Temporal Semantic Centrality for the Analysis of Communication Networks

Damien Leprovost, Lylia Abrouk, Nadine Cullot, David Gross-Amblard

Temporal Semantic Centrality for the Analysis of Communication Networks

Damien Leprovost[1] Lylia Abrouk[1] Nadine Cullot[1]
David Gross-Amblard[2]

[1] Le2i CNRS Lab, University of Bourgogne, Dijon, France
{firstname.lastname}@u-bourgogne.fr

[2] Irisa CNRS Lab, University of Rennes 1, France
{firstname.last_name}@irisa.fr

Abstract

Understanding communication structures in huge and versatile online communities becomes a major issue. In this paper we propose a new metric, the *Semantic Propagation Probability*, that characterizes the user's ability to propagate a concept to other users, in a rapid and focused way. The message semantics is analyzed according to a given ontology. We use this metric to obtain the *Temporal Semantic Centrality* of a user in the community. We propose and evaluate an efficient implementation of this metric, using real-life ontologies and data sets.

Keywords : semantic analysis, centrality, community, communication network, ontology

Résumé

De nos jours, la compréhension des communautés en ligne devient un enjeu majeur du Web. Dans cet article nous proposons une nouvelle mesure, la Probabilité de Propagation Sémantique (*SPP*), qui caractérise la capacité de l'utilisateur à propager un concept sémantique à d'autres utilisateurs, d'une manière rapide et ciblée. La sémantique des messages est analysée selon une ontologie donnée. Nous utilisons cette mesure pour obtenir la Centralité Sémantique Temporelle (*TSC*) d'un utilisateur dans une communauté. Nous proposons et évaluons une expérimentation de cette mesure, en utilisant une ontologie et des données réelles issues du Web.

Mots-clefs : analyse sémantique, centralité, communauté, réseau de communication, ontologie

1 Introduction

With the advent of the collaborative Web, each website can become a place for expression, where users' opinions are exchanged and points of view are discussed. User messages are valuable for the site owner: in addition to a proof of interest for the website or its products, they allow the owner to understand users'judgments and expectations. However, if this reasoning is humanly manageable on a small number of messages, it is reckless for larger systems, handling thousands of users posting thousands of messages per month.

Nowadays, users and community profiling is a growing challenge [2]. Many approaches have been developed in the domain of online community analysis. Initial methods relied on a basic relationship between users like friendship in social networks or answers or citations in communication networks (like forums or emails). For communication networks, the semantics of the message itself is progressively taken into account [7]. In parallel, recent works [25] incorporate the temporal dynamic of messages, but without their semantics.

In this paper we consider as a communication network any system where users are able to exchange messages, such as forums, tweets, mailboxes, etc. In this context, we first propose a method for the identification of hot topics and thematic communities. These topics are identified within user messages using a target *ontology*, which can be generic or specialized for a given domain.

We then present a method for the discovery of central users who play an important role in the communication flow of each community. For this purpose we introduce new semantic measures called the *Semantic Propagation Probability* (SPP) and *Temporal Semantic Centrality* (TSC) that take into account both semantics and communication timestamps *at once*.

A potential limitation of using ontology is to limit a priori the set of topics of interest, what may prevent the discovery on new topics. But the main advantages is to focus the analysis on a known domain that can be extended at will, but in a controlled way. A basic example is to understand the behavior of a forum according to brand product ontology. Another advantage is to rely on the permanently increasing set of generic or specialized ontologies that are linked to other resources or services.

The paper is organized as follows. We present hot topics and community identification in Section 2 and our metric in Section 3. We show our exper-

iments in Section 4. Section 5 discusses the obtained results and Section 6 covers related approaches. Finally, Section 7 concludes.

2 Communication Networks, Thematic Communities

2.1 Overview

We reason according to an ontology $O = (C, is - a)$, where S is a set of concepts and $is-a$ is the subsumption relation. We equip C with a semantic similarity measure $d_C(c, c')$ between two concepts c and c' of C. Let δ be a similarity threshold. We say that two concepts are similar if their distance d_C is smaller than δ (the choice of d_C and δ will be discussed in depth in Section 4).

We consider a communication network $G = (U, S)$, where U is a set of users and $S \subseteq U \times U \times \mathbb{N}$ is the timed directed *send* relation of a message $m = (u, v, t)$ from user u to user v at time t. We take $\mathbb{N}$ as a clock for the sake of simplicity. Perfectly simultaneous messages are possible in this model, and their occurrence is taken into account[1]. This simple model assumes that the originator and receptor of a given message are known. While realistic for email-based communication networks, its applicability to forums where posts are submitted in a communication flow, will be discussed later on. The *content* function maps a message $m = (u, v, t)$ to its plain textual content $content(m)$. In order to focus on concepts in C, the $content_C$ function maps m to the set of concepts of C which appear in $content(m)$. This function encompasses details like stemming.

The aim of this approach is to identify central users acting on major topics of the communication network. Fist, we start by considering hot topics of this network. Then, we identify thematic communities built around them, and last we apply the proposed semantic centrality method to identify the central users of these communities. Figure 1 gives a global view of the method. We analyze on-line forums using a crawler and specific wrappers, then extract concepts from user posts according to a predefined ontology. These concepts are used to summarize user profiles and to identify communities. The target ontology contains concepts, which are the considered topics for the communication network. Users profiles can be identified according to the concepts

[1]By the way, due to the huge traffic of e.g. tweets per seconds, a lot of messages are likely to be simultaneous, whatever the chosen time precision.

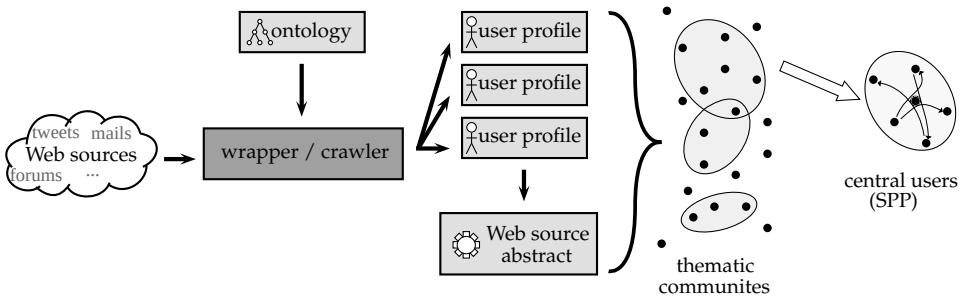

Figure 1: Overall view of the method

of the ontology. Hot topics are the concepts the most present in the users profiles. Then thematic communities are discovered considering these users profiles. Last, the SPP-Central users are detected using the proposed method described in Section 3.

2.2 Identifying hot concepts

2.2.1 Overview

The first ste of our method is to determine the hot topics of the communication network. In this method, hot topics are viewed as a subset of concepts of the target ontology O [15]. We associate with each user a *semantic profile*, that contains the number of occurrences of each ontology concept in the user's posts. This way, hot topics will be the top-n concepts most present in users' profiles.

2.2.2 User profiling

For each user u, according to a concept c on the ontology, the pre-profile of this user u relatively to this concept c, noted $preprofile_u(c)$, could be defined, for a first definition, as the total number of occurrences of c in the user's posts. But in addition, we consider the question/answer context. When a user replies to another, he indeed places his message in the context of the original message. We use the *cite* relationship to enrich the user pre-profile, we then define $preprofile_u(c)$ as the sum of all occurrences of concept c in the posts of user u and its cited posts:

$$preprofile_u(c) = \sum_{p \in post(u)} \big(occurence_c(p) + \sum_{p' : p \in cite(p')} occurrence_c(p')\big).$$

User profiles can be enriched incrementally over the future contributions. We abstract all these information in the user profile $profile_u$, as the current summarization of user u's semantic interests, applying two distinct operations:

- adding well-covered concepts, by generalization,
- deleting nonrelevant concepts.

The first generalization step allows for highlighting the cover of a concept by a user who manipulates its subconcepts. For inner nodes, we consider that a user u who manipulates a *significant part* of the direct child concepts $c_1, \ldots, c_k$ of a concept c, also manipulates c. We define the *coverage* $cov_u(c)$ of a concept c:

$$cov_u(c) = \frac{|\{c_i : is-a(c_i, c) \text{ and } profile_u(c_i) > 0\}|}{|\{c_i : is-a(c_i, c)\}|}.$$

The coverage threshold is materialized by $\delta_{coverage} \in [0, 1]$. Then, if the coverage is good, the profile of a user for this concept is incremented by the average of the subconcepts:

$$profile_u(c) = \begin{cases} preprofile_u(c) + \frac{\sum_{c':is-a(c',c)} profile_u(c')}{|c':is-a(c',c)|} & \text{if } cov_u(c) \geq \delta_{coverage}, \\ preprofile_u(c) & \text{otherwise.} \end{cases}$$

The second step deletes concepts from the profile when their weight is below a minimum weight. This minimum weight is relative to the sum of user's contribution weights, and defined by the threshold $\delta_{relevance}$. That is, a concept c is deleted if

$$\frac{profile_u(c)}{\sum_{c' \in C} profile_u(c')} < \delta_{relevance}.$$

This way, if a user covers a significant part of the sub-concepts of a parent concept, the parent concept score is increased (even if this concept is never used explicitly). The proposed method is also *contextual*, as it takes into account the question/answer structure of the forum and post citations. Finally, it is an *incremental* method: profiles are updated while new posts are emitted.

2.2.3 Hot topics

At the communication network level, we aggregate all the user profiles to build a system profile. Hot concepts are the top-n concepts which are most present in users' profiles. A full description of the profile construction of the system is available in our previous work [15].

2.3 Building thematic communities

Once hot concepts are well identified, our goal is to divide the communication network G into k thematic communities $G_1 \ldots, G_k$, each G_i being labeled with one set of concepts $L_i \subseteq C$. We will filter users according to their semantic profiles. These profiles already encompass semantic deduction through the addition of the well-covered concepts as described previously. In order to control the number of thematic communities, we allow users to be gathered according to their common and similar concepts. The similarity of two concepts of the target ontology O is measured using a semantic distance. Various definition of semantic distances have been proposed so far (e.g. [13, 10]). We rely here on the Wu-Palmer distance [26] restricted to concepts *hierarchies* (trees), which has already been applied to similar cases [4]. The similarity is defined with respect to the distance between two concepts in the hierarchy, and also by their position relative to the root. The semantic similarity between concepts c_1 and c_2 is

$$sim_{Wu\&Palmer}(c_1, c_2) = \frac{2 * depth(c)}{depth(c_1) + depth(c_2)} \ ,$$

where c is the nearest top edge of c_1 and c_2 and $depth(x)$ the number of edges between x and the root.

As stated in the beginning of this section, two concepts c_1 and c_2 will be considered as similar if $d_C(c_1, c_2) \leq \delta$, where δ is the similarity threshold.

$$d_C(c_1, c_2) = 1 - sim_{Wu\&Palmer}(c_1, c_2).$$

We then turn to thematic communities. Let $N_i^+(G_i)$ be the in-degree of community G_i, that the number of posts from members of G_i to members of G_i which contain concepts (similar to) a concept in L_i. Conversely, let $N_i^-(G_i)$ be its out-degree, that is the number of posts from members of G_i to members outside G_i which contain concepts (similar to) a concept in L_i. We can now define a thematic community:

Definition 1. *A set $G_i \subseteq G$ is a thematic community on concepts $L_i \subseteq C$, if, when restricting G_i to posts that contain a concept (similar to) a concept in L_i, the in-degree of G_i is greater than its out-degree (thus, $N_i^+(G_i) > N_i^-(G_i)$).*

Traditional approaches by Flake et al. [6] and various optimizations [11, 12, 22, 5] allow us to effectively group users linked by a binary relation in communities. We take a leaf out of them to define a cutting method, given the resulting simplification of the Definition 1.

For each community G_i, we maintain for each user u , two sets of messages $N_i^+(u)$ and $N_i^-(u)$, representing respectively communications inside G_i and communications outside G_i, with concepts similar to L_i. A message m_k is considered by default in $N_i^-(u)$. Each message m_k to user u is considered initially as unhandled. So, we add the message to $N_i^-(u)$. After that, if one or more message m_l is emitted from u, with $d(m_l, m_k) \leq \delta$.

At any time, communities are $G_i = (U_i, S_i)$, where

$$U_i = u \in U, N_i^+(u) \leq N_i^-(u)$$

and

$$S_i \subseteq U_i \times U \times \mathbb{N} \ .$$

Algorithm 1 and 2 presents this community clustering.

Algorithm 1 Message

Require: message m, concepts $L_1, \ldots, L_i, \ldots, L_k$, δ

for all $c \in L_i, c \in context(m)$ **do**
  if m is incoming **then**
    $N_i^-(u) = N_i^-(u) \cup m$
  else
    for all m_λ to u with $d(m, m_\lambda) \leq \delta$ **do**
      $N_i^+(u) = N_i^+(u) \cup m \cup m_\lambda$
      $N_i^-(u) = N_i^-(u) - m$
    end for
  end if
end for

Algorithm 2 Communities

Require: $G = (U, S)$, $L_1, \ldots, L_i, \ldots, L_k$

```
for all G_i do
    for all u ∈ U do
        if N_i^+(u) ≤ N_i^-(u) then
            U_i = U_i ∪ u
        end if
    end for
end for
```

3 Temporal Semantic Centrality

3.1 Dispersion and Lag

Inside a thematic community labeled by concepts L_i, all users are known to discuss frequently about topics of L_i or similar topics. We would like to rank these users according to their centrality, i.e. to identify the most important information participants inside the community. In this proposal, we base our ranking on *both semantics and time*. We define a *temporal semantic centrality*, using a concept-driven measure, the *semantic propagation probability*, denoted as SPP in the sequel. Globally speaking, this measure aims at capturing:

- how focused are the answers of a user according to an input post,
- how fast are these answers, relatively to the general pace of the community.

Users with a high SPP are more likely to answer or relay messages, semantically relevant to the community.

Let us consider an oriented communication

$$u \rightarrow_t u' \rightarrow_{t'} u''$$

which means that there exists in the communication graph G a message $m = (u, u', t)$ from u to u' at time t, and a messages $m' = (u', u'', t')$ from u' to u'' at time t'. For $t' > t$, m' can be seen as a relay of m in a very broad sense. Globally speaking, user u' is impacted (in various ways) by the reception of m before sending m'. Also, the content of m' can be related to

m or completely independent from it. We will measure this relation so that it depends on the *semantic dispersion* of the sent message, and its *lag*.

Noted $dispersion_c(m)$, the *dispersion* of a message m according to concept c is the ratio between the minimum semantic distance between c and concepts in m, and the maximum semantic distance between c and the concepts of the target ontology:

$$dispersion_c(m) = \frac{\min_{c' \in content(m)} d_C(c, c')}{\max_{c' \in C} d_C(c, c')}.$$

If the message uses concept c ($c \in content(m)$) then $dispersion_c(m) = 0$. Observe also that the dispersion is at most 1. For the special case where the message has no relevant concept (when $content(m)$ is empty), we consider that $dispersion_c(m) = 1$.

Similarly, we define the *lag* between a message received by u_i at time t_{i-1} and a message sent by u_i at time t_i as the duration between them, *relatively to the natural pace of the community*. Indeed, some news-focused or work-oriented communities suppose a rapid pace from its users (say hours, minutes, at most 2 days), while some technical communities may consider a month a natural duration for a specific topic.

The $meanpace_{L_i}$ of a community labeled by L_i is the average of the duration of message transmission between users of the community labeled by L_i:

$$meanpace_{L_i} = avg_{m=(u,u',t),m'=(u',u'',t') \text{ with } u,u',u'' \in G_i, t'>t}(t' - t).$$

The *lag* between two message $m = (v, u, t)$ and $m' = (u, v', t')$, relative to the mean pace $meanpace_{L_j}$ of community G_j labeled by concepts L_j is defined by:

$$lag(m, m') = \begin{cases} \infty \text{ if } t' \leq t, \\ \frac{t'-t}{meanpace_{L_j}} \text{ otherwise.} \end{cases}$$

Note that the infinite lag is used to enforce communication chains with an increasing timestamp and to discard simultaneous messages ($t = t'$).

3.2 Semantic Propagation Probability and Temporal Semantic Centrality

We can now turn to the definition of the *Semantic Propagation Probability (SPP)*. The SPP of user u according to messages m and m' is defined by:

$$SPP_c(u, m, m') = \frac{(1 - dispersion_c(m) \times dispersion_c(m'))}{1 + lag(t, t')}.$$

For example, a user receiving a message talking about c and sending a message about c immediately after (that is $t' \approx t$ in our discretized model), has a SPP_c arbitrary close to 1.

Finally, the temporal semantic centrality $TSC_{L_i}(u)$ of user u within the community labeled by L_i is computed on all incoming and sent messages of u:

$$TSC_{L_i}(u) = avg_{c \in L_i}\Big(\sum_{m=(u,u',t) \in G} \sum_{m'=(u',u'',t') \in G, t' > t} SPP_c(u, m, m')\Big).$$

For a given concept, we sum the SPP_c of u to promote users with numerous good communications. In the sum definition, we take into account all the future messages m' after m and do not restrict our attention to the next one. Indeed, a user will not necessarily answer or forward a message immediately, but will probably interwine answers to several messages. For the overall thematic set L_i, we take the average of the SPP_c, in order to favor users that cover concepts in L_i well.

3.3 Approximation for efficiency

In our implementation of SPP_c, the semantic distance is computed in two phases. An initial phase, done once per ontology, builds an index matching each concept to its ancestor and depth in the ontology. In the second phase, for a new message with at most k distinct concepts, the computation of its dispersion according to concept c requires k queries to the index. The overall computation time is then $O(k.M)$, where M is the total number of hot concepts.

Computing the TSC naively is a time consuming operation, as :

1. the ontology may be extremely large,

2. all incoming messages have to be matched with all potential outcoming messages.

For the first difficulty, we focus on the identified hot concepts, and compute the set of concepts in the relevant neighborhood of at least one of them (that is, with a semantic distance smaller than the prescribed relevance threshold). This drastically reduces the set of concepts to consider when a new message has to be checked. If a new hot concept is identified, we update this relevant set accordingly.

For the second difficulty, it should be observed that a message can impact the TSC only during a short time window, due to the lag function. Outside this window, the TSC contribution is close to zero. This suggests a sliding-window algorithm, where only a finite number of messages is kept in main memory. Outcoming messages are then compared to messages in this window, as depicted in Algorithm 3. We now illustrate our model with these improvements in our experimental section.

Algorithm 3

Require: $G = (U, S)$, message m, lag-relevance threshold δ

for all new message m **do**
  for all $u \in recipient(m)$ **do**
    add m to $INBOX(u)$
    delete from $INBOX(u)$ messages m' with $lag(m', m) \leq \delta$
  end for
  $s = sender(m)$
  for all $m' \in INBOX(s)$ **do**
    compute each $SPP_c(s, m, m')$
    update $TSC(s)$
  end for
end for

4 Experiments

4.1 Data sets

We have taken as a data source the Enron Email data set[2] for its complete communication network with a send relation and precise timestamps. This data set consists in emails collected from about 150 users, mostly senior management of Enron, made public by US federal authorities during its investigation on Enron scandal. The set contains a total of about 500'000 messages.

We have performed an initial cleaning over the set, in order to delete messages with an incorrect timestamp. If 99,87 % of the email set was stamped from 1997 to 2002 (date of the federal investigation), the entire set contains mails stamped from 1970 to 2044 that we do not take into account[3]. The final size of our set is 494'910 mails. Figure 2 shows time dispersion of the set.

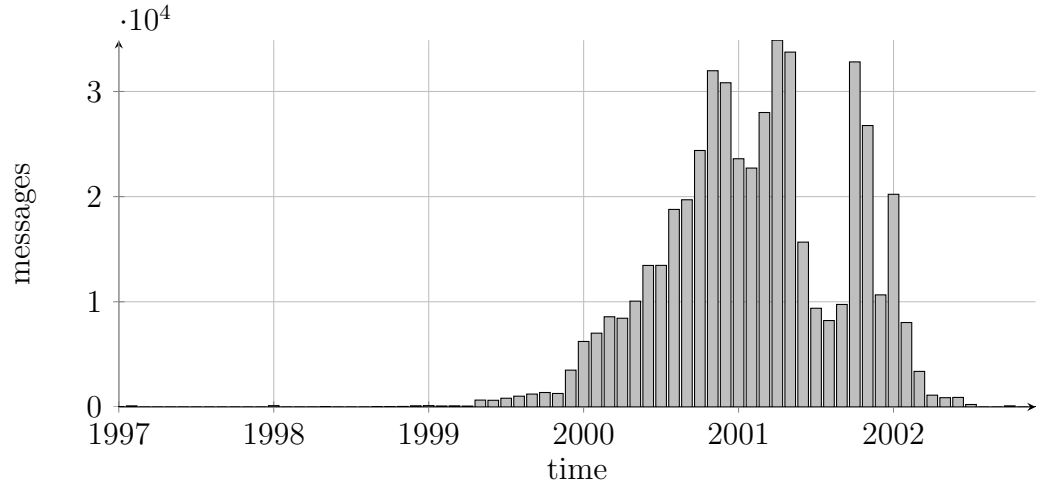

Figure 2: Amout of mails per month

4.2 Ontology

To understand the semantic content of messages, we use WordNet as an ontology, with the *hypernym* relation playing the role of the $is - a$ relation,

[2]Available at `http://www.cs.cmu.edu/~enron/`

[3]These emails were spams or other malformed bot-mails like server messages, etc.

and the *entity* synset as root. We perform a relational mapping of the resulting ontology. This allows us to browse the ontology and to calculate the semantic distance between concepts in constant time. In addition, the use of the synset of WordNet allows us to lift the ambiguity of meaning, as shown by Table 1. The close common ancestor detected, `digit`, is not a source of confusion.

concept 1	**concept 2**	**common ancestor**	**similarity**
dog	cat	animal	0.571
Persian cat	Egyptian cat	domestic cat	0.888
thumb	little finger	digit	0.778
seven	two	digit	0.857
seven	little finger	entity	0

Table 1: Example of computation of semantic similarity

4.3 Communities

As explained in the model, we parse every mail, and extract their main topics. We generalize and summarize them, to obtain the top concepts of the system. Based on our computed semantic similarities, we extract and cluster the main community topics, as shown in Table 2.

rank	**concepts**
#1	{market,services,providence,questioning,management}
#2	{forward,informant,attache,reporter}
#3	{pleasing,contraction}
#4	{subjectivity}
#5	{energy,gas}
#6	{time,change}
#7	{company,business}
#8	{newness}
#9	{thanks}
#10	{power}

Table 2: Concept clusters of communities

4.4 Centrality

Based on this clusters, we compute SPP and centralities for each community. Tables 3 and 4 show results for two of them.

login	$N^+ - N^-$	centrality	position
kate.symes	4310	**5438**	Employee
kay.mann	14332	**3208**	Assistant General Counsel
vince.kaminski	8432	**1170**	Managing Director for Research
		...	
steven.kean	4571	**348**	Vice President & Chief of Staff
		...	
enron.announcements	7284	**0**	Mailing list

Table 3: Centralities of #1{market,services,...} community

login	$N^+ - N^-$	centrality	position
kay.mann	1884	**2810**	Employee
vince.kaminski	2456	**1335**	Managing Director for Research
tana.jones	650	**810**	Employee
		...	
steven.kean	1203	**272**	Vice President & Chief of Staff
		...	
enron.announcements	2477	**0**	Mailing list

Table 4: Centralities of #5{energy,gas} community

It is interesting to note that the centrality does not appear to be directly related to activity (set of posts) within the community. The best example is the announcement address. Despite a strong activity in each of the identified communities, it does not have any centrality. This reflects the fact that if it writes to all, no one communicate with it. It is therefore absent of any communication path identified.

In a second step, it is also interesting to note the role of senior managers. Although their communication is important, and their centrality honorable, they are rarely well positioned in our ranking. This can be explained by their position in the company. As leaders, they are often the start or the end of the communication chain. That is why the best centrality is often held by an employee.

5 Discussion

5.1 Community analysis

The implementation of our model on the Enron data set allows us to compare our results with the reality of this company and its communication network. An interesting point about this is that although the data set contains a high proportion of spam, not any content of this style has emerged from the analysis. This is a great advantage of taking into account the semantic centrality compared to simple raw frequencies: Although the messages are dispatched in large quantities, the total lack of interest that relates users to their content makes them virtually non-existent within the "useful" content of communication that we extract.

In addition, we portray a reality of the corporate communication. If the leaders are of course always present in discussions about their centers of activity and responsibility, they are not, however, the heart of the communication. We speculate that central employees in this model seem to be those responsible for secretarial outsourced tasks: requiring strong two-ways communications, such tasks become the centers. But the lack of data on staff assignments in the data set does not allow us to validate this conclusion further.

5.2 Properties of TSC

The Temporal Semantic Centrality has various interesting properties. First, it should be observed that a user forwarding received emails systematically will be granted a high TSC. Indeed, this centrality does not measure information addition to a message, but the probability to transmit information efficiently. We identified in this respect the forwarding robot of Enron emails as a central "user". This robot is central as it represents a efficient way of propagating messages.

Second, we do not favor explicitely co-occurrences of concepts in emails. For example, it seems natural to weight higher a user who conveys concepts $\{a, b\} \in L_i$ in a unique message m_1 rather than a user conveying a then b in two distinct messages m_2 and m_3. But the definition of SPP takes this co-occurrence into account, as m_1 will contribute twice with the same lag, and m_2 (resp. m_3) will contribute once, with a longer lag (unless m_2 and m_3 are simultaneous, which is unlikely).

5.3 Incremental aspects

Our approach can be interpreted both as off-line and on-line. The off-line interpretation allows to set a given communication network, then to extract its hot topics, to identify users and their communities, and finally to rank them according to the temporal semantic centrality. This approach enables the detection of hot topics that are representative on the whole data set, and to perform the community analysis accordingly.

But it is noteworthy that our algorithms can be implemented in an incremental way: when a new message is acquired by the system (say a post or an email), the user profile and the current list of hot topics can be updated, without a complete recomputation on the whole message log. Also, the SPP and TSC computation can be updated for users concerned by acquired messages. This approach implies that a new hot topic c can appear at a time t during message analysis, and that the centrality according to c must be understood as "after instant t". For example, the topic "federal investigation" for the Enron data set may appear as hot at a given date, but users talking about this topic before this date will not be considered as central. The main advantage of this approach is to enable both topics and centrality monitoring in real time. Also, a message does not require to be materialized after its treatment, which can be crucial for rate intensive monitoring tasks (e.g. Twitter).

6 Related Work

By the emergence of collaborative Web, community of users is a contemporary subject of studies [16, 21]. The new challenge is to detect such activities, thereby defining commonality, and clustering users based on their affinities [1].

Models have been proposed to modelize users' influence applying data mining techniques [20], but they do not take semantics into account. Several studies have focused on the importance of comment activity on blogs or news sites [9, 17]. Sometimes more important than the initial news article, comments have a social role, like staying in contact with friends or meeting new people. Previous works allows extracting emergent structures of discussion within exchange of comments on blogs, in order to determine, for example, popular topics, or those that generate most conflicts of opinion [18], or relational implications between users [3, 19]. Similar methods were also tested on comment-sets from news sites, combining various methods of text mining

(information retrieval, natural language and machine learning) in order to improve the accuracy of detection of these discussion structures [23]. This information is considered useful to increase the meaning of the initial article, but do not focus on the authors of these comments, and on what can be inferred about themselves.

Different approaches focus on mapping the user interests to an ontology also exists [8, 24], based on the user's Web browsing experience. Our method relies on richer users contributions (posts), with a common ontology for all users.

Previous works also consider a notion of semantic centrality [14], in the context of query rewriting. In this work, betweeness is computed on a binary "knows" relation. The semantic similarity is between users ontologies (not posts), and no temporal aspect is taken into account.

A recent work [7] obtains a ranking by computing the betweenness centrality on the communication graph. In this approach, there is an unoriented edge between two users if they exchanged a message once. The centrality of a user u is then the number of shortest paths between any pair of users v, v' *passing through* u, divided by the total number of shortest paths. Hence, betweenness centrality focuses on users playing a great role on the communication structure of the community. But this previous work does not explore the exchanged topics on these shortest paths, nor the speed of the considered communications. Moreover, computing shortest paths is known to be computation intensive.

In [25], the authors study various centrality metrics that incorporate temporal aspects. We agree on various of their observations, like the prominent role of secretaries in the Enron communication graph. We differ from their approach by the incorporation of a structured semantics, and the incremental possibilities of our computations.

7 Conclusion

We presented in this paper an approach to detect central users in a communication network by building semantic-driven communities and evaluating message quality. For this purpose, we have introduced a new measure, the *Semantic Propagation Probability* to take into account semantic accuracy and time delay.

As a future direction, we will consider the transformations that a message undergoes in a communication path, in order to find the user's position (adviser, accountant, etc.), or determine the user's capabilities (computation, correction).

References

[1] S. Amer-Yahia, L. Lakshmanan, and C. Yu. Socialscope: Enabling information discovery on social content sites. In *Conference on Innovative Data Systems Research (CIDR)*, Sep 2009.

[2] M. Bilenko and M. Richardson. Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 413–421, New York, NY, USA, 2011. ACM.

[3] M. De Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts. Inferring relevant social networks from interpersonal communication. In *International conference on World wide web (WWW)*, pages 301–310, New York, NY, USA, 2010. ACM.

[4] E. Desmontils and C. Jacquin. Indexing a web site with a terminology oriented ontology. In *SWWS'01: International Semantic Web Working Symposium*, pages 181–198. IOS Press, 2002.

[5] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In *WWW'07: Proceedings of the 16th international conference on World Wide Web*, pages 461–470, New York, NY, USA, 2007. ACM.

[6] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD'00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160, New York, NY, USA, 2000. ACM.

[7] H. Fuehres, K. Fischbach, P. Gloor, J. Krauss, and S. Nann. Adding taxonomies obtained by content clustering to semantic social network analysis. *On Collective Intelligence, Advances in Intelligent and Soft Computing*, 76, 2010.

[8] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1:219–234, December 2003.

[9] M. Gumbrecht. Blogs as 'protected space'. In *WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, pages 5+, 2004.

[10] G. Hirst and D. St Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, 1998.

[11] N. Imafuji and M. Kitsuregawa. Effects of maximum flow algorithm on identifying web community. In *WIDM'02: Proceedings of the 4th international workshop on Web information and data management*, pages 43–48, New York, NY, USA, 2002. ACM.

[12] H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 661–669, New York, NY, USA, 2005. ACM.

[13] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33, 1997.

[14] J. J. Jung. Query transformation based on semantic centrality in semantic social network. *Journal of Universal Computer Science*, 14(7):1031–1047, 2008.

[15] D. Leprovost, L. Abrouk, and D. Gross-Amblard. Discovering implicit communities in web forums through ontologies. *Web Intelligence and Agent Systems: An International Journal*, 10:93–103, 2011.

[16] A. S. Maiya and T. Y. Berger-Wolf. Sampling community structure. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 701–710, New York, NY, USA, 2010. ACM.

[17] E. Menchen-Trevino. Blogger motivations: Power, pull, and positive feedback. *Internet Research 6.0*, 2005.

[18] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *WWW06 Workshop on the Weblogging Ecosystem*, 2006.

[19] M. Mitrović, G. Paltoglou, and B. Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02005, 2011.

[20] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 61–70, New York, NY, USA, 2002. ACM.

[21] M. Roth, A. B. David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom. Suggesting friends using the implicit social graph. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, New York, NY, USA, 2010. ACM.

[22] J. D. S. Lozano and A. Arenas. Community detection in a large social dataset of european projects. In *Workshop on Link Analysis, Counterterrorism and Security (SIAM on Data mining 2006)*, 2006.

[23] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *ACM international workshop on Web information and data management (WIDM)*, pages 97–104, New York, NY, USA, 2007. ACM.

[24] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 525–534, New York, NY, USA, 2007. ACM.

[25] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, SNS '10, pages 3:1–3:6, New York, NY, USA, 2010. ACM.

[26] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.