Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation

Mario Cataldi
Università di Torino
Torino, Italy
cataldi@di.unito.it

Luigi Di Caro
Università di Torino
Torino, Italy
dicaro@di.unito.it

Claudio Schifanella
Università di Torino
Torino, Italy
schi@di.unito.it

ABSTRACT

Twitter is a user-generated content system that allows its users to share short text messages, called *tweets*, for a variety of purposes, including daily conversations, URLs sharing and information news. Considering its world-wide distributed network of users of any age and social condition, it represents a low level news flashes portal that, in its impressive short response time, has the principal advantage.

In this paper we recognize this primary role of Twitter and we propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. First, we extract the contents (set of terms) of the tweets and model the term life cycle according to a novel aging theory intended to mine the emerging ones. A term can be defined as *emerging* if it frequently occurs in the specified time interval and it was relatively rare in the past. Moreover, considering that the importance of a content also depends on its source, we analyze the social relationships in the network with the well-known Page Rank algorithm in order to determine the authority of the users. Finally, we leverage a navigable topic graph which connects the emerging terms with other semantically related keywords, allowing the detection of the emerging topics, under user-specified time constraints. We provide different case studies which show the validity of the proposed approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Selection process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithm

Keywords

Topic detection, Text analysis, Aging theory

MDMKDD'10, July 25th, Washington, DC, USA

1. INTRODUCTION

Twitter is a popular microblogging service that enables the users to send and read short text messages (up to 140 characters), commonly known as *tweets*. After its launch on July 2006, Twitter users have increased rapidly; on december 2009, they were estimated as 75 million worldwide with around 6.2 million new accounts per month (basically 2-3 per second), which makes Twitter one of the fastest-growing web sites in the world [1]. Moreover, in contrast with other popular social networks, most of Twitter users are adults; according to a demographic report[2], 88% of the users in United States are older than 18, defining a heterogeneous network of authors providing a very diverse set of contributions.

In this system, as information producers, people post tweets for a variety of purposes, including daily chatter, conversations, sharing information/URLs and reporting news, defining a continuous real-time status stream about every argument. Considering this aspect, one of the founders of *Twitter.com*, Evan Williams, defined the service as follow:

> *What we have to do is deliver to people the best and freshest most relevant information possible. We think of Twitter as it's not a social network, but it's an information network. It tells people what they care about as it is happening in the world.*

Following this strategy, Twitter itself recently emphasized their news and information network strategy by changing the question (Figure 1) it asks to the users for status updates from "What are you doing?" to "What's happening?".

Considering all these aspects, Twitter defines a low level infomation news flashes portal. Obviously, even if this system can not represent a serious alternative to the authoritative information media, considering the number of its authors and the impressive response time of their contributions, Twitter can provide a real-time system that can also predate the best newspapers in informing the web community about the emerging topics. In fact, the most important information media always need a certain amount of time to react to a news event; i.e. a professional journalist requires time, external collaborators and/or technology support to provide a professional report. However, a common Twitter user can easily report, as asked by the system, what is happening in front of her eyes, without any concern about

[1] http://themetricsystem.rjmetrics.com/2010/01/26/new-data-on-twitters-users-and-engagement/
[2] http://palatnikfactor.com/2010/01/29/twitter-demographic-report-who-is-really-on-twitter/

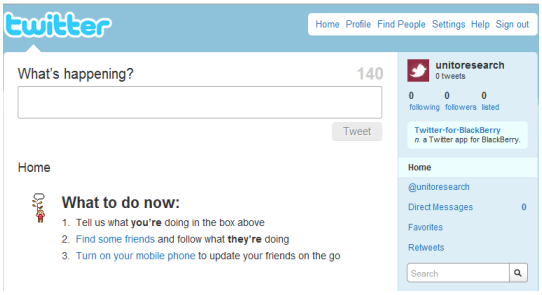

Figure 1: Current Twitter interface for updating the user's status.

her readers or her writing style. This characteristic makes Twitter probably the fastest, low level, information service in the world.

In this paper we recognize this primary information role of Twitter and provide a new method to extract the emerging topics by analyzing in real-time the emerging terms expressed by the community. The general idea is that a topic can be defined as *emergent* in a considered time interval, if it has been extensively treated within it but rarely in the past. In particular, we rely on a 5-steps process:

- we extract and formalize the user-generated content expressed by the tweets (considering all the languages) as vectors of terms with their relative frequencies;
- we define a directed graph of the active authors based on their social relationships, and calculate their authority by relying on the well-known Page Rank algorithm [25];
- for each term, we model its life cycle according to a novel aging theory that leverages the users authority in order to study its usage in a specified time interval;
- we select a set of emerging terms by ranking the keywords depending on their life status (defined by an energy value);
- we finally create a navigable *topic graph* which links the extracted emerging terms with their relative co-occurrent terms in order to obtain a set of *emerging topics.*

In our system, as in the literature ([16],[7],[23]), a topic is defined as a coherent set of semantically related terms that express a single argument. Therefore, for each considered time interval (the duration of the intervals is set by the user), the system is able to retrieve the most relevant emerging topics discussed by the community.

The paper is organized as follows: in Section 2 we present the current state of the art on content aggregation, reccomendation, trend analysis and social monitoring. We will then analyze in detail the proposed 5-steps method by formalizing our assumptions and providing real examples (Sections 3, 4, 5, 6 and 7). In conclusion, we evaluate the proposed approach with real case studies (Section 8).

2. RELATED WORK

The enormous amount of contents generated by web users in the last decade is creating 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. A first issue when dealing with large and heterogeneous data sources to be communicated to the users is the aggregation of them through filtering and merging techniques. Considering Twitter as source of text news, *TweetTabs* [2] and *Where What When* [5] simply aggregate messages and links through attractive interfaces. In general, clustering techniques help in finding groups of similar content that can be further filtered using labeling techniques like [30].

While these nice-looking websites simply aggregate messages or links, one of the most explored tasks when mining streams of text entries coming from social media, is recommendation of topics, URLs, friends, and so forth. So far, two main approaches have been studied: collaborative filtering and content-based techniques. While the first one selects and proposes the contents looking at what similar users have already selected [17], the second one analyzes the semantics of the content without considering its origin [19]. More recently, hybrid approaches have been proposed ([24, 8]). Recommendation systems can differ on what they recommend. In [13] authors propose a system to recommend URLs based on the construction of a user profile. [12] evaluates different algorithms for recommending people that share common keywords. [21] presents algorithms for recommendations of tags in folksonomy-based systems.

Another interesting task when dealing with huge and time-sensitive text contents is the analysis of their trends. This can be important for mining irregular or periodic patterns, or simply to monitor how a specific content behaves over time. Focusing on Twitter-based approaches, Trendistic [1] and Twopular [4] 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 [18]. In [33] the authors represent social text streams as multi-graphs, where each node represents a social actor and each edge represents the information flow between two actors. Events are extracted by combining clustering techniques with graphs analysis. [26, 15] present a system which uses curve analysis of frequencies for automatic segmentation of topics. [14] tracks tags over time in terms of context drifting.

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. In [9] authors 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 exponentially proportional to the ratio of documents that contain the term in the channel itself. In [32] authors 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. [29] consid-

ers Twitter as a social sensor for detecting large scale events like heartquakes, typhoons and traffic jams. The authors analyze the contexts of such keyword in order to discriminate them as positive or negative (the sentence "Someone is shaking hands with my boss" should be captured as negative eventhough it contains the term "shake"). In [6], the authors analyze Twitter messages 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. In comparison with these approaches, our technique is able to mine emerging topics under a broader view, without considering user's profile, preferences or specific events.

The authors in [11] were the first to present an aging theory based on a biological metaphor. Using this approach, the work presented in [31] is able to rank topics from online news streams through the concept of *burstiness*. The *burstiness* of a term, already introduced in [20], 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.

3. CONTENT EXTRACTION

As in most information retrieval (IR) systems, the analysis process starts with the real-time extraction of the relevant keywords (also called terms in the paper) from the stream of tweets. As explained in the introduction, we search for emerging topics on Twitter community in a given time interval: thus, given a time range r set by the user (depending on the preferred topic detection frequency), we define the t-th considered interval I^t as

$$I^t =< i_t, i_t + r >$$

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 $n = \left|TW^t\right|$ 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 relevant 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 and/or stemming; in fact, our system considers all the languages in which Twitter's users update their status. The idea is to leverage the capillary Twitter users network, using their world-wide dislocation, in order to be able to retrieve in real-time relevant news. In fact, we believe that the flow 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 protests in Iran, after the 2009 Iranian presidential election[3], had been generated in Iran itself and then, due to the global political and social importance of this event, it had been also commented by users of different countries (in this case, starting from Asia and flowing to the rest of the word). Thus, in order to be able to quickly catch a 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 unrelevant 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 5.

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 $w_{j,x}$ of the $x-th$ vocabulary term in $j-th$ tweet by using the augmented normalized term frequency [28]:

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

where $tf_{j,x}$ 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

$$\vec{tw_j} = \{w_{j,1}, w_{j,2}, ..., w_{j,v}\}$$

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

4. USER AUTHORITY

While the contents themeselves 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 espliciting 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; 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 degree of influence/importance 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 Al Gore (2,120,106 followers[4]) represents a strongly authoritative twitter user, simply because each of his words can be

[3]This event has also been nicknamed the "Twitter Revolution" because of the protesters' reliance on Twitter and other social-networking Internet sites to communicate with each other.

[4]Updated on 17th April of 2010.

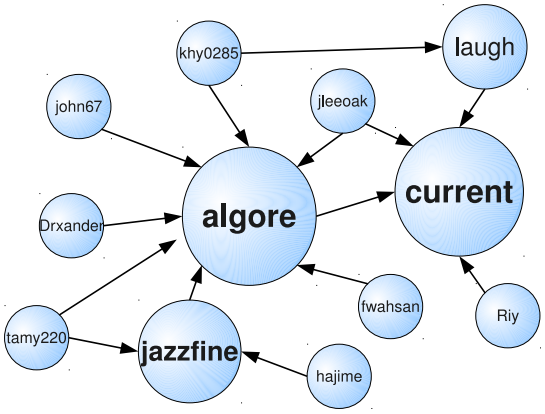

Figure 2: The authority value computation: a sample of the "Al Gore" community. The size of the nodes highlights their importance in the considered community.

instantly read by thousands of other users, and can influence their normal text stream activity. Moreover, the concept of "authority" 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 Al Gore 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, we can refer to the well known PageRank algorithm [25] as the reference approach. In fact, PageRank calculates the authority of each page by analyzing the topological graph of the considererd web entities. Considering this method, the authority of a user depends on the number and the authority of its followers. Hence, given a user $u_i \in U$, its *authority* is computed as follow:

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

where $d \in (0,1)$ is a dumping factor[5], $follower(u_i)$ is a function that returns the set of users following u_i and $following(u_j)$ returns the set of user that u_j follows. Authority values are calculated using an iterative algorithm, where, at the initial instant, each authority value is initialized to:

$$auth^0(u_i) = \frac{1}{|U|}$$

At each step, the algorithm recomputes the authority values as:

$$auth^t(u_i) = d \times \sum_{u_j \in follower(u_i)} \frac{auth^{t-1}(u_j)}{|following(u_j)|} + (1-d)$$

[5]The dumping factor d, introduced by the authors in [25], represents the probability that a "random surfer" of the graph G moves from a user to another; it is usually set to 0.85.

The process ends when a convergence condition is satisfied.

In Figure 2 an example of user authority computation on the input graph, obtained by performing a graph sampling process [22] in which the "Al Gore" vertex represents the starting point, is depicted . User authority values are visually represented by the circle sizes. In this case, "Al Gore" is the most influential user, since it has the highest number of followers. Moreover, its authority is propagated to the "current" user – the media company led by Al Gore– considering that "Al Gore" belongs to the set of its followers; this scenario confers to "current" an authority value comparable to "Al Gore", even if it has a lower number of followers.

5. CONTENT AGING THEORY

Generally speaking, an emerging keyword can be viewed as a semantic unit which links to a very recent news event. The goal of capturing such filtered knowledge relies on an accurate modeling of both the chronological sequences of tweets and the authors. Thus, we propose a content aging theory to automatically identify coherent discussions through a life cycle-based content model.

Many conventional clustering and classification strategies can not 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, similarly to existing approaches [11], we propose a metaphor where each term is seen as a living organism; in contrast to the approach proposed in [11], we analyze the terms life cycles by distinguishing among different time intervals in order to highlight when a term becomes important in the community. Thus, we rely on the same methapor as [11] but we significantly change its meaning in order to stress the temporal definition of each considered term.

The life cycle of a keyword can be considered as analogous to the one of a living being: with abundant nourishment (i.e., related tweets), its life cycle 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 *energy*, which indicates the *vitality* status of the keyword and can qualify the keyword's usage. In fact, a high energy value implies that the term is becoming important in the considered community, while a low energy value implies that it is currently becoming out of favor.

In this section we present a statistical analysis of the life cycle of the contents to quantitatively and qualitatively measure the usage of each term into the Twitter community.

5.1 Content Nutrition

Considering this biological metaphor, the contribuition in terms of nutrition of each nurishment changes depending on its chemical composition; for example, each food brings a different calory contribution depending on its ingredients. Therefore, in our case, we use the concept of authority introduced in Section 4 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

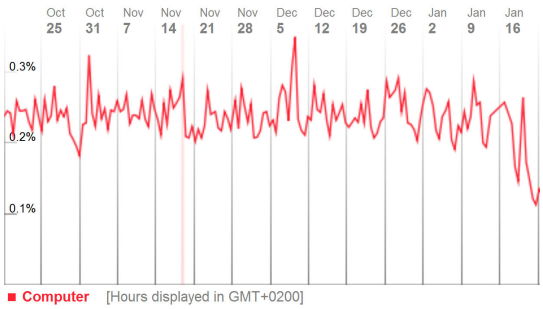

Figure 3: Statistical usage of the term "computer" (provided by Trendistic [1]) in Twitter from October 2009 to January 2010.

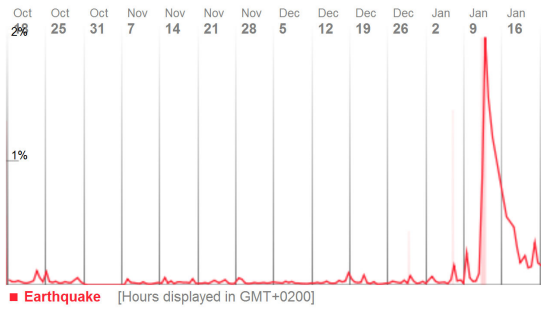

Figure 4: Statistical usage of the term "earthquake" (provided by Trendistic [1]) in Twitter from October 2009 to January 2010; the pick represents the catastrophic earthquake occurred in Haiti on 12 January 2010.

define the amount of nutrition as

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

where $w_{k,j}$ represents the weight of the term k in the tweet vector $\vec{tw_j}$ (thus, $tw_j[k]$), the function $user(tw_j)$ returns the author u of the tweet tw_j and $auth(u)$ returns the authority value associated to u.

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 and also the authority of each single user that reports k. This way, the system quantifies the usage of each term by evaluating its frequency and qualifies its relevance by analyzing the influence of the authors into the Twitter community.

5.2 Content Energy

Once obtained the nutrition of a semantic unit (i.e. a term), we need to map it into a value of *energy*. The energy value of a term indicates its effective 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.

In detail, having for each keyword k its amount of nutrition $nutr_k^t$ in a time interval I^t, it is possibile to rank the hottest terms only considering their related nutrition value.

DEFINITION 5.1. *A term can be defined as* ***hot*** *if its usage is extensive within the considered time interval.*

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's usage to analyze this property. Therefore,

DEFINITION 5.2. *We define a keyword as* ***emergent*** *if it results to be* hot *in the considered time interval but not in the previous ones.*

Namely, we analyze the current nourishment in comparison to the ones build up in the previous time intervals.

Let consider the examples shown in Figure 3 and 4; considering the time frame from October 2009 to January 2010, the term "computer" is relatively more used by the twitter community than the term "earthquake". Thus, according to our definition, the term "computer" can be considered *hotter* than "earthquake". On the other hand, considering the specific time instant defined by the day 13th January 2010, due to their global usages, the term "earthquake" can be easily seen as *emerging* keyword, while there is no significant change in the usage of term "computer".

Obviously, it is not necessary to consider the complete usage history of each keyword: in fact, if for example the keyword "Iran" has been extensively used in 2009, it does not mean that it can not become again important in the community for another related fact in 2010. 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 can not 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 keyword 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 life cycles, and defines the *history worthiness* of the resulting emerging keywords.

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

$$energy_k^t = \sum_{x=t-s}^{t} \left(((nutr_k^t)^2 - (nutr_k^x)^2) \cdot \frac{1}{t-x} \right)$$

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 key-

word 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.

6. SELECTION OF EMERGING TERMS

In the previous section we presented our approach to model the aging of contents for an automatic extraction of emerging terms. In this section we propose a supervised and an unsupervised technique to select a limited set of relevant terms that emerge in the considered time interval.

6.1 Supervised 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 energy 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 dislocated users reported it on a big scale, as for Haiti hearthquake), she probably prefers to avoid such contents that are only relatively emerging (for example, a term 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

$$drop^t = \delta \cdot \frac{\sum_{k \in K^t}(energy_k^t)}{|K^t|}$$

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

$$\forall k \in K^t, k \in EK^t \iff energy_k^t > drop^t$$

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

6.2 Unsupervised Selection

The second approach considers a completely automatic model that does not involve any user interaction. This unsupervised approach is based on the idea that it could be very hard for a user 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 desidered cardinality of emerging keywords. Moreover, depending on the temporal context, it could be necessary to set differently the threshold value. Thus, we leverage an unsupervised ranking model that dynamically sets the *critical drop* as in [10]. This cut-off is adaptively computed as follows. It:

1. first ranks the keywords in descending order of energy value previously calculated in Section 5.2.
2. computes the *maximum drop* in match and identifies the corresponding drop point.
3. 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-steps process, the keywords ranked better than the point of critical drop are defined as emerging keywords on time interval I^t and are collected in EK^t.

7. 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 examinating 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", "Usa", "Obama" and "McCain", while in a more recent time frame, as for example February 2010, the term could be related to a sports event by other keywords as "superbowl"–due to the championship game of the National Football League (NFL)– "football" or "New Orleans Saints" –the name of the team whose won the final game in 2010–.

7.1 Correlation Vector

Thus, in order to express the topics related to the retrieved emerging keywords, we analyze the time frame in which all the tweets have been generated, and analyze the semantic relantionships 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 weigthed 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 negative evidence against the relationship. In detail, we treat each keyword k as a query and the set of the tweets containg 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 [27]:

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

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*

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

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

7.2 Topic Graph

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

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

Therefore $\rho_{k,z}$ represents the relative weight of the keyword k in the corresponding vector $\vec{cv}_k^t$, i.e. 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 (similarly to Section 6.2) 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).

7.3 Topic Detection 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 6, 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.

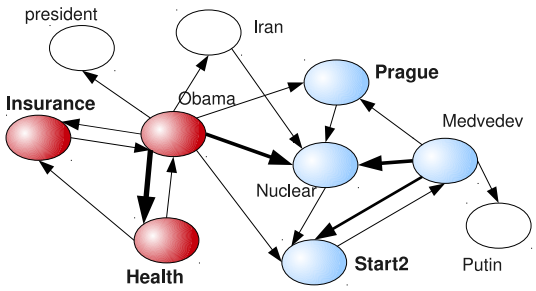

Figure 5: A Topic graph with two Strongly Connected Components (in red and yellow) representing two different emerging topics: labels in bold represent emerging keywords while the thickness of an edge represents the semantical relationship between the considered keywords.

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 $ET^t = \{ET_1^t, ..., ET_n^t\}$ 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.

In Figure 5 is depicted 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 "health", "start2" and other popular keywords, like "Obama", "Medvedev" and "nuclear", that are constantly used by twitter users and do not represent statistically emerging terms. In fact, terms like "Obama" or "Medvedev" represent very popular terms always reported by the users in Twitter as in any other information sources.

Notice that, with this approach, not only 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-occurr 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 (but even in this case, we believe that she will probably share some common terms).

Thus, as 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

$$rank_{ET_z^t} = \frac{\sum_{k \in K_z^t}(energy_k^t)}{|K_z^t|}$$

that leverages the average energy 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.

7.4 Topic Label

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 stricly 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 decribed in Section 6.2) that permits to select the most representive 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, 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[6].

8. EXPERIMENTS AND EVALUATIONS

In this section, we evaluate the proposed method by analyzing real case studies. In particular, we conducted several experiments by monitoring the twitter community during the period included between 13th and 28th of April 2010.

Considering the enormous amount of Twitter users and the impressive number of generated tweets, in our experiments we have monitored a stream which consists of random samples among all public messages. This access level provides a statistically significant input for data mining and research applications [3]; in fact, considering this approach, we have evaluated more than 3 millions of tweets (an average of 10k tweets per hour), which included more than 300k different keywords.

Considering the topic detection method introduced in this paper, we analyze different experiments: we initially analyze a real case study by evaluating the emerging topics retrieved by the system within the considered time interval. Then, we vary the number of considered time slots (Section 5.2) in order to study how this parameter affects the quality of the retrieved topics. Finally, we study the difference between the supervised and unsupervised seletions methods (Section 6), evaluating their impact on the resulting emerging topics.

8.1 Real case study

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

Date	Emerging Topics
15-04-2010	{**eyjafjallajökull**, volcano, airports, iceland, close}[7]
18-04-2010	{**kaczynski**, president, funeral, volcano}[8]
20-04-2010	{**activist**,dorothy, height, death}[9]
20-04-2010	{**rockies**, president, team, dead}[10]
21-04-2010	{**samaranch**, president, barcelona, honor, died}[11]

Table 1: The emerging topics retrieved by the system based on the five most emerging terms (in the time period included between 13th and 28th April 2010). The labels in bold represents emerging terms. In the footnotes we link the related news reports provided by professional information sources.

As previously reported, we initially evaluate the proposed approach by analyzing the retrieved emerging topics within the considered 15 days. For this experiment, we consider the unsupervised selection method and we set the preferred time range r as 15 minutes (Section 3). As explained in the introduction, considering the impressive response time of the users, if continuously monitored with small time ranges, Twitter can also predate the most popular news sources in informing the community about emerging news events. However, it is possible to set higher time range values: in this case, the resulting topics will be statiscally 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 200 (with the selected time range, it means that we basically consider 2 solar days); in fact this time slot size is sufficient to avoid such terms that are significant in a daily perspective.

In Table 1 we show the retrieved topics based on the five most emerging terms (i.e., the terms with the highest nutrition 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. "eyjafjallajökull", the name of a volcano in Iceland); in fact, they represent keywords that are generally very unusual in the community and only emerge in correspondence to unexpected events. Although, by analyzing the co-occurrences information in the tweets reporting such terms (as explained in Section 7.2), it is possibile to link them to other popular keywords (as "airports", in the topic leaded by "eyjafjallajökull") that have very common usages in the community.

8.2 History Worthiness

As reported in Section 5.2, an *emerging* term is defined by taking into account its usage in a limited number of previous time slots. However, depending on the selected number of considered time intervals, the retrieved topics can significantly differ. Thus, in order to study the impact of this parameter on the retrieved topics, we analyze two differ-

[7]http://www.guardian.co.uk/world/blog/2010/apr/15/volcano-airport-disruption-iceland
[8]http://news.bbc.co.uk/2/hi/8627857.stm
[9]http://news.yahoo.com/s/ap/20100420/ap_on_re_us/us_obit_height
[10]http://www.reuters.com/article/idUSTRE63J53B20100420
[11]http://www.nytimes.com/2010/04/22/sports/22samaranch.html

Date	Emerging Topics (s=100)
20-04-2010	{**morning**,early, tuesday, sleep}
20-04-2010	{**music**, album ,video, stereo}
20-04-2010	{**laundry**, citrus ,urgent, liquid}
20-04-2010	{**profile**, facebook ,post, link}

(a)

Date	Emerging Topics (s=200)
20-04-2010	{**activist**,dorothy, height, death}
20-04-2010	{**rockies**, president, team, dead}

(b)

Table 2: The emerging topics retrieved at 20th April 2010 by considering (a) a daily history worthiness ($s = 100$) and (b) a 2-days history worthiness. The labels in bold represents emerging terms.

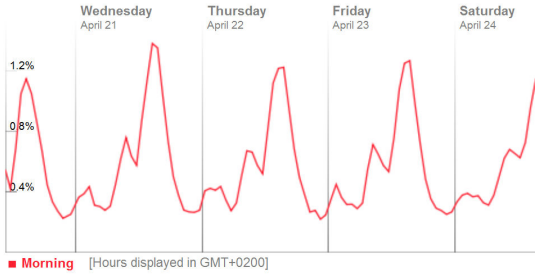

Figure 6: Statistical usage in Twitter of the term "morning" (provided by Trendistic [1]) in consecutive days.

ent number of considered slots, $s = 100$ and $s = 200$ (by considering again a time range r of 15 minutes and the unsupervised selection method).

In Table 2(a) and (b) we present the results at the 20th of April 2010: the most emerging topics obtained by setting $s = 100$ represent common daily activities by 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 "morning" have very standard usages in the community due to the fact that they represent periodic events. We guess that the users systematically use such terms in correspondence to their natural occurrence. For example, in Figure 6 we show the usage of the term "morning" in consecutive days; the system can report a pick if it only takes into account a 24-hours history; however, if it considers a relatively higher time frame, it recognizes a constant usage in time (with picks in the morning and lower usage during the afternoon and the evening) that can lower the energy value of this keyword. Thus, the life status of a keyword strictly depends on the considered number of time intervals (Section 5.2) and this value directly affects the temporal relevance of the retrieved topics.

8.3 Supervised vs Unsupervised Selection

In Section 6 we reported two different selection methods (unsupervised and supervised) to indentify emerging terms. In this last experiment we evaluate the impact of each of them in the retrieved topics by analyzing the example proposed in Table 1: each of the considered five emerging terms ("eyjafjallajökull", "kaczynski", "activist", "rockies" and "samaranch") was indentified as *emergent* in different time intervals.

Date	Emerging Terms	Energy value ($total\ avg = 5.6517$)
15-04-2010	**eyjafjallajökull**	7773.7575
15-04-2010	**bieber**	147.1661
15-04-2010	**wellington**	115.3432
15-04-2010	**betezy**	76.7339
15-04-2010	**diaper**	55.3219

Table 3: The five most emerging terms (and their related energy values) at 3:15pm (GMT+02) on the 15th of April 2010. The total average energy value was 5.6517 (among all the keywords typed within the considered time interval).

In order to understand how they have been considered as emerging, we need to analyze their related time intervals and their associated energy values. Let consider the time interval related to the term "eyjafjallajökull": in Table 3, we show the five most emerging terms and their energy values retrieved by the system at 3:15pm (GMT+02) on the 15th of April 2010. We notice that, considering the unsupervised approach (Section 6.2), only the term "eyjafjallajökull" is considered as emergent. In fact, the system identifies the difference in terms of energy values between "eyjafjallajökull" and the second most emerging terms ("bieber" – a popular young singer –) as the critical drop and only considers as emerging such keywords that are ranked better than the critical drop (thus, in this case, only "eyjafjallajökull").

However, considering the supervised selection method (Section 6.1), 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 energy value is 1000 times higher than the total average energy value), only the term "eyjafjallajökull" is selected as emergent. Instead, with $\delta = 100$, also the terms "bieber", "wellington" and "betezy" –an australian bookmaker website– will be considered as emergent (and analyzed using the topic graph to retrieve the related topics).

9. CONCLUSIONS

In this paper we presented a novel approach to detect in real-time emerging topics on Twitter. We formalized the keyword life cycle leveraging a novel aging theory intended to mine terms that frequently occur in the specified time interval and they are relatively rare in the past. We also studied the social relationships in the user network in order to quantify the importance of each analyzed content. Finally, we formalized a keyword-based topic graph which connects the emerging terms with their co-occurrent ones, allowing the detection of emerging topics under user-specified time constraints. In conclusion we provided case studies that show the effectiveness of the proposed approach and explicit the usages of each introduced parameters.

10. REFERENCES

[1] Trendistic. http://trendistic.com/.
[2] Tweet tabs. http://tweettabs.com/.
[3] Twitter API. http://apiwiki.twitter.com/.
[4] Twopular. http://twopular.com/.
[5] Where-what-when. http://where-what-when.husk.org/.
[6] S. Abrol and L. Khan. Twinner: understanding news queries with geo-content using twitter. In *GIR '10:*

Proceedings of the 6th Workshop on Geographic Information Retrieval, pages 1–8, New York, NY, USA, 2010. ACM.

[7] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[8] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40:66–72, 1997.

[9] K. K. Bun, M. Ishizuka, and B. M. Ishizuka. Topic extraction from news archive using tf*pdf algorithm. In *Proceedings of 3rd Int'l Conference on Web Informtion Systems Engineering (WISE 2002) ,IEEE Computer Soc*, pages 73–82. WISE, 2002.

[10] M. Cataldi, C. Schifanella, K. S. Candan, M. L. Sapino, and L. D. Caro. Cosena: a context-based search and navigation system. In *MEDES*, pages 218–225. ACM, 2009.

[11] C. C. Chen, Y.-T. Chen, Y. S. Sun, and M. C. Chen. Life cycle modeling of news events using aging theory. In *ECML*, pages 47–59, 2003.

[12] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 201–210, New York, NY, USA, 2009. ACM.

[13] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information. Atlanta, USA, 2009. ACM Press.

[14] L. Di Caro, K. S. Candan, and M. L. Sapino. Using tagflake for condensing navigable tag hierarchies from tag clouds. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1069–1072, New York, NY, USA, 2008. ACM.

[15] A. Favenza, M. Cataldi, M. L. Sapino, and A. Messina. Topic development based refinement of audio-segmented television news. In *NLDB '08*, pages 226–232, Berlin, Heidelberg, 2008. Springer-Verlag.

[16] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu. Time-dependent event hierarchy construction. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 300–309, New York, NY, USA, 2007. ACM.

[17] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[19] A. Hassan, D. Radev, J. Cho, and A. Joshi. Content based recommendation and summarization in the blogosphere. *International AAAI Conference on Weblogs and Social Media*, 2009.

[20] Q. He, K. Chang, and E.-P. Lim. Using burstiness to improve clustering of topics in news streams. *Data Mining, IEEE International Conference on*, 0:493–498, 2007.

[21] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *PKDD 2007*, pages 506–514, Berlin, Heidelberg, 2007. Springer-Verlag.

[22] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM.

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

[24] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering. In *In Proceedings of the 2001 SIGIR Workshop on Recommender Systems*, 2001.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[26] Y. Qi and K. S. Candan. Cuts: Curvature-based development pattern analysis and segmentation for blogs and other text streams. In *HYPERTEXT '06*, pages 1–10, New York, NY, USA, 2006. ACM.

[27] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003.

[28] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[29] M. O. Takeshi Sakaki and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW 2010*, 2010.

[30] P. Treeratpituk and J. Callan. Automatically labeling hierarchical clusters. In *dg.o '06: Proceedings of the 2006 international conference on Digital government research*, pages 167–176, New York, NY, USA, 2006. ACM.

[31] C. Wang, M. Zhang, L. Ru, and S. Ma. Automatic online news topic ranking using media focus and user attention based on aging theory. In *CIKM '08*, pages 1033–1042, New York, NY, USA, 2008. ACM.

[32] Y. Wu, Y. Ding, X. Wang, and J. Xu. On-line hot topic recommendation using tolerance rough set based topic clustering. *Journal of Computers*, 5(4), 2010.

[33] Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, pages 1501–1506. AAAI Press, 2007.