Enhanced Filter Feature Selection Methods for Arabic Text Categorization

Abdullah Saeed Ghareb, The National University of Malaysia, Bangi, Malaysia
Azuraliza Abu Bakara, The National University of Malaysia, Bangi, Malaysia
Qasem A. Al-Radaideh, Yarmouk University, Jordan, Jordan
Abdul Razak Hamdan, The National University of Malaysia, Bangi, Malaysia

ABSTRACT

The filtering of a large amount of data is an important process in data mining tasks, particularly for the categorization of unstructured high dimensional data. Therefore, a feature selection process is desired to reduce the space of high dimensional data into small relevant subset dimensions that represent the best features for text categorization. In this article, three enhanced filter feature selection methods, Category Relevant Feature Measure, Modified Category Discriminated Measure, and Odd Ratio2, are proposed. These methods combine the relevant information about features in both the inter- and intra-category. The effectiveness of the proposed methods with Naïve Bayes and associative classification is evaluated by traditional measures of text categorization, namely, macro-averaging of precision, recall, and F-measure. Experiments are conducted on three Arabic text datasets used for text categorization. The experimental results showed that the proposed methods are able to achieve better and comparable results when compared to 12 well known traditional methods.

KEYWORDS
Arabic Text Categorization, Associative Classification, Feature Selection, Naïve Bayes

1. INTRODUCTION

The volume of text in its digital form is continuously increasing, and it has become an area of research interest to investigate and construct new techniques and methods that can handle and control this considerable amount of text. Text categorization techniques have become one of the key technologies for text pattern recognition and text organization; however, these techniques struggle with the large dimensionality of text because there is a huge number of text features. Therefore, the feature selection (FS) process can be of benefit to categorization techniques, and it is used as a categorization pre-process for dimensionality reduction and relevant feature selection that can simplify the categorization process and identify precisely the pattern of text features (Mesleh, 2011; Uguz, 2011).

Text categorization is a forecasting process of text document categories that categorizes text documents into one/multiple pre-defined categories based on the extracted knowledge from text documents (Manning and Schutze, 1999). In recent years, many categorization methods were proposed for categorizing text of different languages such as: K-nearest neighbour (KNN) (Abu Tair & Baraka, 2013; Jiang et al., 2012), support vector machine (SVM) (Joachims, 1998; Mesleh, 2011), naïve Bayes (NB) (Chen et al., 2009; Hattab & Hussein, 2013) decision tree (Harrag et al., 2010), and associative

DOI: 10.4018/IJIRR.2018040101

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
classification (AC) (Al-Radaideh et al., 2011; Chaing et al., 2008; Ghareb et al., 2012; Srvidhya & Anitha, 2009; Khorsheed & Al-Thubaity, 2013).

The high dimensionality of the feature space is a major problem in Arabic text categorization due to the existence of noisy and irrelevant features, which adversely affect the categorization performance and degrades computer resource (Abu Tair & Baraka, 2013; Sharef et al., 2012; Zahran & Kanaan, 2009). Therefore, the feature selection process is frequently used to reduce the high dimensionality of text and select the most informative features for Arabic text categorization. It can be defined as “a process that chooses an optimal subset of features according to certain criterion” (Liu & Motoda, 1998). Feature selection techniques can be categorized into groups based on their functionality for feature evaluation. As cited in the literature, feature selection methods for Arabic text categorization can be divided into filtering feature selection and meta-heuristic/wrapper approaches. The filter feature selection methods are frequently used for Arabic text categorization because of their efficiency and simplicity. They include chi-square, odd ratio, document frequency, information gain, and many others (Mesleh, 2011; Harrag et al., 2010).

This paper is intended to enhance Arabic text categorization on two dimensions: text representation and dimensionality reduction with feature selection. With respect to these dimensions, the lack of important Arabic words identification and the intensive text representation are observed problems in related works that applied for the Arabic language. The related work investigates how to represent Arabic text documents in an efficient way for Arabic text categorization and how to represent a compact feature vector that describes the important content of the text document. The major challenges are the intensive text representation and the lack of important features identification that efficiently describe the content of Arabic text documents and contribute to improving Arabic text categorization. Therefore, the consideration of Arabic word categories and their strength should be considered as a potential way to avoid such problems and represent a summary of Arabic text documents without affecting the important content, which can then help to improve the feature selection and categorization tasks. Also, due to the high dimensionality of Arabic textual data, mainly due to the complex structure of the Arabic language, feature selection has the potential to bring about the dimensionality reduction that is required to accelerate the learning process and make the categorization results more accurate by removing irrelevant and noisy information. Thus, proposing new or modifying available feature selection methods to develop more efficient Arabic text categorization model is an important issue in text mining research.

In this paper, the objective of the proposed methods is to construct a knowledge-based model that can categorize Arabic text documents into their proper category based on knowledge extraction in the construction phase. The proposed methods enhanced two well-known categorization techniques (NB and AC) from the feature selection perspective: the large amount of features in the training dataset is treated by using feature selection and ranking measures, where each feature is given a score to determine its importance in the training dataset and its relevancy for the text categorization process, and then a subset of relevant features is selected based on its ranking to be used for categorization learning. In this stage, we enhance the feature subset selection process by introducing new and modified feature selection measures that combine the useful information about text features. The new feature selection method is called the category relevant feature measure (CRFM) and a modification is applied to enhance two well-known feature selection methods, the class discriminating measure (CDM) (Chen et al., 2009) and the odd ratio (OR) (Mladenic & Grobelnik, 1999). We modified these measures because they have some limitations, particularly in respect of numerical errors and dependency on document frequency; the enhanced measures are called the modified class discriminating measure (MCDM) and OR2, respectively. The justifications for doing this enhancement are discussed in section 3.

Moreover, 12 feature selection methods are selected from relevant related works and compared against the proposed feature selection methods. These methods are: information gain (IG), Chi-Square (CHI), mutual information (MI), document frequency (DF) (Yang & Pedersen, 1997), term frequency-inverse document frequency (TF-IDF) (Salton & Buckley, 1988; GSS Galavotti et al., 2000), precision
Related Content

Semantic Mediation in MedPeer: An Ontology-Based Heterogeneous Data Sources Integration System
www.igi-global.com/chapter/semantic-mediation-in-medpeer/198643?camid=4v1a

Comparative Study Between a Swarm Intelligence for Detection and Filtering of SPAM: Social Bees vs. Inspiration From the Human Renal
www.igi-global.com/chapter/comparative-study-between-a-swarm-intelligence-for-detection-and-filtering-of-spam/197694?camid=4v1a
Fuzzy Opinion: Detection of Opinion Based on SentiWordNet Dictionary by Using Fuzzy Logic
[www.igi-global.com/article/fuzzy-opinion/161658?camid=4v1a](www.igi-global.com/article/fuzzy-opinion/161658?camid=4v1a)

An Efficient Innovative Approach Towards Color Image Enhancement
[www.igi-global.com/article/an-efficient-innovative-approach-towards-color-image-enhancement/193247?camid=4v1a](www.igi-global.com/article/an-efficient-innovative-approach-towards-color-image-enhancement/193247?camid=4v1a)