Information Fusion 28 (2016) 90–98

Contents lists available at ScienceDirect

Information Fusion

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

Web news mining in an evolving framework

José Antonio Iglesias *, Alexandra Tiemblo, Agapito Ledezma, Araceli Sanchis

University Carlos III of Madrid, 28911, Leganes, Madrid, Spain

ARTICLE INFO

Article history:
Available online 30 July 2015

Keywords:
Big data
Web news mining
Evolving fuzzy systems

ABSTRACT

Online news has become one of the major channels for Internet users to get news. News websites are daily overwhelmed with plenty of news articles. Huge amounts of online news articles are generated and updated everyday, and the processing and analysis of this large corpus of data is an important challenge. This challenge needs to be tackled by using big data techniques which process large volume of data within limited run times. Also, since we are heading into a social-media data explosion, techniques such as text mining or social network analysis need to be seriously taken into consideration.

In this work we focus on one of the most common daily activities: web news reading. News websites produce thousands of articles covering a wide spectrum of topics or categories which can be considered as a big data problem. In order to extract useful information, these news articles need to be processed by using big data techniques. In this context, we present an approach for classifying huge amounts of different news articles into various categories (topic areas) based on the text content of the articles. Since these categories are constantly updated with new articles, our approach is based on Evolving Fuzzy Systems (*EFS*). The *EFS* can update in real time the model that describes a category according to the changes in the content of the corresponding articles. The novelty of the proposed system relies in the treatment of the web news articles to be used by these systems and the implementation and adjustment of them for this task. Our proposal not only classifies news articles, but it also creates human interpretable models of the different categories. This approach has been successfully tested using real on-line news.

1. Introduction

Modern society generates huge amounts of information every day, especially in digital format, which obstruct the storage and further processing and analysis. Big data can be defined as a scale of data set that goes beyond existing database management tool capabilities of data collection, storage, management, and analysis capabilities [1]. Although the most common trait of *big data* is Volume, it is typically defined by three *Vs* (Volume, Variety and Velocity).

In addition, big data can be classified taking into account the data type:

1. Structured (data are organized into a predefined data schema).
2. Semi-structured (data does not require a schema definition but the data includes metadata).
3. Unstructured (data are stored in an unstructured form without any defined data schema).

There are many different applications in which big data techniques are applicable: data mining, predictive analytics, geoanalysis, natural language processing and pattern recognition. Also, we are heading into a social-media data explosion. In this connection, web-based applications encounter big data frequently, such as *social computing*, Internet text and documents or Internet search indexing. For this reason, there are some techniques which need to be seriously taken into account such as social network analysis and text mining.

In particular, an important application of the text mining which could be also related with the social big data is: web news mining. Since news websites are daily overwhelmed with plenty of news articles, an important part of the huge amounts of *new* information produced each day is generated by the on-line newspapers. For this reason, automatic systems which can treat, analyze and classify web news articles are essential not only for those systems which manage web news articles but also for user recommendation tasks.

According to the Statistical Report on Internet Development released by China Internet Network Information Center (CNNIC) in July 2014 [2], the number of online news users in China had reached 503 million by the end of January 2014 (a growth of 98.60 millions from June 2012), and the utilization ratio of online

* Corresponding author.
E-mail addresses: jiglesia@inf.uc3m.es (J.A. Iglesias), 100073062@alumnos.uc3m.es (A. Tiemblo), ledezma@inf.uc3m.es (A. Ledezma), masm@inf.uc3m.es (A. Sanchis).

http://dx.doi.org/10.1016/j.inffus.2015.07.004
1566-2535/

news was 79.6%. In this report, the *online news* is the third most used network application by Internet users. The first application in this ranking is the *Instant messaging* (89.3%) and the second one is the *Search engine* (80.3%). However, news is the most frequently searched content by the Internet users using both computers and mobile phones. In addition, this report suggests that online news has become one of the major channels for Internet users to get news and its utilization ratio has remaining high due to the following reasons: (1) in the era of mobile Internet, it is one of major activities of Internet users to read news in their fragmented time, (2) Internet users can get news through more channels, and (3) all news media vied with each other to make inroads into the mobile Internet.

In the era of big data and because of this explosion of information from news websites, extracting knowledge from news articles becomes an interesting challenge. To that end, we need text mining techniques which can extract relevant information from this kind of unstructured type text data. In addition, online news is a special type of public information mainly because there are many news sources and the update of the news is very fast. News mining tools, techniques, and algorithms are strongly emerging during these times. There are many techniques which help to analyze the overflow of information and extract value knowledge from on-line news sources. However, since this information is continuously growing and changing, these techniques have to skim and search for information much more than they had to do in the past.

During the last years, there have been many approaches related with classification, clustering, categorization and summarization of news articles [3–8]. If we consider those approaches which classify news into predefined categories, all of them use a statistic classifier over time. However, the news articles of the different categories change constantly and these changes should be considered in the model of the classifier. For this reason, we propose an approach in which the categories are not predefined but they are updated in an evolving manner according the new news articles and categories obtained. This aspect makes our approach an ideal alternative in this environment.

The presented approach is based on Evolving Fuzzy Systems (EFS) [9] which allows not only update the structure and parameters of an evolving classifier but also cope with huge amounts of web news and process data in on-line and real time – which is essential in this (web) environment. EFS approaches have been successfully applied in many other different areas [10–15] and for big data problems [16].

The remainder of the paper is organized as follows: Section 2 describes existing researches and approaches related with the area of big data and web news mining. Sections 3–5 describe our proposed (evolving) approach for the classification of web news. Section 6 presents the results and analysis of the evaluation of our approach. Finally, Section 7 draws the conclusions and future work guidelines.

2. Background and related work

Many different scientific fields have become highly data-driven with the development of computer science. Social computing [17], astronomy [18] or bioinformatics [19] are some examples of these fields.

Big data uses different techniques to efficiently process large volume of data within limited run times. Because of the most common trait of big data is Volume, the most important challenge is scalability when we deal with the big data analysis tasks. In this sense, incremental algorithms have good scalability property [20,21]. If we focus on the disciplines of data mining and machine learning, we should consider that big data mining is more challenging compared with traditional data mining algorithms [22].

However, in the big data era, we have to consider that the most common format of information storage is text such as web pages, emails, documents or social media. For this reason, text analysis or text mining is a powerful technique at that time. The term *text mining* or *Knowledge Discovery from Text* (KDT) was mentioned for the first time in 1995 by Feldman et al. [23]. They propose to structure the text documents by means of information extraction, text categorization, or applying NLP techniques as pre-processing step before performing any kind of KDTs.

Text mining, also known as text data mining[24], can be defined as the analysis of semi-structured or unstructured text data. As the text is in unstructured form, it is quite difficult to deal with it. In fact, text mining is a much more complex task than data mining [25] as it involves dealing with text data which are inherently unstructured and fuzzy. Thus, the goal of the text mining is to turn text information into numbers so that data mining algorithms can be applied. It arose from the related fields of data mining, artificial intelligence, statistics, databases, library science, and linguistics. As it is detailed in [3], since text mining is a multidisciplinary field, this term has been used to describe different applications such as text categorization [26,27], prediction [28,29], text clustering [30,31], association discovery [32,33] and finding patterns in text databases [34].

In the text mining area, Twitter is considered as a rich source of information for text analysis. In [35], the authors find similarities between tweets before the World Cup started. The high-value social audience from Twitter is identified through text-mining methods [36]. In this case, the Twitter content of an account owner and its list of followers are analyzed. A survey on text mining and sentiment analysis for unstructured web data is presented in [37]. Mathioudakis et al. [38] propose *TwitterMonitor*, a system which detects topic trends in real time and provides meaningful analytics that synthesize an accurate description of each topic. Kim et al. propose in [39] a spatio-temporal trend detection and related keyword recommendation scheme for tweets called *TwitterTrends*. These scheme can identify keyworkds and recommmend related keywords at a given location and time.

Other application of the text mining is: Web news mining. This term describes the analysis of web news and is a special type of public information which has special characteristics [40]. The existence of numerous reliable news sources and fast news updates are two important differences. For this reason, new approaches, technologies and tools need to be developed in order to achieve the different goals proposed in this area.

During the last years, there have been many approaches related with web news mining and news exploration systems. In [6], the authors describe the use of data mining techniques to analyze web news. It is concluded from that study that web mews mining at the terms level serves as a powerful technique to manage knowledge encapsulated in large web news collection. As in our approach, the authors analyze web news by using text mining. However, that research only implements the process of terms extraction from the web news. Our approach, not only analyzes web news but also classifies them in a specific topic.

In [41] the authors propose a flexible topic-driven framework for news exploration. It performs news mining at the topic level and presents news information with topics, entities and relations derived from the news data. Also, the authors consider that in order to facilitate an in-depth analysis of the news it is necessary to extract structured information (ideally, identifying *who, what, whom, when, where* and *why* [42]). In [43], it is presented an endeavor aiming at construction of a real-time event extraction system for border security-related intelligence gathering from online news. In [44] a quantitative method that