A Survey on Improving the Clustering Performance in Text Mining for Efficient Information Retrieval

S.Saranya[#1], R.Munieswari[*2]
M.E Scholar, Department of Computer Science & Engineering,
Kumaraguru College of Technology, Coimbatore, Tamilnadu, INDIA

Abstract **--- In recent years, the development of information systems in every field such as business, academics and medicine has led to increase in the amount of stored data year by year. A vast majority of data are stored in documents that are virtually unstructured. Text mining technology is very helpful for people to process huge information by imposing structure upon text. Clustering is a popular technique for automatically organizing a large collection of text. However, in real application domains, the experimenter possesses some background knowledge that helps in clustering the data. Traditional clustering techniques are rather unsuitable of multiple data types and cannot handle sparsity and high dimensional data. Co-clustering techniques are adopted to overcome the traditional clustering technique by simultaneously performing document and word clustering handling both deficiencies. Semantic understanding has become essential ingredient for information extraction, which is made by adopting constraints as a semi-supervised learning strategy. This survey reviews on the constrained co-clustering strategies adopted by researchers to boost the clustering performance. Experimental results using 20-Newsgroups dataset shows that the proposed method is effective for clustering textual documents. Furthermore, the proposed algorithm consistently outperformed all the existing constrained clustering and coclustering methods under different conditions.**

Keywords **--- Clustering Techniques, Co-Clustering, Constrained Clustering, Semisupervised Learning, Text Mining.**

I. INTRODUCTION

Every day, people encounter a large amount of information and store or represent it as data, for further analysis and management. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is evolved in a multidisciplinary field, including database technology, machine learning, artificial intelligence, neural network, information retrieval, and so on. In principle data mining should be applicable to the different kind of data and databases used in many different applications, including relational databases, transactional databases, data warehouses, object- oriented databases, and special application- oriented databases such as spatial databases, temporal databases, multimedia databases, and time- series databases. Data mining is also known as Knowledge Discovery in Data (KDD) [24]. Basically there are different types related to data mining, they are: text mining, web mining, multimedia mining, object mining and spatial data mining.

Text mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. Text mining can be also defined similar to data mining, information extraction and knowledge discovery process model. In Text mining [21], the selection of characteristics and also the influence of domain knowledge and domain-specific procedures play an important role.

A. *Information Retrieval (IR)***:** Information retrieval is the finding of documents which contain answers to questions and without focus to answers itself. Methods are used for the automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval.

B. *Natural Language Processing (NLP)*: The general goal of NLP is to achieve a better understanding of natural language by use of computers. It employs simple and durable techniques for the fast processing of text. In addition, linguistic analysis techniques are used among other things for the processing of text.

C. *Information Extraction (IE):* The goal of information extraction methods is the extraction of specific information from text documents. These are stored in data base-like patterns for further use.

In order to obtain all words that are used in a given text, a *tokenization* process is required, i.e. a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. The set of different words obtained by merging all text documents of a collection is called the *dictionary* of a document collection (bag-of-words representation). In order to reduce the size of the dictionary filtering and lemmatization or stemming methods can be adopted. *Filtering* methods remove words like articles, conjunctions, prepositions from the dictionary and the same is used for the documents. *Lemmatization* methods try to map verb forms to the infinite tense and nouns to the singular form. Since this tagging process is usually quite time consuming and still error-prone, in practice frequently stemming methods are applied. *Stemming* methods try to build the basic forms of words, i.e. strip the plural 's' from nouns, the 'ing' from verbs, or other affixes. A linguistic preprocessing can be used to enhance the available information about terms. They perform the following methods: (a) *Part-of-speech tagging (POS)* determines the tagging of part of speech, (b) *Text chunking* aims at grouping adjacent words in a sentence; (c) *Word Sense Disambiguation (WSD*) tries to resolve the ambiguity in the meaning of single words or phrases. (d) *Parsing* produces a full parse tree of a sentence.

Successful applications of text mining methods in quite diverse areas are patent analysis, text classification in news agencies, bioinformatics, spam filtering, explorative data analysis, information visualization, text summarization and topic detection studies.

1.1 Clustering Techniques

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining [16].

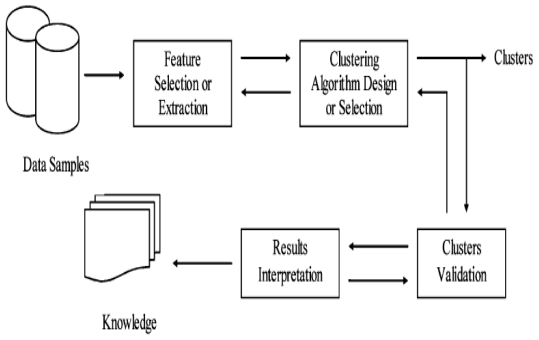

Fig.1 Stages of clustering task.

The different stage of clustering activity is shown in Fig.1. The preprocessed data sample is initially used for further clustering task. Either feature selection or extraction can be used to obtain an appropriate set of features to use in clustering. Pattern proximity is usually measured by a distance function defined on pairs of patterns. Euclidean distance, Minkowski distance, Manhattan distance and Supremum distance are used to calculate the dissimilarity between data objects. Whereas Cosine similarity, Pearson correlation, Bregman divergence, Mahalanobis distance used for similarity measure between data objects. All the metrics are chosen carefully based on feature types. The clusters are generated is assessed for cluster validity. Experts in the relevant fields interpret the data partition. Further experiments can be made to guarantee the reliability of extracted knowledge. To evaluate the quality of clustering measures adopted are Statistical measures, Mean Square Error, Silhouette Coefficient, purity, entropy and other such measures. Normalized Mutual Information (NMI) is clustering evaluation measure is suitable for document clustering. The clustering techniques are classified as in Fig 2. It is broadly categorized as [25]:

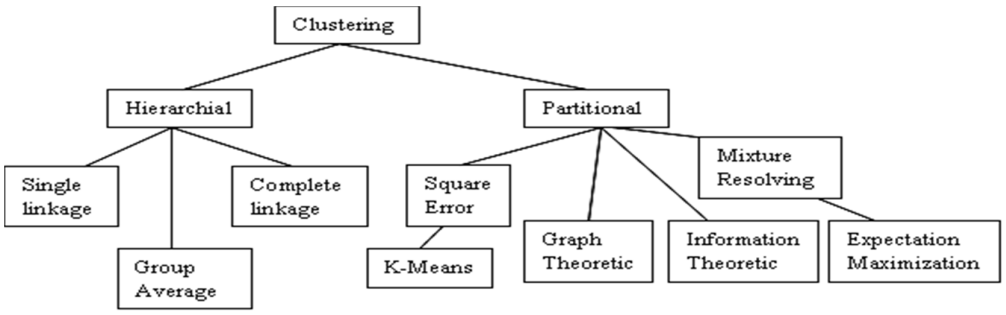

Fig.2 Taxonomy of clustering approaches

Partitional Clustering (division data objects into non-overlapping subsets or clusters) and *Hierarchical clustering* (set of nested clusters organized as a hierarchical tree which maintains class-subclass relationship).Several clustering algorithms used so far are

(a)Hierarchical comprises Agglomerative (Single linkage, complete linkage, group average linkage, median linkage, centroid linkage, Ward's method) and Divisive.
(b)Squared Error-Based (Vector Quantization) comprising K-means.
(c)Mixture Densities comprises Gaussian mixture density decomposition (GMDD), Expectation Maximization (EM).
(d)Graph Theory-Based comprising Chameleon, Delaunay triangulation graph (DTG).
(e)Fuzzy clustering comprises Fuzzy C-Means (FCM), Mountain Method (MM), and Fuzzy C-Shells (FCS).
(f)Neural Networks-Based comprises Learning Vector Quantization (LVQ), Self Organizing Feature Map (SOFM), Self-Splitting Competitive Learning (SPLL).
(g) Large-Scale Data sets comprising CLARANS, BIRCH, DBSCAN, and DENCLUE.

1.2 Semi-Supervised Learning

Unsupervised learning is a class of problems in which one seeks to determine how data are organized. Data mining methods employed this learning strategy to preprocess data. It is distinguished from supervised learning (and reinforcement learning) in that the learner is given only unlabeled examples. Unsupervised learning helps to understand data.

Supervised learning deduces a function from training data. The training data consist of pairs of input objects (typically vectors) and desired outputs. The output of the function can be a continuous value (called regression), or can predict a class label of the input object (called classification). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output). To achieve this, the learner has to generalize from the presented data to unseen situations in a "reasonable" way. Supervised learning enables predictive model testing.

Semi-supervised learning is a technique that makes use of both labeled and unlabeled data for training -typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Unlabeled data, when used in conjunction with a small amount of labeled data, can sometimes produce considerable improvement in learning accuracy. It is adopted in document clustering for significant increase in clustering performance by accuracy and time.

II. Review on Clustering Strategies in Text Mining

The review on clustering strategies falls under three categories: coclustering, constrained coclustering with unsupervised 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 information efficiently. 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, co-clustering 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.

Coclustering 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 K-means and co-clustering algorithms such as information theoretic [13] and minimum sum-squared residue co-clustering [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 also handles missing value prediction and compression of categorical data matrices. Kullback-Leibler divergence (KL-divergence) on text is defined on two multinomial distributions and has proven to be very effective in co-clustering text [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 co-clustering. 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 c-means (FCM) clustering algorithms defined in paper [34] are the well-known and powerful methods in cluster analysis.

2.2 Constrained Co-Clustering with unsupervised constraints

Generally clustering (unsupervised learning) doesn't use information (e.g. labels) as to where each instance should be placed within partition. This lead to Constrained clustering which is an approach to semi supervised learning. Constrained clustering is used to increase document cluster performance by guiding algorithm towards appropriate data partitioning. Constraints are got from background knowledge which handles semantic relationship. Generally pair wise constraints like must-link, cannot-link constraints are adopted, but interval constraints can also be used along for co-cluster discovery in ordered dimensions [23]. Since generating constraints manually or partially is time consuming, thereby using unsupervised method is found to be better. A Penalized Matrix Factorization (PMF) algorithm for constrained semisupervised clustering is used to co-cluster dyadic and multi –type data sets with inter-type and intra-type relationship information constraints [30]. Semisupervised NMF (SS-NMF) based framework to incorporate prior knowledge into heterogeneous data coclustering. Some well-established approaches such as probability based coclustering, information-theoretical coclustering, and spectral coclustering can be considered as variations of this method under certain conditions [7]. A SCM(spectral constraint modeling) is proposed to find the optimal co-clusters by incorporating penalty to the co-clustering assignments that violate the constraints. It is formulated as a new trace minimization problem for finding the globally optimal solution [27]. It uses constraint matrix for bipartite graph representation of co-clusters.

Without using unsupervised learning methods, knowledge from word side can influence the clustering of documents by using non-negative matrix factorization (NMF) model [17]. Sentiment classification refers to the task of automatically identifying whether a given piece of text expresses positive or negative opinion towards a subject at hand [18]. It uses a standard approach to manually label documents with their sentiment orientation and then apply off-the-shelf text classification techniques.

Natural language text contains much information that is not directly suitable for automatic analysis by a computer [21]. The main task is to extract parts of text and assign specific attributes to it. As an example consider the task to extract executive position changes from news stories:"Robert L. James, chairman and chief executive officer of McCann-Erickson, is going to retire on July 1st. He will be replaced by John J. Donner, Jr., the agency's chief operating officer." In this case we have to identify the following information: Organization (McCann-Erickson), position (chief executive officer), date (July 1), outgoing person name (Robert L. James), and incoming person name (John J. Donner, Jr.).

For automatic generation of document constraints, the overlapping named entities concept is used [36]. One such application used for document constraints extraction is Named Entity (NE) extractor. Named Entity Extractor is an information extraction tool which uses recognition of known entity names (for people and organizations, place names, numerical expressions etc).For automatic generation of word constraints, semantic distance is used. WordNet is a lexical database that groups English words to set of synonyms called Synsets. It is open source (open multilingual WordNet) that

combines both dictionary and thesaurus. But the Pairwise constraints generated from data source may be noisy and inaccurate. To handle the situation a generalized maximum entropy model is proposed to learn from noisy side information [35]. Thus with the help of automatic generated constraints into coclustering is proved to have increased performance better than traditional constrained clustering strategies.

2.3 Semi-Supervised Clustering

Semi-supervised clustering methods: Semi-supervised clustering with labeled seeding points and Semi-supervised clustering with labeled constraints.

An initial seed clusters generated using labeled data as well as the using constraints generated from labeled data to guide the clustering process [3]. It uses Seeded-Kmeans and Constrained-Kmeans semi-supervised clustering algorithms that use labeled data to form initial clusters and constrain subsequent cluster assignment. Both methods can be viewed as instances of the EM algorithm, where labeled data provides prior information about the conditional distributions of hidden category. A detailed analysis of performance degradation of more unlabeled data in situations where labeled data can be useful to classification, so this leads to better understanding of semisupervised learning by focusing on maximum-likelihood estimators and generative classifiers [10]. Expectation-Maximization (EM) to learn classifiers that take advantage of both labeled and unlabeled data [22]. EM is a class of iterative algorithms for maximum likelihood or maximum a posteriori estimation in problems with incomplete data. EM performs hill-climbing in data likelihood space, finding the classifier parameters that locally maximize the likelihood of both the labeled and the unlabeled data.

Cobweb is a hierarchical incremental algorithm that uses background knowledge about pair of instance to constrain their clustering which suffer with drawback that majority of assumptions made are false. To handle the situation, K-means clustering is used [29]. K-means is partitional batch algorithm which uses background knowledge about instance level constraints. Inorder to incorporate constraints into clustering COP-K-Means (COnstrained Pairwise -K Means) algorithm was proposed to handle soft constraints. Each clustering algorithms requires good metric for better performance. The choice of distance metric is based on the application for clustering uses learning strategies [32]. Integrating constraints and metric learning in semi-supervised clustering in a uniform, principled framework using K-means and EM algorithm showed better results in paper [6]. It uses MPCK-Means (Metric Pairwise Constraints) where metric used is Euclidean distance and Pairwise constraints are must-link and cannot-link. A probabilistic model on semi-supervised clustering based on Hidden Markov Random Field (HMRF) modeling for combining distance measure and constraints is shown in Figure 3 [36][5][2]. An efficient learning method that exploits the sequential structure is the Hidden Markov models (HMM) were successfully used for named entity extraction [4], which is a probabilistic model to incorporate supervision into prototype based clustering. A probabilistic clustering based on Gaussian mixture models (GMM) of the data distribution that express clustering preferences in a prior distribution over assignments of data points to clusters [20]. This prior penalizes cluster assignments according to the degree with which they violate the preferences and also the model parameters are found to fit with the expectation-maximization (EM) algorithm. In special Cases for bridging the knowledge gap between in-domain and out-of-domain documents is handled using coclustering based classification (CoCC) algorithm [11]. A constrained information-theoretic coclustering (CITCC) algorithm that combines the benefits of information-theoretic co-clustering and constrained clustering [28]. It uses a two-sided hidden Markov random field (HMRF) to model data with both the document and word constraints and alternating expectation maximization (EM) algorithm to optimize the constrained coclustering model.

Yangqiu Song et al. [36] proposed an approach called constrained information-theoretic co-clustering (CITCC) which is extension of paper [28]. It integrates constraints into the information theoretic co-clustering (ITCC) framework where KL-divergence is adopted to better model textual data. The constraints are modeled with two-sided hidden Markov random field (HMRF) regularizations as shown in Fig.3 and alternating expectation maximization (EM) algorithm to optimize the model. Unsupervised constraints are included using Named Entities (NE) for document constraints and WordNet for word constraints for improving clustering performance. The proposed approach additionally handles two-class problem and experiments on 20 newsgroups(NG20) dataset shows significant increase in accuracy of constraints generation, thereby leading to improved clustering performance. The performance of CITCC is evaluated against various clustering algorithms such as Kmeans, constrained Kmeans (CKmeans), Semi-NMF (SNMF), constrained SNMF (CSNMF), Tri-factorization of Semi-NMF (STriNMF), constrained STriNMF (CSTriNMF) and ITCC, using normalized mutual information (NMI)-based measure. The quality of the derived unsupervised document constraints was quite high (95.6 percent) when increasing number of

overlapping NEs. Also under the non-parametric Mann-Whitney U test, CITCC performed significantly better than ITCC and the constrained version took advantage of the NE constraints to improve its clustering performance over non-constrained version.

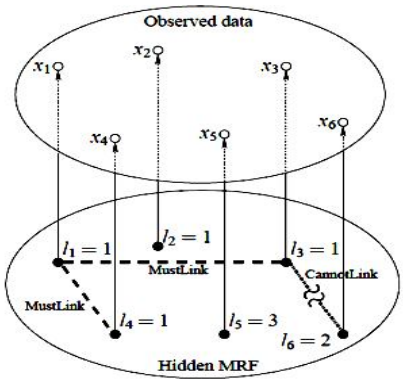

Fig 3. Hidden Markov Random Field

For unsupervised word constraints, a predefined threshold value of WordNet distance is set (0.05 to 0.5) and results show that number of constraints increased significantly on increasing the threshold of WordNet distance. The clustering results were better when the threshold was smaller, e.g., smaller than 0.1 but as the threshold increased the derived constraints became noisy hurting the performance of constrained clustering. This shows the need for better unsupervised word constraint generated.

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: coclustering, constrained coclustering 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.

REFERENCES

[1] Banerjee.A, Dhillon.I, Ghosh.J,. Merugu.S, and Modha.D.S (2007), "A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986.

[2] Basu S., Bilenko M., and Mooney R.J. (2004), "A Probabilistic Framework for Semi-Supervised Clustering," Proc. SIGKDD, pp. 59-68.

[3] Basu.S, Banerjee A., and Mooney R.J. (2002), "Semi-Supervised Clustering by Seeding," Proc. 19th Int'l Conf. Machine Learning (ICML), pp. 27-34.

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

[5] Bilenko M. and Basu S.(2004), "A Comparison of Inference Techniques for Semi-Supervised Clustering with Hidden Markov Random Fields," Proc. ICML Workshop Statistical Relational Learning (SRL '04).

[6] 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.

[7] 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.

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

[9] 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).

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

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

[19] 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.

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

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

[22] 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.

[23] 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.

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

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

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

[27] 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.

[28] 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).

[29] Wagstaff K., Cardie C., Rogers S., and Schro¨ dl S. (2001), "Constrained Kmeans Clustering with Background Knowledge," Proc. 18th Int'l Conf. Machine Learning (ICML), pp. 577-584.

[30] Wang F., Li T., and Zhang C. (2008), "Semi-Supervised Clustering via Matrix Factorization", Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 1-12.

[31] Wang.P, Domeniconi.C , and Laskey K.B. (2009), "Latent Dirichlet Bayesian Co-Clustering," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pp. 522-537.

[32] Xing E.P., Ng A.Y., Jordan M.I., and Russell S.J. (2002), "Distance Metric Learning with Application to Clustering with Side-Information," Proc. Advances in Neural Information Processing Systems Conf., pp. 505-512.

[33] Xu.W, Liu.X, and Gong.Y (2003), "Document Clustering Based on Non-Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 267-273.

[34] Yang M.-S. (1993),"A Survey of Fuzzy Clustering", Mathl. Comput. Modelling Vol. 18, No.11, pp. 1-16.

[35] Yang T., Jin R., and Jain A.K. (2010), "Learning from Noisy Side Information by Generalized Maximum Entropy Model," Proc. Int'l Conf. Machine Learning (ICML), pp. 1199-1206.

[36] Yangqiu Song, Shimei Pan, Shixia Liu, Furu Wei (2013), "Constrained Text Coclustering with Supervised and Unsupervised Constraints", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 6.