Expert Systems With Applications 57 (2016) 21–36

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A step forward for Topic Detection in Twitter: An FCA-based approach

Juan Cigarrán*, Ángel Castellanos, Ana García-Serrano

Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

ARTICLE INFO

Article history:
Received 11 February 2015
Revised 7 March 2016
Accepted 8 March 2016
Available online 19 March 2016

Keywords:
Formal Concept Analysis
Stability
Topic Detection
Online Reputation Management

ABSTRACT

The Topic Detection Task in Twitter represents an indispensable step in the analysis of text corpora and their later application in Online Reputation Management. Classification, clustering and probabilistic techniques have been traditionally applied, but they have some well-known drawbacks such as the need to fix the number of topics to be detected or the problem of how to integrate the prior knowledge of topics with the detection of new ones. This motivates the current work, where we present a novel approach based on Formal Concept Analysis (FCA), a fully unsupervised methodology to group similar content together in thematically-based topics (i.e., the FCA formal concepts) and to organize them in the form of a concept lattice. Formal concepts are conceptual representations based on the relationships between tweet terms and the tweets that have given rise to them. It allows, in contrast to other approaches in the literature, their clear interpretability. In addition, the concept lattice represents a formalism that describes the data, explores correlations, similarities, anomalies and inconsistencies better than other representations such as clustering models or graph-based representations. Our rationale is that these theoretical advantages may improve the Topic Detection process, making them able to tackle the problems related to the task. To prove this point, our FCA-based proposal is evaluated in the context of a real-life Topic Detection task provided by the Replab 2013 CLEF Campaign. To demonstrate the efficiency of the proposal, we have carried out several experiments focused on testing: (a) the impact of terminology selection as an input to our algorithm, (b) the impact of concept selection as the outcome of our algorithm, and; (c) the efficiency of the proposal to detect new and previously unseen topics (i.e., topic adaptation). An extensive analysis of the results has been carried out, proving the suitability of our proposal to integrate previous knowledge of prior topics without losing the ability to detect novel and unseen topics as well as improving the best Replab 2013 results.

1. Introduction

Every day millions of user-generated content are added to the Web. Information related to companies, persons, products and services makes up a large part of that content. These data are especially interesting for mining user opinions on some entity (item, product, company, celebrity service, etc.). The set of user opinions about an entity is also referred to as entity reputation and the tracking and management of these opinions to gain some insight into the reputation of the entities is known as Online Reputation Management (ORM). Companies are especially interested in knowing what their customers think about them for the early detection of events potentially damaging to the company reputation.

ORM is implemented by analysing all the information on an entity or company on the Web. Among all the web sources containing these data, social networks are one of the most interesting. In this regard, the vast majority of studies conducted in the field of ORM are based on Twitter data because of the easy access to the data (through its API), the immediateness of the content (many times faster in reflecting news than the mainstream media) or the huge amount of data available (Laboreiro, Sarmento, Teixeira, & Oliveira, 2010), among other reasons.

ORM involves several related tasks (Amigó et al., 2013a): Filtering, Polarity, Topic Detection or Alert Detection. The work proposed in this paper focuses on Topic Detection and the different aspects involved in it. Topic Detection refers to the finding of topics in data on a company, product or service. These topics will be valuable, for instance, in identifying "trendy-opinion" streams, divide content or users into interest groups or warn about some risk to the reputation or the entity, based on the appearance of a controversial topic. Topic Detection research started years ago, mainly motivated by the interest in managing information contained in datasets. One of the first forums was the Topic Detection and Tracking (TDT) Forum, held within TREC (Fiscus & Doddington, 2002). TDT pursued

* Corresponding author. Tel.: +34 91 398 7620.
E-mail addresses: juanci@lsi.uned.es (J. Cigarrán), acastellanos@lsi.uned.es (Á. Castellanos), agarcia@lsi.uned.es (A. García-Serrano).

http://dx.doi.org/10.1016/j.eswa.2016.03.011
0957-4174/

the discovery and threading together of topically related content in broadcast news. The works proposed within the scope of the TDT task have proven to be relatively satisfactory for Topic Detection in regular textual content (Allan et al., 1998). However, from these seminal works, the focus has moved to Social Network data sources, and especially Twitter. Different methodologies have been proposed in the state of the art to tackle these new environments. In the following we explain the most noteworthy approaches proposed up to date.

The first approaches conducted for Topic Detection successfully applied classification techniques in different scenarios (Bengel, Gauch, Mittur, & Vijayaraghavan, 2004; Kumaran & Allan, 2004; Wayne, 2000). However, dealing with tweets involves some considerations that potentially limit the performance of the traditional classification algorithms. Some of them are explained in Anta, Chiroque, Morere, and Santos (2012): the existence of special signs (i.e., abbreviations, emoticons or hashtags), use of slang, brevity of the content (limited by Twitter) or spelling mistakes. In spite of these issues, classification-based approaches have been widely applied. Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) propose a tweet categorization (e.g., news, opinions, personal messages, events and deals) and a classification algorithm using Twitter related metadata about the authors (e.g. name, information in the profile). Phan, Nguyen, and Horiguchi (2008) try to deal with classification drawbacks by using external sources (Wikipedia and MEDLINE) to expand the tweet content in order to increase the features available. Other solutions proposed also include: tweet tokenization (Laboreiro et al., 2010), stemming, spelling analysis and use of dictionaries (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). This supervised methodology has been also studied in recent researches from the point of view of Expert Systems. In this regard, Vo and Ock (2015) propose the classification of scientific documents according to their title, which is enriched with topics detected in scientific datasets (e.g., DBLP, LNCS) and Wikipedia. In the same line, Uysal (2016) proposes an Improved Global Feature Selection Scheme (IGFSS) to be applied for Text Classification.

In addition to the Twitter-dependant problems, classification algorithms present another problem that limits their application: it is a supervised methodology. It means that the algorithms need to be trained in order to be able to classify new inputs. This methodology has a high performance in classifying content according to the features already seen in the training set. But, if new content presents new features, these will not be taken into account for the classification process. In the field of Topic Detection it means that if new topics including new features, unseen in the training set, appear, they will not be correctly classified.

On the other hand, clustering techniques follow an unsupervised methodology. Thus, they do not need to be trained in order to compute the categorization, so discovering the topic will not be restricted to the training data. In the context of Topic Detection, several works make use of clustering techniques. Some of the most noteworthy clustering-based approaches are: that in Sankaranarayanan, Samet, Teitler, Lieberman, and Sperling (2009), where the authors present the Tweet Stand system to cluster tweets into trending news; that presented in Vakali, Giatsoglou, and Antaris (2012), also focused on detecting “trendy” content, in which the authors propose a clustering framework (called Cloud4Trends) including contextual information on the tweets' authors. The interesting aspect of this latter work is that the clustering-based Topic Detection is only used as a previous step for the later detection of “trendy” news. Nevertheless, the data to be clustered rarely presents a flat hierarchy. To deal with this aspect, hierarchical clustering methodologies have been proposed in the context of Topic Detection. For instance, Zeng, Wu, and Wang (2010) propose an extension of the traditional hierarchical clustering algorithms based on topic grain computation. The same authors propose in Zeng, Duan, Wang, and Wu (2011) an extension of this approach based on the learning of semantic grain topic models by applying Discrete Cosine Transform.

In the context of the Replab Campaign (Amigó et al., 2013a) some clustering proposals have been also presented, focusing on the application of Topic Detection approaches in a real environment. The UAMCLyR group proposed an approach based on a novel term selection methodology for the clustering algorithm (Sánchez-Sánchez, Jiménez-Salazar, & Luna-Ramírez, 2013). The UNED ORM group proposes a tweet expansion using Wikipedia content to refine the operation of the clustering algorithms (Spina et al., 2013). Hierarchical clustering methodologies have also been proposed in this context, like the work in Spina, Gonzalo, and Amigó (2014), where the authors propose the learning of similarity functions for their latter application in a Hierarchical Agglomerative Clustering.

In the context of unsupervised approaches, the Topic Detection has also been presented as a matrix factorization problem; in particular, the factorization of the document-term matrix. In this regard, Non-Negative Matrix Factorization (NMF) has been applied to the Topic Detection task (Arora, Ge, & Moitra, 2012) and it has been demonstrated to regularly produce more coherent topics than other methodologies (O'Callaghan, Greene, Carth, & Cunningham, 2015). NMF is a technique for decomposing a non-negative matrix V into two non-negative factors W and H. In the context of Topic Detection, W and H are reduced k factors of the original document term matrix (V), where the k factors are considered the latent topics to be detected.

Nevertheless, in recent years, probabilistic methodologies, and especially LDA (Blei, Ng, & Jordan, 2003), have emerged as almost a de facto standard for Topic Detection. The application of probabilistic techniques, also known as probabilistic Topic Modelling (pTM), tries to find the latent semantic space to group content together (in topics), according to the shared latent semantic. In fact, some pTM approaches especially adapted to the specific Twitter environment have been developed. For instance, in Yang and Rim (2014) the so-called Trend Sensitive LDA is proposed to reflect the changes in the trends over time or in Ramage, Dumais, and Liebling (2010) wherein the authors use the labelled information of a tweet to detect its latent topics by means of a methodology called Labeled-LDA. Some other interesting works in the literature which apply probabilistic methods for Topic Detection are (AlSumait, Barbara, & Domeniconi, 2008; Anthes, 2010; Huang, Yang, Mahmood, & Wang, 2012; Zeng, Duan, Cao, & Wu, 2012) or (Guo, Xiang, Chen, Huang, & Hao, 2013).

Although clustering and especially probabilistic techniques have been broadly applied to the Topic Detection task, some open questions remain. As has been previously raised by Chemudugunta and Steyvers (2007), probabilistic based methodologies like LDA tend to over-generalize; that is, to generate quite generic topics. Another important issue is: How many topics are there, and consequently, how many clusters should be generated? Guo et al. (2013). This problem has been addressed by means of, for example, the analysis of the kernel matrix for clustering algorithms (Honarkhah & Caers, 2010) or the so-called Hierarchical Dirichlet Processes (HDP) for probabilistic methods (Teh, Jordan, Beal, & Blei, 2006). In this sense, in other areas of research some novel approaches have also been proposed in relation to hierarchical representations, like, for instance, that presented by Yang, Wen, Kinshuk, Chen, and Sutinen (2015), which proposes a Bayesian Topic Modelling applied to the multi-document summarization task. Zeng et al. (2012) noticed another problem related to probabilistic methodologies. They prove that the Dirichlet prior on multinomial parameters, commonly applied in LDA, is not enough for providing well-formed topic descriptions. To address this issue, they present a topic modelling approach to improve the standard LDA framework. In more detail,