Journal of Network and Computer Applications 38 (2014) 16–21

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Text analysis for detecting terrorism-related articles on the web

Dongjin Choi [a], Byeongkyu Ko [a], Heesun Kim [b], Pankoo Kim [a,*]

[a] Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea
[b] Department of Internet & Multimedia Engineering, Konkuk University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 31 October 2012
Received in revised form
3 March 2013
Accepted 8 May 2013
Available online 18 May 2013

Keywords:
Text anlaysis
Document classification
Homeland security
Data mining

ABSTRACT

Classifying web documents is considered as one of the most important tasks to reveal the terrorism-related documents. Internet provides a lot of valuable information to the users and the amount of web contents is progressively increasing. This makes it very difficult to identify potentially dangerous documents. Simply extracting keywords from documents is not enough to classify the contents. To build automated document classification systems, many techniques have been studied so far, but they are mostly statistical and knowledge-based approaches. These methods, however, do not yield satisfactory results because of complexity of natural languages. To overcome this deficiency, we propose a method to use word similarity based on WordNet hierarchy and n-gram data frequency. This method was tested with the sampled New York Times articles by querying four distinct words from four different areas. Experimental results show our proposed method effectively extracts context words from the text and identifies terrorism-related documents.

1. Introduction

Text data is the most common content type on the web when it comes to author's opinion. Recently, following the progress of wireless internet and smartphone devices, the amount of data on the web is dramatically increasing with no constrain to time or location. There are more than eight billion web pages indexed by Google, Bing and Yahoo (http://www.worldwidewebsize.com). Such information wealth makes it possible to satisfy large a variety of user requests. Today, one can get almost any information on the web (Baek et al., 2008). Many researchers in computer sciences are committed to find more useful and efficient methods for providing appropriate results to users' demands. However, this huge amount of information can not only confuse people trying to find what they need, but also pose a serious security threat. Web documents use HTML for displaying information in the web browser. In other words, web data have no semantic information. Therefore, when people look for information on the web, they have to spend a lot of time to understand whether the returned web results are relevant or not (Hwang et al., 2011c). This is a heavy burden on humans, so it is needed that machines could do that automatically. Since people can easily access web data, personal information such as phone numbers and email address are leaking via the web (Hunker and Probst, 2011). Besides, documents such as Microsoft Compound Document File Format may contain personal information, too such as authors name, organizational information of users involved, and more (Castiglione et al., 2007). Currently, the Portable Document Format (PDF) was the most popular document format to submit or share document with other. However, it has diverse privacy related information which might cause information leakage problem (Castiglione et al., 2010). Moreover, terrorists are exchanging information by using internet infrastructure such as email to expand their organization networks. For example, it was found by using terror-related web information that 'Hamburg Cell' was mainly responsible for the preparation of the 11th September attacks against the United States (Corbin, 2002). This proves that it is possible to detect terrorism by analyzing web data to prevent further terrorist attacks. However, by simply extracting keywords and context words it is not possible to detect reliably terrorism-related activities. Many researchers used statistical methods such as Term Frequency (TF) or knowledge base, such as WordNet (Salton and Buckley, 1988; Hwang et al., 2011a,b,c; Kong et al., 2005). However, there is a limitation that the precision rate is not reliable based on word frequency and knowledge bases due to the fact that human written language is more than word frequency. Also when we applied knowledge based approach, the results will depend on the precision of the knowledge bases. In order for computers to understand human language, lots of experiments have been made by using *Bayes* theory (Pavlov et al., 2004), decision trees (Lewis and Ringuette, 1994), Latent Semantic Analysis (LSA) (Yu et al., 2008), Support Vector Machine (SVM) (Barzilay and Brailovsky, 1999) and so on. However, understanding texts is still a challenging and difficult task for computers.

[*] Corresponding author. Tel.: +82 10 9824 2100; fax: +82 62 230 7636.
E-mail addresses: dongjin.choi84@gmail.com (D. Choi), byungkyu.ko@gmail.com (B. Ko), aznturbo@konkuk.ac.kr (H. Kim), pkkim@chosun.ac.kr (P. Kim).

1084-8045/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.05.007

Therefore, we suggest a method of obtaining context words from training terrorism-related articles by using WordNet hierarchy. After that, bigram data frequency is calculated from context words sets. We trained four types of articles which are 'terror,' 'crime,' 'cancer,' and 'health.' The experiment aimed at distinguishing articles about terrorism by using *Keselj* distances (Keselj et al., 2003), which is one of the most famous n-gram based similarity measurement and Context Weight similarity (Choi et al., 2012; Loia et al., 2009).

This paper is organized as follows: in Section 2 we discuss some related works; Section 3 explains the suggested method; Section 4 outlines experimental results; and finally Section 5 presents a conclusion to this work and makes suggestions for the future work.

2. Related works

Data mining based on text analysis is considered as one of the key problems for many homeland security initiatives. Text analysis is used to discover unknown, valid patterns and relationships in large data sets. Even text analysis has a great potential to identifying unknown text documents, there is a limitation that human written language is still complicated for machine to understand semantic meanings of it. Over the years, many studies have been made by using statistical methods to represent documents into meaningful sequences, such as TF-iDF. This is the most basic method for determining which words are significant in the given text data set. However, performance state based on TF-iDF is not acceptable if the amount of data is too small (Li and Guo, 2010). Moreover, this approach depends on the bag of words to calculate TF-iDF value. In order to overcome this deficiency, researches started to use knowledge bases, such as WordNet (Miller, 1995) developed by Cognitive Science Laboratory of Princeton University or Wikipedia, which provides semantic relations among concepts. Hwang studied the possibility to determine semantic similarities and context information from abstracts in Wikipedia documents (Hwang et al., 2011b). His research proved that WordNet contains valuable information to build a semantic network between the words. Because WordNet provides the hierarchy of concepts as shown in Fig. 1, it is possible to measure how concepts are semantically related to one another. Choi studied text data sets in order to annotate images in web news by using WUP measurement in WordNet (Choi and Kim, 2012). This research assumed that text data around image is likely to describe the image itself. The study has proved that context words can be extracted by using WUP method. However, it does not tell the extracted context words can represent given documents precisely. It only can grasp possible words to discover potential key concepts to reduce human efforts for reading all sentences in given documents.

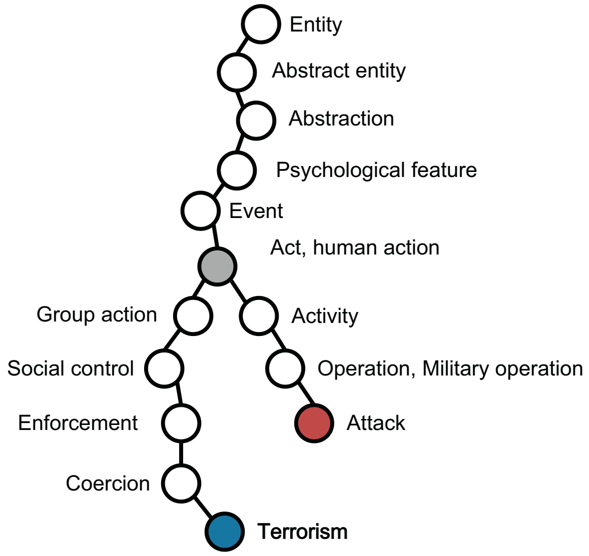

Fig. 1. Concept hierarchy between 'terrorism' and 'attack' in WordNet.

There is another approach that uses phrases rather than individual words. The n-gram is a co-occurrence statistics data which collect every adjacent word from training text documents. Because of linguistic advantages of n-grams, it is possible to extract core features of training documents written in natural language. Also, n-grams can be used for speech recognition (Siu and Ostendorf, 2000), topic discovery and filling incomplete sentences (Choi et al., 2011b) because the combination of n-grams was powerful in text categorization. N-grams can be expressed into two different ways. First is a method for extracting the n-gram by each adjacent English character from documents. Second is the n-gram by each adjacent word. First method has a big problem that the size of n-gram is too huge that it takes lots of times. Moreover, it is hard to find semantic meanings based on the first method. Accordingly, the second method is more used. Bigrams and Trigram are the most popular data types in terms of performance. In order to detect terrorist-related text articles, we built bigram data frequency from a bag of words sets, which were extracted by comparing WUP similarity between the title and the body contents of the given articles.

3. Text analysis for detecting terrorism-related articles

In this paper, we suggest a method for detecting terrorism-related text articles on the web by using the text analysis method. To this end we apply WordNet hierarchies for extracting context words from training documents and build bigram data frequencies for classifying unknown text data.

3.1. Extracting context words from training documents

In this section, we explain how context words are extracted from the given training text articles by using WordNet hierarchy. WordNet is one of the most famous knowledge bases created and maintained by the Cognitive Science laboratory of Princeton University under direction of psychology professor George A. Miller from 1985. It contains valuable information that helps determine semantic relationships between the words. For example, it provides synonyms, coordinate terms, hypernyms, hyponyms, and so on. These concepts hierarchy and semantic networks can be applied to determine semantic distance between the words. Figure 2 illustrates extracting context words from training documents by using WUP distance in WordNet (Wu and Palmer, 1994).

First, we collected the New York Times web articles obtained by using article search API[1] from January 1, 2000 to October 15, 2012 corresponding to queries 'terror,' 'crime,' 'cancer,' and 'health.' The number of articles was: 'terror' 14,172, 'crime' 23,001, 'cancer' 22,899, and 'health' 23,819. After deleting special characters, we obtained title and body parts of articles. The title in articles can be considered as the most important sentence to represent given training articles. The body part of the article expands the subject in detail. This is the main reason why we only considered the title and body part of articles. There is another fact that noun types of word have significant meaning to describe subjects or objects of sentences. Therefore, we only extract nouns from title and body.

[1] ⟨http://developer.nytimes.com/docs/read/article_search_api⟩.