Journal of King Saud University – Computer and Information Sciences xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Arabic Web page clustering: A review

Hanan M. Alghamdi [a,*], Ali Selamat [b]

[a] Department of Computer Science, College of Computing Alqunfudah, Umm Al-Qura University, Alqunfudah, Saudi Arabia
[b] Faculty of Computing, Universiti Teknologi Malaysia, UTM, Johor Bahru, Johor 81310, Malaysia

ARTICLE INFO

Article history:
Received 27 January 2017
Revised 22 May 2017
Accepted 12 June 2017
Available online xxxx

Keywords:
Feature selection
Feature reduction
K-means
Review
Text clustering
ARABIC Web page

ABSTRACT

Clustering is the method employed to group Web pages containing related information into clusters, which facilitates the allocation of relevant information. Clustering performance is mostly dependent on the text features' characteristics. The Arabic language has a complex morphology and is highly inflected. Thus, selecting appropriate features affects clustering performance positively. Many studies have addressed the clustering problem in Web pages with Arabic content. There are three main challenges in applying text clustering to Arabic Web page content. The first challenge concerns difficulty with identifying significant term features to represent original content by considering the hidden knowledge. The second challenge is related to reducing data dimensionality without losing essential information. The third challenge regards how to design a suitable model for clustering Arabic text that is capable of improving clustering performance. This paper presents an overview of existing Arabic Web page clustering methods, with the goals of clarifying existing problems and examining feature selection and reduction techniques for solving clustering difficulties. In line with the objectives and scope of this study, the present research is a joint effort to improve feature selection and vectorization frameworks in order to enhance current text analysis techniques that can be applied to Arabic Web pages.

Contents

1. Introduction . . . 00
2. Text clustering . . . 00
   2.1. Applications of text clustering . . . 00
   2.2. Related works in text clustering . . . 00
   2.3. Related works on Arabic Web text clustering techniques . . . 00
   2.4. Challenges of Arabic Web page analysis using clustering . . . 00
3. Feature selection methods . . . 00
   3.1. Document frequency (DF) . . . 00
   3.2. Information gain (IG) . . . 00
   3.3. Chi-square (CHI) . . . 00
   3.4. Term strength (TS) . . . 00
   3.5. Term contribution (TC) . . . 00
   3.6. Limitations of feature selection methods . . . 00
4. Dimensionality reduction methods . . . 00
   4.1. Principal component analysis (PCA) . . . 00

* Corresponding author.
E-mail addresses: hmhghamdi@uqu.edu.sa (H.M. Alghamdi), aselamat@utm.my (A. Selamat).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.06.002
1319-1578/

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

4.2. Probabilistic latent semantic analysis (PLSA) . . . 00
4.3. Latent semantic analysis (LSA) . . . 00
4.4. Limitations of dimensionality reduction methods. . . . 00
5. Feature hybridization methods . . . 00
6. Conclusion . . . 00
Acknowledgements . . . 00
References . . . 00

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 to manage this vast information, classify relevant data 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., 2013a; 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.

In text clustering, input documents are combined in groups according to the identified content similarity among the documents. Text clustering facilitates the processes of navigating, summarizing, and organizing vast and unorganized information and also finding content from unknown text (Ahmed and Tiun, 2014). Therefore, it is significant to review research in this area to analyse work done in the Arabic text clustering field. This will help identify gaps in literature on Arabic text clustering.

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 analyse 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 with Arabic Web page clustering and to propose solutions.

This paper is aimed to review these studies and explain the solutions applied to overcome the respective difficulties. Thus, the proposed work presents a review of Web page clustering based on Arabic text.

This paper is divided according to the main challenges discussed in previous studies regarding Web page clustering based on Arabic text. The challenges involve term representation, dimensional reduction and improving clustering performance. Each section presents a literature review on each challenge and the solution tasks.

This paper is organised as follows: Section 2 offers a general overview of the text clustering problems and applications, and related works on Arabic text clustering along with the challenges. Section 3 clarifies feature selection methods based on text clustering to solve the term representation issue. Section 4 illustrates the dimensional reduction issue. Section 5 considers means of designing a suitable model for clustering Arabic text that is capable of improving clustering process performance.

2. Text clustering

Text mining is a part of content mining. Web text mining techniques involve processing collections of Web texts and locating identical kinds of knowledge within unstructured data (Backialakshmi, 2015). Content mining deals with how to utilize data stored in text in a suitable machine-understandable form for automatic processing (Kamde and Algur, 2011). Therefore, the purpose of text mining is to transform unstructured textual data, extract meaningful numeric values from text and consequently make the information contained in text reachable to a variety of data mining technique applications (Backialakshmi, 2015). Research in the text mining area addresses different topics, such as information extraction, text summarization, text representation, text classification and document clustering.

Text clustering is the process of organizing a set of text documents to be clustered according to similarities. The aim is to discover natural document groupings, as text clustering achieves an overview of the classes or topics in a corpus (Steinbach et al., 2000).

The applications of nature inspired algorithms used in clustering include weather identification (Djallel DIlmi et al., 2017; Zhang et al., 2017), email spam filtering (Alsmadi and Alhami, 2015; Sahoo et al., 2017a; Zhiwei et al., 2017), SMS spam detection (Nagwani and Sharaff, 2015), stock market prediction (Astudillo et al., 2016; Bansal, 2017; Peachavanish, 2016), online customer review examination (Sahoo et al., 2017b; Stoica and Özyirmidokuz, 2015; Yakut et al., 2015), scientific articles indexing (Wang and Koopman, 2017), document kind identification (Lee et al., 2017; Nuovo et al., 2017), and so forth.

Clustering methodology is valuable for pattern analysis, grouping, decision-making, and machine learning situations, as well as image segmentation, data mining, pattern classification and document retrieval. With little information available about input data, clustering methodology is suitable for discovering relations between data points to assess the data.

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

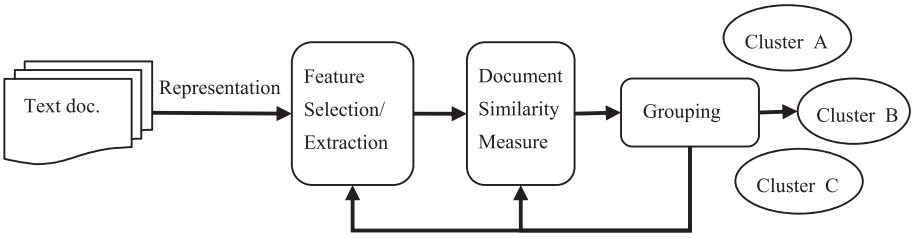

Fig. 1. Clustering stages.

As a result, a single cluster contains text documents much more highly similar to each other than to those in different clusters. Clustering is an unsupervised learning process because its properties or class memberships are unknown in advance (Andrews and Fox, 2007). A typical text clustering algorithm involves the following stages (Fig. 1) (Jain and Dubes, 1988):

i Stage 1: Document representation as an option to include feature extraction or selection methods.
ii Stage 2: Determining and calculating the document similarity measure.
iii Stage 3: Applying clustering or grouping rules.

Fig. 1 shows the sequence of clustering steps, consisting of a feedback path, as the grouping process results could influence consequent feature selection or extraction and similarity calculations (Jain and Murty, 1999). Document representation describes all inputs involved in the clustering algorithm, which are (i) the number of clusters, (ii) the number of documents to be clustered, and (iii) the number, type and weight of features that assist with the clustering process

The first stage is using a feature selection or extraction technique. Feature selection is aimed at identifying the effective set of presented features to be included in clustering, such as DF, IG, TC, etc. Feature extraction is intended to reduce the input features by using approaches like PCA and LSI. In text clustering, the goal of using feature selection or feature extraction is to optimize the clustering ability and computational efficiency by removing irrelevant and noisy terms (features) that carry insufficient information to help with the text clustering process. One or both of these techniques can be used to acquire an appropriate set of features for clustering (Jain and Murty, 1999).

The second stage entails measuring the similarities between input documents. Document similarity is measured by a distance function calculated between pairs of documents. The grouping step can be done by different algorithms for different document categorization stages.

The last stage involves building a clustering model of the input text using the best subset of features selected through feature selection and extraction, and evaluating the performance. The clustering algorithm is divided into hard, soft or fuzzy clustering. In hard clustering, each document belongs exactly to one cluster, which means one document cannot be assigned to two different clusters. An example of hard clustering is *k*-mean clustering. On the other hand, soft clustering assigns a degree of membership of each document to the output clusters. In this case, one document can belong to more than one cluster according to its membership level. Examples of soft clustering are Fuzzy C-Means and Expectation-Maximization (EM) algorithms.

2.1. Applications of text clustering

Clustering is found to be highly beneficial in several contexts and disciplines. Thus, it is applied widely owing to its usefulness as one of the steps in exploratory data analysis. Among the applications of clustering are as follows:

Web Page Clustering: The number of Web pages available is growing rapidly, thus requiring a method of organizing information efficiently and automatically. Clustering methods can automatically categorize Web pages into different topical classes (Thanh and Yamada, 2011). When Web pages are combined based on a similar class, it becomes easier for the research engine to limit the search to a class that contains the required information (Gourav, 2011).

Document Summarization: This is the process of building an abstract representation of an entire document (Froud et al., 2013a). It is very difficult for humans to manually summarize large text documents. An automated summarization system is encouraged that can incur less effort and reduce time consumption. The task of clustering is to select and retrieve related sentences and remove redundancies in the document summarization process (Fejer and Omar, 2015).

Sentiment Analysis: The purpose of sentiment analysis is to determine the emotional polarity of a writer with respect to a specific topic (Li and Wu, 2010). The clustering method requires some domain information either from the domain experts or any semantic repository to group the input text according to similar topical classes (Turney and Pantel, 2010). In this case, a sentimental value is calculated and assigned to each group of text (Gryc and Moilanen, 2010). Therefore, document clustering can be used to extract background knowledge of textual content (Sun et al., 2011).

2.2. Related works in text clustering

To create a significant clustering algorithm, unlimited running text needs to be transformed into structured information by utilizing a document representation model. VSM is broadly utilized for this kind of goal, which receives BOW through which to express the words. With this methodology, text content is represented as a vector specifically featuring a space using a word index, whereby every different vector value corresponds to an occasion or even lack of a chosen feature. The most commonly employed features associated with VSM are words, whereas other systems might utilize characters and phrases as features (Zhang and Zhang, 2006).

Selecting the most relevant and appropriate features for input data representation can greatly impact clustering accuracy improvement (Shaban, 2009; Turney and Pantel, 2010). The enriched representation scheme should reveal existing relations between concepts and assist with accurate similarity measurements to attain better clustering performance. Gabrilovich (2006) proposed a feature generator with the help of Wikipedia to improve document representation by analyzing input documents and mapping them onto relevant concepts. In another study, Shaban (2009) adopted meaning-based text representation to represent the input documents and measure the similarity between documents. The representation scheme entails gathering syntactic and semantic features and it illustrates the space of commonality between documents more explicitly. Gharib et al. (2012) recom-

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

mended a semantic document clustering methodology that involves using part of WordNet lexical categories along with the Self Organizing Map (SOM) neural network. The end goal is to represent more relevant features in VSM and improve the performance of document clustering.

Clustering requires a feature selection method with good performance to overcome the problem of selecting irrelevant features to represent textual data (Patel and Zaveri, 2011; Said et al., 2009). In recent studies (Antony et al., 2016; Chen, 2015; Sutar, 2015) researchers have recommended using a feature selection method to obtain salient features that result in clustering improvement. Sutar (2015) added correlation techniques to the feature selection method to remove irrelevant and redundant features among datasets. Chen (2015) proposed a feature selection method without the need to explore the input data. Only mutual information criteria are used based on the nearest and farthest neighbors to identify relevant features rather than visit the space of all possible feature subsets heuristically.

A hybrid model based on feature selection (DF, MI, IG, CHI) recommended by Li and Zhang (2012) is intended to join the points of interest of different feature selection models to enhance textual clustering. However, they did not acknowledge the semantic similarities between terms nor determine how to consolidate these similarities with feature selection.

Among the problems of text clustering methods is high dimensionality as a result of the enormous number of variables involved in text clustering. Implicating all terms found in a dataset in the clustering process results in a high number of dimensions in the vector representation of documents. Consequently, high-dimensional data decreases the efficiency of clustering methods and capitalizes on execution time.

A number of studies have suggested a low-dimensional VSM algorithm intended to reduce high-dimensional data using dimensionality reduction methods such as Principal Component Analysis (PCA) (e.g., Farahat and Kamel, 2011; Napoleon and Pavalakodi, 2011) or feature selection techniques such as CHI or Mutual Information (MI) (e.g., Li and Zhang, 2012). Napoleon and Pavalakodi (2011) raised the exactness of a *k*-means algorithm in high-dimensional datasets by utilizing PCA; nonetheless, they were not comprehensively ready to include semantic similarity between the terms. Interestingly, for Farahat and Kamel (2011) misuse of semantic relations occurred between terms with GVSM hybrid vector representation, whereby they mapped the statistical correlations between terms onto latent spaces (latent component indexing (LSI) or PCA). With their strategy, the viability of clustering functional processes improved; however, huge-scale datasets consequently need a distributed implementation of the complex calculation of semantic kernels, in contrast to VSM.

A hybrid classification technique was suggested by Isa et al. (Isa et al., 2008, 2009a; Lee et al., 2012) to decrease dimensionality by means of utilizing a probability distribution of the categories in a document. The categories are the vectors used to represent the document and afterwards to encourage the classifier to accept these distributions. Their model achieves comparative valuable alertness regarding accuracy, mainly due to the fundamental elimination of time taken. Moreover, the Bayes equation utilized as a weighting scheme contains some inadequacies mainly by failing to offer sufficient ability to recognize categories. The ability of this procedure to address categories with a significant amount of well-known keywords is limited in accordance with the lack of proficiency to recognize correct categories and comprise document information (Zhou et al., 2010). In addition, the varieties of probability distributions of terms in a respective document are not recognized (Guru et al., 2010).

Although background knowledge is reported in literature studies (Gharib et al., 2012; Hu et al., 2008; Jing et al., 2011; Park and Lee, 2012; Thanh and Yamada, 2011) regarding the improvement of document clustering effectiveness, there are some limitations. Wikipedia-based methods are not very simple to use when handling situations of synonymy and polysemy, and are in fact convoluted when mapping the initial text to the correct concepts. Besides, the powerless focus of WordNet-based strategies can be attributed to the presence of noise as well as unrelated data for documents in a specific area. This occurs since it is a general lexicon and employs common words that are occasionally terms representative of a document.

The algorithms described above have some or fewer drawbacks. Some such algorithms can eliminate irrelevant features but fail to handle high-dimensional features while others can remove irrelevant features and keep in view high-dimensional features. Text clustering is a useful technique for many applications and different languages (Amine et al., 2013; Froud et al., 2013a; Gharib et al., 2012; Jing et al., 2010; Liu et al., 2011; Sharma and Gupta, 2012; Thanh and Yamada, 2011). The next section presents a discussion on related works and issues pertaining to Arabic Web page clustering techniques.

2.3. Related works on Arabic Web text clustering techniques

The growth of Arabic Web pages with large amounts of text that hold unorganized informative data urge the necessity to adopt solutions to wisely manage such textual data (Elarnaoty et al., 2012). Due to the unstructured character of these texts, machines cannot efficiently understand valuable knowledge.

Compared to the large number of research works and resources available in English, the problem of high-dimensional data and lack of relevant features in Arabic documents has been studied much less (Froud et al., 2012, 2013a; Harrag et al., 2010; Karima et al., 2012; Sahmoudi et al., 2013).

Table 1 lists previously published studies on Arabic text clustering.

Table 1 contains five main columns: 'Study', 'Application', 'Clustering method', 'Evaluation' and 'Integrated method.' The 'Study' column represents the study reference, 'Application' explains the implemented study domain, 'Clustering method' symbolizes the method applied for clustering Arabic text in the study, and the 'Integrated method' column indicates the method utilized in the study to enhance clustering. 'Evaluation' is divided into four sub-columns, comprising 'Precision', 'Recall', 'F-measure' and 'Purity' measurements. These sub-columns show the evaluation measurements used in each study and performed by each data source using a few structure and content mining techniques.

Working with word or phrase structure in the Arabic language can enhance document clustering. Pre-processing and stemming methods such as a root-based stemmer or light stemmer can be used to obtain relevant features (Al-Anzi and AbuZeina, 2015; Ashour et al., 2012; Bsoul and Mohd, 2011; Ghanem, 2014; Harrag et al., 2010). Bsoul and Mohd (2011) examined the impact of using an Arabic root-based stemmer (ISRI) with different similarity measures, and suggested that stemming with ISRI improves clustering quality. Similar to their study, Ashour et al., (2012), Ahmed and Tiun (2014) and Ghanem (2014) carried out comparative studies on light stemming, root-based stemming and no stemming. The studies suggest that light stemming is more appropriate than root-based stemming or no stemming using precision and recall evaluation measures. Pre-processing steps are necessary to eliminate noise and keep only useful information to enhance document clustering performance (Ahmed and Tiun, 2014; Al-Omari, 2011).

On the other hand, some researchers have reported that using stemming to identify relevant features for Arabic text clustering may negatively affect the clustering results (Froud et al., 2013a;

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

Table 1
Summary of related works on Arabic text clustering.

Study	Application	Clustering method	Evaluation				Integrated method
			Precision	Recall	F-measure	Purity	
Ghwanmeh (2005)	Information retrieval system	Hierarchical *k*-means (HKM)	√				Hierarchical initial set
Fejer and Omar (2015)	Text summarization	*K*-means with hierarchical clustering		√			Keyphrase extraction
Amine et al. (2013)	Web pages clustering	*K*-means			√		Pre-processing (Stemming, stop-words removal)
Froud et al. (2013a)	Text summarization	*K*-means				√	LSA
Ashour et al. (2012)	Document clustering	*K*-means	√	√	√		Stemming
Al-Omari (2011)	Document clustering	*K*-means				√	Stemming
Sahmoudi et al. (2013)	Web pages clustering	Agglomerative Hierarchical clustering algorithm				√	Keyphrase extraction
Al-sarrayrih and Al-Shalabi (2009)	Document clustering and browsing	Frequent Itemset-based Hierarchical Clustering			√		N-grams
Ghanem (2014)	Web pages clustering	*K*-means	√	√	√		Pre-processing (term pruning, stemming and normalization)
Abuaiadah (2016)	Web pages clustering	*K*-means, Bisect *k*-Means				√	Pre-processing (Stemming, stop-words removal)
Al-Anzi and AbuZeina (2016)	Document clustering	EM, SOM, and *k*-Means	√	√			LSI
Alruily et al. (2010)	Document clustering	SOM					Rule-based approach (intransitive verbs and propositions)

Ashour et al., 2012; Amine et al., 2013). In particular, Arabic stemmers tend to produce high stemming error ratios (Al-Shammari and Lin, 2008). Root-based stemmers produce over-stemming errors because Arabic is a highly inflected and complex morphological language (Ashour et al., 2012), while light stemmers (Larkey et al., 2007) suffer from under-stemming errors. According to Al-Anzi and AbuZeina (2015, 2016), Al-Omari (2011) and Said et al. (2009) 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). Additionally, applying stemming only in clustering is not efficient because documents must be distinguished from each other according to category, while stemming of abstract words can lead to problems with wrongly distinguishing documents.

Other researchers (Fejer and Omar, 2015; Froud et al., 2013b; Sahmoudi et al., 2013; El-beltagy, 2006) have suggested using a keyphrase extraction algorithm based on the Suffix Tree (ST) data structure to improve clustering results by identifying the appropriate features. However, the manual assignment of keyphrases may be very time consuming 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 that particular keyphrase and difficulty arises in selecting the most relevant keyphrases (Sahmoudi and Lachkar, 2016).

Other approaches have been recommended to address the problem of high dimensionality in traditional clustering algorithms for Arabic text (Al-sarrayrih and Al-Shalabi, 2009; Awajan, 2015a, 2015b). This problem results from the large number of variables involved in text clustering methods. All terms found in the documents are included in the clustering process, which leads to a very large number of dimensions in the document vector representation. Therefore, high-dimensional data reduces clustering algorithm efficiency and maximizes execution time.

A novel approach, FIHC (Frequent Itemset-based Hierarchical Clustering), was proposed by Al-sarrayrih and Al-Shalabi (2009) to obtain the most frequently shared itemsets among document sets in clusters. They used N-grams based on word level and character level Trigrams and Quadgrams to extract the most frequent itemsets. They obtained promising results using N-grams based on word-level clustering of Arabic language text. However, a problem with the FIHC method is the number of word occurrences in a document as part of the clustering criteria (Backialakshmi, 2015). Alruily et al. (2010) combined information extraction with the SOM clustering method to help extract types of crime from documents in the crime domain. They applied this method with a rule-based approach using the dependency relation between some intransitive verbs and prepositions. They proved that the proposed method has the ability to extract keywords based on syntactic principles.

In some literature it is recommended to use probabilistic topic models for text representation to improve Arabic language clustering (Amine et al., 2013; Froud et al., 2013a; Al-Anzi and AbuZeina, 2016). The main purpose of topic modeling is to achieve machine-understandable and semantic explanations of Web text content in order to extract knowledge rather than unrelated information. Topic models are based on estimating the probability distributions of multiple topics in a document over a set of words. There are many probabilistic topic models such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). These models capture correlated words in the corpus with a low-dimensional set of multinomial distributions called "topics" and provide short descriptions of documents. Therefore, researchers use such models to extract important topics from large text (Ayadi et al., 2014; Lu et al., 2011; Sriurai, 2011). In addition, Amine et al. (2013) recommended LDA as a suitable model to deal with the morphological and syntactic characteristics of the Arabic language.

A study by Amine et al. (2013) highlighted the influence of the morpho-syntactic characteristics of the Arabic language on document clustering performance. They compared LDA with *k*-means clustering by applying both techniques on a set of Arabic documents. The authors suggested that using probabilistic topic models such as LDA provides substantial performance improvement over *k*-means. Froud et al. (2013a) applied LSA to produce Arabic summaries used to represent documents in VSM and cluster them in order to enhance Arabic document clustering (Froud et al., 2010).

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

Latent Semantic Indexing (LSI) was utilized by Al-Anzi and AbuZeina (2016) to group similar unlabeled documents into a pre-specified number of topics. They compared three different clustering methods: Expectation-Maximization (EM), Self-Organizing Map (SOM), and *k*-means algorithm. According to their research, LSI is recommended for labeling documents as well as improving clustering results. Awadalla and Alajmi (2011) suggested using synonym merging to preserve feature semantics as a way to solve the problem of feature synonym exclusion during the feature selection process.

Arabic text analysis is challenging (Al-Khalifa and Al-Wabil, 2007) due to the complicated morphological characteristics of Arabic words and sentences (Beseiso et al., 2010; Zitouni et al., 2010). Moreover, the in-depth analysis of large volumes of Web documents is also challenging (SAM, 2009) and consequently, an appropriate feature reduction and selection technique is required. In addition, misselecting and misrepresenting relevant features is always a concern with Arabic text analysis techniques (Awajan, 2015b). In other words, the fundamental challenges of Arabic text clustering regard the selection of features to best represent input text and designing feature vector models with the ability to discriminate based on the predefined information required for a particular clustering method (Ghanem, 2014). To date, work on enhancing document representation models for Arabic content clustering by reconciling semantic relations and lessening heightened dimensionality and runtime utilization is quite scarce, while outcomes present certain limitations. Therefore, a better technique for clustering Arabic textual data with a suitable feature selection and reduction design is highly desired. The next section explains the main issues and challenges with analysing Arabic Web pages using clustering methods.

2.4. Challenges of Arabic Web page analysis using clustering

There are three main challenges in applying text clustering to Arabic Web page content. The first challenge concerns the difficulty with identifying significant term features to represent original content by considering the hidden knowledge. Hidden knowledge is found in input text, such as semantic information and category relations. The Arabic language has a complex morphology and is highly inflected (Ashour et al., 2012). Thus, selecting appropriate features affects clustering performance positively. In order to further clarify the term representation issue, feature selection methods based on text clustering are examined in detail in Sect. .3.

The second challenge is related to data dimensionality reduction without losing essential information. Online textual data are numerous and contain features with several dimensions, which leads to complexity throughout the clustering process. For this reason, employing a proper method to discover essential information automatically from textual documents may provide the right features that function correctly to optimize clustering accuracy. Across text classification, vectorizing a document by estimating probability distribution is a successful means of dimension reduction employed to preserve processing time (Isa et al., 2008, 2009a, b; Lee et al., 2012). In order to further illustrate the dimensional reduction issue, feature extraction methods based on text clustering are closely examined in Sect. .4.

The third challenge regards how to design a suitable model for clustering Arabic text that is capable of improving clustering process performance. Clustering performance is mostly dependent on the features' characteristics (Jain and Murty, 1999). A Web page clustering technique is only effective when appropriate feature selection and feature reduction are integrated with a proper clustering method (Ghanem, 2014). However, improving clustering performance requires computational algorithms that adapt appropriate feature selection or reduction methods with well-established clustering approaches capable of achieving higher performance (Jain and Murty, 1999).

According to

Table 1, different approaches have been applied for clustering Arabic text. The majority of listed studies report using k-means (Froud et al., 2013a; Amine et al., 2013; Ashour et al., 2012; Ghanem, 2014). Ghanem (2014) strongly recommended implementing k-means for Arabic text clustering. However, Said et al.'s (2009) study demonstrated that Arabic text clustering performance can be further improved if it is adapted to appropriate feature selection and extraction methods. In order to identify the most appropriate clustering approaches, some of the available methods based on text 'clustering are explained in Sect. .5.

3. Feature selection methods

There are three types of features: irrelevant, strongly relevant and weakly relevant features (Ghwanmeh, 2005). Irrelevant features can be eliminated without affecting clustering performance while strongly relevant features contain helpful information and removing these will decrease clustering performance. Weakly relevant features contain information valuable for clustering, although they are not necessarily related to other words.

The feature selection method aims to eliminate irrelevant or redundant features, and keep features that contain reliable, useful information within a corpus (Seo et al., 2004). Feature selection is very beneficial, as it cuts down the operation time of clustering approaches. When removing unnecessary features, this results in small size data sets. In addition, it improves clustering accuracy by removing meaningless features and keeping the significant features in clustered text documents. The identified features are helpful for data clustering in maximizing the distance between clusters and minimizing the distance within clusters (Chen, 2015). Consequently, the machine memory size required to process the corpus is minimized.

Feature selection has been successfully applied in many real applications, such as in pattern recognition for Object-Based Land Cover Mapping of Unmanned Aerial Vehicle Imagery (Ma et al., 2017), or to improve the quality of ground-truth datasets for digital pathology (Marée, 2017), in text categorization for proper identification of student dataset (Rajeswari and Juliet, 2017) or to identifies terrorism-related documents (Choi et al., 2014; Sabbah et al., 2016a), in making investment decisions such as for stock prediction (Tsai and Hsiao, 2010), in image processing for early diagnosis of diseases (Adeli et al., 2017; Yang et al., 2017), and so forth.

Feature selection algorithms are classified into two primary categories: algorithms based on the filter model and those based on the wrapper model (Yu and Liu, 2004). The filter model depends on the common characteristics of the input text to estimate and select the subset of features without involving any other algorithm. A relevance metric is considered in feature selection (Chen, 2015), whereby a feature is either dependent (related, constant, reliable, significant or helpful) on the objective class or temporarily independent of the other features. The filter approach is widely used in the text clustering field, whereby features are chosen by score matrices like Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Chi-square (CHI) and Term Strength (TS). On the other hand, the wrapper model requires a pre-specified learning algorithm to be trained and search for features. It is aimed to successfully visit the space of all possible feature subsets to select the best feature subset resulting in performance improvement.

Feature selection methods have been proven to be valuable for text categorization and clustering (Dong et al., 2006; Mesleh,

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

2007a,b, 2008; Simanjuntak et al., 2010). A feature selection method developed by Abbasi et al. (2008) demonstrates the importance of stylistic and syntactic features in opinion classification. The authors demonstrated that using such approach improves the identification of the main features used for each sentiment class. Feature selection methods used regularly for Arabic text clustering are DF, IG, CHI, TS, and TC.

DF, IG, and CHI are confirmed to be correlated in terms of extracting any significant terms for text classification (Yang and Pedersen, 1997). TS and TC outperform DF in text clustering according to a comparative study of unsupervised feature selection (Liu et al., 2003). TC exhibits more advantages as an unsupervised feature selection method (Liu et al., 2005). Each feature selection method is detailed below.

3.1. Document frequency (DF)

DF is recognized as one of the more effective feature selection approaches with the text classification function (Dong et al., 2006). DF is the simplest among methods and has lower cost, but its performance level is similar to the CHI and IG feature selection methods (Yang and Pedersen, 1997). DF obtains documents in a corpus that contain a specific term and excludes all documents that do not contain the assigned term. This means the term is weighted according to its frequency of appearance in more than a single document of the corpus. Therefore, if the collection has ten documents and term A appears in four documents, the weight of term A is four. This technique follows the essential assumption that uncommon terms are either non-informative for predicting the category, or not important in the overall performance (Xu and Chen, 2010). Some terms are deleted if they do not match a lower predefined DF threshold. Based on this assumption, terms with lower DF are considered as noise to the document representation and are not valuable to the clustering process.

3.2. Information gain (IG)

The IG method computes the information c_i number for category prediction according to the non-occurrence of a term t in a document d (Xu and Chen, 2010). A hybrid method combining Document Frequency and Information Gain performs best in term selection used by the KNN classifier for Arabic text (Syiam and Fayed, 2006). In addition, IG is used to enhance Support Vector Machine (SVM) learning algorithms, as done by Chen (2008), and exceeds the standard SVM model. For each category, $i \in [1, M]$ where M is the number of categories in a corpus. The IG for a term t is defined in Eq. (1) (Mesleh, 2008):

$$IG(t) = -\sum_{i=1}^{M} p(c_i).\log\ p(c_i) + P(t)\sum_{i=1}^{M} p(c_i|t).\log\ p(c_i|t) + p(\bar{t})\sum_{i=1}^{M} p(c_i|\bar{t}).\log\ p(c_i|\bar{t}) \quad (1)$$

where $p(c_i)$ is the chance that a document variable d relates to class c_i, $p(c_i|t)$ is the probability of class c_i given that document d does not contain a term t, while $p(c_i|\bar{t})$ is the probability of a class c_i when document d contains term t.

3.3. Chi-square (CHI)

The CHI method is used to evaluate the lack of independence relating to text features and text categories. High Arabic text classification effectiveness is evident when using CHI with SVM (Mesleh, 2007a,b; Thabtah et al., 2009). Essentially, CHI feature selection is proven to be a suitable method of classifying Arabic text. The mathematical definition of Chi-square is given in Eq. (2) (Mesleh, 2008), where t represents a term, c represents a category, and $P(\bar{c})$ is calculated as the number of documents not related to class c divided by the total number of training documents.

$$CHI(t,c) = \frac{N.[P(t,c).P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)]^2}{P(t)P(\bar{t})P(c)P(\bar{c})} \quad (2)$$

3.4. Term strength (TS)

The TS method was first introduced by Wilbur and Sirotkin (1992) for use as a stop word reduction means in text retrieval. Later, Yang (1995) applied TS for text categorization purposes. TS can have high computation complexity when used with high numbers of documents, which may lead to difficulty with parameter tuning (Liu et al., 2003). This approach is based on estimating term strength, where strong terms are relatively informative and shared by related documents (Do and Hui, 2006). This method involves two steps:

i Compute the similarities of all documents in a corpus as pairs using the cosine similarity metric value *sim (d_i, dj)* of the two documents. If *sim (di, dj)* exceeds the predefined threshold, then *di* and *dj* are considered similar.
ii The term strength for term t is calculated according to the conditional probability that term t appears in document d_i when it occurs in document d_j as follows:

Table 2
Summary of single and hybrid feature selection methods.

Reference	Feature selection method	Method of combination	Application
Wang et al. (2007)	Word category distinguishing ability and IG	IG after category distinguishing	Chinese text sentiment classification
Chantar and Corne (2011)	BPSO and KNN	KNN after BPSO	Arabic document categorization
Habib et al. (2006)	DF and IG	IG after DF	Arabic document classification
Sabbah et al. (2016b)	TF, DF, IDF, TF–IDF, Glasgow, and Entropy	Union and symmetric difference	Arabic Web page classification
Thabtah et al. (2009)	CHI	Not applicable	Arabic text categorization
Zahran and Kanaan (2009)	PSO	Not applicable	Arabic text categorization
Al-Harbi et al. (2008)	CHI	Not applicable	Arabic text classification
Syiam and Fayed (2006)	DF, IG	IG after DF	Arabic text categorization
Said et al. (2009)	MI, IG, and DF	Not applicable	Arabic text categorization
Li and Zhang (2012)	DF, MI, IG, and CHI	Union	Text classification
Tsai and Hsiao (2010)	PCA and GA	Union, intersection, and multi-intersection strategies	Stock prediction
Awadalla and Alajmi (2011)	DF, TFIDF, CHI, IG, and MI	Not applicable	Arabic document classification

*PCA: Principal Component Analysis, GA: Genetic Algorithm, PSO: Particle Swarm Optimization, KNN: K nearest neighbour, IG: Information Gain, MI: Mutual Information, DF: Document Frequency, TFIDF: Term Frequency–Inverse Document Frequency, CHI: Chi-square, BPSO: Binary PSO.

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

$$TS(t) = p(t \in d_i | t \in d_j) \quad with\ i \neq j \tag{3}$$

3.5. Term contribution (TC)

TC appears to be a favourite unsupervised selection method for text clustering owing to its lower computational cost (Liu et al., 2003). Almeida et al. (2009) used TC as a term selection means to implement text clustering with the *k*-means algorithm. It overcomes the disadvantage of DF of not considering the term's distribution among the corpus, especially when term *t* occurs frequently (Osinski, 2004). It considers the term's contribution to the corpus and is calculated as the overall term's contribution to the documents' similarities. The mathematical calculation provided by Liu et al. (2003) is shown in Eq. (4), where $w_{(t,d_i)}$ denotes the weighting of term *t* in document d_i using TFIDF, and *N* is the total number of documents in the corpus.

$$TC(t) = \sum_{i=1}^{N} \sum_{j=1 j \neq i}^{N} w_{(t,d_i)} \times w_{(t,d_j)} \tag{4}$$

3.6. Limitations of feature selection methods

In terms of Arabic document clustering and based on Sect. .2.3 above, most studies on Arabic text clustering (Al-sarrayrih and Al-Shalabi, 2009; Amine et al., 2013; Fejer and Omar, 2015; Froud et al., 2010, 2013a; Ho et al., 2003) do not use DF, TFIDF, TS, CHI, TS and TC as feature selection methods in clustering algorithms. On the other hand, other classification and categorization applications for Arabic text that heavily use feature selection methods are shown in Table 2 (Al-Harbi et al., 2008; Awadalla and Alajmi, 2011; Habib et al., 2006; Said et al., 2009; Syiam and Fayed, 2006; Thabtah et al., 2009; Zahran and Kanaan, 2009). A feature selection method serves to obtain an appropriate set of features for use in clustering (Jain and Murty, 1999). From a clustering point of view, removing irrelevant features will not negatively influence clustering accuracy but it will decrease the required storage and processing time (Alelyani et al., 2016). Accordingly, there is a need to enrich the application of feature selection methods for Arabic text clustering.

In addition, Dai et al. (2003), Sabbah et al. (2016b), Tsai and Hsiao (2010) and Li and Zhang (2012) found that appropriate combinations of different types of features can function better than a single type of feature. The concept of combining different feature forms has been applied in Arabic text classification tasks (Syiam and Fayed, 2006; Chantar and Corne, 2011; Habib et al., 2006; Sabbah et al., 2016b) as shown in Table 2.

Habib et al. (2006) did a research and concluded that a hybrid approach is a preferable feature selection method for Arabic text classification tasks over a single feature selection method. Their research was motivated by the fact that categorizing Arabic text using a feature selection method which considers all categories can offer higher results. However, this may affect some documents such that they fail to show any terms in the set of selected features, as found in their study (Syiam and Fayed, 2006). On the other hand, feature selection dependent on a single document can be used to mitigate this problem, but it produces a lower classification rate. Hence, to balance between these feature selection methods, a hybrid approach was recommended (Habib et al., 2006). The feature sets generated based on different ideas on relevant features are combined into one hybridized feature set to benefit from the strength of each method while complementing for the weaknesses of other methods in order to obtain better identifiers for text clustering as well as improve clustering performance. For instance, CHI evaluates the lack of independence between terms and categories, MI estimates the compactness of terms with categories, while TF–IDF determines the relevant density of a given word in a single document and DF represents the number of term occurrences in different documents. The feature selection hybridization method facilitates identifying a feature set that strengthens the term weight according to the category's popularity and weakens the term weight according to the category's unpopularity within the document.

4. Dimensionality reduction methods

A method of feature reduction is a learning procedure that captures hidden topics within a given document according to the relationship between topics and words, as well as words and words. It reveals the topics in a corpus without providing any predefined categories. It is based on the definition that every document is a combination of topics and every word within a document belongs to a topic. Furthermore, a single document may contain several words generated by different topics (Wallach, 2006). A feature reduction task is sometimes referred to as a feature extraction task (SAM, 2009; Abbasi et al., 2008a). Feature reduction concerns reducing data complexity to a simpler form of information. A few works have been carried out with the purpose of using feature extraction as topic modeling with Arabic text, as pointed out by Brahmi et al. (2011). The present study employs Principal Component Analysis (PCA) as applied by Buntine et al. (2004) and Buntine and Jakulin (2004), and Probability Latent Semantic Analysis (PLSA) as applied by Lu et al. (2011) and Zhou et al. (2014). Meanwhile, LSA was utilized by Dumais et al. (1997) and Landauer et al. (1998). The resulting hidden knowledge from these models can be used to enhance the clustering algorithm. The feature reduction models are detailed below.

4.1. Principal component analysis (PCA)

PCA models a topic at the word level inside a document, making the topic model a discrete analogue to PCA (Buntine, 2009). This signifies that PCA can also be used to learn topics from a set of documents. Liu et al. (2011) studied PCA for clustering topics. Different extended models of PCA are suggested in the literature regarding the same purpose. Discrete PCA known to build independent components was suggested by Buntine et al. (2004) as a better topic model for Web data. Perkio et al. (2004) used Multinomial Principal Component Analysis (MPCA) as topic modeling to detect topic trends. PCA follows the steps below (Salehi and Ahmadi, 1993):

i Consider X is a $N \times d$ data matrix with one row vector x_n per data point.
ii Subtract mean *x* from each row vector x_n in X. The subtracted mean is the average across every dimension.
iii Then compute the covariance matrix of X.
iv Determine the eigenvectors and eigenvalues of the defined covariance matrix.
v The eigenvector with the highest eigenvalue is considered the principal component of the data set.
vi Multiply the eigenvector with largest eigenvalue by the standardized original data. The first principal component (PC1) is the linear combination of the standardized data with the first eigenvector and is used as the weight.

4.2. Probabilistic latent semantic analysis (PLSA)

The main purpose of PLSA is text document analysis and retrieval. PLSA is an unsupervised way of discovering latent topics or latent events within the text. It is used to automatically group text

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

topics or events from the text and extract keywords (semantics) without using any prior knowledge. PLSA has an advantage in modeling in comparison to LSA. PLSA defines a proper generative data model that can interpret a document in the Probabilistic Latent Semantic space as multinomial word distributions (Xu et al., 2008). It is deemed a better choice for model selection and complexity control. As a result, PLSA is appropriate for clustering text extracted from Web sites using different text presentations and weighting schemes.

In the PLSA model, each document doc_d (with a weight of the word $term_t$) is associated with an unobserved topic variable z, where the documents are represented using VSM, each document is represented as a collection of vectors $doc_d = \{W_{1,d}, W_{2,d}, \cdots\cdots, W_{t,d}\}$, $d = 1, 2, 3 \cdots N$, N is the overall number of documents in the corpus and $W_{t,d}$ is the weight of $term_t$ in document doc_d. Thus, $V_{t,d} = v(doc_d, term_t)$, where $v(doc_d, term_t)$ indicates the weight $W_{t,d}$ of the word $term_t$ in the document doc_d.

In each iteration, a topic $z = 1, 2 \ldots k \sim p(z)$ is chosen first, followed by a document and a word that are independent of each other but both are dependent on the topic, as $doc_d \sim p(doc_d|z_k)$, $term_t \sim p(term_t|z_k)$. A joint probability model $p(term, doc)$ is defined by the combination in Eq. (5) as stated by Xu et al. (2008):

$$p(term, doc) = \sum_{z=1}^{k} p(z)p(doc|z)p(term|z) \tag{5}$$

where $p(term|z)$ and $p(doc|z)$ represent the probabilities of the selected word/document respectively on latent class variable z. Then this model estimates with the selected estimation algorithm. One of the commonly known algorithms for estimating parameters in PLSA is Expectation-Maximization (EM) (Cheng et al., 2011).

EM estimates $p(term|z)$ and $p(doc|z)$ in the model with the following equations

Expectation – which calculates the probability of the following posterior

$$p(z|doc, term) = \frac{p(z)p(doc|z)p(term|z)}{\sum_{i=1}^{k} p(z_i)p(doc|z_i)p(term|z_i)} \tag{6}$$

Maximization – which estimates the following terms:

$$p(term|z) \propto \sum_{d=1}^{N} v(doc_d, term)p(z|doc_d, term) \tag{7}$$

$$p(doc|z) \propto \sum_{t=1}^{M} v(doc, term_t)p(z|doc, term_t) \tag{8}$$

$$p(z) \propto \sum_{d=1}^{N}\sum_{t=1}^{M} v(doc_d, term_t)p(z|doc_d, term_t) \tag{9}$$

The output is a new reduced matrix $N \times k$ of the class-conditional probability of every document as in Eq. (10):

$$\begin{bmatrix} p(doc_1|z_1) & \cdots & p(doc_1|z_k) \\ \vdots & \ddots & \vdots \\ p(doc_N|z_1) & \cdots & p(doc_N|z_k) \end{bmatrix} \tag{10}$$

where N is the total number of documents enclosed in the extracted Web page, k is the total number of topics corresponding to the topic categories in the extracted Web page and M is the total number of words. The two EM steps (Eqs. (6), (7), (8) and (9)) are repeated until convergence is reached.

Once the model is trained, it can be said that $p(term|z)$ are the topics. Each topic is defined by a word multinomial since the topics seem to have distinct semantic meanings. From $p(doc|z)$ and $p(z)$ it is possible to compute $p(z|doc) \propto p(doc|z)p(z)$, where $p(z|doc)$ is the topic weight for document doc_d.

4.3. Latent semantic analysis (LSA)

LSA is meant to measure the semantic similarity of a corpus, as it represents the text corpus in a more semantic manner (Mihalcea et al., 2006). Latent Semantic Indexing (LSI) is the application of LSA in the information retrieval field. LSA estimates the similarities between documents in which words appear and creates a new reduced representation of the text according to the relationships found between the words (Landauer et al., 1998). It is supposed that words in identical documents tend to have related meanings, thus bringing documents with similar topical content close to one another in the resulting space. LSA is reported in the literature for improving the performance of information retrieval and filtering. Paulsen and Ramampiaro (2009) used LSI with k-means clustering to retrieve related documents from a collection of documents. On the other hand, Lucia et al. (2007) used LSA as an information retrieval method to define the dissimilarities between Web pages and then utilized clustering algorithms to group related Web pages. LSA follows Singular Value Decomposition (SVD) of the term-document matrix of a textual collection as shown in Fig. 2.

LSA can be implemented using the SVD of the original term-document matrix X $(N \times d)$ as in Eq. (11) below.

$$A = USV^T \tag{11}$$

A word vector is represented in matrix U, a document vector is represented in matrix V, and S is a diagonal matrix containing the singular values of D. LSA retains single k-largest ($k < r = \text{rank}(A)$) singular triples (Ampazis and Perantonis, 2004).

4.4. Limitations of dimensionality reduction methods

Limitations of the explained dimensionality reduction methods are reported by Hoenkamp (2011), Prasad and Bruce (2008), Sabbah et al. (2016b), SAM (2009), and Thomas (2011). Using SVD with PCA and LSA incurs a limitation with the bottom effect on categorical data (Hoenkamp, 2011). Hence, these methods are not efficient representations for categorical data. The discussed feature reduction models are not human-readable; consequently, it is difficult to apply the resulting topics further such as in clustering annotation. Moreover, determining the number of topics for each model is based on heuristics and requires expertise. Dimensional reduction can also only be achieved if the original variables are correlated (Prasad and Bruce, 2008).

Other methods have been proposed in different studies for the purpose of reducing feature dimensionality without the use of

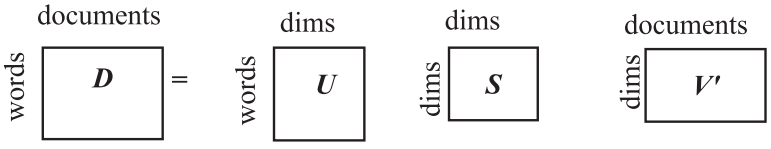

Fig. 2. LSA matrix factorization.

PCA, LSA or PLSA (Isa et al., 2009b; Li et al., 2008; Zhang et al., 2008). The problem of high-dimensional features is sometimes a result of how the features are represented. Accordingly, an enhanced representation model is a solution to reduce document vector dimensionality, as reported in different studies (Zhang et al., 2010; Awajan, 2015a, 2015b; Hotho et al., 2003; Barresi et al., 2008). Another way a specific factor or correlation can be addressed along with the feature reduction method is to describe the term-category dependency more accurately (Isa et al., 2009b; Li et al., 2008) or to represent the word distribution based on the semantic context (Awajan, 2015a,b; Gharib et al., 2012).

According to Amine et al. (2013) and Lu et al. (2011) there are generally two ways to use a feature reduction method for document clustering. In the first approach, the feature reduction method acts as a topic model to reduce the document representation dimensionality (from high-dimensional representation of documents (word features) to low-dimensional representation of topic features), after which a standard clustering algorithm like *k*-means is applied in the new representation. The other approach uses feature reduction methods more directly.

Using a feature reduction method as a topic model is a process of detecting valuable knowledge hidden within the data. The topic model is integrated with a clustering approach to group the provided text content into unrelated groups based on the content similarity score calculated. Text documents that address similar topics are grouped together. Document clustering and topic modeling are related and can benefit each other. On one hand, topic models can determine latent semantics embedded in the document corpus and the semantic information can be much more useful for recognizing document groups than raw term features. In classic document clustering approaches, documents are usually represented with a BOW model, which is purely based on raw terms and is not enough to capture all semantics. Topic models are able to place words with similar semantics into the same group called a topic, where synonymous words are treated as the same. Under topic models, the document corpus is projected into a topic space, which reduces the similarity measure noise and the corpus grouping structure can be identified more effectively.

The most straightforward basic term used to represent a text document is a word. In various text clustering cases a word is a meaningful unit of little ambiguity that can be overcome by considering the background information. The representation scheme reflects existing relations between concepts (structures and semantics) and assists with accurate similarity measurements that result in improved clustering performance. Enriching text representations with knowledge integrated with accurate similarity measures will enhance clustering process output. Across text clustering, vectorizing a document by estimating the probability distribution is a successful dimension reduction means employed to preserve processing time (Isa et al., 2008, 2009a; Lee et al., 2012).

Isa et al. (2008, 2009a) and Lee et al. (2012) recommended an approach to decrease the dimensions by estimating the probability distribution of categories in a document. The probability distribution of categories in a document are the vectors used to represent the document and afterwards to encourage the classifier to accept these distributions. The naive Bayes is used to vectorize raw text data based on probability values and SOM serves for automatic clustering based on previously vectorized data. The text representation consisting of the document probability distribution is annotated to various predefined categories using the Bayes formula (Dai et al., 2003). Naive Bayes is used to vectorize input text because it is considered term-category dependent based on the calculation of the probability distribution of the document related to the current categories in the corpus.

A low-dimensional semantic VSM method was proposed by Awajan (2015a, 2015b). Their method uses the extracted semantic information to build a word-context matrix to represent the word distribution across contexts and to transform the text into a VSM representation based on word semantic similarity. Lexical semantic resources such as Arabic WordNet and named entity gazetteers are involved in the proposed method.

Analyzing Arabic text is challenging (Al-Khalifa and Al-Wabil, 2007) due to the morphological characteristics of Arabic words and sentences (Beseiso et al., 2011). Developing a feature reduction method for Arabic text involves discriminating and deep semantic processing. Moreover, categorizing is a powerful tool to manage large numbers of text documents. By grouping text documents into sets of categories, it is possible to efficiently maintain or search for needed information (Hu et al., 2009).

To adapt the feature reduction method as a topic model to Arabic text clustering, the dimensionality problem must be solved along with term-semantic correlation and term-category dependency. Although the above studies did not use feature reduction methods for clustering Arabic Web page text, the studies did suggest proposing a feature reduction method that is capable of improving model performance.

5. Feature hybridization methods

Feature hybridization involves combining different attribute features to form a new feature set. In addition, different ways of selecting and extracting relevant features may produce different

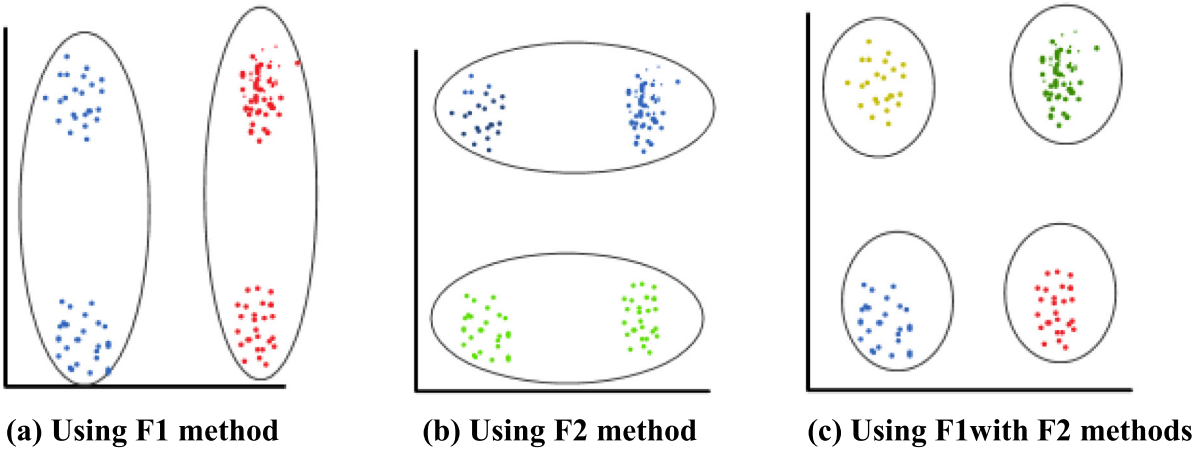

Fig. 3. Single or multiple feature selection methods with clustering.

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

clustering results (Alelyani et al., 2016). In Fig. 3 (c), four clusters are the output of using two feature selection and reduction methods (F1 and F2), while Fig. 3 (a) and Fig. 3 (b) demonstrate two clusters when using either a feature selection or feature reduction method only (F1 or F2). Consequently, combining multiple feature selection and reduction methods may result in different clustering, which can significantly aid with the discovery of hidden knowledge in the input text.

There is always a difference in clustering performance when using full data set representation or computerized feature selection and extraction methods. Feature selection removes input variables that have no significance to model performance, but it may also eliminate some variables that hold valuable information (Awajan, 2015a). On the other hand, feature extraction summarizes the original variables but transforms them into smaller sets to retain as much information as possible (i.e., information that represents real content).

To optimize model performance, some researchers have suggested including a feature hybridization process that combines multiple feature selection and extraction methods to identify the more representative variables. For example, Selamat et al. (2011) introduced a hybrid feature extraction and selection approach to solve the high-dimensionality problem during Web page identification. SAM (2009) used the feature hybridization concept for improving the filtering accuracy of illicit Web content, while Sabbah et al. (2016b) employed a hybrid of Glasgow and Entropy term weighting with TF, DF, IDF, and TF–IDF for Dark Web classification. Although these studies did not focus on feature hybridization for Arabic Web page clustering purposes, they did highlight that using feature hybridization to obtain the most representative features can enhance model performance.

6. Conclusion

This work entailed a review of Web text clustering studies and it was found there are three main challenges to Arabic text analysis based on clustering approaches. First is the complexity in identifying significant term features that better describe the original contents. Second is the large amount of Web documents with thousands of term features that constantly form high-dimensionality features, which results in difficulties during the clustering process. Third is the need for appropriately designed feature selection and reduction methods that are compliant to Arabic text clustering. To overcome these challenges, this study reviewed feature selection and feature reduction methods for improving clustering approaches that implement text analysis.

Future work concerns deeper analysis to try different methods to overcome the three main challenges. There are some ways of improvement arising from this review which should be pursued. Firstly, it is recommended to investigate thoroughly a way of integrating a hybrid feature selection scheme between different feature selection methods to overcome the difficulty of identifying the most informative words within a set of documents for clustering. This may improve the ability for term representation for high similarity Web content and allow better identification of significant term features.

Secondly, it is recommended to examine a method to reduce the dimension of representation of documents by projecting the high-dimensional data into lower-dimensional space. This will hopefully eliminate any misleading term features, with a view to producing definitive and small features. This should allow better interpretation of documents and remove any ambiguity caused by high-dimensional dataset used in clustering process.

Thirdly, it is suggested to enhance clustering performance by fully adapt feature selection and feature extraction methods in the design of clustering approach. However, it may be possible to extract the semantic features from high-dimensional Arabic Web pages and to cluster these pages according to the similarities of their features. These mechanisms need to be further investigated and examined for their ability to optimize the clustering.

Acknowledgements

The authors would like to extend their thanks to the 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.

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.

Abuaiadah, D., 2016. Using bisect K-means clustering technique in the analysis of arabic documents. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 15, 1–13.

Adeli, E., Wu, G., Saghafi, B., An, L., Shi, F., Shen, D., 2017. Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson's disease. Sci. Rep. 7, 41069.

Ahmed, M.H., Tiun, S., 2014. K-means based algorithm for islamic document clustering. Int. J. Islam. Appl. Comput. Sci. Technol. 2, 1–8.

Al-Anzi, F.S., AbuZeina, D., 2015. Stemming impact on arabic text categorization performance: a survey. In: Inf. Commun. Technol. Accessibility. IEEE, Marrakech, Morocco, pp. 1–7.

Al-Anzi, F.S., 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. IIE, Bangkok, Thailand, pp. 1–4.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M.S., Al-Rajeh, A., 2008. Automatic arabic text classification. In: The International Conference on the Statistical Analysis of Textual Data, pp. 77–84. Lyon, France.

Al-Khalifa, H., Al-Wabil, A., 2007. The arabic language and the semantic Web: challenges and opportunities. In: The International Symposium on Computers and Arabic Language & Exhibition. Riyadh, Saudi Arabia, pp. 1–9.

Al-Omari, O., 2011. Evaluating the effect of stemming in clustering of arabic documents. Acad. Res. Int. 1, 284–291.

Al-sarrayrih, H., Al-Shalabi, R., 2009. Clustering arabic documents using frequent itemset-based hierarchical clustering with an N-Grams. In: The International Conference on Information Technology, pp. 1–8. Amman, Jordan.

Al-Shammari, E., Lin, J., 2008. Towards an error-free arabic stemming. In: Proceeding of the 2nd ACM Workshop on Improving Non English Web Searching – iNEWS '08. ACM Press, California, USA, pp. 9–15.

Alelyani, S., Tang, J., Liu, H., 2016. Feature selection for clustering: a review. In: Aggarwal, C.C., Reddy, C.K. (Eds.), Data Clustering: Algorithms and Applications. CRC Press, pp. 29–60.

Ali, N.G., Omar, N., 2014. Arabic keyphrases extraction using a hybrid of statistical and machine learning methods. In: International Conference on Information Technology and Multimedia. IEEE, Putrajaya, Malaysia, pp. 281–286.

Almeida, L.G.P., Vasconcelos, A.T.R., Maia, M.A.G., 2009. A simple and fast term selection procedure for text clustering. Intell. Text Categ. Clust. 164, 47–64.

Alruily, M., Ayesh, A., Al-Marghilani, A., 2010. Using self organizing map to cluster arabic crime documents. In: International Multiconference on Computer Science and Information Technology. IEEE, Wisla, Poland, pp. 357–363.

Alsmadi, I., Alhami, I., 2015. Clustering and classification of email contents. J. King Saud Univ. – Comput. Inf. Sci. 27, 46–57.

Alsulami, B.S., Abulkhair, M.F., Essa, F.A., 2012. Semantic clustering approach based multi-agent system for information retrieval on Web. Int. J. Comput. Sci. Netw. Secur. 12, 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. (Eds.), Modeling Approaches and Algorithms. Springer International Publishing Switzerland, pp. 37–46.

Ampazis, N., Perantonis, S.J., 2004. LSISOM – a latent semantic indexing approach to self-organizing maps of document collections. Neural Process. Lett. 19, 157–173.

Andrews, N.O., Fox, E.A., 2007. Recent Developments in Document Clustering (No. TR-07-35).

Antony, D.A., Singh, G., Leavline, E.J., Priyanka, E., Sumathi, C., 2016. Feature selection using rough set for improving the performance of the supervised learner. Int. J. Adv. Sci. Technol. 87, 1–8.

Ashour, O., Ghanem, A., Wesam, M., 2012. Stemming effectiveness in clustering of arabic documents. Int. J. Comput. Appl. 49, 1–6.

Astudillo, C.A., Poblete, J., Resta, M., Oommen, B.J., 2016. A cluster analysis of stock market data using hierarchical SOMs. In: Kang, B.H., Bai, Q. (Eds.), AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart,

TAS, Australia, December 5–8, 2016, Proceedings. Springer International Publishing, Cham, pp. 101–112.
Awadalla, M.H., Alajmi, A.F., 2011. Dewy index based arabic document classification with synonyms merge feature reduction. Int. J. Comput. Sci. Issues 8, 46–54.
Awajan, A., 2015a. Semantic similarity based approach for reducing arabic texts dimensionality. Int. J. Speech Technol. 19, 191–201.
Awajan, A., 2015b. Semantic vector space model for reducing arabic text dimensionality. In: Digital Information and Communication Technology and Its Applications. IEEE, Beirut, Lebanon, pp. 129–135.
Ayadi, R., Maraoui, M., Zrigui, M., 2014. Latent topic model for indexing arabic documents. Int. J. Inf. Retr. Res. 4, 29–45.
Backialakshmi, P., 2015. Knowledge discovery in text mining with big data using clustering based word sequence. Int. J. Adv. Res. Datamining Cloud Comput. 3, 35–52.
Bansal, A., 2017. Improved K-mean clustering algorithm for prediction analysis using classification technique in data mining. Int. J. Comput. Appl. 157, 35–40.
Barresi, S., Nefti, S., Rezgui, Y., 2008. A concept based indexing approach for document clustering. In: The IEEE International Conference on Semantic Computing. IEEE Computer Society, Washington, DC, USA, pp. 26–33.
Beseiso, M., Ahmad, A.R., Ismail, R., 2010. A survey of arabic language support in semantic Web. Int. J. Comput. Appl. 9, 24–28.
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.
Brahmi, A., Ech-Cherif, A., Benyettou, A., 2011. Arabic texts analysis for topic modeling evaluation. Inf. Retr. Boston 15, 1–21.
Bsoul, Q.W., Mohd, M., 2011. Effect of ISRI stemming on similarity measure for arabic document clustering. In: The Asia Information Retrieval Societies Conference, pp. 584–593. Dubai, United Arab Emirates.
Buntine, W., 2009. Estimating likelihoods for topic models. In: Asian Conference on Machine Learning: Advances in Machine Learning. Springer-Verlag, Nanjing China, pp. 51–64.
Buntine, W., Jakulin, A., 2004. Applying discrete PCA in data analysis. In: Conference on Uncertainty in Artificial Intelligence. AUAI Press Arlington, Banff, Canada, pp. 59–66.
Buntine, W., Perttu, S., Tuulos, V., 2004. Using discrete PCA on Web pages. In: Proceedings of the Workshop Statistical Approaches to Web Mining, pp. 99–110. Pisa, Italy.
Chantar, H.K., Corne, D.W., 2011. Feature subset selection for arabic document categorization using BPSO-KNN. In: The World Congress on Nature and Biologically Inspired Computing. IEEE, Salamanca, Spain, pp. 546–551.
Chen, C.H., 2015. Feature selection for clustering using instance-based learning by exploring the nearest and farthest neighbors. Inf. Sci. (Ny) 318, 14–27.
Chen, H., 2008. IEDs in the dark Web: genre classification of improvised explosive device Web pages. In: IEEE International Conference on Intelligence and Security Informatics. IEEE, Taipei, Taiwan, pp. 94–97.
Cheng, W., Ni, X., Sun, J., Jin, X., 2011. Measuring opinion relevance in latent topic space. In: International Conference on Social Computing, Privacy, Security, Risk, and Trust. IEEE Computer Society, Minneapolis, Minnesota, USA, pp. 323–330.
Choi, D., Ko, B., Kim, H., Kim, P., 2014. Text analysis for detecting terrorism-related articles on the web. J. Netw. Comput. Appl. 38, 16–21.
Dai, P., Iurgel, U., Rigoll, G., 2003. A novel feature combination approach for spoken document classification with support vector machines. In: Multimedia Information Retrieval Workshop, pp. 1–5. Toronto, Canada.
Djallel DIlmi, M., Mallet, C., Barthes, L., Chazottes, A., 2017. Data-driven clustering of rain events: microphysics information derived from macro-scale observations. Atmos. Meas. Tech. 10, 1557–1574.
Do, T., Hui, S., 2006. Associative feature selection for text mining. Int. J. Inf. Technol. 12, 59–68.
Dong, H., Hui, S.C., He, Y., 2006. Structural analysis of chat messages for topic detection. Online Inf. Rev. 30, 496–516.
Dumais, S., Letsche, T., Littman, M., Landauer, T., 1997. Automatic cross-language retrieval using latent semantic indexing. AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval, 115–132.
El-beltagy, S.R., 2006. KP-miner: a simple system for effective keyphrase extraction. In: Innovations in Information Technology. IEEE Computer Society, Dubai, UAE, pp. 1–5.
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.
Farahat, A.K., Kamel, M.S., 2011. Statistical semantics for enhancing document clustering. Knowl. Inf. Syst. 28, 365–393.
Fejer, H.N., Omar, N., 2015. Automatic Arabic text summarization using clustering and keyphrase extraction. J. Artif. Intell. 8, 293–298.
Froud, H., Benslimane, R., Lachkar, A., Ouatik, S.A., 2010. Stemming and similarity measures for arabic documents clustering. In: International Symposium on Communications and Mobile Network. IEEE, pp. 1–4.
Froud, H., Lachkar, A., 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. Adv. Comput. An Int. J. 3, 55–67.
Froud, H., Lachkar, A., Ouatik, S., 2013a. Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. Int. J. Data Min. Knowl. Manage. Process 3, 79–95.
Froud, H., Sahmoudi, I., Lachkar, A., 2013b. 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, pp. 243–256. Dubai, UAE.
Gabrilovich, E., 2006. Feature Generation for Textual Information Retrieval Using World Knowledge. Israel Institute of Technology.
Ghanem, O., 2014. Evaluating the effect of preprocessing in arabic documents clustering. Islamic University, Gaza, Palestine.
Gharib, T.F., Fouad, M.M., Mashat, A., Bidawi, I., 2012. Self organizing map_based document clustering using WordNet ontologies. Int. J. Comput. Sci. 9, 88–95.
Ghwanmeh, S., 2005. Applying clustering of hierarchical K-means-like algorithm on arabic language. Int. J. Inf. Technol. 3, 467–471.
Gourav, B., 2011. Similarity measures of research papers and patents using adaptive and parameter free threshold. Int. J. Comput. Appl. 33, 9–13.
Gryc, W., 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. Vrije University, Amsterdam, pp. 47–70.
Guru, D.S., Harish, B.S., Manjunath, S., 2010. Symbolic representation of text documents. In: Third Annual ACM Bangalore Conference. ACM, New York, NY, pp. 1–4.
Habib, M.B., Fayed, Z.T., Gharib, T.F., 2006. A hybrid feature selection approach for arabic documents classification, Egypt. Comput. Sci. J. 28, 1–7.
Harrag, F., El-Qawasmah, E., Al-Salman, A.M.S., 2010. Comparing dimension reduction techniques for arabic text classification using BPNN algorithm. In: Integrated Intelligent Computing (ICIIC), pp. 6–11. Bangalore.
Hoenkamp, E., 2011. Trading spaces: on the lore and limitations of latent semantic analysis. In: Amati, G., Crestani, F. (Eds.), Advances in Information Retrieval Theory. Springer, Berlin Heidelberg, pp. 40–51.
Hotho, A., Staab, S., Stumme, G., 2003. Wordnet improves text document clustering. Semantic Web Workshop, 541–544.
Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., Chen, Z., 2008. Enhancing text clustering by leveraging Wikipedia semantics. In: The International Conference on Research and Development in Information Retrieval. ACM Press, New York, USA, pp. 179–188.
Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X., 2009. Exploiting Wikipedia as external knowledge for document clustering. In: International Conference On Knowledge Discovery And Data Mining. ACM, Paris, France, pp. 389–396.
Isa, D., Hong, L., Kallimani, V.P., RajKumar, R., 2009a. Text document pre-processing using the bayes formula for classification based on the vector space model. Comput. Inf. Sci. 1, 79–90.
Isa, D., Kallimani, V.P., Lee, L.H., 2009b. Using the self organizing map for clustering of text documents. Expert Syst. Appl. 36, 9584–9591.
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.
Jain, A., Murty, M., 1999. Data clustering: a review. ACM Comput. Surv. 31, 255–323.
Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall.
Jing, L., Ng, M.K., Huang, J.Z., 2010. Knowledge-based vector space model for text clustering. Knowl. Inf. Syst. 25, 35–55.
Jing, L., Yun, J., Yu, J., Huang, J., 2011. High-order co-clustering text data on semantics-based representation model. In: Advances in Knowledge Discovery and Data Mining, pp. 171–182. Shenzhen, China.
Kamde, P.M., Algur, D.S.P., 2011. A survey on Web multimedia mining. Int. J. Multimed. Its Appl. 3, 72–84.
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.
Landauer, T., Foltz, P., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Process. 25, 259–284.
Larkey, L., Ballesteros, L., Connell, M., 2007. Light stemming for arabic information retrieval. In: Soud, A., van den Bosch, A., Neumann, G. (Eds.), Arabic Computational Morphology. Springer, pp. 221–243.
Lee, C.-J., Hsu, C.-C., Chen, D.-R., 2017. A hierarchical document clustering approach with frequent itemsets. Int. J. Eng. Technol. 9, 174–178.
Lee, L.H., Rajkumar, R., Isa, D., 2012. Automatic folder allocation system using bayesian-support vector machines hybrid classification approach. Appl. Intell. 36, 295–307.
Leopold, E., Kindermann, J., 2002. Text categorization with support vector machines. How to represent texts in input space? Mach. Learn. 46, 423–444.
Lewis, D.D., 1990. Representation quality in text classification: an introduction and experiment. In: Workshop on Speech and Natural Language, pp. 288–295. Stroudsburg, PA, USA.
Li, N., Wu, D.D., 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis. Support Syst. 48, 354–368.
Li, R.Z., Zhang, Y..Sen., 2012. Study on the method of feature selection based on hybrid model for text classification. Adv. Mater. Res. 433–440, 2881–2886.
Li, Y., Luo, C., Chung, S., 2008. Text clustering with feature selection by using statistical data. IEEE Trans. Knowl. Data Eng. 20, 641–652.
Liu, L., Kang, J., Yu, J., Wang, Z., 2005. A comparative study on unsupervised feature selection methods for text clustering. In: International Conference on Natural Language Processing and Knowledge Engineering. IEEE, Wuhan, China, pp. 597–601.
Liu, T., Liu, S., Chen, Z., 2003. An evaluation on feature selection for text clustering. In: International Conference on Machine Learning. AAAI Press, Washington, DC, USA, pp. 488–495.
Liu, Y., Wu, C., Liu, M., 2011. Research of fast SOM clustering for text information. Expert Syst. Appl. 38, 9325–9333.
Lu, Y., Mei, Q., Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf. Retr. Boston. 14, 178–203.

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

Lucia, A..De., Risi, M., Tortora, G., 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. IEEE Computer Society, Washington, USA, pp. 65–72.

Ma, L., Fu, T., Blaschke, T., Li, M., Tiede, D., Zhou, Z., 2017. Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. ISPRS Int. J. Geo-Inf. 6, 21.

Marée, R., 2017. The need for careful data collection for pattern recognition in digital pathology. J. Pathol. Inf. 1, 1–4. Publ. by Wolters Kluwer-Medknow.

Mesleh, A., 2007a. Chi square feature extraction based svms arabic language text categorization system. J. Comput. Sci. 3, 430–435.

Mesleh, A.M., 2007b. Support vector machines based arabic language text classification system: feature selection comparative study. In: Int. Conf. On Applied Mathematics. Springer, Cairo, Egypt, pp. 11–16.

Mesleh, A.M., 2008. Support vector machines based Arabic language text classification system: feature selection comparative study. Adv. Comput. Inf. Sci. Eng., 11–16

Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In: The National Conference on Artificial Intelligence. American Association for Artificial Intelligence, Boston, Massachusetts, pp. 775–781.

Nagwani, N.K., Sharaff, A., 2015. SMS spam filtering and thread identification using bi-level text classification and clustering techniques. J. Inf. Sci. 43, 75–87.

Napoleon, D., Pavalakodi, S., 2011. A new method for dimensionality reduction using K-means clustering algorithm for high dimensional data set. Int. J. Comput. Appl. 13, 41–46.

Nath, S., 2006. Crime pattern detection using data mining. In: International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, Hong Kong, pp. 41–44.

Nuovo, D.I., Hirsch, L., Nuovo, A. Di, 2017. Document clustering with evolved search queries. In: Evolutionary Computation. IEEE, San Sebastián, Spain.

Osinski, S., 2004. Dimensionality Reduction Techniques For Search Results Clustering. The University of Sheffield.

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

Patel, D., Zaveri, M., 2011. A review on Web pages clustering techniques. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (Eds.), Trends in Network and Communications. Springer, Berlin Heidelberg, pp. 700–710.

Paulsen, J.R., Ramampiaro, H., 2009. Combining latent semantic indexing and clustering to retrieve and cluster biomedical information: a 2-step approach. In: NIK-2009 Conference. Trondheim, pp. 131–142.

Peachavanish, R., 2016. Stock selection and trading based on cluster analysis of trend and momentum indicators. Lect. Notes Eng. Comput. Sci. 1, 317–321.

Perkio, J., Buntine, W., Perttu, S., 2004. Exploring independent trends in a topic-based search engine. In: International Conference on Web Intelligence. IEEE Computer Society, Beijing, China, pp. 664–668.

Prasad, S., Bruce, L.M., 2008. Limitations of principal components analysis for hyperspectral target recognition. Geosci. Remote Sens. Lett. IEEE 5, 625–629.

Qi, X., Christensen, K., Duval, R., Fuller, E., Spahiu, A., Wu, Q., Zhang, C.-Q., 2010. A hierarchical algorithm for clustering extremist Web pages. In: International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, Odense, Denmark, pp. 458–463.

Rajeswari, R.P., Juliet, K., 2017. Text classification for student data set using naive bayes classifier and KNN classifier. Int. J. Comput. Trends Technol. 43, 8–12.

Sabbah, T., Selamat, A., Selamat, M.H., Ibrahim, R., Fujita, H., 2016a. Hybridized term-weighting method for Dark Web classification. Neurocomputing 173, 1908–1926.

Sabbah, T., Selamat, A., Selamat, M.H., Ibrahim, R., Fujita, H., 2016b. Hybridized term-weighting method for dark Web classification. Neurocomputing 173, 1908–1926.

Sahmoudi, I., Froud, H., Lachkar, A., 2013. A new keyphrases extraction method based on suffix tree data structure. Int. J. Database Manag. Syst. (IJDMS) 5, 17–33.

Sahmoudi, I., Lachkar, A., 2016. Towards a linguistic patterns for arabic keyphrases extraction. In: International Conference on Information Technology for Organizations Development. IEEE, Fez, Morocco, pp. 1–6.

Sahoo, G., Pawar, P., Malvi, K., Jaladi, A., Khithani, K., 2017a. E-mail spam detection with speech tagging. Int. J. Innov. Res. Comput. Commun. Eng. 5, 8669–8674.

Sahoo, G., Pawar, P., Malvi, K., Jaladi, A., Khithani, K., 2017b. Analysis of customer churn by big data clustering. Int. J. Innov. Res. Comput. Commun. Eng. 5, 6157–6162.

Said, D.A., Wanas, N.M., Darwish, N.M., Hegazy, N.H., 2009. A study of text preprocessing tools for arabic text categorization. In: International Conference on Arabic Language Resources and Tools. The MEDAR Consortium, Cairo, Egypt, pp. 330–336.

Salehi, F., Ahmadi, A., 1993. Principal components analysis. In: Methods in Molecular Biology. Idea Group Inc, pp. 527–547.

SAM, L.Z., 2009. Enhanced Feature Selection Method for Illicit Web Content Filtering. Universiti Teknologi Malaysia.

Selamat, A., Lee, Z.S., Maarof, M.A., Shamsuddin, S.M., 2011. Improved web page identification method using neural networks. Int. J. Comput. Intell. Appl. 10, 87.

Seo, Y., Ankolekar, A., Sycara, K., 2004. Feature Selection for Extracting Semantically Rich Words.

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

Sharma, S., Gupta, V., 2012. Recent developments in text clustering techniques. Int. J. Comput. Appl. 37, 14–19.

Simanjuntak, D.A., Ipung, H.P., Lim, C., Nugroho, A.S., 2010. Text classification techniques used to facilitate cyber terrorism investigation. In: International Conference on Advances in Computing, Control, and Telecommunication Technologies. IEEE Computer Society, Jakarta, Indonesia, pp. 198–200.

Sriurai, W., 2011. Improving text categorization by using a topic model. Adv. Comput. An Int. J. 2, 21–27.

Steinbach, M., Karypis, G., Kumar, V., others, 2000. A comparison of document clustering techniques. In: World Text Mining Conference. Boston, pp. 525–526.

Stoica, E.A., Özyirmidokuz, E.K., 2015. Mining customer feedback documents. Int. J. Knowl. Eng. 1, 68–71.

Sun, J., Wang, X., 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, pp. 12–16.

Sutar, S., 2015. Feature selection algorithm using fast clustering and correlation measure. Int. Res. J. Eng. Technol. 2, 236–241.

Syiam, M., Fayed, Z., 2006. An intelligent system for arabic text categorization. Int. J. Intell. Comput. Inf. Sci. 6, 1–19.

Ho, T.B., Kawasaki, S., Nguyen, N.B., 2003. Documents clustering using tolerance rough set model and its application to information retrieval. In: Szczepaniak, P. S., Segovia, J., Kacprzyk, J., Zadeh, L.A. (Eds.), Studies In Fuzziness And Soft Computing: Intelligent Exploration of the Web. Physica-Verlag HD, pp. 181–196.

Thabtah, F., Eljinini, M., Zamzeer, M., Hadi, W., 2009. Naive Bayesian based on Chi Square to categorize Arabic data. Commun. IBIMA 10, 4–6.

Thanh, N., Yamada, K., 2011. Document representation and clustering with WordNet based similarity rough set model. Int. J. Comput. Sci. 8, 1–8.

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

Tsai, C.-F., Hsiao, Y.-C., 2010. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decis. Support Syst. 50, 258–269.

Turney, P., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.

Wallach, H., 2006. Topic modeling: beyond bag-of-words. In: International Conference on Machine Learning. ACM, Pittsburgh, USA, pp. 977–984.

Wang, S., Koopman, R., 2017. Clustering articles based on semantic similarity. Scientometrics 111, 1017–1031.

Wang, S., Wei, Y., Li, D., Zhang, W., Li, W., 2007. A hybrid method of feature selection for Chinese text sentiment classification. In: International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Computer Society, Haikou, China, pp. 435–439.

Wilbur, W., Sirotkin, K., 1992. The automatic identification of stop words. J. Inf. Sci. 18, 45–55.

Xu, C., Zhang, Y., Zhu, G., Rui, Y., Lu, H., Huang, Q., 2008. Using webcast text for semantic event detection in broadcast sports video. IIEEE Trans. Multimed. 10, 1342–1355.

Xu, Y., Chen, L., 2010. Term-frequency based feature selection methods for text categorization. In: International Conference on Genetic and Evolutionary Computing. IEEE Computer Society, Shenzhen, China, pp. 280–283.

Yakut, I., Turkoglu, T., Yakut, F., 2015. Understanding customer's evaluations through mining airline reviews. Int. J. Data Min. Knowl. Manage. Process 5, 1143–1148.

Yang, F., Hamit, M., Yan, C.B., Yao, J., Kutluk, A., Kong, X.M., Zhang, S.X., 2017. Feature extraction and classification on esophageal X-ray images of Xinjiang Kazak Nationality. J. Healthc. Eng., 1–11

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. ACM, Seattle, USA, pp. 256–263.

Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization. In: International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 412–420.

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224.

Zahran, B.M., Kanaan, G., 2009. Text feature selection using particle swarm optimization algorithm. World Appl. Sci. J. Special Issue Comput. IT 7, 69–74.

Zhang, W., Yoshida, T., Tang, X., 2008. Text classification based on multi-word with support vector machine. Knowledge-Based Syst. 21, 879–886.

Zhang, W., Yoshida, T., Tang, X., Wang, Q., 2010. Text clustering using frequent itemsets. Knowledge-Based Syst. 23, 379–388.

Zhang, Y., Moges, S., Block, P., 2017. Does objective cluster analysis serve as a useful precursor to seasonal precipitation prediction at local scale? Application to western Ethiopia. Hydrol. Earth Syst. Sci. Discuss., 1–18

Zhang, Y., Zhang, Q., 2006. A text classifier based on sentence category VSM. In: The Pacific Asia Conference on Language. Information and Computation. Wuhan, China, pp. 244–249.

Zhiwei, M., Singh, M.M., Zaaba, Z.F., 2017. Email spam detection: a method of metaclassifiers stacking. In: The 6th International Conference on Computing and Informatics, pp. 750–757. Kuala Lumpur, Malaysia.

Zhou, X.F., Liang, J.G., Hu, Y., Guo, L., 2014. Text document latent subspace clustering by PLSA factors. In: International Joint Conferences on Web Intelligence and Intelligent Agent Technologies. IEEE, Warsaw, Poland, pp. 442–448.

Zhou, Y., Yang, Y., Peng, W., Ping, Y., 2010. A novel term weighting scheme with distributional coefficient for text categorization with support vector machine.

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002

In: The Conference on Information Computing and Telecommunications. IEEE, Beijing, China, pp. 182–185.

Zitouni, A., Damankesh, A., Barakati, F., Atari, M., Watfa, M., Oroumchian, F., 2010. Corpus-based arabic stemming using N-grams. In: The Sixth Asia Information Retrieval Societies Conference. Springer, Berlin Heidelberg, Taipei, Taiwan, pp. 280–289.

Further reading

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

Please cite this article in press as: Alghamdi, H.M., Selamat, A. Arabic Web page clustering: A review. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.06.002