Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.61 – 64

SURVEY ARTICLE

Survey on Clustering Algorithm for Sentence Level Text

Mr. Snehal Raundal, Prof. C. R. Barde

Department of Computer Engineering, R.H.Sapat College Nashik, Maharashtra & University of Pune, India

Department of Computer Engineering, R.H.Sapat College Nashik, Maharashtra & University of Pune, India

snehalraundal@gmail.com; erchandubarde@gmail.com

Abstract— Clustering is an extensively studied data mining problem in the text domains. The difficulty finds numerous applications in customer segmentation, classification, collaborative filtering, visualization, document organization, and indexing. In text mining, clustering the sentence is one of the processes and used within general text mining tasks. Several clustering methods and algorithms are used for clustering the documents at sentence level. In this article, the sentence level based clustering algorithm is discussed as a survey. The main goal of this survey is to present an overview of the sentence level clustering techniques. This demonstration of these techniques is used to obtain the efficient scheme for clustering for sentence level text. We can obtain the more efficient method or we may propose the new technique to overcome the problems in these existing approaches. This survey article is intended to provide easy accessibility to the main ideas for non-experts.

Keywords— Sentence level clustering, Sentence Similarity, ranking, clustering of sentences, Median Fuzzy CMeans Clustering

I. INTRODUCTION CLUSTERING TECHNIQUES

In many text processing activities, Sentence clustering plays an important role. For instance, various authors have argued that incorporating sentence clustering into extractive multi-document summarization helps avoid problems of content overlap, leading to better coverage [1], [2], [3], [4]. On the other hand, sentence clustering can also be used within more general text mining tasks. For instance, regard as web mining [5], 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 these documents we would intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; though, other clusters may contain information pertaining to the query in some way previously unknown to us, and in such a case we have successfully mined new information. Irrespective of the specific task (e.g., summarization, text mining, etc.), most documents will contain interrelated themes or topics, and many sentences will be related to some degree to a number of these. Nevertheless, clustering text at the sentence level poses specific challenges not present when clustering larger segments of text, such as documents. We now highlight some important differences between clustering at these two levels, and examine some existing approaches to fuzzy clustering.

Text mining mainly depends on geometric examination of a phrase, word or term. Sentence level clustering is an application of text classification[6]. 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 [7]. It is an important tool in text mining and knowledge discovery. It is able to automatically group similar textual objects together enables one to discover hidden similarity and key concepts, also use 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 [8]. In text clustering, it is important to note that selecting important features, which present the text data properly, has a critical effect on the output of the clustering algorithm. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially.

To discriminate it from attribute data, also refer to such data as relational data. A huge range of hierarchical clustering algorithms can also be applied. The vector space model has been flourishing in IR because it is able to effectively capture much of the semantic content of document-level text. This is since documents that are semantically correlated are likely to include lots of words in common, and consequently are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence. Conversely, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at the document.

II. CLUSTERING TECHNIQUES

A. *An algorithm for fuzzy-based sentence-level document clustering for micro-level contradiction analysis*

Contradiction 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 [9]. It is a means to identifying Outlier documents that do not confirm to the overall sense conveyed by different documents. In the existing techniques perform document-level comparisons, ignoring the sentence-level semantics, often leading to loss of information. Applications in different domains like Defence and Healthcare require high levels of accuracy and identification of micro-level contradictions are vital. In this paper they propose an algorithm for identifying contradictory documents using sentence-level clustering technique along with an optimization feature. A visualization scheme is also suggested to present the results to an end user.

B. *Median Fuzzy C-Means for Clustering Dissimilarity Data*

Median clustering is a powerful methodology for prototype based clustering of similarity/dissimilarity data [10]. In this contribution we combine the median c-means algorithm with the fuzzy c-means approach, which is only applicable for vectorial (metric) data in its original variant. The resulted median fuzzy c-means approach we prove convergence and investigate the behaviour of the algorithm in several experiments including real world data from psychotherapy research.

C. *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. More specifically, here propose a new fuzzy relational algorithm, based on the popular fuzzy C-means (FCM) algorithm, which does not require any particular restriction on the relation matrix. Here describe the application of the algorithm to four real and four synthetic data sets, and show that this algorithm performs better than well-known fuzzy relational clustering algorithms on all these sets.

D. *Clustering Using Parts-of-Speech*

Clustering algorithms are used in many Natural Language Processing (NLP) tasks. They have proven to be popular and effective tools to use to discover groups of similar linguistic items. In this exploratory paper, propose a new clustering algorithm to automatically cluster together similar sentences based on the sentences

part-of-speech syntax. This algorithm generates and merges together the clusters using a syntactic similarity metric based on a hierarchical organization of the parts-of-speech. Here demonstrate the features of this algorithm by implementing it in a question type classification system, in order to determine the positive or negative impact of different changes to the algorithm.

E. *Embedded graph based sentence clustering*

In this paper, a document summarization framework for storytelling is proposed to extract essential sentences from a document by exploiting the mutual effects between terms, sentences and clusters. There are three phrases in the framework: sentence clustering, sentence ranking and document modelling. The story document is modelled by a weighted graph with vertexes that represent sentences of document. The sentences are clustered into different groups to find the not developed topics in the story. To alleviate the influence of unrelated sentences in clustering, an embedding process is used to optimize the document model. The sentences are rank according to the mutual effect between terms, sentence and clusters, and high-ranked sentences are selected to comprise the summarization of the document. The experimental results on the Document Understanding Conference (DUC) data sets demonstrate the effectiveness of the proposed method in document summarization. The results show that the embedding process for sentence clustering renders the system more robust with respect to different cluster numbers.

F. *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 (QA) and Text Summarization. In this paper, treat event detection as a sentence level text classification problem. Overall, here compare the performance of discriminative versus generative approaches to this task: namely, a Support Vector Machine (SVM) classifier versus a Language Modelling (LM) approach. Here also investigate a rule-based method that uses handcrafted lists of 'trigger' terms derived from WordNet. Two datasets are used in the experiments to test each approach on six different event types, i.e., Die, Meet, Transport, Attack, Injure, and Charge-Indict.

G. *Sentence Clustering using Similarity Word-Sequence Kernels*

In this paper, present a novel clustering approach based on the use of kernels as similarity functions and the C-means algorithm. Several word sequence kernels are defined and extended to verify the properties of similarity functions. Afterwards, these monolingual word-sequence kernels are extended to bilingual word-sequence kernels, and applied to the task of monolingual and bilingual sentence clustering. The motivation of this proposal is to group similar sentences into clusters so that specialized models can be trained for each cluster, with the purpose of reducing in this way both the size and complexity of the initial task. Here also provide empirical evidence for proving that the use of bilingual kernels can lead to better clusters, in terms of intra-cluster perplexities.

III. PROBLEMS IN SENTENCE LEVEL CLUSTERING

A. Some of the issues during the clustering the sentence in the documents

- Using the Similarity measure in some clustering algorithm can be measured in terms of word co-occurrence may be valid at the document level, the assumed similarity measures 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.
- The length of sentences is short and the content it contains is limited, the similarity traditionally used for document clustering is no longer suitable for sentence clustering. Special treatment for measuring sentence similarity is necessary.

IV. CONCLUSIONS

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, with the intention 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 article, 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. Fuzzy clustering approaches can be used to improve the overall performance of the clustering approaches.

ACKNOWLEDGEMENT

We sincerely thanks to all the people who helped us to write this review paper. We also like to thank all the open source software developer communities and the researchers for publishing their research work as a guideline.

REFERENCES

[1] V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and K.R. McKeown, "SIMFINDER: A Flexible Clustering Tool for Summarization," Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.

[2] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.

[3] D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J., vol. 40, pp. 919-938, 2004.

[4] R.M. Aliguyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization," Expert Systems with Applications, vol. 36, pp. 7764- 7772, 2009.

[5] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

[6] Amanda Rachel Hutton, B.S. "Using Sentence-Level Classification to Predict Sentiment at the Document-Level" May 2012.

[7] Karypis, George, Vipin Kumar and Michael Steinbach. 2000. *A Comparison of Document Clustering Techniques*. KDD workshop on Text Mining.

[8] J.Durga, D.Sunitha, S.P.Narasimha, B.Tejeswini Sunand "A Survey on Concept Based Mining Model using Various Clustering Techniques" International Journal of Advanced Research in Computer Science andSoftware Engineering 2012.

[9] 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 Advances in Computing, Communications and Informatics 2012.

[10] T. Geweniger, D. Zu¨ hlke, B. Hammer, and T. Villmann, "Median Fuzzy C-Means for Clustering Dissimilarity Data," Neurocomputing, vol. 73, nos. 7-9, pp. 1109-1116, 2010.

