Information Fusion 20 (2014) 129–135

Contents lists available at ScienceDirect

Information Fusion

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

Detecting trends on the Web: A multidisciplinary approach

Rodrigo Dueñas-Fernández [a], Juan D. Velásquez [a,*], Gaston L'Huillier [b]

[a] Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box: 8370439, Santiago, Chile
[b] Groupon Inc., 3101 Park Blvd., Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form 19 November 2013
Accepted 6 January 2014
Available online 2 February 2014

Keywords:
Trend detection
Web opinion mining
Topic modeling

ABSTRACT

This paper introduces a framework for trend modeling and detection on the Web through the usage of Opinion Mining and Topic Modeling tools based on the fusion of freely available information. This framework consists of a four step model that runs periodically: crawl a set of predefined sources of documents; search for potential sources and extract topics from the retrieved documents; retrieve opinionated documents from social networks for each detected topic and extract sentiment information from them. The proposed framework was applied to a set of 20 sources of documents over a period of 8 months. After the analysis period and that the proposed experiments were run, an F-Measure of 0.56 was obtained for the detection of significant events, implying that the proposed framework is a feasible model of how trends could be represented through the analysis of documents freely available on the Web.

1. Introduction

In the current state of the art, the only widely used method for inferring trends and trying to assert if a topic is becoming a trend is to run surveys over a sparse set of individuals. Nevertheless, the use of surveys for such purpose has some caveats that need to be addressed in order for it to become a more useful tool than what it is nowadays. For example, it is highly likely that the background and attitude of the interviewers interfere with the results of the survey itself, thus making the results biased [1]. Therefore, there is an unfulfilled need to complement the existing methodologies for trend detection and any other traditional approach focused on gathering knowledge regarding events that are occurring daily. In particular, the usage of the freely available information that exists on the Web allows us to access a significant number of people freely expressing their opinion [2] that could not have been reached otherwise.

In this study we propose a generic framework that allows using the data available on the Web towards the detection and modeling of trends over time. With that objective in mind, a multidimensional approach was taken, where a topic will be monitored over time according to how much media coverage it gets and how the users of social networks react to it. Thus, the proposed framework consists of two main components, the one focused on extracting topics and monitoring their evolution in the traditional media, and the second one in charge of extracting valuable information about how users feel and the opinions they publish on social networks.

To the best of our knowledge, there are no unified methodologies that try to tackle the trend detection problem from a multidisciplinary approach, i.e. that take into consideration both the problems of Topic Modeling and Opinion Mining. Nevertheless, every problem mentioned above within the framework of trend detection has been approached over the last years by different disciplines. The branch of knowledge that involves both issues is Information Retrieval, which focuses on retrieving documents from the Web and process them to be able to analyze the data contained within them.

Under its practical and theoretical frameworks, information retrieval has presented several techniques that allow the retrieval and processing of relevant documents. In terms of detecting the topics of such documents and the specific events or features about which opinions have been expressed, text mining and natural language processing communities have developed a vast set of models to determine what are the topics being discussed across a collection of documents [3]. Finally, the field of Web Opinion Mining has presented several approaches to represent the polarity of documents posted on the Web by their users [4], whether they come from traditional media or social media, the latter usually having less structured content. The main contribution of this work is to integrate these disciplines into one unified framework that allows us to monitor trends on the Web using information present in both the traditional media and the social media (such as social networks).

* Corresponding author. Tel.: +56 2 2978 4834; fax: +56 2 2689 7895.
E-mail addresses: rduenas@ing.uchile.cl (R. Dueñas-Fernández), jvelasqu@dii.uchile.cl (J.D. Velásquez), gaston@groupon.com (G. L'Huillier).
URL: http://wi.dii.uchile.cl/ (J.D. Velásquez).

http://dx.doi.org/10.1016/j.inffus.2014.01.006
1566-2535/© 2014 Elsevier B.V. All rights reserved.

This paper is organized as follows. In Section 2 a brief summary of related research is provided. Section 3 describes the proposed methodology for detecting trends on the Web. Section 4 outlines the experiments performed with the proposed methodology and Section 5 provides some conclusions and suggests future research.

2. Related work

Although there is no unique definition of what a trend is, how it should be represented and how a topic becomes a trend, several approaches have been proposed by multiple authors to detect trends or the evolution of topics over time in specific areas. For example, applications in politics are presented in [5,6] and finance in [7]. Due to the broad definition of what a trend is, there have not been a significant number of attempts to develop generic frameworks that go beyond a singular application domain, or even monitor how topics evolve in multiple areas. A basic example is presented in [8], where the main focus of their research is to show an approach towards building a trend detection framework on top of a cloud computing architecture, rather than proposing a framework capable of retrieving documents and deciding whether a topic presented in several documents over time reflects a trend or not.

In terms of information retrieval, there is a vast amount of literature that provide some insights about the different types of data and how this data should be handled [9]. Research has been done in information retrieval frameworks for the detection of trends in blogging [10], microblogging (e.g. Twitter) [11] and social networking sites (e.g. Facebook) [12]. Once the data has been retrieved and stored, the textual information has to be processed in such a way that the underlying patterns are extracted for further usage. In this domain, keyword-based analysis approaches have been proposed in the specific context of Web usage mining [13,14]. Also, several approaches focused on the detection of unknown events or determining the impact of news online [15,16] have been proposed.

However, one of most relevant techniques used in recent years for modeling how events evolve over time are topic models [17]. A topic model can be considered as a probabilistic model that relates documents and words through variables, which represent these main topics, inferred from the text itself. In this context, a document can be considered as a mixture of topics, represented by probability distributions that generate the words that belong to a document that contains these topics. The process of inferring the latent variables, or topics, is the key component of this model, whose main objective is to learn the distribution of the underlying topics from text in a given corpus of text documents. A main topic model is the latent Dirichlet allocation (LDA) [18]. LDA is a Bayesian model in which latent topics of documents are inferred from estimated probability distributions over a training data set.

Opinion mining and sentiment analysis is a field whose objective is to consider a collection of opinionated documents and determine the orientation (positive, negative, and objective) of an opinion about a particular aspect of an entity at a given time [19]. In terms of trends detection, this task is fundamental in order to identify whether the trending topics are being generated with a certain opinion orientation. In this work, the opinion mining step will be focused in the usage of lexicon-based algorithms. Algorithms that are based on the use of lexicons can be found in [1,2], which according to the research presented in [20] can return valuable information in the context of mining opinions of documents retrieved from microblogging sites. It should be noted that opinion mining algorithms that focus on the classification of polarity face several challenges, such as irony and sarcasm linguistics [21,22] and also the amount of text available in a document [23].

3. A methodology for trend detection on the Web

In this chapter, the methodology proposed to detect trends on the Web is presented. First, the definition of the problem to solve and every term used throughout this paper are detailed. Next, some of the main text analysis techniques used during the development of the proposed methodology for detection of trends on the Web are discussed. Finally, the methodology itself and the main contribution of this work are described.

3.1. Problem definition and general notation

In the following, the term *Trend* will be presented together with its ontological and linguistic representation. In this context, a *trend* will be defined as a given event whose impact on a system as a whole, is above the average over a certain period of time.

The problem of detecting trends on the Web is described. Several research areas are focused on modeling the so-called collective behavior in order to, for example, predict how important events will develop, which politician will win a debate, which football team will win a match and so on.

Even though existing methodologies to predict trends and monitor their evolution over time have been successfully applied to a large variety of problems, there's a vast amount of information not being used, created by users on the Internet, where the act of expressing one's opinion or feelings is not restrained by the common issues that are found in the standard methodologies such as limited time and biased answers based on the person running the interview.

The objective of this research is to tackle what will be called as the *trend detection problem*, which is defined as:

Definition 1 (*Trend Detection Problem*). Given a set of topics, to determine if the way they behave over time makes them qualify as a trend.

In this research the following definition of *trend* will be used:

Definition 2 (*Trend*). A trend is a given event or topic whose impact on a system as a whole, is above the average over a certain period of time. Furthermore, for a system composed by a chain of events, a trend is defined by its expected future behavior given how it behaved in the past and how it reacts to external stimulus.

In this study, a *factual document* is a document that contains no opinion whatsoever and refers to one or more events. On the other hand, the term *opinionated document* refers to any opinionated document whose subject is an event. Examples of these types of documents are tweets and opinion columns in journals, among other personal opinions expressed on the Web by one of its users.

As the detection of trends in most useful scenarios is always framed within a certain domain of knowledge, a similar approach will be taken in our methodology, where the set of websites to be crawled for documents is defined beforehand and they are expected to belong to specific domain of knowledge. Each of these sources of documents will be referred to as *feeds*.

In order to be able to create a more descriptive model of topic evolution, an information fusion [24] approach was taken, in which their evolution is measured by a multidimensional analysis based on information retrieved from *factual documents* and *opinionated documents* extracted from several sources. The proposed methodology for detecting and modeling trends on the Web consists of four main steps that are executed periodically and then complemented by the visualization of the extracted data. These steps are: