ENHANCED ONTOLOGY-BASED TEXT CLASSIFICATION ALGORITHM FOR STRUCTURALLY ORGANIZED DOCUMENTS

SUHA SAHIB OLEIWI

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2015

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
UUMCollege of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Pengelasan Teks (TC) merupakan asas yang penting dalam dapatan semula maklumat dan perlombongan teks. Fungsi utama TC adalah untuk menentukan kelas teks mengikut kepada jenis label yang diberi lebih awal. Kebanyakan algoritma TC menggunakan istilah dalam mewakili dokumen yang tidak mengambil kira hubungan di antara istilah tersebut. Algoritma ini mewakili dokumen dalam satu ruangan di mana setiap perkataan diandaikan menjadi satu dimensi. Hal ini menyebabkan terjadinya kedimensian tinggi yang akan memberi kesan negatif terhadap prestasi pengelasan. Objektif kajian ini adalah untuk merangka algoritma pengelasan teks dengan mewujudkan ciri vektor yang sesuai dan mengurangkan dimensi data yang akan meningkatkan ketepatan pengelasan. Kajian ini menggabungkan ontologi dan perwakilan teks untuk pengelasan dengan membangunkan lima algoritma. Algoritma pertama dan kedua iaitu Vektor Bercirikan Konsep (CFV) dan Vektor Bercirikan Struktur (SFV), akan mewujudkan ciri vektor untuk menggambarkan dokumen tersebut. Algoritma ketiga iaitu Pengelasan Teks Berasaskan Ontologi (OBTC) dibangunkan untuk mengurangkan kedimensian kumpulan–kumpulan latihan. Algoritma keempat dan kelima iaitu Pengelasan Teks_Vektor Bercirikan Konsep (CFV_TC) dan Pengelasan Teks_Vektor Bercirikan Struktur (SFV_TC) akan mengelaskan dokumen tersebut kepada kumpulan–kumpulan pengelasan yang berkaitan. Algoritma yang dicadangkan ini telah diuji menggunakan data set dari lima dokumen saintifik yang berbeza yang dimuat turun dari pelbagai perpustakaan digital dan repository. Hasil pengujian pengelasan teks daripada algoritma CFV_TC dan SFV_TC menunjukkan nilai purata kepersisan, dapatan semula, ukuran-f dan ketepatan adalah lebih baik berbanding dengan pendekatan SVM dan RSS. Kajian ini menyumbang kepada bidang penyelidikan dalam dapatan maklumat dan perlombongan teks untuk mendapatkan dokumen yang lebih relevan melalui penggunaan ontologi dalam pengelasan teks.

Kata kunci: Klasifikasi teks, Ontologi, Struktur, Dokumen berstruktur.

Abstract

Text classification (TC) is an important foundation of information retrieval and text mining. The main task of a TC is to predict the text's class according to the type of tag given in advance. Most TC algorithms used terms in representing the document which does not consider the relations among the terms. These algorithms represent documents in a space where every word is assumed to be a dimension. As a result such representations generate high dimensionality which gives a negative effect on the classification performance. The objectives of this thesis are to formulate algorithms for classifying text by creating suitable feature vector and reducing the dimension of data which will enhance the classification accuracy. This research combines the ontology and text representation for classification by developing five algorithms. The first and second algorithms namely Concept Feature Vector (CFV) and Structure Feature Vector (SFV), create feature vector to represent the document. The third algorithm is the Ontology Based Text Classification (OBTC) and is designed to reduce the dimensionality of training sets. The fourth and fifth algorithms, Concept Feature Vector_Text Classification (CFV_TC) and Structure Feature Vector_Text Classification (SFV_TC) classify the document to its related set of classes. These proposed algorithms were tested on five different scientific paper datasets downloaded from different digital libraries and repositories. Experimental obtained from the proposed algorithm, CFV_TC and SFV_TC shown better average results in terms of precision, recall, f-measure and accuracy compared against SVM and RSS approaches. The work in this study contributes to exploring the related document in information retrieval and text mining research by using ontology in TC.

Keywords: Text classification, ontology, structural, structured documents.

Acknowledgement

It gives me great pleasure to express my gratefulness to everyone who contributed in completing this thesis. It was my pleasure to study under Associate Professor Dr. Azman Yasin's supervision. I'm so grateful for his support during the last five years. I am so grateful for his all assistants that he gave me through these years. There are no words to express my gratitude for his guidance in helping me to achieve my goal. Without his valuable support, my thesis would not have been possible. I would like to tell him that thank you so much for everything you have been done for me to reach my goal. I would like to thank my co-supervisor Dr. Nor Idayu Mahat for her progressive thinking and her open mind. Her continuous advice and significant comments helped develop my work successfully.

To my father, whose surname I proudly carry – I am forever appreciative. I want to tell him thanks for all things you supported me and make me strong to across this stage of my life. To my mother, who gave me life and prayed for me all the time, may Allah continuously bless her with good health. To my sisters Sahar and Rafah, I would like to tell them thanks for your feelings and supporting. To my dear brothers Ali and Hassanin, thanks for their love and support. To my Husband Ghassan, who gave me power and patience during the last five years of study, I thank his from the bottom of my heart. I would also like to thank my two young babies Mohammed and Zainab, without whom my goal would not have been achieved. I dedicate this work to my family. I'm so glad to study at Universiti Utara Malaysia (UUM). During my time in UUM, I have gained a lot of friends, and studying there was like being in my hometown. My sincere gratitude to all of them for all the encouragement during my study. I want to tell all of them thank you so much for everything you help me.

Table of Contents

List of Tables

List of Figures

List of Appendices

CHAPTER ONE
INTRODUCTION

1.1 Background

Text categorization is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world (Kaur & Jyoti, 2013). In the recent years, TC has gained tremendous attention and rapidly developed. Today, TC is widely used in applications such as “automatic indexing” for "Boolean information retrieval" systems, "document organization", "text filtering", and "word-sense disambiguation" (Rafi, et al, 2012; Shimodaira, 2014).

According to (Calvo, Lee, & Li, 2006), TC reduces the time required to classify vast amounts of documents without the need for experts. While TC methods may vary in terms of accuracy and computation efficiency, TC methods generally save time and expense required to perform TC. Classification algorithms can be used to extract models describing important data classes.

There are several algorithms used to classify text such as "k-nearest neighbors" (KNN), "naïve Bayes" (NB), and "Support Vector Machines" (SVM) (Patra & Singh, 2013). To build a classifier in text classification there is need to define set of example as training set. These sets are labelled with pre-defined classes (Li & Liu, 2003). Often, a data set sample contains both positive and negative examples of a concept to induce a classification rule use machine learning algorithm (Aytug, Boylu, & Koehler, 2006).

The contents of the thesis is for internal user only

References

Achananuparp, P., Zhou, X., Hu, X., & Zhang, X. (2008). Semantic representation in text classification using topic signature mapping. *Paper presented in IEEE International Joint Conference on Neural Networks IJCNN*, 1034 – 1040.

Agarwal, S., Singhal, A., & Bedi, P. (2012). Classification of RSS feed news items using ontology. *Paper presented in IEEE 14th Intelligent Systems Design and Applications (ISDA)*, 491 – 496.

Aggarwal, C., C., Zhai, & ChengXiang. (2012). Mining Text Data. *Chapter six of the book XII, 524 p*, 182.

Ahmed, N., Khan, S., Latif, K., Masood, A., & Elberrichi, Z. (2008). Extracting semantic annotations and their correlation with document components. *Paper presented in IEEE 4th International Conference on Merging Technologies, ICET* 32 – 37.

Ajgalik, M., Barla, M., & Bielikova, M. (2013). From Ambiguous Words to Key-Concept Extraction. *Paper presented in IEEE 24th International Workshop on Database and Expert Systems Applications (DEXA)*, 63 - 67.

Almeida, T. A., Yamakami, A., & Almeida, J. (2009). Evaluation of Approaches for Dimensionality Reduction Applied with Naive Bayes Anti-Spam Filters. *Paper presented in IEEE International Conference on Machine Learning and Applications, ICMLA*, 517-522.

Aytug, H., Boylu, F., & Koehler, G. J. (2006). Learning in the Presence of Self-Interested Agents. *Paper presented in IEEE International Conference of the 39th Annual Hawaii, Vol. 7*, 1-7.

Basu, T., & Murthy, C. A. (2012). Effective Text Classification by a Supervised Feature Selection Approach. *Paper presented in IEEE 12th International Conference on Data Mining Workshops (ICDMW), 918* – 925.

Bhatia, N., & Vandana, S. (2010). Survey of Nearest Neighbor Techniques. *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 2,* 302-305.

Bin, L., Jun, L., Min, Y., J., & Ming, Z. Q. (2008). Automated Essay Scoring Using the KNN Algorithm. *Paper presented in IEEE International Conference on Computer Science and Software Engineering, Vol. 1*, 735 - 738.

Bleik, S., Mishra, M., Huan, J., & Song, M. (2013). Text Categorization of Biomedical Data Sets Using Graph Kernels and a Controlled Vocabulary. *Paper presented in IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 10 (Issue 5),* 1211 - 1217.

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models. *In Workshop on Text Learning held at ICML*.

Brank, J., Mladenić, D., & Grobelnik, M. (2010). Large-scale Hierarchical Text Classification Using SVM and Coding Matrices. *In: Large-Scale Hierarchical Classification Workshop of ECIR, 28 – 31 March, Milton Keynes, UK*.

Calvier, F. c.-E., Planti´e, M., Dray, G. e., & Ranwez, S. (2013). Ontology Based Machine Learning for Semantic Multiclass Classification. *Author manuscript,*

published in "TOTH: Terminologie & Ontologie: Théories et Applications 2013, Chambéry: France.

Calvo, R. A., Lee, J.-M., & Li, X. (2006). Managing content with automatic document classification. *Journal of Digital Information*, 1-15.

Celik, K., & Gungor, T. (2013). A comprehensive analysis of using semantic information in text categorization. *Paper presented in IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 1 – 5.

Chang, Y.-H. (2007). Automatically Constructing a Domain Ontology for Document Classification. *Paper presented in IEEE International Conference on Machine Learning and Cybernetics, Vol. 4*, 1942 - 1947.

Chang, Y.-H., & Huang, H.-Y. (2008). An Automatic Document Classifier System based on Naíve Bayes Classifier and Ontology. *paper presented in IEEE International Conference on Machine Learning and Cybernetics, Vol. 6*, 3144 - 3149.

Che, C., & Teng, H. (2009). Document representation combining concepts and words in Chinese text categorization. *Paper presented in IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 1-5.

Chirawichitchai, N., Sanguansat, P., & Meesad, P. (2009). A Comparative Study on Feature Weight in Thai Document Categorization Framework, 257-266.

Cunhua, L., Yun, H., & Zhaoman, Z. (2010). An event ontology construction approach to web crime mining. *Paper presented in IEEE Seventh*

International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 5, 2441 - 2445.

Cunningham, P. & Delany, S. J. (2007). K-Nearest Neighbour Classifiers. *Technical Report UCD-CSI-2007-4.*

Debole, F., & Sebastiani, F. (2003). Supervised Term Weighting for Automated Text Categorization. *Proceedings of the 2003 ACM symposium on Applied computing*, 784-788.

Deisy, C., Gowr, M., Baskar, S., Kalaiarasi, S. M. A., & Ramraj, N. (2010). A Novel Term Weighting Scheme Midf for Text Categorization. *Journal of Engineering Science and Technology, Vol. 5,* 94 - 107.

Devare, M., Rikert, J., C., Caruso, B., Lowe, B., Chiang, K., & McCue, J. (2007). Connecting People, Creating a Virtual Life Sciences Community. *D-Lib Magazine, ISSN 1082- 9873, Vol. 13, No. 7/8.*

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification. Journal of Machine Learning 1265-1287.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *In First International Workshop on Multiple Classifier Systems 2000, Cagliari, Italy, Vol. 1857 of Lecture Notes in Computer Science, Springer*,1–15.

Dollah, R. B., & Aono, M. (2011). Ontology based Approach for Classifying Biomedical Text Abstracts. *International Journal of Data Engineering (IJDE), Vol. 2 (Issue 1),* 1-15.

Elberrichi, Z., Amel, B., & Malika, T. (2012). Medical Documents Classification Based on the Domain Ontology MeSH. *The International Arab Journal of e-Technology, Vol. 2, No. 4*, 210-215.

Erenel, Z., Altincay, H. & Varoglu, E. (2011). Explicit Use of Term Occurrence Probabilities for Term Weighting in Text Categorization. *Journal of information science and engineering 27*, 819-834.

Fang, J., Guo, L., Wang, X., & Yang, N. (2007). Ontology-Based Automatic Classification and Ranking for Web Documents. *Paper presented in IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, Vol. 3*, 627 - 631.

Fang, J., Guo, L., & Niu, Y. (2010). Documents Classification by Using Ontology Reasoning and Similarity Measure. *Paper presented in IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 4*, 1535 - 1539.

FAO Ontology Portal Prototype Fishery (2004). Development of Multilingual Domain Ontologies Fishery Ontology.

Fu, Z., Chen, C., Gong, Y., & Bie, R. (2008). A Comparison Study: Web Pages Categorization with Bayesian Classifiers. *Paper presented in 10th IEEE International Conference on High Performance Computing and Communications, HPCC '08*, 789 - 794.

Gang, X., & Jiancang, X. (2009). Performance Analysis of Chinese Webpage Categorizing Algorithm Based on Support Vector Machines (SVM). *Paper*

presented in IEEE Fifth International Conference on Information Assurance and Security, IAS '09, Vol. 1, 231 - 235.

Gardner, D., Akil, H., Ascoli, G., A., Bowden, D., A., Bug, W., Duncan, E., Donohue, David, H., Goldberg, Grafstein, B., Grethe, J., S., Gupta, A., Halavi, M., Kennedy, D., N., Marenco, L., Martone, M., E., Miller, P., L., Müller, H., L., Robert, A., Shepherd, G., M., Sternberg, P., W., Essen, D., C., V., & Williams, R., W. (2008). The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience. *Published in final edited form as: Neuroinformatics, Vol. 6 (3)*, 149–160.

Garrido, A. L., Gomez, O., Ilarri, S., & Mena, E. (2011). NASS: News Annotation Semantic System. *Paper presented in 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 904 - 905.

Genkin, A., Madigan, D., & Lewis, D. D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *American Statistical Association and the American Society for Quality Vol. 49, No. 3*, 291-304.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *ACM Digital Library Knowledge Acquisition - Special issue: Current issues in Knowledge Modeling Vol. 5*, 199 – 220.

Guhan, T., & Selvarajan, S. (2014). A Survey on the Suitability of ANN Based Classification Algorithms for Multidimensional Data Classification. *International Journal of Computer Science information and Engineering Technologies*, 1-5.

Ha-Thuc, V., & Renders, J.-M. (2011). Large-Scale Hierarchical Text Classification without Labelled Data. Proceedings *of the fourth ACM international conference on Web search and data mining, WSDM '11,* 685-694.

Haifeng, L., Shousheng, L., & Zhan, S. (2010). An improved KNN text categorization on skew sort condition. *Paper presented in IEEE International Conference on Computer Application and System Modeling (ICCASM) Vol. 7,* 182-186.

Halloran, J. (2009). Classification: Naive Bayes vs Logistic Regression. *University of Hawaii at Manoa EE 645, Fall. 2009,* 1-24.

Han, J., Kamber, M., & Pei, J. (2013). Data Mining: Concepts and Techniques. *Book, Chapter 9, Classification: Advanced Methods, University of Illinois at Urbana-Champaign & Simon Fraser University.*

Harrag, F., El-Qawasmah, E., & Al-Salman, A. M. S. (2010). Comparing Dimension Reduction Techniques for Arabic Text Classification Using BPNN Algorithm. *Paper presented in IEEE First International Conference on Integrated Intelligent Computing (ICIIC),* 6-11.

Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009). Improving Arabic text categorization using decision trees. *Paper presented in IEEE First International Conference on Networked Digital Technologies, NDT '09,* 110-115.

Haruechaiyasak, C., Jitkrittum, W., Sangkeettrakarn, C., & Damrongrat, C. (2008). Implementing News Article Category Browsing Based on Text

Categorization Technique. *Paper presented in IEEE International Conference on Web Intelligence and Intelligent Agent Technology WI-IAT '08*, 143 - 146.

He, D., & Wu, X. (2006). Ontology-Based Feature Weighting for Biomedical Literature Classification. *Paper presented in IEEE International Conference on Information Reuse and Integration*, 280-285.

He, J., Tan, A. H., & Tan, C. L. (2000). A comparative study on Chinese text categorization methods. *In PRICA 2000 Workshop on Text and Web Mining, Melbourne, Australia*, 24–35.

Hong-wei, Z., Jian-fang, C., & Feng, S.-q. (2010). An improved text feature selection method based on key words. *paper presented in IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), Vol. 2*, 293 - 297.

Hong, K. (2014). Improving the Estimation of Word Importance for News Multi-Document Summarization. *Technical Reports (CIS). Paper 989*.

Howard, B., W., Soonho, K., & Hagan, D. (2005). A Crop-Pest Ontology for Extension Publication. *5th Conference of the European Federation for Information Technology in Agriculture.*

Hui, D., & Siqing, Y. (2010). An improved feature weighting algorithm for Chinese text classification. *paper presented in IEEE International Conference on Computer Application and System Modeling (ICCASM), Vol. 6*, 433-436.

Hur, A., B. & Weston, J. (2010). A User's Guide to Support Vector Machines. *Department of Computer Science, Colorado State University, DOI: 10.1007/978-1-60327-241-4_13 Source: PubMed.*

Huth, J., Brogan M., Dancik B., Kommedahl T., Nadziejka D., Robinson P., & Swanson W. (1994). *Scientific format and style: The CBE manual for authors, editors, and publishers. Cambridge: Cambridge University Press*. p. 825.

Islam, M. R., & Islam, M. R. (2008). An effective term weighting method using random walk model for text classification. *Paper presented in IEEE 11th International Conference on Computer and Information Technology, ICCIT,* 411 - 414.

Jiang, H., Li, P., Hu, X., & Wang, S. (2009). An improved method of term weighting for text classification *paper presented in IEEE International Conference on Intelligent Computing and Intelligent Systems ICIS* 294 - 298

Jin, Y., Xiong, W., & Wang, C. (2010). Feature selection for Chinese Text Categorization based on improved particle swarm optimization. *Paper presented in IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE),* 1 – 6.

Joachims, T. (1998). Text categorization with support vector machines. *In European Conference on Machine Learning (ECML).*

Kadhim, M. H., & Omar, N. (2012). Automatic Arabic Text Categorization using Bayesian learning. *Paper presented in 7th IEEE International Conference on Computing and Convergence Technology (ICCCT),* 415 - 419.

Kaur, H., & Jyoti, K. (2013). Design and Implementation of Hybrid Algorithm for e-news Classification. *International journal of computers and technology, Vol. 12, No. 1*, 3178-3186.

Kehagias, A., Petridis, V., Kaburlasos, V. G., & Fragkou, P. (2001). A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms *Journal of Intelligent Information Systems Vol. 21*, 227-247.

Khan, A., Baharudin, B., & Khan, K. (2010). Semantic based features selection and weighting method for text classification. *Paper presented in IEEE International Symposium in Information Technology (ITSim), Vol. 2*, 850 – 855.

Khan, A., Baharudin, B., & Khan, K. (2012). Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification. *paper presented in IEEE Second International Conference on Computer Engineering and Applications (ICCEA), Vol. 2*, 398 - 403.

Kim, H. J., & Chang, J. (2007). Integrating Incremental Feature Weighting into Naive Bayes Text Classifier. *Paper presented in IEEE International Conference on Machine Learning and Cybernetics, Vol. 2*, 1137 – 1143.

Kruse, R., Rosner, D., & Nakhaeizadeh, G. (2001). Enhancing Text Classification to Improve Information Filtering. *Dissertation zur Erlangung des akademis chen Grades, Promotions colloquium: Magdeburg, den 07. December 2001*.

Korada, N., K., Kumar, N., S., P., Deekshitulu, Y., V., N., H. (2012). Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert

System. *International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.3*, 63-75.

Kumar, R. (2011). Research methodology: A step-by-step guide for beginners (3rd). *Thousand Oaks, CA: Sage Publications Inc.*

Lee, D.-l., Yang, S.-Y., & Hsu, C.-L. (2008). Ontology-supported webpage classifier for scholar's webpages in ubiquitous information environment. *Paper presented in First IEEE International Conference on Ubi-Media Computing*, 523 - 528.

Li, J. (2013). An approach to Meta feature selection. *Paper presented in IEEE 26th Annual Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1 – 4.

Li, X., & Liu, B. (2003). Learning to Classify Texts Using Positive and Unlabeled Data *In: Proceedings of the 19th international joint conference on artificial intelligence*.

Li, Y., & Chen, C. (2012). Research on the feature selection techniques used in text classification. *Paper presented in IEEE 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 725 – 729.

Li, Y., & Hu, D. (2009). Study on the Classification of Mixed Text Based on Conceptual Vector Space Model and Bayes. *Paper presented in IEEE International Conference on Asian Language Processing, IALP '09*, 269 - 272.

Liu, J. N. K., He, Y.-L., Lim, E. H. Y., & Wang, X.-Z. (2013). A New Method for Knowledge and Information Management Domain Ontology Graph Model.

Paper presented in IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 43 (Issue 1), 115-127.

Liu, Z., & Yang, J. (2011). A Feature Selection Simultaneously Based on Intra-category and Extra-Category for Text Categorization. *Paper presented in IEEE International Conference of Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 2*, 178-181.

Lord, L., (2010). Components of an Ontology. An Ontology Tutorial, Computing Science at Newcastle University.

Lu, Z., Shi, H., Zhang, Q., & Yuan, C. (2009). Automatic Chinese text categorization system based on mutual information. *Paper presented in IEEE International Conference on Mechatronics and Automation, ICMA*, 4986 - 4990

Luo, X., Ohyama, W., Wakabayashi, T., & Kimura, F. (2011). A Study on Automatic Chinese Text Classification. *Paper presented in IEEE International Conference on Document Analysis and Recognition (ICDAR)*, 920 - 924.

Luong, H. P., Gauch, S., & Wang, Q. (2009). Ontology-Based Focused Crawling. Paper *presented in IEEE International Conference on Information, Process, and Knowledge Management, eKNOW '09*, 123 - 128.

Ma, L., Ofoghi, B., Watters, P., & Brown, S. (2009). Detecting Phishing Emails Using Hybrid Features. *Paper presented in IEEE Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, UIC-ATC '09.* , 493 – 497.

Malone, J. & Parinson, H. (2010). Refrence and application Ontologies. *European Bioinformatics Institute, Cambridge, CB10 1SD, UK.*

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., & Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics.15; 26 (8) 10.1093*, 1112-1118.

Manning, C., D., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval. *Book ISBN: 0521865719. Cambridge University Press.*

Mathy, F. (2010). Assessing Conceptual Complexity and Compressibility Using Information Gain and Mutual Information. *Tutorials in Quantitative Methods for Psychology*, Vol. 6 (1), 16-30.

Maleki, M. (2010). Utilizing Category Relevancy Factor for text categorization. *Paper presented in IEEE 2nd International Conference on Software Engineering and Data Mining (SEDM)*, 334 - 339.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *ISBN: 9780521865715*.

Manuja, M., & Garg, D. (2014). Intelligent text classification system based on self-administered ontology. *Turkish Journal of Electrical Engineering & Computer Sciences*.

Meena, M. J., & Chandran, K. R. (2009). Classifying Text with Statistically Selected Features to Closely Related Classes. *Paper presented in IEEE International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom '09*, 297- 301.

Mesleh, A. M., & Kanaan, G. (2008). Support vector machine text classification system: Using Ant Colony Optimization based feature subset selection. *Paper*

presented in IEEE International Conference on Computer Engineering & Systems, ICCES, 143 - 148.

Mohaqeqi, M., Soltanpoor, R., & Shakery, A. (2009). Improving the Classification of Unknown Documents by Concept Graph. *Paper presented in IEEE 14th International CSI Computer Conference (CSICC'09), 259 - 264.*

Mohsenzadeh, M. , Mohaqeqi, M. , Soltanpoor, R. (2010). A New Approach for Better Document Retrieval and Classification Performance Using Supervised WSD and Concept Graph. *Paper presented in IEEE First International Conference on Integrated Intelligent Computing (ICIIC)*, 32 - 38.

Moschitti, A., & Basili, R. (2004). Complex Linguistic Features for Text Classification. *A comprehensive study Lecture Notes in Computer Science, Vol. 2997*, 181-196.

Mouratis, T., & Kotsiantis, S. (2009). Increasing the Accuracy of Discriminative of Multinomial Bayesian Classifier in Text Classification. *Paper presented in Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT '09*, 1246 – 1251.

Mousavi, H., Gao, S., & Zaniolo, C. (2013). Discovering attribute and entity synonyms for knowledge integration and semantic web search. *CSD/UCLA. Los Angeles, computer science Department technical report,* 1-12.

Munteanu, D. (2007). A Quick Survey of Text Categorization Algorithms. *The Annals of "Dunarea de Jos" University of Galati Fascicle ISSN 1221-454X*, 35-42.

Negoita, Marcia, G. H. (2004). Knowledge-Based Intelligent Information and Engineering Systems. *8th International Conference, KES 2004, Wellington, New Zealand,* September 20–25, 2004.

Nguyen, G. S., Gao, X., & Andreae, P. (2011). Phoneme Based Representation for Vietnamese Web Page Classification. *Paper presented in IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1,* 15 – 22.

Nguyen, T. T., Chang, K., & Hui, S. C. (2011). Supervised term weighting for sentiment analysis. *Paper presented in IEEE International Conference on Intelligence and Security Informatics (ISI),* 89 – 94.

Noh, S., Seo, H., Choi, J., Choi, K., & Jung, G. (2003). Classifying Web pages using adaptive ontology. *Paper presented in IEEE International Conference on Systems, Man and Cybernetics Vol. 3,* 2144 - 2149

Nuipian, V., Meesad, P., & Boonrawd, P. (2011). A comparison between keywords and key-phrases in text categorization using feature section technique. *Paper presented in IEEE 9th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering),* 156 – 160.

Pang, X.-L., Feng, Y.-Q., & Jiang, W. (2007). An Improved Document Classification Approach with Maximum Entropy and Entropy Feature Selection. *Paper presented in IEEE International Conference on Machine Learning and Cybernetics, Vol. 7,* 3911 - 3915.

Patra, A., & Singh, D. (2013). A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification

Algorithms. *International Journal of Computer Applications, Vol. 75, No.7,* 14-18.

Pawar, P. Y., & Gawande, S. H. (2012). A Comparative Study on Different Types of Approaches to Text Categorization. *Paper presented in IEEE International Journal of Machine Learning and Computing, Vol. 2, No. 4*, 295-301.

Pei, Z., Zhou, Y., Liu, L., Wang, L., Lu, Y., & Kong, Y. (2010). A mutual information and information entropy pair based feature selection method in text classification. *Paper presented in IEEE International Conference on Computer Application and System Modeling (ICCASM), Vol. 6*, 258-261.

Ping, Y., Zhou, Y. j., Yang, Y. X., & Peng, W. p. (2010). A novel term weighting scheme with distributional coefficient for text classification with support vector machine. *Paper presented in IEEE on Youth Conference International Information Computing and Telecommunications (YC-ICT)*, 182-185.

Polpinij, J. (2009). An ontology-based text processing approach for simplifying ambiguity of requirement specifications. *Paper presented in IEEE Asia-Pacific Services Computing Conference, APSCC*, 219 - 226.

Poyraz, M., Ganiz, M.C.,Akyokus, S. ; & Gorener, B. (2012). Exploiting Turkish Wikipedia as a semantic resource for text classification. *IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA),* 1-5.

Pote, R., M., & Akarti, S. P. (2014). Study of multiclass classification for imbalanced biomedical data. *International Journal of Application or Innovation in Engineering & Management (IJAIEM). Vol. 3, Issue 9*, 158-162.

Qin, T., Liu, T.-Y., Zhang, X.-D., Wang, D.-S., & Li, H. (2008). Global Ranking Using Continuous Conditional Random Fields. *Institution Microsoft Research Tech Report Number MSR-TR-156,* 1-8.

Rafi, M., Hassan, S., & Shaikh, M. S. (2012). Content-based Text Categorization using Wikitology. 1-9.

Raghunathan, P. (2003). Fast semi-automatic generation of ontologies and their exploitation. *Technical Report, Kansas State University*.

Rizvi, S. R. A., & Wang, S. X. (2010). DT-Tree: A Semantic Representation of Scientific Papers. *Paper presented in IEEE 10th International Conference on Computer and Information Technology (CIT)*, 1280 - 1284.

Roussey, C., Pinet, F., Kang, M., A. & Corcho, O. (2011). An introduction to ontologies and ontology engineering. *Ontologies in Urban Development Projects Advanced Information and Knowledge Processing Vol. 1, 2011, 9-38.*

Rujiang, B., & Junhua, L. (2009a). Improving Documents Classification with Semantic Features. *Paper presented in IEEE Second International Symposium on Electronic Commerce and Security, ISECS '09, Vol. 1,* 640 - 643.

Rujiang, B., & Junhua, L. (2009b). A Novel Conception Based Texts Classification Method. *Paper presented in IEEE International e-Conference on Advanced Science and Technology, AST '09*, 30 - 34.

Sajgal´ık, M. a., Barla, M., & Bielikov´a, M. a. (2013). From ambiguous words to key-concept extraction. *Paper presented in IEEE 24th International Workshop on Database and Expert Systems Applications*, 63-67.

Salkohe, G. (2006). Examples of ontology applications. *Seventh Agricultural Ontology Service Workshop Bangalore, India.*

Salton, G. (1971). The SMART retrieval system. *Experiments in Automatic Document Processing, Prentice-Hall, Upper Saddle River, NJ.*

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, Vol. 18, No. 11*, 613-620.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, Vol. 34, No. 1,* 1–47.

Shahi, A. M., Issac, B., & Modapothala, J. R. (2012). Enhanced intelligent text categorization using concise keyword analysis. *Paper presented in IEEE International Conference on Innovation Management and Technology Research (ICIMTR)*, 574 - 579.

Sharma, A., & Kuh, A. (2008). Class document frequency as a learned feature for text categorization. *Paper presented in IEEE World Congress on Computational Language Processing and Knowledge Engineering*, 2988 – 2993.

Shimodaira, H., (2014). Text Classification using Naïve Bayes. *Paper presented in Learning and Data Note*, 7, 1-9.

Singh, U., Goyal, V., & Rani, A. (2014). Disambiguating Hindi Words Using N-Gram Smoothing. *An International Journal of Engineering Sciences, Issue June 2014, Vol. 10, ISSN: 2229-6913*, 1-4.

Sini, M., Salokhe, G., Pardy, C., Albert, J., Keizer, J., & Katz, S. (2007). Ontology-based Navigation of Bibliographic Metadata: Example from the Food, Nutrition and Agriculture Journal. *Food and Agriculture Organization of the United Nations, Rome, Italy.*

Shein, K. P. P., & Nyunt, T. T. S. (2010). Sentiment Classification Based on Ontology and SVM Classifier. *Paper presented in IEEE Second International Conference on Communication Software and Networks, ICCSN '10*, 169 – 172.

Song, M.-H., Lim, S.-Y., Kang, D.-J., & Lee, S.-J. (2005). Automatic classification of Web pages based on the concept of domain ontology. *Paper presented in IEEE 12th Asia-Pacific Software Engineering Conference, APSEC '05.*

Soucy, P., & Mineau, G. W. (2005). Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. *In Proceedings of the 19th International Joint Conference on Artificial Intelligence IJCAI*, 1130-1135.

Tiun, S., Abdullah, R., & Kong, T. E. (2001). Automatic Topic Identification Using Ontology Hierarchy. *Paper published in Springer Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, Vol. 2004*, 444 – 453.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research,* 141-188.

Tong, S., Koller, D. (1998). Support Vector Machine Active Learning with Applications to Text Classification. *The Seventeenth International Conference on Machine Learning (ICML-00), Stanford, California*287-295.

Uchyigit, G. (2012). Experimental evaluation of feature selection methods for text classification. *Paper presented in IEEE 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1294 - 1298.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review AIAI-TR 191, Vol. 11, No 2*, 1-63.

Varela, P., N. (2012). Sentiment Analysis. *Dissertation submitted for obtaining the degree of Master in Electrical and Computer Engineering*.

Verleysen, M., Rossi, F., & François, D. (2009). Advances in Feature Selection with Mutual Information *Similarity-Based Clustering*, 52-69.

Wang, B. B., McKay, R. I., Abbass, H. A., & Barlow, M. (2002). Learning text classifier using the domain concept hierarchy. *Paper presented in IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions Vol. 2*, 1230-1234.

Wang, D., & Jiang, L. (2007). An improved attribute selection measure for decision tree induction. *Paper presented in IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD Vol. 4*, 654 – 658.

Wang, L., Liu, Z.-t., Wang, Y., Sun, R., & Liu, H. F. (2009). Event Feature and Personality-Event - Ontology Based for Classifying Chinese Web Pages.

Paper presented in IEEE Second International Workshop on Computer Science and Engineering, WCSE '09 Vol. 2, 555 - 557.

Wang, N., Wang, P., & Zhang, B. (2010). An improved TF-IDF weights function based on information theory. *Paper presented in IEEE International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE), Vol. 3*, 439 – 441.

Warintarawej, P., Laurent, A., Pompidor, P., Cassanas, A., & Laurent, B. (2011). Classifying Words: A Syllables-Based Model. *Paper presented in IEEE 22nd International Workshop on Database and Expert Systems Applications (DEXA)*, 208 - 212.

Wei, G., Gao, X., & Wu, S. (2010). Study of text classification methods for data sets with huge features. *Paper presented in IEEE 2nd International Conference on Industrial and Information Systems (IIS), Vol. 1*, 433 - 436.

Wei, Z., Guo-He, F., & Zheng, N. (2011). An Improved KNN Text Classification Algorithm Based on Clustering. *Paper presented in IEEE International Conference on Internet Technology and Applications (iTAP)*, 1-5.

Wen, J., & Li, Z. (2007). Semantic Smoothing the Multinomial Naive Bayes for Biomedical Literature Classification. *Paper presented in IEEE International Conference on Granular Computing, GRC*, 648.

Wibowo, A., Handojo, A., & Halim, A. (2011). Application of Topic Based Vector Space Model with WorldNet. *Paper presented in IEEE International Conference of Uncertainty Reasoning and Knowledge Engineering (URKE), Vol. 1*, 133-136.

Wu, G., & Liu, K. (2009). Research on Text Classification Algorithm by Combining Statistical and Ontology Methods. *Paper presented in IEEE Computational Intelligence and Software Engineering CiSE*, 1-4.

Xi, L., Hang, D., & Mingwen, W. (2012). An Effective Feature Selection Tool for Text Classification. *Paper presented in IEEE Fourth International Conference on Multimedia Information Networking and Security (MINES)*, 254 - 257.

Xia, F., Jicun, T., & Zhihui, L. (2009). A Text Categorization Method Based on Local Document Frequency. *Paper presented in IEE Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD '09, Vol. 7*, 468 – 471.

Xia, T., Chai, Y., & Wang, T. (2012). Improving SVM on web content classification by document formulation. *Paper presented in IEEE International Conference on Computer Science & Education (ICCSE)*, 110 - 113.

Xiao, S., Shi, Z., Liu, K., & Lv, X. (2010). A kind of Vector Space Representation Model based on Semantic in the field of English Standard Information 2010. *Paper presented in IEEE International Conference of Computational Intelligence and Security (CIS)*, 582 - 585.

Xiaoming, D., & Yan, T. (2013). Improved mutual information method for text feature selection. *Paper presented in IEEE 8th International Conference on Computer Science & Education (ICCSE)*, 163 - 166.

Xiaoyue, W., & Rujiang, B. (2009). Applying RDF Ontologies to Improve Text Classification. *Paper presented in IEEE International Conference on*

Computational Intelligence and Natural Computing, CINC '09, Vol. 2, 118 - 121.

Xu, Y. (2012). A comparative study on feature selection in Chinese Spam Filtering. *Paper presented in IEEE 6th International Conference on Application of Information and Communication Technologies (AICT)*, 1-6.

Xue, C., Qiu, Q.-Y., Feng, P.-E., & Yao, Z.-N. (2010). An automatic classification method for patents. *Paper presented in IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) Conference on, Vol. 4*, 1497 – 1501.

Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., and Chen, Z. (2006). Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing. *IEEE transactions on knowledge and data engineering, vol. 18, No. 2*, 1-14.

Yang, Y., Joachims, T. (2008). Text Categorization. *Scholarpedia*, 3(5):4242.

Yang, J., & Liu, Z. (2011). A feature selection based on deviation from feature centroid for text categorization. *Paper presented in IEEE International Conference of Intelligent Control and Information Processing (ICICIP), Vol. 1*, 180 – 184.

Yang, M., & Chen, H. (2012). Partially Supervised Learning for Radical Opinion Identification in Hate Group Web Forums. *Paper presented in IEEE International Conference of Intelligence and Security Informatics (ISI)*, 96-101.

Yang, Y., & Pederson, J. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning ICML '97*, *412*-420.

Yuan, M., Ouyang, Y. X., & Xiong, Z. (2013). A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets. *Journal of Information Science and Engineering*, 99-114.

Yunhe, W., Yuan, G., & Chao, X. (2013). Manifold Learning Method for Large Scale Dataset Based on Gradient Descent. *Article published by Atlantis Press ISSN: 1951-6851,* 1187-1194.

Yusof, N., & Hui, C. J. (2010). Determination of Bloom's cognitive level of question items using artificial neural network. *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10,* 866-870.

Zhan, Y., & Chen, H. (2012a). Feature extended short text categorization based on theme ontology. *Paper presented in IEEE 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 702 - 705.

Zhan, Y., & Chen, H. (2012b). Feature extended short text categorization based on theme ontology. *Paper presented in IEEE Fuzzy Systems and Knowledge Discovery (FSKD)*, 702 – 705.

Zhang, B., Xu, M., & Wu, M. (2012). Research on web filtering technology based on the dual feature selection. *Paper presented in 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, 675 - 679.

Zhang, G. P. (2000). Neural Networks for Classification: A Survey. *Paper presented in IEEE Transactions on Systems, man, and cybernetics, Vol. 30, No. 4,* 451-462.

Zhang, H., & Song, H.-t. (2006). Fuzzy Related Classification Approach Based on Semantic Measurement for Web Document. *Paper presented in Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops,* 615 - 619.

Zhang, W., Yoshida, T., & Tang, X. (2008). TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization. *Paper presented in IEEE International Conference on Systems, Man and Cybernetics SMC,* 108 – 113.

Zhang, X., Zhou, M., Dong, L., & Ye, N. (2009). Design of Chinese Text Categorization Classifier Based on Attribute Bagging. *Paper presented in IEEE International Conference on Business Intelligence and Financial Engineering, BIFE '09,* 201 - 204.

Zhang, W., Yoshida, T., Tang, X. & Ho, T. B. (2009). Improving effectiveness of mutual information for substantively multiword expression extraction. *Journal of Expert Systems with Applications 36*, 10919–10930.

Zhanga, W., Yoshidaa, T., & Tangb, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems, Vol. 21,* 879-886.

Zhanguo, M., Jing, F., Liang, C., Xiangyi, H., & Yanqin, S. (2011). An Improved Approach to Terms Weighting in Text Classification. *Paper presented in*

IEEE International Conference on Computer and Management (CAMAN), 1 – 4.

Zhanguo, M., Jing, F., Xiangyi, H., Yanqin, S., & Liang, C. (2011). Improved Terms Weighting Algorithm of Text. *Paper presented in IEEE International Conference of Network Computing and Information Security (NCIS), Vol 2*, 367 – 370.

Zhu, D., & Xiao, J. (2011). A Variety of tf-idf Term Weighting Strategy in Document Categorization. *Paper presented in IEEE Seventh International Conference on Semantics Knowledge and Grid (SKG)*, 83 – 90.

Zuo, J., Wan, M., & Ye, H. (2011). Text Classification Model Based on Markov Network Distance. *Journal of Computational Information Systems, Vol. 7: 9,* 3368-3375.