GRAPH BASED TEXT REPRESENTATION FOR DOCUMENT CLUSTERING

ASMA KHAZAAL ABDULSAHIB

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

SCHOOL OF COMPUTING
COLLEGE OF ARTS AND SCIENCES
UNIVERSITI UTARA MALAYSIA

2015

PERMISSION TO USE

In presenting this dissertation in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this dissertation in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my dissertation.

Requests for permission to copy or to make other use of materials in this dissertation, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

ABSTARK

Kemajuan dalam teknologi digital dan World Wide Web telah membawa kepada peningkatan document digital yang digunakan untuk pelbagai tujuan seperti penerbitan dan Perpustakaan digital. Fenomena ini telah menimbulkan kesedaran untuk mewujudkan teknik-teknik yang lebih berkesan untuk membantu dalam pencarian dan pencapaian teks. Salah satu tugas yang paling diperlukan ialah pengkelompokkan yang boleh mengkategorikan dokumen secara automatik kepada kumpulan yang bermakna. Pengkelompokkan adalah satu tugas yang penting dalam perlombongan data dan pembelajaran mesin. Ketepatan kelompok bergantung erat pada pemilihan kaedah perwakilan teks. Kaedah tradisional memodelkan perwakilan dokumen teks dalam bentuk bag perkataan yang menggunakan teknik frekuensi istilah frekuensi dokumen indeks (TFIDF). Kaedah ini mengabaikan hubungan dan makna perkataan di dalam dokumen. Akibatnya masalah *sparsity* dan semantik yang lazim dalam dokumen teks tersebut tidak dapat diselesaikan . Dalam kajian ini , masalah *sparsity* dan semantik dikurangkan dengan mengusulkan kaedah perwakilan teks berdasarkan graf iaitu graf ketergantungan dengan tujuan untuk meningkatkan ketepatan pengkelompokkan dokumen. Skim perwakilan graf ketergantungan dihasilkan menerusi pengumpulan analisis sintaks dan semantik. Sampel daripada dataset 20 kumpulan berita telah digunakan dalam kajian ini. Dokumen-dokumen teks mengalami pra- pemprosesan dan *parsing* sintaks untuk mengenal pasti struktur ayat. Kemudian semantik perkataan dimodelkan menggunakan graf ketergantungan. Graf ketergantungan yang dihasilkan kemudian digunakan dalam proses analisis kelompok. Teknik K-means telah digunakan dalam kajian ini. Hasil kelompok berdasarkan graf ketergantungan dibandingkan dengan kaedah popular perwakilan teks iaitu TFIDF dan teks perwakilan berasaskan Ontologi. Hasil kajian menunjukkan bahawa graf ketergantungan menghasilkan keputusan baik yang melebihi kedua-dua TFIDF dan teks perwakilan berasaskan Ontologi. Ini membuktikan bahawa kaedah perwakilan teks yang dicadangkan mampu memberi hasil pengkelompokan dokumen yang lebih tepat.

ABSTRACT

Advances in digital technology and the World Wide Web has led to the increase of digital documents that are used for various purposes such as publishing and digital library.This phenomenon raises awareness for the requirement of effective techniques that can help during the search and retrieval of text. One of the most needed tasks is clustering, which categorizes documents automatically into meaningful groups. Clustering is an important task in data mining and machine learning. The accuracy of clustering depends tightly on the selection of the text representation method. Traditional methods of text representation model documents as bags of words using term-frequency index document frequency (TFIDF). This method ignores the relationship and meanings of words in the document. As a result the sparsity and semantic problem that is prevalent in textual document are not resolved. In this study, the problem of sparsity and semantic is reduced by proposing a graph based text representation method, namely dependency graph with the aim of improving the accuracy of document clustering. The dependency graph representation scheme is created through an accumulation of syntactic and semantic analysis. A sample of 20 news group, dataset was used in this study. The text documents undergo pre-processing and syntactic parsing in order to identify the sentence structure. Then the semantic of words are modeled using dependency graph. The produced dependency graph is then used in the process of cluster analysis. K-means clustering technique was used in this study. The dependency graph based clustering result were compared with the popular text representation method, i.e. TFIDF and Ontology based text representation. The result shows that the dependency graph outperforms both TFIDF and Ontology based text representation. The findings proved that the proposed text representation method leads to more accurate document clustering results.

KEYWORDS

Text Representation scheme, Dependency Graph, Document Clustering

ACKNOWLEDGEMENT

بسم الله الرحمن الرحيم

In the name of Allah the most gracious the most merciful First and foremost, Praise to Allah, Lord of the Worlds and prayers and peace are upon the master of messengers the Prophet Mohammed. Our leader in this life until the closing.

I would like to convey my deepest gratitude to my supervisor, Dr. SITI SAKIRA KAMARUDDIN for all continuous guidance and advices given to me in writing up of this dissertation.

Next I would like to thank University Utara Malaysia (UUM) staff. Especially, School of Computing staff for their cooperation with me.

Especial thanks to my husband Amjad Majed and my kids (Noor, Mohammed and Haidr) for their constant support and encouraged me and sacrifice during the production of this dissertation. I would like to thank my father and mother for Permanent their prayer for me to finish this work. Lastly, I want to thank my country, Iraq for the material and moral support for getting the Master Certificate.

DEDICATIONS

To the Most Merciful…

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

الرَّحْمَنُ (1) عَلَّمَ الْقُرْآنَ (2) خَلَقَ الْإِنْسَانَ (3) عَلَّمَهُ الْبَيَانَ (4). سورة الرحمن

In the name of Allah, Most Gracious, Most Merciful.

[1] (Allah) Most Gracious! [2] It is He Who has taught the Quran.

[3] He has created man: [4] He has taught him speech (and

Intelligence). *Quran* 55:1-4.

TABLE OF CONTENTS

C H A P T E R TWO: LITERATURE REVIEW 11

LIST OF FIGURES

LIST OF TABLES

List of Appendices

CHAPTER ONE

INTRODUCTION

1.1 DOCUMENT CLUSTERING

Document clustering is considered a vital technology in the era of internet. It's an essential technique in mining underlying structures in text document data sets. Furthermore, this is a very interesting research topic that has influenced a number of researchers and practitioners from a number of fields, including data mining, machine learning, and information retrieval due to its fundamental role in many of real-world applications (Andrews & Fox, 2007). Text clustering means finding the groups that are related to each other. These groups are collected together in an unstructured formal document. In fact, clustering becomes very famous for its ability to offer an exceptional way of digesting in addition to generalize a good quantity of information. The extracting appropriate feature is considered the basis of clustering. Clustering text documents into category groups is a necessary step in the mining of abundance text data on the Web, indexed and retrieval or incorporate information systems and extract proper feature (concept) of a problem area. Text documents are often represented as high-dimensional, sparse vectors and complex semantics (Dhillon, et al., 2001& Jing, et al., 2005).

In existing clustering methods, a document is often represented as "bag of words" (in BOW model), N-grams (in suffix tree document model), or TF-IDF without considering the natural language relationships between the words (Wang et al.,2011).

The contents of the thesis is for internal user only

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms Mining text data (pp. 77-128): Springer.

Andrews, N. O., & Fox, E. A. (2007). Recent developments in document clustering. Computer Science, Virginia Tech, Tech Rep.

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S .(2005) .An Ontology-based Framework for Text Mining. Paper presented at the LDV Forum.

Balmas, F. (2004). Displaying dependence graphs: a hierarchical approach. Journal of Software Maintenance and Evolution: Research and Practice, 16(3), 151-185.

Balmas, Francoise (2001) Displaying dependence graphs: a hierarchical approach, [1] wcre, p. 261, Eighth Working Conference on Reverse Engineering (WCRE'01).

Beck, F., & Diehl, S. (2013). On the impact of software evolution on software clustering. Empirical Software Engineering, 18(5), 970-1004.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: O'Reilly Media, Inc.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class based n-gram models of natural language. Comput. Linguist., 18(4):467{479, 1992.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). Introduction to information retrieval. An Introduction To Information Retrieval, 151-177.

Chen, Y., & Tu, L. (2007). Density-based clustering for real-time stream data. Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Cui, X., & Potok, T. E. (2005). Document clustering analysis based on hybrid PSO+ K-means algorithm. Journal of Computer Sciences (special issue), 27, 33.

Cimiaon,p.,Hotho,A.,& Staab,S.(2005).Learning Concept Hierarchies from text corpora Using Formal Concept Analysis. J.Artif. Intell.Res(JAIR),24,305-339.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Ann Arbor MI, 48113(2), 161-175.

Chakravarthy, S., Venkatachalam, A., & Telang, A. (2010 .(A graph-based approach for multi-folder email classification. Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on.

Chakrabarti, S. (2003). Mining the Web: Discovering knowledge from hypertext data: Morgan Kaufmann.

Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S., & Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. Statistica Sinica, 12(1), 241-262.

Dolamic, L., & Savoy, J. (2008). Stemming approaches for East European languages Advances in Multilingual and Multimodal Information Retrieval (pp. 37-44): Springer.

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1-2), 143-175.

Dietrich, J., Yakovlev, V., McCartin, C., Jenson, G., & Duchrow, M. (2008). Cluster analysis of Java dependency graphs. Paper presented at the Proceedings of the 4th ACM symposium on Software visualization.

Davidov, D., & Rappoport, A. (2008). Classification of Semantic Relationships between Nominals Using Pattern Clusters. Paper presented at the ACL.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the KDD.

Eikvil, L. (1999). Information extraction from world wide web-a survey.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Fares, M., Oepen, S., & Zhang, Y. (2013). Machine learning for high-quality tokenization replicating variable tokenization schemes Computational linguistics and intelligent text processing (pp. 231-244): Springer.

Gil-García, R., Badia-Contelles, J. M., & Pons-Porrata, A. (2006). A general framework for agglomerative hierarchical clustering algorithms. Paper presented at the Pattern Recognition, 2006. ICPR 2006. 18th International Conference on.

Gardner, S. (2007). Ontology-based information management system and method: Google Patents.

Gil-Garcia, R., & Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. Pattern Recognition Letters, 31(6), 469-477.

Giller, G. L. (2012). The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity. Available at SSRN 2167044.

Han, X., Sun, L., & Zhao, J. (2011). Collective entity linking in web text: a graph-based method. Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Huang, C., Simon, P., Hsieh, S., & Prevot, L. (2007). Rethinking Chinese Word Segmentation: Tokenization. Character Classification, or Word break Identification.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107-145.

Huang, X., & Wu, Q. (2013). Micro-blog commercial word extraction based on improved TF-IDF algorithm. Paper presented at the TENCON 2013-2013 IEEE Region 10 Conference (31194).

Huang, J. Z., & Ng, M. (2006). Text clustering: algorithms, semantics and systems. history, 8, 3.

Hammouda, K. M., & Kamel, M. S. (2004). Efficient phrase-based document indexing for web document clustering. Knowledge and Data Engineering, IEEE Transactions on, 16(10), 1279-1296.

Hjørland, B. (2010). The foundation of the concept of relevance. Journal of the American Society for Information Science and Technology, 61(2), 217-237.

Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6), 773-786.

Harish, B., Guru, D., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2), 110-119.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2), 177-196.

Hotho, A., Staab, S., & Stumme, G. (2003). Ontologies improve text document clustering. Paper presented at the Data Mining, 2003. ICDM 2003. Third IEEE International Conference on IEEE.

Hu, J., Xiong, C., Shu, J., & Zhou, X. (2009). A novel text clustering method based on TGSOM and fuzzy K-means. Paper presented at the Education Technology and Computer Science, 2009. ETCS'09. First International Workshop on IEEE.

Huang, J., Sun, H., Song, Q., Deng, H., & Han, J. (2013). Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network. Knowledge and Data Engineering ،IEEE Transactions on, 25(8), 1876-1889.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. JASIS, 47(1), 70-84.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

Jing, L., Ng, M. K., Xu, J., & Huang, J. Z. (2005). Subspace clustering of text documents with feature weighting k-means algorithm Advances in Knowledge Discovery and Data Mining (pp. 802-812): Springer.

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. Knowledge and Data Engineering, IEEE Transactions on, 19(8), 1026-1041.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data: Prentice-Hall, Inc.

Kaur, M., & Kaur, N. (2013). Web Document Clustering Approaches Using K-Means Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, (35).

Kyle, K., Crossley, S., Dai, J., & McNamara, D. S. (2013). Native Language Identification: A Key N-gram Category Approach. NAACL/HLT 2013, 242.

Karaa, A., & Ben, W. (2013). A NEW STEMMER TO IMPROVE INFORMATION RETRIEVAL. International Journal of Network Security & Its Applications, 5(4).

Kisilevich, S., Mansmann, F., & Keim, D. (2010). P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. Paper presented at the Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application.

Klusch, M., Lodi, S., & Moro, G. (2003). Distributed clustering based on sampling local density estimates.Paper presented at the IJCAI.

Liao, W.-k., Liu, Y., & Choudhary, A. (2004). A grid-based clustering algorithm using adaptive mesh refinement. Paper presented at the 7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining.

Liu, M., & Yang, J. (2012). An improvement of TFIDF weighting in text categorization. International Proceedings of Computer Science & Information Technology, 47.

Luo, D., Ding, C., & Huang, H. (2010). Towards structural sparsity: An explicit l2/l0 approach. Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on IEEE.

Liu, Q., Deng, M., Shi ،Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers & Geosciences, 46, 296-309.

Mahdabi, P., & Crestani, F. The effect of citation analysis on query expansion for patent retrieval. Information Retrieval, 1-18.

Matteucci, M. (2008). A tutorial on clustering algorithms. See at: http://home. dei. polimi. it/matteucc/Clustering/tutorial/html/index. html

Mitchell, B. S., & Mancoridis, S. Clustering module dependency graphs of software systems using the bunch tool.

Montes-y-Gómez, M., López-López, A., & Gelbukh, A. (2000). Information retrieval with conceptual graph matching. Paper presented at the Database and Expert Systems Applications.

Moreira, A., Santos, M. Y., & Carneiro, S. (2005). Density-based clustering algorithms–DBSCAN and SNN. University of Minho–Portugal, Version 1.0, 25.07.

Náther, P. (2005). N-gram based Text Categorization. Lomonosov Moscow State University.

Ma, J., Xu, W., Sun, Y.-h., Turban, E., Wang, S., & Liu, O. (2012). An ontology-based text-mining method to cluster proposals for research project selection. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42(3), 784-790.

Pandit, S. (2008). On a robust document classification approach using TF-IDF scheme with learned, context-sensitive semantics.

Parikh, M., & Varma, T. (2014). Survey on Different Grid Based Clustering Algorithms. International Journa.(2)2 ،

Parimala, M., Lopez, D., & Senthilkumar, N. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology, 31(1).

Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques. Information processing & management, 43(3), 752-768.

punitha, v. s. s. c. (2012). Approaches to Ontology Based Algorithms for Clustering Text Documents. Int.J.Computer Technology&Applications, 3 (5), 1813-1817.

Pandit, S. (2008). On a robust document classification approach using TF-IDF scheme with learned, context-sensitive semantics.

Priss, U. (2006) "Formal concept analysis in information science." ARIST, 40.1: 521-543.

Patel, C., Hamou-Lhadj, A., & Rilling, J. (2009). Software clustering using dynamic analysis and static dependencies. Paper presented at the Software Maintenance and Reengineering, 2009. CSMR'09. 13th European Conference on IEEE.

Popat, K. (2013). Word Clustering for Data Sparsity: A Literature Survey.

Paterlini, S., & Krink, T. (2006). Differential evolution and particle swarm optimisation in partitional clustering. Computational Statistics & Data Analysis, 50(5), 1220-1247. Text Documents. Int.J.Computer Technology & Applications, 3 (5), 1813-1817.

Qu, Q., Qiu, J., Sun, C., & Wang, Y. (2008). Graph-based knowledge representation model and pattern retrieval. Paper presented at the Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on IEEE.

Qadi, A. E., Aboutajedine, D., & Ennouary, Y. (2010). Formal concept analysis for information retrieval. arXiv preprint arXiv:1003.1494.

Rajaraman, K., & Tan, A.-H. (2002). Knowledge discovery from texts: a concept frame graph approach. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management.

Rajaraman, K., & Tan, A.-H. (2003). Mining semantic networks for knowledge discovery. Paper presented at the Data Mining, 2003. ICDM 2003. Third IEEE International Conference on IEEE.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. Paper presented at the Proceedings of the First Instructional Conference on Machine Learning.

Roma, V., Bewoor, M., & Patil, S. (2013). Evaluator and Comparator: Document Summary Generation based on Quantitative and Qualitative Metricsfor International Journal of Scientific & Engineering Research.

Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook: Springer.

Siti S.K.(2011).Frame work for deviation detection in text:Thesis,Universiti Kebangsaan Malaysia,Bangi.

Shiga, M., & Mamitsuka, H. (2012). A variational bayesian framework for clustering with multiple graphs. Knowledge and Data Engineering, IEEE Transactions on, 24(4), 577-590.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information processing & management, 33(2), 193-207.

Schedl, M. (2012). # nowplaying Madonna: a large-scale evaluation on estimating similarities between music artists and between movies from microblogs. Information Retrieval, 15(3-4), 183-217.

Shaban, K., Basir, O., & Kamel, M. (2006). Document mining based on semantic understanding of text Progress in Pattern Recognition, Image Analysis and Applications (pp. 834-843): Springer.

Suchanek, F. M., Ifrim, G., & Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Saad, F. H., de la Iglesia, B., & Bell, D. G. (2006). A Comparison of Two Document Clustering Approaches for Clustering Medical Documents. Paper presented at the DMIN.

Sowa, J. F., & Way, E. C. (1986). Implementing a semantic interpreter using conceptual graphs. IBM Journal of Research and Development, 30(1), 57-69.

Stumme, G. (2002). Formal concept analysis on its way from mathematics to computer science Conceptual Structures: Integration and Interfaces (pp. 2-19): Springer.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

Sowa, J.F. ,(1976) “Conceptual Graphs for a Database Interface”, IBM J. R&D, pp 336-357 vol.20.

Shaban, K. (2006). A semantic graph model for text representation and matching in document mining. Citeseer.

Schenker, A., Last, M., Bunke, H., & Kandel, A. (2003). Classification of web documents using a graph model. Paper presented at the Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on IEEE.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation AI 2006: Advances in Artificial Intelligence (pp. 1015-1021): Springer.

Tackstrom, O.,Ryan McDonald, and Jakob Uszkoreit (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 477{487, Montr_eal, Canada.

Tasdemir, K., & Merényi, E. (2011). A validity index for prototype-based clustering of data sets with complex cluster structures. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEETransactions on, 41(4), 1039-1053.

Tar, H. H., & Nyunt, T. T. S. (2011). Ontology-Based Concept Weighting for Text Documents. Paper presented at the International Conference on Information Communication and Management IPCSIT.

Tarabalka, Y., Benediktsson, J. A., & Chanussot, J. (2009). Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. Geoscience and Remote Sensing, IEEE Transactions on, 47(8), 2973-2987.

Theodoridis, S., & Koutroubas, K. (1999). Feature generation II. Pattern Recognition, 233-270.

Theodoridis, S., Pikrakis, A., Koutroumbas, K., & Cavouras, D. (2010). Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach: Academic Press.

Tong, T. (2011). Semantic frameworks for document and ontology clustering. University of Missouri--Kansas City.

Uszkoreit, J., & Thorsten Brants. Distributed word clustering for large scale class-based language modeling in machine translation. In Proceedings of ACL-08: HLT, Columbus, Ohio, 2008.

Velmurugan, T., & Santhanam, T. (2010). Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. Journal of computer science, 6(3), 363.

Wang, Y., Ni, X., Sun, J.-T., Tong, Y., & Chen, Z. (2011). Representing document as dependency graph for document clustering. Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management.

Wang, L., & Liu, X. (2008). A new model of evaluating concept similarity. Knowledge-Based Systems, 21(8), 842-846.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16(3), 645-678.

Yun-tao, Z., Ling, G., & Yong-cheng, W. (2005). An improved TF-IDF approach for text classification. Journal of Zhejiang University SCIENCE A, 6(1), 49-55.

Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management.

Zimmermann, T., & Nagappan, N. (2008). Predicting defects using network analysis on dependency graphs. Paper presented at the Proceedings of the 30th international conference on Software engineering.