Journal of King Saud University – Computer and Information Sciences (2014) **26**, 388–397

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Arabic web pages clustering and annotation using semantic class features

Hanan M. Alghamdi [a,b,*], **Ali Selamat** [b,c], **Nor Shahriza Abdul Karim** [d]

[a] *Faculty of Computer Science, Umm Al-Qura University, Al-Gunfdh, Saudi Arabia*
[b] *Faculty of Computing, Universiti Teknologi Malaysia, UTM, Johor Bahru, Johor 81310, Malaysia*
[c] *UTM-IRDA Digital Media Center of Excellence, Universiti Teknologi Malaysia, UTM, Johor Bahru, Johor 81310, Malaysia*
[d] *Computer & Information Science Department, Prince Sultan University, 66833 Rafha Street, Riyadh 11586, Saudi Arabia*

Available online 28 September 2014

KEYWORDS

k-Means;
Semantic similarity;
Text clustering;
Arabic webpage

Abstract To effectively manage the great amount of data on Arabic web pages and to enable the classification of relevant information are very important research problems. Studies on sentiment text mining have been very limited in the Arabic language because they need to involve deep semantic processing. Therefore, in this paper, we aim to retrieve machine-understandable data with the help of a Web content mining technique to detect covert knowledge within these data. We propose an approach to achieve clustering with semantic similarities. This approach comprises integrating *k*-means document clustering with semantic feature extraction and document vectorization to group Arabic web pages according to semantic similarities and then show the semantic annotation. The document vectorization helps to transform text documents into a semantic class probability distribution or semantic class density. To reach semantic similarities, the approach extracts the semantic class features and integrates them into the similarity weighting schema. The quality of the clustering result has evaluated the use of the purity and the mean intra-cluster distance (MICD) evaluation measures. We have evaluated the proposed approach on a set of common Arabic news web pages. We have acquired favorable clustering results that are effective in minimizing the MICD, expanding the purity and lowering the runtime.

1. Introduction

The growth of Arabic web pages and the great amount of text contained in them, which hold unorganized informative data, urge the necessity to adopt solutions that can wisely manage these textual data (Elarnaoty et al., 2012). Because of the unstructured character of these texts, valuable knowledge cannot be efficiently understood by machines.

Many studies have been conducted to classify related information and to support the manipulation of texts available on the Internet. Document clustering is the most common

* Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, UTM, Johor Bahru, Johor 81310, Malaysia.
E-mail addresses: hanani.alghamdi@gmail.com (H.M. Alghamdi), aselamat@utm.my (A. Selamat), nshahriza@pscw.psu.edu.sa (N.S. Abdul Karim).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.002

technique utilized in categorizing web pages that contain related information into one group (Froud et al., 2013). This technique speeds up the process of allocating documents with similar information.

In addition, the production of semantic metadata identified by textual content appears to be a way to reveal the hidden knowledge (Faria et al., 2013). Extracting semantic features help to capture more understanding about the given documents based on the semantic similarities between them (Chang and Lee, 2011).

The semantic annotation is defined as the procedure of indexing and retrieving valuable information from documents and creating annotation on top of the documents' contents. The aim of this process is to provide data that can be understood by humans and machines.

In this paper, we aim to retrieve machine-understandable data with the help of a web content mining technique to detect covert knowledge within these data. In addition, we try to find semantic similarities between the web pages and cluster them based on the similarities. We extract the semantic annotation with the assistance of Arabic VerbNet[1](Mousser, 2010) as a way to produce a picture about the knowledge contained as suggested by Malik and Rizvi (2011). By using this technique, we will be able to annotate the resulting clusters based on the semantic features found in their contents. This paper is organized as follows: in Section 2, we explain the related works. Section 3 discusses the proposed model. Section 4 explains the study set-up. Section 5 presents the experimental results. Finally, Section 6 gives the results' discussion and conclusions.

2. Related works

The current web page analysis techniques differ according to the classification levels used (sentence, phrase, or document level) or the types of features considered for the techniques used. According to Abbasi et al. (2008), the types of features observed are (1) syntactic, which concerns with the structure of the word where the semantic orientation of words is considered and (2) stylistic, which focuses on the word style (Abbasi et al., 2008).

A study in sentiment text mining has been very much confined to the Arabic language (Farra et al., 2010). Analysis of the Arabic text is challenging because of the morphological characteristics of the Arabic words and sentences (Al-Khalifa and Al-Wabil, 2007; Beseiso et al., 2011). Developing a machine-understandable system for the Arabic language involves discriminated and deep semantic processing.

Farra et al. suggest that the Arabic text sentiment mining approach is on two sides, i.e., the sentence level and document level. In their study, they used the identified polarities of the sentences to classify the general polarity of the document (Farra et al., 2010). Abbasi et al. used the syntactic and stylistic features together to categorize the opinions in multilingual (English and Arabic) web forums (Abbasi et al., 2008). However, the semantic features are not considered in the classifying process. Froud et al. (2010) investigated the impacts of stemming on the Arabic text document clustering (Froud et al., 2010). The study concludes that the representation of the documents and the preprocessing can make the documents smaller and the clustering faster.

Other studies have focused on approaches to classify documents according to semantic similarities but with languages other than Arabic. An approach for clustering documents according to semantic information by determining the similarities between documents is proposed in Shaban (2009). The approach is composed of the semantic components to provide an accurate similarity measure between documents. Thus, the approach can be used to solve document clustering problems. Eventually, the approach produces effective document clustering that is able to recognize the meaning and structures of text in documents.

Semantic annotation can be used as a guide to understand and classify the document and reveal informative knowledge. In Nguyen et al. (2009) and Park and Lee (2012), the authors proposed a framework for clustering and labeling with the hidden topics of web documents. By revealing the hidden topics and preparing them for annotating clusters, more meaningful clusters can be produced, and the quality of clustering can be improved.

To the best of our knowledge, classifying documents according to semantic similarities for retrieving information from Arabic web pages is limited. In this research, we intend to extract the semantic features from Arabic web pages and cluster these pages according to the similarities of these features. We consider that a word that carries very strong semantic information can disclose hidden knowledge.

In the proposed method, we did not use any machine translation tools that may cause the loss of meaning or some semantic distortions that result from the wrong choice of words and language models (Larkey et al., 2004). Instead, we used available lexical resources for Arabic text to process Arabic language, such as Arabic VerbNet. The tool offers systematic investigation of the semantic/syntactic aspects of the morphological system. According to Hawwari et al. (2013), Arabic VerbNet is one of the lexical resources for Arabic verbs that provides large coverage for Arabic verb taxonomy with semantic aspects of the morphological system. The work of Mousser (2010), which is based on an English VerbNet project (Kipper et al., 2008), is a representation of Levin's syntactic alterations into Arabic. In this research, we used Arabic VerbNet to find the semantic similarities between web pages. This resource gives essential information about the syntax and semantics of Arabic verbs by applying the concept of verb-classes. The current version of the work by Mousser (2010) has 202 classes populating 4707 verbs and 834 frames. These frames consider alternations where the verbs can appear. Every class is a hierarchical structure, providing syntactic and semantic information about verbs and pre-allocating them to subclasses.

3. Proposed model

The proposed model, as shown in Fig. 1, performs clustering with the semantic similarities of Arabic Web pages and produces document vectorization according to semantic features (density or the probability distribution) with the help of Arabic VerbNet lexical, and then it finds the semantic annotation of the resulting clusters.

It contains two main phases: (1) extracting semantic features and document vectorization to group Arabic Web pages

[1] VerbNet is available for download http://ling.uni-konstanz.de/pages/home/mousser/files/Arabic_verbnet.php.

Figure 1 Proposed model.

according to the semantic similarities and (2) showing the semantic annotation. There are three steps in this method, which are: (1) extraction of semantic class features, (2) document vectorization, and (3) clustering and annotation.

This model works as follows: the model takes unannotated documents (to be classified), it will recognize all opinion words by using semantic feature extraction, and then it will aggregate all of the words to give a semantic annotation to the document.

3.1. Extraction of semantic class feature

The task of feature extraction is to find a semantic correspondence between one or more semantic classes and a document. For example, consider the document contents and semantic class repository as shown in Fig. 2. This task must extract the features of semantic class found in the document in terms of the attribute/value. Therefore, it combines and refines the document's terms and maps them to the target semantic class.

For each text, the verbs are extracted and grouped together based on the semantic class. The process follows the rule-based POS tagging constructed in Al-Shalabi and Kanaan (2004). Each word entry is tagged as a noun, a verb or stopwords. In this study, verbs will only be used to assign the semantic class.

The verb can be related to its semantic class using the lexicon-based task (Ahmed, 2009) in Arabic VerbNet (Mousser, 2010). In this task, each semantic class has a set of verb lexicon associated with it. This task is simply a search through a lexicon list. If the verb is found, an appropriate class is assigned to the extracted word or term in the document. By the end of this task, it will produce the features in vector space using the word and semantic class as the vectors. Consequently, each of the extracted documents may have more than one semantic class.

3.1.1. Using Arabic VerbNet

To generate semantic features from input verbs, an analyzer was implemented as part of our model to generate these features automatically. Arabic VerbNet frames contain the description of the syntactic and semantic information of each verb. The semantic meaning of the verb, such as the cause, emotional state, motion, and made of, are connected with each frame (Mousser, 2011).

The analyzer will extract the verbs found on Arabic VerbNet and the semantic frames related to these verbs and build a local frame database to be used with our model. The semantic information found on extracted frames will be used as a semantic class in our model. Therefore, the semantic classes will be related to verbs according to the local frames extracted from Arabic VerbNet. The semantic class of the verb can be defined as a semantic feature of this verb.

For each verb resulting from POS tagging, the analyzer will look up the verb in the local frame database and relate this verb to the appropriate semantic class. Because of polysemy, each verb can be related to more than one semantic class.

3.2. Document vectorization

The document vectorization task is shown in Fig. 3 that represents a document using the semantic class feature extracted from the previous step. The term "semantic class" means the classes of verbs as obtained from the VerbNet. The task carries out document vectorization, which converts each document text into vectors that characterize the contained semantic class features through the exploitation of the semantic class density or the probability distribution.

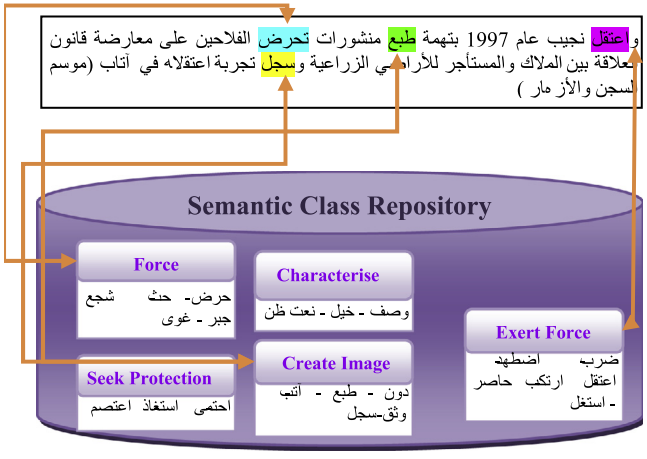

Figure 2 Process of extracting features of semantic classes.

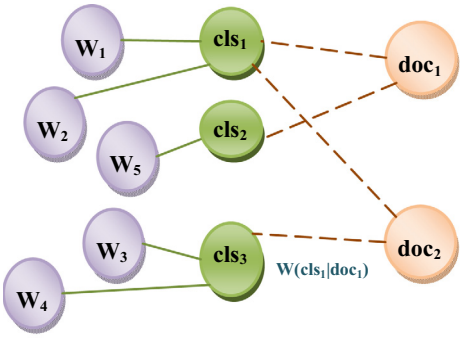

Figure 3 Document vectorization.

This task uses the semantic class features to provide a series of probabilities to which the document can be assigned according to a pre-specified set of semantic classes that are based on semantic class feature extraction. The utilization of vectorization in our method helps to transform text documents into a semantic class probability distribution or semantic class density in the vector space. As a result, these vectorizations can be used to calculate the semantic similarity between web pages.

3.2.1. Semantic class density

The distribution of the semantic class within a webpage may provide extra implicit knowledge. Semantic class density assumes that a semantic class's density in sampled documents is a good approximation of its density in the complete database. The relevant documents have approximately the same density as semantic classes.

The semantic class density is the average frequency of one class over the set of documents. In this step, the weight of the semantic class cls_i in the document doc_d is calculated as a formula for merging the class frequency of each semantic class and the total frequencies of all of the semantic classes to obtain the total weight of class (cls_i) and is as follows:

$$W(cls_i|doc_d) = \frac{\mathcal{O}(cls_i)}{\sum_{t \epsilon doc_d} \mathcal{O}(cls_t)} \quad (1)$$

where $\mathcal{O}(cls_i)$ is the total occurrence of class (cls_i) in document (doc_d) and it is calculated as a sum of frequencies over all words as follows:

$$\mathcal{O}(cls_i) = \sum F(w_e) \Rightarrow w_e \in cls_i \text{ and } w_e \in doc_d \quad (2)$$

If the density value of a semantic class cls_1 in a document doc is computed as, $W(cls_1|doc_d)$ then each document doc_d has x weights of semantic classes, which will represent the (document vectorization) $\mathrm{DV}(doc_d) = W(cls_1|doc_d)$, $W(cls_2|doc_d)$, $W(cls_3|doc_d)$,...$W(cls_x|doc_d)$.

3.2.2. Semantic class probability distribution

Every document can be represented by its probability distribution on the semantic classes as a feature vector space. The Bayes formula can be employed to calculate the probabilities to which the document can be assigned according to a predefined set of classes.

We calculate the probability of the documents for each semantic class (cls_1, cls_2, cls_3,...cls_x) using Eq. (3) (Isa et al., 2008; Mohammad et al., 2007). If the probability value for a document doc_d to be assigned to a semantic class cls_1 is computed as $P(cls_1|doc_d)$, then each document doc_d has x probability distributions $P(cls_1|doc_d)$, $P(cls_2|doc_d)$, $P(cls_3|doc_d)$,...$P(cls_x|doc_d)$ as shown in Table 1. The detailed description of the Bayesian vectorization is given in Isa et al. (2008) as follows:

$$P(cls_i|w_t) = \frac{P(w_t|cls_i) * P(cls_i)}{P(w_t)} \quad (3)$$

where w_t is the word extracted from document doc_d, z refers to the total number of words in doc_d, cls_i refers to the semantic class number i and x is the total number of available semantic classes.

3.3. Clustering and annotation

The document vectorization is designed as an input to the k-means clustering for classification purposes. In this model, the document vectorization task is employed as a text representation model, where the document vectorization represents a text as understandable machine vectors and aids in the creation analysis and classification system. The outcomes of the clustering step are clusters with semantic informative data that can be used to add semantic annotation to the resulting clusters.

Table 1 Semantic class probability distribution calculation.

	$P(cls_1\|doc_d) = \frac{\sum_{t=1}^{n} P(cls_1\|w_t)}{x}$	$P(cls_i\|doc_d) = \frac{\sum_{t=1}^{n} P(cls_i\|w_t)}{x}$	$\mathrm{doc_d}$ vectors
doc_1	$\frac{P(cls_1\|w_5) + P(cls_1\|w_2) + P(cls_1\|w_1)}{3}$	$\frac{P(cls_i\|w_3) + P(cls_i\|w_2) + \cdots\cdots}{z}$	$P(cls_1\|doc_1), ...P(cls_i\|doc_1)$
doc_2	$\frac{P(cls_1\|w_3) + P(cls_1\|w_1) + P(cls_1\|w_2) + P(cls_1\|w_5)}{4}$	$\frac{P(cls_i\|w_3) + P(cls_i\|w_1) + \cdots\cdots}{z}$	$P(cls_1\|doc_2), ...P(cls_i\|doc_2)$
doc_d	$\frac{P(cls_1\|w_1) + P(cls_1\|w_2) + P(cls_1\|w_4)}{3}$	$\frac{P(cls_i\|w_1) + P(cls_i\|w_2) + \cdots\cdots}{z}$	$P(cls_1\|doc_d), ...P(cls_x\|doc_d)$

3.3.1. Clustering

In the clustering step, the squared Euclidean distance (Cha, 2007; Deza and Deza, 2006) is used to present the degree of closeness or separation of the target document to the chosen cluster. Clustering with the squared Euclidean distance metric is faster than clustering with the regular Euclidean distance (Fabbri et al., 2008). The squared Euclidean distance between two distributions $P(d_i)$ and $P(d_r)$ is calculated in Eq. (4) as follows:

$$Sim_{sqe} = \sum_{i,r=1}^{d} (P(d_i) - P(d_r))^2 \quad (4)$$

Afterward, the k-means will cluster the vectorized doc_d to the suitable cluster that contains a similar attribute based on the probability distribution or density of the semantic classes. The documents found in every cluster are amassed depending on the likeness of the semantic classes recognized in each of them.

3.3.2. Semantic annotation

Semantic annotation can be defined as the process of inserting semantic tags in a document that allows the documents to be processed either by humans or by using automated software agents. The semantic annotation for Arabic web page is assigned based on the highest semantic feature relevance scores as a way to produce a picture about the knowledge contained and its semantics in the domain (Malik and Rizvi, 2011). The extracted feature from the first task helps specify the most relevant semantic annotation.

To annotate the related clusters, the appropriateness scores for every semantic feature discovered in the cluster are figured. Then, the annotation of the cluster is the mixture of the five topics with the highest scores. Hence, clusters with high scores demonstrate that the document is more related to the semantic meaning of this topic (Deng, 2011). To measure the relevance score between cluster k and semantic topic Stj, two methods based on the mean score among cluster k and mean ratio among the corpus (Smith and Tesic, 2006) are used. These methods are explained as follows:

- Mean score among cluster k: This method calculates the average of the semantic topic weight to determine the relation of this topic to the cluster k. As a result, those semantic topics with a weight stronger than the mean of cluster k are the most important topics of this cluster sorted by weight from the most important to the least important. Let the mean value of semantic topic j belong to cluster k as in Eq. (5), where μSt is the mean value of semantic topic j among the corpus.

$$SL_j^k = \mu St_j^k \quad (5)$$

- Mean ratio among corpus (MRAC): This method is based on measuring the significance score of the semantic topic to all other documents found in the corpus (Smith and Tesic, 2006). Let the main ratio of the semantic topic j belong to cluster k as in Eq. (6), where μSt_i is the mean value of semantic topic j among the corpus.

$$SL_j^k = \frac{\mu St_j^k}{\mu St_j} \quad (6)$$

Table 2 Arabic dataset.

Category name	Number of documents
Political news	194
Economic news	133
Sports news	126
Social news	60
Cultural news	101
Technology & Science news	139
Total	753

4. Experimental set-up and evaluation

The goal of this assessment is to figure out the effect of the recommended model, which aims to find the semantic similarities between web pages and extract the semantic annotation on the cluster quality and performance. The default number of clusters is set as the same number of pre-assigned categories in the dataset and its multiples.

When we execute the tests, we need to prepare the collected web pages for the classification algorithms. The pre-processing stage is intended to gather and extract related web pages and to diminish the noise terms (undesired term inside the content) to simplify the technique of measuring the weighting of features. This phase consists of collecting the URL seeds of web sites as a dataset using a web extractor agent and a text pre-processing stage that incorporates tokenization and normalization, tagging, and stemming. The pre-processing phase is depicted clearly in Alghamdi and Selamat (2012). The remainder of this section clarifies the evaluation criteria and the test results.

4.1. Datasets

In this study, we have collected corpus from the archives of online Arabic newspapers because there are no common Arabic datasets available by which to test the proposed model. These newspapers are namely Al-Akhbar[2], Alhayat[3], Aldostor[4], Gomhuria online[5], Akhbar Alarab.Net[6], Alriyadh[7] and Saudi Times[8]. These online newspapers are commonly used for many applications related to Arabic text language (Alsaleem, 2011, 2013; Karima et al., 2012; Saleh and Al-Khalifa, 2009). The collected datasets contain 753 documents with a dissimilar length of words. There are six categories to which the documents belong as explained in Table 2. We employed a web extractor agent (Easy Web Extract version 2.7[9]) to extract the textual data from these web pages.

[2] Al-Akhbar online news available at http://www.al-akhbar.com/.
[3] Alhayat online news available at http://alhayat.com/.
[4] Aldostor online news available at http://dostor.org/.
[5] Gomhuria online available at http://www.gomhuriaonline.com/.
[6] Akhbar Alarab.Net online news available at http://akhbaralarab.net/.
[7] Alriyadh online newspaper available at http://www.alriyadh.com/section.home.html.
[8] Saudi Times online newspaper available at http://www.saudi-times.net/Default.aspx.
[9] Easy Web Extract webpage (http://webextract.net/).

4.2. Evaluation criteria

The quality of the clustering result using the above datasets is evaluated using three evaluation measures, namely, purity measure, mean intra-cluster distance (MICD) and Davies–Bouldin index (DBI). These measures are widely used to evaluate the performance of unsupervised classification algorithms (Chawla and Gionis, 2013; Forsati et al., 2013; Huang, 2008; Rana et al., 2013). These evaluation measures are computed as follows:

- Purity measure: this measure is used to estimate the coherence of a resulting cluster. Our approach evaluates the degree to which a cluster encloses documents from a particular category. The purity of a single cluster C_i of size e_i is formally defined in Eq. (7):

$$\text{Purity}(C_i) = \frac{1}{e_i}\max_h e_i^h \tag{7}$$

  where $\max_h e_i^h$ represents the main category in cluster C_i and e_i^h that corresponds to the number of documents in cluster C_i that annotates to category h. In an optimal cluster that consists of group documents from a single category, its purity rate is one (Huang, 2008).
- Mean intra-cluster distance (MICD): This measure is the distance between data vectors and its cluster center where the low MICD signifies a compact cluster and the one with a high MICD signifies a loose cluster (Rana et al., 2013). The MICD is calculated in Eq.(8):

$$\text{MICD} = \sum_{c_i \epsilon C_K} \frac{\|c_i - \mu_K\|}{N_K} \tag{8}$$

- Davies–Bouldin index (DBI): This measure aims to find well-separated, compact clusters. It takes into account within cluster vectors the variance and distance between clusters centers (Demiriz et al., 1999). The smaller value of DBI shows a better clustering result. It has been found to be among the best indices (Arbelaitz et al., 2013; Rendón and Abundez, 2011). The DBI is calculated in Eq.(9):

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{j=1\cdots k, j\neq i}\left\{\frac{diam(c_i) + diam(c_j)}{\|\mu_i - \mu_j\|^2}\right\} \tag{9}$$

  where the diameter of a cluster is defined as:

$$diam(c_i) = \frac{1}{N_i}\sum_{x \epsilon c_i} \|c_i - \mu_i\|^2 \tag{10}$$

  where $diam(c_i)$ and $diam(c_j)$ are the average distances of all data vectors in clusters i and j to their respective cluster centroids in Eq. (10), μ_i is the center of cluster c_i consisting of N_i points and $\|\mu_i - \mu_j\|^2$ is the Euclidean distance between these centroids.

A strong structure and good clustering have a small MICD (similar data vectors are grouped together), smaller BDI (well separated compact clusters) and high purity.

5. Results

The results of applying the proposed model are gathered from two experiments. The first experiment is performed to show the fitness of the proposed solutions compared to the standard k-means. The second experiment is performed to illustrate the output of the semantic annotation process, where we can label the resulting clusters.

Fig. 4 demonstrates the results of the proposed approaches using a purity evaluation. Document vectorization (semantic class probability distribution or semantic density) with k-means is adequate for creating more coherent clusters that are well divided in relation to the categories. Document vectorization with k-means signifies that the clusters have high purity scores.

Fig. 5 shows the comparison results based on an MICD evaluation. The low value of MICD means that all points in the cluster are close to each other. The resulting clusters using vectorization (semantic class probability distribution or semantic density) with k-means appear to be more compact. The proposed model tends to outperform the standard k-means.

The comparison results based on the DBI evaluation are shown in Fig. 6. The smaller value of DBI signifies that there

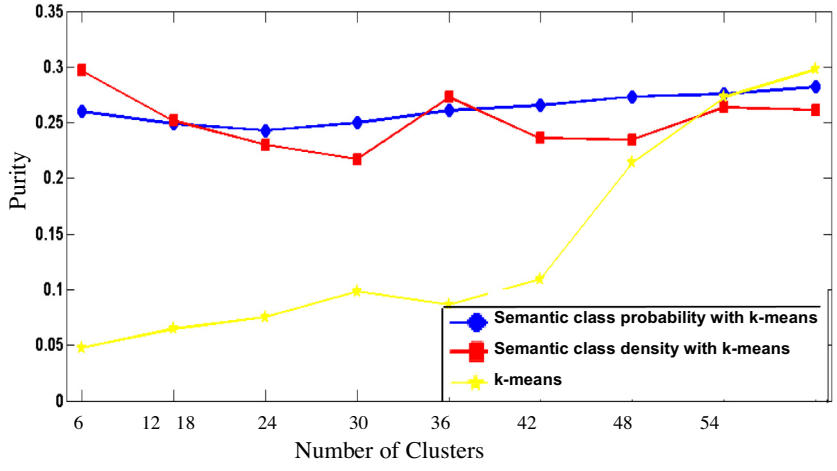

Figure 4 Purity results.

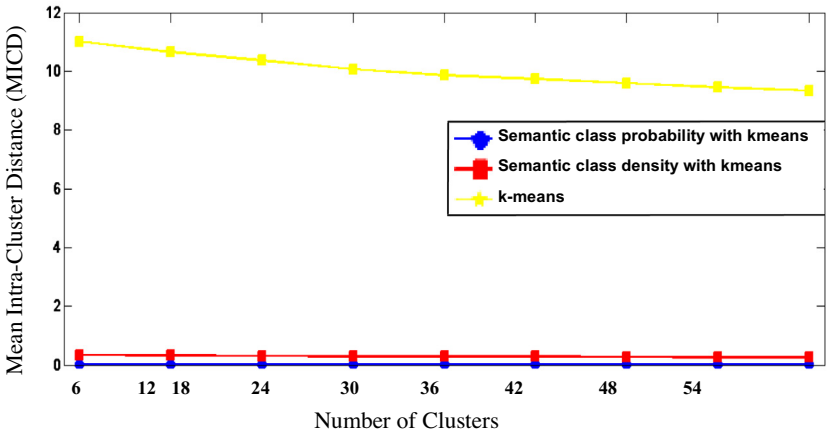

Figure 5 Mean intra-cluster distance (MICD) results.

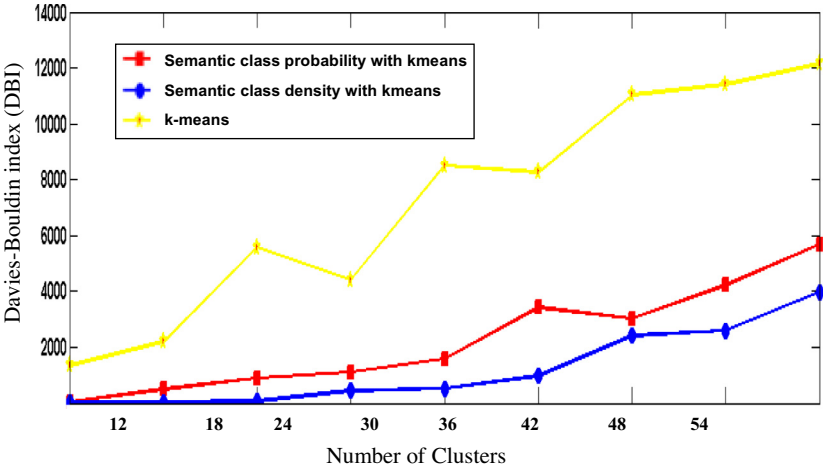

Figure 6 Davies-Bouldin index (DBI) results.

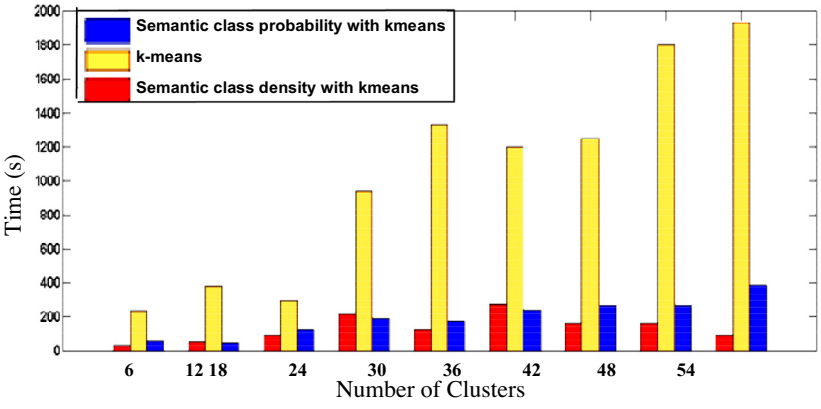

Figure 7 Time Consumed.

is a good separation distance between clusters and that the distances between points in the cluster and its center are small. The resulting clusters using vectorization for either semantic class probability distribution or semantic density with *k*-means appear to have a smaller value of DBI, which means the proposed approach outperforms the standard *k*-means.

The consumed time for comparing approaches is shown in Fig. 7, where the time elapsed is measured in seconds. We can see that the runtime rises gradually when the number of clusters increases. The runtime of the standard *k*-means is substantially longer than the time consumed by the other two solutions. In contrast, the time elapsed using document

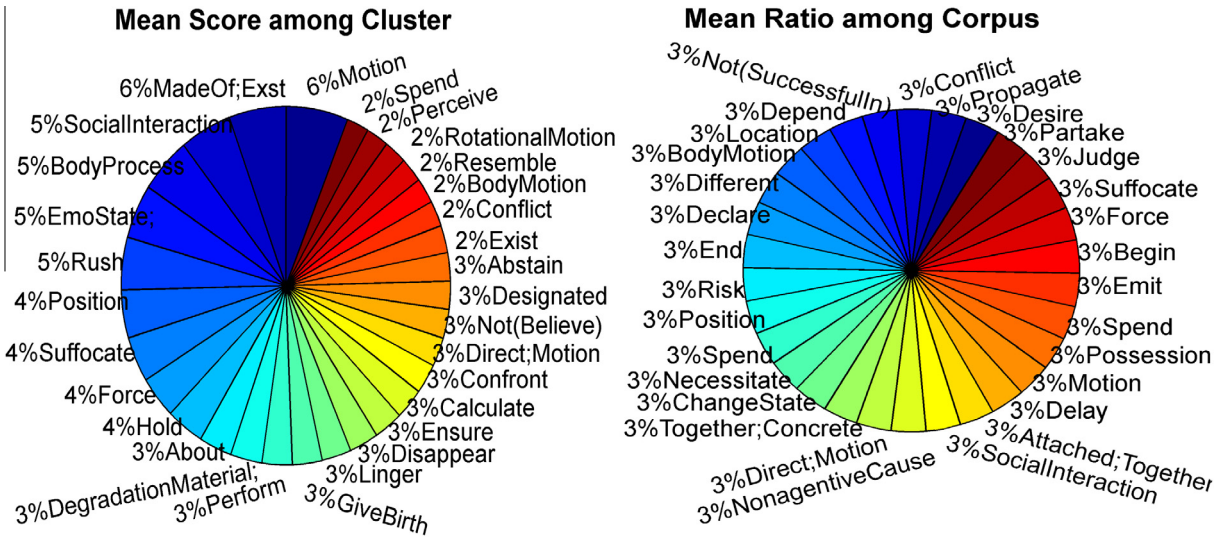

Figure 8 Sample of semantic annotation using semantic class density vectorization with *k*-means.

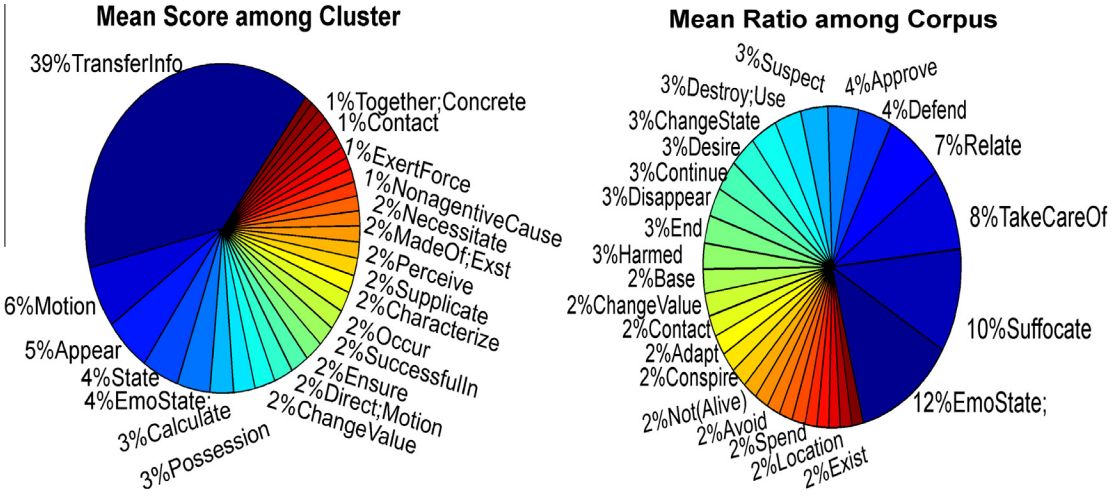

Figure 9 Sample of semantic annotation using semantic class probability distribution vectorization with *k*-means.

vectorization (semantic class probability distribution or semantic density) with *k*-means is substantially shorter. The vectorization with semantic density consumed 85 s to simply classify 753 documents into 54 clusters using this solution.

The semantic annotations for cluster number 6 when using (semantic class density or semantic class probability distribution) vectorization with *k*-means are shown in Fig. 8 and Fig. 9, respectively. Each cluster is represented by the five highest and relevant semantic features along with the percentage of each of them. The relevance score is based on two variables: mean score of the topic among cluster *k* and mean ratio of the topic among the corpus.

6. Discussion and conclusion

The presented results show that the proposed solutions are reasonably accurate and fast. Through the proposed document vectorization solutions with *k*-means, we have succeeded in increasing the purity and decreasing the MICD and BDI compared to the standard *k*-means algorithm. Furthermore, we managed to lower the runtime using the proposed solutions. Next, using the proposed document vectorization with *k*-means, the dimension of the documents is reduced from 753×4681 to 753×131. Hence, the document vectorization in our method helps to transform text documents into a semantic class probability distribution or semantic class density in the vector space.

Moreover, the document vectorization allows us to represent the web pages according to semantic class features, which are later used to calculate the semantic similarity between web pages. The semantic annotations in these web pages reveal informative exposition about the communications used within these pages. As shown in Fig. 8 and Fig. 9, respectively, we can see that each cluster can be labeled with five semantic features with different percentages according to the verbs found in each cluster. Consequently, the suggested solutions are able to show the semantic features shared between similar web pages that are grouped together in one cluster.

We believe that using the proposed approach is a promising technique to classify Arabic web pages according to the semantic similarities between them with a low runtime and an accurate performance. This approach is meant to enhance the document representation models for text clustering based on semantic similarities. For future work, we plan to utilize the proposed approach to extract the semantic orientation of Arabic web pages related to terrorism and extremism.

Acknowledgments

The authors would like to extend their thanks to the Ministry of Education, Malaysia, Universiti Teknologi Malaysia (UTM) under Vot 03H02, Umm Al-Qura University (UQU) and the Ministry of Higher Education, Saudi Arabia for supporting this research.

References

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26, 1–34.

Ahmed, Z., 2009. Domain Specific Information Extraction For Semantic Annotation. Charles University of Prague, Czech Republic and University of Nancy, France.

Alghamdi, H.M., Selamat, A., 2012. Topic detections in arabic Dark websites using improved vector space model. In: 4th Conference on Data Mining and Optimization (DMO). Langkawi, Malaysia, pp. 6–11.

Al-Khalifa, H., Al-Wabil, A., 2007. The Arabic language and the semantic web: challenges and opportunities. In: The 1st International Symposium on Computers and Arabic Language & Exhibition, Riyadh, Saudi Arabia.

Alsaleem, S., 2011. Automated Arabic text categorization using SVM and NB. Int. Arab J. e-Technol. 2, 124–128.

Alsaleem, S.M., 2013. Neural networks for the automation of arabic text categorization. In: International Conference on Computer Applications Technology (ICCAT), pp. 1–6.

Al-Shalabi, R., Kanaan, G., 2004. Constructing an automatic lexicon for Arabic language. Int. J. Comput. Inf. Sci. 2, 114–128.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recognit. 46, 243–256.

Beseiso, M., Ahmad, A.R., Ismail, R., 2011. An Arabic language framework for semantic web. In: International Conference on Semantic Technology and Information Retrieval. IEEE, Putrajaya, Malaysia, pp. 7–11.

Cha, S., 2007. Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. 1, 300–307.

Chang, Y., Lee, K., 2011. Bayesian feature selection for sparse topic model. In: IEEE International Workshop on Machine Learning for Signal Processing. IEEE, Beijing, China, pp. 1–6.

Chawla, S., Gionis, A., 2013. *k*-Means: a unified approach to clustering and outlier detection. In: the13th SIAM International Conference on Data Mining. Austin, Texas, USA, pp. 189–197.

Demiriz, A., Bennett, K., Embrechts, M., 1999. Semi-supervised clustering using genetic algorithms. In: Artificial Neural Networks in Engineering (ANNIE-99). ASME Press, pp. 809–814.

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

Deza, M.-M., Deza, E., 2006. Chapter 14 – distances in probability theory. In: Dictionary of Distances. Elsevier, pp. 176–188.

Elarnaoty, M., AbdelRahman, S., Fahmy, A., 2012. A machine learning approach for opinion holder extraction in Arabic language. Int. J. Artif. Intell. Appl. 3, 45–63.

Fabbri, R., Costa, L.D.F., Torelli, J.C., Bruno, O.M., 2008. 2D Euclidean distance transform algorithms. ACM Comput. Surv. 40, 1–44.

Faria, L., Akbik, A., Sierman, B., Ras, M., 2013. Automatic preservation watch using information extraction on the Web: a case study on semantic extraction of natural language for digital preservation. In: 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal.

Farra, N., Challita, E., Assi, R.A., Hajj, H., 2010. Sentence-level and document-level sentiment mining for Arabic texts. In: IEEE International Conference on Data Mining Workshops, IEEE Computer Society, pp. 1114–1119.

Forsati, R., Mahdavi, M., Shamsfard, M., Meybodi, M.R., 2013. Efficient stochastic algorithms for document clustering. Inf. Sci. 220, 269–291.

Froud, H., Benslimane, R., Lachkar, A., Ouatik, S.A., 2010. Stemming and similarity measures for Arabic documents clustering. In: 5th International Symposium on I/V Communications and Mobile Network (ISVC), IEEE, pp. 1–4.

Froud, H., Sahmoudi, I., Lachkar, A., 2013. An efficient approach to improve Arabic documents clustering based on a new keyphrases extraction algorithm. Comput. Sci., 243–256.

Hawwari, A., Zaghouani, W., O'Gorman, T., Badran, A., Diab, M., 2013. Building a lexical semantic resource for Arabic morphological patterns. In: International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), IEEE, pp. 1–6.

Huang, A., 2008. Similarity measures for text document clustering. In: The New Zealand Computer Science Research Student Conference (NZCSRSC'08), Christchurch, New Zealand.

Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R., 2008. Text document preprocessing with the Bayes formula for classification using the support vector machine. Trans. Knowl. Data Eng. 20, 1264–1272.

Karima, A., Zakaria, E., Yamina, T.G., 2012. Arabic text categorization: a comparative study of different representation modes. J. Theor. Appl. Inf. Technol. 38, 1–5.

Kipper, K., Korhonen, A., Ryant, N., Palmer, M., 2008. A large-scale classification of English verbs. Lang. Resour. Eval. J. 42, 21–40.

Larkey, L.S., Feng, F., Connell, M., Lavrenko, V., 2004. Language-specific models in multilingual topic tracking. In: Special Interest Group on Information Retrieval (SIGIR). ACM, Sheffield, UK, pp. 402–409.

Malik, S.K., Rizvi, S., 2011. Information extraction using web usage mining, web scrapping and semantic annotation. In: International Conference on Computational Intelligence and Communication Networks, IEEE Computer Society, pp. 465–469.

Mohammad, S., Resnik, P., 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, pp. 326–333.

Mousser, J., 2010. A large coverage verb taxonomy for Arabic. In: Seventh Conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta, pp. 2675–2681.

Mousser, J., 2011. Classifying Arabic verbs using sibling classes. In: International Workshop on Computational Semantics. Oxford, UK, pp. 355–359.

Nguyen, C., Phan, X., Horiguchi, S., 2009. Web search clustering and labeling with hidden topics. ACM Trans. Asian Lang. Inf. Process. 8, 37.

Park, S., Lee, S.R., 2012. Text clustering using semantic terms. Int. J. Hybrid Inf. Technol. 5, 135–140.

Rana, S., Jasola, S., Kumar, R., 2013. A boundary restricted adaptive particle swarm optimization for data clustering. Int. J. Mach. Learn. Cybern. 4, 391–400.

Rendón, E., Abundez, I., 2011. Internal versus external cluster validation indexes. Int. J. Comput. Commun. 5.

Saleh, L.M.B., Al-Khalifa, H., 2009. AraTation: an Arabic semantic annotation tool. In: Proceedings of the 11th International Conference on Information Integration and Web-Based Applications and Services, ACM, pp. 447–451.

Shaban, K., 2009. A semantic approach for document clustering. J. Softw. 4, 391–404.

Smith, J.R., Tesic, J., 2006. Semantic labeling of multimedia content clusters. In: International Conference on Multimedia and Expo. IEEE, Toronto, Canada, pp. 1493–1496.