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/

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, $\mathrm{Max}\, f_{zi}$:

$$\mathrm{TF}_{ti} = \frac{f_{ti}}{(\max_z f_{zi})}. \tag{1}$$

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.

2.2. Graph-based text representation

A graph $G = (V,E)$ can be defined as a structure that specifies relationships between the elements of a collection, where V corresponds to the set of elements, called *vertices* or *nodes*, and E is the set of relations among vertices, called *edges*. In graph-based text representation, a text can be seen as a collection of interconnected words [28–32]. Thus, a textual graph is a graph in which the set of vertices corresponds to the collection of terms, and the set of edges provides the relationships among terms obeying some criteria, such as the proximity between the terms or the existence of a given relationship of these terms in the same sentence, such as a co-occurrence relationship. Some of main graph-based models to represent texts include the works of [28–32]. Among these works, the work of [28] is discussed in more detail below, for it directly inspired the technique proposed in this paper due to its simplicity and the capacity to incorporate contextual information.

The textual graph proposed by [28] captures information about the order, frequency and context of terms. To capture the notion of order, lists of pre-processed terms in the same order in which they appear in the original document are obtained. In the textual graph formed, each vertex corresponds to a term or concept, and is represented in a unique way. Each edge is constructed based on the proximity and co-occurrence relationships between concepts within a given window, which is moved along the list. The capture of frequency information in the co-occurrence relationships between concepts is based on the procedure described previously. Their fundamental difference is the assignment of a weight to the edges as the graph is built. This weight is based on the frequency with which two concepts occur together (co-occurrence), and the frequency of individual occurrence in each concept. It is possible to assign the context information to the concepts from the order of terms retained and their co-occurrence frequency. Edges establish bidirectional relationships between vertices, being assigned a transition probability in each direction [28].

2.3. Centrality measures

Centrality measures are discriminative properties of the importance of a vertex in a graph, and are directly related to the structure of the graph. The identification of central vertices is a major task in graph analysis [33]. In this context, the notion of importance or centrality of a vertex can be related to the *degree*, *closeness*, *eccentricity*, and many other features of a given vertex [33].

The *degree centrality* (C_i^D) defines the importance of a vertex i according to the number of edges incident upon it [33]. This number quantifies the interaction level of an element with others in the network. This measurement value can be obtained by $C_i^D = deg_i$, where deg_i is the number of edges connected to vertex i.

The *closeness centrality* (C_i^C) of a vertex i is defined as the inverse of its *farness*, where the fairness of a vertex i is defined as the sum of its distance to all other vertices. Therefore, the more central a vertex, the lower its total distance to all other vertices [34]:

$$C_i^C = \frac{1}{\sum_{i,j\in V} d_{ij}}, \tag{3}$$

where d_{ij} is the shortest path between i and j.

Analogously to the closeness, there is also the *eccentricity centrality* (C_i^E), which defines the distance d_{ij} between a vertex i and the other vertices based on the largest geodesic distance [35]. A high geodesic distance represents a greater eccentricity, as in

$$C_i^E = \max_{i,j\in V}\{d_{ij}\}. \tag{4}$$

Since a text is represented by a graph and centrality measures are a possibility for highlighting important nodes (terms) according to the structure of the graph, these are used here as weighting criteria for words in a document.

2.4. Keyword extraction

A *keyword* can be understood as the smallest unit of one or more terms which summarizes and identifies the subject of a text [36–38]. One of the main tasks in the analysis of large textual databases is the discovery of keywords within a document collection, which is a process known as *keyword extraction*. This task is a core technology for many text mining applications, such as indexing, summarization, classification, clustering, topic detection and tracking, among others [10–14]. The simplest, and probably the most effective, solution for this task is the assignment of keywords by a person that reads the full document. However, when a massive number of documents is continuously received, such as in Twitter tracking applications, humanly reading tweets in search for keywords becomes an unfeasible and error prone task [38].

In this context, it becomes necessary to design and implement techniques capable of automatically extracting keywords. Such techniques might be divided into [37]: (1) *corpus-oriented*, requiring a collection of documents; or (2) *single document-oriented*, performed on individual documents. Despite their differences, these techniques follow two basic steps [38]: (1) define heuristics, such as a measure of similarity, term frequency or predetermined relationships between words; and (2) locate and define a set of one or more words that describe the topic of a text with precision. Given this, according to [36], the applied heuristics might be classified into *simple statistical*, *linguistics*, *machine learning* and *hybrid* approaches. The

simple statistical approaches do not require any learning mechanism and use simple statistical information of words to define the keywords, such as *word frequency* [39], *word co-occurrence* [40], TFIDF [12], etc. The linguistic approaches are those that use linguistic features of words, such as *parsing* [41], *lexical analysis* [42], *discourse analysis* [42], etc.

The machine learning approaches use keywords extracted from documents by means of a training process and apply them to a machine-learning model in order to find keywords in new presented documents. Among the well-known methods are the *naïve bayes* [34], *support vector machines* [36], KEA [45], and GenEx [46]. The hybrid approaches combine one or more techniques.

The method to be proposed in this paper is a graph-based and corpus-oriented approach, because the keywords will be extracted from a set of tweets represented as a graph of messages posted in Twitter. This text representation model allows the capture of statistical information, such as term co-occurrence frequency, centrality measures, and aggregation of linguistic features.

2.5. Related works

The previous section reviewed graph-based text representation and keyword extraction methods, listing some of the main works from the literature about these subjects. The present paper brings together these two fields so as to propose a Twitter keyword extraction method based on texts represented as graphs. This section provides a brief overview of the main works that combine these two approaches.

Proposed by [47], the KeyGraph is a method for graph-based document representation and keyword extraction. This method has the terms as vertices and co-occurrence relationships between pairs of these terms as edges, such as most of the works that rely on graph structures. First, the method creates an initial graph by establishing co-occurrence relationships between terms after stopwords removal. Then, the words joining two maximally connected subgraphs are identified. Finally, among the words found in the second step, those that appear in many maximally connected components are taken as keywords. Experiments were performed with a collection of 5900 documents in the artificial intelligence domain. The authors used *precision* and *recall* to evaluate KeyGraph, and compared it with TFIDF and N-Gram based methods.

In [29] *centrality measures* are employed in the definition of keywords for individual documents, which are transformed into textual graphs. Each word is represented by a single vertex and each edge corresponds to a pair of terms to which it is assigned a dissimilarity value. This value is given by the inverse of the co-occurrence frequency between a pair of terms, which corresponds to the number of sentences in which both occur. After the graph is constructed, the centrality measures are calculated for each vertex and a ranking is generated. Vertices that occupy the top ranking positions correspond to the potential keywords of the document. In his experiments, the author collected news stories from 64 news magazines in India, covering the categories environment, economy, defense, health and cinema. The average size of these stories was 1352 words and 8208 characters. As an evaluation criterion for the quality of the obtained keywords, the author compared the keywords from the algorithm with the words in news stories headlines, which also might be seen as keywords.

In [48], a method based on both supervised and unsupervised approaches was proposed. In the supervised approach, classification algorithms were trained over a summarized collection of documents and, thus, a keyword identification model was induced. The unsupervised approach consisted of the application of the HITS algorithm to a textual graph, from which the top ranked nodes were taken as keywords. The experiments were driven for the base of news articles from the *Document Understanding Conference* 2002 (DUC2002), which contains 566 documents in English. The performance was evaluated in terms of *accuracy*, *true positive rate* (TP), *false positive rate* (FP), *precision*, *recall* and *F-measure*.

The work proposed by [49] consisted of a method that received as input an ontology and a plain text document, and returned as output a set of contextualized keywords of the document. For the ontology entry, the model used a *termino-ontological resource*, TOR, which used the categorization from Wikipedia. The TOR resource was used in the creation of a directed weighted graph that represents each word or composition of words, that is, *n*-grams. These words and *n*-grams match the entries of the TOR by means of a string comparison, e.g. if the match is positive, a graph is built having the word and *n*-grams as its leaves. Thus, the relation between the vertices in this graph was based on a hierarchical relation to a concept of the TOR resource. After this graph was constructed, it was merged into a consolidated graph that represented the whole text. The model used Wikipedia as TOR resource and was applied to bases such as Wikiversity[1] and UNIT (a French acronym for engineering and technology digital university). The evaluation was made by taking into account the *recall*, *precision* and *F-measure* scores.

A common feature between the method to be proposed here, called TKG, and the techniques reviewed above is the principle of word co-occurrence in a text, from which the edges of a graph are given. However, the main differences to these techniques are:

1. The way in which the textual graph is built. The edge assignment heuristics of TKG are based on the text representation model from [28], which has low computational cost and is simple to be implemented, because it does not require any external information such as a pre-defined keyword list for training.

[1] https://pt.wikiversity.org/wiki/P%C3%A1gina_principal.

2. The methods of [47,29] focus on individual documents. An individual document is able to present one or more subjects in its content, requiring only the location of discriminative keywords without the information from a corpus. The TKG, on the other hand, which can also work with individual documents, is focused on tweet collections. As the tweets are short documents, normally written in informal language, they not always have useful information when individually analyzed [52].
3. The construction of the textual graph is also different from that of [48,49]. In [48], the graph is built according to the model proposed in [31], which establishes directed edges without weights. On the other hand, TKG defines undirected edges and presents different possibilities of weighting. The technique of [49] is the most different from TKG in terms of graph construction, because it requires external information that relies on TOR resources.

Although not directly comparable with the method to be proposed here, the KEA (*Keyphrase Extraction Algorithm*) is one of the most well-known algorithms for extracting keyphrases from text documents [45]. Thus, a brief description of how it works will be presented here. The KEA operation is performed in four steps: extracting candidate *n*-grams; extracting computer features; building the model; and extracting keyphrares. The extracting candidates step extracts *n*-grams of a predefined length (e.g. 1–3 words) that do not start or end with a stopword; a thesaurus can be inserted, in this case it only collects those *n*-grams that match the thesaurus terms. In the extraction of computer features step, for each candidate it is computed four values: TF-IDF; the percentage of documents preceding the first occurrence of the term in the document; the length of a phrase (i.e., the number of its component words); and the node degree of a candidate phrase (i.e., the number of phrases in the candidate set that are semantically related to this phrase). In the building model step a computational model is built from a set of text documents and the corresponding keywords. The extraction step selects the keywords by computing probabilities generated in the Naïve Bayes training [1].

3. TKG: a graph-based technique to extract keywords from tweets

The technique proposed in this paper, named TKG (standing for *Twitter Keyword Graph*), consists of three sequential steps (Fig. 1): (1) document pre-processing; (2) textual graph building from preprocessed tweets; and (3) keyword extraction, which involves the calculation of the centrality measures for each vertex (token), the ranking of these vertices, and the selection of keywords based on their rank.

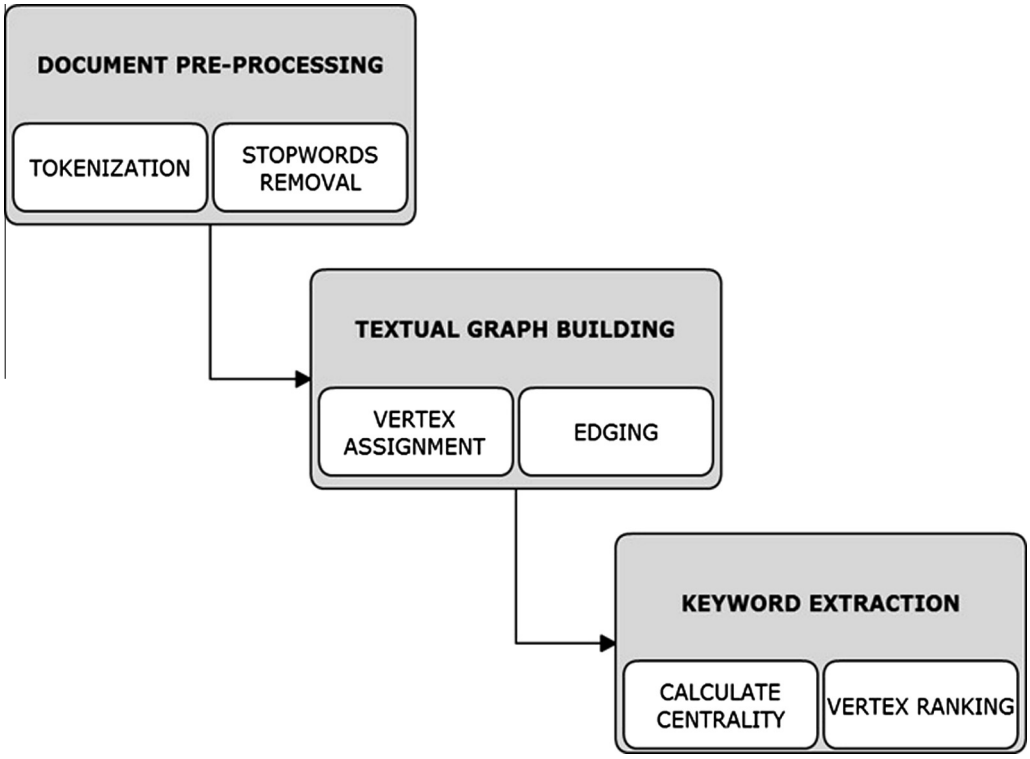

Fig. 1. Main processes of the proposed method (TKG).

3.1. Document pre-processing

Each tweet collected is seen as a document $\mathbf{d} \in D$, where $D = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N\}$ is the set (corpora) of N documents collected. In the *preprocessing* step, a document $\mathbf{d}_i \in D$ is tokenized; each *token* or *term* is a basic unit of the document. Then, the stopwords are removed and the document $\mathbf{d}_i$ is represented by its attribute vector $\mathbf{d}_i = \mathbf{t}^i = \{t_1^i, t_2^i, \ldots, t_K^i\}$, of K terms, $t_j \in T$, $j = 1, \ldots, K$. Note that K is the number of tokens extracted from the whole collection of documents.

3.2. Textual graph-building

In the *textual graph building* step a token corresponds to a vertex and an edge defines the co-occurrence relationship between tokens. Given the token set T, it is possible to define the vertices and edges of the graph $G = (V, E)$, where $V = \{t_1, t_2, \ldots, t_K\}$ is the set of vertices and $E = \{e_{1,1}, e_{1,2}, \ldots, e_{I,J}\}$ the set of edges. In the *vertex assignment* step, a complete reading of the attribute vector is performed and one vertex is created for each token. In the *edging* step, the definition of the edges can be performed according to two heuristics, inspired by the model proposed in [28], based on co-occurrence:

1. *Nearest neighbor edging* (TKG_1): in the first heuristics the edges are established by pairs of tokens read in the same sequence in which they appear in the original document. These tokens are read from the first to the penultimate, so that an edge $e_{i,j}$ is generated for each token i and its immediate successor j.
2. *All neighbors edging* (TKG_2): the second heuristics, though also uses a notion of proximity between the terms, is not limited to the following ones. Edges connecting the terms are created among all tokens that occur in a window of arbitrary size. Thus, for a given token and each following token within the window, an edge is created. The reading of these tokens also occurs in the same sequence in which they appear in the original document. In the context of the present work, as a tweet can have at most 140 characters, it does not contain a number of tokens as large as news, technical reports, scientific papers, or other types of documents. Besides, after preprocessing, this number is even smaller. This small number of tokens makes it possible to set the window size the same as the attribute vector size.

The two heuristics for edge creation also include the *weighting* information of an edge given by the frequency with which the tokens co-occur; in other words, the edge co-occurrence frequency $f_{i,j}$ between the tokens i and j. Thus, the weight w_{ij} of an edge might be assigned according to the three following ways:

1. *Same weight edges* (W^1): all edges are assigned the same weight $w_{ij} = 1$.
2. *Weight as co-occurrence frequency* (W^F): the weight of an edge is seen as the co-occurrence frequency between the nodes i and j it connects, that is, $w_{ij} = f_{ij}$.
3. *Weight as inverse co-occurrence frequency* ($W^{1/F}$): the edge weight corresponds to the inverse of the co-occurrence frequency of the nodes it connects, given by $w_{ij} = 1/f_{ij}$.

3.3. Keyword extraction

After the graph is built, the centrality measures described previously (degree centrality C_i^D, closeness centrality C_i^C, and eccentricity C_i^E) can be calculated for each vertex $v_i \in V$ in G, and a rank for each measure is generated. With this, the n best ranked vertices are selected as keywords.

4. Experiments and discussion

The proposed technique, TKG, is evaluated according to its properties: edging heuristics (*nearest-neighbor edging*, named TKG_1; *all neighbors edging*, named TKG_2); edge-weighting possibilities (*same weight assignment*, W^1; *weight as co-occurrence frequency*, W^F; *weight as inverse co-occurrence frequency*, $W^{1/F}$); and centrality measures to be assigned to vertices (C_i^D, C_i^E and C_i^C). These properties are summarized in Table 1.

Given the TKG properties, the experiments to be conducted in this paper are divided into *preliminary validation*, *application to Portuguese tweet collections*, and *computational scalability*. The first one aims to validate the proposed technique by comparing its results with those presented in [29]. The second one investigates the performance of TKG when applied to the task of extracting a set of keywords from tweet collections written in Portuguese, in a way that this set might be able

Table 1
Variations of TKG properties.

Edging	Weighting	Centrality
TKG_1	W^1	C^D
TKG_2	W^F	C^E
	$W^{1/F}$	C^C

to provide a good representation of the overall collection. The third and last experiments assess how the three methods, TKG, TFIDF, and KEA scale in relation to the dataset size.

4.1. Prelimirary validation

The technique proposed by [29], here called *P2007*, is similar to the TKG approach in the sense that both rely on a graph-based text representation and the use of centrality measures to identify relevant nodes (words). The main differences between them involve the textual graph building process (mainly the vertex assignment process and edging criteria); the absence of stemming in TKG preprocessing (the reduction of words to their root is undesirable for TKG, because it may result in loss of information about their context [50]); and the number and type of focal documents (while *P2007* performs keyword extraction in individual documents, TKG extracts representative terms from tweet corpora). It is also important to remark that a tweet is different from a conventional document (e.g. business reports, news or academic papers), basically due to its reduced length, informality, slangs, use of hashtags, emoticons, mentions, and other elements specific to tweets.

In this preliminary analysis, the configurations of TKG properties were applied to the same text as *P2007* in his original work [29]. This text is an article published in November 21st, 2006 in the TIME Magazine. Its title is "Nepal, rebels sign peace accord". For sake of comprehension, the published text is fully transcribed below.

> *Nepal's government and Maoist rebels have signed a peace accord, ending 10 years of fighting and beginning what is hoped to be an era of peaceful politics in the Himalayan kingdom. In a ceremony, Nepali Prime Minister Girija Prasad Koirala and Maoist leader Prachanda signed the agreement on Tuesday, which brings the rebels into peaceful multiparty democratic politics.*
>
> *"The politics of violence has ended and a politics of reconciliation has begun," Koirala said after the signing. Last week, the Maoists agreed to intern their combatants and store their arms in camps monitored by the United Nations. Nepal's Maoist rebels have been fighting an armed rebellion for 10 years to replace the monarchy with a republic. More than 13,000 people have been killed in the fighting. According to the agreement, any use of guns by the rebels will be punished. The democratic government and the Maoists have agreed to hold elections in June 2007 for constituent assembly that will decide the fate of the monarchy.*
>
> *"This is a historic occasion and victory of all Nepali people," Chairman of the Communist Party of Nepal Prachanda said at the signing ceremony, witnessed by political leaders, diplomats, bureaucrats and the media. "A continuity of violence has ended and another continuity of peace has begun," Koirala said. "As a democrat it was my duty to bring non-democrats into the democratic mainstream. That effort is moving ahead towards success." The peace agreement is an example for the whole world since it is a Nepali effort without outside help," he added. The challenge Nepal now faces is holding constituent assembly elections in a peaceful manner.*
>
> *Meanwhile, Maoist combatants continued to arrive in seven camps across the country Tuesday, albeit without United Nations monitoring. A tripartite agreement between the government, Maoists and the U.N. has to be signed before the U.N. can be given a mandate to monitor arms and combatants. "I hope that we will quickly be able to reach tripatriate agreement on the full modalities for the management of arms and armies clarifying essential detail," said Ian Martin, Special Representative of the United Nations Secretary General in Nepal. The Maoists will now join an interim parliament and an interim government, as early as next week, following the agreement.*

The sample text was pre-processed by the following steps: division of the text into sentences delimited by punctuation characters (each sentence might be seen as a tweet); stopwords removal; and stemming. These steps lead the original text to a very similar set of 97 tokens upon which [45] applied his technique. After pre-processing, the TKG_1 and TKG_2 graphs were constructed following the steps described previously, being generated 145 edges by TKG_1 and 940 edges by TKG_2. For both graphs, the edges may assume the weights W^1, W^F and $W^{1/F}$.

The results for the TKG techniques are given by the ranking scheme based on the centrality measures used in *P2007*. The author used the eccentricity C^E and closeness C^C measures to define an initial ranking of 16 positions. After that, if some words have equal centrality values, the degree centrality C^D is used to break these ties, such that high degree words are considered more relevant. After that, a ranking of 10 positions is generated. Here, the notation $C^E|C^D$ corresponds to the ranking of 16 positions given by C^E, followed by the tie break performed by the C^D order, while $C^C|C^D$ is the analogous process for C^C followed by C^D.

Table 2 compares the results of TKG_1 and TKG_2 variations with that of *P2007*. The common keywords between a TKG technique and *P2007* were highlighted in bold. The results showed some variations of terms. This occurs, basically, due to the edging and weighting schemes performed by the techniques.

The three results of TKG_1 according to $C^E|C^D$ presented 5, 5 and 4 terms in common with *P2007*, whilst those of TKG_2 presented 3, 2 and 7 common terms. The common terms for the $C^C|C^D$ results of TKG_1 were 8, 9 and 7, respectively, whilst the results for TKG_2 were 9, 7 and 9. These results suggest that the $C^E|C^D$ scheme is more sensitive to the graph topology than $C^C|C^D$. The first method recovered about half the terms found by *P2007*, whilst the second one, $C^C|C^D$, presented similar results as those of *P2007*, diverging in only a few terms for each comparison.

This preliminary experiment suggests the usefulness of centrality measures for keyword extraction in textual graphs. In this particular case, the closeness centrality showed to be more stable in relation to the graph topology. It is also important to remark that in the keyword extraction problem, most of the time the order (rank) is not of primary importance.

4.2. Application to tweet collections in portuguese

The goal of the second set of experiments is to evaluate the TKG performance and compare it with other approaches, KEA and TFIDF, from the literature, and human keyword extraction. In this analysis, all algorithms performed the task of

Table 2
Comparison of TKG variations with P2007.

Top-10	$C^E \vert C^D$						P2007
	TKG_1			TKG_2			
	W^1	W^E	$W^{1/E}$	W^1	W^E	$W^{1/E}$	
1	***maoist***	***nepal***	***nepal***	***nepal***	***polit***	***nepal***	*nepal*
2	*arm*	***maoist***	***maoist***	***peac***	*week*	***agreement***	*agreement*
3	*combat*	***sign***	*arm*	***sign***	***prachanda***	***peac***	*peac*
4	***nepal***	***rebel***	***rebel***	*week*	*tripatriat*	***maoist***	*sign*
5	***sign***	*govern*	*fight*	*tripatriat*	*special*	***sign***	*polit*
6	***rebel***	*fight*	*combat*	*special*	*secretary*	***polit***	*maoist*
7	*Tuesday*	*combat*	*govern*	*secretari*	*representative*	*interim*	*rebel*
8	*monitor*	*Tuesday*	*agre*	*representativ*	*reach*	*govern*	*leader*
9	***leader***	*monitor*	*stor*	*Tuesday*	*quickly*	*arm*	*prachanda*
10	*week*	***leader***	***sign***	*year*	*Tuesday*	***rebel***	*ceremoni*
1	***maoist***	***sign***	***maoist***	***nepal***	***nepal***	***nepal***	*nepal*
2	***rebel***	***peac***	***rebel***	***maoist***	***maoist***	***maoist***	*peac*
3	***sign***	***agreement***	***govern***	***govern***	***govern***	***agreement***	*maoist*
4	***govern***	***maoist***	***nepal***	***agreement***	***sign***	***peac***	*agreement*
5	***nepal***	***nepal***	***sign***	***sign***	***arm***	***sign***	*sign*
6	***peac***	***govern***	***peac***	***polit***	***polit***	***rebel***	*polit*
7	***agreement***	***rebel***	*agre*	***peac***	*democrat*	***polit***	*rebel*
8	*combat*	***polit***	***agreement***	***arm***	*Tuesday*	***govern***	*arm*
9	*prachanda*	*ceremoni*	*combat*	*hope*	*prachanda*	***arm***	*govern*
10	***leader***	***leader***	*prachanda*	***rebel***	***leader***	*interim*	*leader*

extracting keywords from three Brazilian TV shows, each with 100 tweets. The first topic, T_1, discusses a show called "*Trofeu Imprensa*"; the second one, T_2, a reality show called "*A Fazenda*"; and the third one, T_3, refers to children's support donation show called "*Crianca Esperanca*". The full list of tweets can be found in Appendix A.

In the present paper, KEA was trained with tweets about Brazilian TV shows. 100 tweets of each TV show were collected and their keywords were defined manually. The Portuguese stopword list was inserted in the extracting candidates step, and no thesaurus was used. 10 keywords from each TV show were extracted, and 10 keywords joining all tweets.

In principle, there is no "correct" set of keywords for a given document, not even humans may agree on the keywords they extract from documents. Therefore, to assess the performance of the proposed algorithm and the others from the literature, the following methodology was adopted:

1. Three persons (humans) were invited to suggest an unspecified number of keywords from the documents. After that, the intersection of these sets for each show is determined: these new sets contain the relevant documents, {*Relevant*}, for each show.
2. The TKG variations are compared with the TFIDF and KEA methods, having the human-defined sets as references. Here, the TKG configurations were denoted as follows:

$$TKG_{\text{"EDGING"}} \left| W^{\text{"WEIGHTING"}} \right| C^{\text{"CENTRALITY"}}, \tag{5}$$

Table 3
Keyword sets suggested by each human reader for each TV show. In bold are those keywords proposed by all readers (intersection set).

Reader 1	Show 1	***mel***, ***fronckowiak***, ***micael***, ***borges***, ***trofeu***, ***imprensa***, ***sbt***, ***emissoras***, *trofeuimprensa, paula, fernandes,* ***premio***, *restarnotrofeuimprensa, lombardi,* ***premiacao***, *rebeldes, silvio, santos*
	Show 2	***fazenda***, ***record***, ***xepa***, *andressa, reality, show, afazenda,* ***barbara***, *evans, rede, foraxepa, denise, rocha, novela, juliana, silveira, monique, trevisol*
	Show 3	***crianca***, ***esperanca***, ***globo***, ***dinheiro***, *wikileaks, criancaesperanca,* ***ronaldo***, ***emagrecer***, *novelas, jorge, mateus, unesco,* ***bb***, *doacoes, luan, santana, ivete*
Reader 2	Show 1	***mel***, ***fronckowiak***, *chay, suede,* ***micael***, ***borges***, ***sbt***, ***emissoras***, ***premiacao***, ***trofeu***, ***imprensa***, *edicao, rebelde, juntas,* ***premio***, *compartilhe*
	Show 2	***fazenda***, *afazenda,* ***record***, ***barbara***, *assistindo, assistir, vendo, roca, banho, mateus,* ***xepa***
	Show 3	***crianca***, ***esperanca***, ***globo***, ***dinheiro***, ***ronaldo***, ***emagrecer***, *milhao, millhoes, mesmice, menos, sonega, jogada, impostos, luan, santana,* ***bb***
Reader 3	Show 1	***trofeu***, ***imprensa***, ***sbt***, ***mel***, ***fronckowiak***, *chay, suede,* ***micael***, ***borges***, *hoje,* ***emissoras***, *trofeuimprensa,* ***premio***, ***premiacao***, *restarnotrofeuimprensa*
	Show 2	***fazenda***, ***record***, ***barbara***, *assistir, evans, dona,* ***xepa***, *reality, show, votacao, rede, foraxepa, factor, denise*
	Show 3	***globo***, ***crianca***, ***esperanca***, ***dinheiro***, *wikileaks,* ***bb***, *criancaesperanca, doaram,* ***ronaldo***, ***emagrecer***, *documento, unesco, sonega, ivete*

where, according to Table 1, "EDGING" may assume the TKG_1 or TKG_2 heuristics, "WEIGHTING" may assume the W^1, W^F or $W^{1/F}$ heuristics, and "CENTRALITY" may assume the C^D, C^E or C^C centrality measures. The comparisons are performed according to the Top-10 rankings from the TKG configurations, TFIDF and KEA. The centrality measures adopted in this second set of experiments for the TKG rank creation were C^C and C^E.

3. The well-known Precision, Recall, and F-measure from information retrieval were used as evaluation metrics [51]:

$$\mathrm{Pr} = \frac{|\{Relavant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}, \tag{6}$$

$$\mathrm{Re} = \frac{|\{Relavant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}, \tag{7}$$

$$F - measure = F = 2 * \frac{\mathrm{Pr}^{*}\mathrm{Re}}{(\mathrm{Pr} + \mathrm{Re})}. \tag{8}$$

As the number of keywords retrieved by each method was defined as 10, to compute Pr the number of relevant documents retrieved was equal to the number of keywords proposed by the method that appear in at least one of the human lists. By contrast, to compute Re the number of relevant documents was considered to be the number of keywords in the intersection set of the three human lists.

Table 3 shows the sets of keywords suggested by each reader for each set of documents (TV show). The common keywords (intersection) among these sets are highlighted in bold; they compose the set of relevant keywords, {*Relevant*} to calculate recall.

Tables 4 and 5 show the results obtained by the TKG variations, TFIDF, and KEA. Keywords over a gray background are those that match the set of relevant keywords, and those highlighted in bold type appear in at least one of the

Table 4
Summary of the results for all methods applied to each set of tweets. Keywords over a gray background match the set of relevant keywords, and those printed in gray appear in at least of the sets proposed by humans. Precision, Recall and *F*-measure are calculated as in Eqs. (5)–(7), respectively. In this case, the TKG configurations have used closeness (C^C) as centrality measure.

	$TKG_1\|W^1\|C^C$	$TKG_1\|W^F\|C^C$	$TKG_1\|W^{1/F}\|C^C$	$TKG_2\|W^1\|C^C$	$TKG_2\|W^F\|C^C$	$TKG_2\|W^{1/F}\|C^C$	*TFIDF*	*KEA*
1	sbt	sbt	imprensa	trofeu	sbt	imprensa	**hoje**	sbt
2	trofeu	mel	trofeu	imprensa	trofeu	trofeu	**trofeuimprensa**	trofeu
3	imprensa	melhor	sbt	sbt	imprensa	sbt	daqui	imprensa
4	percam	trofeu	**trofeuimprensa**	**hoje**	perder	**hoje**	**juntas**	**trofeuimprensa**
5	**trofeuimprensa**	ganhou	**juntas**	**trofeuimprensa**	receber	**trofeuimprensa**	emissoras	**hoje**
6	mel	imprensa	**hoje**	percam	hrs	**juntas**	percam	daqui
7	**hoje**	partir	emissoras	mel	**restartnotrofeuimpensa**	emissoras	mel	emissoras
8	participacao	percam	**silvio**	fronckowiak	premio	daqui	micael	**juntas**
9	melhor	**hoje**	participacao	micael	banda	percam	**chay**	participacao
10	ganhou	estarao	daqui	**chay**	ganhou	mel	imprensa	percam
Pr	60%	50%	80%	90%	50%	80%	80%	70%
Re	40%	40%	40%	60%	40%	50%	40%	40%
F	48.00%	44.44%	53.33%	72.00%	44.44%	61.54%	53.33%	50.91%

Show 2: A Fazenda

	$TKG_1\|W^1\|C^C$	$TKG_1\|W^F\|C^C$	$TKG_1\|W^{1/F}\|C^C$	$TKG_2\|W^1\|C^C$	$TKG_2\|W^F\|C^C$	$TKG_2\|W^{1/F}\|C^C$	*TFIDF*	*KEA*
1	record	record	fazenda	record	fazenda	record	record	fazenda
2	fazenda	fazenda	rede	fazenda	record	fazenda	fazenda	record
3	hoje	agora	record	afazenda	programa	**rede**	**afazenda**	**rede**
4	agora	comercial	dormir	barbara	**afazenda**	**afazenda**	**rede**	comecar
5	**rede**	demora	ofuro	hoje	yudi	**denise**	**denise**	**afazenda**
6	bota	hoje	**mateus**	mulher	sente	barbara	barbara	comeca
7	demora	proibido	explica	gente	peao	dormir	**dormir**	barbara
8	comercial	der	motivos	falando	atividade	**rocha**	hoje	foto
9	**assiste**	gente	**factor**	**denise**	fudeu	peoa	mulher	**assistir**
10	explica	voc	papo	voc	estreia	ofuro	falando	hoje
Pr	40%	20%	40%	40%	30%	70%	70%	60%
Re	50%	50%	50%	50%	50%	75%	75%	75%
F	44.44%	28.57%	44.44%	44.44%	37.50%	72.41%	72.41%	66.67%

Show 3: Crianca Esperanca

	$TKG_1\|W^1\|C^C$	$TKG_1\|W^F\|C^C$	$TKG_1\|W^{1/F}\|C^C$	$TKG_2\|W^1\|C^C$	$TKG_2\|W^F\|C^C$	$TKG_2\|W^{1/F}\|C^C$	*TFIDF*	*KEA*
1	esperanca	globo	crianca	esperanca	esperanca	esperanca	dinheiro	crianca
2	globo	esperanca	esperanca	crianca	crianca	crianca	**menos**	esperanca
3	crianca	**criancaesperanca**	**menos**	globo	**criancaesperanca**	globo	**milhoes**	pagar
4	**criancaesperanca**	ontem	dinheiro	**milhoes**	globo	dinheiro	globo	dinheiro
5	dinheiro	crianca	globo	dinheiro	merda	**menos**	**criancaesperanca**	**menos**
6	**jogada**	**jogada**	pagar	**criancaesperanca**	ganhador	pagar	pagar	globo
7	engana	engana	bb	pedir	daquela	**milhoes**	emagrecer	novelas
8	**menos**	pergunta	destina	programa	chamada	emagrecer	destina	emagrecer
9	**milhoes**	dar	pedir	**menos**	sbt	bb	pedir	**milhoes**
10	arrecada	acerta	**wikileaks**	gente	ruim	ronaldo	**wikileaks**	ronaldo
Pr	80.00%	50.00%	70.00%	70.00%	40.00%	90.00%	70.00%	80.00%
Re	57.14%	42.86%	71.43%	57.14%	42.86%	100.00%	42.86%	85.71%
F	66.66%	46.16%	70.71%	62.92%	41.38%	94.74%	53.17%	82.76%

Table 5
Summary of the results for all methods applied to each set of tweets. Keywords over a gray background match the set of relevant keywords, and those printed in gray appear in at least of the sets proposed by humans. Precision, Recall and *F*-measure are calculated as in Eqs. (5)–(7), respectively. In this case, the TKG configurations have used eccentricity (C^E) as centrality measure.

	$TKG_1\|W^1\|C^E$	$TKG_1\|W^F\|C^E$	$TKG_1\|W^{1/F}\|C^E$	$TKG_2\|W^1\|C^E$	$TKG_2\|W^F\|C^E$	$TKG_2\|W^{1/F}\|C^E$	*TFIDF*	*KEA*
1	programa	**juntas**	programa	usem	**trofeuimprensa**	trofeu	**hoje**	sbt
2	melhor	emissoras	melhor	ultima	sbt	imprensa	**trofeuimprensa**	trofeu
3	imprensa	sbt	imprensa	**trofeuimprensa**	ultima	sbt	daqui	imprensa
4	twitteiros	restart	chato	trofeu	telesenna	**hoje**	**juntas**	**trofeuimprensa**
5	**trofeuimprensa**	receber	trofeu	transmitido	saudoso	**trofeuimprensa**	emissoras	**hoje**
6	trofeu	recebdno	sbt	tag	saudade	percam	percam	daqui
7	transmitido	perder	**trofeuimprensa**	**suede**	santana	**juntas**	mel	emissoras
8	**silvio**	partir	entrevistas	sbt	sair	emissoras	micael	**juntas**
9	sbt	obrigado	**juntas**	saudoso	roubalheira	daqui	**chay**	participacao
10	risus	melhor	**hoje**	saudade	rolaouenrola	micael	imprensa	percam
Pr	40%	30%	60%	40%	20%	80%	80%	70%
Re	30%	20%	30%	20%	10%	50%	40%	40%
F	34.29%	24%	40%	26.67%	13.33%	61.54%	53.33%	50.91%

Show 2: A Fazenda

	$TKG_1\|W^1\|C^E$	$TKG_1\|W^F\|C^E$	$TKG_1\|W^{1/F}\|C^E$	$TKG_2\|W^1\|C^E$	$TKG_2\|W^F\|C^E$	$TKG_2\|W^{1/F}\|C^E$	*TFIDF*	*KEA*
1	record	record	fazenda	record	record	record	record	fazenda
2	der	proibido	**rede**	yudi	paciencia	fazenda	fazenda	record
3	demora	hoje	record	**votacao**	horario	**rede**	**afazenda**	**rede**
4	voc	comercial	dormir	vorazes	certo	minutos	**rede**	comecar
5	vdd	**banho**	ofuro	volume	atrasado	**afazenda**	**denise**	**afazenda**
6	tds	arrumar	**mateus**	voltar	yudi	**denise**	barbara	comeca
7	tanto	agora	papo	voc	**votacao**	dormir	**dormir**	barbara
8	seculo	**afazenda**	motivos	verrugas	vorazes	barbara	hoje	foto
9	sambando	voc	**factor**	vdd	voltar	**rocha**	mulher	**assistir**
10	saiba	vdd	vontade	tt	verrugas	raios	falando	hoje
Pr	10%	30%	50%	20%	20%	70%	70%	60%
Re	25%	25%	50%	25%	25%	75%	75%	75%
F	14.29%	27.27%	50%	22.22%	22.22%	72.41%	72.41%	66.67%

Show 3: Crianca Esperanca

	$TKG_1\|W^1\|C^E$	$TKG_1\|W^F\|C^E$	$TKG_1\|W^{1/F}\|C^E$	$TKG_2\|W^1\|C^E$	$TKG_2\|W^F\|C^E$	$TKG_2\|W^{1/F}\|C^E$	*TFIDF*	*KEA*
1	**criancaesperanca**	globo	setembro	sbt	ronaldo	globo	dinheiro	crianca
2	**unesco**	voc	**criancaesperanca**	pedir	menos	esperanca	**menos**	esperanca
3	setembro	venha	**unesco**	**milhoes**	globo	**criancaesperanca**	**milhoes**	pagar
4	pablo	tuiteiro	arrecadado	merda	esperanca	crianca	globo	dinheiro
5	menos	sonegacao	destina	menos	emagrecer	dinheiro	**criancaesperanca**	**menos**
6	globo	**sonega**	globo	globo	dinheiro	menos	pagar	globo
7	esperanca	respeito	dinheiro	ganhador	crianca	pagar	emagrecer	novelas
8	chato	rede	menos	esperanca	bb	**milhoes**	destina	emagrecer
9	arrecadado	pergunta	doacao	dinheiro	volta	emagrecer	pedir	**milhoes**
10	ama	ordem	crianca	destina	vizinho	**wikileaks**	**wikileaks**	ronaldo
Pr	40%	20%	50%	40%	70%	80%	70.00%	80.00%
Re	28.57%	14.29%	42.86%	42.86%	100%	85.71%	42.86%	85.71%
F	33.33%	16.67%	46.15%	41.38%	82.35%	82.76%	53.17%	82.76%

human sets (in addition to the other detached ones). The former are used to compute recall (Re), and the latter to compute precision (Pr). The results presented in these tables show that the all neighbors edging scheme (TKG_2) when combined with the inverse co-occurrence frequency was superior to the nearest neighbor edging (TKG_1), giving the best results in two out of three cases for these configurations when using the C^C centrality measure and in all the three cases for configurations using C^E centrality. It is also observed that the TFIDF method found better results than KEA in two of the three tweets set.

4.3. Computational scalability

To assess the computational scalability of the algorithms; that is, how their computational time increases in relation to the dataset size, a new set of experiments was performed. A dataset with 50,000 tweets related to the program "*A Fazenda*" was taken and sampled as follows: {0.1, 1, 10, 20, 50 k} tweets. For each of these five datasets sizes, the running time of the algorithms was calculated and stored for comparison. The algorithms were run in a PC, Win 7, Core i7 2.2 GHz, 8 GB RAM. TFIDF, KEA and the graph generation of TKG were run in Java SE 1.7, and the centrality measures of TKG were run using Wolfram Mathematica 9.0.

Fig. 1 shows the computational time, in seconds, of the six variations of TKG compared with TFIDF and KEA. By comparing Fig. 2(a) with (d), and (b) with (e), it is noted that TKG_1 is almost twice faster than TKG_2 when combined with the same weight and co-occurrence frequency weighting schemes. This was partially expected, because TKG_1 uses a nearest neighbor edging scheme, whilst TKG_2 employs an all neighbors edging scheme, thus requiring more computation. The homogeneity in the difference may be explained by the small and restricted tweet length. For the inverse co-occurrence frequency weighting scheme of TKG, it is observed that the algorithm scales poorly for larger datasets and, in all cases (Fig. 2(c) and (f)), showed to be more computationally intensive than TFIDF and KEA. It is worth noting however, that TFIDF and KEA presented exponential growth in all experiments, whilst TKG was only exponential for $W^{1/F}$.

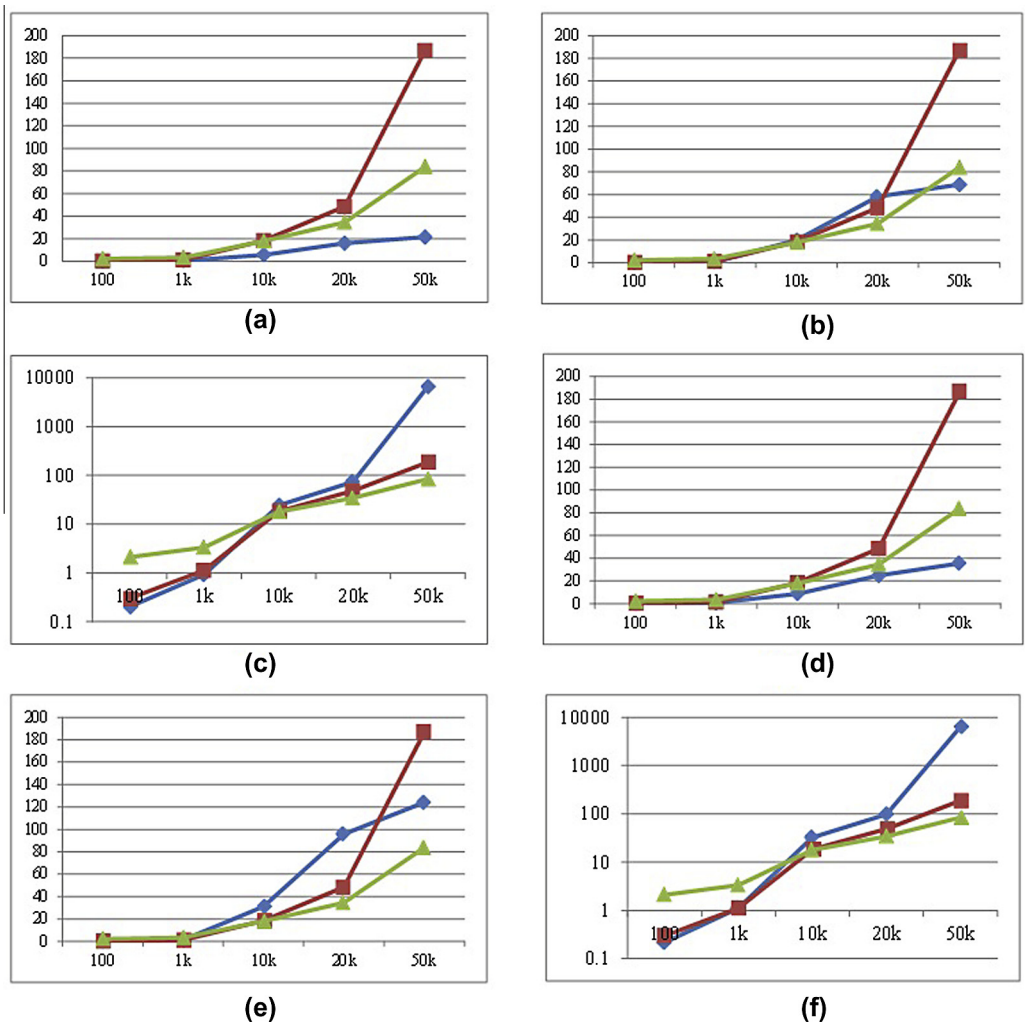

Fig. 2. Running time of the algorithms for instances of different sizes: {100,1,10,20,50 k} tweets. (a) $TKG_1|W^1|C^C$. (b) $TKG_1|W^F|C^C$. (c) $TKG_1|W^{1/F}|C^C$. (d) $TKG_2|W^1|C^C$. (e) $TKG_2|W^F|C^C$. (f) $TKG_2|W^{1/F}|C^C$. Legend: ◇ TKG; □ TFIDF; Δ KEA.

5. Conclusions and future work

This paper proposed a keyword extraction technique of Twitter messages in which tweets are represented as graphs. This kind of representation allows the application of many techniques from graph theory, link analysis, among others, to determine important patterns and trends, such as keywords. It also opens room for the application of graph mining techniques which can provide better performance when compared with text mining techniques that rely on the VSM model.

The proposed approach, called TKG, relies on three phases: text pre-processing; graph building; and keyword extraction. The text pre-processing methods used are standard ones from the text-mining literature, such as tokenization, stemming, and stopwords removal. Graph building takes each token as a vertex and edging can be performed in one of two ways: using a nearest-neighbor approach, or using an all neighbors approach. Finally, keywords are extracted from the graph by applying, in cascade, some graph centrality measures.

To assess the performance of the algorithm, it was initially applied to a single text from the literature and its results were compared with those from the literature. This preliminary experiment served the purpose of understanding the influence of each variation used in the method proposed. After this experiment, a new set of tests was performed. This time, TKG was compared with the TFIDF approach and the KEA algorithm. The benchmark was a human keyword extraction performed by three different users. The keyword (relevant) set for each of the three sets of tweets evaluated was constructed by taking the intersection of the keywords proposed by each of the human readers. This allowed us to compute the recall for each algorithm being compared. As only the Top-10 keywords extracted by each method were accounted for, to determine their precision the set of relevant keywords was that composed of the keywords proposed by at least one human reader. Finally, a new set of experiments was designed so as to investigate the computational scalability of the three algorithms. In this case,

five sets of tweets of increasing size were used and the computational time necessary to run the algorithms was recorded and compared.

The results obtained in these experiments showed that building the graph using an all neighbors edging scheme invariably provided superior performance, and assigning weights to the edges based on the weight as the inverse co-occurrence frequency was superior in two thirds of the cases. Closeness centrality was the keyword selection method chosen by the preliminary experiment, and was used in one of the variations of the second set of experiments. TKG showed to be faster for all its variations when compared with TFIDF and KEA, with the exception of the weighting scheme based on the inverse co-occurrence frequency. It is important to remark, however, that the use of TKG is much simpler than KEA, for instance, because it does not require any external knowledge, e.g. a model provided by training. Its scalability is achieved due to the way in which the edging processes receive a vector of tokens and assigns vertices and edges in order to build the graph, which requires basically local updates in the structure.

One possible future work is to apply other centrality measures not discussed here, individually or in combination. Further studies might also be directed to refining the structure of the graph using some heuristics, such as the elimination of noisy edges or vertices, what may reduce the computational running time of the algorithm as well. A clear difficulty of the proposed technique is to find the best configuration of TKG properties. Studies in this direction might also be explored in the future.

Acknowledgments

The authors thank Mackenzie University, Mackpesquisa, CNPq, Capes (Proc. n. 9315/13-6) and FAPESP for the financial support.

Appendix A. Tweet collections

TROFEU IMPRENSA

1. hoje estarao mel fronckowiak chay suede micael borges trofeu imprensa sbt percam hein
2. assistir trofeu imprensa
3. daqui pouco todas emissoras juntas sbt trofeuimprensa
4. hoje paula fernandes sbt recebdno premio trofeu imprensa
5. hoje estarao mel fronckowiak chay suede micael borges trofeu imprensa sbt assistir
6. hoje trofeu imprensa sbt participacao percam tag restartnotrofeuimpensa
7. daqui pouco todas emissoras juntas sbt trofeuimprensa
8. daqui pouco todas emissoras juntas sbt trofeuimprensa
9. gente hoje estarao chay suede mel fronckowiak micael borges trofeu imprensa transmitido sbt
10. daqui pouco todas emissoras juntas sbt trofeuimprensa
11. daqui pouco todas emissoras juntas sbt trofeuimprensa
12. daqui pouco todas emissoras juntas sbt trofeuimprensa
13. daqui pouco todas emissoras juntas sbt trofeuimprensa
14. saudade ultima abertura trofeu imprensa locucao saudoso lombardi trofeuimprensa sbt
15. hoje trofeu imprensa sbt participacao percam tag restartnotrofeuimprensa
16. trofeu imprensa edicao premiacao neste domingo divulgacao sbt neste domingo
17. trofeu imprensa edicao premiacao neste domingo divulgacao sbt neste domingo
18. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
19. daqui pouco todas emissoras juntas sbt trofeuimprensa
20. hoje trofeu imprensa sbt participacao perde hein
21. daqui pouco todas emissoras juntas sbt trofeuimprensa
22. falta hora comecar trofeu imprensa assistir
23. neste domingo luan santana trofeu imprensa partir sbt perca
24. trofeu imprensa sbt percam
25. daqui pouco todas emissoras juntas sbt trofeuimprensa
26. hoje trofeu imprensa sbt participacao percam tag restartnotrofeuimprensa
27. daqui pouco todas emissoras juntas sbt trofeuimprensa
28. hoje trofeu imprensa sbt participacao percam tag restartnotrofeuimprensa
29. hoje participam transmissao trofeu imprensa
30. hoje trofeu imprensa sbt participacao percam tag restartnotrofeuimprensa
31. falta hora comecar trofeuimprensa ansiosos mel fronckowiak recendo premio perca sbt
32. daqui pouco todas emissoras juntas sbt trofeuimprensa
33. trofeu imprensa categorias rebelde ganhou orgulhos addictedtheyluar melhor novela
34. gente hoje estarao chay suede mel fronckowiak micael borges trofeu imprensa transmitido sbt

35. daqui pouco todas emissoras juntas sbt trofeuimprensa
36. hoje trofeu imprensa sbt participacao perde hein
37. melhor premiacao brasileira trofeu imprensa risus
38. daqui pouco todas emissoras juntas sbt trofeuimprensa
39. daqui pouco todas emissoras juntas sbt trofeuimprensa
40. daqui pouco todas emissoras juntas sbt trofeuimprensa
41. daqui pouco todas emissoras juntas sbt trofeuimprensa
42. noite trofeu imprensa canal sbt
43. daqui pouco todas emissoras juntas sbt trofeuimprensa
44. daqui pouco todas emissoras juntas sbt trofeuimprensa
45. hoje atores cantora banda estarao trofeuimprensa
46. falta hora comecar trofeu imprensa assistir
47. trofeu imprensa twitteiros comentam sobre trofeu imprensa exibido sbt porquetts
48. galerinha linda liguem hoje trofeu imprensa sbt espalha
49. trofeu imprensa sbt mel fronckowiak
50. esquecam chay suede micael borges mel fronckowiak estarao trofeu imprensa transmitido sbt
51. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
52. obrigado sbt trofeuimprensa agora ficar ligadinho ganho telesenna
53. ano roubalheira trofeu imprensa sbt ganhou premios globo palhacada
54. esquecam trofeu imprensa sbt hoje hrs percam
55. pensei comecar trofeu imprensa passar programa chato silvio santos
56. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
57. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
58. hoje participam transmissao trofeu imprensa
59. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
60. trofeu imprensa categorias rebelde ganhou orgulhos addictedtheyluar melhor novela
61. noite trofeu imprensa canal sbt
62. rebeldeoficial trofeu imprensa categorias banda novela rebelde ganhou
63. hoje participam transmissao trofeu imprensa
64. hoje maior premiacao brasileira trofeu imprensa trofeuimprensa
65. hoje noite sbt trofeu imprensa eternos rebeldes chay mel micael percam chamelmichojenotrofeuimprensa
66. daqui pouco todas emissoras juntas sbt trofeuimprensa
67. trofeu imprensa twitteiros comentam sobre trofeu imprensa exibido sbt porquetts
68. trofeu imprensa categorias rebelde ganhou orgulhos addictedtheyluar melhor novela
69. hoje estarao record trofeu imprensa
70. hoje trofeu imprensa mel fronckowiak chay suede micael borges sbt perder hein
71. galerinha linda liguem hoje trofeu imprensa sbt espalha
72. assistir trofeu imprensa
73. daqui pouco todas emissoras juntas sbt trofeuimprensa
74. agora dar role canais voltar sbt trofeuimprensa
75. perca edicao trofeu imprensa silvio santos compartilhe
76. daqui pouco todas emissoras juntas sbt trofeuimprensa
77. perca edicao trofeu imprensa silvio santos compartilhe
78. daqui pouco todas emissoras juntas sbt trofeuimprensa
79. perca edicao trofeu imprensa silvio santos compartilhe
80. agoraetarde ganhou trofeu imprensa melhor programa entrevistas parabens danilogentili equipe merecido
81. assistir trofeu imprensa
82. hoje sbt trofeu imprensa participacao micael borges mel fronckowiak chay suede percam
83. hoje mel chay micael estarao trofeu imprensa sbt receber premios percam
84. perca edicao trofeu imprensa silvio santos compartilhe
85. perca edicao trofeu imprensa silvio santos compartilhe
86. hoje sbt trofeu imprensa participacao micael borges mel fronckowiak chay suede percam
87. assistir trofeu imprensa
88. hoje restart sbt trofeu imprensa percaam
89. perca edicao trofeu imprensa silvio santos compartilhe
90. daqui pouco todas emissoras juntas sbt trofeuimprensa
91. logo apos percam trofeu imprensa silvio santos jornalistas escolhendo melhores
92. sair trofeu imprensa sbt melhor
93. agora dar role canais voltar sbt trofeuimprensa rolaouenrola
94. liguem hoje trofeu imprensa sbt espalha usem tag luansantananotrofeuimprensasbt
95. perca edicao trofeu imprensa silvio santos compartilhe

96. hoje participam transmissao trofeu imprensa
97. daqui pouco todas emissoras juntas sbt trofeuimprensa
98. galerinha linda liguem hoje trofeu imprensa sbt espalha
99. trofeu imprensa categorias rebelde ganhou orgulhos addictedtheyluar melhor novela
100. liguem hoje trofeu imprensa sbt espalha usem tag luansantananotrofeuimprensasbt

A FAZENDA

1. agora vamos assistir estreia factor fazenda record fudeu
2. vendo xepa comeca fazenda
3. atividade barbara yudi ex peao sente falta fazenda rede record
4. assisti fazenda durmi bjss
5. enfim sair ficar assistindo fazenda
6. dias assisto fazenda
7. olhar fazenda dormir
8. barraqueira andressa mostrou reality show pessoa baixa afazenda foraandresa
9. tomar banho arrumar coisas assistir fazenda dormir
10. hoje roca record explica afazenda
11. hoje roca record explica afazenda
12. ficar vendo comecar fazenda
13. acho nen aguentar assistir fazenda
14. hjj votacao caramba record explica coisas direito afazenda
15. quero assistir fazenda
16. barbara evans mostra corpao mulher delicioso banho foto fazenda rede record
17. acobtecendo fazenda vivo
18. jantar gostozinho assistir fazenda
19. pensando assistir fazenda sono zero
20. realmente existem assistem fazenda
21. barbara evans mostra corpao mulher delicioso banho foto fazenda rede record
22. todas propagandas record iguais credo cicatricure point verrugas assepxia ex namorada pedrao afazenda
23. alguem passar link passa fazenda vivo pai vendo jogo deixar
24. gente vendo record net hoje muentto lenta acompanhar voc tt afazenda
25. novela dona chepa poderia acabar cedo fazenda demora tanto comecar poxa vida viu
26. raios louca mateus fazenda rede record
27. raios louca mateus fazenda rede record
28. raios louca mateus fazenda rede record
29. leva pessoa ficar comentando fazenda twitter sexta globo reporter
30. acaba logo novela comeca fazenda
31. assiste fazenda assisto factor sentiram diferenca
32. novela chata acaba ava oxe acaba logo quero fazenda foraxepa afazenda
33. assisto bate papo fazenda motivos suporto falando mal barbara denise
34. assistir fazenda agora crucifiquem
35. gzuis semanas vejo namorado bati record
36. esquecido vdd programas record fraquinhos rs
37. atriz record igualzinha voc
38. record sambando cara demora comecar afazenda
39. assistir fazenda dormir
40. tds novelas record mesmos atores
41. raios louca mateus fazenda rede record
42. novelinha record
43. claro gente decifrar fala peoa sobre denise rocha ofuro fazenda rede record
44. claro gente decifrar fala peoa sobre denise rocha ofuro fazenda rede record
45. ignorem materias postar print materias desnecessarias fazem paciencia neh nhacc risos
46. assistir fazenda
47. sono acho aguentar assistir fazenda
48. desenhar decifrar fala peoa sobre denise rocha ofuro fazenda rede record
49. desenhar decifrar fala peoa sobre denise rocha ofuro fazenda rede record
50. assistir fazenda sono deixa
51. record merda minutos comercial novela pacaba
52. assisto bate papo fazenda motivos suporto falando mal barbara denise
53. factor voltou slkhsjdk socorro oq vontade assistir fazenda factor

54. record bota fazenda comeca comeca quase absurdo
55. odeio gente assiste afazenda fica falando mal denise sabe tratam
56. voc assiste fazenda qse msm sendo porquera
57. olhar fazenda dormir whats beijos
58. coloco globo volume normal troco canal coloco record parece volume maximo shit
59. tirem criancas sala programa baixarias comecar fazenda
60. record bota fazenda comeca comeca quase absurdo
61. ignorem materias postar print materias desnecessarias fazem paciencia neh nhacc risos
62. argh curto fazenda preparando voice globo sony
63. agora dormir feliz assistir fazenda
64. mudar canal comecar fazenda
65. gracas record agora proibido assistir record casa
66. vontade assistir fazenda hoje mpn claudialeitte
67. record comeca programas horario certo atrasado minutos paciencia
68. sala assistir fazenda
69. demi lembrei record hoje falando selena mostrando fotos derepente apareceu foto demi what
70. imaginando remake ausurpadora record atrizes pobres balacobaco melhor pensar
71. saiba ira terminar contrato atores atrizes record
72. sono acho olhar fazenda hoje
73. assisto bate papo fazenda motivos suporto falando mal barbara denise
74. edicao programas brasileiros melhores record pasmem thexfactor
75. noticias juliana silveira estrelar especial ano record saiba
76. partiu assistir fazenda boa noite bjss fika otima noite durmam curtir compartilhar minuto
77. olhar fazenda dormir amores amoo
78. ganhar dinheiro assistir fazenda penso assistir
79. record gentileza atrasar vivo afazenda
80. falta fazenda comeca record sempre apelano
81. propagandas ordem rederecord cicatricure assepxia point verrugas mulher cafe leite eudora friboi ex namorada pedrao afazenda
82. album fotos renember record apresenta lua blanco outros atores principais
83. propagandas ordem rederecord cicatricure assepxia point verrugas mulher cafe leite eudora friboi ex namorada pedrao afazenda
84. propagandas ordem rederecord cicatricure assepxia point verrugas mulher cafe leite eudora friboi ex namorada pedrao afazenda
85. reclama cantando gritinhos assistindo fazenda entendo
86. fazenda comecar record
87. tiram supernatural programacao sabem perdem record audiencia
88. queria escolhi esperar discutir continuar vendo record
89. biscoitos cobertos chocolate comer vendo fazenda hue assiste
90. assisti fazenda der comercial tomar banho lavar belo
91. globo crepusculo sbt harry potter record voc dever passar jogos vorazes
92. julianne trevisol confirmada serie record
93. assistir fazenda ganhar rs
94. ruim record comercial demora seculo voltar
95. barbara evans fica pensativa sozinha area externa sede afazenda
96. assistir fazenda hihi
97. monique evans fila maquiagem record falar honestidade programa
98. pai senhor assistindo jogo serie pai fazenda lei papis
99. programacao record certinha
100. julianne trevisol confirmada serie record

CRIANCA ESPERANCA

1. doaram crianca esperanca naoo ruim voc globo condicoes gente daria milhoes pedir
2. globo venha pedir dinheiro crianca esperanca sendo milhao meio dar big brother
3. globo sonegou milhoes reais ano sabem equivalente edicoes crianca esperanca
4. doc vazado wikileaks globo destina arrecadado criancaesperanca unesco
5. globo dinheiro tudoo menos crianca esperanca
6. globo dinheiro menos crianca esperanca
7. globo mata programacao futuro criancas crianca esperanca
8. globo dinheiro menos crianca esperanca

9. assisti crianca esperanca causa ivete
10. globo dinheiro menos crianca esperanca
11. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
12. enjoei globo dinheiro menos crianca esperanca
13. tuiteiro globo dinheiro menos crianca esperanca
14. globo dinheiro menos crianca esperanca
15. mundo assistindo crianca esperanca assistindo rainha festa uva
16. blog globo acerta tirar crianca esperanca mesmice
17. globo dinheiro menos crianca esperanca
18. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
19. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
20. crianca esperanca achei ofensivo programa retira favor globo ordem
21. rede globo sonega impostos deveriam publico cujo explora crianca esperanca bonito gente
22. assisti crianca esperanca jorge mateus awn lindoss
23. globo paga milhoes ronaldo emagrecer pedir dinheiro crianca esperanca
24. comeca palhacada global crianca esperanca funciona menos populacao doa dinheiro globo
25. documento vazado wikileaks globo destina arrecadado criancaesperanca unesco
26. globo dinheiro menos crianca esperanca
27. doc vazado wikileaks globo destina arrecadado criancaesperanca unesco
28. globo dinheiro menos crianca esperanca
29. globo dinheiro menos crianca esperanca
30. puutz acho perdendo crianca esperanca
31. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
32. globo milhoes rica crianca esperanca arrecada milhoes doacoes
33. assistindo crianca esperanca tirando fotos divos lindos jorge mateus fotos entra
34. globo dinheiro menos crianca esperanca
35. concordo assino embaixo assistindo crianca esperanca
36. documento vazado wikileaks globo destina arrecadado criancaesperanca unesco
37. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
38. acho quiser ligo sabe crianca esperanca precisa ficar falando
39. globo dinheiro menos crianca esperanca
40. assistir crianca esperanca fiquei preguica
41. globo dinheiro menos crianca esperanca
42. fiz merecer ficar sabado noite casa assistindo crianca esperanca
43. foda globo dando milhoes maluco emagrecer vir pedir dinheiro crianca esperanca
44. programa aberta enche saco crianca esperanca
45. globo dinheiro menos crianca esperanca
46. globo dinheiro menos crianca esperanca
47. doc vazado wikileaks globo destina arrecadado criancaesperanca unesco
48. globo sonegou milhoes reais ano sabem equivalente edicoes crianca esperanca
49. assistindo crianca esperanca luan luansantananocriancaesperanca
50. globo sonegou milhoes reais ano sabem equivalente edicoes crianca esperanca
51. alguem globo paga ronaldo emagrecer pede dinheiro crianca esperanca
52. globo milhoes reais ronaldo emagrecer fica arrecadando dinheiro crianca esperanca
53. globo dinheiro menos crianca esperanca
54. globo dinheiro menos crianca esperanca
55. assistindo crianca esperanca posso ceu
56. real olhar crianca esperanca tarde entro
57. globo dinheiro menos crianca esperanca
58. entendendo luan santana fazendo show entrar crianca esperanca parada vivo confusao
59. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
60. foda globo dando milhoes maluco emagrecer vir pedir dinheiro crianca esperanca
61. moral globo dinheiro menos criancaesperanca
62. globo dinheiro menos crianca esperanca
63. uhul maratona novela crianca esperanca amp altas sabado
64. volta casa filho mae pergunta filha assistir crianca esperanca
65. convidou sair mentira convidou sermos duas assistindo crianca esperanca juntas
66. complicado ama criancaesperanca globo
67. incrivel globolixo programa serginho groisman ensinando jovens libertinagem sexual crianca esperanca
68. globo crianca esperanca ontem dar comparar presente corinthians corinthians anos
69. crianca esperanca desculpa acobertar sonegacao globo

70. crianca esperanca programa serio destina doacoes instituicoes carentes exemplo palmeiras
71. pablo criancaesperanca chato partiu sbt
72. crianca esperanca arrecada milhoes doacoes voc burro globo mente
73. gente entendi globo dinheiro menos crianca esperanca parem kibar
74. assistindo crianca esperanca agora nenhum cantor agradou
75. gente entendi globo dinheiro menos crianca esperanca parem kibar
76. gata vamos casa assistir crianca esperanca
77. assistindo crianca esperanca luan luansantananocriancaesperanca
78. crianca esperanca arrecada milhoes doacoes voc burro globo mente
79. globo dinheiro menos crianca esperanca
80. sobre crianca esperanca globo desnecessario
81. desculpa sociedade acredito nesse papo crianca esperanca engana globo
82. perguntarem dormi tarde hoje falar assistindo crianca esperanca
83. sbt poderia passar harry potter record titanic fuder crianca esperanca hahaa
84. eduardoazeredo globo milhoes rica crianca esperanca arrecada milhoes doacoes
85. gata vamos casa assistir crianca esperanca
86. ultima noite exemplifica penso respeito globo crianca esperanca
87. balada assistindo crianca esperanca
88. gosta crianca esperanca continua assistindo
89. doem crianca esperanca jogada globo deduzir impostos milhoes gordo rico emagrecer
90. globo dinheiro pagar bb pagar ronaldo emagrecer pagar novelas menos crianca esperanca
91. globo dinheiro menos crianca esperanca
92. gente bonita twitter assisti crianca esperanca incrivel amanha denovo
93. uhul maratona novela crianca esperanca amp altas sabado
94. globo dinheiro menos crianca esperanca
95. fav assistiu crianca esperanca jorge mateus luan santana
96. obrigado acompanhar durante repleto solidariedade doacao setembro criancaesperanca
97. ouvindo crianca esperanca causa vizinho chatice total hipocrisia programa
98. globo doasse ganha semana comerciais precisava crianca esperanca
99. pergunta globo dinheiro bb dar crianca esperança
100. globo milhoes ganhador daquela merda chamada bb pedir dinheiro crianca esperanca criancaesperanca

References

[1] J.H. Kietzmann, K. Hermkens, I.P. McCarthy, B.S. Silvestre, Social media? Get serious! Understanding the functional building blocks of social media, Bus. Horizons 54 (3) (2011) 241–251, http://dx.doi.org/10.1016/j.bushor.2011.01.005. ISSN 0007-6813.
[2] A.M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of social media, Bus. Horizons 53 (1) (2010) 59–68.
[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, Finding high-quality content in social media, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008, pp. 183–194.
[4] S. Asur, B.A. Huberman, Predicting the future with social media, in: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, IEEE, 2010, pp. 492–499.
[5] L. Safko, The Social Media Bible: Tactics, Tools, and Strategies for Business Success, John Wiley & Sons, 2010.
[6] M. Yoshida, S. Matsushima, S. Ono, I. Sato, H. Nakagawa, ITC-UT: tweet categorization by query categorization of on-line reputation management, in: Conference on Multilingual and Multimodal Information Access Evaluation, 2010.
[7] P.S. Earle, D.C. Bowden, M. Guy, Twitter earthquake detection: earthquake monitoring in a social world, Ann. Geophys. 54 (6) (2011) 708–715.
[8] A. Bermingham, A. Smeaton, On Using Twitter to Monitor Political Sentiment and Predict Election Results, Sentiment Analysis where AI meets Psychology, 2011, pp. 2–10.
[9] C.D. Corley, D.J. Cook, A.R. Mikler, K.P. Singh, Text and structural data mining of influenza mentions in web and social media, Int. J. Environ. Res. Publ. Health 7 (2) (2010) 596–615.
[10] R. Feldman, J. Sanger, The Text Mining Handbook Advanced Approaches in Analysing Unstructured Data, [S.l.]: Cambridge, 2007.
[11] M.W. Berry, M. Castellanos (Eds.), Survey of Text Mining, Springer, New York, 2004.
[12] A.M. Cohen, W.R. Hersh, A survey of current work in biomedical text mining, Briefings Bioinf. 6 (1) (2005) 57–71.
[13] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, Ldv Forum 20 (1) (2005) 19–62.
[14] V. Gupta, G.S. Lehal, A survey of text mining techniques and applications, J. Emerg. Technol. Web Intell. 1 (1) (2009) 60–76.
[15] L. Hirschman, H.S. Thompson, Overview of evaluation in speech and natural language processing, in: Ron Cole (Ed.), Survey of the State of the Art in Human Language Technology, Cambridge Studies in Natural Language Processing Series, vol. XII–XIII, Cambridge University Press, New York, NY, USA, 1997, pp. 409–414.
[16] G.G. Chowdhury, Natural language processing, Annu. Rev. Inf. Sci. Technol. 37 (1) (2003) 51–89.
[17] C.D. Manning, Foundations of statistical natural language processing, in: H. Schütze (Ed.), MIT Press, 1999.
[18] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, 1999.
[19] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill, 1983.
[20] W.B. Frakes, R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms, Prentice Hall, 1992.
[21] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, vol. 1, Cambridge University Press, 2008. p. 6.
[22] G. Salton, C.S. Yang, C.T. Yu, A theory of term importance in automatic text analysis, J. Am. Soc. Inf. Sci. (1975).
[23] Datasift, Browse Data Sources – Twitter, 2012. [Online]. Available: <http://datasift.com/source/6/twitter>. [Accessed24October 2013].
[24] M.A. Russell, Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. O'Reilly Media Inc., 2013.
[25] D. Ediger, K. Jiang, J. Riedy, D.A. Bader, C. Corley, R. Farber, W.N. Reynolds, Massive social network analysis: mining twitter for social good, in: 2010 39th International Conference on Parallel Processing (ICPP), IEEE, 2010, pp. 583–593.

[26] C. Zhang, H. Wang, Y. Liu, Y. Wu, Y. Liao, B. Wang, Automatic keyword extraction from documents using conditional random fields, J. Comput. Inf. Syst. (2008) 1169–1180.
[27] J.L. Gross, J. Yellen, Graph Theory and Its Applications, second ed., Chapman & Hall/CRC, 2006.
[28] W. Jin, R.K. Srihari, Graph-based text representation and knowledge discovery, in: Proceedings of the 2007 ACM Symposium on Applied, Computing, vol. 7, 2007, pp. 807–811.
[29] G.K. Palshikar, Keyword extraction from a single document using centrality measures, Pattern Recogn. Mach. Intell. 4815 (2007) 503–510.
[30] F. Zhou, F. Zhang, B. Yang, Graph-based text representation model and its realization, in: 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), vol. 1, No. 8, 2010, p. 21–23.
[31] A. Schenker, M. Last, H. Bunke, Classification of web documents using a graph model, in: Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR2003), Computer Society Press, Scotland, 2003.
[32] S. Hensman, Construction of conceptual graph representation of texts, in: Proceedings of Student Research Workshop at HLT-NAACL, Boston, 2004, p. 49–54.
[33] J. Nieminen, On the centrality in a graph, Scand. J. Psychol. 15 (1974) 332–336.
[34] S. Wasserman, K. Faust, D. Iacobucci, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, 1995.
[35] P. Hage, F. Harary, Eccentricity and centrality in networks, Soc. Networks 17 (1995) 57–63.
[36] K. Zhang, H. Xu, J. Tang, J.Z. Li, Keyword extraction using support vector machine, in: Proceedings of the Seventh International Conference on Web-Age, Information Management (Waim2006), 2006.
[37] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic Keyword Extraction from Individual Documents, Text Mining: Applications and Theory, 2010.
[38] B. Lott, Survey of Keyword Extraction Techniques, UNM Education, 2012.
[39] H.P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM J. Res. Dev. (1957).
[40] Y. Matsuo, M. Ishizuka, Keyword extraction from a single document using word co-occurrence statistical information, Int. J. Artif. Intell. Tools 4 (2004).
[41] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, p. 216–223.
[42] G. Erckan, I. Cicekli, Using lexical chains for keyword extraction, Inf. Processing Manage. (2007).
[43] S.F. Dennis, The design and testing of a fully automatic indexing–searching system for documents consisting of expository text, in: G. Schecter (Eds.), Information Retrieval: A Critical Review, 1967.
[44] E. Frank, G.W. Paynter, I.H. Witten, Domain-specific keyphrase extraction, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence, 1999.
[45] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, C.G. Nevill-Manning, KEA practical automatic keyphrase action, in: Proceedings of the 4th ACM Conference on Digital Library (DL'99), Berkeley, CA, USA, 1999, p. 254–226.
[46] P.D. Turney, Learning to Extract Keyphrases from Text, NRC Technical Report ERB-1057, National Research Council, Canada, 1999, p. 1–43.
[47] Y. Ohsawa, N.E. Benson, M. Yachida, KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, in: Proceedings. IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998, ADL 98, p. 12–18.
[48] M. Litvak, M. Last, Graph-based keyword extraction for single-document summarization, in: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, 2008, p. 17–24.
[49] C.A. Chahine, N. Chaignaud, J.Ph. Kotowicz, J.P. Pécuche, Context and keyword extraction in plain text using a graph representation, in: Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems, vol. 8, 2008, p. 692–696.
[50] G. Kowalski, Information Retrieval Architecture and Algorithms, Springer, US, 2011.
[51] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Academic Press, 2001.
[52] W.X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, X. Li, Topical keyphrase extraction from Twitter, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), Association for Computational Linguistics, vol. 1, Stroudsburg, PA, USA, 2011, p. 379–388.