Applied Mathematics and Computation 240 (2014) 308–325

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

A keyword extraction method from twitter messages represented as graphs

Willyan D. Abilhoa *, Leandro N. de Castro

Mackenzie Presbyterian University, Natural Computing Laboratory, São Paulo, Brazil

ARTICLE INFO

Keywords:
Knowledge discovery
Text mining
Keyword extraction
Graph theory
Centrality measures
Twitter data

ABSTRACT

Twitter is a microblog service that generates a huge amount of textual content daily. All this content needs to be explored by means of text mining, natural language processing, information retrieval, and other techniques. In this context, automatic keyword extraction is a task of great usefulness. A fundamental step in text mining techniques consists of building a model for text representation. The model known as vector space model, VSM, is the most well-known and used among these techniques. However, some difficulties and limitations of VSM, such as scalability and sparsity, motivate the proposal of alternative approaches. This paper proposes a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures for finding the relevant vertices (keywords). To assess the performance of the proposed approach, three different sets of experiments are performed. The first experiment applies TKG to a text from the Time magazine and compares its performance with that of the literature. The second set of experiments takes tweets from three different TV shows, applies TKG and compares it with TFIDF and KEA, having human classifications as benchmarks. Finally, these three algorithms are applied to tweets sets of increasing size and their computational running time is measured and compared. Altogether, these experiments provide a general overview of how TKG can be used in practice, its performance when compared with other standard approaches, and how it scales to larger data instances. The results show that TKG is a novel and robust proposal to extract keywords from texts, particularly from short messages, such as tweets.

1. Introduction

Social media refers to a set of different web sites that allow users to create, share and exchange content, such as social network sites, blogs, microblogs, video shares, bookmarks, among others [1–5]. The content generated is important in many research areas because there is information about various subjects in different contexts created from the users' point of view. Examples of applications of social media data mining include helping individuals and organizations to discover the acceptance level of products [6], the detection of disasters and anomalies [7], the forecast of the performance of politicians in election campaigns [8], the monitoring of diseases [9], to name but a few potential applications.

When the database is composed of written documents, methods based on text mining [10–14], natural language processing [15–17], and information retrieval [18–21] are usually applied. In the specific case of text mining approaches,

* Corresponding author.
E-mail addresses: abilhoa.willyan@gmail.com (W.D. Abilhoa), lnunes@mackenzie.br (L.N. de Castro).

http://dx.doi.org/10.1016/j.amc.2014.04.090
0096-3003/© 2014 Elsevier Inc. All rights reserved.

documents are represented using the well-known vector space model [22], which results in sparse matrices to be dealt with computationally. Besides, when the target application involves Twitter messages, as is the case of the present research, this problem becomes even worse. Due to the short texts (140 characters), informality, grammatical errors, buzzwords, slangs, and the speed with which real-time content is generated, approximately 250 million messages posted daily [23], effective techniques are required [24,25].

Keyword extraction is the task of finding the words that best describe the subject of a text. Its applications include indexing, summarization, topic detection and tracking, among others [26–44]. This paper proposes a technique to extract keywords from collections of Twitter messages based on the representation of texts by means of a graph structure [27–31], from which it is assigned relevance values to the vertices based on graph centrality measures [27,33]. The proposed method, named TKG, is assessed using three different sets of experiments. First, it is applied to a text from the Time magazine and compared with the results from the literature. Then, tweets from three different TV shows are taken, and TKG is applied and compared with the TFIDF method and the KEA algorithm, having human-based keyword extraction as benchmark. Several variations of TKG are proposed, including different forms of building the graph connections and weighting them, plus different centrality measures to extract keywords from the graphs. Finally, the three algorithms are applied to tweets of increasing size and their computational running time is measured and compared. These experiments were designed so as to provide a general overview of how TKG can be used in practice, its performance when compared with other standard approaches, and how it scales to larger data instances. The experiments performed showed that some variations of TKG are invariably superior to others and the other algorithms for the problems tested. Also, it was observed that most TKG variations scale almost linearly with the size of the data set, contrary to the other approaches that scaled exponentially.

The paper is organized as follows. Section 2 gives provides a gentle introduction to about the problems of text representation and keyword extraction, and briefly reviews the main works from the literature on text representation by means of graphs and keyword extraction. Section 3 introduces the proposed technique, and Section 4 covers the experiments performed, results obtained and discussions. The paper is concluded in Section 5 with a general discussion about the work and perspectives for future research.

2. Text representation and keyword extraction: a brief overview

This paper proposes a graph-based text representation for keyword extraction from tweets. To achieve this goal, concepts from text mining, graph-based document representation, centrality measures in graphs, and the keyword extraction problem have to be understood. In addition to providing some basics of all these subjects, this section provides a brief review of the main works from the literature that propose similar approaches.

2.1. The vector space model

In many text mining applications, the well-known model for text representation, called *vector-space model* (VSM), is the most used [10–14]. The VSM consists of building a numerical matrix **M** in which lines correspond to the vector form of documents, and columns correspond to the words from a dictionary. Thus, given a set of N documents $D = \{\mathbf{d}_1,\ldots,\mathbf{d}_N\}$ and a dictionary of K words, or tokens, $T = \{t_1,\ldots,t_K\}$, a weight w_{ij} is assigned to the element $m_{ij} \in \mathbf{M}_{N\times K}$, $i = 1,\ldots,N$; $j = 1,\ldots,K$. This weight assumes a value usually related to the frequency of a given word in a document or set of documents. The most commonly used frequencies are the *binary*, *absolute*, *relative* and *weighted*. The binary frequency assumes a value 1 if a word occurs in a document, otherwise it assumes 0. The absolute frequency is the number of occurrences of a word in a document. The relative frequency, TF_{ti}, corresponds to the absolute frequency of a word t divided by the maximal absolute frequency of any word in the document, Max f_{zi}:

$$\mathrm{TF}_{ti} = \frac{f_{ti}}{(\operatorname*{Max}_{z} f_{zi})}. \tag{1}$$

The weighted frequency TF-IDF, or simply TFIDF, is the product between the relative frequency TF_{ti} and a modifying value called *inverse document frequency*, denoted by IDF_t, which weights the importance of a word by its frequency for an individual document and the overall document collection. Therefore, a word that appears in many documents has a potentially lower weight than a word that appears in certain distinct documents:

$$\begin{aligned} &\mathrm{TF}-\mathrm{IDF} = \mathrm{TF}_{ti} \times \mathrm{IDF}_t \\ &\mathrm{IDF}_t = \log\left(\frac{N}{\mathrm{DF}_t}\right) \end{aligned} \tag{2}$$

where DF_t is the number of documents containing a word t, and N is the number of documents in the collection.

In the VSM model, matrix **M** needs to be rebuilt whenever a new word has to be included in the dictionary, because **M** gains a new column with the weight corresponding to this word. This may be a problem when working with continuously incoming data, such as those provided by social media services. In such cases, instead of working with numerical matrices, one possibility is to represent texts as graphs, which enable to capture key text features in terms of frequency and relationships of words.