Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 15 Issue 2 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Nomenclature and Contemporary Affirmation of the Unsupervised Learning in Text and Document Mining

By Annaluri Sreenivasa Rao & Prof. S. Ramakrishna

MREC Hyderabad, India

Abstract- Document clustering is primarily a method applied for an uncomplicated, document search, analysis and review of content or is a process of automatic classification of documents of similar type categorized to relevant clusters, in a clustering hierarchy. In this paper a review of the related work in the field of document clustering from the simple techniques of word and phrase to the present complex techniques of statistical analysis, machine learning etc are illustrated with their implications for future research work.

Keywords: *document classification, document clustering, similarity measure, accuracy, classifiers, clustering algorithms.*

GJCST-C Classification : *H.2.8*

NOMENCLATUREANDCONTEMPORARYAFFIRMATIONOFTHEUNSUPERVISEDLEARNINGINTEXTANDDOCUMENTMINING

Strictly as per the compliance and regulations of:

Nomenclature and Contemporary Affirmation of the Unsupervised Learning in Text and Document Mining

Annaluri Sreenivasa Rao [α] & Prof. S. Ramakrishna [σ]

Abstract- Document clustering is primarily a method applied for an uncomplicated, document search, analysis and review of content or is a process of automatic classification of documents of similar type categorized to relevant clusters, in a clustering hierarchy. In this paper a review of the related work in the field of document clustering from the simple techniques of word and phrase to the present complex techniques of statistical analysis, machine learning etc are illustrated with their implications for future research work.

Keywords: document classification, document clustering, similarity measure, accuracy, classifiers, clustering algorithms.

I. Introduction

Document clustering [1], [2], [3], [4] techniques find relevance in a wide range of tasks from a simple search with a few terms to vast information retrieval processes. The early document clustering techniques used were developed for typically enhancing information retrieval systems [5], were designed to find documents according to the query type, however could not perform the task of creating a query, generate a synopsis of the documents, or provide an interface to the search results. The progress of internet, digital libraries, news sources and company-wide intranets has made available huge volumes of text documents. The tremendous increase in the already quantum size of web data and the classification of the web documents into relevant and moderate number of clusters has led to the development of large number of web clustering engines and high performing clustering algorithms.

The process of document clustering involves four stages which are,

i) Data collection, crawling to accumulate the documents, indexing the set of documents in a structured fashion, filtering of data with techniques of tokenization, stop words removal and stemming, lemming etc.
ii) preprocessing where the data is represented in suitable form, vector etc. and measurable factors applied to determine the similarity,
iii) Document clustering where a clustering technique and an efficient clustering algorithm are identified for clustering based on preset criteria and
iv) Post processing involving applications of business and scientific requirements adaptation of the document clustering technique.

The applications of document clustering are of diverse nature such as,

i) Creation of document taxonomies
ii) IR process of search, accessing and collection [6], Similar documents identification, review and classification of results [7], automatic topic extraction [8], content summarization
iii) Recommendation System,
iv) Search Optimization, etc. For instance the processes are used enormously in the data classification process such as Google Web Directory, Social media data classification etc.

The clustering techniques though being studied since several years, still face many of the same challenges. These challenges [9,10] of document clustering are mostly of,

i) Huge volume of data,
ii) The high dimensionality of the feature space,
iii) A feasible clustering method in terms of constraints such as cluster quality and performance and
iv) Representing the results in an effective browsing interface. The current challenges associated with text clustering are the requirement of dynamic clustering techniques to incrementally update clusters as new data is added [11,12]. For instance the social media has to generate user specific content [13] instantly and this requires real time data clustering methodologies.

The remainder of this paper is organized as follows. In Section 2 we discuss the "Taxonomy" of document clustering, in Section 3 the "Contemporary literature work of clustering techniques" are evaluated and Section 4 gives the "Conclusion" of the paper.

II. Taxonomy

The clustering functionality can be expressed as a function comprising of a document set mapped to a

Author α : Associate Professor at Department of Computer Science, MRCE, Hyderabad, Telangana State, India.
e-mail: annaluri.rao@gmail.com
Author σ : Prof. S. Ramakrishna Professor, Department of Computer Science, Sri Venkateswara University, Tirupathi, Andhra Pradesh, India.
e-mail: drsramakrishna@gmail.com

set of clusters. Based on specified constraints the minimum and maximum of the function defines the clustering difficulty and algorithms applied over the similarity criteria determine the clustering quality.

The preprocessing step of clustering for finding the document similarity is determined with methods based on the following strategies, (i) phrase or pair-wise methodology, (ii) tree form data depiction, (iii) component dependent data depiction, (iv) semantic relation dependent documents depiction, (v) concept and feature vector dependent depiction.

The clustering methods of are generally of two types, 1) Word patterns and phrases based 2) Feature based.

The clustering methods algorithms are mostly of two types 1) hierarchical methods and 2) partitioning methods (non hierarchical) [14, 15, 16]. The hierarchical algorithms for clustering represent data sets as a cluster tree and are of two types 1-1) agglomerative [17] 1 - 2) divisive hierarchical clustering methods. Partitional clustering algorithms [17] are of two types, 2-1) iterative 2 - 2) single pass methods. K means and its variants etc. are the popular partitioning methods. The hierarchical clustering algorithms are considered efficient than the remaining algorithms [18] however due to their inherent complexness they are not applicable to huge document sets.

The techniques for determining inter-cluster similarity in classification [19 20] ex. single link and for enhancing the value of the clusters where the cluster size differs or fluctuates by a huge factor [17], especially in case of high performing clustering algorithms have been studied widely in recent years.

The widely used document clustering methods are Spectral Clustering, LSI dependent cluster development and NMF technique based clustering. The Spectral clustering methods [21] are LPI, LSI etc. Latent semantic indexing (LSI) [22] a feature extraction approach [23] tries to optimize the documents space compared to the given document and is a widely used linear document indexing method [24]. LSI is inapplicable for processes with a high range of documents [24] and similarly spectral clustering when used in a large dimensional space the dimensionality reduction is very costly which limits its usability.

The word patterns and phrases based approaches are the traditional strategies where the clustering is dependent on the documents features such as words, phrases and sequences [25, 26]. These methods are of four types, 1-1) Clustering with Frequent Word Patterns 1-2) Application of Word Clusters in Document Clusters 1-3) Co-clustering Words and Documents, Co-clustering with graph partitioning and Information- Theoretic Co-clustering 1-4) Clustering based on Frequent Phrases. The technique VSM is used in almost all the document clustering methods used nowadays [27]. The vector space model is a data model for representing the terms related to the words in a document as a feature vector.

The features based clustering approaches are of two types 2-1) Feature Extraction 2-2) Feature Selection.

The Feature Extraction approaches are based on the algorithm of two types i) linear and ii) nonlinear techniques. The models of linear type algorithms are unsupervised PCA, OCA, MMC etc. The examples of non linear algorithms are LLE, Laplacian Eigenmaps, and ISOMAP etc. The linear methods show better operational performance in contrast to nonlinear approaches, however underperform in the clustering of huge and complicated data of the internet. The feature extraction technique finds applications in the fields of IR based on human language learning ability, comparing reviewed and submitted papers, of various languages or networks and filter of data. Feature selection algorithms are of two types, 2-2-1) Feature Ranking that is metric based and 2-2-2) Subset Selection from the possible features. The feature selection algorithms are of two categories, i) supervised and ii) unsupervised. The supervised feature selection algorithms are the most researched as well as used and they are IG, CHI, and MI. The unsupervised methods that are most popular are, i) DF-based selection dependent on term strength and ranking dependent on entropy or term contribution, ii) LSI-based method and iii) NMF based method. These techniques of unsupervised approach such as, decision trees, statistics, NLP and ML are being used in BI or analytics, in neural networks for developing AI or bio neural networks, for developing systems of AI that are rule based for intelligent content development, database development, information retrieval and automatic grouping of web documents with Enterprise Search engines or open source software's in web mining or text mining.

The strategies of feature selection used mostly are i) wrapper, ii) filter and iii) embedded methods [28] however a study [29] has shown, the methods of supervised feature selection dependent on algorithms using the filter metric IG, are most efficient over others techniques.

III. Contemporary Affirmation of the Recent Literature

An approach of bisecting k-means algorithm proposed by Steinbach, M, Karypis, G, & Kumar, V [14] breaks up a large cluster into small clusters repetitively to generate k numbers of clusters of huge similarity for filtering the clusters and collecting similar texts based on the method.

A technique called CCA [30] widely used in the emerging technologies of ML etc applies correlation for measuring the similar features in a document. However, CCA has its own limitations in clustering.

An approach of spectral clustering based on graph partitioning strategy called LPI [31] proposed however fails in feature selection and comprises of the existing problems of distance based clustering documents.

An approach for document clustering called Frequent Term based Clustering or HFTC [32] is a topic of extensive research. However it is not scalable for huge data or of documents.

A technique known as Hierarchical Document Clustering using Frequent itemsets (FIHC) approach proposed by Fung, B., Wang, K., Ester, M, is discussed in [33]. The strategy of FIHC though performs better than HFTC underperforms in clustering efficiency when compared to existing approaches such as UPGMA and Bisecting K-means.

The TDC algorithm technique based on closed frequent itemsets for clustering is proposed by Yu, H., Searsmith, D., Li, X., Han, J [34]. The algorithm performs better compared to HFTC and FIHC however the use of closed itemsets makes it avoidable.

A strategy of Hierarchical Clustering using Closed Interesting Itemsets, referred to as HCCI proposed by Malik, H.H., Kender, J.R [35], is the best clustering method available. However the technique may cause information loss.

An approach based on PSSM histogram by Gad and Kamel [36] combines the text semantic with the process of incremental clustering and measures the similarity of the documents for adjusting the insertion order of the documents in the cluster for quality.

An improved incremental clustering technique for an efficient clustering algorithm proposed by Gavin and Yue [37] improves categorization of web data incrementally. The method based on cluster specific multiple information anew document is assigned to a cluster.

An approach for improving text clustering mining by Shehata, S, Fakhri, K, & Mohamed S, S. [38] outperforms the existing techniques such as HAC, k-NN etc.

A progressive clustering algorithm by Liu, Y, Ouyang, Y, Sheng, H, & Xiong, Z. (2008) [39] based on Cluster Average Similarity Area determines the cluster coherence and progressively assigns the new data items to the clusters.

A technique for enhancing the clustering functionality based on the partial disambiguation of words by means of their PoS [40] is recommended by the developers as the approach finds the inefficiency of considering synonyms and hypermy my for selecting the right sense of the word disambiguated solely by PoS tags.

The CFWS technique proposed by Y. LI, and S.M. Chung, enhances the capability to process the document, considering the word sequences apart from the words [41].

The technique of non linear representation of the data by J.B. Tenenbaum, V. de Silva, and J.C. Langford [42] keeps specific local data simultaneously based on the optimization factors however is associated with high complexity.

A study of the approaches for reducing the complexity of feature extraction based on a new technique called approximation algorithm [43], [44], [45] is found to be good.

A software for automatically retrieving information from websites by Zamir O Etzioni [46] is designed for websites comprising of vast amount of data

The approach of integrating clustering and feature selection for text clustering based on the semantic relation of the text documents with ontology was proposed by Thangamani.M and P.Thangaraj in [47]. The approach minimizes dimensionality and improves feature selection.

The clustering technique, for finding the clustering quality based on WordNet [48] phrasal noun and semantic relationships [49] shows better performance with hyperny my based strategy compared to other noun phrases.

A system for determining the ontology related semantic relations of the term or word and associated weight measure is given by Prof. K. Raja, C. Prakash Narayanan [9]. However the technique has dimensionality and other problems.

A description of the task of Ontology based automatic categorizing of web documents [50] and the scope of Ontology in improving the current machine learning and IR approaches is given by Andreas Hotho. The integration of ontology's for combining various information types of multiple resources by Young-Woo et al. in the paper [51].

The process of using domain specific ontology's for enhancing performance of text classification where text learning and IR are used to generate ontology's with minimum user interaction is given in [52, 53].

The methods utilizing Wikipedia ontology for improving primarily the document depiction and cluster quality by Gabrilovich and Markovitch [54] and a further extension provided a structure based on the Wikipedia guidelines and groups [55, 56]. The Wikipedia ontology is most relevant as it is applicable to a large cross section of domains and also restructured on a regular basis.

A technique for feature selection in text clustering based on supervised feature selection on the intermediary clustering outcomes by Xu, J. Xu, B [57] generates a efficient subset for classification. The suggested techniques performance is efficient compared to manual process.

A technique of feature selection dependent on the ACO algorithm by M. Janaki Meena,K.R.

Chandran,J. Mary Brinda," [58] is a unique method. Comparative tests of the approach with existing chi-square and CHIR techniques shows the proposed approach achieves better performance in FS.

An entropy based FS approach i.e. a filter solution [59] tested with various data types that reduces dimensionality and is efficient in finding the subset of major features.

A feature co-selection method called MFCC (multi type feature co-selection), proposed by Shen huang, Zheng Chen, Yong Yu, and Wei-Ying main [60] shows enhanced clusters performance of web documents based on the outcomes of intermediate clustering.

A method to remodel the matrix of data similarity as a bi-stochastic matrix prior to executing algorithms by F. Wang, P. Li, and A. C. K Aonig showed better clustering performance [61].

The techniques of document clustering that are term based for clustering in dynamic environments, is given in [11] by Wang, X, Tang, J, & Liu, H, synonyms and hypermy my by Bharathi and Vengatesan [62], Synonyms and Hyponyms, Nadig, R, Ramanand, J, & Bhattacharyya, P in [12]. These approaches are however not applicable to technically similar documents.

A document clustering approach [63] dependent on phrases and the STC technique by O. Zamir, O. Etzioni, O. Madanim, and R.M. Karp builds the clusters on the common documents suffixes. The method though efficient in cluster quality however is associated with high amount of term redundancy.

A study of the TF-IDF method of clustering [64], term frequency dependent algorithms [65] and a review of clustering algorithms [66] showed that majority of clustering approaches are TF-IDF based, however associated with several problems.

The NMF (Nonnegative Matrix Factorization) technique in text classification [67], improved clustering performance compared to the existing approaches [68], relationship study of NMF techniques with earlier clustering techniques [69], [70] [71]. A review of established techniques of NMF such as multiplicative updates [72], projected gradients [73] though efficient however are associated with the problems of memory for huge datasets streamed and not disk based [74]. To overcome these problems, approaches such as random projections [61, 75] and sketch/sampling algorithms [76] have been proposed. An NMF based technique by Li and Zhu in 2011 [77] for research specific documents minimizes high dimensionality, finds relevant topics for clustering and shows performance efficiency in classification comparatively. A study of the online algorithm based on Nonnegative Matrix Factorization [78], a NMF based method that uses features based on weights and similar cluster property by Sun Park, Dong Un An, Choi Im Cheon [79] performs comparatively more efficiently than the remaining NMF based strategies.

IV. Conclusion

In this paper we analyzed several techniques developed for clustering documents with their applications and relevance in terms of today's requirements. The task of developing perfect strategies for classification of varied forms and types of documents for a near optimal solution or finding accurate ways of assessing the quality of the performed clustering though is impossible and is increasing in its complex nature, the field today deals with extraordinary tasks like granular taxonomies generation, sentiment analysis and document summarization for generating reliable and relevant insights applicable to several fields. In conclusion we can say document clustering is going to be widely studied and will find relevance in a number of newer areas.

References Références Referencias

1. S. Kotsiantis and P. Pintelas, "Recent Advances in Clustering: A Brief Survey," WSEAS Trans. Information Science and Applications, vol. 1, no. 1, pp. 73-81, 2004.
2. W. Xu and Y. Gong, "Document Clustering by Concept Factorization," Proc. Int'l Conf. Research and Development in Information Retrieval, pp. 202-209, July 2004.
3. S. Siersdorfer and S. Sizov, "Restrictive Clustering and Metaclustering for Self-Organizing Document Collections," Proc. Int'l Conf. Research and Development in Information Retrieval, pp. 226-233, July 2004.
4. Brian S. Everitt, Sabine Landau, and Morven Leese. Cluster Analysis. Oxford University Press, fourth edition, 2001.
5. Van Rijsbergen, CJ. Information Retrieval (Secondth ed.). London: Buttersworth. 1989.
6. C. Carpineto, S. Osi´nski, G. Romano, and D. Weiss. A survey of web clustering engines. ACM Comput. Surv., 41(3):1–38, 2009.
7. X. Liu and W. B. Croft. Cluster-based retrieval using language models. In Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 186–193, 2004.
8. J. Silva, J. Mexia, A. Coelho, and G. Lopes. Document clustering and cluster topic extraction in multilingual corpora. In Proceedings of the 1st IEEE International Conference on Data Mining (ICDM), pages 513–520, 2001.
9. Prof. K. Raja, C. Prakash Narayanan, "Clustering Technique with Feature Selection for Text Documents", Proceedings of the Int.Conf. on

Information Science and Applications ICISA 2010 6 February 2010, Chennai, India.

10. Fabrizio Sebastiani "Machine Learning in Automated Text Categorization" ACM Computing Surveys, Vol. 34, No. 1, March 2002
11. Wang, X, Tang, J,& Liu, H Document clustering via matrix representation. In 11th IEEE International Conference on DataMiningICDM2011 (pp. 804–813), 2011.
12. Nadig, R, Ramanand, J, & Bhattacharyya, P. (2008).Automatic evaluation of Word Net synonyms and hypermy my India: Proceedings of ICON-2008, 6th International Conference on Natural Language Processing.
13. A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In Proceedings of the 16th International Conference on World Wide Web (WWW), pages 271{280, 2007.
14. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. KDD Workshop on Text Mining‟ 2000.
15. P. Berkhin. 2004. Survey of clustering data mining techniques [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.
16. Xu Rui.. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3):pp. 634-678, 2005.
17. B.C.M.Fung, K.Wan, M.Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets", SDM‟03.
18. I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning, 42(1):143{175, 2001.
19. M. Khalilian & N. Mustapha (2010), "Data Stream Clustering: Challenges and Issues", In Proceedings of the International Multiconference of Engineers and Computer Scientists IMECS 2010, Hong Kong, pp. 978-988.
20. R. Martinez-Morais, F.J. Alfaro-Cortes, & J. L. Sanchez (2010), "Providing QoS with the Deficit Table Scheduler", IEEE Transactions on Parallel and Distributed Systems, Vol. 21, No. 3, pp. 327-341.
21. A.Y. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," Advances in Neural Information Processing Systems 14,pp. 849-856, Cambridge, Mass.: MIT Press,2001.
22. S.T. Dumais, "Latent Semantic Indexing (LSI) and TREC-2,"Proc.Second Text Retrieval Conf. (TREC),pp. 105-116, 1993.
23. LSA @ CU Boulder, http://lsa.colorado.edu/, 2010.
24. S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis,"J. Am.Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.
25. Hung, C. and Xiaotie, D., "Efficient Phrase-Based Document Similarity for Clustering," IEEE Transaction on Knowledge and Data Engineering, vol. 20, no. September, pp. 1217- 1229, 2008.
26. Soon, M. C. , John, D. H., and Yanjun, L., "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol. 64, pp. 381-404, 2008.
27. Aas, K,& Eikvil, L.(1999). Text Categorisation: A Survey. Technical Report 941. Oslo Norway: Norwegian Computing Center.iteseer.ist.psu.edu/aas99text.html.
28. Barak Chizi (Tel-Aviv University, Israel); Lior Rokach (Ben-Gurion University, Israel); Oded Maimon (Tel-Aviv University, Israel) "a survey of feature selection techniques"1888-1895 pp. John Wang (Ed.) (Montclair State University, USA), DOI:10.4018/978-1- 60566-010-3.ch289, ISBN13:9781605660103, 2009.
29. Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. 14th Int'l Conf. Machine Learning, pp. 412-420, 1997.
30. D.R. Hardoon, S.R. Szedmak, and J.R. Shawe-taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods,"J. Neural Computation, vol. 16, no. 12, pp. 2639-2664, 2004.
31. D. Cai, X. He, and J. Han, "Document Clustering Using Knowledge and Data Eng.,vol. Locality Preserving Indexing,"IEEE Trans. 17, no. 12, pp. 1624-1637, Dec. 2005.
32. Beil, F., Ester, M., Xu, X.: Frequent Term-based Text Clustering, In Proc. of Intl. Conf. on Knowledge Discovery and Data Mining,. (2002).
33. Fung, B.C.M., Wang, K., and Ester, M., "Hierarchical document clustering using frequent Itemsets," Proceedings of SIAM International Conference on Data Mining, 2003.
34. Yu, H., Searsmith, D., Li, X., Han, J.: Scalable Construction of Topic Directory with Nonparametric Closed Termset Mining, In Proc. of Fourth IEEE Intl. Conf.on Data Mining. (2004).
35. Malik, H.H., Kender, J.R.: High Quality, Efficient Hierarchical Document Clustering Using Closed Interesting Itemsets, In Proc. of IEEE Intl. Conf. on Data Mining. (2006).
36. Gad, WK, & Kamel, MS. (2010). Incremental clustering algorithm based on phrase- semantic similarity histogram. Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, 11(14), 2088–2093.
37. Gavin, S, & Yue, X. (2009). Enhancing an incremental clustering algorithm for Web page collections. IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies,81–84.

38. Shehata, S, Fakhri, K, & Mohamed S, S. (2010). An efficient concept-based mining model for enhancing text clustering.IEEE Transactions On Knowledge And Data Engineering, 22(10), 1360–137.
39. iu, Y, Ouyang, Y, Sheng, H, & Xiong, Z. (2008).An Incremental Algorithm for Clustering Search Results, IEEE International Conference on Signal Image Technology and Internet Based Systems(pp. 112–117).
40. Sedding, J., & Kazakov, D. 2004. Wordnet-based text document clustering. 3rd Workshop on Robust Methods in Analysis of Natural Language Data, pp. 104–113.
41. Y. LI, and S.M. Chung. 2005. Text Document Clustering Based on Frequent Word Sequences. In Proceedings of the. CIKM, 2005. Bremen, Germany, October 31- November 5.
42. J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction,"Science,vol. 290, pp. 2319-2323, 2009.
43. J. Weng, Y. Zhang, and W.-S. Hwang, "Candid Covariance-Free Incremental Principal Component Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 25, no. 8, pp. 1034-1040, Aug.2003.
44. K. Hiraoka and M. Hamahira, "On Successive Learning Type Algorithm for Linear Discriminant Analysis,"IEIC Technical Report,(in Japanese), vol. 99, pp. 85-92, 1999.
45. J. Yan, Zhang, B.S. Yan, Z. Chen, W. Fan, Q. Yang, W.Y. Ma, and Q. Cheng, "IMMC: Incremental Maximum, Marginal Criterion,"Proc. 10th ACM SIGKDD,pp. 725-730, 2004.
46. Zamir, O.Etzioni, "Web Document Clustering, A Feasibility Demonstration, " in Proceedings of the 21st International ACM SIGIR Conference on Research and Development. K.Mugunthadevi et al. / International Journal on Computer Science and Engineering (IJCSE)ISSN :
47. Thangamani.M and P.Thangaraj," integrated clustering and feature selection scheme fo textdocuments",J.Comput.Sci.,6:536.541,DOL:10.3844/jcssp.2010.536.541,URL:http://www.thescipub.com/abstract/10.3844/jcssp.2010.536.54.
48. Miller G. 1995. Wordnet: A lexical database for English. CACM, 38(11), pp. 39–41.
49. Zheng, Kang, Kim. 2009. Exploiting noun phrases and semantic relationships for text document clustering. Information Science 179 pp. 2249-2262.
50. Andreas Hotho," Using Ontologies to Improve the Text Custering and Classification Task", Knowledge and Data Engineering Group, University of Kassel, January 14, 2005.
51. Young-Woo Seo, Anupriya Ankolekar, Katia Sycara," Feature Selections for Extracting Semantically Rich Word for Ontology Learning" CMU-RI-TR-04-18. March 2004.
52. B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining", In Proceedings of International Semantic Web Conference (ISWC), pages 264– 278,2002.
53. A. Hotho, S. Staab, and A. Maedche. Ontology-based text clustering. In Proceedings of the IJCAI-2001 Workshop Text Learning: Beyond Supervision, Seattle,USA, August 2001.
54. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, In Proc. of The 20th Intl. Joint Conf.on Artificial Intelligence. (2007)
55. Hu, X., Zhang, X., Lu, C., et al.: Exploiting Wikipedia as External Knowledge for Document Clustering, In Proc. of Knowledge Discovery and Data Mining. (2009).
56. Hu, J., Fang, L., Cao, Y., et al: Enhancing Text Clustering by Leveraging Wikipedia Semantics, In Proc. of 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval. (2008)
57. Xu, J. Xu, B. Zhang, W. Cui, Z. Zhang, W."A new feature selection method for text clustering ",wuhan university journal of natural sciences, 2007, vol 12; number 5, pages 912-916.
58. M. Janaki Meena, K.R. Chandran,J. Mary Brinda," integrating swarm intelligence and statistical data forfeature selection in text categorization" ©2010 International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 11.
59. Manoranjan Dash ,Kiseok Choi ,Peter Scheuermann ,Huan Liu," Feature Selection for Clustering – A Filter Solution" Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)0-7695-1754-4/02 © 2002 IEEE.
60. shen huang, zheng chen, yong yu, and wei-ying ma," multitype features coselection for web document clustering", ieee transactions on knowledge and data engineering, vol. 18, no. 4, april 2006, 1041-4347/06/$20.00 _ 2006 ieee published by the ieee computer society.
61. F. Wang, P. Li, and A. C. KÄonig. Learning a bistochastic data similarity matrix. In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), 2010.
62. Bharathi, G, & Vengatesan, D. (2012). Improving information retrieval using document clusters and semantic synonym extraction.Journal of Theoretical and Applied Information Technology, 36(2), 167–173.
63. O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," Computer Networks, vol. 31, nos. 11-16, pp. 1361-1374, 1999.

64. Salton, G, & Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. Information Processing & Management,24(5), 513–523.
65. Kumar, N, & Srinathan, K. (2009).A New Approach for Clustering Variable Length Documents(Proceedings of the Advanced computing Conference, IEEE, pp. 982–987).
66. Prathima, Y, & Supreethi, KP. (2011). A survey paper on concept based text clustering. International Journal of Research in IT & Management, 1(3), 45–60.
67. D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. Nature ,401: 788-791, 1999.
68. F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. Information Processing and Management, 42(2):373-386, 2006.
69. C. Ding, T. Li, and M. I. Jordan. Convex and seminonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010.
70. C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In Proceedings of the 5th SIAM Int'l Conf. Data Mining (SDM), pages 606- 610, 2005.
71. E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 601-602, 2005.
72. D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In Advances in Neural Information Processing System (NIPS), pages 556-562, 2000.
73. C. J. Lin. Projected gradient methods for nonnegative matrix factorization. Neural Computation,19(10):2756 - 2779.
74. S. Zhong. Efficient streaming text clustering. Neural Networks, 18(5-6):790{798, 2005.
75. F. Wang and P. Li. efficient non-negative matrix factorization with random projections. In Proceedings of the 10th SIAM International Conference on Data Mining (SDM), pages 281-292, 2010.
76. P. Li, K. W. Church, and T. Hastie. One sketch for all: Theory and application of conditional random sampling. In Advances in Neural Information Processing System (NIPS), pages 953-960, 2008.
77. Li, F, & Zhu, Q. (2011). Document clustering in research literature based on NMF and testor theory.Journal of Software, 6(1), 78–82.
78. B. Cao, D. Shen, J. Sun, X. Wang, Q. Yang, and Z. Chen. Detect and track latent factors with online nonnegative matrix factorization. In Proc. International Joint Conference on Artificial Intelligence, pages 2689- 2694, 2007.
79. Sun Park, Dong Un An, Choi Im Cheon, "Document Clustering Method Using Weighted Semantic Features and Cluster Similarity," digitel, pp.185-187, 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, 2010.

This page is intentionally left blank