Information Fusion 28 (2016) 60–74

Contents lists available at ScienceDirect

Information Fusion

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

Time Aware Knowledge Extraction for microblog summarization on Twitter

Carmen De Maio, Giuseppe Fenza, Vincenzo Loia [*], Mimmo Parente

Department of Computer Science, University of Salerno, Fisciano, SA, Italy

ARTICLE INFO

Article history:
Available online 20 June 2015

Keywords:
Microblog summarization
Time awareness
Fuzzy Formal Concept Analysis
Big Data
Social media analytics

ABSTRACT

Microblogging services like Twitter and Facebook collect millions of user generated content every moment about trending news, occurring events, and so on. Nevertheless, it is really a nightmare to find information of interest through the huge amount of available posts that are often noisy and redundant. In the era of Big Data, social media analytics services have caught increasing attention from both research and industry. Specifically, the dynamic context of microblogging requires to manage not only meaning of information but also the evolution of knowledge over the timeline. This work defines Time Aware Knowledge Extraction (briefly TAKE) methodology that relies on temporal extension of Fuzzy Formal Concept Analysis. In particular, a microblog summarization algorithm has been defined filtering the concepts organized by TAKE in a time-dependent hierarchy. The algorithm addresses topic-based summarization on Twitter. Besides considering the timing of the concepts, another distinguishing feature of the proposed microblog summarization framework is the possibility to have more or less detailed summary, according to the user's needs, with good levels of quality and completeness as highlighted in the experimental results.

1. Introduction

Context. Nowadays, microblogging streams are useful to detect and track political events [1], media events [2], and other real world events [3]. Nevertheless, it is really difficult to understand the main aspects of the news or events inquiring these microblogging services. In fact, given a specific topic on Twitter a huge amount of relevant tweets that are redundant or not relevant due to the ambiguity and noise of the social media exists. Furthermore, the dynamic context of microblogging requires to manage not only meaning of information but also the evolution of knowledge over time. To face with this side effect many applications have been realized on Twitter, like Tweetchup (tweetchup.com), Twitalyzer (twitalyzer.com), which provide social media analytics services to detect and track trending topics. Moreover, automatic microblog summarization algorithms that extend Latent Diriclet Allocation (LDA) exist that consider both chronological order of the tweets and their information content, but at two distinct stages [4–6]. In the light of the described scenario, this work defines topic-based microblog summarization framework extending Fuzzy Formal Concept Analysis to manage time relations among tweets and introducing two main distinguishing features, that are specifically: first considering both time and meaning of the tweet at the same time to analyze knowledge evolution over the timeline; and second providing summaries with different level of detail according to the user's needs exploiting the peculiar properties of timed fuzzy lattice.

Problem. Formally, this work tries to face the following problem. Given a topic-focused timestamped tweet stream and a level of *detail d*, the task is aimed to filter and chronologically order tweets in order to produce a Microblog Summary MS_d that provides a complete description of the story covering main concepts describing topic development over the timeline. The proposed framework is able to retrieve more or less detailed summary according to the user's demand in terms of the level of *detail.*

Proposed Solution. This work defines a Time Aware Knowledge Extraction (briefly, TAKE) as a new methodology to solve the problem of topic-based microblog summarization focusing here on Twitter to give experimental evidence. In particular, a temporal extension of Fuzzy FCA in order to arrange resources (i.e., tweets) into a hierarchy of time dependent concepts, that is a timed fuzzy lattice, is defined. This enhancement makes Fuzzy FCA theory more suitable to deal with microblog considering temporal dimension. The final summarization process is achieved taking into account both timestamps and semantics of the tweets.

[*] Corresponding author.
E-mail addresses: cdemaio@unisa.it (C. De Maio), gfenza@unisa.it (G. Fenza), loia@unisa.it (V. Loia), parente@unisa.it (M. Parente).

http://dx.doi.org/10.1016/j.inffus.2015.06.004
1566-2535/© 2015 Elsevier B.V. All rights reserved.

Firstly, temporal peaks of tweet frequency analyzing timestamps and exploiting the Offline Peak-Finding Algorithm (OPAD), proposed in [7], are identified.

Secondly, content will be annotated via sentence wikification that is the practice of representing a sentence with a set of Wikipedia concepts (i.e., entries)[8,9]. Wikify service,[1] provided by University of Waikato, enable us to face with the short length nature of microblogs, specifically we did not rely on co-occurrences of words in the tweets to identify a topic. it is the practice of representing a sentence with a set of Wikipedia concepts. Wikification enables us to recognize sense of main concepts and named entity mentioned in the tweet associating a Wikipedia link and a corresponding weight representing confidence degree of the disambiguation result. Sentence wikification has been already exploited for text summarization [10,9,11] and it reveals to be not compromised by the short nature of the sentence. Wikification is an alternative approach with respect to co-occurring words analysis usually exploited in the literature and less suitable, in our opinion, to address the short nature of the tweet.

Then, taking into account the meaning of the tweet content and time dependences among detected peaks, temporal extension of Fuzzy Formal Concept Analysis [12,13] will be performed in order to arrange tweets into a hierarchy of time dependent concepts, that is a *timed fuzzy lattice*. Finally, a summarization algorithm has been defined exploring resulting timed fuzzy lattice knowledge structure. The algorithm extracts chronologically ordered tweets summarizing main concepts of the story according to their temporal evolution.

Motivation. Analogously to other unsupervised methods, like Topic Modeling and clustering, the FCA theory groups together resources that share their meaning (or topics) that is essentially represented by set of strictly related words (e.g., Bill Clinton, Election, Politics, and so on). Nowadays, Topic Modeling has been widely applied to ontology extraction [14], text categorization [15], topic detection and tracking [16], microblog summarization [4], and so forth. Topic modeling has been described as “a recurring pattern of co-occurring words”. A topic modeling tool looks through a corpus for these clusters of words and groups them together by a process of similarity. Nevertheless, Topic models exploit the co-occurrences of words between documents to find relations between words. Given that documents d_1 and d_2 share the common words $\{w1, w2, w3, w4\}$, we can infer that this densely connected set of words forms a topic and has semantically similar meanings. However, in the case of tweets due to the smaller number of words, there is less likelihood for words to co-occur with one another across different tweets. The words which could be inferred as belonging to the same topics have a weaker co-occurrence relationship with other words. To overcome this drawback, the authors in [4] have assumed that the tweets written around the same time are similar in content and so they can “share” words from other tweets to compensate for their short length. While, in our case we do not make this assumption and each tweet is treated separately considering its time occurrence if more tweets share the meaning that is extracted disambiguating the words with Wikipedia.

Furthermore, LDA (*Latent Dirichlet Allocation*), that is the technique usually exploited to extract topic model, does not extract relations among the identified topics. FCA (and Fuzzy FCA), instead, arranges the tweets into a hierarchical conceptual structure of topics, where the topic is represented by Formal Concept that is composed of words (the intent) shared among different tweets (the extent of the formal concept, Definition 3).

Considering these aspects and our expertise in the area of Fuzzy FCA the proposed approach is in our opinion the right starting way to address social media research challenges, and specifically microblog summarization, proposing an alternative approach compared to the state of the art.

Experimental Results. The proposed framework has been applied on the same tweet streams used in [4] that are focused on some real-world events, such as: Obamacare, Japan Earthquake, and so on. The results have been evaluated considering the following metrics: *Novelty Measurements*, *Text-based Coverage of Wikipedia*, and *Concept-based Coverage of Wikipedia*. The evaluation has been performed by varying the level of *detail d* in $[0-1]$. For all of the used metrics the system produces good performances. Specifically, the algorithm outperforms the results in [4] in terms of *Novelty Measurement* and *Text-based Coverage of Wikipedia*. Furthermore, evaluating *Concept-based Coverage of Wikipedia* setting level of *detail* with values ~ 0.9 (that is a verbose summary), the algorithm outperforms the results shown in [4] in terms of F-Measure, with optimal Recall and comparable values of Precision.

Outline. The manuscript is organized as follows: Section 2 provides an overview of the literature describing some related works; Section 3 introduces the theoretical background, i.e. Fuzzy Formal Concept Analysis; Section 4 introduces the overall framework detailing each phase in the Sections 5–7; finally, Section 8 shows the obtained results and argues the comparison with other existing approaches.

2. Related works

Nowadays, automatic microblog summarization has caught increasing attention from worldwide researchers.

From the time-dependent document summarization point of view, some existing approaches are aimed to address update summarization task defined in TAC (www.nist.gov/tac). Specifically, they emphasize the novelty of the subsequent summary [17]. The proposed approach focuses more on the temporal development of the story (i.e. topic or event) that is stressed by the multitude of the messages posted through microblogging service, i.e. Twitter.

From the microblog summarization point of view, some pioneering approaches working on Twitter exist that are essentially aimed to describe topic extracting list of relevant words or sentences. Specifically, TweetMotif [18] summarizes what’s happening on Twitter providing a list of relevant terms that should explain Twitter topics. [19,5] extract a succinct summary for each topic using a phrase reinforcement ranking approach. [20] explores tweets and linked web contents to discover relevant information about topics. Moreover, [21] generates summaries especially for sport topics. Furthermore, [22] defines frequency and graph based method to select multiple tweets that conveyed information about a given topic without being redundant. Other approaches are based on integer linear programming [23] or clustering to perform the summarization of Evolving Tweet Streams [24]. Other approaches consist of aggregating tweets about specific topic into a visual summaries. These visualizations must be interpreted by users and do not include sentence-level textual summaries. For instance, Visual Backchannel [25] and TwitInfo [7] allow users to graphically browse a large collection of tweets. Specifically, [25] visualizes conversations in Twitter data using topic streams that is visually represented as stacked graphs and TwitInfo [7] uses a timeline-based display that highlights peaks of high tweet activity.

Considering our proposal we find some similarities in [4] and in [6]. Specifically, [4] describes a framework for summarizing events from tweet stream. The authors define two topic models, Decay Topic Model (DTM) and Gaussian DTM, to extract summaries from microblog, and they finally argue that these models outperforms

[1] Publically available at http://wikipedia-miner.cms.waikato.ac.nz/ Let us note that we have exploited a local installation of the Wikipediaminer installation.