Temporal Semantic Centrality for the Analysis of Communication Networks

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

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

1 Introduction

With the advent of the collaborative Web, each website can become a place for expression, where users' opinions are exchanged. User messages are valuable for the site owner: in addition to a proof of interest for the website, 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 [1]. Many approaches have been developed, initialy relied on a basic relationship between users like friendship in social networks or answers / citations in social communication networks (like forums or emails).

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 use 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*

M. Brambilla, T. Tokuda, and R. Tolksdorf (Eds.): ICWE 2012, LNCS 7387, pp. 177–184, 2012.

(*SPP*) and *Temporal Semantic Centrality* (*TSC*) that take into account both semantics and communication timestamps *at once*.

A potential limitation of using an 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 ontologies. 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 experiments in Section 4. Section 5 discusses the obtained results and Section 6 covers related approaches. Finally, Section 7 concludes[1].

2 Communication Networks and Thematic Communities

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

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 . This simple model assumes that the originator and receptor of a given message are known. 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.

Identifying Hot Concepts. The first step of our method is to determine the hot topics of the communication network, as a subset of concepts of O. We associate with each user a *semantic profile*. 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. Due to a lack of space, we do not provide here a full description of the profile construction of the system, which is available in our previous work [8].

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. In order to control the number of

[1] A detailed version of the method is available as a technical report: `http://hal.archives-ouvertes.fr/hal-00692289`

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. We rely here on the Wu-Palmer distance [13] restricted to concepts *hierarchies* (trees), which has already been applied to similar cases [3]. 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. [5] and various optimizations [7,4] 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.

3 Temporal Semantic Centrality

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*

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 $u \in U$ **do**
    if $N_i^+(u) \leq N_i^-(u)$ **then**
      $U_i = U_i \cup u$
    end if
  end for
end for

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*.

The *dispersion* of a message m according to concept c, noted $dispersion_c(m)$, 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 $c \in content(m)$, 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 ($content(m)$ 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 ($t = t'$).

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:

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

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 $O(kM)$, 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 and (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).

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 set $INBOX(u)$ of mesages recently received by u is kept in main memory. Outcoming messages are then compared to messages in this window.

4 Experiments

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.

Ontology. We use WordNet as an ontology, with the *hypernym* relation playing the role of the $is-a$ relation, and the *entity* synset as root. We perform a relational mapping of the resulting ontology.

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. We extract and cluster the main community topics, as shown in Table 1.

Table 1. Concept clusters of communities

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

Temporal Semantic Centrality. Based on this clusters, we compute SPP and centralities for each community. Table 2 shows results for one of them. 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.

Table 2. Centralities of #1{market,services,...} community

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

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

5 Discussion

Community Analysis. The implementation 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, no content of this type has emerged from the analysis. This is a great advantage of taking into account the semantic centrality compared to simple raw frequencies. 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. We speculate that central employees 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.

Properties of TSC. 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).

6 Related Work

Models have been proposed to modelize users' influence applying data mining techniques [11], or centrality metrics [6]. We differ from their approaches by the incorporation of a structured semantics, the role of each user in the communication, and the incremental possibilities of our computations. Several studies have focused on the importance of comment activity on blogs or news sites [9] and highlight the social role of comments. It allows to determine popular topics, conflicts of opinion [10], or relational implications between users [2]. Different approaches focus on mapping the user interests to an ontology [12], based on the user's Web browsing experience. Our method relies on richer users contributions (posts), with a common ontology for all users.

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 like computation, correction, etc.

References

1. Bilenko, M., Richardson, M.: Predictive client-side profiles for personalized advertising. In: ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 413–421. ACM, New York (2011)
2. De Choudhury, M., Mason, W.A., Hofman, J.M., Watts, D.J.: Inferring relevant social networks from interpersonal communication. In: International Conference on World Wide Web (WWW), pp. 301–310. ACM, New York (2010)
3. Desmontils, E., Jacquin, C.: Indexing a web site with a terminology oriented ontology. In: International Semantic Web Working Symposium, pp. 181–198. IOS Press (2002)
4. Dourisboure, Y., Geraci, F., Pellegrini, M.: Extraction and classification of dense communities in the web. In: International Conference on World Wide Web (WWW), pp. 461–470. ACM, New York (2007)
5. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. In: ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 150–160. ACM, New York (2000)
6. Fuehres, H., Fischbach, K., Gloor, P.A., Krauss, J., Nann, S.: Adding Taxonomies Obtained by Content Clustering to Semantic Social Network Analysis. In: Bastiaens, T.J., Baumöl, U., Krämer, B.J. (eds.) On Collective Intelligence. AISC, vol. 76, pp. 135–146. Springer, Heidelberg (2010)
7. Ino, H., Kudo, M., Nakamura, A.: Partitioning of web graphs by community topology. In: International Conference on World Wide Web (WWW), pp. 661–669. ACM, New York (2005)
8. Leprovost, D., Abrouk, L., Gross-Amblard, D.: Discovering implicit communities in web forums through ontologies. Web Intelligence and Agent Systems: An International Journal 10, 93–103 (2011)
9. Menchen-Trevino, E.: Blogger motivations: Power, pull, and positive feedback. Internet Research 6.0 (2005)
10. Mishne, G., Glance, N.: Leave a reply: An analysis of weblog comments. In: WWW 2006 Workshop on the Weblogging Ecosystem (2006)
11. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 61–70. ACM, New York (2002)
12. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: ACM Conference on Information and Knowledge Management, CIKM 2007, pp. 525–534. ACM, New York (2007)
13. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Association for Computational Linguistics (ACL), pp. 133–138. Association for Computational Linguistics, Stroudsburg (1994)