TEXT REPRESENTATION USING CANONICAL DATA MODEL

HIBA JASIM HADI

MASTER OF INFORMATION TECHNOLOGY

COLLEGE OF ARTS AND SCIENCES

UNIVERSITI UTARA MALAYSIA

2016

Permission to Use

In presenting this thesis in fulfilment 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 thesis 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 thesis 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 thesis.

Requests for permission to copy or to make other use of materials in this thesis, 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

ABSTRAK

Pembangunan teknologi digital dan World Wide Web telah membawa kepada peningkatan dokumentasi-dokumentasi digital yang digunapakai untuk pelbagai keperluan contohnya dalam bidang penerbitan yang telah menunjukkan perkaitan dalam meningkatkan kesedaran tentang keperluan teknik yang berkesan yang membantu dalam pencarian dan mendapatkan teks.Persembahan teks memainkan peranan yang amat penting dalam menyampaikan maksud teks dengan lebih bermakna atau tepat. Ketepatan penyampaian sesuatu teks amat bergantung kepada pemilihan kaedah teks itu dipersembahkan. Kaedah tradisional di dalam persembahan teks berdasarkan model dokumen seperti term-frequency invers document frequency (TF-IDF) tidak menitikberatkan hubungan dan makna perkataan di dalam sesuatu dokumen.Oleh itu, masalah sparsiti dan semantik yang merupakan masalah yang dominan di dalam dokumen teks masih belum menemui penyelesaian.Kajian ini mencadangkan bagaimana masalah sparsiti dan semantic dikurangkan dengan penggunaan Canonical Data Model (CDM) untuk meyampaikan teks.CDM distruktur melalui pengumpulan analisis semantik dan sintaksis.20 kumpulan dataset berita telah digunakan untuk menguji kesahihan CDM dalam penyampaian teks dalam kajian ini.Dokumen-dokumen teks akan melalui beberapa proses pra-pemprosesan dan menghuraikan sintaksis untuk mengenal pasti struktur ayat.Dokumen teks akan melalui beberapa langkah pra-pemprosesan dan menghuraikan sintaksis untuk mengenal pasti struktur ayat dan maka kaedah TF-IDF digunakan untuk mewakili teks yang melalui CDM. Ini membuktikan bahawa CDM tepat untuk mewakili teks, berdasarkan pengesahan model melalui kajian bahasa pakar-pakar berdasarkan peratusan kaedah pengukuran persamaan.

KATA KUNCI: Perwakilan Teks, TF-IDF, CDM

ABSTRACT

Developing digital technology and the World Wide Web has led to the increase of digital documents that are used for various purposes such as publishing, in turn, appears to be connected to raise the awareness for the requirement of effective techniques that can help during the search and retrieval of text. Text representation plays a crucial role in representing text in a meaningful way. The clarity of representation depends tightly on the selection of the text representation methods. Traditional methods of text representation model documents such as term-frequency invers document frequency (TF-IDF) ignores the relationship and meanings of words in documents. As a result the sparsity and semantic problem that is predominant in textual document are not resolved. In this research, the problem of sparsity and semantic is reduced by proposing Canonical Data Model (CDM) for text representation. CDM is constructed through an accumulation of syntactic and semantic analysis. A number of 20 news group dataset were used in this research to test CDM validity for text representation. The text documents goes through a number of pre-processing process and syntactic parsing in order to identify the sentence structure. Text documents goes through a number of pre-processing steps and syntactic parsing in order to identify the sentence structure and then TF-IDF method is used to represent the text through CDM. The findings proved that CDM was efficient to represent text, based on the model validation through language experts' review and the percentage of the similarity measurement methods.

Keywords: Text Representation, TF-IDF, CDM

Acknowledgments

"Alhamdulillah', praise be to Allah, The Most Beneficent, The Most Merciful"

It gives me great pleasure to thank and acknowledge all those who have contributed to my education journey whose results have flourished into this thesis.

First and foremost: I would like to extend my gratitude to my supervisors Dr. Azizah Bt Haji Ahmad and DR. Siti Sakira Kamaruddin for their unreserved guidance and counsel rendered from the very beginning to the completion of the research. I appreciate their kindness and support which have manifested in various ways. Without their support, guidance, and help this research would not have been successfully materialized. I cannot fully express my gratitude to the exceptional advice every time I seek enlightenment, the sharing of their knowledge from both theoretical and practical aspects, and their logical way of thinking have provided a good basis for this work. They guided me and polish my ideas, and translate them into workable solutions.

I owe my most heartiest gratitude to my wonderful parents for their patience, understanding, encourage, unlimited support, and prayers for smoothness and blessed journey of my Master, without them, I could not pursue and accomplish this task so, thank you very much.

My sincere appreciation and gratitude to the most important person who have strongly contributed, in various ways, by providing the help and support needed my dearest and nearest friend Atheer Flayh Hassan.

Hiba Jasim Hadi

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

1.1 Background

In the last decade, text has become the most popular tool for communication due to the rapid technological increase. Realizing that extracting useful information from text is not an easy task, there is a need to have an intelligent tool which is able to extract useful information as quick as possible and at a low cost (Jusoh & Alfawareh, 2012) and the most prominent method to handle the task is text mining (Gharehchopogh & Khalifelu, 2011). According to Fleuren and Alkema (2015), text mining is the process of extracting new knowledge from a predefined information by regulating the distance between piece of information into certain meanings.

Text mining is considered a vivid domain for research that changes the stress in text-based information to the level of exploration and analysis from the level of retrieval. It is also one of the famous way to organizing unstructured information (Patil & Saraf, 2013). Summaries of the words are derived from information to make it easy to investigate words used in the documents (Suguna & Gomathi, 2014). Organizations can explore interesting rules, models and patterns from the text in the same manner as data mining searches data in the tables (Jhanji & Garg, 2014).

The most difficult part in text mining is the complication involved in a natural language, i.e., every natural language faces some ambiguous issues in its structure of

REFERENCES

Abdullah, A., & Fadhil, O. (2014). Implementation of Preprocessing Techniques in Datamining.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining Text Data* (pp. 77–128). Springer.

Agnihotri, D., Verma, K., & Tripathi, P. (2014). Pattern and Cluster Mining on Text Data. In *Communication Systems and Network Technologies (CSNT), forth International Conference on communication and network technologies 2014* (pp. 428–432).

Alfawareh, H. M., & Jusoh, S. (2013). Resolving ambiguous preposition phrase for text mining applications. In *Computer Applications Technology (ICCAT), 2013 International Conference on* (pp. 1–5).

Andrews, S., & Polovina, S. (2011). A mapping from conceptual graphs to formal concept analysis. In *Conceptual Structures for Discovering Knowledge* (pp. 63–76). Springer.

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., & Statnikov, A. (2014). A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for Information Science and Technology*, *65*(10), 1964–1987.

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

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, *12*(3), 183–210.

Basu, T., & Murthy, C. A. (2012). Effective text classification by a supervised feature selection approach. In *Data Mining Workshops (ICDMW), 12th International Conference onData Mining Workshops 2012 . IEEE* (pp. 918–925).

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

Bhaisare, R., & Nayyar, V. (2014). Analysis of Effective Pattern Discovery with Text Mining in Business Based Application. *International Journal of Research*, *1*(11), 288–292.

Bloechle, J.-L., Rigamonti, M., Hadjar, K., Lalanne, D., & Ingold, R. (2006). XCDF: a canonical and structured document format. In *Document Analysis Systems VII* (pp. 141–152). Springer.

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S. (2005). An Ontology-based Framework for Text Mining. In *LDV Forum* (Vol. 20, pp. 87–112).

Boroch, N., & Heger, T. (2010). Vehicle environment description and interpretation using conceptual graphs. In *Intelligent Transportation Systems (ITSC), 13th International Conference on Intelligent Transportation Systems 2010 .IEEE* (pp. 1233–1236).

Boyer, K. L., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2014). Editorial: Introduction to the special issue on supervised and unsupervised classification techniques and their applications. *Pattern Recognition Letters*, *41*, 1–2.

Capasso, P. (2008). *BioInView: a system for integrating heterogeneous and distributed biological data sources*. University degli Studi di Napoli Federico II.

Chekima, K., On, C. K., Alfred, R., Soon, G. K., & Anthony, P. (2012). Document categorizer agent based on ACM hierarchy. In *Control System, Computing and Engineering (ICCSCE), International Conference on Control System, Computing and Engineering 2012. IEEE* (pp. 386–391).

Chen, M., Weinberger, K. Q., Sha, F., & Others. (2013). An alternative text representation to TF-IDF and Bag-of-Words. *arXiv Preprint arXiv:1301.6770*.

Chen, X., & Wu, C. (2012). A Text Representation Method Based on Harmonic Series. In *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012* (pp. 1830–1834).

Chen, Y., Sun, Y., & Han, B.-Q. (2015). Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection. *BioMed Research International*.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). Introduction to information retrieval. *Christopher. Cambridge Univ. Press. 2008*, p(544).

Ciesielski, M., Jabir, A. M., & Pradhan, D. K. (n.d.). Canonical Graph-based Representations for Verification of Logic and Arithmetic Designs. *University of Massachusetts, Amherst, MA(01003)*.

Cohen, W. W. (1998). WHIRL: A Word-based Information Representation.

Coldicott, P. A., & Lane, E. (2009). Standard Based Mapping of Industry Vertical Model to Legacy Environments. Google Patents.

Comer, E. (2010). *Canonical Data Model Design Guidelines*.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Darwiche, A. (2011). SDD: A new canonical representation of propositional knowledge bases. In *IJCAI Proceedings International Joint Conference on Artificial Intelligence* (Vol. 22, p. (819)).

Dayananda, P., & Shettar, R. (2011). Survey on Information Retrieval in Semi Structured Data. *International Journal of Computer Applications*, *32*(8), 1–5.

Delugach, H., & Lampkin, B. (2000). Troika: using grids, lattices and graphs in knowledge acquisition. *Working with Conceptual Structures: Contributions to ICCS*, 201–214.

Deshmukh, S. N., Deshmukh, R. R., & Deshmukh, S. N. (2014). Performance Analysis of Different Sentence Oddity Measures Applied on Google and Google News Repository for Detection of Substitution. *International Refereed Journal of Engineering and Science (IRJES)*, *3*(3), 20–25.

Dey, A., & Prukayastha, B. S. (2013). Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach. *International Journal of Computer Applications*, *84*(9), 31–35.

Dietrich, J., Yakovlev, V., McCartin, C., Jenson, G., & Duchrow, M. (2008). Cluster analysis of Java dependency graphs. In *Proceedings of the 4th ACM symposium on Software visualization* (pp. 91–94).

Dietrich, M., & Lemcke, J. (2011). A Refined Canonical Data Model for Multi-schema Integration and Mapping. In *IEEE 8th International Conference on e-Business Engineering (ICEBE), 2011* (pp. 105–110).

Dietrich, M., Weissmann, D., Rech, J., & Stuhec, G. (2010). Multilingual extraction and mapping of dictionary entry names in business schema integration. In *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services* (pp. 863–866).

Dolamic, L., & Savoy, J. (2009). Indexing and stemming approaches for the Czech language. *Information Processing & Management*, *45*(6), 714–720.

Durugkar, S. (2013). Various Issues in Implementing Cross Language Information Retrieval and Enhancing the Efficiency of Meta Search Tool. *International Journal of Emerging Technology and Advanced Engineering*, *3*(2), 605–609.

El-Shishiny, H., & Volkov, P. (2011). Systems and methods for building an electronic dictionary of multi-word names and for performing fuzzy searches in the dictionary. Google Patents.

Fares, M., Oepen, S., & Zhang, Y. (2013). Machine learning for high-quality tokenization replicating variable tokenization schemes. In *Computational Linguistics and Intelligent Text Processing* (pp. 231–244). Springer.

Fleuren, W. W. M., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, *74*, 97–106.

Francis, M., & KN, R. N. (2014). An Algorithm for Plagiarism Detection in

Malayalam Language Documents Using Modified n-gram Model. *National Conference on Indian Language Computing (NCILC 2014)*, (13), 10–15.

Galescu, L., & Allen, J. F. (2001). Bi-directional conversion between graphemes and phonemes using a joint n-gram model. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

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

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *Application of Information and Communication Technologies (AICT), 5th International Conference on Communication Technologies 2011.* (pp. 1–4).

Gonzalez, M. A., Martí, J. R., & Kruchten, P. (2011). A canonical data model for simulator interoperation in a collaborative system for disaster response simulation. In *Electrical and Computer Engineering (CCECE), 24th Canadian Conference on Electrical and Computer Engineering 2011.* (pp. 1519–1522).

Guo, X., Xiang, Y., & Chen, Q. (2011). A vector space model approach to social relation extraction from text corpus. In *Fuzzy Systems and Knowledge Discovery (FSKD), Eighth International Conference on Fuzzy Systems and Knowledge Discovery 2011* (Vol. 3, pp. 1756–1759).

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In *Information Technology and Electrical Engineering (ICITEE), 6th International Conference on Information Technology and Electrical Engineering 2014* (pp. 1–4).

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. *IJCA*, 110–119.

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

Hsiao, D. K., Neuhold, E. J., & Sacks-Davis, R. (2014). *Interoperable Database Systems (DS-5): Proceedings of the IFIP WG2. 6 Database Semantics Conference on Interoperable Database Systems (DS-5) Lorne, Victoria, Australia, 16-20 November, 1992*. Elsevier.

Hsiao, D., Neuhold, E. J., & Sacks-Davis, R. (2014). Interoperability between database models. In *Interoperable Database Systems (DS-5): Proceedings of the IFIP WG2. 6 Database Semantics Conference on Interoperable Database Systems (DS-5) Lorne, Victoria, Australia, 16-20 November, 1992* (p. 101).

Hu, X., & Liu, H. (2012). Text analytics in social media. In *Mining text data* (pp. 385–414). Springer.

Huang, C.-R., Šimon, P., Hsieh, S.-K., & Prévot, L. (2007). Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th Annual Meeting of the Acl on Interactive Poster and Demonstration Sessions* (pp. 69–72).

Huang, X., & Wu, Q. (2013). Micro-blog commercial word extraction based on improved tf-idf algorithm. In *TENCON 2013-2013 IEEE Conference (31194)* (pp. 1–5).

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

Jhanji, D., & Garg, P. (2014). Text Mining. *International Journal Of Scientific Research And Education*, *2*(08).

Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems*, *23*(4), 302–308.

Jusoh, S., & Alfawareh, H. M. (2012). Techniques, Applications and Challenging Issue in Text Mining ues, Applications and Challenging Issue in Text Mining ues, Applications and Challenging Issue in Text Mining. *IJCSI International Journal of Computer Science Issues*, *9*(6), 431–436.

Kamaruddin, S. S., Bakar, A. A., Hamdan, A. R., & Nor, F. M. (2008). Conceptual graph formalism for financial text representation. In *Information Technology, 2008. ITSim 2008. International Symposium on* (Vol. 3, pp. 1–6).

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

Kaur, M., & Kaur, N. (2013). Web Document Clustering Approaches Using K-Means Algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(5), 861–864.

Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications (IJAIA)*, *3*(2), 85–99.

Kumar, P., Khulbe, S., & Dhami, H. S. (2012). Computer Program for Counting the Part of Speeches, Text Narrations by using Secondary Data Algorithm Techniques. *International Journal of Computer Applications*, *51*(11).

Lan, M., Tan, C.-L., Low, H.-B., & Sung, S.-Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Special interest tracks and posters of the 14th International Conference on World Wide Web* (pp. 1032–1033).

Lazim, R. Y., & Kamaludin, A. (2013). Going Native: Indexing architecture of eXist-db_ An Open Source native XML database system. In *TENCON 2013-2013 IEEE Conference (31194)* (pp. 1–4).

Liu, M., & Yang, J. (2012). An improvement of TFIDF weighting in text categorization. *International Proceedings of Computer Science and Information Technology*, 44–47.

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

Malarvizhi, M. P., & Mohana, M. S. (2014). A Survey on Various Candidate Generator Methods for Efficient String Transformation. *COMPUSOFT, An International Journal of Advanced Computer Technology*, *3*(2), 524–528.

Melnikov, M. P., & Vorobkalov, P. N. (2014). Retrieval of Drug-Drug Interactions Information from Biomedical Texts: Use of TF-IDF for Classification. In *Knowledge-Based Software Engineering* (pp. 593–602). Springer.

Miguelanez, E., Brown, K. E., Lewis, R., Roberts, C., & Lane, D. M. (2008). Fault diagnosis of a train door system based on semantic knowledge representation. In *Railway Condition Monitoring, 4th IET International Conference on Railway Condition Monitoring 2008.* (pp. 1–6).

Mitchell, B. S., & Mancoridis, S. (2010). Clustering module dependency graphs of software systems using the bunch tool.

Nawab, R. M. A., Stevenson, M., & Clough, P. (2013). Comparing Medline citations using modified N-grams. *Journal of the American Medical Informatics Association*, amiajnl–2012.

Novak, J. R., Minematsu, N., & Hirose, K. (2013). Failure transitions for joint n-gram models and G2p conversion. In *INTERSPEECH* (pp. 1821–1825).

Palmer, D. D. (2010). Text preprocessing. *HANDBOOK OF*, 9.

Pan, S., Wang, L., & Xia, G. (2012). Mining association rules from consumer product safety cases based on text classification. *JCIT: Journal of Convergence Information Technology*, *7*(9), 422–430.

Panda, G., Panda, B. S., & Giridhar, B. (2015). Text Data Mining with Different Comparisons. *International Journal of Computer Science and Mobile*

Computing, *4*(2), 7–13.

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

Patel, C., Hamou-Lhadj, A., & Rilling, J. (2009). Software clustering using dynamic analysis and static dependencies. In *13th European Conference on Software Maintenance and Reengineering, 2009. CSMR'09.* (pp. 27–36).

Patil, S. P., & Saraf, R. N. (2013). Two-Step Approach for Acquiring Semantic Relations from Textual Web Content. *International Journal of Innovative in Engineering and Technology (IJIET)*, *2*(4), 145–150.

Pedrosa, G. V, & Traina, A. J. M. (2013). From Bag-of-Visual-Words to Bag-of-Visual-Phrases using n-Grams. In *Graphics, Patterns and Images (SIBGRAPI), 26th SIBGRAPI-Conference on Graphics, Patterns and Images 2013* (pp. 304–311).

Qu, Q., Qiu, J., Sun, C., & Wang, Y. (2008). Graph-based knowledge representation model and pattern retrieval. In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 5, pp. 541–545).

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Reshamwala, A., & Mahajan, S. (2015). Automating sequence dataset generating by using SeqGen. In *Communication, Information & Computing Technology (ICCICT), International Conference 2015* (pp. 1–5).

Roebuck, K. (2012). *Enterprise Information Management (EIM): High-impact Strategies-What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Emereo Publishing.

Sangwan, U. (2014). A Search Technique On Databases Based On User Expertise. *International Journal Publication In Academic Education And Research – Computer Science and Engineering*, *1*(1), 1–4.

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation* (pp. 130–143).

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.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47.

Segond, F., Schiller, A., Grefenstette, G., & Chanod, J.-P. (1997). An experiment in semantic tagging using hidden markov model tagging. In *ACL/EACL workshop on automatic information extraction and building of lexical semantic resources for NLP applications*.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50–64.

Shneiderman, B. (1992). *Designing the user interface: strategies for effective human-computer interaction* (Vol. 3). Addison-Wesley Reading, MA.

Siti Mahfuzah, S. (2011). *Conceptual Design Model of Computerized Personal-Decision AID (ComPDA)*. Universiti Utara Malaysia.

Sonawane, S. S., & Kulkarni, P. A. (2014). Graph based Representation and Analysis of Text Document: A Survey of Techniques. *International Journal of Computer Applications*, *96*(19), 1–8.

Specia, L., & Motta, E. (2006). A hybrid approach for extracting semantic relations from texts. In *In. Proceedings of the 2 nd Workshop on Ontology Learning and Population* (pp. 57–64).

Stevens, R., Goble, C. A., & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, *1*(4), 398–414.

Suchanek, F. M., Ifrim, G., & Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 712–717).

Suguna, S., & Gomathi, B. (2014). Comparison between Clustering Algorithms Based On Ontology Based Text Mining Techniques. *International Journal of Advanced Research in Computer Science*, *5*(7).

Sureka, V., & Punitha, S. C. (2012). Approaches to Ontology Based Algorithms for Clustering Text Documents. *International Journal of Circuit Theory and ApplicationsSept-Oct*.

Szyma'nski, J., & Duch, W. (2010). Representation of hypertext documents based on terms, links and text compressibility. In *Neural Information Processing. Theory and Algorithms* (pp. 282–289). Springer.

Truong, M. T., Amblard, F., Gaudou, B., & Sibertin-Blanc, C. (2014). To Calibrate & Validate an Agent-Based Simulation Model-An Application of the Combination Framework of BI solution & Multi-agent platform.

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

Wang, Y., Ni, X., Sun, J.-T., Tong, Y., & Chen, Z. (2011). Representing document as dependency graph for document clustering. In *Proceedings of the 20th ACM International Conference on Information and knowledge Management* (pp. 2177–2180).

Wei, D., Zhang, C. H., & Zhu, X. N. (2014). Service Selection Based on Rule and Statistics Model in WoT Smart Home. In *Applied Mechanics and Materials* (Vol. 681, pp. 244–248).

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, *42*(4), 2264–2275.

Wei, X., Xiaofei, X., Lei, S., Quanlong, L., & Hao, L. (2001). Business intelligence based group decision support system. In *International Conferences on Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 2001* (Vol. 5, pp. 295–300).

Yuan, B., Chen, Q., Xiang, Y., & Wang, X. (2013). Event Detection and Recommendation Based on Heterogeneous Information. In *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012* (pp. 407–416).

Yuanyuan, L., & Jianhu, W. (2011). RESEARCH ON TEXT MINING. *American Journal of Engineering and Technology Research Vol*, *11*(9).

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.

Zhang, G., & Odbal. (2012). Sentence alignment for web page text based on vector space model. In *International Conference on Computer Science and Information Processing (CSIP), 2012* (pp. 167–170).

Zhang, J., & Rasmussen, E. M. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management*, *37*(2), 279–294.

Zhou, L. (2010). An approach of semantic web service discovery. In *Communications and Mobile Computing (CMC), International Conference on ommunications and Mobile Computing 2010* (Vol. 1, pp. 537–540).

Zimmermann, T., & Nagappan, N. (2008). Predicting defects using network analysis on dependency graphs. In *Proceedings of the 30th International Conference on Software Engineering* (pp. 531–540).

Zwietasch, T. (2014). *Detecting anomalies in system log files using machine learning techniques*. Institut für Softwaretechnologie.