# SEMANTIC FEATURE REDUCTION AND HYBRID FEATURE SELECTION FOR CLUSTERING OF ARABIC WEB PAGES

HANAN MUSAFER H. ALGHAMDI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

DECEMBER 2016

To the soul of my Father, may Allah forgive him.

To my beloved Mother, for her ongoing love and prayers to me.

To my Husband, for care, support and encouragement all the time.

And to my Children's, Faisel and Yousef with hope for proud future.

## ACKNOWLEDGEMENT

I would like to acknowledge, and I wish to express my deep gratitude to my supervisor, Professor Ali bin Selamat for guidance, valuable time, technical and friendly dealing through out my study.

I would like to express my profound gratitude to Ministry of Education, Malaysia, Universiti Teknologi Malaysia (UTM), Umm Al-Qura University (UQU) and the Ministry of Higher Education, Saudi Arabia for supporting this research.

Finally but as important, my deepest gratitude is due to my family members and my friends for their loving support. Without their encouragement and understanding it would have been impossible for me to finish this work.

# ABSTRACT

In the literature, high-dimensional data reduces the efficiency of clustering algorithms. Clustering the Arabic text is challenging because semantics of the text involves deep semantic processing. To overcome the problems, the feature selection and reduction methods have become essential to select and identify the appropriate features in reducing high-dimensional space. There is a need to develop a suitable design for feature selection and reduction methods that would result in a more relevant, meaningful and reduced representation of the Arabic texts to ease the clustering process. The research developed three different methods for analyzing the features of the Arabic Web text. The first method is based on hybrid feature selection that selects the informative term representation within the Arabic Web pages. It incorporates three different feature selection methods known as Chi-square, Mutual Information and Term Frequency–Inverse Document Frequency to build a hybrid model. The second method is a latent document vectorization method used to represent the documents as the probability distribution in the vector space. It overcomes the problems of high-dimension by reducing the dimensional space. To extract the best features, two document vectorizer methods have been implemented, known as the Bayesian vectorizer and semantic vectorizer. The third method is an Arabic semantic feature analysis used to improve the capability of the Arabic Web analysis. It ensures a good design for the clustering method to optimize clustering ability when analysing these Web pages. This is done by overcoming the problems of term representation, semantic modeling and dimensional reduction. Different experiments were carried out with *k*-means clustering on two different data sets. The methods provided solutions to reduce high-dimensional data and identify the semantic features shared between similar Arabic Web pages that are grouped together in one cluster. These pages were clustered according to the semantic similarities between them whereby they have a small Davies–Bouldin index and high accuracy. This study contributed to research in clustering algorithm by developing three methods to identify the most relevant features of the Arabic Web pages.

# ABSTRAK

Dalam kajian lepas, data dimensi tinggi dapat mengurangkan kecekapan dalam algoritma pengklusteran. Pengklusteran teks Arab merupakan sesuatu yang mencabar kerana semantik dalam teks melibatkan pemprosesan semantik yang mendalam. Bagi mengatasi masalah ini, pemilihan ciri-ciri dan kaedah pengurangan menjadi penting dalam memilih dan mengenal pasti ciri-ciri yang bersesuaian bagi mengurangkan ruang dimensi yang tinggi. Terdapat keperluan untuk membangunkan reka bentuk yang bersesuaian dalam pemilihan ciri-ciri dan kaedah pengurangan yang akan menyebabkan perwakilan teks Arab yang lebih relevan, bermakna dan kurang bagi memudahkan proses pengklusteran. Kajian ini membangunkan tiga kaedah yang berbeza untuk menganalisis ciri-ciri teks bagi Web Bahasa Arab. Kaedah pertama adalah berdasarkan kepada pemilihan ciri-ciri hibrid yang memilih perwakilan jangka bermaklumat dalam halaman Web Bahasa Arab. Ia menggabungkan tiga kaedah pemilihan ciri yang berbeza yang dikenali sebagai Khi-Kuasa Dua, Maklumat Bersama dan Frekuensi Dokumen Frekuensi Songsang Bertempoh untuk membina sebuah model hibrid. Kaedah kedua merupakan kaedah pemvektor dokumen terpendam yang digunakan untuk mewakili dokumen sebagai taburan kebarangkalian dalam ruang vektor. Ia mengatasi masalah dimensi tinggi dengan mengurangkan ruang dimensi. Bagi mengekstrak ciri-ciri yang terbaik, dua kaedah pemvektor dokumen telah dilaksanakan yang dikenali sebagai pemvektor Bayesian dan pemvektor semantik. Kaedah ketiga adalah analisis ciri-ciri semantik Arab yang digunakan untuk meningkatkan keupayaan analisis Web Bahasa Arab. Ia memastikan reka bentuk terbaik untuk kaedah pengklusteran bagi mengoptimumkan keupayaan pengklusteran apabila menganalisis laman Web ini. Ini dilaksanakan dengan mengatasi masalah perwakilan jangka, pemodelan semantik dan pengurangan dimensi. Penyelidikan yang berbeza telah dijalankan dengan pengklusteran k-cara ke atas dua set data yang berlainan. Kaedah ini dapat menyelesaikan pengurangan dimensi data yang tinggi dan mengenal pasti ciri-ciri semantik yang dikongsi bersama laman Web Bahasa Arab yang dikumpulkan bersama-sama dalam satu kluster. Laman ini telah diklusterkan mengikut persamaan semantik antara mereka di mana mereka mempunyai indeks terkecil Davies-Bouldin dan ketepatan yang tinggi. Kajian ini menyumbang kepada penyelidikan dalam pengklusteran algoritma dengan membangunkan tiga kaedah untuk mengenal pasti ciri-ciri yang paling relevan dalam laman Web Bahasa Arab.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| ACM | - | Association for Computing Machinery |
| AI | - | Artificial Intelligence |
| ASD | - | Average Shortest Distance |
| ASFA | - | Arabic Semantic Feature Analysis |
| ATC | - | Anti-terrorism Coalition |
| ATDS | - | Advanced Terror Detection System |
| BV | - | Bayesian Vectorization |
| BV-SV | - | Hybrdization of Baysian Vectorization and Semantic Vectorization |
| CCA | - | Content and Composition Analysis |
| CHI | - | Chi-square |
| DBI | - | Davies–Bouldin index |
| DF | - | Document Frequency |
| EM | - | Expectation-Maximization Estimator |
| FF | - | Feature Frequency |
| FRS | - | Feature Relation Strength |
| GA | - | Genetic Algorithm |
| H.L. | - | High Level of attribute |
| H_FS | - | Hybrid feature Selection |
| H_FS-BV-SV | - | Hybrdization of Hybrid feature Selection with Baysian Vectorization and Semantic Vectorization |
| IDF | - | Inverse Document Frequency |
| IEDs | - | Improvised Explosive Devices |
| IEEE | - | Institute of Electrical and Electronics Engineers |
| IG | - | Information Gain |
| KL | - | Kullback-Leibler distance |

| | | |
|---|---|---|
| L.A. | - | Latin American groups |
| L.L. | - | Low Level of attribute |
| LDA | - | Latent Dirichlet Allocation |
| LSA | - | Latent Semantic Analysis |
| LSI | - | Latent Semantic Indexing |
| M.E. | - | Middle Eastern groups |
| M.L. | - | Medium Level of attribute |
| MEMRI | - | Middle East Media Research Institute |
| MI | - | Mutual Information |
| N/A | - | Not Applicable |
| PCA | - | Principal Component Analysis |
| SLR | - | Systematic Literature Review |
| SNA | - | Social Network Analysis |
| SV | - | Semantic Vectorization |
| SVa | - | Semantic Vectorization using Semantic Class Density |
| SVb | - | Semantic Vectorization using Semantic Class Probability Distribution |
| SVM | - | Support Vector Machine |
| TC | - | Term Contribution |
| TF | - | Term Frequency |
| TF-BV-SV | - | Hybrdization of Baysian vectorization and semantic vectorization using term frequency feature selection |
| TFIDF | - | Term Frequency–Inverse Document Frequency Weight |
| TFIDF-BV-SV | - | Hybrdization of Baysian Vectorization and Semantic Vectorization using Term Frequency-Invers Document Frequency Feature Selection |
| TF-PLSA | - | Probability Latent Semantic Analysis using Term Frequency Feature Selection |
| TS | - | Term Strength |
| U.S. | - | U.S. Domestic groups |
| VSM | | Vector Space Model |
| WWW | | World Wide Web |

# LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Abundant amounts of Arabic text are currently available on the World Wide Web (WWW) in electronic form. The unorganized information in these textual data (Elarnaoty, *et al.*, 2012) has encouraged various new studies on managing this vast information to classify relevant information and to accordingly enhance the organization of text available on the WWW.

Document clustering is among the methods employed to group documents containing related information into clusters, which facilitates the allocation of relevant information. This technique can efficiently enhance the search process of a retrieval system (Alsulami *et al.*, 2012), aids with the process of identifying crime patterns (Nath, 2006), helps extract types of crimes from documents (Alruily *et al.*, 2010), and can facilitate determining hidden or unknown affiliations within a social network (Qi *et al.*, 2010). Clustering is a method of grouping data items that have similar characteristics, while samples in different groups are dissimilar.

An effectively built clustering algorithm must transform free running text into structured data using a document representation model. The Vector Space Model (VSM) is the most widely used approach for this purpose and adopts Bag-of-Words (BOW) to express text. With VSM, text content is represented as vectors in a specific feature space using a word index, where each vector value corresponds to the occurrence or absence of a selected feature. The most commonly employed features in VSM are words, while other techniques use characters and phrases as features (Zhang and Zhang, 2006).

Although considerable work has been published on Arabic Web page classification, little published research related to Arabic Web page clustering is available (Abuaiadah, 2016; Froud *et al.*, 2013; Ghanem, 2014). Arabic is a morphologically rich (Al-Khalifa and Al-Wabil, 2007) and highly inflectional language (Beseiso *et al.*, 2011); consequently, many clustering algorithms developed for the English language perform poorly when applied to Arabic (Abuaiadah, 2016). Developing a machine-understandable system for Arabic involves discriminating and deeply semantic processing. Accordingly, interest in research on Arabic language processing has been increasing.

## 1.2 Problem Background

The fundamental challenges with clustering Arabic Web pages include identifying the most informative features to best represent original content and designing feature discriminating vectors in order to analyze large volumes of unstructured Arabic text. The performance of text-based systems is highly dependent on the representation of text in the input space (Leopold and Kindermann, 2002; Lewis, 1990). A number of studies have been done to address these difficulties in terms of Arabic Web page clustering and proposing solutions.

A problem with identifying relevant features is derived from treating terms as independent from each other and neglecting the semantic relations and category popularity among terms, which often leads to synonym and polysemy problems (Hu *et al.*, 2008) or missing category problems (Hu *et al.*, 2009). Such problems can produce very low similarity scores for related documents because two samples with a semantic or category relation and two other samples without this relation are grouped similarly. Consequently, the effectiveness of the document clustering method is reduced. Some studies have investigated Arabic text representation and used several approaches based on language-dependent techniques. Bsoul and Mohd (2011), Froud *et al.* (2010), Ashour (2012), Al-Omari (2011), Amine *et al*.(2013), Ahmed and Tiun (2014) and Ghanem (2014) have utilized the word stemming approach to represent manipulated texts. Stemming is the process of removing morphological affixes from *words* to get the *word root*. Other researchers, such as Sahmoudi *et al.* (2013) and El-beltagy (2006) have used keyphrase extraction, which is defined as a process of identifying a set of words or phrases that express a document using the Suffix Tree algorithm, after which these words are used for text representation.

Stemming and keyphrase extraction approaches focus on the morphological aspect of text and ignore the semantics of terms and the semantic or category relations. Manual keyphrase assignment can be time consuming, especially when large volumes of Web pages are involved (Ali and Omar, 2014). Additionally, each generated keyphrase may be attached to a number of keyphrases that are part of this keyphrase, and the difficulty arises in selecting the relevant ones (Sahmoudi and Lachkar, 2016). In the literature, the benefits of using stemming to identify the relevant features for Arabic text clustering are debated. Al-Anzi and AbuZeina (2016), Al-Omari (2011) and Said *et al*. (2009) have reported that stemming is not always beneficial for Arabic text-based tasks, since many terms may be combined with the same root form. In addition, multiple entries may be created in the text representation model for different words that carry the same meaning (Awajan, 2015a). On the other hand, Said *et al*. (2009) demonstrated that using stemming in combination with a good feature selection method improves the performance of Arabic text clustering. Feature selection is aimed at selecting the most relevant subset of existing features without transformation and then using these subset features for text representation. A better feature selection method is desired to identify informative features, and which is able to consider semantic and category relations to represent high-similarity Arabic Web content in computer-understandable form.

The problem of high dimensionality stems from the large number of variables considered in text clustering methods. All terms found in a document are included in the clustering process, which leads to a very large number of dimensions in the vector representation of the document. Therefore, high-dimensional data reduces the efficiency of clustering algorithms and maximizes execution time. Some researchers have suggested solutions for the high dimensionality problem in clustering Arabic

Web page content. Awajan (2015a, 2015b) proposed a semantically enriched and reduced vector space model (VSM). Harrag *et al.* (2010) used a feature selection technique with VSM to reduce high-dimensional data. Their results showed that the DF, TFIDF and LSI techniques are more effective and efficient than stemming techniques. The Frequent Itemset-based Hierarchical Clustering (FIHC) approach proposed by Al-sarrayrih and Al-Shalabi (2009) involves finding frequent word sets, which are then used to cluster documents.

However, FIHC may produce low-quality clusters due to considering the number of word occurrences in a document as part of the clustering criteria (Backialakshmi, 2015). The high dimensionality of document representation using VSM is a potential problem, since not all documents in a collection contain all words used in the representation, and therefore sparseness occurs extensively in the document vectors (Ampazis and Perantonis, 2004; Zhang *et al*., 2010). Furthermore, considering keywords alone cannot capture all the similar information between documents, such as word proximity, semantic features and word distributions among categories (Osinski, 2004; Shaban, 2009). A study by Turney and Pantel (2010) revealed that the main alternative to VSM is a probabilistic model based on creating a probabilistic language model for text vectorization and document clustering according to the measured probability in that model. Accordingly, there is a need for a vectorization technique that transforms existing features into a lower-dimensional space, taking into account background information such as feature probability distribution and semantic information in order to compact and enrich the document representation for clustering.

Another problem faced in clustering Arabic Web pages is the architecture design of the respective clustering method. A challenging task is to realize how to enhance the clustering performance. In general, the aptitude of Arabic Web page clustering is highly based on the input features' characteristics. The performance of a Web page clustering technique is only effective when the appropriate feature selection and feature reduction methods are integrated with a proper clustering method (Ghanem, 2014). Improving clustering performance requires computational algorithms that adapt appropriate feature selection or reduction methods to well-established clustering approaches capable of achieving higher performance (Jain and Murty, 1999). Thus, there is a need to develop a suitable design for feature selection and feature reduction methods for more relevant and reduced Arabic text representation, which can facilitate optimizing the clustering ability for Arabic Web page analysis.

## 1.3 Problem Statement

The ideal Arabic Web page clustering depends on features representative of the content. However, Web pages contain a vast number of distinct features that produce high-dimensional data, which makes the clustering process more difficult. Therefore, it is important to enhance feature selection and reduction methods in order to solve the issue of high-dimensional data and identify the most informative feature set to enhance Arabic Web page clustering performance.

## 1.4 Goal of the study

The goal of this research is to develop and enhance feature reduction and feature selection methods that can be used to improve Arabic Web page clustering.

## 1.5 Research Questions

According to the research problem presented above, the following research questions are introduced:

Q 1. How can Arabic Web pages be represented for optimized clustering?

Q 2. How can the feature sets generated by different feature selection methods be hybridized to obtain the most relevant features?

Q 3. What is an appropriate dimensional reduction algorithm to use in solving the problem of high-dimensional data taking into account semantic and category information?

Q 4. How can a feature selection and reduction method be designed that is adaptable to a clustering approach?

Q 5. Do the proposed methods produce accurate clustering results?

## 1.6 Research Objectives

This study comprises three main objectives.

i. To improve the feature selection ability of Arabic Web page clustering by proposing a hybrid feature selection method.

ii. To propose a latent document vectorization model for enhancing document representation for Arabic Web clustering.

iii. To propose an Arabic semantic feature analysis method by hybridizing the proposed hybrid feature selection with the proposed latent document model to efficiently reduce feature space dimensionality as well as achieve higher clustering performance.

## 1.7 Scope of the Study

The scope of this research is limited to the following:

i. The study focuses on feature selection and reduction methods for Arabic Web page clustering.

ii. Focus is on Arabic text without using any machine translation.

iii. This research specifically focuses on Arabic language textual content obtained from Arabic newspaper and Dark Web site archives.

iv. The *K*-means clustering method is used for Arabic Web content analysis.

v. The evaluation of the proposed methods' ability to cluster Arabic Web pages is based on information retrieval measurements, i.e., purity, DBI and F-measure.

## 1.8 Contributions and Significance of the Study

i. The first contribution is in proposing a hybrid feature selection method that integrates three different selection methods, namely Chi-square (CHI), Mutual Information (MI) and Term Frequency-Inversed Document Frequency (TF–IDF).

ii. The second contribution is in proposing three different latent document vectorization methods. The first method is semantic vectorization using semantic class density (SVa). The second method is semantic vectorization using estimated probability distribution of semantic classes (SVb). The third method is Bayesian vectorization (BV) using estimated probability distribution of categories.

iii. The third contribution is in developing an algorithm called Arabic Semantic Feature Analysis (ASFA) that enhances feature selection and reduction for analyzing Web textual content using the k-means clustering method.

iv. The fourth contribution is in revealing the importance of the feature selection and feature reduction methods for improving Arabic Web page clustering.

v. The fifth contribution is in performing an empirical investigation and revealing how the proposed methods are useful for analyzing Arabic Web pages.

## 1.9 Outline of the Thesis

This chapter provided an overview of the aims of conducting this research. It comprises an introduction, problem statement, objectives, research questions, scope and contributions. The summary and organization of this thesis are as follows:

i. Chapter 1 presents the research with an introduction, problem statement, objectives, research questions, scope and contributions.

ii. Chapter 2 reports a review of literature on Arabic text clustering.

iii. Chapter 3 provides the methodology used to achieve the objectives of this research.

iv. Chapter 4 describes the methodology, implementation and experimental results of the first Arabic Web page text clustering approach with the proposed feature selection method.

v. Chapter 5 demonstrates the methodology, implementation and experimental results of the second Arabic Web page text clustering approach with the document vectorization method.

vi. Chapter 6 presents the methodology, implementation and experimental results of the third Arabic Web page text clustering approach with the proposed hybrid method.

vii. Chapter 7 presents the conclusions of this study.

# REFERENCES


Abbasi, A., and Chen, H. (2005). Applying Authorship Analysis to Extremist-Group Web Forum Messages. *Homeland Security*. *20*(5), 67–75.

Abbasi, A., and Chen, H. (2007). Affect Intensity Analysis of Dark Web Forums. In *The International Conference on Intelligence and Security Informatics*. New Brunswick, NJ: IEEE, 282–288.

Abbasi, A., and Chen, H. (2008). Analysis of Affect Intensities in Extremist Group Forums. In H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor (Eds.), *Intelligence and Security Informatics* (pp. 285–307). Springer US.

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*. *26*(3), 1–34.

Abuaiadah, D. (2016). Using Bisect K-Means Clustering Technique in the Analysis of Arabic Documents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* *15*(3), 1–13.

Ahmed, M. H., and Tiun, S. (2014). K-Means based Algorithm for Islamic Document Clustering. *International Journal on Islamic Applications in Computer Science and Technologies*. *2*(3), 1–8.

Ahmed, Z. (2009). *Domain Specific Information Extraction for Semantic Annotation*. Diploma Thesis, Charles University of Prague and University of Nancy.

Al-Anzi, F. S., and AbuZeina, D. (2015). Stemming Impact On Arabic Text Categorization Performance: A Survey. In *Information and Communication Technology and Accessibility*. Marrakech, Morocco: IEEE, 1–7.

Al-Anzi, F. S., and AbuZeina, D. (2016). Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering. In *International Conference on Engineering Technologies and Big Data Analytics*. Bangkok, Thailand: IIE, 1–4.

Alelyani, S., Tang, J., and Liu, H. (2016). Feature Selection for Clustering : A Review. In C. C. Aggarwal and C. K. Reddy (Eds.), *Data Clustering: Algorithms and Applications* (pp. 29–60). CRC Press.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., and Al-Rajeh, A. (2008). Automatic Arabic Text Classification. In *The International Conference on the Statistical Analysis of Textual Data*. Lyon, France, 77–84.

Ali, N. G., and Omar, N. (2014). Arabic Keyphrases Extraction Using a Hybrid of Statistical and Machine Learning Methods. In *International Conference on Information Technology and Multimedia*. Putrajaya, Malaysia: IEEE, 281–286.

Aljlayl, M., and Frieder, O. (2002). On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In *International conference on Information and Knowledge Management*. New York,USA: ACM, 340–347.

Al-Khalifa, H., and Al-Wabil, A. (2007). The Arabic language and the Semantic Web: Challenges and Opportunities. In *The International Symposium on Computers and Arabic Language and Exhibition*. Riyadh, Saudi Arabia, 1–9.

Almeida, L. G. P., Vasconcelos, A. T. R., and Maia, M. A. G. (2009). A Simple and Fast Term Selection Procedure for Text Clustering. *Intelligent Text Categorization and Clustering*. *164*, 47–64.

Al-Omari, O. (2011). Evaluating The Effect Of Stemming In Clustering Of Arabic Documents. *Academic Research International*. *1*(1), 284–291.

Alrahabi, M., Ibrahim, A., and Desclés, J.-P. (2006). Semantic Annotation of Reported Information in Arabic. In *FLAIRS 2006*. Floride: AAAI Press, 263–268.

Alruily, M., Ayesh, A., and Al-Marghilani, A. (2010). Using Self Organizing Map to Cluster Arabic Crime Documents. In *International Multiconference on Computer Science and Information Technology*. Wisla, Poland: IEEE, 357–363.

Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *International Arab Journal of E-Technology*. *2*(2), 124–128.

Alsaleem, S. M. (2013). Neural Networks for the Automation of Arabic Text Categorization. In *International Conference on Computer Applications Technology (ICCAT)*. Sousse, Tunisia: IEEE, 1–6.

Al-sarrayrih, H., and Al-Shalabi, R. (2009). Clustering Arabic Documents Using Frequent Itemset-based Hierarchical Clustering with an N-Grams. In *The International Conference on Information Technology*. Jordan, Amman, 1–8.

Al-Shalabi, R., and Kanaan, G. (2004). Constructing an Automatic Lexicon for Arabic Language. *International Journal of Computing and Information Sciences*. *2*(2), 114–128.

Al-shammari, E. (2010). Improving Arabic Document Categorization : Introducing Local Stem. In *International Conference on Intelligent Systems Design and Applications*. Cairo, Egypt: IEEE, 385–390.

Al-Shammari, E., and Lin, J. (2008). Towards an Error-Free Arabic Stemming. In *Proceeding of the 2nd ACM workshop on Improving non English web searching - iNEWS '08*. California, USA: ACM Press, 9–15.

Al-shawakfa, E., Al-Badarneh, A., Shatnawi, S., Al-Rabab'ah, K., and Bani-Ismail, B. (2010). A comparison Study of Some Arabic Root Finding Algorithms. *Journal of the American Society for Information Science and Technology*. *61*(5), 1015–1024.

Alsulami, B. S., Abulkhair, M. F., and Essa, F. A. (2012). Semantic Clustering Approach Based Multi-agent System for Information Retrieval on Web. *International Journal of Computer Science and Network Security*. *12*(1), 41–46.

Amine, A., Mohamed, O. A., Bellatreche, L., Biskri, I., Rompré, L., Jouis, C., Achouri, A., Descoteaux, S., Bensaber, B. A. (2013). Clustering with Probabilistic Topic Models on Arabic Texts. In Amine, A., Otmane, A. M., Bellatreche, L. (Ed.) *Modeling Approaches and Algorithms for Advanced Computer Applications* (pp. 37–46). Switzerland: Springer International Publishing.

Ampazis, N., and Perantonis, S. J. (2004). LSISOM – A Latent Semantic Indexing Approach to Self-Organizing Maps of Document Collections. *Neural Processing Letters*. *19*(2), 157–173.

Andrews, N. O., and Fox, E. A. (2007). *Recent Developments in Document Clustering.* Technical Report TR-07-35, Computer Science, Virginia Tech.

Antony, D. A., Singh, G., Leavline, E. J., Priyanka, E., and Sumathi, C. (2016). Feature Selection Using Rough Set For Improving the Performance of the Supervised Learner. *International Journal of Advanced Science and Technology*. *87*, 1–8.

Anwar, T., and Abulaish, M. (2012). Identifying cliques in dark web forums - An agglomerative clustering approach. In *International Conference on Intelligence and Security Informatics*. Washington, D.C., USA: IEEE, 171–173.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*. *46*(1), 243–256.

Ashour, O., Ghanem, A., and M., W. (2012). Stemming Effectiveness in Clustering of Arabic Documents. *International Journal of Computer Applications*. *49*(5), 1–6.

Awadalla, M. H., and Alajmi, A. F. (2011). Dewy index based Arabic Document classification with Synonyms Merge Feature Reduction. *International Journal of Computer Science Issues*. *8*(6), 46–54.

Awajan, A. (2015a). Semantic Similarity Based Approach For Reducing Arabic Texts Dimensionality. *International Journal of Speech Technology*. *19*(2), 191–201.

Awajan, A. (2015b). Semantic Vector Space Model For Reducing Arabic Text Dimensionality. In *Digital Information and Communication Technology and its Applications*. Beirut, Lebanon: IEEE, 129–135.

Ayadi, R., Maraoui, M., and Zrigui, M. (2014). Latent Topic Model for Indexing Arabic Documents. *International Journal of Information Retrieval Research*. *4*(1), 29–45.

Backialakshmi, P. (2015). Knowledge Discovery In Text Mining With Big Data Using Clustering Based Word Sequence. *International Journal Of Advanced Research In Datamining And Cloud Computing*. *3*(3), 35–52.

Barresi, S., Nefti, S., and Rezgui, Y. (2008). A Concept Based Indexing Approach for Document Clustering. In *The IEEE International Conference on Semantic Computing*. Washington, DC, USA: IEEE Computer Society, 26–33.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C. A., and Chen, H. (2013). Evaluating Text Visualization : An Experiment in Authorship Analysis. In *IEEE Intelligence and Security Informatics*. Seattle, Washington,USA, 16–20.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C. a, and Chen, H. (2014). Evaluating Text Visualization For Authorship Analysis. *Security Informatics*. *3*(10), 1–13.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., and Smeaton, A. F. (2009). Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. In *International Conference on Advances in Social Network Analysis and Mining*. Athens Greece: IEEE Computer

Society, 231–236.

Beseiso, M., Ahmad, A. R., and Ismail, R. (2010). A Survey of Arabic Language Support in Semantic Web. *International Journal of Computer Applications*. *9*(1), 24–28.

Beseiso, M., Ahmad, A. R., and Ismail, R. (2011). An Arabic Language Framework for Semantic Web. In *International Conference on Semantic Technology and Information Retrieval*. Putrajaya, Malaysia: IEEE, 7–11.

Bouchard, M., Joffres, K., and Frank, R. (2014). Preliminary Analytical Considerations In Designing A Terrorism And Extremism Online Network Extractor. In Mago, V. K., Dabbaghian, V. (Ed.). Computational Models of Complex Systems, Springer International Publishing, Switzerland. 171-184.

Brahmi, A., Ech-Cherif, A., and Benyettou, A. (2011). Arabic Texts Analysis for Topic Modeling Evaluation. *Information Retrieval*. *15*(1), 1–21.

Bsoul, Q. W., and Mohd, M. (2011). Effect of ISRI Stemming on Similarity Measure for Arabic Document Clustering. In *The Asia Information Retrieval Societies Conference*. Dubai, United Arab Emirates, 584–593.

Buntine, W. (2009). Estimating Likelihoods for Topic Models. In *Asian Conference on Machine Learning: Advances in Machine Learning*. Nanjing China: Springer-Verlag, 51–64.

Buntine, W., and Jakulin, A. (2004). Applying Discrete PCA in Data Analysis. In *Conference on Uncertainty in Artificial iIntelligence*. Banff, Canad: AUAI Press Arlington, 59–66.

Buntine, W., Perttu, S., and Tuulos, V. (2004). Using Discrete PCA on Web Pages. In *Proceedings of the workshop Statistical Approaches to Web Mining*. Pisa, Italy, 99–110.

Cha, S. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*. *1*(4), 300–307.

Chalothorn, T., and Ellman, J. (2013). Affect Analysis of Radical Contents on Web Forums Using SentiWordNet. *International Journal of Innovation, Management and Technology*. *4*(1), 122–124.

Chang, F., and Liu, C. (2012). *Ranking and Selecting Features Using an Adaptive Multiple Feature Subset Method.* Technical Report No. TR-IIS-12-005, Institute of Information Science, Academia Sinica.

Chantar, H. K., and Corne, D. W. (2011). Feature Subset Selection for Arabic Document Categorization using BPSO-KNN. In *The World Congress on Nature and Biologically Inspired Computing*. Salamanca, Spain: IEEE, 546–551.

Chawla, S., and Gionis, A. (2013). K-Means: A Unified Approach To Clustering And Outlier Detection. In *The International Conference on Data Mining*. Austin, Texas, USA: SIAM, 189–197.

Chen, C. H. (2015). Feature selection for clustering using instance-based learning by exploring the nearest and farthest neighbors. *Information Sciences*. *318*, 14–27.

Chen, H. (2006). *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining*. London: Springer.

Chen, H. (2007). Exploring Extremism and Terrorism on the Web: The Dark Web Project. In C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, and X. Cheng (Eds.), *Intelligence and Security Informatics*, 1–20. Springer Berlin Heidelberg.

Chen, H. (2008a). IEDs in the Dark Web: Genre Classification of Improvised Explosive Device Web Pages. In *IEEE International Conference on Intelligence and Security Informatics*. Taipei, Taiwan: IEEE, 94–97.

Chen, H. (2008b). Sentiment and Affect Analysis of Dark Web Forums: Measuring Radicalization on the Internet. In *International Conference on Intelligence and Security Informatics*. Taipei, Taiwan: IEEE, 104–109.

Chen, H. (2009). IEDs in the Dark Web: Lexicon Expansion and Genre Classification. In *Intelligence and Security Informatics Conference*. TX, USA: IEEE, 173–175.

Chen, H. (2011). From Terrorism Informatics to Dark Web Research. In U. K. Wiil (Ed.). *Counterterrorism and Open Source Intelligence* (pp. 317–341). Vienna: Springer.

Chen, H., Qin, J., Reid, E., and Zhou, Y. (2008). Studying Global Extremist Organizations' Internet Presence Using the Dark Web Attribute System. In Chen, H., Reid, E., Sinai, J., Silke, A., Ganor, B. (Ed.). *Terrorism Informatics* (pp. 237–266). US: Springer.

Chen, H., Thoms, S., and Fu, T. (2008). Cyber Extremism in Web 2.0: An Exploratory Study of International Jihadist Groups. In *International Conference on Intelligence and Security Informatics*. Taipei, Taiwan: IEEE, 98–103.

Cheng, W., Ni, X., Sun, J., and Jin, X. (2011). Measuring Opinion Relevance in Latent Topic Space. In *International Conference on Social Computing, Privacy,* *Security, Risk, and Trust*. Minneapolis, Minnesota, USA: IEEE Computer Society, 323–330.

Corbin, J. (2003). *Al-Qaeda: In Search of the Terror Network that Threatens the World*. Thunder Mouth Press/Nation Books.

Dai, P., Iurgel, U., and Rigoll, G. (2003). A Novel Feature Combination Approach for Spoken Document Classification with Support Vector Machines. In *Multimedia Information Retrieval Workshop*. Toronto, Canada, 1–5.

Deelers, S., and Auwatanamongkol, S. (2007). Enhancing K-means Algorithm with Initial Cluster Centers Derived from Data Partitioning Along the Data Axis with the Highest Variance. *International Journal of Electrical and Computer Engineering*. *2*(4), 247–252.

Demiriz, A., Bennett, K., and Embrechts, M. (1999). Semi-Supervised Clustering Using Genetic Algorithms. In *Artificial Neural Networks In Engineering*. St. Louis, Missouri: ASME Press, 809–814.

Deng, X. (2011). *Measuring Influence by Including Latent Semantic Analysis in Twitter Conversations*. University of Agder.

Deza, M. M., and Deza, E. (2016). Distances in Probability Theory. In *Encyclopedia of Distances*. Berlin, Heidelberg: Springer Berlin Heidelberg, 259–274.

Do, T., and Hui, S. (2006). Associative Feature Selection for Text Mining. *International Journal of Information Technology*. *12*(4), 59–68.

Dong, H., Hui, S. C., and He, Y. (2006). Structural Analysis of Chat Messages for Topic Detection. *Online Information Review*. *30*(5), 496–516.

Dumais, S., Letsche, T., Littman, M., and Landauer, T. (1997). Automatic Cross-Language Retrieval using Latent Semantic Indexing. In *Procceding of AAAI Spring Symposuim on Cross- Language Text and Speech Retrieval*. 115–132.

Elarnaoty, M., AbdelRahman, S., and Fahmy, A. (2012). A Machine Learning Approach for Opinion Holder Extraction in Arabic Language. *International Journal of Artificial Intelligence and Applications*. *3*(2), 45–63.

El-beltagy, S. R. (2006). KP-Miner: A Simple System for Effective Keyphrase Extraction. In *Innovations in Information Technology*. Dubai, UAE: IEEE Computer Society, 1–5.

Elovici, Y., Kandel, A., Last, M., Shapira, B., and Zaafrany, O. (2004). Using Data Mining Techniques for Detecting Terror-Related Activities on the Web. *Journal of Information Warfare*. *3*(1), 17–29.

Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., and Kandel, A. (2008). Content-Based Detection of Terrorists Browsing the Web Using an Advanced Terror Detection System (ATDS). In Kantor, P., Muresan, G., Roberts, F., Zeng, D.D.,Wang, Fei-Y., Chen, H., Merkle, R. C. (Ed.). *Terrorism Informatics* (pp. 365–384). US: Springer.

Erisoglu, M., Calis, N., and Sakallioglu, S. (2011). A New Algorithm for Initial Cluster Centers in K-means Algorithm. *Pattern Recognition Letters*, *32*(14), 1701–1705.

Fabbri, R., Costa, L. D. F., Torelli, J. C., and Bruno, O. M. (2008). 2D Euclidean distance transform algorithms. *ACM Computing Surveys*. *40*(1), 1–44.

Farahat, A. K., and Kamel, M. S. (2011). Statistical Semantics for Enhancing Document Clustering. *Knowledge and Information Systems*, *28*(2), 365–393.

Fejer, H. N., and Omar, N. (2015). Automatic Arabic text summarization using clustering and keyphrase extraction. *Journal of Artificial Intelligence*. *8*(1), 293–298.

Forsati, R., Mahdavi, M., Shamsfard, M., and Meybodi, M. R. (2013). Eficient Stochastic Algorithms for Document Clustering. *Information Science*. *220*, 269–291.

Froud, H., Benslimane, R., Lachkar, A., and Ouatik, S. A. (2010). Stemming and Similarity Measures for Arabic Documents Clustering. In *International Symposium on Communications and Mobile Network*. IEEE, 1–4.

Froud, H., Lachkar, A., and Ouatik, S. (2012). A Comparative Study Of Root-Based And Stem-Based Approaches For Measuring The Similarity Between Arabic Words For Arabic Text Mining Applications. *Advanced Computing An International Journal*. *3*(6), 55–67.

Froud, H., Lachkar, A., and Ouatik, S. (2013). Arabic Text Summarization based on Latent Semantic Analysis to Enhance Arabic Documents Clustering. *International Journal of Data Mining and Knowledge Management Process*. *3*(1), 79–95.

Froud, H., Sahmoudi, I., and Lachkar, A. (2013). An Efficient Approach To Improve Arabic Documents Clustering Based On A New Keyphrases Extraction Algorithm. In *Second International Conference on Advanced Information Technologies and Applications*. Dubai, UAE, 243–256.

Fu, T., Abbasi, A., and Chen, H. (2010). A Focused Crawler for Dark Web Forums.

*Journal of the American Society for Information Science*. *61*(6), 1213–1231.

Fu, T., and Chen, H. (2008). Discovery of Improvised Explosive Device Content in the Dark Web. In *International Conference onIntelligence and Security Informatics*. Taipei, Taiwan: IEEE, 88–93.

Gabrilovich, E. (2006). *Feature generation for textual information retrieval using world knowledge*. PhD Thesis, Israel Institute of Technology.

Ghanem, O. (2014). *Evaluating the Effect of Preprocessing in Arabic Documents Clustering*. Master Thesis, Islamic University, Gaza, Palestine.

Gharib, T. F., Fouad, M. M., Mashat, A., and Bidawi, I. (2012). Self Organizing Map_based Document Clustering Using WordNet Ontologies. *International Journal of Computer Science*. *9*(1), 88–95.

Ghwanmeh, S. (2005). Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language. *International Journal of Information Technology*. *3*(3), 467–471.

Gryc, W., and Moilanen, K. (2010). Leveraging Textual Sentiment Analysis with Social Network modelling: Sentiment Analysis of Political Blogs in the 2008 US Presidential Election. In *From Text to Political Positions Workshop*. Amsterdam: Vrije University, 47–70.

Gunal, S. (2012). Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering and Computer Sciences*. *20*(2), 1296–1311.

Guru, D. S., Harish, B. S., and Manjunath, S. (2010). Symbolic Representation of Text Documents. In *Third Annual ACM Bangalore Conference*. New York, NY: ACM, 1–4.

Habib, M. B., Fayed, Z. T., and Gharib, T. F. (2006). A Hybrid Feature Selection Approach for Arabic Documents Classification. *Egyptian Computer Science Journal*. *28*(3), 1–7.

Harrag, F., El-Qawasmah, E., and Al-Salman, A. M. S. (2010). Comparing Dimension Reduction Techniques for Arabic Text Classification Using BPNN Algorithm. In *Integrated Intelligent Computing (ICIIC)*, 6–11. Bangalore.

Hawwari, A., Zaghouani, W., O’Gorman, T., Badran, A., and Diab, M. (2013). Building a Lexical Semantic Resource for Arabic Morphological Patterns. In *International Conference on Communications, Signal Processing, and their Applications*. Sharjah, United Arab Emirates: IEEE, 1–6.

He, J., Weerkamp, W., Larson, M., and Rijke, M. (2009). An Effective Coherence

Measure to Determine Topical Consistency in User-Generated Content. *International Journal on Document Analysis and Recognition (IJDAR)*. *12*(3), 185–203.

Hoenkamp, E. (2011). Trading Spaces: On The Lore And Limitations Of Latent Semantic Analysis. In G. Amati and F. Crestani (Eds.). *Advances in Information Retrieval Theory*. (pp. 40–51). Berlin: Springer Berlin Heidelberg.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *International ACM SIGIR Conference On Research And Development In Information Retrieval*. Berkeley, CA, USA: ACM, 50–57.

Hotho, A., Staab, S., and Stumme, G. (2003). Wordnet improves Text Document Clustering. In *Semantic Web Workshop*. 541–544.

Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., and Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In *The International Conference on Research and Development in Information Retrieval*, 179–188. New York, USA: ACM Press.

Hu, X., Zhang, X., Lu, C., Park, E. K., and Zhou, X. (2009). Exploiting Wikipedia As External Knowledge For Document Clustering. In *International Conference On Knowledge Discovery And Data Mining*. Paris, France: ACM, 389–396.

Huang, A. (2008). Similarity Measures for Text Document Clustering. In *The New Zealand Computer Science Research Student Conference*. Christchurch, New Zealand, 49–56.

Is, A. C. E., and Woodward, W. A. (2006). Comparing One or Two Means Using the t-Test: With SPSS Examples. In *Statistical Analysis Quick Reference Guidebook*. SAGE Publications, 47–76.

Isa, D., Hong, L., Kallimani, V. P., and RajKumar, R. (2009). Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model. *Computer and Information Science*. *1*(4), 79–90.

Isa, D., Kallimani, V. P., and Lee, L. H. (2009). Using the self organizing map for clustering of text documents. *Expert Systems with Applications* *36*(5), 9584–9591.

Isa, D., Lee, L. H., Kallimani, V. P., and RajKumar, R. (2008). Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine. *Transactions on Knowledge and Data Engineering*. *20*(9), 1264–1272.

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Jain, A., and Murty, M. (1999). Data Clustering: A Review. *ACM Computing Surveys (CSUR)*. *31*(3), 255–323.

Javed, K., Babri, H. A., and Saeed, M. (2012). Feature Selection Based On Class-Dependent Densities For High-Dimensional Binary Data. *IEEE Transactions on Knowledge and Data Engineering*. *24*(3), 465–477.

Jing, L., Ng, M. K., and Huang, J. Z. (2010). Knowledge-based Vector Space Model for Text Clustering. *Knowledge and Information Systems*. *25*(1), 35–55.

Jing, L., Yun, J., Yu, J., and Huang, J. (2011). High-Order Co-clustering Text Data on Semantics-Based Representation Model. In *Advances in Knowledge Discovery and Data Mining*. Shenzhen, China, 171–182.

Kamde, P. M., and Algur, D. S. P. (2011). A Survey on Web Multimedia Mining. *The International Journal of Multimedia and Its Applications (IJMA)*. *3*(3), 72–84.

Kanaan, G., and Al-Shalabi, R. (2009). A Comparison of Text-Classification Techniques Applied to Arabic Text. *Journal of the American Society for Information Science and Technology*. *60*(6), 1836–1844.

Karima, A., Zakaria, E., and Yamina, T. G. (2012). Arabic Text Categorization: A Comparative Study Of Different Representation Modes. *Journal of Theoretical and Applied Information Technology*. *38*(1), 1–5.

Kaur, M., and Kaur, U. (2013). Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection. *International Journal of Advanced Research in Computer Science and Software Engineering*. *3*(7), 1454–1459.

Kchaou, Z., and Kanoun, S. (2008). Arabic Stemming with Two Dictionaries. In *International Conference on Innovations in Information Technology*. Al-Ain, UAE: IEEE, 688–691.

Khalifa, H. (2013). *New Techniques for Arabic Document Classication*. PhD Thesis, Heriot-Watt University.

Khoja, S. (1999). *Stemming Arabic Text.* PhD Thesis, Lancaster University, UK.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal*. *42*(1), 21–40.

L'Huillier, G., Rios, S. A., Alvarez, H., and Aguilera, F. (2010). Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web. *ACM*

*SIGKDD Explorations Newsletter*. *12*(2), 66–73.

Laender, A. H. F., Ribeiro-Neto, B. a., da Silva, A. S., and Teixeira, J. S. (2002). A Brief Survey of Web Data Extraction Tools. *ACM SIGMOD Record*. *31*(2), 84–93.

Landauer, T., Foltz, P., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*. *25*(2–3), 259–284.

Larkey, L., Ballesteros, L., and Connell, M. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *SIGIR'02*. Tampere,Finland: ACM, 275–282.

Larkey, L., Ballesteros, L., and Connell, M. (2007). Light Stemming for Arabic Information Retrieval. In A. Soud, A. van den Bosch, and G. Neumann (Eds.). *Arabic Computational Morphology* (pp. 221–243). Springer Netherlands.

Larkey, L., and Connell, M. E. (2001). Arabic Information Retrieval at UMass in TREC-10. In *Text Retrieval and Evaluation Conference*. Gaithersburg, Maryland: Department of Commerce, National Institute of Standards and Technology, 562–570.

Larkey, L. S., Feng, F., Connell, M., and Lavrenko, V. (2004). Language-specific Models in Multilingual Topic Tracking. In *Special Interest Group on Information Retrieval (SIGIR)*. Sheffield, UK.: ACM, 402–409.

Larson, C., and Chen, H. (2009). Dark Web Forums Portal: Searching and Analyzing Jihadist Forums. In *IEEE International Conference on Intelligence and Security Informatics*. Richardson, TX, USA: IEEE, 71–76.

Last, M., Markov, A., and Kandel, A. (2008). Multi-lingual Detection of Web Terrorist Content. In H. Chen and C. C. Yang (Eds.). *Intelligence and Security Informatics* (pp. 79–96). Springer Berlin Heidelberg.

Lee, L. H., Rajkumar, R., and Isa, D. (2012). Automatic Folder Allocation System using Bayesian-support Vector Machines Hybrid Classification Approach. *Applied Intelligence*. *36*(2), 295–307.

Leopold, E., and Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*. *46*(1–3), 423–444.

Lewis, D. D. (1990). Representation Quality in Text Classification: An Introduction and Experiment. In *Workshop on Speech and Natural Language*. Stroudsburg, PA, USA, 288–295.

Li, N., and Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*. *48*(2), 354–368.

Li, R. Z., and Zhang, Y. Sen. (2012). Study on the Method of Feature Selection Based on Hybrid Model for Text Classification. *Advanced Materials Research*. *433*, 2881–2886.

Li, Y., Luo, C., and Chung, S. (2008). Text clustering with feature selection by using statistical data. *IEEE Transactions on Knowledge and Data Engineering*. *20*(5), 641–652.

Lim, C., Eng, K. I., and Nugroho, A. S. (2010). Implementation of Intelligent Searching Using Self-Organizing Map for Webmining Used in Document Containing Information in Relation to Cyber Terrorism. In *International Conference on Advances in Computing, Control, and Telecommunication Technologies*. Jakarta, Indonesia: IEEE Computer Society, 195–197.

Liu, L., Kang, J., Yu, J., and Wang, Z. (2005). A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering. In *International Conference on Natural Language Processing and Knowledge Engineering*. Wuhan, China: IEEE, 597–601.

Liu, T., Liu, S., and Chen, Z. (2003). An Evaluation on Feature Selection for Text Clustering. In *International Conference on Machine Learning*. Washington, DC, USA: AAAI Press, 488–495.

Liu, Y., Wu, C., and Liu, M. (2011). Research of Fast SOM Clustering for Text Information. *Expert Systems with Applications*. *38*(8), 9325–9333.

Lliu, X., Ma, F., and Lin, H. (2011). Topic Detection with Hypergraph Partition Algorithm. *Journal of Software*. *6*(12), 2407–2415.

Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*. *14*, 178–203.

Lucia, A. De, Risi, M., Tortora, G., and Scanniello, G. (2007). Clustering Algorithms and Latent Semantic Indexing to Identify Similar Pages in Web Applications. In *Proceedings of the 9th IEEE International Workshop on Web Site Evolution*. Washington, USA: IEEE Computer Society, 65–72.

Luo, C., Li, Y., and Chung, S. M. (2009). Text document clustering based on neighbors. *Data and Knowledge Engineering*. *68*(11), 1271–1288.

Lv, X., Wang, T., Shi, S., and Li, S. (2010). The Key Technology of Topic Detection based on K-means. In *International Conference on Future Information Technology and Management Engineering*. Changzhou, China: IEEE, 387–390.

Machova, K., Szaboova, A., and Bednar, P. (2007). Generation of a Set of Key Terms Characterising Text Documents. *Journal of Information and Organizational Sciences*. *31*(1), 1–7.

Malik, S. K., and Rizvi, S. (2011). Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. In *International Conference on Computational Intelligence and Communication Networks*. Gwalior, India: IEEE Computer Society, 465–469.

Mesleh, A. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *Journal of Computer Science*. *3*(6), 430–435.

Mesleh, A. M. (2007). Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study. In *Int. Conf. On Applied Mathematics*. Cairo, Egypt: Springer, 11–16.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *The National Conference on Artificial Intelligence*. Boston, Massachusetts: American Association for Artificial Intelligence, 775–781.

Mohammad, S., Resnik, P., and Hirst, G. (2007). TOR , TORMD : Distributional Profiles of Concepts for Unsupervised Word Sense Disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Stroudsburg, PA, 326–333.

Monay, F., Quelhas, P., Gatica-Perez, D., and Odobez, J. M. (2006). Constructing Visual Models With A Latent Space Approach. In Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (Eds.). *Lecture Notes in Computer Science* (pp. 115–126). Springer Berlin Heidelberg.

Montañés, E., Quevedo, J., Combarro, E. F., Diaz, I., and Ranilla, J. (2007). A Hybrid Feature Selection Method for Text Categorization. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. *15*(2), 133–151.

Mousser, J. (2010). A Large Coverage Verb Taxonomy for Arabic. In *International Language Resources and Evaluation Conference*. Valetta, Malta: European Language Resources Association, 2675–2681.

Mousser, J. (2011). Classifying Arabic verbs using sibling classes. In *International*

*Workshop on Computational Semantics*. UK: Oxford , 355–359.

Napoleon, D., and Pavalakodi, S. (2011). A New Method for Dimensionality Reduction using K- Means Clustering Algorithm for High Dimensional Data Set. *International Journal of Computer Applications*. *13*(7), 41–46.

Nath, S. (2006). Crime Pattern Detection Using Data Mining. In *International Conference on Web Intelligence and Intelligent Agent Technology*. Hong Kong: IEEE Computer Society, 41–44.

Nguyen, C., Phan, X., and Horiguchi, S. (2009). Web Search Clustering and Labeling with Hidden Topics. *ACM Transactions on Asian Language Information Processing*. *8*(3), 1–40.

Nwesri, A. F. a., Tahaghoghi, S. M. M., and Scholer, F. (2007). Answering English Queries in Automatically Transcribed Arabic Speech. In *International Conference on Computer and Information Science*. Melbourne, Australia: IEEE Computer Society, 11–16.

Osinski, S. (2004). *Dimensionality Reduction Techniques For Search Results Clustering*. Master Thesis, The University of Sheffield.

Park, S., and Lee, S. R. (2012). Text Clustering using Semantic Terms. *International Journal of Hybrid Information Technology*. *5*(2), 135–140.

Patel, D., and Zaveri, M. (2011). A Review on Web Pages Clustering Techniques. In D. C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, and D. Nagamalai (Eds.), *Trends in Network and Communications*. Springer Berlin Heidelberg, 700–710.

Patil, G. A., Manwade, K. B., and Landge, P. S. (2012). A Novel Approach for Social Network Analysis and Web Mining for Counter Terrorism. *International Journal on Computer Science and Engineering (IJCSE)*. *4*(11), 1816–1825.

Patil, G. A., Manwade, K. B., and Landge, P. S. (2013). A Novel Approach for Recognized and Overcrowding of Terrorist Websites. *International Journal of Engineering Trends and Technology*. *4*(3), 463–470.

Paulsen, J. R., and Ramampiaro, H. (2009). Combining Latent Semantic Indexing and Clustering to Retrieve and Cluster Biomedical Information: A 2-step Approach. In *NIK-2009 conference*. Trondheim, 131–142.

Pedersen, T., and Kulkarni, A. (2005). Identifying Similar Words and Contexts in Natural Language with SenseClusters. In *The National Conference on Artificial InItelligence*. Pittsburgh, PA: American Association for Artificial Intelligence, 1694–1695.

Perkio, J., Buntine, W., and Perttu, S. (2004). Exploring Independent Trends in a Topic-based Search Engine. In *International Conference on Web Intelligence*. Beijing, China: IEEE Computer Society, 664–668.

Prasad, S., and Bruce, L. M. (2008). Limitations of Principal Components Analysis for Hyperspectral Target Recognition. *Geoscience and Remote Sensing Letters. 5*(4), 625–629.

Prentice, S., Taylor, P. J., Rayson, P., Hoskins, A., and O'Loughlin, B. (2010). Analyzing the Semantic Content and Persuasive Composition of Extremist Media: A Case Study of Texts Produced During the Gaza Conflict. *Information Systems Frontiers*. *13*(1), 61–73.

Qi, X., Christensen, K., Duval, R., Fuller, E., Spahiu, A., Wu, Q., and Zhang, C.-Q. (2010). A Hierarchical Algorithm for Clustering Extremist Web Pages. In *International Conference on Advances in Social Networks Analysis and Mining*. Odense, Denmark: IEEE Computer Society, 458–463.

Qin, J., Zhou, Y., and Chen, H. (2010). A Multi-region Empirical Study on the Internet Presence of Global Extremist Organizations. *Information Systems Frontiers*. *13*(1), 75–88.

Qin, J., Zhou, Y., Reid, E., Lai, G., and Chen, H. (2007). Analyzing Terror Campaigns on the Internet: Technical Sophistication, Content Richness, and Web Interactivity. *International Journal of Human-Computer Studies*. *65*(1), 71–84.

Qu, S., Wang, S., and Zou, Y. (2008). Improvement of Text Feature Selection Method Based on TFIDF. In *International Seminar on Future Information Technology and Management Engineering*. Leicestershire, United Kingdom: IEEE, 79–81.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In *International Conference on Machine Learning*. New Brunswick:NJ, USA: Rutgers University, 1–4.

Rana, S., Jasola, S., and Kumar, R. (2010). A Hybrid Sequential Approach for Data Clustering using K-Means and Particle Swarm Optimization Algorithm. *International Journal of Engineering, Science and Technology*. *2*(6), 167–176.

Ray, S., and Turi, R. H. (1999). Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. In *International Conference on Advances in Pattern Recognition and Digital Techniques*. Boca

Raton, FL, USA: IEEE Computer Society, 137–143.

Redmond, S. J., and Heneghan, C. (2007). A Method for Initialising the K-means Clustering Algorithm using kd-trees. *Pattern Recognition Letters*. *28*(8), 965—973.

Reid, E., Qin, J., Zhou, Y., Lai, G., Sageman, M., Weimann, G., and Chen, H. (2005). Collecting and Analyzing the Presence of Terrorists on the Web: A Case Study of Jihad Websites. In P. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F.-Y. Wang, H. Chen, and R. C. Merkle (Eds.). *Intelligence and Security Informatics* (pp.402–411). Springer Berlin Heidelberg.

Ren, W., and Han, K. (2014). Sentiment Detection of Web Users Using Probabilistic Latent Semantic Analysis. *Journal of Multimedia*. *9*(10), 1194–1200.

Rendón, E., and Abundez, I. (2011). Internal versus External cluster validation indexes. *International Journal of Computer and Communication*. *5*(1), 27–34.

Ricca, F., Pianta, E., Tonella, P., and Girardi, C. (2008). Improving Web Site Understanding with Keyword-based Clustering. *Journal of Software Maintenance and Evolution: Research and Practice*. *20*(1), 1–29.

Roberts, N. C. (2010). Tracking and Disrupting Dark Networks: Challenges of Data Collection and Analysis. *Information Systems Frontiers*. *13*(1), 5–19.

Rodner, E., and Denzler, J. (2009). Randomized probabilistic latent semantic analysis for scene recognition. In Bayro-Corrochano, E., Eklundh, J.(Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (pp. 945–953). Springer Berlin Heidelberg.

Rokach, L., and Maimon, O. (2005). Clustering Methods. In O. M. and L. Rokach (Ed.). *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). New York, NY: Spinger.

Saad, E., Awadalla, M., and Alajmi, A. (2010). Arabic Verb Pattern Extraction. In *International Conference on Information Science, Signal Processing and their Applications*. Kuala Lumpur, Malaysia: IEEE, 642–645.

Sabbah, T., and Selamat, A. (2014). Modified Frequency-Based Term Weighting Scheme for Accurate Dark Web Content Classification. In A. Jaafar, N. Mohamad Ali, S. A. Mohd Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, J. Nursuriati, T. M. T. Sembok (Eds.). *Information Retrieval Technology* (pp. 184–196). Springer International Publishing.

Sabbah, T., and Selamat, A. (2015). Hybridized feature set for accurate Arabic dark web pages classification. In H. Fujita and G. Guizzi (Eds.). *Intelligent Software Methodologies, Tools and Techniques* (pp. 175–189). Springer International Publishing.

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., and Fujita, H. (2016). Hybridized Term-weighting Method for Dark Web Classification. *Neurocomputing*, *173*(3), 1908–1926. JOUR.

Sahmoudi, I., Froud, H., and Lachkar, A. (2013). A New Keyphrases Extraction Method Based On Suffix Tree Data Structure. *International Journal of Database Management Systems ( IJDMS ). 5*(6), 17–33.

Sahmoudi, I., and Lachkar, A. (2016). Towards A Linguistic Patterns For Arabic Keyphrases Extraction. In *International Conference on Information Technology for Organizations Development* Fez, Morocco: IEEE, 1–6..

Said, D. A., Wanas, N. M., Darwish, N. M., and Hegazy, N. H. (2009). A Study of Text Preprocessing Tools for Arabic Text Categorization. In *International Conference on Arabic Language Resources and Tools*. Cairo, Egypt: The MEDAR Consortium, 330–336.

Saitta, S., Raphael, B., and Smith, I. F. C. (2007). A Bounded Index for Cluster Validity. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Leipzig, Germany: Springer, 174–187.

Saleh, L. M. B., and Al-Khalifa, H. (2009). AraTation: An Arabic Semantic Annotation Tool. In *International Conference on Information Integration and Web-based Applications and Services*. Kuala Lumpur, Malaysia: ACM, 447–451.

SAM, L. Z. (2009). *Enhanced Feature Selection Method for Illicit Web Content Filtering*. PhD Thesis, Universiti Teknologi Malaysia.

Scanlon, J. R. (2014). *Automatic Detection and Forecasting of Violent Extremist Cyber-Recruitment*. Master Thesis, University of Virginia.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47.

Selamat, A., Subroto, I. M. I., and NG, C.-C. (2009). Arabic Script Web Page Language Identification Using Hybrid-KNN Method. *International Journal of Computational Intelligence and Applications*. *8*(3), 315–343.

Seo, Y., Ankolekar, A., and Sycara, K. (2004). *Feature Selection for Extracting Semantically Rich Words*. Technial Report No. CMU-RI-TR-04-18, Carnegie

Mellon University.

Shaban, K. (2009). A Semantic Approach for Document Clustering. *Journal of Software*. *4*(5), 391–404.

Sharma, S., and Gupta, V. (2012). Recent Developments in Text Clustering Techniques. *International Journal of Computer Applications*. *37*(6), 14–19.

Simanjuntak, D. A., Ipung, H. P., Lim, C., and Nugroho, A. S. (2010). Text Classification Techniques Used to Faciliate Cyber Terrorism Investigation. In *International Conference on Advances in Computing, Control, and Telecommunication Technologies*. Jakarta, Indonesia: IEEE Computer Society, 198–200.

Smith, J. R., and Tesic, J. (2006). Semantic Labeling of Multimedia Content Clusters. In *International Conference on Multimedia and Expo*. Toronto,Canda: IEEE, 1493–1496.

Soni, G. (2013). *An automatic email mining approach using semantic non-parametric K-Means++ clustering*. Master Thesis, University of Windsor.

Sriurai, W. (2011). Improving Text Categorization By Using A Topic Model. *Advanced Computing: An International Journal (ACIJ)*. *2*(6), 21–27.

Steinbach, M., Karypis, G., Kumar, V., and others. (2000). A Comparison of Document Clustering Techniques. In *World Text Mining Conference*. Boston, 525–526.

Sun, J., Wang, X., and Yuan, C. (2011). Annotation-aware Web Clustering Based on Topic Model and Random Walks. In *International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE, 12–16.

Sutar, S. (2015). Feature Selection Algorithm Using Fast Clustering and Correlation Measure. *International Research Journal of Engineering and Technology (IRJET)*. *2*(7), 236–241.

Syiam, M., and Fayed, Z. (2006). An Intelligent System for Arabic Text Categorization. *International Journal of Intelligent Computing and Information Science*. *6*(1), 1–19.

Taghva, K., Elkhoury, R., and Coombs, J. (2005). Arabic Stemming without a Root Dictionary. In *International Conference on Information Technology: Coding and Computing*. Las Vegas, Nevada: IEEE Computer Society, 152–157.

Tan, F. (2007). *Improving Feature Selection Techniques for Machine Learning*. PhD Thesis, Georgia State University.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Cluster Analysis: Basic Concepts and Algorithms. In P.-N. Tan, M. Steinbach, and V. Kumar (Eds.). *Introduction to Data Mining* (pp. 487–568). Pearson.

TB, H., S, K., and NB, N. (2003). Documents Clustering Using Tolerance Rough Set Model and its Application to Information Retrieval. In Szczepaniak, P. S., Segovia, J., Kacprzyk, J., Zadeh, L. (Ed.) .*Studies In Fuzziness And Soft Computing:Intelligent exploration of the web* (pp. 181–196). Physica-Verlag HD.

Thabtah, F., Eljinini, M., Zamzeer, M., and Hadi, W. (2009). Naive Bayesian based on Chi Square to categorize Arabic data. *Communications of the IBIMA*. *10*, 4–6.

Thanh, N., and Yamada, K. (2011). Document Representation and Clustering with WordNet Based Similarity Rough Set Model. *International Journal of Computer Science*. *8*(5), 1–8.

Thomas, S. W. (2011). Mining software repositories using topic models. In *International Conference on Software engineering*. New York, USA: ACM Press, 11–38.

Tomar, G., Singh, M., Rai, S., Kumar, A., Sanyal, R., and Sanyal, S. (2013). Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation. *International Journal of Computer Science Issues*. *10*(5), 127–133.

Tsai, C.-F., and Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*. *50*(1), 258–269.

TU, N. C. A. M. (2008). *Hidden Topic Discovery Toward Classification And Clustering In Vietnamese Web Documents*. Master Thesis, Viet Nam National University.

Turney, P., and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*. *37*, 141–188.

Uysal, A. K., and Gunal, S. (2012). A Novel Probabilistic Feature Selection Method for Text Classification. *Knowledge-Based Systems*. *36*, 226–235.

Wallach, H. (2006). Topic Modeling: Beyond Bag-of-Words. In *International Conference on Machine Learning*. Pittsburgh, USA: ACM, 977–984.

Wang, S., Wei, Y., Li, D., Zhang, W., and Li, W. (2007). A Hybrid Method of

Feature Selection for Chinese Text Sentiment Classification. In *International Conference on Fuzzy Systems and Knowledge Discovery*. Haikou, China: IEEE Computer Society, 435–439.

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wilbur, W., and Sirotkin, K. (1992). The Automatic Identification of Stop Words. *Journal of Information Science*. *18*(1), 45–55.

Wilson, A. T., and Chew, P. A. (2010). Term Weighting Schemes for Latent Dirichlet Allocation. In *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*. Los Angeles, California: Association for Computational Linguistics, 465–473.

Wu, S., and Chow, T. W. S. (2004). Clustering of The Self-Organizing Map Using A Clustering Validity Index Based on Inter-Cluster and Intra-Cluster Density. *Pattern Recognition*. *37*(2), 175–188.

Xu, C., Zhang, Y., Zhu, G., Rui, Y., Lu, H., and Huang, Q. (2008). Using Webcast Text for Semantic Event Detection in Broadcast Sports Video. *IIEEE Transactions on Multimedia*. *10*(7), 1342–1355.

Xu, J., and Chen, H. (2008). The Topology of Dark Networks. *Communications of the ACM*. *51*(10), 58–65.

Xu, J., Chen, H., Zhou, Y., and Qin, J. (2006). On the Topology of the Dark Web of Terrorist Groups. In *The International Conference on Intelligence and Security Informatics*. San Diego, USA: Springer Berlin Heidelberg, 367–376.

Xu, Y., and Chen, L. (2010). Term-Frequency Based Feature Selection Methods for Text Categorization. In *International Conference on Genetic and Evolutionary Computing*. Shenzhen, China: IEEE Computer Society, 280–283.

Yang, C., Tang, X., and Gong, X. (2011). Identifying Dark Web Clusters with Temporal Coherence Analysis. In *International Conference on Intelligence and Security Informatics*. Beijing, China: IEEE, 167–172.

Yang, L., Liu, F., Kizza, J. M., and Ege, R. K. (2009). Discovering Topics from Dark Websites. In *IEEE Symposium on Computational Intelligence in Cyber Security*. IEEE, 175–179.

Yang, Y. (1995). Noise Reduction in A Statistical Approach to Text Categorization. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, USA: ACM, 256–263.

Yang, Y., and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 412–420.

Yong-qing, W., Pei-yu, L., and Zhu, Z. (2008). A Feature Selection Method based on Improved TFIDF. In *International Conference on Pervasive Computing and Applications*. Alexandria, Egypt: IEEE, 94–97.

Yu, L., and Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*. *5*, 1205–1224.

Zahran, B. M., and Kanaan, G. (2009). Text Feature Selection using Particle Swarm Optimization Algorithm. *World Applied Sciences Journal Special Issue of Computer and IT*. *7*, 69–74.

Zhang, D., Jing, X., and Yang, J. (2006). *Biometric Image Discrimination Technologies*. Idea Group Inc.

Zhang, D., and Li, S. (2011). Topic Detection based on K-means. In *International Conference on Electronics, Communications and Control*. Mudanjiang, China: IEEE, 2983–2985.

Zhang, W., Yoshida, T., and Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*. *21*(8), 879–886.

Zhang, W., Yoshida, T., Tang, X., and Wang, Q. (2010). Text Clustering Using Frequent Itemsets. *Knowledge-Based Systems*. *23*(5), 379–388.

Zhang, Y., Zeng, S., Huang, C. N., Fan, L., Yu, X., Dang, Y., Larson, C., Denning, D., Roberts, N., Chen, H. (2010). Developing a Dark Web collection and infrastructure for computational and social sciences. In *International Conference on Intelligence and Security Informatics*. Vancouver, Canada: IEEE, 59–64.

Zhang, Y., and Zhang, Q. (2006). A Text Classifier Based on Sentence Category VSM. In *The Pacific Asia Conference on Language, Information and Computation*. Wuhan, China, 244–249.

Zhou, X. F., Liang, J. G., Hu, Y., and Guo, L. (2014). Text Document Latent Subspace Clustering by PLSA Factors. In *International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*. Warsaw, Poland: IEEE, 442–448.

Zhou, Y., Qin, J., Lai, G., and Chen, H. (2007). Collection of U.S. Extremist Online Forums: A Web Mining Approach. In *The Hawaii International Conference on System Sciences*. Waikoloa, USA: IEEE, 1–10.

Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. (2005). US Domestic Extremist Groups on the Web: Link and Content Analysis. *IEEE Intelligent Systems*. 44–51.

Zhou, Y., Yang, Y., Peng, W., and Ping, Y. (2010). A Novel Term Weighting Scheme With Distributional Coefficient For Text Categorization With Support Vector Machine. In *The Conference on Information Computing and Telecommunications*. Beijing, China: IEEE, 182–185.

Zhu, X. (2007). *Advanced NLP : Latent Topic Models Probabilistic Latent Semantic Analysis ( pLSA )*. Lecturer Notes, University of Wisconsin–Madison.

Zitouni, A., Damankesh, A., Barakati, F., Atari, M., Watfa, M., and Oroumchian, F. (2010). Corpus-Based Arabic Stemming Using N-Grams. In *The Sixth Asia Information Retrieval Societies Conference*. Taipei, Taiwan: Springer Berlin Heidelberg, 280–289.