© 2019 IJSRSET | Volume 6 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : 10.32628/IJSRSET196173

# A Review on Improving the Clustering Performance in Text Mining

**Ashwini Harishchandra Ghonge[1], Prof. Vijaya Kamble[2]**

[1]M. Tech, Computer Science Engineering, Guru Nanak Institute of Engineering and Technology, Nagpur, Maharashtra, India

[2]Computer Science Engineering, Guru Nanak Institute of Engineering and Technology, Nagpur, Maharashtra, India

## ABSTRACT

Lately, the improvement of data frameworks in each field, for example, business, scholastics, and the drug have prompted increment in the measure of put away information step by step. A lion's share of information is put away in archives that are practically unstructured. Content mining innovation is exceptionally useful for individuals to process immense data by forcing structure upon content. Clustering is a well-known procedure for automatically sorting out a huge gathering of content. Nonetheless, in genuine application spaces, the experimenter has some foundation learning that helps in clustering the information. Customary clustering strategies are somewhat unsatisfactory of various information types and can't deal with sparsity and high dimensional information. Co-clustering strategies are received to defeat the customary clustering strategy by at the same time performing report and word clustering taking care of the two insufficiencies. Semantic comprehension has turned out to be a fundamental element for data extraction, which is made by receiving requirements as a semi-supervised learning technique. This overview audits on the compelled co-clustering techniques embraced by specialists to support the clustering execution.

**Keywords :** Clustering Techniques, Co-Clustering,Constrained Clustering, Semi-supervised Learning, Text Mining.

## I. INTRODUCTION

Consistently, individuals encounter a lot of data and store or speak to it as information, for further investigation and the board. Information mining is the act of automatically looking substantial stores of information to discover examples and patterns that go past basic examination. Information mining utilizes modern numerical calculations to fragment the information and assess the likelihood of future occasions. Information mining is developed in a multidisciplinary field, including database innovation, machine learning, man-made brainpower, neural system, data recovery, etc. On a basic level information mining ought to be pertinent to the distinctive sort of information and databases utilized in a wide range of utilizations, including social databases, value-based databases, information stockrooms, object-situated databases, and exceptional application-arranged databases, for example, spatial databases, worldly databases, multimedia databases, and time-arrangement databases. Information mining is otherwise called Knowledge Discovery in Data (KDD) [24]. Essentially there are diverse sorts identified with information mining, they are: text

mining, web mining, multimedia mining, object mining and spatial information mining.

Text mining, generally equal to text investigation, alludes to the way toward getting brilliant data from text. Text mining or learning discovery from text (KDT)deals with the machine upheld investigation of text. It utilizes techniques from data recovery, data extraction just as characteristic language preparing (NLP) and connects them with the calculations and strategies for KDD, information mining, machine learning and insights. Text mining can be likewise characterized like information mining, data extraction and learning discovery process model. In Text mining [21], the choice of attributes and furthermore the impact of space learning and area explicit methods assume a critical job.

**A. Data Retrieval :** Information recovery is the finding of archives which contain answers to questions and without center to answers itself. Techniques are utilized for the automatic handling of text information and comparison to the given inquiry. Data recovery in the more extensive sense manages the whole scope of data handling, from information recovery to learning recovery.

**B. Natural Language Processing (NLP):** The general objective of NLP is to accomplish a superior comprehension of common language by utilization of computers. It utilizes straightforward and strong techniques for the quick preparing of text. What's more, phonetic investigation techniques are utilized in addition to other things for the preparing of text.

**C. Data Extraction (IE)** : The objective of data extraction techniques is the extraction of explicit data from text archives. These are put away in information base-like examples for further use. So as to get all words that are utilized in a given text, a tokenization procedure is required, for example a text record is part into a flood of words by evacuating all accentuation marks and by supplanting tabs and other non-text characters by single void areas. The arrangement of various words acquired by consolidating all text archives of a collection is known as the lexicon of a record collection (pack of-words portrayal). So as to decrease the span of the word reference separating and lemmatization or stemming techniques can be received. Sifting techniques evacuate words like articles, conjunctions, relational words from the lexicon and the equivalent is utilized for the archives. Lemmatization techniques attempt to delineate structures to the unending tense and things to the solitary structure. Since this labeling procedure is normally very time consuming and still mistake inclined, by and by much of the time stemming techniques are connected. Stemming techniques endeavour to construct the fundamental types of words, for example strip the plural 's' from things, the 'ing' from action words, or different appends.

A phonetic preprocessing can be utilized to upgrade the accessible data about terms. They play out the accompanying techniques: (a) Part-of-discourse labeling (POS) decides the labeling of grammatical form, (b) Text piecing goes for gathering contiguous words in a sentence; (c) Word Sense Disambiguation (WSD) endeavors to determine the vagueness in the significance of single words or expressions. (d) Parsing produces a full parse tree of a sentence. Fruitful uses of text mining strategies in very differing zones are patent investigation, text grouping in news offices, bioinformatics, spam sifting, explorative information examination, data representation, text outline and point location ponders.

## II. LITERATURE SURVEY

The review on clustering strategies falls under threecategories: co-clustering, constrained co-clustering withunsupervised constraints and semi-supervised clustering.

2.1 Co-Clustering

Most of the traditional clustering algorithms aim at clustering homogeneous data, which is contrary to many real world applications. Also there exists close relationships between different types of data, and it is difficult for the traditional clustering algorithms to utilize that relationship informationefficiently. It cannot handle missing data (or empty clusters or Sparsity), dimensionality reduction and computational inefficient clustering algorithms for inference been used. The existing document clustering methods are Agglomerative clustering, partitional k-means algorithm, Projection based LSA (Latent Semantic Indexing), Self-Organizing Maps (SOM), multidimensional scaling, Singular Valued Decomposition (SVD) etc. Example methodology is generating document -word frequency which is thereby complex for computation and processing. Consequently, coclustering techniques aims to cluster different types of data simultaneously by making efficient use of the relationship information i.e. examines both document and word relationship simultaneously. They follow thereby another paradigm than the classical cluster algorithm as k-means which only clusters elements of the one dimension on the basis of their similarity to the second one, e.g. documents based on terms.

Co-clustering can be done using matrix or graph as a good representation of document-word pair. For graph theoretic approach, bipartite spectral graph partitioning can be used to handle the problem of dimensionality reduction and Sparsity of data. But many effective heuristic methods exist, such as, the Kernighan-Lin (KL) and the Fiduccia-Mattheyses (FM) algorithms. However, both the KL and FM algorithms search in the local vicinity of given initial partitioning and have a tendency to get stuck in local minima. The novel idea of modeling the document collection as a bipartite graph between documents and words, using which the simultaneous clustering problem can be posed as a bipartite graph partitioning problem [12]. To solve the partitioning problem, a new spectral co-clustering algorithm enjoys some optimality properties; it is shown that the singular vectors solve a real relaxation to the NP-complete graph bipartitioning problem and finds global optimal solution. But algorithm results show that sparsity is still present and it is difficult to recover original classes. With a similar philosophy,

Gao et al. [15] proposed Consistent Bipartite Graph Co-partitioning (CBGC) using semi definite programming for high-order data coclustering and applied it to hierarchical text taxonomy preparation. Due to the nature of graph partitioning theory, these algorithms have the restriction that clusters from different types of objects must have one-to-one associations. More recently, Long et al. [19] proposed Spectral Relational Clustering (SRC), to perform heterogeneous coclustering. SRC provides more flexibility by lifting the requirement of one-to-one association in graph-based coclustering. However, to obtain data clusters, all the before mentioned graph theoretical approaches require solving an Eigen-problem, which computationally is not efficient for large-scale data sets. Using matrix representation is deemed to be best to handle document clustering since generating clusters row wise and column wise is computationally efficient than handling graph. In application of gene expression data, an expression matrix is generated that uses combination of genes and conditions, the enables automatic discovery of similarity based on subset of attributes and overlapped grouping for better representation of genes with multiple functions [8]. But the empty clusters handled in [8] are inefficient because of usage of random number for missing data replacement and also algorithm used is not good in cases like NP-hardness. On motivation of [8], a concept proposed in [9] that uses mean squared residue to simultaneously cluster genes and conditions handling empty clusters and local minima problems. It uses iterative non-overlapping algorithm that uses k * l co-clusters simultaneously (k rows and l columns) rather than

one co-cluster at a time and uses local search strategy to avoid empty clusters and local minima, the algorithm suffers from a drawback of anti-correlation. Nonnegative matrix factorization (NMF) is widely used to approximate high dimensional data comprising nonnegative components i.e. to extract concepts/topics from unstructured text documents. In [33] it is shown that Non-negative Matrix Factorization (NMF) outperforms spectral methods in document clustering achieving higher accuracy and efficiency, but still achieves only local minima of objective function. The co-occurrence frequencies can also be encoded in co-occurrence matrices and then matrix factorizations are utilized to solve the clustering problem [14]. Ding et al. in [14] uses bi-orthogonal 3-factor NMF (BiORNM3F) clustering algorithm to rigorously cluster documents and compare its performance with other standard clustering algorithms, where documents are represented using the binary vector-space model and each document is a binary vector in the term space. But in measures of entropy the algorithm is no better than k-means algorithm. In paper [1], Bregman co-clustering is used for matrix approximation which is measured in terms of distortion measure. A minimum Bregman information (MBI) principle that simultaneously generalizes the maximum entropy and standard least squares principles, leads to a matrix approximation that is optimal among all generalized additive models in a certain natural parameter space is used. Analysis based on this principle yields an elegant meta algorithm, special cases of which include most previously known alternate minimization-based clustering algorithms such as Kmeans and co-clustering algorithms such as information theoretic [13] and minimum sum-squared residue coclustering [9]. Bregman divergences constitute a large class of distortion measures including the most commonly used ones such as squared Euclidean distance, KL-divergence, Itakura- Saito distance, I-divergence etc. Bregman co-clustering alsohandles missing value prediction and compression of categorical data matrices. Kullback-Leibler divergence (KLdivergence) on text is defined on two multinomial distributions and has proven to be very effective in coclusteringtext [1]. The paper [26] overcomes the drawback of Generative Mixture Model (GMM) by proposing Bayesian Co-Clustering (BCC) model allowing mixed membership in row and column clusters and also introduces separate Dirichlet distributions as Bayesian priors over mixed membership. BCC handles sparse matrices and efficiently handles different data types. To optimize the model Expectation Maximization (EM) style algorithm was proposed to preserve dependencies among entries in same row/column and parameters could be learned using maximum likelihood estimation. The paper [31] is variation of [26] that use collapsed Gibbs sampling and collapsed variation inference for parameter estimation. Latent Dirichlet Bayesian Co-Clustering (LDCC) approach assumes Dirichlet priors for row- and column-clusters, which are unobserved in the data contingency matrix. The collapsed Gibbs sampling and collapsed variation Bayesian algorithms help to learn more accurate likelihood functions than the standard variation Bayesian algorithm which can lead to higher predictive performance. The paper [13] uses theoretical formulation to obtain useful information on performing coclustering. It uses Optimal co-clustering strategy that minimize loss of mutual information by using Joint probability distribution between two discrete random variable i.e. rows and columns. Relative entropy called Kullback- Leibler (KL) divergence is used to maximize the mutual information for hard clusters. In NG20 (20 Newsgroups) application, it reports that 45% of documents are cross posted making boundaries between newsgroups fuzzy. While most classical clustering algorithms assign each datum to exactly one cluster, thus forming a crisp partition of the given data, fuzzy clustering allows for degrees of membership, to which a datum belongs to different clusters. This approach is frequently more stable in application using text. The fuzzy cmeans (FCM) clustering algorithms defined in paper [34] are the well-known and powerful methods in cluster analysis.

## III. CONCLUSION

This survey focuses to provide the clustering techniques adopted in text mining. For deeper understanding of clustering in text mining, it is necessary to handle each and every process in its life cycle for achieving better results. The clustering methodology of choice depends on type of application domain and also based on expected results. This review focuses on three major categories: co-clustering, constrained co-clustering with unsupervised constraints and semi-supervised clustering. Every category is determined for particular purpose which aims to improve the clustering performance by quality and accuracy of clusters and constraints generated.

## IV. REFERENCES


[1] Banerjee.A, Dhillon.I, Ghosh.J,. Merugu.S, and Modha.D.S (2017), "A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix

[2] Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986.

[3] Basu S., Bilenko M., and Mooney R.J. (2014), "A Probabilistic Framework for Semi-Supervised Clustering," Proc. SIGKDD, pp. 59-68.

[4] Basu.S, Banerjee A., and Mooney R.J. (2012), "Semi- Supervised Clustering by Seeding," Proc. 19th Int'l Conf. Machine Learning (ICML), pp. 27-34.

[5] Bikel D., Schwartz R., and Weischedel R. (1999)," An algorithm that learns what's in a name", Machine learning, 34:211–231.

[6] Bilenko M. and Basu S.(2004), "A Comparison of Inference Techniques for Semi-Supervised Clustering with Hidden Markov Random Fields," Proc. ICML

[7] Workshop Statistical Relational Learning (SRL '04).

[8] Bilenko.M, Basu.S, and Mooney R.J. (2004), "Integrating Constraints and Metric Learning in Semi-Supervised Clustering," Proc. 21st Int'l Conf. Machine Learning (ICML), pp. 81-88.

[9] Chen Y., Wang L., and Dong M.(2010), "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co-Clustering," IEEE Trans. Knowledge and Data Eng., vol.22, no. 10, pp. 1459-1474.

[10] Cheng Y. and Church G.M. (2000), "Biclustering of Expression Data," Proc. Int'l System for Molecular Biology Conf. (ISMB), pp. 93-103.

[11] Cho H., Dhillon I.S., Guan Y., and Sra S. (2004), "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," Proc. Fourth SIAM Int'l Conf.Datamining (SDM).

[12] Cozman F.G., Cohen I., and Cirelo M.C. (2003), "Semi- Supervised Learning of Mixture Models," Proc. Int'l Conf. Machine Learning (ICML), pp. 99-106.

[13] Dai W., Xue G.-R., Yang Q., and Yu Y. (2007), "Co- Clustering Based Classification for Out-of-Domain Documents," Proc. 13th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 210- 219.

[14] Dhillon I.S. (2001), "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD), pp. 269-274.

[15] Dhillon.I.S, Mallela.S, and Modha D.S.(2003), "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98.

[16] Ding C., Li.T, Peng.W, and Park.H (2006), "Orthogonal Nonnegative Matrix T-Factorizations for Clustering," Proc. 12th ACM SIGKDD Int'l Conf. KnowledgeDiscovery and Data Mining, pp. 126-135.

[17] Gao.B, Liu T.-Y., Feng G., Qin T., Cheng Q.-S. And Ma W.-Y. (2005) ,"Hierarchical Taxonomy Preparation for Text Categorization Using Consistent Bipartite Spectral Graph Co

partitioning," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 9, pp. 1263- 1273.

[18] Jain.A, Murty.M, and Flynn.P (1999), "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323.

[19] Li T., Ding C., Zhang Y., and Shao B. (2008), "Knowledge Transformation from Word Space to Document Space," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 187-194.

[20] Li T., Zhang Y., and Sindhwani V.(2009), "A Non- Negative Matrix Tri- Factorization Approach to Sentiment Classification with Lexical PriorKnowledge," Proc. Joint Conf. (ACL-IJCNLP), pp. 244- 252.

[21] Long.B, Wu X., Zhang Z. and Yu. P.S (2006), "Spectral Clustering for Multi-Type Relational Data," Proc. 23rd Int'l Conf. Machine Learning, pp. 585-592.

[22] Lu Z. and Leen T.K. (2007), "Penalized Probabilistic Clustering," Neural Computation, vol. 19, no. 6, pp. 1528-1567.

[23] Michael W. Berry and Malu Castellanos (2007),"Survey of Text Mining: Clustering, Classification, and Retrieval", Springer, Second Edition.

[24] Nigam K., McCallum A.K., Thrun S., and Mitchell T.M. (2000), "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, vol. 39, no. 2/3, pp. 103-134.

[25] Pensa R.G. and Boulicaut J.-F.(2008), "Constrained Co- Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 25-36.

[26] Revathi.T, Sumathi.P (2013)," A Survey on Data Mining using Clustering Techniques", International Journal of Scientific & Engineering Research Volume 4, Issue 1.

[27] Rui Xu, Donald Wunsch II (2005)," Survey of Clustering Algorithms", IEEE Transactions On Neural Networks, Vol. 16, NO. 3, pp. 645-678.

[28] Shan.H and A. Banerjee.A (2008), "Bayesian Co-Clustering," Proc. IEEE Eight Int'l Conf. DataMining (ICDM), pp. 530-539.

[29] Shi X., Fan W., and Yu P.S. (2010), "Efficient Semi- Supervised Spectral Co-Clustering with Constraints," Proc. IEEE 10th Int'l Conf. Data Mining (ICDM), pp. 1043-1048.

[30] Song Y., Pan S., Liu S., Wei F., Zhou M.X., and Qian W. (2010), "Constrained Co-Clustering for Textual Documents," Proc. Conf. Artificial Intelligence (AAAI).

**Cite this article as :**

Ashwini Harishchandra Ghonge, Prof. Vijaya Kamble, " A Review on Improving the Clustering Performance in Text Mining, International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 6, Issue 1, pp.380-385, January-February-2019.

Journal URL : http://ijsrset.com/IJSRSET196173