Expert Systems with Applications 42 (2015) 6472–6485

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Unsupervised topic discovery in micro-blogging networks

Carlos Vicient *, Antonio Moreno

*Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA), Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain*

ARTICLE INFO

*Article history:*
Available online 13 April 2015

*Keywords:*
Knowledge-based systems
Semantic Web
Micro-blogging
Twitter

ABSTRACT

Unsupervised automatic topic discovery in micro-blogging social networks is a very challenging task, as it involves the analysis of very short, noisy, ungrammatical and uncontextual messages. Most of the current approaches to this problem are basically syntactic, as they focus either on the use of statistical techniques or on the analysis of the co-occurrences between the terms. This paper presents a novel topic discovery methodology, based on the mapping of hashtags to WordNet terms and their posterior clustering, in which semantics plays a centre role. The paper also presents a detailed case study in the field of Oncology, in which the discovered topics are thoroughly compared to a golden standard, showing promising results.

## 1. Introduction

Micro-blogging services such as Twitter constitute one of the most successful kinds of applications in the current Social Web. Every day more than 500 million tweets are sent, providing up to date information about any imaginable domain of knowledge (Twitter, 2014). Each tweet is a string of up to 140 characters that may basically contain text, links, user mentions and *hashtags* (strings preceded by the # symbol with which users tag their messages). In the last years there has been a growing interest in the design and development of tools that allow users to analyse large unstructured repositories of user-tagged data in order to discover and extract meaningful knowledge from them (Aiello et al., 2013; Teufl & Kraxberger, 2011). The determination of the main topics of interest in a collection of tweets may be a useful first step to sort them and address the problems of data visualisation, semantic (not keyword-based) information retrieval, information extraction, detection of users with similar interests, hashtag recommendation, etc. (Bhulai et al., 2012; Cotelo, Cruz, & Troyano, 2014; Kywe, Hoang, Lim, & Zhu, 2012). One of the main uses of hashtags is the categorisation of tweets, because (ideally) all the tweets that share the same hashtag should somehow refer to the same topic (e.g. the tweets with the hashtag #WorldCup2014 are related to facts, events, comments or opinions about the Football World Cup in Brazil in 2014). Thus, one of our working hypotheses is that the automated clustering of the hashtags present in a set of tweets may lead to a straightforward discovery of its main topics. However, grouping hashtags automatically in an unsupervised way turns out to be a very complex task, even if all the tweets belong to a certain domain of discourse (e.g. Oncology, the area of the case study developed in Section 4).

There are two main reasons that hamper the construction of groups of related hashtags. The first one is that users can freely annotate tweets without any restriction in their choice of hashtags. This means that, by nature, hashtags are unstructured and unlimited. They lack any form of explicit organisation or normalisation and, as a consequence, retrieval tasks and classification methods have to deal with basic problems like synonymy (different hashtags might have been used for the same concept, e.g. #illness and #disease) or polysemy (the same tag can have different meanings in different contexts, e.g. the term #operation may refer to "surgical treatment" but also to "the act of causing to function", "an action", etc.). There may also be lexically similar hashtags that do not have exactly the same meaning (#pharmaceuticals, #pharmaceutical, #pharmacy, #pharmacology, #pharma); thus, standard stemming techniques used in Natural Language Processing may lead to wrong results. Moreover, tags may also be acronyms (#HIV – human immunodeficiency virus, #AIDS – acquired immunodeficiency syndrome), named entities (#MayoClinic, #AustinCancerCenter), a combination of several words (#HighBloodPressure), an expression of a feeling (#CancerSucks), or just invented words or even pure nonsense. All these issues present a big challenge and most of the topic discovery methods described in the current literature are not able to deal with them.

The second reason, as will be shown in the next section, is that current hashtag clustering methods are mostly based on a

* Corresponding author. Tel.: +34 977256563; fax: +34 977559710.
*E-mail addresses:* carlos.vicient@urv.cat (C. Vicient), antonio.moreno@urv.cat (A. Moreno).

http://dx.doi.org/10.1016/j.eswa.2015.04.014
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

syntactic analysis of their co-occurrence (Vicient & Moreno, 2013). This kind of analysis presents several problems, just to name a few:

- As tweets are very short, it is uncommon to use more than one hashtag in a tweet; in fact, some studies indicate that roughly 16% of them contain at least one hashtag (Mazzia & Juett, 2011). Therefore, the hashtag co-occurrence matrix is usually very sparse.
- A purely syntactic analysis will always treat a polysemic hashtag in the same way, without distinguishing its different meanings.
- Synonymous hashtags will hardly co-occur and they will not be assigned to the same cluster. For example, the terms "car" and "automobile" will unlikely appear together in the same sentence (especially with a length up to 140 characters).
- The meaning of acronyms will not be taken into account.
- The components of a multi-word hashtag will not be separately considered (e.g. the relation between #Cancer and #LungCancer will not be obvious, as they will just be treated as two different strings).
- General concepts and named entities will be analysed in the same way, as mere strings of characters.

The main hypothesis of this paper is that the incorporation of semantic information, i.e. the analysis of the actual meaning of the hashtags, may help to alleviate these issues and to make a better clustering of them, which will lead to an improved identification of the topics underlying a tweet set. The linkage between a term (e.g. a hashtag) and its meaning (a concept in a background knowledge structure, typically a domain ontology) is called *semantic annotation* according to the Semantic Web paradigm (Berners-Lee & Hendler, 2001). Having solved this task, one may apply an ontology-based semantic similarity measure to group related terms.

The new topic discovery method proposed in this paper is thus based on the semantic annotation of hashtags supported by well-known knowledge repositories like WordNet and Wikipedia. The contributions of this paper are threefold:

- A novel procedure to link hashtags to WordNet synsets is defined.
- A new methodology to perform an automatic unsupervised semantic clustering of the set of hashtags contained on a given set of tweets is proposed.
- It is explained how to analyse the resulting hierarchy in order to identify the classes that are really significant, filtering the huge amount of noise present in a hashtag set.

The rest of the paper is structured as follows. The next section comments previous works related to topic detection in social networks. Section 3 explains the new methodology of analysis, which is composed of three basic steps: mapping hashtags to concepts, clustering hashtags according to the semantic similarity between their associated concepts, and filtering the relevant classes of hashtags. Section 4 presents an application of the methodology to a corpus of tweets related to Oncology, in which encouraging results have been obtained. The final section discusses the work and comments future lines of work.

## 2. Related work

This section provides a survey of the three basic kinds of techniques that have been proposed to detect the main topics of interest within a set of messages exchanged in a social network (Aiello et al., 2013).: probabilistic models, document-pivot approaches and feature-pivot methods. The following subsections comment in more depth the main characteristics of the three basic kinds of approaches and introduce the more recent related work in this area.

### 2.1. Probabilistic models

Probabilistic topic modelling algorithms aim to discover, based on historical data, the hidden thematic structure in large archives of documents. They analyse (by means of statistical methods) the words that appear in a document in order to discover the underlying topics. The main advantage of these methods is that they do not require any prior annotation or labelling of the documents because topics are supposed to emerge directly from the analysis of the original documents.

The simplest and well-known topic model is the *Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Jordan, 2012), although other statistical models, such as Hidden Markov Models, have been proposed to discover topics on collections of document (Sista, Schwartz, Leek, & Makhoul, 2002). LDA is a statistical model of document collections that aims to capture the intuition that documents exhibit multiple topics (understanding a topic as a distribution over a fixed vocabulary). Usually, in a collection of related documents, they share the same set of general topics but each document by itself exhibits those topics in different proportions. The observed variables are the words of the documents, the hidden variables are the topic structure and the generative process defines a joint probability distribution over both the observed and hidden random variables. The *posterior* distribution is the conditional distribution of the hidden variables given the values of the observed variables. Unfortunately, it may not be directly computed because the number of possible topic structures is exponentially large.

As probabilistic models of language are typically driven by long-term dependencies between words, they use LDA to extract *semantic concepts*, understood as probability distributions over words that tend to co-occur. However, it may be intuitively realised that, due to the particular characteristics of tweets, co-occurrence-based models will not provide as good results in Twitter as in the study of standard long documents. In Rajani, McArdle, and Baldridge (2014) they propose a variant of LDA, called the *Author-Recipient-Topic model*, in which the probabilistic distributions of words are conditioned to the document's authors and recipients. This model is shown to present better results than LDA when the number of topics is large (over 300).

Other works that use probabilistic models to analyse Twitter messages are *TWITOBI* (Kim & Shim, 2011) and its extension *TWILITE* (Kim & Shim, 2014). They propose a recommendation system for Twitter, using probabilistic modelling based on LDA and matrix factorization, which recommends the top-K users to follow and the top-K tweets to read for a user. In *TWITOBI*, the model estimates the probability that a user *u* generates a word *w* in his/her tweets, whereas *TWILITE* is an algorithm that estimates the topic preference distributions of users to generate tweet messages as well as the latent factor vectors of users to establish friendship relations. Ma, Sun, Yuan, and Cong (2014) propose the use of a related mechanism, *Probabilistic Latent Semantic Analysis*, to discover the probabilistic distribution of words and hashtags for each topic. Ramage, Dumais, and Liebling (2010) present a supervised learning model called *Labelled LDA* that maps the content of the Twitter feed into four dimensions: events, ideas, things or people (substance), social communication (social), personal updates (status) and broader trends of language use (style). The posts of individual users can be mapped to one of these four categories, giving a mechanism to characterise them.

One of the main shortcomings of LDA-based models is their high computational cost when they have to manage a large dataset; in consequence, they must be parallelized to scale with the