed set of vocabulary, the proposed method dynamically forms a joint word set only using all the distinct words in the pair of sentences. For every sentence, a raw semantic vector is derived with the assistance of a lexical database. A word order vector is formed for every sentence, again using information from the lexical database. Since every word in a sentence contributes change to meaning of the complete sentence. The significance of a word is weighted by using information content derived from a corpus. By combining the raw semantic vector with information content from the corpus, a semantic vector is obtained for each of the two sentences. Semantic similarity computed based on the two semantic vectors. An order similarity is determined using the two order vectors. At last using the sentence similarity is derived by combining semantic similarity and order similarity.

G. Median Fuzzy C-Means Clustering

Median clustering is a powerful methodology for prototype based clustering of similarity/dissimilarity data [7]. In this contribution combine the median C means algorithm with the fuzzy c-means approach which is only applicable for vectorial (metric) data in its original variant. For the resulted median fuzzy C means approach here prove convergence and

investigate the behavior of the algorithm in several experiments including real world data from psychotherapy research.

H. Correlation Similarity for Document Clustering

Document clustering is to automatically group related documents into clusters. It is most important tasks in machine learning and artificial intelligence and has received much attention in recent years. Based on different distance measures, a number of methods have been proposed to handle document clustering. A typically and widely used distance measure is Euclidean distance. The k-means method is used the Euclidean distance that minimizes the sum of the squared Euclidean distance between data points and their corresponding cluster centers. Since the document space is all time of high dimensionality, it is preferable to determine a low-dimensional representation of the documents to reduce computation complexity.

Propose a new document clustering method [8] based on correlation preserving indexing, which externally considers the manifold structure embedded in the similarities between the documents. It aims to determine an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. The similarity-measure-based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely distributed the documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents CPI can effectively detect the intrinsic semantic structure of the high-dimensional document space. At this point, it is similar to Latent Dirichlet Allocation (LDA) which attempts to capture significant intra document statistical structure via the mixture distribution model.

I. Fuzzy Approach for Multitype Relational Data Clustering

Generally, pairwise relation could be described by similarities or dissimilarities between each pair of objects in the given dataset. In this work [9], only consider a similarity-type relation, which means the larger the value of the relationship between two objects and the more similar the two objects. Co clustering is treated as bipartite graph partitioning by measuring the singular value decomposition of the data matrix.

Multitype relational data may form various structures, depending on the availability of relations. A star-structure is a special case where relations only exist between the central type and several attribute types. It is possible to transform a multi type relational data into one of the basic data representation forms and then use an existing approach to get the clusters of

objects the interested type. However, meaningful information may be lost during data transformation. Moreover, clustering on each type of objects individually loses the chance of mutual improvement among clusters of different object types and is unable to capture the interrelated patterns among different types which may be of interest in some data-mining applications

III. CONCLUSION

Sentence Clustering is one of the conventional data mining strategies is an unsubstantiated knowledge pattern. Normally, the text document clustering 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 study 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. Different algorithms are used to find the solutions to the above problems.

REFERENCES

1. Andrew Skabar, Member, IEEE, and Khale Abdalgader "Clustering Sentence -Level Text Using a Novel Fuzzy Relational Clustering Algorithm" IEEE Trans. Knowledge and Data Eng, vol.25, no. 8,pp. 1138-1150, No.1,January 2013.

2. P. Corsini, F. Lazzerini, & F. Marcellon "A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C Means Algorithm, Soft Computing, vol. 9, pp. 439-447, 2005.

3. Richard Khoury "Sentence Clustering Using Part-of Speech I.J. Information Engineering And Electronic Business 2012,1,1-9.

4. R.Vasanth Kumar Mehta, B Sankarasubramaniam., S. Rajalakshmi "An algorithm for fuzzy-based sentence-level document clustering for micro level contradiction analysis" Proceeding ICACCI'12 Proceedings of the International Conference on Advance in Computing, Communications and Informatics 2012.

5. Martina Naughton, Nicola Stokes, and Joe Carthy "Sentence Level Event Classification in Unstructure Texts" Journal Information Retrieval archive Volume 13 Issue 2, April 2010 Pages 132-156.

6. Y. Li, D. McLean, Z.A. Bandar, J.D O'Shea, and K Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 8, pp. 1138-1150, Aug. 2006.

7. T. Geweniger, D. Zu¨ hlke, B. Hammer, and T. Villmann, "Median Fuzzy C- Means for Clusterin Dissimilarity Data," Neuro computing, vol.73, nos. 7-9, pp.1109-1116, 2010.

8. Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang, "Document Clustering in Correlation Similarity Measure Space", IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012.

9. Jian-Ping Mei, and Lihui Chen, "A Fuzzy Approach for Multitype Relational Data Clustring", IEEE TRANSACTION ON FUZZY SYSTEMS, VOL. 20, NO. 2, APRIL 2012.