Personalized Emerging Topic Detection Based on a Term Aging Model

MARIO CATALDI, École Centrale Paris
LUIGI DI CARO and CLAUDIO SCHIFANELLA, Università di Torino

Twitter is a popular microblogging service that acts as a ground-level information news flashes portal where people with different background, age, and social condition provide information about what is happening in front of their eyes. This characteristic makes Twitter probably the fastest information service in the world. In this article, we recognize this role of Twitter and propose a novel, user-aware topic detection technique that permits to retrieve, in real time, the most emerging topics of discussion expressed by the community within the interests of specific users. First, we analyze the topology of Twitter looking at how the information spreads over the network, taking into account the authority/influence of each active user. Then, we make use of a novel term aging model to compute the burstiness of each term, and provide a graph-based method to retrieve the minimal set of terms that can represent the corresponding topic. Finally, since any user can have topic preferences inferable from the shared content, we leverage such knowledge to highlight the most emerging topics within her foci of interest. As evaluation we then provide several experiments together with a user study proving the validity and reliability of the proposed approach.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-bases services*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Social network analysis, Twitter, topic detection and tracking, trends, aging theory

ACM Reference Format:
Cataldi, M., Di Caro, L., and Schifanella, C. 2013. Personalized emerging topic detection based on a term aging model. ACM Trans. Intell. Syst. Technol. 5, 1, Article 7 (December 2013), 27 pages.
DOI: http://dx.doi.org/10.1145/2542182.2542189

1. INTRODUCTION

Microblogging today has become a very popular communication system among users. In fact, due to the short format of messages and the accessibility of microblogging platforms, users tend to shift from traditional communication tools (such as blogs, traditional Web sites, and mailing lists) to microblogging services. Billions of messages are appearing daily in these services such as Twitter[1], Tumblr[2], Facebook[3], etc. The authors of those messages share contents about their private life, exchange opinions on a variety of topics, and discuss a wide range of information news.

[1] http://www.twitter.com.
[2] http://www.tumblr.com.
[3] http://www.facebook.com.

Authors' addresses: M. Cataldi (corresponding author), MAS Laboratory, Ecole Centrale Paris, France; email: mario.cataldi@ecp.fr; L. Di Caro and C. Schifanella, Dipartimento di Informatica, Universita di Torino, Italy.

DOI: http://dx.doi.org/10.1145/2542182.2542189

Among all the existing platforms, after its launch on July 2006, Twitter became the most popular *microblogging system*; on February 2012, its number of users has been estimated to be about 500 million world-wide with around a half million new accounts per day, which makes Twitter one of the fastest growing Web sites in the world. Moreover, in contrast with other popular social networks, such as Facebook or Google Plus, most of its users are adults; in fact, according to a demographic report[4], 88% of the users are older than 18, which makes the service likely oriented to information rather than other social aspects.[5]

In fact, as information producers, people post tweets (text messages up to 140 characters) for a variety of purposes, including daily chatter, conversations, share of information/URLs, news reports, etc., defining a continuous real-time information stream about every argument.

For this reason, even if Twitter cannot represent an alternative to the authoritative information media, considering the number of its users and the impressive response time of their contributions, it represents a sort of real-time news sensor that can also predate the best newspapers in informing the Web community about emerging topics and trends. In fact, the most important information media always need a certain amount of time to react to a news event; that is, professional journalists require time, collaborators, and/or technology support to provide a professional report. However, within Twitter, a user can easily report, in 140 characters, what is happening in front of her eyes, without any concern about the readers or the writing style. This characteristic makes Twitter probably one of the fastest, low-level information services in the world.

In this article we recognize this informative role of Twitter and we present an extension of the approach proposed by Cataldi et al. [2010] to extract, in real time, the most emerging topics expressed by the community along the interests of a specific user.

More in detail, considering an active user, we initially analyze her interests by extracting and formalizing the content of her generated tweets. We then model the social community as a directed graph of the active authors based on their social relationships, and calculate their authority by relying on the well-known PageRank algorithm [Page et al. 1998]. Therefore, we monitor the stream of information expressed by the entire network by studying the lifecycle of each term according to a novel aging model that also leverages the reputation of each author. We therefore select the set of most emerging keywords by dynamically ranking the terms depending on their life status (defined through a burstiness value). Finally, we represent each topic (expressed as minimal set of terms, as in Allan [2002] and Makkonen et al. [2004]) related to each emergent term by constructing and analyzing a keyword graph which links the extracted emerging terms with all their co-occurrence keywords.

At this point, in order to personalize the list of retrieved emerging topics, we analyze the temporal time frames in which the user has been active and analyze her generated content to estimate the user's interests according to this temporal information. This time-aware information is finally used to highlight the topics that best match the interests of the user.

The article is organized as follows: in Section 2 we present the current state-of-the-art on content aggregation, recommendation, trend analysis, social monitoring, and content personalization by reporting a short summary of the existing approaches. We then analyze step by step the proposed personalized topic detection method by formalizing our assumptions and providing real case scenarios (Section 3).

[4] http://palatnikfactor.com/2010/01/29/twitter-demographic-report-who-is-really-on-twitter/.

[5] A deeper analysis of the Twitter network is also provided in Poblete et al. [2011].

Finally, we evaluate the presented technique by reporting a various range of real case studies and analyze its validity (from a user point of view) with a user study (Section 4). Conclusions end the article.

2. RELATED WORK

In the last decade the enormous amount of contents generated by Web users created new challenges and new research questions within the data mining community. In this section we present an overview of those works which share part of our techniques, ideas, and motivations. In particular, we initially report related work about aggregation, recommendation, and propagation of information from large-scale social networks. We then survey the related work on automatic detection of events within user-generated environments and, finally, we analyze the current state-of-the-art about personalization and user context analysis.

2.1. Aggregation, Propagation and Recommendation through Social Networks

A first issue when dealing with large and heterogeneous data sources is the aggregation of the content through filtering and merging techniques. Considering Twitter as a source of text data, many Web services like *TweetTabs*[6] and *Where What When*[7] aggregate messages and links through user-friendly interfaces. In general, clustering techniques help in finding groups of similar contents that can be further filtered using labeling techniques [Treeratpituk and Callan 2006].

While these services simply aggregate messages and/or links, one of the most explored tasks in mining of text entries streams from social media is the recommendation of topics, URLs, friends, and so forth. So far, two main high-level approaches have been studied: collaborative filtering and content-based techniques. While the first aims at selecting and proposing content by looking at what similar users have already selected (as in Goldberg et al. [1992]), the second analyzes the semantics of the content without considering its origin (as in Hassan et al. [2009]). More recently, hybrid approaches have been also proposed by Melville et al. [2001] and Balabanovic and Shoham [1997].

Recommendation systems can differ on what they recommend: URLs (as in Chen et al. [2010]), users that share same interests (as in Chen et al. [2009]) and tags in folksonomy-based systems (like the approach proposed in Jäschke et al. [2007]).

Another issue when dealing with huge, time-sensitive, and user-generated text content is the analysis of how such information spreads through the blogosphere or a social network. Generally speaking, research on flows of information in networks initially started from the analogy with the spread of a disease in a social environment. This model is based on the following disease lifecycle: someone is first susceptible to the disease and then, if exposed to the disease by an infectious contact, he becomes infected (and infectious) with some probability. In general, there exist two main approaches to model propagation, namely threshold models [Granovetter 1978] and cascade models [Goldenberg et al. 2001]. While the first one treats the problem as a chain reaction of influence where each node in the network obeys a monotone activation function, in the second one nodes close to each other have some chance to influence each other as well.

In Gruhl et al. [2004], the authors studied the dynamics of Web environments at topic and user level, inducing propagation networks from a sequence of posts in the blogosphere. Goyal et al. [2010] studied how to learn influence probabilities from a log of past propagations in the *Flickr*[8] community. Yang and Leskovec [2010] developed the linear influence model, where the influence functions of individual nodes govern the

[6] http://tweettabs.com.

[7] http://where-what-when.husk.org.

[8] http://www.flickr.com.

overall rate of diffusion through the network. Cha et al. [2010] present a comparison between three measures of influence: indegree, retweets, and mentions in Twitter, investigating the dynamics of users' influence across topics and time. Still, Yang and Counts [2010] proposed a model to capture properties of information diffusion like speed, scale, and range.

Focusing on Twitter-based approaches, *Trendistic*[9] and *Twopular*[10] represent two examples from which it is possible to analyze the trends of some keywords along a timeline specified by the user. In general, several studies examined topics and their changes across time in dynamic text corpora. The general approach orders and clusters the documents according to the timestamps, analyzing the relative distributions (see also Griffiths and Steyvers [2004]). Zhao et al. [2007] represent social text streams as multigraphs, where each node represents a social actor and each edge represents the information flow between two actors. In Qi and Candan [2006] and Favenza et al. [2008], the authors present a system which uses curve analysis of frequencies for automatic segmentation of topics.

Wu et al. [2010] use the tolerance rough set model to enrich the set of feature words into an approximated latent semantic space, from which they extract hot topics by a complete-link clustering. Abrol and Khan [2010] analyze tweets in order to predict whether the user is looking for news or not, and determine keywords that can be added to her Web search query. Asur et al. [2011] explore the longevity of trending topics on Twitter, and analyze the role of users in the emergence of trends. The role of the users, their collaboration, and the influence among them has been also extensively studied in social networks [Katz et al. 1997], from qualitative studies on cooperation behaviors [Crane 1969; Chubin 1976; Shapin 1981] to more quantitative approaches [de Beaver and Rosen 1979; Melin and Persson 1996]. The latter includes collaboration network-based studies (because a social network can be easily seen as a social network where people form teams to produce some results), which are generally aiming at understanding the structural determinants and patterns of collaboration [Newman 2001; Barabasi et al. 2002; Schifanella et al. 2012; Di Caro et al. 2012]. Such networks have been deeply analyzed by looking at properties like network topology, size, and evolution [Moon et al. 2010; Hou et al. 2008].

Chen et al. [2003] were the first to present an aging theory based on a biological metaphor. Using this approach, the work presented by Wang et al. [2008] is able to rank topics from online news streams through the concept of *burstiness*. The burstiness of a term, already introduced by He et al. [2007], is computed with a χ-statistic on its temporal contingency table. Although this work shares our goal, it is based on the concepts of user attention and the media focus, whereas our proposed approach is independent from them and it is assumed to be less complex and more general.

2.2. Event Identification in Social Networks

Identifying events in real time on Twitter is a challenging problem, due to the heterogeneity and immense scale of the data. In fact, as already reported in the Introduction, users post messages with a variety of purposes. For this, while many contents are not specifically related to any particular real-world event, informative event messages nevertheless abound.

Regarding the classification of single tweets, Sriram et al. [2010] define a typology of five generic classes of tweets (news, events, opinions, deals, and private messages) in order to improve information filtering and recognize events. The research works proposed in Backer et al. [2011, 2010] and Sankaranarayanan et al. [2009]

[9] http://trendistic.com.
[10] http://twopular.com.

focused on identifying events in social media in general, and on Twitter. Recent works on this social network started to process data as a stream with the goal of identifying events of a particular type (e.g., news events, earthquakes, etc.). Petrović et al. [2010] identify the first Twitter message associated with an event in order to analyze the starting point of the related information flow.

The real-time social content can also be seen as a sensor that captures what is happening in the world: similarly to the recommendation task, this can be exploited for a zero-delay information broadcasting system that detects emerging concepts. Generally, all the techniques rely on some measure of importance of the keywords. Bun et al. [2002] present the $TF * PDF$ algorithm which extends the well-known $TF - IDF$ to avoid the collapse of important terms when they appear in many text documents. Indeed, the *IDF* component decreases the frequency value for a keyword when it is frequently used. Considering different newswire sources or channels, the weight of a term from a single channel is linearly proportional to the term's frequency within it, while it is exponentially proportional to the ratio of documents that contain the term in the channel itself. Lampos and Cristianini [2012] presented a supervised learning system for the extraction of events from unstructured textual information that relies on geo-tagged Twitter posts. Takeshi Sakaki and Matsuo [2010] consider Twitter as a social sensor for detecting large-scale events like earthquakes, typhoons, and traffic jams. The authors analyze the context of such keywords in order to discriminate them as positive or negative (the sentence "Someone is shaking hands with my boss" should be captured as negative even though it contains the term "shake").

In AlSumait et al. [2008] the authors present a system called OLDA (online topic model) which permits to automatically capture the thematic patterns and identifies emerging topics of text streams and their changes over time. However, in contrast with our approach, the system detects topics of discussion without any knowledge about user properties and/or preferences.

2.3. Content Personalization

Since our system includes a module for the personalization of the emerging topics to be retrieved from Twitter, we also cover here the relevant works that have been done in this field. Most of the literature refers to this task as "personalization of search results", or "user-driven personalization" (as in the works proposed in Sugiyama et al. [2004], Teevan et al. [2005b], Ziegler et al. [2005], and Noll and Meinel [2007]). As stated by Teevan et al. [2005a], the motivation behind the interest around this area of research is based on the reasonable assumption that different users generally expect different information even with the same query. There obviously exist several approaches of facing such a task. For instance, depending on the domain, one may be interested in *reranking* the results based on their relevance [Noll and Meinel 2007; Teevan et al. 2005b], rather then to *diversify* them (as in Lin et al. [2010], Agrawal et al. [2009], Radlinski and Dumais [2006], and Wedig and Madani [2006]). Still, the actual personalization of contents (whether they are search results or emerging topics as in our case) could be made by leveraging different kinds of data: users' contents [Cantador et al. 2010; Xu et al. 2008], ontologies [Han et al. 2010; Sieg et al. 2007; Gauch et al. 2003], and users, social network and activity [Wang and Jin 2010; Carmel et al. 2009]. Given this brief overview, our system can be classified as a reranking approach that makes use of the users' contents, that is, their posted tweets. In addition to this simple scheme, we also wanted to take into account the temporal aspect associated to the tweets in order to weight more what has been recently posted, and the other way around.

3. SEARCHING EMERGING TOPICS ALONG THE USERS' INTERESTS

In this section we report, in detail, our method for analyzing, in real time, the dynamic stream of information expressed by the Twitter community and retrieving the most emerging topics within the user's interests. For this, we first formalize the set of tweets, generated within a specific time interval, as a set of keyword vectors, and we then make use of a term aging model to monitor the usage of each keyword over time. Moreover, we leverage the social reputation of the Twitter users to balance the importance of the information expressed by the community. Finally, we take into account the user context, provided by the generated set of tweets, to highlight the most emerging topics within her interests. In the next sections we will report these steps in detail.

3.1. Real-Time Vectorization of Tweets

As in most Information Retrieval (IR) systems, the process starts with the real-time extraction of the relevant keywords (also called terms in the article) from the stream of tweets. Thus, given a time range r set by the system (depending on the preferred topic detection frequency; see also Section 3.2.2), we define the t-th considered interval I^t as

$$I^t = < i_t, i_t + r >, \tag{1}$$

where i_t is the starting instant of the t-th considered time interval (and $i_0 = 0$ represents the first considered instant). Thus, we extract the corpus TW^t, with $n = |TW^t|$ text tweets extracted during the time interval I^t, and we associate to each tweet tw_j a representative *tweet vector*, $\vec{tw_j}$, that formalizes the information extracted from it.

Each component of the vector $\vec{tw_j}$ represents a weighted term extracted from the related tweet tw_j. As opposed to common systems, we do not perform any preliminary phase of stop-word elimination; in fact, our system considers all the languages in which Twitter's users update their status. The idea is to leverage the Twitter users' network, using their world-wide extension, in order to be able to retrieve in real time relevant news. In fact, we believe that the stream of information directly rises in the geographical origin of the event and expands its influence proportionally to its global importance; for example, the first news reports about the revolutionary wave of demonstrations and protests occurring in the Arab world in 2011, also known as "Arab Spring"[11], had been initially generated in Tunisia and Egypt and then, due to the global political and social importance of these events, they had been also commented by users of different countries and continents. Thus, in order to be able to quickly catch relevant news from this world-wide information network, we do not have to discriminate the information based on the language or the country in which it has been generated. Obviously, this approach has the significant disadvantage of maintaining all the keywords, including stop words, typos, and irrelevant terms; however, we believe that it is possible to recognize this noise by adapting standard text analysis methods that consider inverse frequency-like techniques. This information refinement step is applied in Section 3.2.

Considering this idea, we preserve not only all the keywords, but we also try to highlight such keywords that appear less frequently but could be highly relevant for one topic. Therefore, we calculate the weight $w_{x,j}$ of the x-th vocabulary term in j-th tweet by using the augmented normalized term frequency [Salton and Buckley 1988]

$$w_{x,j} = 0.5 + 0.5 \cdot \frac{tf_{x,j}}{tf_j^{max}}, \tag{2}$$

[11] These protests have also been known for the massive use of Twitter post because of the protesters' reliance on Twitter and other social networking Internet sites to communicate with each other.

where $tf_{x,j}$ is the term frequency value of the x-th vocabulary term in j-th tweet and tf_j^{max} returns the highest term frequency value of the j-th tweet.

Thus, for each tweet tw_j, a tweet vector

$$t\vec{w}_j = \{w_{1,j}, w_{2,j}, \ldots, w_{v,j}\} \tag{3}$$

is defined, where K^t is the vocabulary (set of keywords) of the corpus in the time interval I^t and $v = |K^t|$ is its size.

At the end of this step, the knowledge expressed by each collected tweet in the considered time interval has been formalized as a weighted tweet vector.

3.2. Analyzing the Stream of Information by Taking into Account Temporal Conditions

In this section we analyze the extracted tweets stream in order to study the semantic relationships that exist among the keywords reported by the community in a given time interval.

Generally speaking, a term can be viewed as a semantic unit which can potentially link to a news event. The goal of capturing such correlation relies on an accurate modeling of both the chronological sequences of tweets and the authors. Following this intuition, we propose a content aging theory to automatically identify coherent discussions through a lifecycle-based content model.

Many conventional clustering and classification strategies cannot be applied to this problem due to the fact that they tend to ignore the temporal relationships among documents (tweets in our case) related to a news event. Relying on this temporal feature, we propose a metaphor where each term is seen as a living organism. The lifecycle of a keyword can be considered as analogous to the one of a living being: with abundant nourishment (i.e., related tweets), its lifecycle is prolonged; however, a keyword or a live form dies when nourishment becomes not sufficient.

Relying on this analogy, we can evaluate the usage of a keyword by its *burstiness*, which indicates the *vitality* status of the keyword and can qualify the keyword's usage. In fact, a high burstiness value implies that the term is becoming important in the considered community, while a low burstiness value implies that it is currently becoming out of favor.

Considering this biological metaphor, the contribution in terms of nutrition of each nourishment changes depending on its chemical composition; for example, each food brings a different calorie contribution depending on its ingredients. Therefore, in our case, we use the concept of authority to define the *quality* of the nutrition that each tweet gives to every contained keyword. This way, different tweets containing the same keyword generate different amount of nutrition depending on the representativeness of the author in the considered community.

Thus, considering a keyword $k \in K^t$ and the set of tweets $TW_k^t \in TW^t$ containing the term k at time interval I^t, we define the amount of nutrition as

$$nutr_k^t = \sum_{tw_j \in TW_k^t} w_{k,j} * rep(user(tw_j)), \tag{4}$$

where $w_{k,j}$ represents the weight of the term k in the tweet vector $t\vec{w}_j$ (thus, $tw_j[k]$), the function $user(tw_j)$ returns the author u of the tweet tw_j, and $rep(u)$ returns the reputation value associated to u by using some reputation evaluation system.

Thus, considering a keyword k used by the community in the time interval I^t, this nutrition formula evaluates the usage of this term by considering its frequency in the tweets that mention it as well as the reputation of each single user that reported k[12].

In the next section we will report the details for the estimation of the reputation of an author.

3.2.1. Reputation of the Users.

While the contents themselves constitute the entire semantics from where we want to extract emerging facts, a fundamental issue in the treatment of such knowledge is the importance of the source. In Twitter, the origin of the messages is a set of users whose huge heterogeneity was already prefaced in the Introduction. Figuring out a level of importance of a specific source (i.e., a Twitter user) represents a key point towards a well-advised filtering and weighting of the contents.

A Twitter user can follow the text stream of other users by making explicit the social relationship of *follower*. On the other hand, a user who is being followed by another user does not necessarily have to reciprocate the relationship by following her back, which makes the graph of the network directed. This social model enables us to define an author-based graph $G(U, F)$ where U is the set of users and F is the set of directed edges (i.e., the follower relation); thus, given two users u_i and u_j, the edge $<u_i, u_j>$ exists only if u_i is a follower of u_j.

Thus, we measure the reputation of each user by analyzing the connectivity in G; in particular, since users tend to follow people that suppose to be interesting (for example because they share the same topics of interest), we can assume that a user with a high number of followers (incoming edges) represents an influential information source into this social community. For example, we believe that most of the people easily agree that Barack Obama (with more than 10 million followers) represents a strongly authoritative Twitter user, simply because each of his words can be instantly read by thousands of other users, and can influence their normal text stream activity. Moreover, the concept of "reputation" can be also extended by taking into account the fact that the importance of a user is also related to the degree of importance of its followers; considering for example Barack Obama again, each of the users followed by him assumes more importance based on the influence of this authoritative relationship. For all these considerations, this scenario can be easily compared to the problem of topological-based computation of Web pages authority in large hypertextual systems. In particular, as in alternative works [Weng et al. 2010; Bakshy et al. 2011; Kwak et al. 2010], we can refer to the well-known PageRank algorithm [Page et al. 1998] for this task. More in detail, PageRank calculates the authority of each page by analyzing the topological graph of the considered Web entities. Following this strategy, the reputation of a user depends on the number and the reputation of its followers. Hence, given a user $u_i \in U$, its *reputation* is computed as

$$rep(u_i) = d \times \sum_{u_j \in follower(u_i)} \frac{rep(u_j)}{|following(u_j)|} + (1 - d), \tag{5}$$

[12] Please notice that the authoritativeness of the users who first talk about an emerging topic could be not necessarily high. However, considering the dynamicity of the Twitter network, the information spreads very quickly and reaches in a limited time window (if the information is significant, in some sense, for the active users) a very large number of authors with different authority values. Thus, the authority of the users is not the only essential parameter. If a news item is massively commented by not only authoritative users, the system will yet be able to detect the emerging keywords and retrieve the related topics. In other words, as we will comment in the experimental section, a keyword is labeled as emergent based on the rapidity of its spread over the network. In fact, as reported in Asur et al. [2011], "factors such as user activity and number of followers do not contribute strongly to trend creation and its propagation." The authority of the users only helps to further highlight this fact.

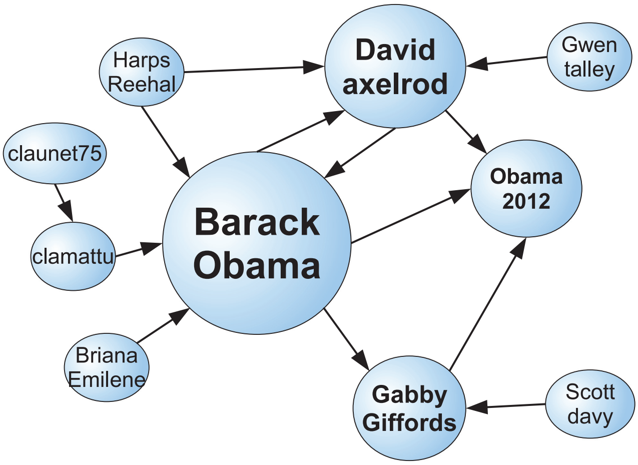

Fig. 1. The reputation value computation: a sample of the “Barack Obama” community. The size of the nodes represents their importance in the considered community.

where $d \in (0, 1)$ is a dumping factor[13], $follower(u_i)$ is a function that returns the set of users following u_i, and $following(u_j)$ returns the set of users that u_j follows.

In Figure 1 an example of user reputation computation on the input graph, obtained by performing a graph sampling process [Leskovec and Faloutsos 2006] in which the “Barack Obama” vertex represents the starting point, is depicted. User reputation values are visually represented by the circle sizes. In this case, “Barack Obama” is the most influential user, since it has the highest number of followers (more than 10 million). Moreover, its reputation is propagated to the “davidaxelrod” user—the Twitter account of David Axelrod, an American political consultant—because of the follower relation by “Barack Obama”; this scenario confers to “davidaxelrod” a high reputation value, even if it has a significantly lower number of followers (∼60k followers).

3.2.2. Computing Term Burstiness Values. Once obtaining the nutrition of a term, we need to map it into a value of *burstiness*. The burstiness value of a term indicates its actual contribution (i.e., how much it is emergent) in the corpus of tweets. Our idea is that the temporal information associated to the tweets can be used as discriminant function in that sense. The term *burstiness* used in this article closely resembles the concept of *burstiness* introduced in many related works, such as Glance et al. [2004], Liang et al. [2010], and Chen et al. [2003]. In detail, we define the burstiness of a term as the ratio of the frequency of the number of occurrences of a keyword on the current time period with respect to its occurrences on previous time slots. Please notice that the term burstiness has been used also with respect to many different factors (occurrences in different documents, sentences, and/or position in a text); in this article we refer to this concept only with respect to temporal conditions.

In our work, having for each keyword k its amount of nutrition $nutr_k^t$ in a time interval I^t, it is possible to rank the hottest terms only considering their related nutrition value. A term can be defined as *hot* if its usage is extensive within the considered time interval.

[13] The dumping factor d, introduced by the authors in [Page et al. 1998], represents the probability that a “random surfer” of the graph G moves from a user to another; it is usually set to 0.85.

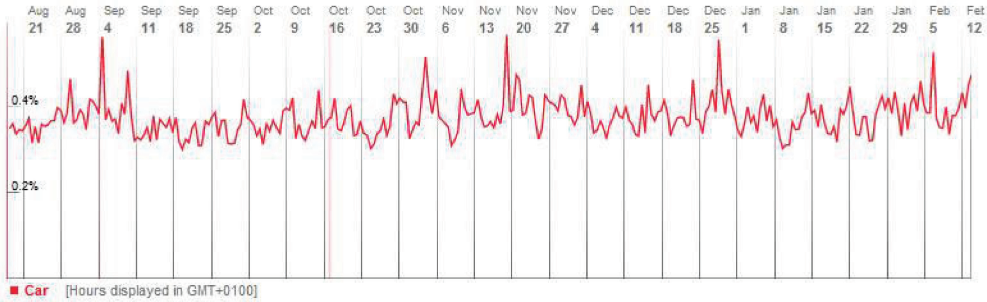

Fig. 2. Statistical usage of the term “car” (provided by Trendistic, https://twitter.com/trendistic) in Twitter from August 2011 to February 2012.

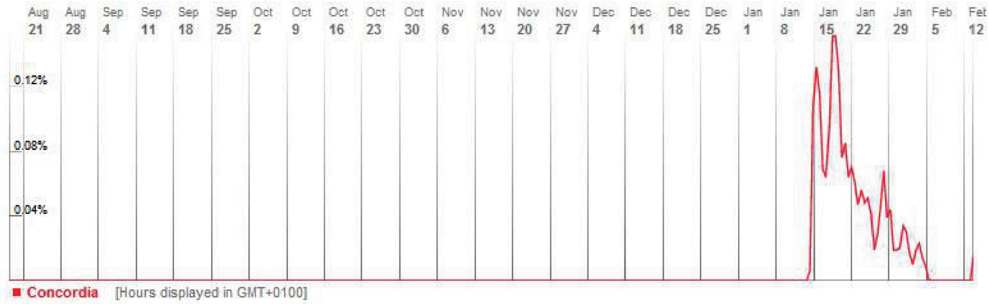

Fig. 3. Statistical usage of the term “Concordia” (provided by Trendistic, https://twitter.com/trendistic) in Twitter from August 2011 to February 2012; the peak represents the catastrophic sinking of the Italian boat “Costa Concordia” occurring in Italy on 13 January 2012.

However, as explained in the Introduction, in this step we are interested in detecting the *emerging* terms during the considered time interval I^t. Therefore, we need to introduce a temporal evaluation of each keyword usage to analyze this property. For this, we define a keyword as *emergent* if it results to be *hot* in the considered time interval but not in the previous ones. In other words, we analyze the keywords, life-cycles by comparing their nutrition values obtained on the considered time frame with the usage of the same terms in the past time intervals. Namely, we analyze the current nourishment in comparison to the ones built up in the previous time intervals.

Let us consider the examples shown in Figures 2 and 3; considering the time frame from August 2011 to February 2012, the term “car” is, obviously, significantly more used by the Twitter community than the term “concordia”. Thus, according to our definition, the term “car” can be considered *hotter* than “Concordia”. On the other hand, considering the specific time instant defined by the day 13th January 2012, the term “concordia” can be easily seen as *emerging* keyword, while there is no significant change in the usage of term “car” (because of the extensive use of the term which can be correlated to the sinking of the cruise ship “Costa Concordia” on 13 January 2012).

Obviously, it is not necessary to consider the complete usage history of each keyword: in fact, if, for example, the keyword “Iraq” has been extensively used in past intervals, it does not mean that it cannot become again important in the community for another event in future time frames. However, if its nutrition value stays constant during closer time intervals (for example, two intervals in the same day), it means that the community is probably still referring to the same news event. In this case, according to our definition, even if the keyword can be considered as *hot*, it cannot be referred as *emerging* due to this temporal discrimination.

It is important to notice that this temporal parameter strictly affects the type of the detected emerging keywords retrieved by the system. Let us consider, for example, the keywords reported by only one user through her tweets: if we only consider a short keyword's history (for example, by taking into account only such intervals included in 24 hours), the system will only detect such keywords that emerge in a daily perspective (referring to her daily activities, related for example to her job or hobbies). Otherwise, if we consider a longer history (i.e., all the intervals included in a calendar year), the resulting emerging keywords will represent globally relevant activities that modify the general daily trends (for example, unexpected facts or events).

Therefore, we introduce a parameter s, where $0 < s < t$, that limits the number of previous time slots considered by the system to study the keywords' lifecycles and defines the *history worthiness* of the resulting emerging keywords.

Now given a keyword k, it is possible to calculate its burstiness value at the time interval I^t as

$$burst_k^t = \sum_{x=t-s}^{t-1} \left(\left(\left(nutr_k^t\right)^2 - \left(nutr_k^x\right)^2 \right) \cdot \frac{1}{log(t-x+1)} \right), \tag{6}$$

where $nutr_k^x$ represents the nutrition obtained by the keyword k during the interval time I^x.

This formula permits to quantify the usage of a given keyword k with respect to its previous usages in a limited number of time intervals. In fact, considering two distinct time intervals I^x and I^t, with $x < t$, this formula quantifies the difference in terms of usage of a given keyword, by considering the difference of nutritions received in the time frames I^x and I^t, and taking also into account the temporal distance among the two considered intervals.

3.3. Selection of Emerging Terms

In the previous section we presented our approach to model the lifecycle of the content for an automatic extraction of emerging terms.

In this section we propose a user-driven and a fully automatic technique to select a limited set of relevant terms that emerge in the considered time interval.

At the end of this step, given a set of tweets TW^t generated on the time interval I_t, we calculate a limited set of keywords EK^t that are emerging on the considered time interval.

3.3.1. User-Driven Keywords Selection. The first approach for the selection of emerging terms relies on a user-specified threshold parameter. Our initial assumption is that, given two keywords with very high burstiness values, they can be considered as emerging or not depending on the user evaluation. Indeed, if a user wants to be informed only about the most emerging events (i.e., when world-wide distribution of users reported it on a big scale, as for disasters and/or world-wide tragedies), she probably prefers to avoid such contents that are only relatively emerging (for example, topics referring to a less globally important news event).

In order to do that, we introduce a *critical drop* value that allows the user to decide when a term is *emergent*. In particular, the critical drop is defined as

$$drop^t = \delta \cdot \frac{\sum_{k\in K^t}(burst_k^t)}{|K^t|}, \tag{7}$$

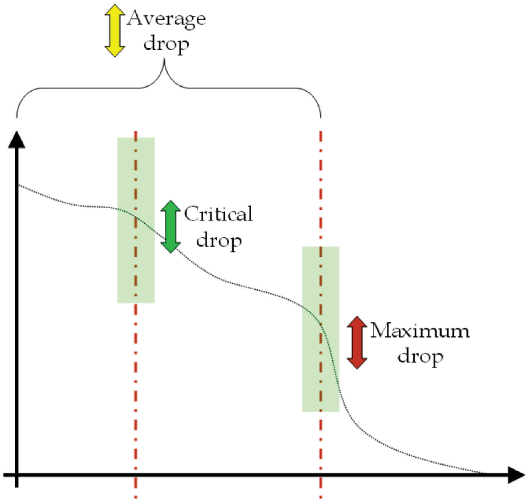

Fig. 4. The critical drop cut-off: the maximum drop is the highest variation in the ordered list of weights (red mark). The average drop (between consecutive entities) is the average difference between those items that are ranked before the identified maximum drop point (yellow mark). The first drop which is higher than the computed average drop is called the critical drop (green mark).

where $\delta \geq 1$. It permits to set the critical drop by also taking into account the average burstiness value. Therefore, we define the set of emerging keywords EK^t as

$$\forall k \in K^t, k \in EK^t \iff burst_k^t > drop^t. \tag{8}$$

It is possible to notice that the cardinality of EK^t is directly proportional to the value of δ.

3.3.2. Automatic Keywords Selection. An intrinsic limitation of the ranking approach introduced in Section 3.3.1 is that it could be very hard for a user/application to set a proper δ value. In fact, it can be a hard task to numerically quantify a threshold from an abstract perception as the desired cardinality of emerging keywords. Moreover, depending on the temporal context, it could be necessary to set the threshold value differently. Thus, we leverage, as explained in Cataldi et al. [2009], a fully automatic ranking model that dynamically sets the *critical drop*. In particular, following this strategy, for each keyword expressed by the community, the most emergent are those whose burstiness is above an adaptively computed critical point (Figure 4). Intuitively, to preserve only the keywords with very high burstiness, we consider only those terms whose burstiness is higher than the average burstiness of the most discussed terms. In order to cope with this, we provide a completely automatic model that works as follows.

(1) The system first ranks the keywords in descending order of burstiness value previously calculated in Section 3.2.2.
(2) It computes the *maximum drop* between consecutive entries and identifies the corresponding drop point.
(3) It computes the *average drop* (between consecutive entities) for all those keywords that are ranked before the identified maximum drop point.
(4) The first drop which is higher than the computed average drop is called the *critical drop*.

At the end of this 4-step process, the keywords ranked before the critical drop are defined as emerging keywords on time interval I^t and are collected in set called EK^t.

3.4. Leveraging Users' Context for Personalization Purposes

While the formula presented in Section 3.2.2 permits to monitor in real time the stream of information expressed by the Twitter community, it does not take into account any user preference and/or context. For this, we introduce a normalization that uses the content generated by the user (i.e., the tweets/retweets generated by her) to personalize the list of emerging topics.

More in detail, we first analyze the temporal time frames in which the user has been active within the Twitter environment. For this, we calculate the entire user's *life activity* as the time range included between the instant of her first tweet (or retweet) and the current instant.

From this, we segment the user's life activity into time intervals by taking into account her frequency of posting. The rationale is that any user has different habits and abilities with respect to the social network environment and expresses herself at different rhythm. Thus, we aim to take into account these different behaviors by analyzing the personal attitude with respect to the posting activity.

For this, we divided the user's life activity into numbered intervals, by segmenting the user life time based on the number of tweets generated by the user. Thus, with the function $interval(tw_j, u_i)$, we are able to determine in which interval, relatively to the considered user, each tweet has been generated. The lower the interval value, the more recent the tweet with respect to her posting frequency. Note that the value of the $interval$ function ranges from 1 to k, where k is the farthest time interval while 1 indicates the most recent one. This permits to take into account *when* the user expresses an interest to a topic (by reporting a correlated tweet) and therefore numerically quantify her interest into the topic according to this temporal information.

At this point, given the set of tweets TW_{u_i} generated by the user u_i, we define the context of the user u_i as her vocabulary K_{u_i} containing the set of terms typed by the user in any of her tweets (or retweets). From this, we calculate the weight of each term z in K_{u_i}, within the context of u_i, by considering the relative time interval in which it has been generated. In detail, it is calculated as

$$p_z^{u_i} = \sum_{tw_j \in TW_{u_i}} \frac{w_{z,j}}{log(interval(tw_j, u_i) + 1)}, \tag{9}$$

where $w_{z,j}$ is the weight of the z-th vocabulary term in the j-th tweet by using the augmented normalized term frequency [Salton and Buckley 1988] (calculated as explained in Section 3.1).

At this step, we can contextualize the importance of each retrieved emerging term (as explained in Section 3.3), by taking into account the importance of the term within the context of the considered user.

The problem of matching user contexts with emerging terms lies in the shortness of the tweets, that likely makes the contexts vocabulary abridged as well. For this reason, the evaluation of the similarity between these two entities needs to be carefully chosen. In fact, the naive solution of calculating exact matches between context terms and emerging terms avoids to consider both morphological variations and similarities between different words with shared meanings. In that sense, we made use of the system proposed by Ponzetto and Strube [2007] to estimate words' similarities based on statistically significant co-occurrences within large corpora, like Wikipedia. An advantage of using such a resource relies on its multi-language nature (as in our system).

In detail, taking a pair of terms as input, Ponzetto and Strube [2007] proposed to retrieve the Wikipedia articles they refer to, and compute a semantic similarity between the two terms based on the paths, in the Wikipedia categorization graph, that

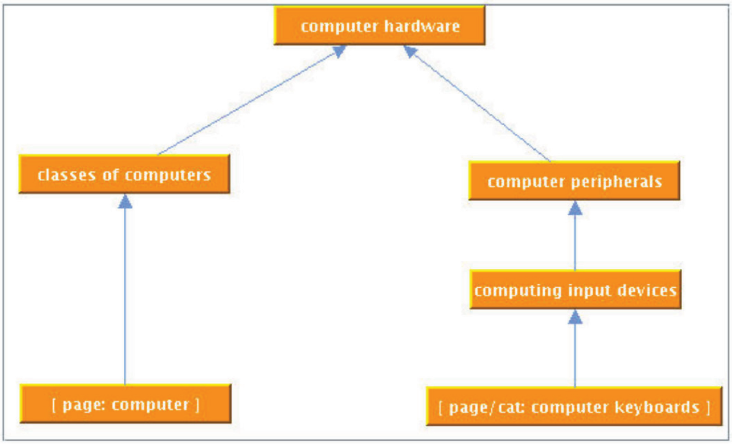

Fig. 5. The shortest path in Wikipedia between the terms “keyboard” and “computer”, which provides the base for calculating the semantic similarity between the terms (0.74 in this case).

connect these articles. This way, we calculate the burstiness of an emerging keyword k relatively to the user u_i as

$$burst^t_{k,u_i} = burst^t_k * \left(1 + \frac{\sum_{z \in K_{u_i}} (sim(k,z) \cdot p^{u_i}_z)}{|K_{u_i}|}\right), \tag{10}$$

where $|K_{u_i}|$ is the cardinality of the user context vocabulary, while sim is the function that returns a $[0,1]$-similarity value between the emerging keyword and a context keyword, according to the method presented by Ponzetto and Strube [2007]. Note that, if the semantic relatedness between the two terms results unknown (for example, one of them is not included in the considered Wikipedia dataset), it is estimated as 0. On the other hand, if $k = z$, the similarity returned by the system is equal to 1[14]. An example, related to the semantic similarity between the term “keyboard” and “computer”, is shown in Figure 5.

Following this strategy, if an emerging term results outside the interests of the user (because no semantic paths emerged, within the Wikipedia dataset, between the terms within the user context and the emerging one), we do not lower the overall burstiness of the considered emerging keyword. In fact, we only proportionally increase the burstiness of the emerging terms when they resulted semantically close to some of the interests of the user. We believe that this strategy permits to preserve the most discussed arguments and therefore inform the users about the most important news/topics expressed by the community, even if they are not specifically interested in the domain. Following this high-level assumption, when global events happen and they are extensively commented in Twitter, all the users can remain informed even if they are not specifically interested in the domain. This assumption will be further commented within the experimental analysis (Section 4.2).

[14] Please notice that, considering that the semantic similarity of each pair of terms is based on the external knowledge base of Wikipedia, this information is precomputed offline and it is periodically updated in, order to reflect the novel information introduced in Wikipedia.

3.5. From Emerging Terms to Emerging Topics

Considering the given corpus of tweet TW^t, extracted within the time interval I^t, in this step we study the semantic relationships that exist among the keywords in K^t in order to retrieve the topics related to each emerging term.

In our system we define a *topic* as a minimal set of a terms semantically related to an emerging keyword. Thus, in order to retrieve the emerging topics, we consider the entire set of tweets generated by the users within the time frame I^t, and we analyze the semantical relationships that exist among the keywords by examining the co-occurrences information.

Let us consider, for example, the keyword "*victory*" in a given set of tweets: this term alone does not permit to express the related topic. In fact, considering as a time frame November 2008, the related topic can be easily defined by the association with other keywords (among the most used) as "elections", "Us", "Obama", and "McCain", while in another time frame, as for example February 2012, the term could be related to a sports event by other keywords such as "superbowl" (due to the championship game of the National Football League (NFL)), "football", or "New York Giants" (the name of one of the teams who played the final game in 2012).

Thus, in order to express the topics related to the retrieved emerging keywords, we consider the time frame in which all the tweets have been generated and analyze the semantic relationships among the keywords based on the co-occurrences information.

We formalize this idea by associating to each keyword $k \in K^t$ a *correlation vector* $\vec{cv}_k^t$, formed by a set of weighted terms, that defines the relationships that exist among k and all the other keywords in the considered time interval.

In other words, we compute the degree of correlation between a keyword k and another keyword z by using the set of documents containing both terms as positive evidence of the relationship between the two keywords, and the set of documents containing only one of them as positive evidence against the relationship. In detail, we treat each keyword k as a query and the set of the tweets containing the keyword TW_k^t as the *explanation* of this term in the time interval I^t.

Intuitively, this is analogous to treating: (a) the keyword k as a query and (b) the set of tweets containing k as relevance feedback on the results of such query. Recognizing this, we identify the correlation weight $c_{k,z}^t$, between k and another keyword z at time I^t relying on a probabilistic feedback mechanism [Ruthven and Lalmas 2003]

$$c_{k,z}^t = log \frac{r_{k,z}/(R_k - r_{k,z})}{(n_z - r_{k,z})/(N - n_z - R_k + r_{k,z})} \times \left| \frac{r_{k,z}}{R_k} - \frac{n_z - r_{k,z}}{N - R_k} \right|, \tag{11}$$

where:

—$r_{k,z}$ is the number of tweets in TW_k^t containing the keywords k and z;
—n_z is the number of tweets in the corpus containing the keyword z (it is equal to $|TW_z^t|$);
—R_k is the number of tweets containing k (it is equal to $|TW_k^t\|$); and
—N is the total number of tweets.

Notice that the first term increases as the number of the tweets in which k and z co-occur increases, while the second term decreases when the number of tweets containing only the keyword z increases.

Thus, given a term k, we associate a so-called *correlation vector*

$$\vec{cv}_k^t = \langle c_{k,1}, c_{k,2}, \ldots, c_{k,v} \rangle, \tag{12}$$

which represents the relationships that exist between k and the other v keywords (with $v = |K^t|$) at the time interval I^t.

At this point we leverage information vehiculated by the *correlation vectors* in order to identify the topics related to the emerging terms retrieved during the considered time interval. In order to do that, we construct a keyword-based topic graph in the form of a directed, node-labeled, edge-weighted graph, $TG^t(K^t, E, \rho)$, as follows.

—Let K^t be a set of vertices, where each vertex $k \in K^t$ represents a keyword extracted during the time interval I^t.
—For all $k \in K^t$ and $z \in K^t$ such that $\vec{cv}_k^t[z] \neq 0$, there exists an edge $\langle k, z \rangle \in E$ such that

$$\rho(\langle k, z \rangle) = \rho_{k,z} = \frac{\vec{cv}_k^t[z]}{\left\| \vec{cv}_k^t \right\|}.$$

Therefore $\rho_{k,z}$ represents the relative weight of the keyword k in the corresponding vector $\vec{cv}_k^t$, namely the role of the keyword z in the context of the keyword k.

Finally, the complete graph $TG^t(K^t, E, \rho)$ is thinned by applying a locally adaptive edge thinning algorithm. For each $k \in K^t$, we consider the set of all outgoing edges and we apply an adaptive cut-off (as explained in Section 3.3.2) based on the corresponding weights. This process ensures that only those edges that represent the strongest relationships are maintained (note that, since the graph is directed and the thinning process is asymmetrical, it is possible that TG^t will contain the edge $\langle k, z \rangle$ but not vice versa).

3.6. Topic Detection, Labeling and Ranking

Since in our system each topic is defined as a set of semantically related keywords, we leverage the topological structure of the topic graph TG^t to detect the emerging topics into the Twitter community. In order to do that, we consider the set of emerging keywords EK^t, computed as described in Section 3.3, and we search for the Strongly Connected Components (SCC) rooted on them in TG^t.

Therefore, given a keyword k that represents a vertex within the topic graph TG^t, we find the set of vertices S reachable from k through a path, simply applying a Depth-First Search (DFS) visit (or any other similar algorithm). Then, we repeat the process on the same topic graph TG^t with reversed edges in order to find the set of vertices T that can reach k through a path. The strongly connected component EK_k^t is formed by all the vertices within the intersection between T and S. The complexity of this process is linear.

Thus, for each emerging keyword $z \in EK^t$, we define the related *emerging topic* as a subgraph $ET_z^t(K_z, E_z, \rho)$ representing a set of keywords semantically related to the term z within the time interval I^t. Considering the entire set of emerging keywords EK^t, in this step we compute the corresponding set of emerging topics as $ET^t = \{ET_1^t, \ldots, ET_n^t\}$ of strongly connected components. It is important to notice that the number of the retrieved emerging topics can be lower than the number of the emerging keywords ($n \leq |EK^t|$); in fact two emerging keywords can belong to the same emerging topic.

At the end of this step, the set of keywords K_z^t belonging to the emerging topic ET_z^t is calculated by considering as starting point in TG^t the *emerging* keyword z, but also contains a set of common terms semantically related to z but not necessarily included in EK^t.

Figure 6 is depicted as an example in which each topic is represented by a different color. As it is possible to notice, each strongly connected component contains both *emerging* terms (labeled in bold) like “Concordia”, “Wulff”, and other popular keywords, like “Merkel”, “sea”, and “EU”, that are constantly used by Twitter users and do not represent statistically emerging terms. In fact, terms like “Merkel” or “EU” represent

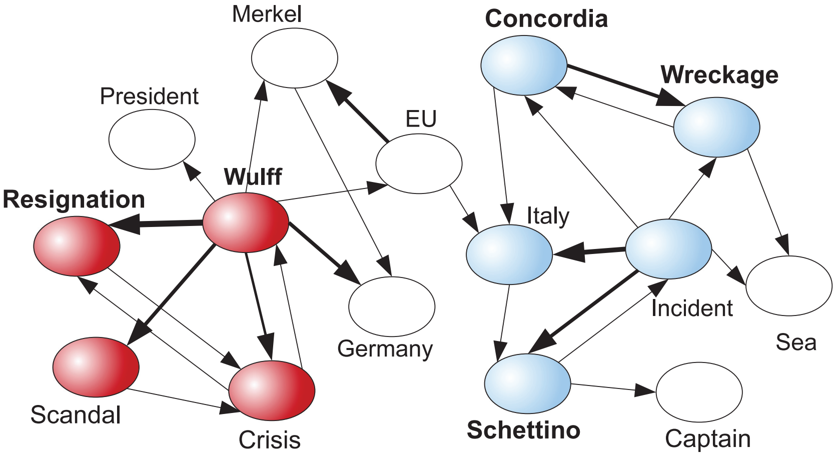

Fig. 6. A topic graph with two strongly connected components (in red and blue) representing two different emerging topics: labels in bold represent emerging keywords while the thickness of an edge represents the semantical relationship between the considered keywords.

very popular terms always reported by the users in Twitter as in any other information sources.

Notice that, with this approach, not only do we retrieve such terms that directly co-occur with the emerging terms but we can also retrieve those which are indirectly correlated with the emerging ones (by co-occurring with keywords that they themselves co-occur with the emerging terms). In fact, the considered topic graph leverages the information contained in all the tweets, even those that do not report emerging terms; indeed a user can always report an emerging topic simply using synonyms.

Thus, as the last step, we need to establish an order among the retrieved topics in order to guide the user in understanding which topic is *more* emergent in the considered time frame. In order to do that, we introduce a *ranking* value as

$$rank_{ET_z^t} = \frac{\sum_{k \in K_z^t}(burst_k^t)}{\left|K_z^t\right|} \tag{13}$$

that leverages the average burstiness of the terms in K_z^t to define the importance of the topic led by the emerging keyword $z \in EK^t$. Finally, using this value we are able to rank the retrieved topics in descending order of importance.

At this point, the system has found a set of topics ET^t emerging within the time interval I^t. Nevertheless, it is now necessary to carefully select a minimal set of keywords (belonging to the considered topic) to represent each retrieved emerging topic to the user.

In fact, we believe that too many keywords could represent an information overload for the user; on the other hand, the cardinality of the retrieved emerging topics strictly depends on the topological structure of the topic graph TG^t.

In order to avoid this problem, given a topic $ET_z^t \in ET^t$ and the related keywords K_z^t, we apply an unsupervised keyword ranking mechanism (as described in Section 3.3.2) that permits to select the most representative keywords for each cluster.

Notice that, even using this adaptive cut-off, there is no guarantee to obtain a relatively small set of keywords representatives. Thus, we also introduce a numerical threshold χ, set by the user/application, that limits the representative keywords of each topic. This threshold can be set depending on visualization constraints of the device and/or the user preferences.[15]

4. EXPERIMENTS: CASE AND USER STUDIES

In this section, we evaluate the proposed method by analyzing real case scenarios and user studies. In particular, we conducted several experiments by monitoring the Twitter community during the period included between 10th January 2012 and 4th July 2012. In our experiments we have monitored a stream which consists of two random samples (taken in two different time intervals) among all public messages. This access level provides a statistically significant input for data mining and research applications[16]; in fact, we evaluated more than 15 millions of tweets, generated by ~1 million Twitter users, which included more than 600k different keywords.[17]

The main aim of the experimental evaluation was twofold: from one side analyze, through examples and case studies, the impact and the proper setup of each parameter proposed within the presented technique. On the other hand, considering that our approach is also based on the idea to take into account the users' interests, we asked different users to provide a real feedback about the retrieved personalized topics. Moreover, as in Lu et al. [2012], we compared the retrieved topics to the ones reported by Associated Press Web site[18], evaluating the precision of the presented methods.

In particular, for the first task, we analyzed different experimental setup strategies: we initially considered a real case study by evaluating the emerging topics retrieved by the system within two different time intervals. Then, we varied the number of considered time intervals (Section 3.2.2) in order to determine how this parameter affects the quality of the retrieved topics. We also compared the presented keywords selection methods (Section 3.3) evaluating their impact on the resulting emerging topics.

For the second objective of this experimental evaluation, we asked human users to subjectively judge the topics and the personalization strategy through a questionnaire. Please notice that the evaluation of a personalized topic detection strategy results in a very complex task; in fact no manually annotated dataset exists for this problem and the unique possible ground truth is provided by the users. Based on these considerations, we asked different users, with different background, age, and interests, to reply to a very diverse set of questions that aim at covering the problems faced by our approach. We also analyzed the advantages and disadvantages of the proposed approach highlighting possible future works and strategies for improvements.

4.1. Experiments Based on Real Case Scenarios

As previously reported, we initially evaluated the proposed approach by analyzing the retrieved emerging topics within the considered time period. For this experiment we considered the automatic selection method and we set the time range r as 30 minutes[19] (Section 3.1). As explained in the Introduction, considering the impressive response time of the users, our assumption is that, if continuously monitored within small time ranges, Twitter can also predate the most authoritative news sources in informing the community about emerging news events. Obviously, it is possible to set higher time

[15] We used $\chi = 5$ as default value.

[16] http://apiwiki.Twitter.com.

[17] The sampling rate of the used standard Twitter account is 1% over an average of 200 millions per day.

[18] http://www.ap.org.

[19] In our experimental evaluation we set r equal to 30 in order to adapt our system to the high dynamicity of Twitter users. This is clearly in agreement with the experimental results shown in Asur et al. [2011]. In fact, in this work, the authors revealed that there are few topics that last for longer times, while most topics decay in about 20–40 minutes.

Table I. The Emerging Topics Retrieved by the System Based on the Five Most Emerging Terms (in the time period included between January 10th and February 20th 2012)

Date	Emerging Topics
13-01-2012	{**concordia**, cruise, disaster, schettino, sea}[20]
13-02-2012	{**adele**, grammy, 2012, perry, kate}[21]
06-02-2012	{**jubilee**, uk, england, diamond, elizabeth, II}[22]
01-02-2012	{**port**, killed, match, said}[23]
04-02-2012	{**protest**, kremlin, putin, government,}[24]

range values: in this case, the resulting topics will be statistically more significant (a higher number of authors certifies the importance of the argument) but the advantage in terms of time with respect to the traditional information source will vanish. The number s of time slots considered was 400 (with the selected time range, we considered more than a week); in fact this time slot size is sufficient to avoid such terms that are significant in a daily/weekly perspective (see also Section 4.1.1).

Using this experimental setup, we totally collected 147 emerging topics. At this point, in order to test the precision of the proposed method, we asked to two external domain experts to manually compute the precision of the proposed method[25]. For this, as in Lu et al. [2012], we considered as baseline the RSS (Really Simple Syndication) of the Associated Press Web site. The resulting average precision was 0.892, proving the effectiveness of the proposed system in detecting relevant topics of discussion (because they deserved to be commented in authoritative information services).

In Table I we also show the retrieved topics based on the five most emerging terms (i.e., the terms with the highest burstiness values). The emerging terms are visualized in bold. We also link the professional news articles of the retrieved news topics. It is possible to notice that the emerging terms are always the most specific ones (i.e., "concordia", the name of the Italian cruise ship that partially sank on the night of 13 January 2012; in fact, they represent keywords that are generally very unusual in the community and only emerge in correspondence to unexpected events. However, by analyzing the co-occurrences information in the tweets reporting such terms (as explained in Section 3.5), it is possible to link them to other popular keywords (such as "sea", in the topic led by "concordia") that have very common usages in the community.

4.1.1. History Worthiness. As reported in Section 3.2.2, a term is defined as *emerging* by evaluating its usage in a limited number of previous time slots. However, depending on the selected numbers of considered time intervals, the retrieved topics can significantly differ. Thus, in order to study the impact of this parameter, we analyze two different

[20] http://www.bbc.co.uk/news/world-europe-16558910.

[21] http://www.nytimes.com/2012/02/14/arts/music/at-the-54th-grammy-awards-everything-old-is-praised-again.html.

[22] http://www.guardian.co.uk/uk/queen-diamond-jubilee.

[23] http://www.bbc.co.uk/news/world-middle-east-16845841.

[24] http://www.time.com/time/world/article/0,8599,2106183,00.html.

[25] Notice that we do not compute any recall value. In fact, recall is strictly dependent on the nature and the focus of the considered ground truth (CNN, AP, some online newspaper, etc.). For example, some important news article in an authoritative newspaper could report a very deep political analysis of some high-level global political situation that is not referred to any specific very recent topic (or can be linked to many specific events). In fact, considering that Twitter is a user-generated content depository, it contains a wide range of comments and posts about high-level/global events but very few (also for the nature of this microblogging service) in-depth analyses of the news. For this, counting how many news articles are detected as emerging topics results in not being significant for the considered task.

Table II. The Emerging Topics Retrieved at 20th April 2010 by considering (a) a Daily History Worthiness ($s = 50$) and (b) a 7-Days History Worthiness

Date	Emerging Topics ($s = 50$)
20-01-2012	{**tonight**, working, dead, work}
20-01-2012	{**weekend**, ready , start, friday}
20-01-2012	{**weather**, nice , cold, morning, then}
20-01-2012	{**school**, day , later, sleep, morning}

(a)

Date	Emerging Topics ($s = 300$)
20-01-2012	{**etta**, james, death, today, rip}
20-01-2012	{**army**, thx, day}

(b)

The labels in bold represent emerging terms.

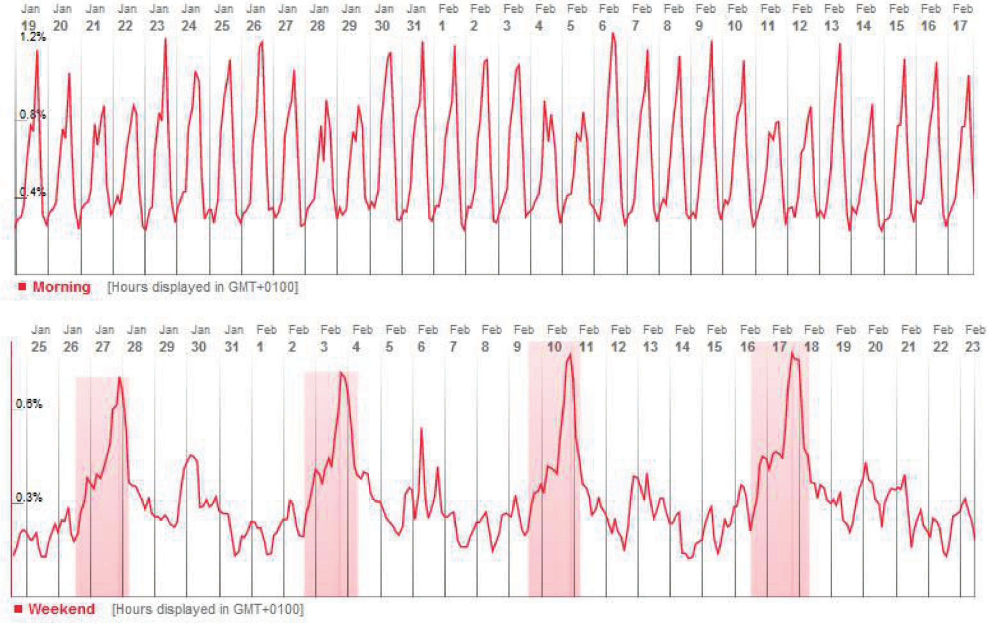

Fig. 7. Statistical usage in Twitter of the terms “morning” and “weekend” (provided by Trendistic, https://twitter.com/trendistic) in consecutive days and/or weeks.

numbers of considered slots, $s = 50$ and $s = 400$ (by considering again a time range r of 30 minutes and the automatic selection method).

In Table II(a) and (b) we present the results obtained on January 20th 2012: the most emerging topics obtained by setting $s = 50$ represent common daily activities from a user perspective. Most of the terms, indeed, can be considered as emerging only if the system does not take into account comparable time intervals. Terms like “weekend” have very standard usages in the community due to the fact they represent periodic events. We guess that users systematically use such terms in correspondence to their natural occurrence. For example, in Figure 7 we show the usage of the terms “morning” and “weekend” in consecutive days; the system reported a peak if it only takes into account a 24-hours history (or 7-days period); however, if it considers a relatively

Table III. The Five Most Emerging Terms (and their related burstiness values) Retrieved by the Proposed Temporal-Based Topic Detection System within the Considered Time Interval

Date	Emerging Terms	Burstiness value (*total avg = 135.67*)
13-01-2012	**concordia**	193461.83
13-01-2012	**carre**	9677.42
13-01-2012	**sarah**	6764.79
13-01-2012	**dark**	3284.39
13-01-2012	**holmes**	1345.25

The total average burstiness value was 135.67 (among all the keywords typed within the considered time interval).

higher time frame, it recognizes a constant pattern in time. Thus, the life status of a keyword strictly depends on the considered number of time intervals (Section 3.2.2) and this value directly affects the temporal relevance of the retrieved topics.

4.1.2. User-Selected Threshold vs Automatic Ranking Mechanism. In Section 3.3 we reported two different selection methods (user-selected threshold and automatic) to identify emerging terms. In this experiment we evaluate the impact of each of them in the retrieved topics by analyzing the example proposed in Table I: each of the considered five emerging terms (“concordia”, “adele”, “jubilee”, “port”, and “protest”) was identified as *emergent* in different time intervals.

In order to understand why they have been considered as emerging, we need to analyze the considered intervals and their associated burstiness values. Let consider the time interval related to the term “concordia”: in Table III we show the five most emerging terms and their burstiness values retrieved by the system on January 13th 2012. We notice that, considering the automatic approach (Section 3.3.2), only the term “concordia” has been considered as emergent. In fact, the system identified the difference in terms of burstiness values between “concordia” and the second most emerging terms (“carre”—which refers to the writer John Le Carré, —) as the critical drop point and only considered as emerging such keywords that are ranked better than the critical drop.

However, considering the user-driven selection method (Section 3.3.1), the retrieved emerging terms depend on the δ value set by the user. In fact, considering for instance $\delta = 1000$ (i.e., each term is identified as emergent only if its burstiness value is 1000 times higher than the total average burstiness value), only the term “concordia” is selected as emergent. Instead, with $\delta = 10$, also the terms “carre”, “sarah”, and “dark” will be considered as emergent (and analyzed using the topic graph to retrieve the related topics).

4.2. User Study: Evaluating Personalization and Topic Detection Strategies

In order to analyze the personalization strategy of the proposed topic detection technique, we conducted a user study by asking multiple Twitter users to reply to a questionnaire when evaluating the emerging topics, retrieved by the system, within their interests. The users represent a various range of ages, backgrounds, jobs, interests, and education level and they have intermediate Web ability (they are not computer scientists or social networks analysts). Among all, 64 Twitter users replied to our call.

For this experiment, we asked the users to analyze and report their opinions about the most emerging topics within two different time intervals; the first one from February 6th 2012 to February 20th 2012 and the second one from June 25th 2012 to July 4th 2012. In order to also test the reliability, we involved in the two tests different users with different foci of interest. The results are shown in Table IV.

Table IV. Two Subjective Questionnaires Run on Different Time Intervals and Involving Different Users

Time Interval	Closeness	Importance	Novelty	Understandability
February 2012	2.94 (0.54)	4.22 (0.56)	4.31 (0.45)	4.44 (0.41)
June 2012	3.21 (0.40)	3.98 (0.36)	4.61 (0.52)	4.12 (0.36)
Avg	3.07	4.1	4.46	4.28

They replied to four questions related to world-wide importance of the retrieved topics, understandability, closeness to their interest, and degree of novelty of the retrieved topics of discussion (variance is reported between parentheses).

As first question, we asked to evaluate, with a 5-point scale ratings, how much the returned topics could be considered within their interests (expressed by their Twitter activity). The higher the ratings, the more adherent the emerging topics to the user's interests. The result showed an average rate of 3.07, highlighting an intermediate level of pertinence of the retrieved topics to their topic interests. Note that, even if this result can appear lower than desired, it permits to report an important consideration. The personalization approach proposed in Section 3.4 does not filter the topics that most emerged when they are not within the user interests; in fact, it only highlights the discussions closer to the interests of the user but it does not filter the emerging ones that are not explicitly within them. In other words, this strategy permits to inform the users about the most important news/topics expressed by the community even if they are not specifically interested in the domain. With this approach, when global news events happen (and therefore they are extensively discussed within the community) the user can remain informed even if she is not specifically interested in the domain. Let us consider, for example, the case of the Haiti earthquake (happening in 2010) and/or the tsunami that hit Japan in 2011. Twitter collected a huge amount of data coming from people in directly affected areas [Acar and Muraki 2011]. We can easily believe that very few people are specifically interested in monitoring these information domains per se, but every person could desire to be informed about these global events when they happen.

To better evaluate this assumption, we asked to the user to report her opinion (with a 5-point scale rating) about the overall importance of the reported topics. For this reason, we asked the user the following question: “Can this be considered the reported topics of general and world-wide interest?” The user reported her opinion for each of the presented topics. The average value was 4.1, demonstrating that the system was able to identify world-wide emerging discussions that can be considered interesting to any user, independently from the specific domain interests.

Moreover, the third question investigated the ability of the proposed mechanism to predate traditional information systems (newspapers, television, blogs, etc.). Specifically, we asked the user if she already read about the reported discussion, within alternative information channels, before being informed by our system (“Have you already heard about the topics reported by the system?”). The users expressed, again, their responses with a 5-point scale rating (the higher the rate, the more unknown the related news). The feedback of the users within the first time interval, with an average rate of 4.46, clearly showed the ability of the proposed topic detection and tracking mechanism to predate, in real time, the traditional information systems in reporting fresh, emergent information news.

Finally, each user also replied to the question “Were the reported topics easy to understand?” The users replied that the reported topics were, in average, very “easy to understand” (with an average subjective rate of 4.28), highlighting the fact that the proposed topic construction strategy (Section 3.5) is able to associate a minimal set of terms to each most emerging keyword that can effectively summarize the overall discussion.

Please also notice that the results are consistent between the two considered time intervals, highlighting the ability of the proposed approach to automatically detect emerging topics in different conditions, with different users, and with different domains of interests.

In conclusion, we can summarize these results as follows: as initially supposed, the proposed topic detection mechanism is able to predate the classic information channels in informing users about the most discussed news events, in a way that resulted in being both compact and understandable from a human point of view. In fact, the user study highlighted the fact that the topic construction mechanism provides sufficient details (i.e., number of keywords) to understand the core of the related event/discussion. Moreover, even when the reported topics were not within the users' interests, they were important to be reported for their global importance.

5. CONCLUSIONS AND FUTURE WORKS

This work is part of a wider ongoing research activity on social networks, topic detection, tracking, and user modeling. Starting from our previous work aimed at detecting real-time emerging topics on Twitter [Cataldi et al. 2010], we focused our attention on how the interests of the users can be exploited to contextualize such mined knowledge. More in detail, we based our approach on the assumption that everything that appears to be of high interest for all Twitter users world-wide should be of interest to anyone. Thus, we did not intend to leverage user profiles as central bases for the contextualization of the extracted topics. In fact, our idea was that user contents and interests only need to be accentuated and added to world-level important events.

In this work, we first formalized the lifecycle of a term occurring in the tweets through an aging theory that aimed at modeling the temporal usage of each keyword expressed by the community. The assumption was that important topics start from little information (i.e., single terms) that acquire unexpected importance within the user-generated data. Then, we took into accounting for the social relationships in the Twitter network (i.e., follower and following relationships) to estimate an authority level of the authors, thus normalizing their propagation of information in the network. Finally, we constructed the topics as minimal sets of keywords co-occurring with high emerging ones. In addition, we considered the user context (given by the posted tweets) to highlight those topics that better match her specific interests.

Within our experimentations, we found that our approach was able to extract important events even before the most important information media, that usually needed a certain amount of time to react. We presented real case scenarios that showed the effectiveness of our proposal and explained the usages of the introduced parameters. Then, we carried out two user studies aimed at highlighting the validity of our user-aware topic detection mechanism for monitoring, in real time, the Twitter information network and its latent informative role.

The related work presented in Section 2 showed that Twitter is the ideal scenario for the study of real-time information spreading phenomena. Moreover, we think that our approach could be used in the communication network analysis of both more similar microblogs, like Weibo[26], and systems like Facebook or, more in general, in online forums.

Considering possible future works on this direction, we aim at facing some challenging problems and limits that we have encountered throughout this work. Among all, the proposed approach is not able to distinguish between a news event and irrelevant

[26] http://www.weibo.com.

hot topics (e.g., discussions on facts about celebrities). In this sense, we will study how to distinguish them through an event and entity detection technique, taking into account their presentation to the user (separately, and when requested). Moreover, for what concerns the computation of the user authority, it is important to note that our proposal can support other existing approaches, based on PageRank [Weng et al. 2010] rather than different algorithms and information, like retweets [Cha et al. 2010]. The work presented in Castillo et al. [2011], for instance, offers interesting possibilities in that sense.

Finally, we would aim at evaluating the existing correlation between the popularity of a user (intended as the capacity to influence other users to post related tweets, measured in some way) and the information spread. In detail, our goal would be to investigate on this aspect and analyze its impact in research approaches for topic detection and tracking techniques on information networks.

REFERENCES

ABROL, S. AND KHAN, L. 2010. Twinner: Understanding news queries with geocontent using twitter. In *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10)*. ACM Press, New York, 1–8.

ACAR, A. AND MURAKI, Y. 2011. Twitter for crisis communication: Lessons learned from japan's tsunami disaster. *Int. J. Web Based Communities 7,* 3, 392–402.

AGRAWAL, R., GOLLAPUDI, S., HALVERSON, A., AND IEONG, S. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. ACM Press, New York, 5–14.

ALLAN, J. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, Norwell, MA.

ALSUMAIT, L., BARBARA, D., AND DOMENICONI, C. 2008. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. 3–12.

ASUR, S., HUBERMAN, B. A., SZABO, G., AND WANG, C. 2011. Trends in social media: Persistence and decay. In *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*. AAAI Press.

BAKSHY, E., HOFMAN, J. M., MASON, W. A., AND WATTS, D. J. 2011. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. ACM Press, New York, 65–74.

BALABANOVIC, M. AND SHOHAM, Y. 1997. Fab: Content-based, collaborative recommendation. *Comm. ACM 40*, 66–72.

BARABASI, A. L., JEONG, H., NEDA, Z., RAVASZ, E., SCHUBERT, A., AND VICSEK, T. 2002. Evolution of the social network of scientific collaborations. *Physica A: Statist. Mech. Appl. 311,* 3–4, 590–614.

BECKER, H., NAAMAN, M., AND GRAVANO, L. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. ACM Press, New York, 291–300.

BECKER, H., NAAMAN, M., AND GRAVANO, L. 2011. Beyond trending topics: Realworld event identification on twitter. In *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*. AAAI Press.

BUN, K. K., ISHIZUKA, M., AND ISHIZUKA, B. M. 2002. Topic extraction from news archive using tf*pdf algorithm. In *Proceedings of 3rd International Conference on Web Information Systems Engineering (WISE'02)*. 73–82.

CANTADOR, I., BELLOGIN, A., AND VALLET, D. 2010. Content-based recommendation in social tagging systems. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*. ACM Press, New York, 237–240.

CARMEL, D., ZWERDLING, N., GUY, I., OFEK-KOIFMAN, S., HAREL, N., RONEN, I., UZIEL, E., YOGEV, S., AND CHERNOV, S. 2009. Personalized social search based on the user's social network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM Press, New York, 1227–1236.

CASTILLO, C., MENDOZA, M., AND POBLETE, B. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM Press, New York, 675–684.

CATALDI, M., DI CARO, L., AND SCHIFANELLA, C. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining (MDMKDD'10).* ACM Press, New York, 4:1–4:10.

CATALDI, M., SCHIFANELLA, C., CANDAN, K. S., SAPINO, M. L., AND DI CARO, L. 2009. Cosena: A context-based search and navigation system. In *Proceedings of the International Conference on Management of Emergent Digital* EcoSystems *(MEDES'09).* ACM Press, New York, 33:218–33:225.

CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th AAAI International Conference on Weblogs and Social Media (ICWSM'10).* AAAI Press, 10–17.

CHEN, C. C., CHEN, Y.-T., SUN, Y. S., AND CHEN, M. C. 2003. Life cycle modeling of news events using aging theory. In *Proceedings of the 14th European Conference on Machine Learning (ECML'03).* Springer, 47–59.

CHEN, J., GEYER, W., DUGAN, C., MULLER, M., AND GUY, I. 2009. Make new friends, but keep the old: Recommending people on social networking sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*. ACM Press, New York, 201–210.

CHEN, J., NAIRN, R., NELSON, L., BERNSTEIN, M., AND CHI, E. 2010. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10).* ACM Press, New York, 1185–1194.

CHUBIN, D. E. 1976. The conceptualization of scientific specialties. *Sociol. Quart. 17,* 4, 448–476.

CRANE, D. 1969. Social structure in a group of scientists: A test of the "invisible college" hypothesis. *Amer. Sociol. Rev. 3*, 335–352.

DE BEAVER, D. AND ROSEN, R. 1979. Studies in scientific collaboration. *Scientometrics 1,* 2, 133–149.

DI CARO, L., CATALDI, M., AND SCHIFANELLA, C. 2012. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics 93,* 3, 583–607.

FAVENZA, A., CATALDI, M., SAPINO, M. L., AND MESSINA, A. 2008. Topic development based refinement of audio-segmented television news. In *Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB'08).* Springer, 226–232.

GAUCH, S., CHAFFEE, J., AND PRETSCHNER, A. 2003. Ontology-based personalized search and browsing. *Web Intell. Agent Syst. 1*, 219–234.

GLANCE, N. S., HURST, M., AND TOMOKIYO, T. 2004. BlogPulse: Automated trend discovery for weblogs. In *Proceedings of the Workshop on the Weblogging Ecosystem*. ACM Press, New York.

GOLDBERG, D., NICHOLS, D., OKI, B. M., AND TERRY, D. 1992. Using collaborative filtering to weave an information tapestry. *Comm. ACM 35,* 12, 61–70.

GOLDENBERG, J., LIBAI, B., AND MULLER, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Lett. 12,* 3, 211–223.

GOYAL, A., BONCHI, F., AND LAKSHMANAN, L. V. 2010. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10).* ACM Press, New York, 241–250.

GRANOVETTER, M. 1978. Threshold models of collective behavior. *Amer. J. Sociol. 83,* 6, 1420–1443.

GRIFFITHS, T. L. AND STEYVERS, M. 2004. Finding scientific topics. *Proc. Nat. Acad. Sci. 101,* 1, 5228–5235.

GRUHL, D., GUHA, R., LIBEN-NOWELL, D., AND TOMKINS, A. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04).* ACM Press, New York, 491–501.

HAN, X., SHEN, Z., MIAO, C., AND LUO, X. 2010. Folksonomy-based ontological user interest profile modeling and its application in personalized search. In *Proceedings of the 6th International Conference on Active Media Technology (AMT'10).* Springer, 34–46.

HASSAN, A., RADEV, D. R., CHO, J., AND JOSHI, A. 2009. Content based recommendation and summarization in the blogosphere. In *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI Press, 34–41.

HE, Q., CHANG, K., AND LIM, E.-P. 2007. Using burstiness to improve clustering of topics in news streams. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'07).* 493–498.

HOU, H., KRETSCHMER, H., AND LIU, Z. 2008. The structure of scientific collaboration networks in scientometrics. *Scientometrics 75,* 2, 189–202.

JASCHKE, R., MARINHO, L., HOTHO, A., SCHMIDT-THIEME, L., AND STUMME, G. 2007. Tag recommendations in folksonomies. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07).* Springer, 506–514.

KATZ, J. S., KATZ, J. S., MARTIN, B. R., AND MARTIN, B. R. 1997. What is research collaboration? *Res. Policy 26*, 1–18.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM Press, New York, 591–600.

Lampos, V. and Cristianini, N. 2012. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol. 3,* 4, 72:1–72:22.

Leskovec, J. and Faloutsos, C. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM Press, New York, 631–636.

Liang, X., Chen, W., and Bu, J. 2010. Bursty feature based topic detction and summarization. In *Proceedings of the 2nd International Conference on Computer Engineering and Technology (ICCET'10),* vol. 6. 49–253.

Lin, G.-L., Peng, H., Ma, Q.-L., Wei, J., and Qin, J.-W. 2010. Improving diversity in web search results re-ranking using absorbing random walks. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC'10),* vol. 5. 2116 –2421.

Lu, R., Xu, Z., Zhang, Y., and Yang, Q. 2012. Life activity modeling of news event on twitter using energy function. In *Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'12)*. Springer, 73–84.

Makkonen, J., Ahonen-Myka, H., and Salmenkivi, M. 2004. Simple semantics in topic detection and tracking. *Inf. Retr. 7,* 3–4, 347–368.

Melin, G. and Persson, O. 1996. Studying research collaboration using coauthorships. *Scientometrics 36*, 363–377.

Melville, P., Mooney, R. J., and Nagarajan, R. 2001. Content-boosted collaborative filtering. In *Proceedings of the SIGIR Workshop on Recommender Systems*. ACM Press, New York, 16–23.

Moon, S., You, J., Kwak, H., Kim, D., and Jeong, H. 2010. Understanding topological mesoscale features in community mining. In *Proceedings of the 2nd International Conference on Communication Systems and Networks (COMSNETS'10).* 1–10.

Newman, M. E. J. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E 64,* 1.

Noll, M. G. and Meinel, C. 2007. Web search personalization via social bookmark. In *Proceedings of the 6th International Semantic Web and 2nd Asian Semantic Web Conference (ISWC'07/ASWC'07).* Springer, 367–380.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference (WWW'98)*. ACM Press, New York, 161–172.

Petrovic, S., Osborne, M., and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT'10).* 181–189.

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. 2011. Do all birds tweet the same? Characterizing twitter around the world. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11).* ACM Press, New York, 1025–1030.

Ponzetto, S. P. and Strube, M. 2007. An api for measuring the relatedness of words in wikipedia. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07).* 49–52.

Qi, Y. and Candan, K. S. 2006. Cuts: Curvature-based development pattern analysis and segmentation for blogs and other text streams. In *Proceedings of the 17th Conference on Hypertext and Hypermedia (HYPERTEXT'06).* ACM Press, New York, 1–10.

Radlinski, F. and Dumais, S. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06).* ACM Press, New York, 691–692.

Ruthven, I. and Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. *Knowl. Engin. Rev. 18,* 2, 95–145.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag. 24, 5,* 513–523.

Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., and Sperling, J. 2009. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM Press, New York, 42–51.

Schifanella, C., Caro, L. D., Cataldi, M., and Aufaure, M.-A. 2012. The dindex: A web environment for analyzing dependences among scientific collaborators. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12).* ACM Press, New York, 1520–1523.

SHAPIN, S. 1981. Laboratory life. The social construction of scientific facts. *Med. History 25,* 3, 341–342.

SIEG, A., MOBASHER, B., AND BURKE, R. 2007. Web search personalization with ontological user profiles. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07).* ACM, New York, 525–534.

SRIRAM, B., FUHRY, D., DEMIR, E., FERHATOSMANOGLU, H., AND DEMIRBAS, M. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR'10).* ACM Press, New York, 841–842.

SUGIYAMA, K., HATANO, K., AND YOSHIKAWA, M. 2004. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04).* ACM Press, New York, 675–684.

TAKESHI SAKAKI, M. O. AND MATSUO, Y. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).* ACM Press, New York, 851–860.

TEEVAN, J., DUMAIS, S., AND HORVITZ, E. 2005a. Beyond the commons: Investigating the value of personalizing web search. In *Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA'05).* 84–92.

TEEVAN, J., DUMAIS, S. T., AND HORVITZ, E. 2005b. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR'05).* ACM Press, New York, 449–456.

TREERATPITUK, P. AND CALLAN, J. 2006. Automatically labeling hierarchical clusters. In *Proceedings of the International Conference on Digital Government Research.* ACM Press, New York, 167–176.

WANG, C., ZHANG, M., RU, L., AND MA, S. 2008. Automatic online news topic ranking using media focus and user attention based on aging theory. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08).* ACM Press, New York, 1033–1042.

WANG, Q. AND JIN, H. 2010. Exploring online social activities for adaptive search personalization. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10).* ACM Press, New York, 999–1008.

WEDIG, S. AND MADANI, O. 2006. A large-scale analysis of query logs for assessing personalization opportunities. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06).* ACM Press, New York, 742–747.

WENG, J., LIM, E.-P., JIANG, J., AND HE, Q. 2010. Twitterrank: Finding topic sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10).* ACM Press, New York, 261–270.

WU, Y., DING, Y.,WANG, X., AND XU, J. 2010. On-line hot topic recommendation using tolerance rough set based topic clustering. *J. Comput. 5,* 4, 549–556.

XU, S., BAO, S., FEI, B., SU, Z., AND YU, Y. 2008. Exploring folksonomy for personalized search. In *Proceedings of the 31st Annual ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR'08).* ACM Press, New York, 155–162.

YANG, J. AND COUNTS, S. 2010. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10).* 355–358.

YANG, J. AND LESKOVEC, J. 2010. Modeling information diffusion in implicit networks. In *Proceedings of the 10th International Conference on Data Mining (ICDM'10).* 599–608.

ZHAO, Q., MITRA, P., AND CHEN, B. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the 22nd National Conference on Artificial Intelligence.* 1501–1506. AAAI Press.

ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A., AND LAUSEN, G. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW'05).* ACM Press, New York, 22–32.

Received February 2012; revised July 2012; accepted October 2012