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

King Saud University

Journal of King Saud University – Computer and Information Sciences

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

Arabic web pages clustering and annotation using semantic class features

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

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

Available online 28 September 2014

KEYWORDS

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

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

1. Introduction

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

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

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

Production and hosting by Elsevier

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

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

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

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

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

2. Related works

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

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

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

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

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

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

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

3. Proposed model

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

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

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