Decision Support Systems 57 (2014) 64–76

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Exploiting poly-lingual documents for improving text categorization effectiveness

Chih-Ping Wei [a], Chin-Sheng Yang [b,*], Ching-Hsien Lee [a], Huihua Shi [c], Christopher C. Yang [d]

[a] *Department of Information Management, National Taiwan University, Taipei, Taiwan, ROC*
[b] *Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, ROC*
[c] *Infrastructure & System Department I, Information Technology Division (AUT), AU Optronics Corporation, Hsinchu Science Park, Hsinchu, Taiwan, ROC*
[d] *College of Computing and Informatics, Drexel University, Philadelphia, PA, USA*

ARTICLE INFO

Article history:
Received 17 April 2012
Received in revised form 2 July 2013
Accepted 2 August 2013
Available online 15 August 2013

Keywords:
Text mining
Text categorization
Poly-lingual text categorization
Feature reinforcement
Document management

ABSTRACT

With the globalization of business environments and rapid emergence and proliferation of the Internet, organizations or individuals often generate, acquire, and then archive documents written in different languages (i.e., poly-lingual documents). Prevalent document management practice is to use categories to organize this ever-increasing volume of poly-lingual documents for subsequent searches and accesses. Poly-lingual text categorization (PLTC) refers to the automatic learning of text categorization models from a set of preclassified training documents written in different languages and the subsequent assignment of unclassified poly-lingual documents to predefined categories on the basis of the induced text categorization models. Although PLTC can be approached as multiple, independent monolingual text categorization problems, this naïve PLTC approach employs only the training documents of the same language to construct a monolingual classifier and thus fails to exploit the opportunity offered by poly-lingual training documents. In this study, we propose a feature-reinforcement-based PLTC (FR-PLTC) technique that takes into account the training documents of all languages when constructing a monolingual classifier for a specific language. Using the independent monolingual text categorization (MnTC) approach as a performance benchmark, the empirical evaluation results show that our proposed FR-PLTC technique achieves higher classification accuracy than the benchmark technique. In addition, our empirical results suggest the superiority of the proposed FR-PLTC technique over its counterpart across a range of training sizes.

1. Introduction

With advances in and the proliferation of information and networking technologies, organizations increasingly participate in or shift to the Internet environment to conduct business transactions, gather marketing and competitive intelligence information from various online sources, and facilitate information and knowledge sharing within or beyond organizational boundaries. Such e-commerce and knowledge management applications generate and maintain a tremendous amount of textual documents (or documents for short) in organizational repositories. To facilitate subsequent access to these documents, the use of categories to manage the ever-increasing volume of documents often occurs at both organizational and individual levels. Text categorization deals with the assignment of documents to appropriate categories on the basis of their contents [2,7,8,40]. Central to text categorization is the automatic learning of a text categorization model using a set of preclassified documents that serve as training examples. The induced text categorization model then can classify (or predict) the particular category (or categories) to which a new document belongs.

Various text categorization techniques have been proposed [1,2,7,8,18,19,22,34,39,40,42], but most of them focus on monolingual documents (i.e., all documents written in the same language) for both learning a text categorization model and assigning (or predicting) new documents into appropriate categories. Because of the trend of globalization, an organization or individual often generates, acquires, and then archives documents written in different languages (i.e., poly-lingual documents). Assume that the languages involved in a repository include L_1, L_2, …, L_s, where $s \geq 2$. That is, the set of poly-lingual documents contains some documents in L_1, some in L_2, …, and some in L_s. Consider the following scenarios: A division in a multinational corporation receives poly-lingual documents from other divisions and uses them in its routine activities. Or a financial analyst scrutinizing the global investment market needs to collect and archive financial reports and news that are in effect poly-lingual. Such scenarios are even more prevalent in countries with more than one official language. For example, Chinese and English are official languages of Hong Kong; French and English for Canada; Chinese, Malay, Tamil, and English for Singapore; and Dutch, French, and German for Belgium. In such poly-lingual environments, if organizations or individuals have already organized their poly-lingual

* Corresponding author. Tel.: +886 3 463 8800x2799; fax: +886 3 435 2077.
E-mail addresses: cpwei@im.ntu.edu.tw (C.-P. Wei), csyang@saturn.yzu.edu.tw (C.-S. Yang), chinghsienlee@gmail.com (C.-H. Lee), Sally.Shih@auo.com (H. Shi), chris.yang@drexel.edu (C.C. Yang).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.08.001

documents into existing categories and want to use this set of preclassified documents as a training set to construct text categorization models and then to classify newly received poly-lingual documents into appropriate categories, they face a poly-lingual text categorization (PLTC) problem.

Formally, PLTC pertains to learning text categorization models from a training set of poly-lingual documents (some written in language L_1, some in L_2, …, and some in L_s, where $s \geq 2$, but each training document only written in one language), each of which is preclassified into a predefined category (C_1, C_2, …, or C_n), and then assigning unclassified documents written in L_1, …, or L_s into appropriate categories. Because of the availability of training documents in each language, PLTC can be approached simply as multiple, independent monolingual text categorization problems. That is, given a set of poly-lingual preclassified documents as training examples, we can construct a monolingual text categorization model (i.e., classifier) for each language on the basis of the training examples of that respective language. When a new document written in a specific language arrives, we select the corresponding classifier to predict the appropriate category for the target document. However, this naïve PLTC approach employs only the training documents of the same language to construct a monolingual classifier and ignores all training documents of other languages. Thus, if the training documents of a language (e.g., L_i) are less representative of the target semantic space of the predefined categories, especially when the training set of the language is small-sized, the effectiveness of the induced classifier for that language would not be satisfactory. If the representativeness of the training documents of another language (e.g., L_j) with respect to the predefined categories is greater than that of the training documents of L_i, the training documents of L_j can be beneficial to improve the effectiveness of the classifier for L_i. Because the naïve PLTC approach constructs each monolingual classifier independently, it fails to exploit this opportunity offered by poly-lingual training documents. In addition, if the training documents of the target language (e.g., L_i) contain features (i.e., terms) that are non-semantic-bearing for the predefined categories, the inclusion of those noisy features into the induced classifier will degrade its effectiveness. However, terms in another language (e.g., L_j) that linguistically correspond to the noisy features in L_i may not have any discriminating power according to the training documents of L_j. In this case, a cross-check of features occurring in the training documents of different languages (i.e., referred to as feature reinforcement in this study) can help remove the noisy features of individual languages and hence improve the effectiveness of the classifier for each language.

Most existing text categorization techniques deal with monolingual text categorization [1,2,7,8,18,19,22,27,34,39,40,42], and some prior studies address the challenge of cross-lingual text categorization (i.e., learning from a set of training documents written in one language and then classifying new documents in a different language) [4,9,23,24,33]. However, prior research has not paid much attention to PLTC. This study therefore is motivated by the importance of providing PLTC support to organizations and individuals in increasingly global, multilingual environments. Specifically, we propose a feature-reinforcement-based PLTC (FR-PLTC) technique that takes into account the training documents of all languages when constructing a monolingual classifier for a specific language. For the purposes of our intended feasibility assessment and illustration, this study concentrates on only two languages involved in poly-lingual documents. To support linguistic interoperability between training documents in different languages, we rely on a statistical-based bilingual thesaurus, automatically constructed from a collection of parallel documents. Experimentally, we evaluate the effectiveness of our proposed FR-PLTC technique using independent monolingual classifiers built by the aforementioned naïve PLTC approach.

The remainder of this article is organized as follows: In Section 2, we review literature relevant to this study, including existing monolingual, poly-lingual, and cross-lingual text categorization techniques. We depict the detailed development of our proposed FR-PLTC technique in Section 3, including the overall processes and specific designs. Subsequently, we describe the evaluation design and discuss important experimental results in Section 4. Finally, we conclude with a summary and some further research directions in Section 5.

2. Literature review

2.1. Monolingual text categorization techniques

Text categorization refers to the assignment of documents, on the basis of their contents, to one or more predefined categories. Many text categorization techniques have been proposed in the literature [1,2,7,8,18,19,22,27,34,39,40,42] but most of them focus on monolingual documents. Central to text categorization is the automatic learning of a text categorization model from a training set of preclassified documents. The resulting model will then be used to classify or predict the particular category or categories to which a new, unclassified document belongs. The process of (monolingual) text categorization generally includes three main steps: 1) feature extraction and selection, 2) document representation, and 3) induction [1,32].

Feature extraction extracts or identifies terms (or features) from the training documents. Different languages exhibit different grammatical and lexical characteristics that affect how the features in documents are segmented. For example, there exist prominent differences between European languages (e.g., English) and Oriental languages (e.g., Chinese). Term extraction of English documents typically involves lexical analysis, stopword removal, stemming, and term-phrase formation [37]. However, no natural delimiter in the Chinese language marks word boundaries. Additional mechanism, such as the lexical rule-based or the statistical approach, is required to support lexical analysis and term-phrase segmentation for Chinese documents [38].

Following extraction is feature selection, which reduces the size of the feature space, a process that not only improves learning efficiency but also potentially improves learning effectiveness by suppressing potential biases embedded in the original (i.e., non-condensed) feature set [8]. According to the top-k selection method, commonly used in prior research, the k features with the highest selection metric scores are selected to represent each training document. However, previous research varies considerably in the underlying metric used for feature selection. Common examples include TF (term frequency), TF × IDF (IDF denotes inverse document frequency), correlation coefficient, χ^2 metric, and mutual information [8,17,19,22,26].

In the document representation step, each document is represented by a vector space jointly defined by the top-k features selected in the previous step and labeled to indicate its category membership. A review of prior research suggests the prevalence of several representation methods, such as binary (which indicates the presence or absence of a feature in a document), within-document TF, and TF × IDF. Finally, in the induction step, a text categorization model(s) that distinguishes categories from one another on the basis of the set of training documents is constructed. Prevalent supervised learning techniques employed for text categorization include decision-tree induction [34], decision-rule induction [2,7], k-nearest neighbor (kNN) classification [12,18,20,39], neural network [22,35], the Naïve Bayes probabilistic algorithm [1,3,18,19,21], and SVM [8,14,41]. Empirical evaluations of different supervised learning strategies for text categorization can be found in the studies by Sebastiani [27] and Yang and Liu [41].

2.2. Poly-lingual and cross-lingual text categorization techniques

As an emerging research topic, the literature related to PLTC remains very limited. Bel et al. [4] assume that each training document is available in two different languages (i.e., parallel document) and a newly arrived (i.e., unclassified) document is available in one or both languages. Accordingly, they simply construct a single classifier that encompasses terms from both English and Spanish as its features. However, their