TOPIC IDENTIFICATION USING FILTERING AND RULE GENERATION ALGORITHM FOR TEXTUAL DOCUMENT

NURUL SYAFIDAH BINTI JAMIL

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

UNIVERSITY 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
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Maklumat yang disimpan secara digital dalam dokumen teks jarang disusun mengikut tajuk yang spesifik. Keperluan untuk membaca keseluruhan dokumen akan mengurangkan minat dalam pencarian maklumat. Kebanyakkan kaedah pengenalpastian tajuk bergantung kepada kekerapan perkataan yang muncul di dalam teks. Namun, bukan semua perkataan yang mempunyai kekerapan yang tinggi adalah relevan. Fasa penyarian perkataan dalam kaedah pengenalpastian tajuk menghasilkan perkataan yang mungkin mempunyai maksud yang sama yang dikenali sebagai masalah sinonim. Algoritma penyarian dan algoritma penjanaan peraturan diperkenalkan dalam kajian ini untuk mengenalpasti tajuk di dalam dokumen teks. Algoritma penyarian diperkenalkan (PFA) telah menyari perkataan yang paling berkaitan dari teks dan menyelesaikan masalah sinonim dalam kalangan perkataan yang di keluarkan. Algoritma penjanaan peraturan (TopId) diperkenalkan untuk mengenalpasti tajuk bagi setiap ayat berdasarkan perkataan yang dikeluarkan. PFA akan memproses dan menapis setiap ayat berdasarkan kata nama dan kata kunci yang telah ditetapkan untuk menghasilkan perkataan yang bersesuaian untuk tajuk. Peraturan kemudiannya dihasilkan dari perkataan yang dikeluarkan menggunakan pengkelasan berasaskan peraturan. Rekabentuk eksperimen telah dijalankan ke atas 224 ayat Bahasa Inggeris daripada terjemahan Al-Quran yang berkaitan dengan isu wanita. Tajuk yang dikenalpasti oleh TopId dan teknik Set Kasar dibandingkan dan kemudian disahkan oleh pakar. PFA telah berjaya menyari perkataan yang berkaitan berbanding dengan teknik penapisan yang lain. TopId telah mengenalpasti tajuk yang hampir sama dengan tajuk dari pakar dengan ketepatan 70%. Kedua-dua algoritma yang dicadangkan berupaya mengeluarkan perkataan yang berkaitan tanpa kehilangan perkataan yang penting dan mengenalpasti tajuk dalam ayat.

Kata kunci: Pengenalpastian tajuk, Algoritma penyarian, Algoritma penjanaan peraturan, Set Kasar, Ayat Al-Quran.

Abstract

Information stored digitally in text documents are seldom arranged according to specific topics. The necessity to read whole documents is time-consuming and decreases the interest for searching information. Most existing topic identification methods depend on occurrence of terms in the text. However, not all frequent occurrence terms are relevant. The term extraction phase in topic identification method has resulted in extracted terms that might have similar meaning which is known as synonymy problem. Filtering and rule generation algorithms are introduced in this study to identify topic in textual documents. The proposed filtering algorithm (PFA) will extract the most relevant terms from text and solve synonym roblem amongst the extracted terms. The rule generation algorithm (TopId) is proposed to identify topic for each verse based on the extracted terms. The PFA will process and filter each sentence based on nouns and predefined keywords to produce suitable terms for the topic. Rules are then generated from the extracted terms using the rule-based classifier. An experimental design was performed on 224 English translated Quran verses which are related to female issues. Topics identified by both TopId and Rough Set technique were compared and later verified by experts. PFA has successfully extracted more relevant terms compared to other filtering techniques. TopId has identified topics that are closer to the topics from experts with an accuracy of 70%. The proposed algorithms were able to extract relevant terms without losing important terms and identify topic in the verse.

Keyword: Topic identification, Filtering algorithm, Rule generation algorithm, Rough Set, Al-Quran verses.

Acknowledgement

All praise to Allah SWT who gave me strength and patience to finish this study. Alhamdulillah.

Firstly, I would like to express my gratitude to my first supervisor Prof. Dr. Ku Ruhana Binti Ku Mahamud for her support and her willingness to guide me on such an interesting research for my master thesis. I have learned so much from you, Prof. Secondly; I would like to thank my second supervisor Miss Aniza Binti Mohamed Din for providing me with great input and constant encouragement. Thank you for not losing hope on me from the beginning until the end of my study.

Furthermore, my sincere appreciation also goes to Associate Prof. Dr. Faudziah Binti Ahmad, Dr. Siti Sakira Binti Kamaruddin, Dr. Nooraini Binti Yusof, Dr. Yuhanis Binti Yusof and to Dr. Massudi Bin Mahmuddin who assisted and continuously helped me to understand and deepen my knowledge, especially in text mining and text classification related work. I am also truly indebted to Prof. Dr. Rosna Binti Awang-Hashim, Associate Prof. Dr. Norhafezah Binti Yusof, Prof. Dr. Nena Valdez and Associate Prof. Dr. Arminda Santiago for giving me moral support and continuously inspiring me to love my research and make this study even more interesting and joyful.

Not forgotten, gratitude to my beloved parents who never gave up in providing me with love and support to keep me strong. To my respected father, Jamil Bin Yusuff, you are the biggest reason why I am pursuing my study. Thank you for your valuable advice and motivation. To my cheerful mother, Midah Binti Musa, thank you Mama for supporting me to achieve my dream. Without your prayers, none of this work would have been accomplished.

Last but not least, thank you so much to all my lab-mates for our discussions and knowledge sharing session. It has been a great experience to know all of you. I will never forget our moments of tears and laughter in the years of our study. I wish all of you good luck in the future. To those people whom I have not mentioned here, thank you very much.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Appendices

List of Abbreviations

RSAR	Rough Set Attribute Reduction
PFA	Proposed Filtering Algorithm
TopId	Topic Identification
NLP	Natural Language Processing
POS	Part-of-Speech
VSM	Vector Space Model
TF-IDF	Term Frequency-Inverse Diverse Frequency
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
LSI	Latent Semantic Indexing
Pbuh	Praise Be Upon Him
IR	Information Retrieval
HMM	Hidden Markov Model

CHAPTER ONE

INTRODUCTION

The trend to store data in electronic format has increased and requires an efficient effort to organize these important yet beneficial documents (Sumathy & Chidambaram, 2013). Data can be stored in several formats such as image, video and audio. However, text is commonly used to store knowledge and exchange information (Sumathy & Chidambaram, 2013; Jusoh & Alfawaref, 2012; Mahender, 2012 & Aery, Ramamurthy, & Aslandogan, 2003). Text is usually unstructured and not restricted to any specific format (Jusoh & Alfawaref, 2012; Kamaruddin, 2011). The exploration of hidden information from this unstructured text is useful because interesting patterns and valuable knowledge can be discovered from the text (Sumathy & Chidambaram, 2013).

The issue in text mining research domain is to organize a vast amount of textual documents that have no specific topic or category describing the content (Aggarwal & Zhai, 2012; Jusoh & Alfawareh, 2012; Hotto, Nurnbeger & Paab, 2005). Text classification concerns on determining which decision is made on the information. There are many underlying classifiers for text classification such as Decision Tree, Rule-Based, Neural Network, K-Nearest Neighbour and Support Vector Machine (Aggarwal & Zhai, 2012). Topic identification is a method that lies under text mining domain and aims to analyze the text data and assign a correct label as a topic for the text documents (Baghdadi & Ranaivo-Malancon, 2011). However, analyzing texts is tedious, in which the complexity of natural language and misinterpretation that leads to misunderstanding might occur (Sumathy & Chidambaram, 2013; Jusoh &

The contents of the thesis is for internal user only

REFERENCES

Abdullah, Z., Kassim, J. M., & Saad, N. (2009). Pembangunan Perpustakaan Digital: Ayat al-Quran Berkaitan Wanita. *Asia-Pacific Journal of Information Technology and Multimedia*, *6*(1).

Abdullah, A.H.H., & Sudiro, S.R. (2010). Wanita menurut Hamka di dalam Tafsir Al-Azhar: Kajian terhadap Surah An-Nisa'.

Aery, M., Ramamurthy, N., & Aslandogan, Y. A. (2003). *Topic identification of textual data*. Technical report, The University of Texas at Arlington.

Aggarwal, C.C., & Zhai, C.X. (2012). A survey of text classification algorithm.

Ahmad, O., Hyder, I., Iqbal, R., Murad, M. A. A., Mustapha, A., Sharef, N. M., & Mansoor, M. (2013). A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran.*Life Science Journal*, *10*(4).

Ain, Q., & Basharat, A. (2011). Ontology driven information extraction from the holy Quran related documents. *26th IEEEP Students. Seminar 2011. UK*, 41-42.

Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I., & Al-Helwah, N. (2010). An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian Journal for Science and Engineering*, *35*(2), 21.

Ali, N. H., & Ibrahim, N. S. (2012). Porter Stemming Algorithm for Semantic Checking. ICCIT.

Amrani, A., Azé, J., Heitz, T., Kodratoff, Y., & Roche, M. (2004, December). From the texts to the concepts they contain: a chain of linguistic treatments. In *In Proceedings of TREC* (Vol. 4, pp. 712-722).

Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2008). A new document clustering algorithm for topic discovering and labeling. In *Progress in Pattern Recognition, Image Analysis and Applications* (pp. 161-168). Springer Berlin Heidelberg.

Atwell, E., Brierley, C., Dukes, K., Sawalha, M., & Sharaf, A. B. (2011). An Artificial Intelligence approach to Arabic and Islamic content on the internet. In*Proceedings of NITS 3rd National Information Technology Symposium*. Leeds.

Badawi, J. A. (1980). *Status of women in Islam*. Saudi Arabia Foreigners Guidance Center.

Badawi, J. A. (2000). The Status of Woman.

Badr, Y., Chbeir, R., Abraham, A., & Hassanien, A. E. Emergent Web Intelligence: Advanced Semantic Technologies. 2010.

Baghdadi, H. S., & Ranaivo-Malançon, B. (2011). An Automatic Topic Identification Algorithm. *Journal of Computer Science*, *7*(9), 1363.

Bakar, Z. A., & Rahman, N. A. (2003). Evaluating the effectiveness of thesaurus and stemming methods in retrieving Malay translated Al-Quran documents. In *Digital Libraries:*

Technology and Management of Indigenous Knowledge for Global Access (pp. 653-662). Springer Berlin Heidelberg.

Bakus, J., & Kamel, M.S. (2006). Higher order feature selection for text classification. Knowledge Information System. 9,4. 468-491.

Basit, H. A., & Jarzabek, S. (2007, September). Efficient token based clone detection with flexible tokenization. *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering* (pp. 513-516). ACM.

Baqai, S., Basharat, A., Khalid, A., Hassan, A., & Zafar, S. (2009). Leveraging Semantic Web Technologies for Standardized Knowledge Modeling and Retrieval from the Holy Quran and Religious Texts. ACM

Bazan, J. G., Nguyen, H. S., Nguyen, S. H., Synak, P., & Wróblewski, J. (2000). Rough set algorithms in classification problem. In *Rough set methods and applications* (pp. 49-88). Physica-Verlag HD.

Beniwal, S., & Arora, J. (2012, August). Classification and feature selection techniques in data mining. In *International Journal of Engineering Research and Technology* (Vol. 1, No. 6 (August-2012)). ESRSA Publications.

Berkowitz, S. (2010). *U.S. Patent No. 7,805,291*. Washington, DC: U.S. Patent and Trademark Office.

Berry, M. W., & Kogan, J. (Eds.). (2010). Text mining: Application and theory. Wiley.com.

Bigi, B., Brun, A., Haton, J. P., Smaili, K., & Zitouni, I. (2001, November). A comparative study of topic identification on newspaper and e-mail. In *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on* (pp. 238-241). IEEE.

Bolin, H. (2008). Knowledge extraction based on sentence matching and analyzing. *International Symposium on Knowledge Acquisition and Modelling.* 122-126.

Bong, C.H., & Wong, T.K. (2005). An examination of feature selection frameworks in text categorization. Information Retrieval Technology. 3689.

Brun, A., Smaïli, K., & Haton, J. P. (2002). Contribution to topic identification by using word similarity. In *INTERSPEECH*.

Bhumika., Sehra, S.S., & Nayyar, A. (2013). A review paper on algorithms used for text classification. International Journal of Application or Innovation in Engineering & Management (IJAIEM). 2(3).

Chizi, B., Rokach, L., & Maimon, O. (2009). A Survey of Feature Selection Techniques.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493-2537.

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using Encyclopedic knowledge for automatic topic identification. *Proceedings of the Thirteenth Conference on Computational Natural*

Language Learning (CoNLL), 210-218, Boulder, Colorado. Association for Computational Linguistics.

Dalal, M. K., & Zaveri, M.A. (2011). Automatic Text Classification : A Technical Review. In *International Journal of Computer Application*. *28*(2), 37–40.

Debruyne, M., & Verdonck,T. (2010). Robust kernel principal component analysis and classification.

Devasena, C. L., & Hemalatha, M. (2012, March). Automatic text categorization and summarization using rule reduction. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 594-598). IEEE.

Dong, R., Schaal, M., O'Mahony, M.P., & Smyth, B. (2013). Topic extraction from online reviews for classification and recommendation. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.*

Edward, H.Y.L., James, N.K.L., & Raymond, S.T.L. (2011). Text information retrieval. In: *Knowledge Seeker-Ontology Modelling for Information Search and Management.* DOI: 10.1007/978-3-642-17916-7_3

Elder, J., Hill, T., Delen, D., & Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Fuddoly, A., Jaafar, J., & Zamin, N. (2013, November). Keywords Similarity Based Topic Identification for Indonesian News Documents. In *Modelling Symposium (EMS), 2013 European* (pp. 14-20). IEEE.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., & Vanderwende, L. (2009). Using contextual speller techniques and language modeling for ESL error correction. *Urbana*, *51*, 61801.

Gelbukh, A., Espinoza, F. C., & Galicia-Haro, S. N. (Eds.). (2014). *Human-Inspired Computing and its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings* (Vol. 8856). Springer.

Gharaibeh, I.K, & Gharaibeh, N.K. (2012). Towards Arabic Noun Phrase Extractor (ANPE) Using Information Retrieval Techniques. *International Journal of Software Engineering*, *2*(2), 36–42. doi:10.5923/j.se.20120202.04

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., & Tserpes, K. (2012, June). Representation Models for Text Classification: a comparative analysis over three Web document types. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (p. 13). ACM.

Granitzer, M. (2003). *Hierarchical text classification using methods from machine learning* (p. 104).

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Hamzah, M. P., & Sembok, T. M. (2006, February). On Retrieval Performance of Malay Textual Documents. In *Artificial Intelligence and Applications* (pp. 156-161).

Hanum, H. M., Abu Bakar, Z., & Ismail, M. (2013, March). Evaluation of Malay grammar on translation of Al-Quran sentences using Earley algorithm. In *Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on* (pp. 1-4). IEEE.

Harrag, F., El-Qawasmah, E., & Al-Salman, A. M. S. (2011, April). Stemming as a feature reduction technique for Arabic Text Categorization. In *Programming and Systems (ISPS), 2011 10th International Symposium on* (pp. 128-133). IEEE.

Harish, B.S., Guru, D.S., & Manjunath, S. (2010). Representation and classification of text documents: A Brief Review. *IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition.*

Hassan, M. (2013). *Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge* (Doctoral dissertation, University of Waterloo).Wang, X., & Wang, J. (2013). A Method of Hot Topic Detection in Blogs Using N-gram Model. *Journal of Software*, *8*(1), 184-191.

Hassan, M. M., Karray, F., & Kamel, M. S. (2012, July). Automatic document topic identification using wikipedia hierarchical ontology. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on* (pp. 237-242). IEEE.

Hong, T. P., Lin, C. W., Yang, K. T., & Wang, S. L. (2013). Using TF-IDF to hide sensitive itemsets. *Applied Intelligence*, 1-9.

Hotto, H., Nurnberger, A., & Paab, G. (2005). A brief survey of text mining. *Journal for Computational Linguistics and Language Technology.* (20) 1, 9.

Huang, C. C., Tseng, T. L. B., Fan, Y. N., & Hsu, C. H. (2013). Alternative rule induction methods based on incremental object using rough set theory. *Applied Soft Computing*, *13*(1), 372-389.

Huynh, D., Tran, D., Ma, W., & Sharma, D. (2011, January). A new term ranking method based on relation extraction and graph model for text classification. In *Proceedings of the Thirty-Fourth Australasian Computer Science Conference-Volume 113* (pp. 145-152). Australian Computer Society, Inc.

Jain, S., & Pareek, J. (2010). Automatic topic(s) identification from learning material: An ontological approach. *2010 Second International Conference On Computer Engineering And Application.* Pp.358-362.

Janik, M., & Kochut, K. J. (2008, August). Wikipedia in action: Ontological knowledge in text categorization. In *Semantic Computing, 2008 IEEE International Conference on* (pp. 268-275). IEEE.

Janecek, A., Gansterer, W. N., Demel, M., & Ecker, G. (2008, September). On the Relationship Between Feature Selection and Classification Accuracy. In *FSDM* (pp. 90-105).

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005, August). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM.

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.

Khan, A., Baharudin, B. Lee, L.H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology. 1*(1).

Kamaruddin, S.S. (2011). *Framework for deviation detection in text.* Universiti Kebangsaan Malaysia, Bangi.

Kamruzzaman, S. M. (2010). Text Classification using Artificial Intelligence.*arXiv preprint arXiv:1009.4964*.

Kaplan, R. M. (2005). A method for tokenizing text. *Festschrift in Honor of Kimmo Koskenniemi's 60th anniversary. CSLI Publications, Stanford, CA*.

Keller, M., & Bengio, S. (2005). A neural network for text representation. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005* (pp. 667-672). Springer Berlin Heidelberg.

Khedikar, K. A., & Lobo, M. L. Data Mining: You've missed it If Not Used Lewis, D. D. (1991). Evaluating text categorization. *Proceedings of the workshop on Speech and Natural Language - HLT '91*, 312–318. doi:10.3115/112405.112471

Ko, Y., Park, J., & Seo, J. (2002, August). Automatic text categorization using the importance of sentences. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261-283.

Ku-Mahamud, K.R., Ahmad, F., Mohamed Din, A., Wan-Ishak, W.H., Ahmad, F.K., Din, R., & Che Pa, N. (2012). Semantic network representation of female related issues from the Holy Quran. *Knowledge Management International Conference (KMICe), Johor Bharu, Malaysia*, 4-6 July 2012. Pp.714-718.

Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010, October). Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 81-88). ACM.

Li, R., & Wang, Z. O. (2004). Mining classification rules using rough sets and neural networks. *European Journal of Operational Research*, *157*(2), 439-448.

Lin, C. Y. (1997). *Robust automated topic identification* (Doctoral dissertation, University of Southern California).

Mahar, J. A., Shaikh, H., & Memon, G. Q. (2012). A Model for Sindhi Text Segmentation into Word Tokens. *History*, *3*, 37-997.

Mahender, C. N. (2012). Text Clasification and Classifier: A Survey. *International Journal of Artificial Intelligence & Applications (IJAIA). 3*(2), 85–99.

Manacer, M., & Arbaoui, A. (2013). Content extraction of Quran Interpretation (Tafseer) books for digital content creation and distribution.

McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., & Rohlicek, J. R. (1994, April). Approaches to topic identification on the switchboard corpus. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on* (Vol. 1, pp. I-385). IEEE.

Mendes, A. C., & Antunes, C. (2009). Pattern mining with natural language processing: An exploratory approach. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 266-279). Springer Berlin Heidelberg.

Manne, S., & Fatima, S. S. (2011). A novel approach for text categorization of unorganized data based with information extraction. *International Journal on Computer Science and Engineering (IJCSE),* 3, pp. 2846-2854.

Manning, C.D. (2011). Part-of-Speech Tagging from 97% to 100%: Is it Time for Some Linguistic?. *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science.* 6608, pp. 171-189.

Moschitti, A., & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. In *Advances in Information Retrieval*(pp. 181-196). Springer Berlin Heidelberg.

Mukhtar, T., Afzal, H., & Majeed, A. (2012, December). Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran. In*Multitopic Conference (INMIC), 2012 15th International* (pp. 43-46). IEEE.

Na, F., Cai, W. D., & Zhao, Y. (2009). A method based on generation models for analyzing sentiment-topic in text.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An Introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–51. doi:10.1136/amiajnl-2011-000464

Natarajan, P., Prasad, R., Subramanian, K., Saleem, S., Choi, F., & Schwartz, R. (2007). Finding structure in noisy text: topic classification and unsupervised clustering. *International Journal of Document Analysis and Recognition (IJDAR)*, *10*(3-4), 187-198.

Nomponkrang, T., & Woraratpanya, K. (2010, September). Thai-sentence classification using conceptual graph. In *Educational and Information Technology (ICEIT), 2010 International Conference on* (Vol. 2, pp. V2-479). IEEE.

Noordin, M. F., & Othman, R. (2006, January). An information retrieval system for Quranic texts: a proposed system design. In *Information and Communication Technologies, 2006. ICTTA'06. 2nd* (Vol. 1, pp. 1704-1709). IEEE.

Nguyen, H. S., & Skowron, A. (2013). Rough Sets: From Rudiments to Challenges. In *Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam* (pp. 75-173). Springer Berlin Heidelberg.

Nuipan, V., & Meesad, P., & Boonrawd, P. (2012). A comparison between keywords and key-phrases in text categorization using feature section technique.ICT and Knowledge Engineering (ICT & Knowledge Engineering). IEEE.

Okhovvat, M., & Bidgoli, B.M (2011). A hidden Markov model for Persian part-of-speech tagging. *Proceedia Computer Science*, *3*, 977–981. doi:10.1016/j.procs.2010.12.160

Padhy, N., Mishra, D., & Panigrahi, R. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*

Podgorelec, V., & Zorman, M. (2009). *Decision TreesDecision tree* (pp. 1826-1845). Springer New York.

Protaziuk, G., Kryszkiewicz, M., Rybinski, H., & Delteil, A. (2007). Discovering compound and proper nouns. In *Rough Sets and Intelligent Systems Paradigms* (pp. 505-515). Springer Berlin Heidelberg.

Ramanathan, V., & Meyyapan, T. (2013). Survey of text mining. *International Conference on Technology and Business Management.*

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

Riloff, E. (1995, July). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 130-136). ACM.

Saad, S., Salim, N., & Zainal, H. (2009, November). Islamic knowledge ontology creation. In *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for* (pp. 1-6). IEEE.

Said, D. A. (2007). *Dimensionality reduction techniques for enhancing automatic text categorization* (Doctoral dissertation, Faculty of Engineering at Cairo University in Partial Fulfillment of the Requirements for the Degree of MASTER OF SCIENCE in COMPUTER ENGINEERING FACULTY OF ENGINEERING, CAIRO UNIVERSITY GIZA).

Sadiq, A. T., & Abdullah, S. M. (2012, November). Hybrid Intelligent Technique for Text Categorization. In *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on* (pp. 238-245). IEEE.

Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems*, *7*(2), 195-207.

Sagar, B. M., Shobha, G., & Kumar, R. (2009). Solving the Noun Phrase and Verb Phrase Agreement in Kannada Sentences. *International Journal of Computer Theory and Engineering*, *1*(3).

Sagayam, R., Srinivasan, S., & Roshini, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research*. 2,5

Salem, Y. (2009). A generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Sawalha, M., Brierley, C., & Atwell, E. (2012). Predicting Phrase Breaks in Classical and Modern Standard Arabic Text Experimental Dataset : The Quran, 3868–3872.

Sharaf, A. (2009). Knowledge Representation in the Quran. *Interim Report, University of Leeds*.

Sharaf, A., & Atwell, E. (2009). Knowledge Representation of the Quran Through Frame Semantics: A Corpus-Based Approach. *Corpus Linguistics-2009*, 12.

Sebastiani, F. (2005). Text Categorization.

Serrano, J. I., del Castillo, M. D., Oliva, J., & Iglesias, A. (2011). The influence of stop-words and stemming on human text base comprehension. *Proceedings of the European Perspectives on Cognitive Science*.

Sharaf, A. B. M., & Atwell, E. (2009). The Qur'an Annotation for Text Mining.

Silva, C., & Ribeiro, B. (2010). Background on Text Classification. In *Inductive Inference for Large Scale Text Classification* (pp. 3-29). Springer Berlin Heidelberg.

Skorkovská, L., Ircing, P., Pražák, A., & Lehečka, J. (2011, January). Automatic topic identification for large scale language modeling data filtering. In*Text, Speech and Dialogue* (pp. 64-71). Springer Berlin Heidelberg.

Srivastava, A., & Sahami, M. (Eds.). (2010). *Text mining: Classification, clustering, and applications*. CRC Press.

Stein, B., & Zu Eissen, S. M. (2004, June). Topic identification: Framework and application. In *Proc. International Conference on Knowledge Management* (Vol. 400, pp. 522-531).

Stoyanov, V., & Cardie, C. (2008, August). Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 817-824). Association for Computational Linguistics.

Sumathy, K. L., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues–An Overview. *International Journal of Computer Applications*, *80*(4), 29-32.

Syed, Z. S., Finin, T., & Joshi, A. (2008, March). Wikipedia as an Ontology for Describing Documents. In *ICWSM*.

Syiam, M. M., Fayed, Z. T., & Habib, M. B. (2006). An intelligent system for Arabic text categorization. *International Journal of Intelligent Computing and Information Sciences*, *6*(1), 1-19.

T.Sembok, T.M., Abu Bakar, Z., & Ahmad, F. (2011). Experiments in Malay Information Retrieval. 2011 *International Conference on Electrical Engineering and Informatics* 17-19 July. Bandung, Indonesia.

T.Sembok, T.M., Abu Ata, B.M., & Abu Bakar, Z. (2011). A rule-based Arabic stemming algorithm. *Proceedings of the 5th European Conference on European Computing Conference.* P.392-397, April 28-30,2011, Paris, France.

Tiun, S., Abdullah, R., & Kong, T. E. (2001). Automatic topic identification using ontology hierarchy. In *Computational Linguistics and Intelligent Text Processing* (pp. 444-453). Springer Berlin Heidelberg.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141-188.

Uysal, A. K., Gunal, S., Ergin, S., & Sora Gunal, E. (2012). The Impact of Feature Extraction and Selection on SMS Spam Filtering. *Elektronika ir Elektrotechnika*, *19*(5), 67-72.

Patel, F. N., & Soni, N. R. (2012). Text mining: A brief survey. *International Journal of Advanced Computer Research.* (4), 2-7.

Protaziuk, G., Kryszkiewicz, M., Rybinski, H., & Delteil, A. (2007). Discovering compound and proper nouns. In *Rough Sets and Intelligent Systems Paradigms* (pp. 505-515). Springer Berlin Heidelberg.

Van Zaanen, M., & Kanters, P. (2010). Automatic mood classification using Tf * IDF based on lyrics. Proceedings of the 11th International Society for Music Information Retrieval Conference, August 9-13, 2010, Utrecht, Netherlands, pp: 75-80.

Ventura, J., & da Silva, J. F. (2008). *Ranking and extraction of relevant single words in text*. INTECH Open Access Publisher.

Von Denffer, A. (1983). Ulum al Quran. *The Islamic Foundation*,

Wang, X., & Wang, J. (2013). A method of hot topic detection in blogs using N-gram model. Journal of Software, (8),1.

Xu, J., Lu, Q., & Liu, Z. (2012, April). Combining classification with clustering for web person disambiguation. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 637-638). ACM.

Xu, Y., Zhang, D., & Yang, J. Y. (2010). A feature extraction method for use with bimodal biometrics. *Pattern recognition*, *43*(3), 1106-1115.

Yuan, L. (2010). An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing. *Science*, 267-269.

Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013a, March). Ontology semantic approach to extraction of knowledge from Holy Quran. In *Computer Science and Information Technology (CSIT), 2013 5th International Conference on* (pp. 19-23). IEEE.

Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013b). Quranic Verse Extraction base on Concepts using OWL-DL Ontology. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(23), 4492-4498.

Zhang, H., Wang, D., Wu, W., & Hu, H. (2012). Term frequency–function of document frequency: a new term weighting scheme for enterprise information retrieval. *Enterprise Information Systems*, *6*(4), 433-444.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765.

Zhong, N., Li, Y., & Wu, S. T. (2012). Effective pattern discovery for text mining. *Knowledge and Data Engineering, IEEE Transactions on*, *24*(1), 30-44.

Zhao, W. X., Chen, R., Fan, K., Yan, H., & Li, X. (2012, July). A novel burst-based text representation model for scalable event detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 43-47). Association for Computational Linguistics.