An Approach to Event-based Dynamic User Profiling in Social Media

Gabriella Pasi[0000−0002−6080−8170], Marco Viviani[0000−0002−2274−9050], and Matteo Zardoni

Università degli Studi di Milano-Bicocca,
Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)
Edificio U14 – Viale Sarca, 336 – 20126 Milano, Italy
{pasi,marco.viviani}@disco.unimib.it, zardo1992@gmail.com,
WWW home page: http://www.ir.disco.unimib.it/

Abstract. Supporting the user in finding relevant information is one of the major tasks in Information Retrieval. The user-centered approach to IR is based on a user profile that represents the user's interests and preferences, to obtain relevant information beyond the formulation of a query. Nowadays, the content diffused by people in the Social Web represents a source of primary information to build a user profile. For this reason, the aim of this paper is to propose an event-based approach to user profiling in social media, where an event is identified as a named-entity. Furthermore, by the proposed approach, it is possible to consider the user's interests evolution over time.

Keywords: Information Retrieval, User Profiling, Information Evolution, Social Media, User-Generated Content.

1 Introduction

The main objective of Information Retrieval is to support users in the process of finding information relevant to their needs and interests, which are usually formulated by means of a query. Originally, Information Retrieval Systems (IRS)s relied on the so called *system-centered* approach, where the IRS produces the same results to the same query, independently from the user context (the "one size fits all" paradigm). In addition to this, by employing this approach, for a particular user it could not be easy to formulate a need or a complex request using only few words. For these reasons, the *user-centered* approach to Information Retrieval is by now the most frequently applied, since it is aimed at identifying the information relevant to the user beyond the classic formulation of a query, by considering the user's *context* [15]. To this purpose, a *user model* (a.k.a., *user profile*), which is a formal representation of the user's interests and preferences, is employed. According to [12], the phases that have to be considered in the user profile definition are: (*i*) the *acquisition* of the information characterizing the

IIR 2018, May 28-30, 2018, Rome, Italy.

user's context, (*ii*) the *formal representation* of the user profile by means of a formal language, and (*iii*) its *updating*, finalized at learning the changes of the user's preferences in time.

This paper presents a preliminary approach for the definition of dynamic user profiles that exploit the content generated and diffused by users through social media, the so called *User-Generated Content* (UGC). In fact, UGC can be considered as a sort of repository from which the user's interests and preferences can be acquired, and which also allows to capture their dynamic nature [6]. Specifically, concerning the above mentioned acquisition phase, the user's preferences are extracted from social media posts collected from Twitter and Facebook, which can contain both text and other multimedia data. In particular, an approach to *event-based user profiling* is proposed in this paper, to both formally represent the user's preferences and for monitoring their evolution over time. More specifically, an *event* that represents a user's interest in a given period of time is identified by means of a *named-entity*, which is extracted from the user's social media posts by means of a *Named Entity Recognition* (NER) tool. Each named-entity is associated with a *topical representation* that is extracted from the social media posts' content related to the entity in a given time-window. Both the events in the user profile and their topical representations are then *updated* over time by considering new social media posts generated by the user.

2 Background

In the literature, several models to represent user profiles have been proposed, such as *bag-of-words* [7, 13], *vectors* [1], *graphs* [3], *personal ontologies* [5]. The bag-of-words model is one of the most common approaches to user profiling, but in the case of short texts – such as social media posts – it presents some limitations. First, terms usually do not occur more than once within a single post, making it difficult to understand the user's interests based on term frequency [14]; second, the bag-of-words model is not able to capture the semantics of a particular post [1]. To overcome these limitations, the definition of a user profile based on the extraction and recognition of *entities* instead of keywords has been suggested [1], and it constitutes the strategy employed by the approach proposed in the next section.

Another important issue that concerns the definition of a user profile is its updating [16]. Interests related to a particular user change over time, and, consequently, it is necessary to update the user profile accordingly. In the literature, several works have addressed this issue. In [4], a long-term user model, which takes into account the long-term interests of the user, and a short-term user model, which takes into account the interests recently manifested by the user, have been defined. It is also possible to define several user profiles through a vector-based representation, based on what the user has published in consequent periods of time, and to measure the similarity between the profiles [1]. Also the use of 'forgetting' functions has been proposed, in order to take into account the interest drift over time [9]. In the proposed approach, as it will be illustrated

below, the concepts of *Temporal User Profile* and *General User Profile* will be described to reflect the above situations.

3 Event-based Dynamic User Profiling

The proposed approach focuses on the idea that a topical user profile can be centered on the concept of *event*. An event has been defined in the literature as "a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences" [2]. In the proposed approach, the events are automatically extracted from the content generated by the user in a given period of time. An event is intended in this context as a potential interest, and it is represented by means of a *named-entity* automatically extracted from the user's *feed* in a *time-window*, where the feed is intended as the collection of social media posts of a particular user, and the time-window is the period of time on which the feed spans.

The term named-entity was coined by Grishman and Sundheim in [8], in the context of Information Extraction (IE); it corresponds to an information unit that represents a real-world entity, such as a place, a person, a product, an organization, and so on, defined by an appropriate name [11]. In the approach described in this paper, once the named-entities have been identified, a *topical representation* is associated with them, based on the content published by the user. To this purpose, the bag-of-words model is applied: the words are extracted directly from the user's feed (i.e., from Facebook and Twitter posts) associated with each event. Finally, both entities in the user profile and their topical representations have to be updated over time based on the dynamicity of the UGC.

Therefore, the proposed approach can be described according to the following three phases: (*i*) the *UGC acquisition* phase, (*ii*) the *named-entity extraction and user profile formal representation* phase, and (*iii*) the *user profile updating* phase.

3.1 UGC Acquisition

In this phase, the user's social media posts gathered from both Facebook (when available) and Twitter are collected and grouped according to a time-window, the granularity of which can be variable, i.e., one week, 15 days, etc.

In general, each post can contain both text and other multimedia data, but in this paper only textual information is considered. In Facebook, a post can represent each user's status update without particular limits, while in Twitter a post is known as *tweet*, and there are limits in the tweet length; originally, the maximum length was fixed to 140 characters, although this has been made more flexible over time. Currently, the length of a tweet cannot exceed 280 characters, and each tweet can be equipped with a picture and a location.

3.2 Named-entity Extraction and User Profile Formal Representation

As previously illustrated, in this work a user profile is composed of *named-entities* extracted from the user's posts in a specific *time-window*. Formally, let us denote by D_{ui} the feed generated by a user u in a time-window t_i. From D_{ui}, a collection E_i of named-entities is extracted. The extraction can be performed by exploiting a *Named-Entity Recognition* (NER) tool; in the proposed approach, DBpedia Spotlight [10] has been employed. Based on the extracted entities, the proposed approach defines a *Temporal User Profile* $PT_i(u)$ for the time-window t_i. The Temporal User Profile can be seen as a set of pairs $\langle e, \omega_{ei} \rangle$, in which e is an *entity*, and ω_{ei} is the *weight* expressing the *importance* of that entity. Formally:

$$PT_i(u) = \{\langle e, \omega_{ei} \rangle | e \in E_i\}.$$

Given a feed D_{ui}, the weight ω_{ei} is computed as follows:

$$\omega_{ei} = \frac{N_{ei}}{|D_{ui}|},$$

where N_{ei} is the number of posts in D_{ui} in which the entity e is cited, and $|D_{ui}|$ represents the total number of posts in D_{ui}.

To provide a *topical representation* of an entity e, a bag-of-words BoW_{ei} is associated with e, which contains the words extracted from the posts of the feed D_{ui} in which e is cited. Formally:

$$e \rightarrow BoW_{ei} = \{\langle w_1, nocc_{w_1} \rangle, \langle w_2, nocc_{w_2} \rangle, \ldots, \langle w_j, nocc_{w_j} \rangle, \ldots\},$$

where the symbol $\rightarrow$ expresses the association between e and the bag-of-words, and $nocc_{w_j}$ represents the number of occurrences of the word w_j in the posts referring to e.

3.3 User Profile Updating

The Temporal User Profile defined in the previous phase allows to capture the user's interests based on the content s/he published in a given time period. In order to be able to capture the dynamicity of the user's interests over time, a *General User Profile* $P(u)$ is defined in this phase. Specifically, at each time window t_i, an *updated instance* $P_i(u)$ of the General User Profile is created, which takes into account both the current interests of the user and the previous ones. The first instance, i.e., $P_1(u)$, coincides with the Temporal User Profile $PT_1(u)$ associated with the first time-window t_1, the second instance, i.e., $P_2(u)$, considers both the Temporal User Profile $PT_2(u)$ associated with time-window t_2 and $P_1(u)$ related to t_1, and so on, as illustrated in Figure 1. From a formal point of view:

$$P_i(u) = \{\langle e, \nu_{ei} \rangle | e \in \mathbf{E}_i\},$$

where ν_{ei} is the *interest value* of e at the time-window t_i, and $\mathbf{E}_i$ is the set of *all* entities extracted from the user's feed until time-window t_i (included). The value of ν_{ei} in $P_i(u)$ and the bag-of-words associated with e are obtained as follows.

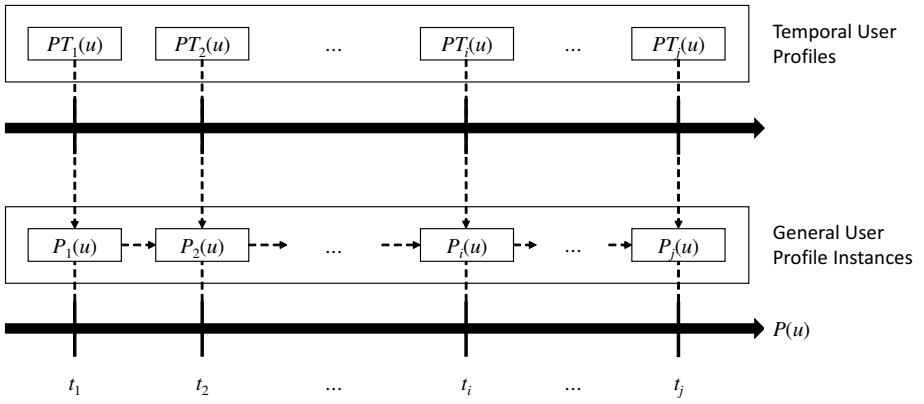

Fig. 1. Representation of the General User Profile construction.

Computing the interest weights At each time-window t_i, the value of ν_{ei} can be computed according to the following simple formula, which have been defined to preliminarily evaluate the approach:

$$\nu_{ei} = \begin{cases} \omega_{ei} & i = 1 \text{ or } \nu_{e(i-1)} = 0 \\ \frac{\nu_{e(i-1)} + \omega_{ei}}{2} & \text{otherwise} \end{cases} \quad (1)$$

where $\nu_{e(i-1)}$ is the interest value of the entity e at time $t_{(i-1)}$.

This means that, if the user had no previous interest in the entity e, the interest value ν_{ei} in $P_i(u)$ assumes the value of ω_{ei}. Otherwise, the value of ν_{ei} is the arithmetic mean between the values $\nu_{e(i-1)}$ and ω_{ei}.

Updating the bag-of-words The bag-of-words related to a given entity has to be updated at each time-window, in order to have, within each instance of the General User Profile, an up-to-date topical representation for the entity. Specifically, for time-window t_i, the bag-of-words associated with an entity e in $P_i(u)$ is constituted by the *union* of the terms related to e from the bag-of-words in the Temporal User Profile $PT_i(u)$ and those from the bag-of-words in $P_{(i-1)}(u)$. If a term appears in the bags-of-words of both $PT_i(u)$ and $P_{(i-1)}(u)$, the term frequency associated with the term in $P_i(u)$ can be computed as the arithmetic mean of the term frequency in $PT_i(u)$ and the term frequency in $P_{(i-1)}(u)$; otherwise, if a term does not appear in $PT_i(u)$, the frequency of the term in $P_i(u)$ can be obtained by applying a decay function to the frequency of the term in $P_{(i-1)}(u)$; a simple solution could be represented by the function $f(tf) = \frac{1}{\lambda} tf$, where λ is a decay parameter, and tf is the term frequency in $P_{(i-1)}(u)$. In this way, within the updated bag-of-words representing e, the terms that have the higher frequency are the ones that better describe the entity, and the terms that are no more employed in recent posts will assume less and less importance over time.

4 Implementation and Preliminary Results

The prototype implementing the approach described in the previous section has been developed in Java 8. In order to crawl both Facebook and Twitter posts, the Twitter4J,[1] version 4.04, and Facebook4J,[2] version 2.4.8, libraries have been employed. Each feed has been indexed by exploiting the Apache Lucene library,[3] version 6.4.0. DBpedia Spotlight[4] has been used for the automatic extraction of named-entities from unstructured text.

In order to preliminarily evaluate if the proposed approach is able to track the evolution of the user's interests over time, a simple analysis on two users's interests has been conducted. Table 1 illustrates the real users that have been considered, i.e., Barack Obama and Coldplay; their feeds are publicly accessible and their interests easily analyzable. Data refer to the period ranging from March 2013 to January 2017. For the two users, a couple of entities (interests) that is worth mentioning are commented in the following.

Table 1. The considered users.

Category	User	Facebook	Twitter	Language	$\|D_u\|$
Politician	Barack Obama	x	x	English	3188
Musician	Coldplay		x	English	3168

In Figures 2 and 3, the bar chart illustrates the value of the weight ω_{ei} associated with the entity in each Temporal User Profile at each time-window t_i, while the line chart illustrates the evolution of the interest value ν_{ei} of the entity in the instances of the General User Profile.

In relation to the user Barack Obama, two entities are analyzed in Figure 2; it emerges that the entity *Patient_Protection_and_Affordable_Care_Act* represents a long-term interest, since its evolution graph in the left line chart shows a growing curve, while the entity *United_States_Senate* only appears during the last part of the tracking period, as illustrated in the right line chart. Since it never occurred before, it is plausible to state that it represents a short-term interest, related to a particular phase of his political career.

Figure 3 illustrates two entities that might represent two short-term interests referred to Coldplay. The interesting observations that emerge from this figure are that the entity *iTunes* occurs especially during the first part of the overall period of social activity analyzed; instead, the evolution trend of the entity *Spotify* is almost the opposite of the *iTunes* one. Considering that Coldplay

[1] http://twitter4j.org/
[2] http://facebook4j.org/
[3] http://lucene.apache.org/core/
[4] http://www.dbpedia-spotlight.org/

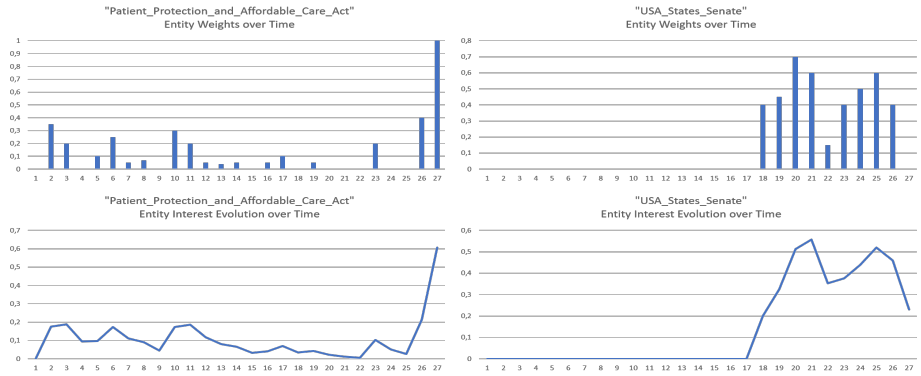

Fig. 2. *Patient_Protection_and_Affordable_Care_Act* and *United_States_Senate* entities.

use their Twitter account mainly for advertisement purposes, it is possible to suppose that for *iTunes* an interest drift occurred because Coldplay decided, at a certain point, to focus more on *Spotify* instead of *iTunes* as a music provider.

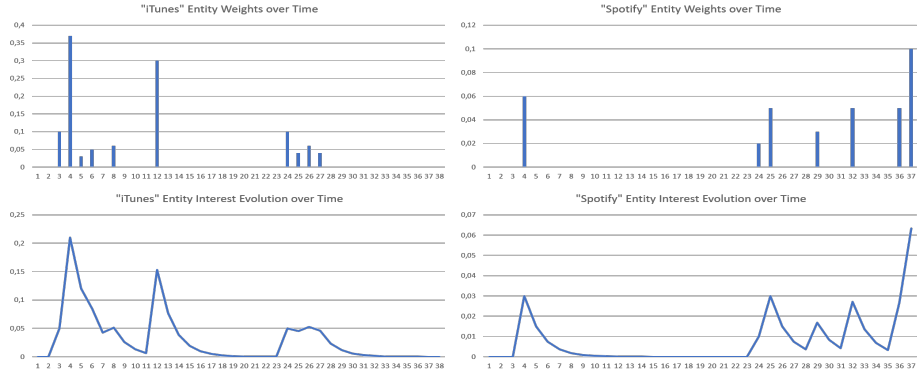

Fig. 3. *iTunes* and *Spotify* entities.

5 Conclusions

In this paper, a preliminary approach for modeling a dynamic user profile in the Social Web by exploiting the User-Generated Content (UGC) spreading through social media has been proposed. The approach is based on an event-based strategy, where the user's interests are represented as named-entities extracted from Facebook and Twitter feeds related to a given period of time. The proposed approach allows to consider the dynamic nature of the user's interests, and to update the user profile accordingly; it could be useful for example to search or

recommend social media contents, in a scenario where opinions and interests change very quickly. In this paper, only preliminary results connected to the evolution of the user's interests over time have been presented. In the future, the model will be refined by considering for example the normalization of interest weights across entities, and extended evaluations will be performed.

References

1. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on Twitter for personalized news recommendations. In *Int. Conf. on User Modeling, Adaptation, and Personalization*, pages 1–12. Springer, 2011.
2. J. Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2012.
3. R. Arezki, P. Poncelet, G. Dray, and D. W. Pearson. Information retrieval model based on user profile. In *Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications*, pages 490–499. Springer, 2004.
4. D. Billsus and M. J. Pazzani. A hybrid user model for news story classification. In *UM99 User Modeling*, pages 99–108. Springer, 1999.
5. S. Calegari and G. Pasi. Personal ontologies: Generation of user profiles based on the YAGO ontology. *Inf. Proc. & Manag.*, 49(3):640–658, 2013.
6. B. Carminati, E. Ferrari, and M. Viviani. A multi-dimensional and event-based model for trust computation in the social web. In *Int. Conf. on Social Informatics*, pages 323–336. Springer, 2012.
7. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.
8. R. Grishman and B. Sundheim. Design of the MUC-6 evaluation. In *Proc. of the 6th Conf. on Message Understanding*, pages 1–11. Association for Computational Linguistics, 1995.
9. I. Koychev and I. Schwab. Adaptation to drifting users interests. In *Proc. of ECML2000 Workshop: Machine Learning in New Inf. Age*, pages 39–46, 2000.
10. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: shedding light on the web of documents. In *Proc.s of the 7th Int. Conf. on Semantic Systems*, pages 1–8. ACM, 2011.
11. D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
12. G. Pasi. Issues in Personalizing Information Retrieval. *IEEE Intelligent Informatics Bulletin*, pages 3–7, 2010.
13. M. J. Pazzani, J. Muramatsu, D. Billsus, et al. Syskill & webert: Identifying interesting web sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.
14. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in Twitter to improve Information Filtering. In *Proc. of the 33rd Int. ACM SIGIR Conf. on Research and development in information retrieval*, pages 841–842. ACM, 2010.
15. L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Knowledge and Information Systems*, 24(1):1–34, 2010.
16. H. Yin, B. Cui, L. Chen, Z. Hu, and X. Zhou. Dynamic user modeling in social media systems. *ACM Trans. on Inf. Syst. (TOIS)*, 33(3):10, 2015.