Discovering Implicit Communities in Web forums through Ontologies

Damien Leprovost, Lylia Abrouk, David Gross-Amblard

Web Intelligence and Agent Systems: An International Journal 10 (2012) 93–103
DOI 10.3233/WIA-2012-0234
IOS Press

Discovering Implicit Communities in Web forums through Ontologies

Damien Leprovost [a], Lylia Abrouk [a] and David Gross-Amblard [a,b]
[a] *LE2I CNRS Lab., University of Bourgogne, Dijon, France*
E-mail: {firstname.lastname}@u-bourgogne.fr
[b] *INRIA Saclay, France*

Abstract. Being a *Community manager* is an emerging employment in social Web companies. His or her role is to monitor communities on a devoted social website, in order to understand new trends or behaviours. He or she also has to discover and attract new potential users of the website in external resources like web forums, that are not necessarily on the same topics nor explicitly defined. In this paper we propose a scalable protocol to monitor on-line communications, like web forums, walls or twits. We provide an analysis method to extract implicit communities and user interests based on the semantics of data exchange and the structure of communications. The method is parameterized by a target vocabulary expressed as an ontology, in order to focus on relevant communities.

Keywords: Semantic analysis, forums, communities, social networks, ontologies

1. Introduction

Attracting a wide community of users is nowadays a common ingredient for the success of a website. Examples are ranging from open-source developers dedicated websites (like Sourceforge) to Brand-related communities managed inside a company (like Apple with its Ping Network) or through groups in generic social-oriented infrastructure (Facebook, Ning, etc.).

In parallel, there is a tremendous growth of job offers for so-called *Online Community Managers*. The corresponding job description includes *Identify and analyze issues, patterns and trends in customer requests & product performance* or *Participate in professional networking by following the prominent bloggers and online writers & attending events*, to name a few[1]. One routine task for the community manager is to monitor internal forums for classical maintenance (organizing topics, banning malevolent users, etc.). This monitoring can also target external forums, as the community manager has to attract new users from other websites. Hence there is a natural need to assist the community manager in browsing external forums and analyzing communities behind them. In this work we focus on web forums for the sake of simplicity, but the same technique applies for any communication system with a send/answer structure like internal emails, posts on accessible Facebook walls, twit/answer/retwit on Twitter, and so forth.

One of the main difficulties of this analysis task is the overwhelming amount of posts the manager is facing. First, on the semantic side, the monitored forums are not necessarily oriented on the topic searched for by the community manager. It is then necessary to focus on a vocabulary of interest in order to identify relevant users. Second, on the practical side, the community discovery shall rely on an incremental structure, that is updated as soon as new posts are extracted from forums. This is a mandatory condition to scale up to a large amount of monitored forums.

In this paper we propose a scalable method for the semantic analysis of Web communication. A few previous work consider the analysis of on-line communications [15], but with a purely statistical approach, and without an a priori knowledge on the target vo-

[1] http://conniebensen.com/2008/07/17/community-manager-job-description/

cabulary. On the contrary, our method rely on one or several ontologies of interest selected by the community manager. The main contribution is a profile model that takes into account the semantic of messages during communications, and relates each user with concepts of the ontology. We propose a scalable method to sum up users contributions by generalizations of concepts according to the ontology. Concept detection is enriched by context propagation along the question/answer structure of the target forum. We assess our method on the comments of a popular on-line newspaper, extracted by wrappers part of our overall analysis plateform WEBTRIBE[2].

The paper is organized as follows. Section 2 gives a brief overview of the related work. Section 3 introduces our analysis model and Section 4 presents the building of our semantic profiles, with an accompanying example. Section 5 describes communities clustering. Section 6 illustrates our method on a real forum and Section 7 concludes.

2. Related work

2.1. Comments analysis

The importance of comment activity on blogs or news sites was the subject of several studies [7,10]. 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 [12]. 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 [16]. 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. On the contrary, our approach is user-centered, based on user similarity by aggregating semantic contributions. The method is also dynamic, as user communities can evolve over time and depend on interests of users.

Different approaches focus on mapping the user interests to an ontology [5,17], based on the user's Web browsing experience. Our method relies on richer users contributions (posts), with a common ontology for all users.

2.2. Implicit Communities

Since the Web birth until now, the community concept has evolved. From the first hyperlinked webpages communities, deduced from the topology of the web formed by the links between pages and sites [8,6,4], the concept of "web communities" is now often understood as user-oriented: a community is a set of users, and is based on the activities of its members, within the collaborative Web [9,14]. The new challenge is to detect such activities, thereby defining commonality, and clustering users based on their affinities [1].

3. Semantic analysis of communications

3.1. Abstracting web communications

Our proposition applies for any structured textual communications where users are identified (this is the case of a vast majority of systems). As a running example we consider the community manager of an healthcare company, and suppose that he or she monitors competing companies' forums to identify interesting communities and important users. Figure 1 shows an example of such communications. We suppose that an automatic crawler has been assigned the task to monitor these forums, that may have been discovered by classical keyword search on the Web, and more specifically by focusing on the classical platforms that power them (for example the query "PhpBB health" is likely to return interesting forums supported by the popular PhpBB[3] tool).

Then, a forum is seen as a set P of posts issued by a set U of users. Users are identified by their visible id (email address, forum id, etc.). We denote by $author(p) \in U$ the unique author of a given post p, and by $post(u) \in P$ the set of user u's posts.

There are several technical annotations or textual conventions for answering a given post. For emails or tweets the users who is answered to is explicitly given. For purely web systems, there exists common situations or practices to express answers. A classical pat-

[2] http://www.damien-leprovost.fr/webtribe

[3] http://www.phpbb.com

```
Alice: I feel a pain in my left arm.
Joey:  Is there a physician on this
       WebSite ?
Bob:   @Alice: Could you be more
       precise ?
Alice: My shoulder hurts.
Bob:   ----------------------------
       I feel a pain in my left arm.
       ----------------------------
       Did you perform a strong move
       recently ?
Alice: Yes, I played tennis
       yesterday.
```

Fig. 1. Posts in a forum

tern is to start the answer with a "@u" pattern, that indicates a message to user u. A second convention is to cite (part of) the answered message. Examples of these patterns are shown on figure 1. In order to keep track of this information, we denote by $cite(p) \subseteq P$ the set of posts answered by post p (we will discuss the computation of $post(p)$ in Section 4).

In order to finely define the axis of the forum semantic analysis, we naturally rely on a domain or generic ontology. Its choice is therefore critical: a specific and detailed domain ontology should be chosen for the analysis of specialized forums, and general ontologies could be preferred for generic forum or initial explorations. A specialized ontology could be the precise names of brand products with their relationships (e.g. an iPhone 4 – 32Go is a kind of iPhone which is a SmartPhone). Generic ontologies are plentiful: Wordnet [11], YAGO [18], DBpedia [2], to name a few. On the community manager's side, choosing the right ontology is a very interesting problem on its own, but which is not the focus in the present paper. In our example the chosen ontology is a medical information ontology like MESH [3], an anatomical ontology like FMA [13] or a thematic cut into a generic ontology.

More formally, we are given an ontology $(C, is-a)$, where $C = \{c_1, \ldots, c_n\}$ is a set of concepts and $is-a$ is the direct subconcept relation structuring the ontology ($is-a(c, c')$ denotes that c is a direct child of c'). The set of concepts of C manipulated by a given post p is denoted by $concept(p)$. This set can be computed by stemming the post p and removing stop-words, and by comparison with the ontology (of the stemmed terms of the ontology). We denote by $occurrence_p(c) \in \{0, 1\}$ the occurrence of a concept c in a post p.

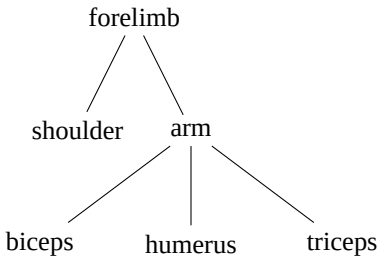

Fig. 2. Target ontology

According to the target ontology of Figure 2, for p the second post of Alice in Figure 1,

$$occurrence_p(\text{"}shoulder\text{"}) = 1,$$

and no other relevant concept appears in this post.

4. Semantic profiles and generalization

Our semantic analysis of a forum is performed in two steps. First, we associate with each user its *semantic profile*, which represents its forum contributions in terms of the ontology. These semantic profiles take into account the question/answer structure of the forum. Then, we aggregate users' contributions according to their semantic profile, in order to sum up his/her activity in a few concepts.

4.1. User profiles

The method shall be seen as an incremental method: as soon as a new post p from a user u is discovered by the crawler, the system updates its profile. The profile of user u according to concept c, $profile_u(c)$ could be for a first definition the total number of occurrences of c in $post(u)$.

This initial definition is not satisfactory: a post may embrace a wider scope than just the words it contains, according to its context. Of course, we can not claim a comprehensive understanding of all contexts, nevertheless, we consider the question/answer context. When a user replies to another, he indeed places its message in the context of the original message, as in Example 1.

Example 1 *We consider two posts p_1 and p_2.*

```
Joe:  I underwent several heart
      operations in the past, and I
```

```
      just moved to Manhattan.
      I seek the address of a good
      cardiologist.
Dave: @Joe There is one good at
      1546, 7th Avenue.
```

Even if the second message has no explicit semantic content on the studied area, we can propagate the semantic content of the first message, identified as its context.

It is noteworthy that, for scalability reasons, the entire post collection cannot be kept. It is thus not possible to assess the $cite$ relation in all situation in practice. A relevant time-window has to be chosen, starting from the post at current time. An $@u$ answer in post p is then interpreted as a citation of the immediately previous post of user u in the chosen time window, or discarded as a citation if no such post is found. Similarly, the text of a current post is compared to all posts in the same time-window. If a significant part of the text appears as an extract of a previous post in the time-window, it is considered as a citation.

Thus, we use the $cite$ relationship to enrich the semantic profile. We then define $profile_u(c)$ as then the sum of all occurrences of concept c in posts of user u and its cited posts:

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

Observe that we chose a non-recursive definition, in order to evaluate posts in a given time-window only, again for scalability reasons. When a new post p from user u is crawled, its cited posts within the time-window are extracted, and $profile_u$ for relevant concepts is incremented.

Without the scalability constraint, a recursive definition of contexts could be envisioned: the temporal ordering of posts guarantees a loop-free recursion, as a post from the past can not cite a post in the future (although some forum systems allows for the modification of a post in the past, this modification should be seen as a new post at modification time).

4.2. User abstracts

User profiles can be enriched incrementally over the future contributions. However, a profile that describes all the user activity often contains information that are not relevant to describe briefly the user, like concepts with rare occurrences compared to others. But if they are not relevant now, we can not discard them immediately because they may become salient later. Indeed, a user can change its activity gradually. We then introduce the user abstract $abstract_u$, as the current summarization of user u's semantic interests. For a concept c, $abstract_u(c)$ is the weight of concept c in the abstracted view of u. Summarization is composed of 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 that manipulates its subconcepts. For a leaf concept c of the ontology, the abstract is simply the profile (no generalization can occur), that is:

$$abstract_u(c) = profile_u(c).$$

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. The significance threshold is materialized by $\delta_{coverage} \in [0, 1]$. Then, if

$$\frac{|\{c_i : is-a(c_i, c) \text{ and } abstract_u(c_i) > 0\}|}{|\{c_i : is-a(c_i, c)\}|} \geq \delta_{coverage},$$

the abstract of c is the average abstract of *all* subconcepts of c:

$$abstract_u(c) = \frac{1}{|c' : is-a(c', c)|} \sum_{c' : is-a(c', c)} abstract_u(c') + profile_u(c).$$

If subconcepts are not well covered, then simply

$$abstract_u(c) = profile_u(c).$$

The second step for our abstract construction simply deletes concepts from the abstract 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

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

Observe that we first generalize concepts, then delete those without relevance. This allow to discover a well-covered concept which is not explicitly used. These notions are illustrated on Example 2.

Example 2 *Considering a local part of the ontology, and the related* $profile_u$ *of Figure 3. For* $\delta_{coverage} = 0.66$ *and* $\delta_{relevance} = 0.5$*, the resulting* $abstract_u$ *appears in Figure 4.*

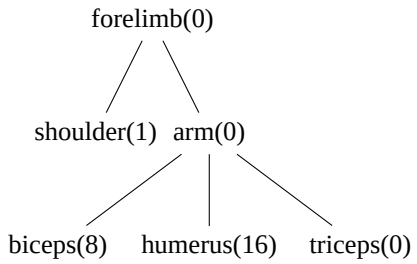

Fig. 3. User profile

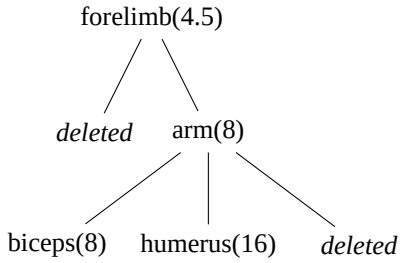

Fig. 4. User abstract, $\delta_{coverage} = 0.66$ and $\delta_{relevance} = 0.5$

As highlighted in the example, generalization produces weights on upper concepts, without removing previous subconcepts weights. By doing so, the awareness of a generalization of the user does not erase the specialties he has. A basic method to compute the abstract in a bottom-up manner. Algorithm 1 presents a global computation on the whole ontology while 2 is incremental: only the impacted concepts or subconcepts are tested and eventually modified (this method starts with a blank $profile$ and $abstract$).

A third version of the algorithm was produced, recursive this time. Algorithm 3 has the advantage that it does stand as a concept only if it is impacted by the change of one of its son concepts. However, for performance reasons, we use for the experiment presented in Section 6, the incremental version. Its overcost, due to the tree parsing, having been neutralized by an abstrac-

Algorithm 1 Generalization

Input: $profile_u, ontology, \delta_{coverage}, \delta_{relevance}$
Output: $abstract_u$

$abstract_u = profile_u$
for all depth d of the ontology, starting at leaves **do**
  for all concept c at depth d **do**
    $abstract_u(c) = profile_u(c)$
    if c is not a leaf **then**
      $S =$ subconcepts of c
      $SU = \{c' \in S : abstract_u(c') \neq 0\}$
      if $\frac{|SU|}{|S|} > \delta_{coverage}$ **then**
        $abstract_u(c) += \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|}$
      end if
      if $\frac{abstract_u(c)}{\sum_{c' \in C} abstract(c')} < \delta_{relevance}$ **then**
        $abstract_u(c) = 0$
      end if
    end if
  end for
end for

Algorithm 2 Incremental generalization

Input: $abstract_u, post, ontology, \delta_{coverage}$
Output: $abstract_u$

for all depth c covered by $post$, starting at leaves **do**
  if c is not a leaf **then**
    $previous = 0$
    $S =$ subconcepts of c
    $SU = \{c' \in S : abstract_u(c') \neq 0\}$
    if $\frac{|SU|}{|S|} > \delta_{coverage}$ **then**
      // Save existing previous value
      $previous = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|}$
    end if
  end if
  $abstract_u(c) += post(c)$re
  if c is not a leaf **then**
    $SU = \{c' \in S : abstract_u(c') \neq 0\}$
    if $\frac{|SU|}{|S|} > \delta_{coverage}$ **then**
      // Update of abstract by adding difference with previous
      $abstract_u(c) += \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|} - previous$
    end if
  end if
end for

Algorithm 3 Recursive Generalization

Input: $abstract_u, concept, value, ontology, \delta_{coverage}$
Output: $abstract_u$

$toParentValue = 0$
if *concept* is not the root **then**
  $S =$ subconcepts of $parent(concept)$
  $SU = \{c' \in S : abstract_u(c') \neq 0\}$
  if $\frac{|SU|}{|S|} > \delta_{coverage}$ **then**
    $toParentValue = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|}$
  end if
end if
$abstract_u(c) += value$
if *concept* is not the root **then**
  $SU = \{c' \in S : abstract_u(c') \neq 0\}$
  if $\frac{|SU|}{|S|} > \delta_{coverage}$ **then**
    $toParentValue = \frac{\Sigma_{c' \in S} abstract_u(c')}{|S|} - toParentValue$
  end if
  if $toParentValue \neq 0$ **then**
    Recursive Generalization($parent(concept)$, $toParentValue$)
  end if
end if

tion of relational ontology. Indeed, we store in a relational database, not the ontology, but only its characteristics of interest. The depth of each concept is known during the abstraction, and the incremental path can then "browsing" the tree, without the cost of such an operation.

5. Clustering Communities

The previous computations enable to deduce the communities of the target forum. We perform this task in the following two steps:

- detect the main concepts, covered primarily by user's contributions;
- cluster users around these concepts.

5.1. Main concepts

We sum up all computed $abstracts$, concept-wise. The resulting *global abstract* can be seen as the abstract of the whole forum, after generalization of contributions. As for each user's abstract, we apply the relevance threshold, $\delta_{relevance}$ to keep only the major concepts of the forum (Algorithm 4).

Algorithm 4 GlobalAbstract

Input: $abstracts, ontology, \delta_{relevance}$

for all concept $c \in C$ **do**
  $globalAbstract(c) = 0$
end for
for all user $u \in U$, concept $c \in C$ **do**
  if $\frac{abstract_u(c)}{\sum_{c' \in C} abstract(c')} \geq \delta_{relevance}$ **then**
    $globalAbstract(c) += abstract_u(c)$
  end if
end for
for all $c \in C$ **do**
  if $\frac{globalAbstract(c)}{\sum_{c' \in C} globalAbstract(c')} < \delta_{relevance}$ **then**
    $globalAbstract(c) = 0$
  end if
end for

5.2. Communities

To each major identified concept, we now attach its main contributors (Algorithm 5).

Algorithm 5 Communities

Input: $abstracts, globalAbstract, \delta_{relevance}$

$Communities = \emptyset$
for all user $u \in U$ **do**
  for all $c \in C$ s.t. $globalAbstract(c) > 0$ **do**
    if $abstracts_u(c) > 0$ **then**
      if $\frac{globalAbstract(c)}{\sum_{c' \in C} globalAbstract(c')} \geq \delta_{relevance}$ **then**
        $Communities(c) += u$
      end if
    end if
  end for
end for

5.3. Social Analysis

By exploiting the previous steps, we can obtain various socials information about users and their behavior:

- main topics of interest: the main interests of a user u are just the sorted list of concepts in $abstract_u$, order by their weight;
- community top-users: top-users define the user group at the center of a community, that is those users who provide the forum with contents closest to the community topic: in other words, those users u with the highest $abstract_u$ score on the summarized community concepts.

This information is illustrated in our experiments.

6. Experiments

6.1. Data sets

As a data source, we have focused on USA Today[4]'s website, an U.S. online newspaper, and more precisely on its Health News section. Health News represent a subpart of the website, and consist in clustered news published by with the *Health* section tag. Each health new comes with identified users comments, with citations and answers features. Focusing on it, we performed a specialized study on this area.

We developed a specialized crawler and a wrapper that includes HTML and JSON parsers. We extracted around 15k of user comments. All these contributions are signed by their authors (authenticated users). All contributions are normalized, and represented as standard XML documents, whose markup declarations is provided by a DTD[5]. Contributions are processed in a stream-oriented way.

number of article	200
number of comments	14978
number of users	3682
min comments per user	1
average comments per user	4
max comments per user	447
1-comment users	1848
min comments per news	0
average comments per news	75
max comments per news	1642

Table 1
News-article Metrics

The statistical analysis of collected data, summarized in Table 1, shows that more than half of the forum users are only identified by the publication of a single comment. Conversely, as shown in Figure 5, a minority of users is responsible for the majority of contributions.

These features are indicative of an open forum, where anyone can participate without being personally bound to the forum (free-riding behavior). The involvement of the majority of users is low. Consequently, a larger volume of data is needed to obtain valid conclusions about communities, as the majority of contributions are not significant. In contrast, in a denser forum, as a platform for software management, members feel strongly involved. The number of contributions required would then be lowered.

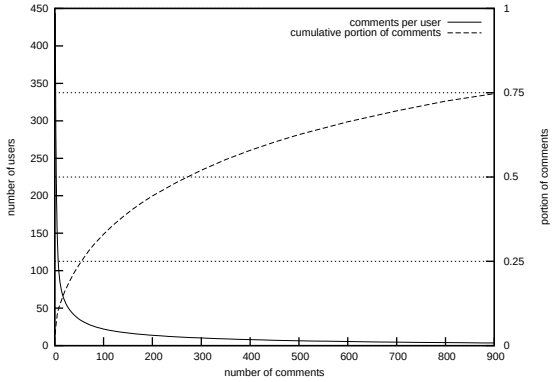

Fig. 5. Message distribution among users

6.2. Concept detection

To analyze these discussions and in agreement with our model, we use an ontology. However, commons medical ontologies, like MESH [3] or FMA [13] do not appear to be suitable for this case. Indeed, they use precise and specialized terms: this level of vocabulary is rarely used by the general public. To overcome this problem, we perform a thematic cut into WordNet [11], a famous lexial database for the English language. We assume as root the concept *body part*, witch results an ontology of words used to describe the differents parts of the human body.

concepts	1824
concept detections	55875
average concept by post	3.73
0-concept users	125
0-concept user rate	3.4%
0-concept comments	994
0-concept comments rate	6.64%

Table 2
Semantic Metrics

We apply the method of concept detection described in Section 4. As shown in Table 2, comments are often poor in semantic content contributions (we discuss later on how to enhance concept detection).

[4]http://www.usatoday.com
[5]available at: http://www.damien-leprovost.fr/webtribe/thread.dtd

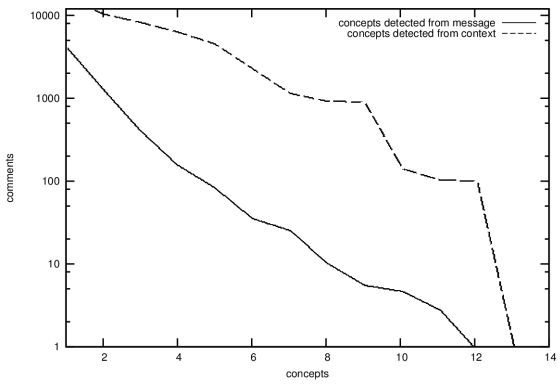

Fig. 6. Detected concepts by source method

6.3. Communities

By applying the algorithms presented in Section 5, we build user profiles, summarize user abstracts and cluster communities. A community is characterized by the number of its users, and its semantic weight (Tables 3, 4, and 5, commented later on).

6.3.1. Taking context(s) into account

We conducted three distinct computations, with three distinct context definitions. On the first one, we computed communities with propagation of the Thread Context (TC): messages are contextualized according to the thread (starting news) to which they respond. Figure 6 shows that taking the context into account enables to significantly increase the concept detection.The resulting communities are shown in Table 3.

rank	main concept	users	weight
1	Heart	338	5098
2	Brain	159	2227
3	Lung	153	1952
4	Heart	23	1104
5	Belly	93	982
6	Neck	110	917
7	Skin	131	913
8	Large inestine	25	168
9	Liver	21	167
10	Heel	3	167

Table 3
Detected Communites (TC)

For the second computation, we took into account only the local context of direct answers, called Anwser Context (AC). Each message in context take the message to which it responds. The resulting communities are shown in Table 4.

rank	main concept	users	weight
1	Belly	225	1677
2	Heart	172	653
3	Lung	121	541
4	Brain	117	413
5	Side	123	395
6	Heel	27	263
7	Liver	48	155
8	Skin	47	139

Table 4
Detected Communites (AC)

For the third one, we used the two contexts simultaneously, Thread and Anwser Contexts (AC&T). The resulting communities are shown in Table 5.

rank	main concept	users	weight
1	Heart	306	5192
2	Brain	167	2368
3	Lung	145	2081
4	Belly	112	1582
5	Neck	111	940
6	Skin	115	877
7	Side	64	699
8	Heel	5	213
9	Liver	21	195
10	Large intestine	25	175
11	Eye	18	151
12	Knee	15	123
13	Hand	13	112

Table 5
Detected Communites (T&AC)

The influence of these different contexts is related to the structure of the analyzed system. In our example, we are working on a system of comments on published news. Consequently, news have a volume of information stronger than the frequent short comments left there by free-riding users. Users also usually respond to the initial news rather than to another comment. That is the reason why, in the case of a news site, the Thread Context is dominating, while the Answer Context is low in information.

Accordingly, AC-calculated communities are weighted relatively low. TC-communities are more robust, and T&AC-communities can be seen as an improvement, but weakly significant. But in a totally different system like a discussion forum, where responses are more important than the topic initiator, this behavior would

probably be reversed, with a dominating importance of AC on TC.

6.3.2. Top-users

In the TC setting, from the largest detected community about *Thorax*, we search for its top users. Table 6 presents the results of this computation (nicknames are anonymized).

rank	nickname	user portion of community	user portion of site
1	li	28,12 ‰	**37,61** ‰
2	gre	27,54 ‰	**35,99** ‰
3	xiu	**20,22** ‰	16,25 ‰
4	BAL	**18,68** ‰	12,63 ‰
5	uknowi	**15,60** ‰	11,98 ‰
6	popo	**14,83** ‰	4,35 ‰
7	brokena	**14,06** ‰	3,41 ‰
8	zoila	**13,29** ‰	3,37 ‰

Table 6

Top-users of *Heart* community

Top-users are sorted, according to the weight they occupy in the community, that is the weight of their contributions. However, we can distinguish two types of community members:

- major contributors of the system. If they do not overlook this topic area, they necessarily deal with a high score in this community. Their rank is a logical consequence of their strong involvement in the whole system. This is felt especially in an open system with unequal proportions of contribution, as explained previously;
- dedicated contributors. Their involvement in the subject of the community is greater than in the rest of the entire system. They are mainly focused on the topic area and do not necessarily need a large number of contributions to figure prominently in the community. The more the system grows and becomes denser, the more this type of user tends to be found in communities.

6.3.3. User main topics

As described in Section 5, detection of user main topics is a simple cut into user abstracts. Table 7 is an example.

rank	concept	portion
1	Heart	26.25%
2	Lung	9.75%
3	Side	7.00%
4	Skin	3.75%
5	Liver	3.00%
6	Hand	1.75%

Table 7

Main topic of user *xiu*

6.4. Order of Magnitude

We sought to know what is the order of magnitude of the number of users we could manage through this method. We take as example a server, relatively common these days, with 10GB of RAM. We consider an ontology of 1024 terms, and storage of integers in two unsigned bytes.

On the one hand, we compute the space occupied by user data (profiles and abstracts), and on the other hand treatment data (the message being analyzed, the contexts that have spread). The size of user data appears to be dependent on the average number of concepts that users manipulate. We declare two possible cases: the worst possible case (always all the concepts used), and the average case observed on USA Today.

Based on these assumptions, the worst possible case allows us to maintain around 268 million of users. With averages of USA Today, that number rises to 11 billion. This is due to the low coverage of the majority of users, typical of a very open system.

6.5. Ontology impact

In order to analyze the impact of the ontology choice, we also tested our data set with a smaller knowledge base. As previously explained, we want to avoid the problem of specialization of the language. So, we built a descriptive ontology draft of the human body[6], based on Wikipedia body description. It represented a first approach using common words from everyday language, including body parts, muscles and bones. But it is interesting to note that with such an ontology, much smaller than WordNet cuts, results are broadly similar. The detection rate is much smaller, but the general appearance of communities is similar. This confirmed that the density of the ontology allows refining the results, but the main contribution lies in the operation of generalization. Indeed, the relationships between concepts can preserve semantics consistency, regardless of the level of accuracy.

[6] `http://www.damien-leprovost.fr/webtribe/HumanAnatomyBasics.owl`

Conversely, the use of the whole WordNet, in addition to performance degradation, presents a scattering of concepts. We identify the most used words in the English language, but these are not consistent with the subject of study intended for the forum: communities then loose their interest. This confirms our need for a relationship between the system analyzed and the chosen ontology.

7. Conclusion and future work

We presented here an analysis method to extract implicit communities from a target communication system, and illustrated the importance of context propagation for understanding communication semantics. These data provide to the Community Manager interesting information about discussed topics, users profiles and behavior inside the targeted system. This information allows the Community Manager to take decisions on how to manage the communities. For a local forum, he can for example decide to split discussion threads according to the identified communities, or to directly talk with identified leaders to improve feedbacks, and so on. All these functionalities will be part of a global Community Management Tool, WEBTRIBE.

The approach can be enhanced in several directions:

- semiotic analysis of contributions. This will allow developing new axis of interpretation of communication. It will allow, for example, to detect disputed claims, opposites users, etc.
- targeting users. A Community Manager can compare his own communities with externals, and deploy strategies to be more attractive or efficient about treat and opportunities from the outside.
- temporal aspects. As our algorithms follow a stream-like style, it should be useful to remember the temporal evolution of our communities, leaderships, etc. This can provide major information, like emerging communities that must be considered, aging communities that need to be overhauled, and so on.

Acknowledgments

Work partially supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC grant Webdam, agreement 226513. (http://webdam.inria.fr/).

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] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009.

[3] M. C. Díaz-Galiano, M. A. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ure na-López. Integrating mesh ontology to improve medical information retrieval. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 601–606, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] 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.

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

[6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *HYPERTEXT'98: Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 225–234, New York, NY, USA, 1998. ACM.

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

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA'98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.

[9] 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.

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

[11] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

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

[13] C. Rosse and J. L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, December 2003.

[14] 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.

[15] T. Schoberth, J. Preece, and A. Heinzl. Online communities: A longitudinal analysis of communication activities. In *Hawaii International Conference on System Sciences*, volume 7, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[16] 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.

[17] 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.

[18] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.