View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by International Journal of Science Engineering and Advance Technology (IJSEAT)

International Journal of Science Engineering and Advance Technology,IJSEAT, Vol 3, Issue 3,March - 2015 | ISSN 2321-6905

International Journal of Science Engineering and Advance Technology

A New Clustering Technique On Text In Sentence For Text Mining

[1]B.Lakshmi Narayana,S.Phani Kumar
[1,2]Dept. of CSE, PACE Institute of Technology And Sciences., Ongole, Prakasam District,AP, India

Abstract: Clustering is a commonly considered data mining problem in the text domains. The problem finds numerous applications in customer segmentation, classification, collaborative filtering, visualization, document organization, and indexing. In this paper, the sentence level based clustering algorithm is discussed as a survey. The survey explains about the problems in clustering in sentence level and the solutions to overcome these problems. 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 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. In this algorithm single object may belong to more than one cluster.

I. Introduction

Sentence clustering acts 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.

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.

II. 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).

III. Literature Survey

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.

IV. Problem Definition

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.

V. Proposed Approach

This 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.

VI. Proposed Methodology

User Module

The user login and register for the specific query search, NLP Request and to cluster sentence level text using FRECCA algorithm.

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.

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.

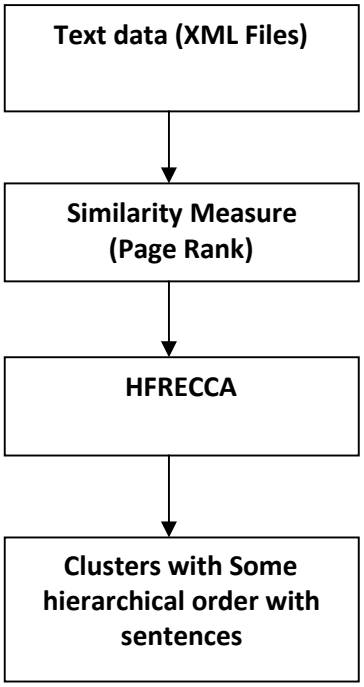
Page Rank

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.

VII. System Architecture



VIII. Conclusion

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.

IX. 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] H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBM Journal of Research and Development, vol. 2, pp.159-165. 1958.

[7] G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences" American Documentation, vol. 12, pp.139-143.1961.

[8] Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization MIT Press. 1999.

[9] H. P. Edmundson., "New methods in automatic extracting" Journal of the Association for Computing Machinery 16 (2). pp.264- 285.1969.

[10] R. O. Duda, P. H. Hart, and D. G. Stock, Pattern Classification. New York: Wiley, 2001.

[11] U. von Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol. 17, no. 4, 2007.

[12] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1537–1544.

[13] K. Zhang, I.W. Tsang, and J. T.Kwok, "Maximum margin clusteringmade practical," in Proc. 24th Int. Conf. Mach. Learning, 2007, pp. 1119–1126.

[14] F.Hoppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis.New York: Wiley, 1999.

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

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

[17] 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.

[18] C.D. Manning, P. Raghavan, and H. Schu¨ tze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.

[19] 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.

[20] C. Fellbaum, Word Net: An Electronic Lexical Database. MIT Press, 1998.

Authors:

Mr. BADUGU LAKSHMI NARAYANA is a student of Pace Institute of Technology and Sciences, Ongole. Presently he is pursuing his M.Tech(CSE) from this college and he received his B.Tech from SSN Engineering College, affiliated to Jawaharlal Nehru Technological University, Kakinada in the year 2012. His area of interest includes Object Oriented Programming language and Computer Networks, all current trends and techniques in Computer Science.

Mr S. PHANI KUMAR well known author and excellent teacher Received B.Tech in Jawaharlal Nehru Technological University Hyderabad (JNTUH) and M.Tech(WT) from Jawaharlal Nehru Technological University Hyderabad. he is working as Assistant Professor in the deportment of CSE. He has four years of teaching experience in various engineering colleges. To his credit couple of publications both national and international conferences/journals. His area of interest includes in Web Technologies and other advances in computer Applications.