Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Unfolding social content evolution along time and semantics

Carmen De Maio [a], Giuseppe Fenza [b], Vincenzo Loia [b,*], Francesco Orciuoli [b]

[a] Dipartimento di Ingegneria dell'Informazione, Elettrica e Matematica Applicata, University of Salerno, 84084 Fisciano (SA), Italy

[b] Dipartimento di Scienze Aziendali - Management & Innovation Systems, University of Salerno, 84084 Fisciano (SA), Italy

HIGHLIGHTS

- The defined approach faces with a generalized version of TDT problem to track any kind of entity or scope along time and semantic dimensions.
- The proposed framework relies on Temporal Fuzzy Concept Analysis that enable us to analyze at the same stage both time relations among tweets and their semantic meaning of their content.
- Description Logic enable us to decouple the tweet analysis and understanding from the definition of the tracking scope (user's query processing) enabling the system to support different kind of analytics perspectives at run time.

ARTICLE INFO

Article history:
Received 29 March 2016
Received in revised form
11 May 2016
Accepted 27 May 2016
Available online xxxx

Keywords:
Temporal Concept Analysis
Social media
Data mining
Fuzzy Formal Concept Analysis
Description Logic

ABSTRACT

In the context of social media, the unstructured and dynamic nature of exchanged data and the information overload contribute to the growth of the number of research works proposing methods to improve performance of intelligent analytics services considering both time and semantics of the shared content. The presented paper focuses on the definition of a knowledge tracking framework to answer questions, such as "What is the semantic evolution of a topic (or news) along the time?", "How did we arrive to a specific event?", "What is the evolution of the topics of interest of a user?", and so on. Our interest is about the elicitation of temporal patterns revealing the evolution of concepts along the time from a social media data stream; we focus on Twitter. Such patterns can be extracted at different levels of abstraction by considering different-sized time intervals and different scopes driven by the conceptualization of users' queries. To address the proposed aim, we extend Temporal Concept Analysis and we use Description Logic to reason on semantically represented tweet streams. The evaluation activity reveals promising results from both sides quantitative and qualitative.

1. Introduction

Context. The explosion of social media, the unstructured and dynamic nature of the exchanged data and the information overload stress the urgent need to define innovative and intelligent analytics services able to turn data into actionable knowledge, such as topic detection and tracking, microblog search, microblog summarization, and so forth. Nowadays, microblogging streams are useful to detect and track political events [1], media events [2], and other real world events [3]. Indeed, the improvement of social media services, for instance, to reveal important temporal patterns describing users' interests or topic development along the time, requires to take into account the temporal relations among shared content. We advice that time awareness has assumed a crucial role with the microblog to support any kind of information extraction service, such as recommendation provisioning, information retrieval, language modeling, auto-completion during query formulation, and so on. For instance, incorporating the temporal dimension in the concept-based query expansion method has a significant positive impact on searching content in microblogs. On the light of the aforementioned needs, in this paper we define a framework to elicit time-aware knowledge structure describing life cycle of an entity (e.g., user's interest, topic, etc.), namely the *tracking scope*, considering the posts exchanged on Twitter along the timeline.

Problem. Formally, this work tries to face with the following problem. Given a timestamped tweet stream, a *tracking scope* e (e.g. a user, group of them, topic, and so on) and a time requirements in terms of *level of granularity* $\Delta t \in [0, 1]$, the problem is to carry out the main temporal paths of concepts

* Corresponding author.
E-mail addresses: cdemaio@unisa.it (C. De Maio), gfenza@unisa.it (G. Fenza), loia@unisa.it (V. Loia), forciuoli@unisa.it (F. Orciuoli).

http://dx.doi.org/10.1016/j.future.2016.05.039
0167-739X/

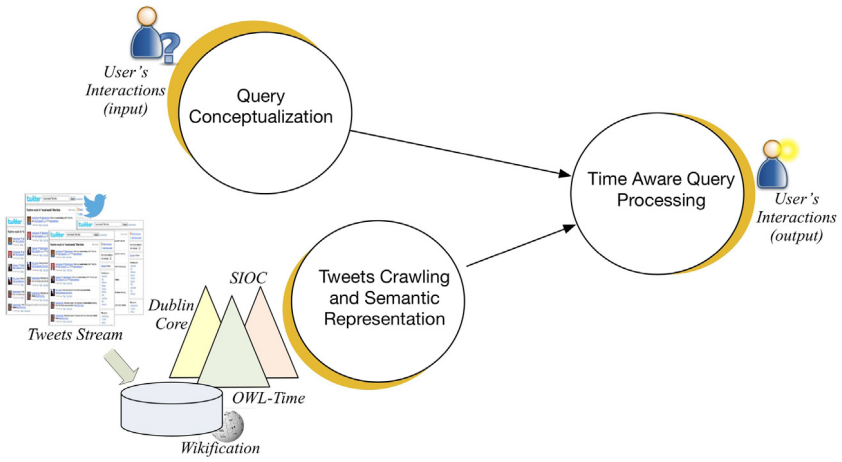

Fig. 1. Overall approach.

describing the semantic evolution of the scope along the time corresponding to the tweets related to the selected *tracking scope e*.

Proposed solution. The proposed solution uses the extension of the theory of Temporal Concept Analysis (TCA) [4], that is Temporal Fuzzy Concept Analysis (TFCA) that carries out temporal paths of Fuzzy Concepts extracted via Fuzzy Formal Concept Analysis (FFCA). Indeed, TFCA allows us to derive a time-aware and hierarchical knowledge structure to filter in/out temporal relations among groups of tweets to summarize the evolution of the scope of interest with different levels of granularity. The framework relies on semantic enriched tweet stream by exploiting third-party semantic API and well-known semantic web ontologies, e.g., SIOC (Semantically-Interlinked Online Communities), SKOS (Simple Knowledge Organization Systems). Fig. 1 gives an overview of the proposed approach. It consists of two preliminary phases in which the tweet stream is semantically represented (*Tweets Crawling and Semantic Representation*) and the user's request is caught and processed in order to build a processable query (*Query Conceptualization*). These phases are prerequisites for the query processing accomplished by the third phase (*Time-Aware Query Processing*) that provides results to the user.

Motivations. The proposed work extends the goals of the topic detection and tracking (TDT) [5]. Analogously to [6] in which authors generalize the definition of the topic to track, we generalize the goal of TDT to track several kinds of entities (or scope) along the timeline. We are interested in a broader definition of the tracking scope that might refer to each item whose life cycle is described by the set of timestamped subset of tweets in the available tweet stream, for instance the tweets created by an author (or group of them), ones that share the same hashtags or classified in the same topic, and so forth. This allows us to provide general knowledge tracking services to answer questions, such as "What is the semantic evolution of a topic (or news)?", "How did we arrive to a specific event?", "What is the evolution of the topic interest of a user along timeline?", and "How can I expand user query considering time-aware correlation among concepts?". Despite to the limit of 140-character messages there is a huge amount of information enclosed in a single tweet, in fact, each shared entity (hashtags, mentions, urls, etc.) may be a source of information that is interesting to track in order to understand influence, behavior, and so on [7,8].

Contributions. The main contributions are summarized as follows:

- this paper faces with a generalized version of TDT problem to track and unfold semantics of shared entities along the time;
- the methodology relies on Temporal Fuzzy Concept Analysis that enables us to analyze at the same stage both time relations among tweets and semantic meaning of their content;
- the architecture decouples the analysis of the tweets from the definition of the tracking scope (user's query processing) enabling the system to report different kinds of analytics at query time.

Experimental results. The proposed framework has been evaluated considering both qualitative and quantitative measures that reveal promising results. On one hand, qualitative results are obtained performing a survey with users. On the other hand, quantitative results are obtained evaluating an *ad hoc* defined measure of Tracking Scope Coverage performed on the resulting temporal paths. The workflow has been executed on the queries that are essentially focused on the following tracking scopes: popular person Vincenzo De Luca, during the Italian regional election of May 2015; the hashtag *#expo2015*, during the Italian Expo of last year; and old news of Japan Earthquake 2011. The results have been evaluated by varying the *time granule* size (i.e., level of granularity Δt).

Outlines. The manuscript is organized as follows: Section 2 describes the state of the art pointing out related works corresponding to different perspectives; Section 3 describes a theoretical background of the overall workflow; Section 4 describes the semantic modeling in the triplestore of the tweets; Section 5 details the query processing workflow; then, Section 6 provides an overview of the framework emphasizing the main phases of the overall workflow. The experimental results are given in Section 7. Finally, conclusion and future directions are argued.

2. Related works

In the literature, several research works focus on the study of the behavioral evolution of a topic on social networks from theories in sociology and epidemiology. The structural properties in the spread of information in networks, including the blogspace [9], have attracted much attention [10,11]. Recently, with the explosion of microblog, the investigation about topic development over the social media stream is becoming crucial to address analytics need, but also to improve search and other information services; Twitter has caught increasing attention from worldwide researchers. In [12], the authors did a study on hashtags in Twitter and found that there is a significant variation in the ways that widely used hashtags on different topics spread. In [13], the authors investigated the correlation between information propagation and user interests in Twitter. In [14], the authors postulate that the influence structure varies across topics. In [6], the topic tracking problem is conceived as a language modeling problem where topic models are trained using hashtags in the tweet stream.

Please cite this article in press as: C. De Maio, et al., Unfolding social content evolution along time and semantics, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.05.039