Social Networking, 2014, 3, 86-93
Published Online February 2014 (http://www.scirp.org/journal/sn)
http://dx.doi.org/10.4236/sn.2014.32011

A Process to Support Analysts in Exploring and Selecting Content from Online Forums

Darlinton Carvalho[1], Ricardo Marcacini[2], Carlos Lucena[1], Solange Rezende[2]
[1]Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil
[2]Institute of Mathematical and Computer Sciences, University of Sao Paulo, São Carlos, Brazil
Email: darlinton@acm.org, rmm@icmc.usp.br, lucena@inf.puc-rio.br, solange@icmc.usp.br

Received December 15, 2013; revised January 12, 2014; accepted February 7, 2014

ABSTRACT

The public content increasingly available on the Internet, especially in online forums, enables researchers to study society in new ways. However, qualitative analysis of online forums is very time consuming and most content is not related to researchers' interest. Consequently, analysts face the following problem: how to efficiently explore and select the content to be analyzed? This article introduces a new process to support analysts in solving this problem. This process is based on unsupervised machine learning techniques like hierarchical clustering and term co-occurrence network. A tool that helps to apply the proposed process was created to provide consolidated and structured results. This includes measurements and a content exploration interface.

KEYWORDS

Qualitative Analysis of Online Forums; Explore and Select the Online Forums Content; Machine Learning; Hierarchical Clustering; Terms Co-Occurrence Network; Consolidated and Structured Results

1. Introduction

Online communities [1] are important meeting points for people on the Internet, leaving records of their discussions in the online forums. The public content increasingly available in these forums enables researchers to study society in new ways. Influenced by ethnographic principles, Kozinets [2] says that the Netnography observation happens in users' natural habitat, being relevant for its users, detailed and contextualized in community. The content can be retrieved in non-intrusive ways, enabling opportune, effective, and efficient processing. In this way, users are not summoned to participate in a reactive fashion (e.g. online surveys), which enables the analysis of spontaneously constructed manifestations.

An opportunistic approach that explores social media in order to conduct studies is presented in [3]. This approach relies especially on a qualitative analysis of online forums. A technique that helps to perform this analysis can be the Discourse of the Collective Subject, a qualitative technique with roots in the Theory of Social Representations [4]. The aim of this technique is to identify collectives aggregated by central ideas, and describe them by means of a discourse woven as a patchwork from the collective members' speeches, synthesizing the discourse as one collective subject. In order to achieve some interesting results, this technique requires considerable effort from analysts, which is a common characteristic of qualitative analysis methods.

However, most content of online forums is not related to analysts' interests in the context of a study. Consequently, analysts face the following problem: how to efficiently explore and select the content to be analyzed? This article introduces a new process to support analysts in solving this problem.

In order to solve this problem, we propose a process performing automatic and intelligent organization of the texts presented in a forum. This is achieved by unsupervised machine learning techniques, solving the tasks of identifying a term co-occurrence network, and hierarchical text clustering. This article also presents TorchSR, a tool based on the proposed process created to support analysts. This tool provides consolidated and structured results, including a content exploration interface.

The remainder of this article is organized as follows: literature related to this work is presented in Section 2. The proposed process is described with its creation rationale in Section 3. The TorchSR tool, based on the proposed process, is shown in Section 4. Section 5 gives the example of a study that would benefit from the proposed process. Section 6 concludes this article.

2. Related Work

2.1. Online Forums as Input for Analysis

Some experts in the scientific community claim that a new scientific field is at rise: the coming of the age of computational social science [5]. Taking the field of health science for example, we present some studies that consider social media as input for analysis.

Lasker *et al.* [6] studied the role of an online community for people with primary biliary cirrhosis through the content analysis of a mailing list. Despite the underlying technology for content sharing, a mailing list is basically an online forum. Accordingly, our work considers a broad definition of online forum [1] as any online system where people can discuss things and share content. Whitehead [7] provides an integrated review of the literature on methodological and ethical issues in Internet-mediated research in the field of health.

Using Facebook, Greene *et al.* [8] performed a qualitative evaluation of communication by patients with diabetes. For instance, they found that "*approximately two-thirds of posts included unsolicited sharing of diabetes management strategies, over* 13% *of posts provided specific feedback to information requested by other users, and almost* 29% *of posts featured an effort by the poster to provide emotional support to others as members of a community*." Bender, Jimenez-Marroquin, and Jadad [9] analyzed the content of Breast Cancer Groups on Facebook, finding 620 breast cancer groups containing a total of 1,090,397 members. The groups were created for fundraising (277 of 620 or 44.7%), awareness (38.1%), product or service promotion related to fundraising or awareness (61.9%), or patient/caregiver support (46.7%).

Madeira [10] investigated on online communities about transformations in the power relationship between physician and patient for her PhD thesis. This investigation mainly considered discussions fostered in online forums and an analysis of content through a discourse analysis of their content, using the Discourse of the Collective Subject technique [4].

2.2. Methods for Automatic and Intelligent Organization of Text Collections

Due to the need to extract useful knowledge from the increasing growth of online text repositories, methods for automatic and intelligent organization of text collections have received great attention in the research community. The use of topic hierarchies is one of the most popular approaches for such organization, allowing users to interactively explore the collection, guided by topics that indicate the contents of the documents available.

The extraction of topic hierarchies is based on unsupervised and semi-supervised learning methods from text collections. The hierarchical clustering strategy can be classified as agglomerative or divisive. In agglomerative hierarchical clustering, initially each document is a singleton cluster. For each of the following iterations, the closest pair of clusters is unified until they form only one cluster. In the other strategy, the divisive hierarchical clustering starts with a cluster containing all documents, which are iteratively divided into smaller clusters until there remains only a singleton cluster. Experimental evaluations show that the algorithm UPGMA (agglomerative) and the Bisecting k-means (divisive) achieve the best results in textual data [11]. However, it is noteworthy that these clustering algorithms were proposed to solve general-purpose tasks. Consequently, several studies have investigated text clustering approaches for specific applications. For instance, there are works in construction of digital libraries of patents [12], search engines [13], web mining [14], and, more recently, analysis of online communities and social networks [15].

In particular, text clustering for online forums analysis has recently gained the attention of the research community because of the need to organize automatically the huge volume of texts published daily. Most of the existing works found in the literature investigate extraction of user profiles [16], event detection from social data [17], and recommendation tasks [18]. The problem of content selection from online forums is an aspect that, to the best of our knowledge, has not been explored so far by other researchers. The proposed process adds to this body of work. This article consolidates and extends the seminal results [19] that were presented at the first Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), a satellite event of the XXXII Brazilian Computer Society Conference in July of 2012. The details of how this process attempts to solve this task are presented in the next section.

3. The Proposed Process

Here we present the proposed process in order to support analysts in exploring and selecting content from online forums. This process is divided into three steps as follows: 1) Term Co-occurrence Network, 2) Hierarchical Clustering, and 3) Post and Topics Recommendation. **Figure 1** shows the whole process at glance.

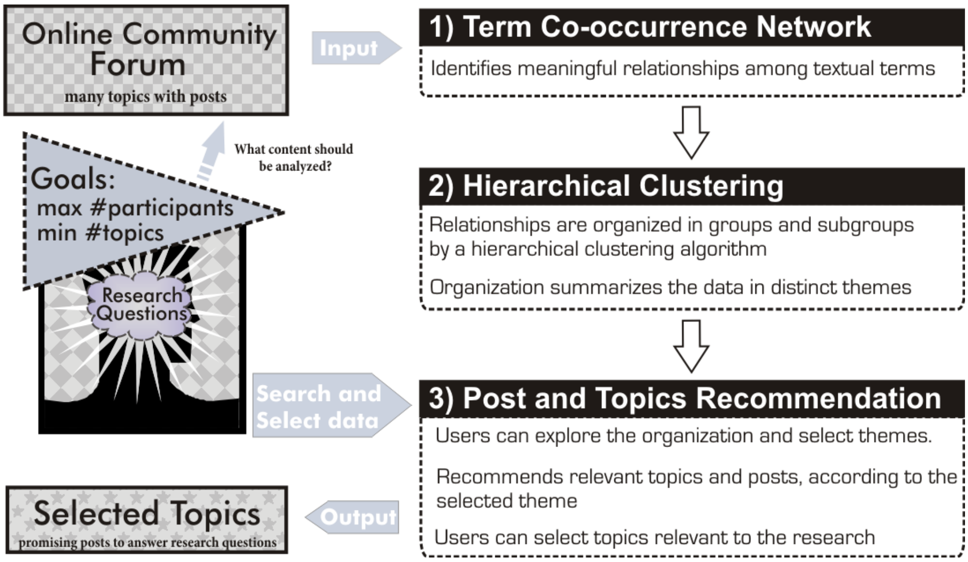

Online Community Forum
many topics with posts
Input
1) Term Co-occurrence Network
Identifies meaningful relationships among textual terms
What content should be analyzed?
Goals:
max #participants
min #topics
Research Questions
2) Hierarchical Clustering
Relationships are organized in groups and subgroups by a hierarchical clustering algorithm
Organization summarizes the data in distinct themes
Search and Select data
3) Post and Topics Recommendation
Users can explore the organization and select themes.
Recommends relevant topics and posts, according to the selected theme
Users can select topics relevant to the research
Selected Topics
promising posts to answer research questions
Output

Figure 1. **The proposed process to support analysts.**

In the first step, a term co-occurrence network is used to identify meaningful relationships among text terms. Next, in the second step, the relationships from the term co-occurrence network are organized in clusters and sub clusters by a hierarchical clustering algorithm. This organization summarizes the textual data in distinct themes. At last, in the third step, analysts can explore the organization and select themes. After selecting a theme of interest, topics and posts relevant to this theme are presented to the analysts.

In order to properly describe the proposed process, we need to define a structured text representation model, a similarity measure among documents, and a clustering strategy [20]. The vector space model is one of the most common structures for text representation. In this model, each document is represented by a vector of terms $d = \{t_1, t_2, \cdots, t_m\}$, where each term t_i has a value associated, such as word frequency, with its relevance (weight) to the document. A document cluster $G = \{d_1, d_2, \cdots, d_n\}$ also has a representation in the vector space model that is defined by the centroid C_G in Equation (1), which means the vector of all documents from G.

$$C_G = \frac{1}{|G|}\sum_{i=1}^{n} d_i \tag{1}$$

The similarity between two documents (or document clusters) represented in the vector space model is usually calculated by the cosine measure. This is shown in Equation (2). In this measure, let d_i and d_j be two documents, then the cosine angle has value 1 when the two documents are identical and value 0 when they do not share any term (orthogonal vectors). In some cases, it is useful to adapt the cosine similarity measure to a dissimilarity measure by using the equation $dis(d_i,d_j) = 1 - cos(d_i, d_j)$.

$$cos\left(d_i, d_j\right) = \frac{d_i \dot{d}_j}{\left|d_i\right|\left|d_j\right|} \tag{2}$$

The process execution has as input a text collection $\mathrm{D} = \{P_1, P_2, \cdots, P_n\}$, composed by n textual posts P, that are retrieved from an online forum. Thus, a post P_i has a representation in vector space model as text document. The representation of a topic T in the vector space model is obtained by Equation (3), where T_{set} is the set with all posts that belong to the topic T. Following next, each step of the process execution is detailed.

$$T = \frac{1}{\left|T_{set}\right|}\sum_{\forall P \in T_{set}} P \tag{3}$$

3.1. Term Co-Occurrence Network

A term co-occurrence network is defined by a graph **GRAPH** (**V**, **E**, **W**), where **V** is the vertex set, **E** is the set of edges that connect two vertices. Finally, **W** is the weight set associated to the edges, identifying the strength of the relationship.

The **vertices** are the terms in the textual collection, more specifically, terms selected to represent each document in the vector space model. The co-occurrence between two terms identifies the **edges** of the graph. For that reason, two terms are connected by an edge if there exists a meaningful co-occurrence between them. The co-occurrence between two terms is considered meaningful if the frequency of this co-occurrence is greater than a defined threshold (*i.e.* minimal frequency value).

In general, the edges' **weight** is given in numeric values used to identify the relationship strength between two terms. However, in this work, a centroid is used to identify this relationship. The centroid allows a concise representation of a document set in the vector space model. Therefore, let $e = \{t_i, t_j\}$ be an edge that connects the terms t_i and t_j, then a centroid to the edge e is defined according to Equation (4):

$$w(e) = C(t_i \cap t_j) \tag{4}$$

in which $C(t_i \cap t_j)$ is the centroid that represents the document subset (post subset) with both terms t_i and t_j. In this way, the term co-occurrence network, as applied in this work, can be seen as a structure with two main characteristics:

- Capability to identify meaningful relationships among terms from the online forum text, based on the co-occurrence frequency; and
- Capability to extract posts' subset (represented by centroids), in which the pairs of terms (edges) are used to describe these posts' content.

The term co-occurrence network is a useful structure for analyzing text collections. Furthermore, when combined with visualization and clustering tools, it allows the exploration of existent themes in the texts through an interactive user interface.

3.2. Hierarchical Clustering

The term co-occurrence network, in general, contains all relationships showing relevant co-occurrence. Thus, the goal of the hierarchical clustering from the term co-occurrence network is to summarize the existent relationships in the networks in term clusters. In the hierarchical clustering step, each pair of terms (edges), represented by its centroid, is seen as an object by the clustering algorithm. So, it is possible to use the same cosine similarity measure and traditional hierarchical clustering algorithms, like UPGMA and Bisecting K-means.

The hierarchical clustering allows the thematic organization of the posts in cluster and sub clusters, in a way that similar posts can belong to the same clusters. Analysts can visualize the information in different levels of granularity, allowing them to explore iteratively the textual content of the online forum. This thematic organization has an important role to the analysts, as it allows them to perform an exploratory search. Usually, analysts have little prior knowledge about the data from the online forum being studied, especially at the beginning of the analysis. Nevertheless, analysts can rely on the available content labeling, *i.e.*, each post cluster has a descriptor set (terms of the co-occurrence network) that contextualizes and gives a meaning to the clusters, as a guide to their search.

It is noteworthy that the thematic organization is related to the hypothesis that if an analyst is interested in a post belonging to some theme (and, consequently, the topic), he/she must be interested in other posts (and topics) on this same theme. Therefore, the theme organization provides a promising organization to find similar relevant content.

3.3. Post and Topics Recommendation

The thematic organization works as a topic taxonomy, in which analysts can select a theme of interest (among the possibilities). The selected theme is used to recommend topics and posts from the online forum to the analyst.

The topic and post recommendation is achieved through a ranking strategy. From a theme selected by the analyst, the ranking of topics and posts is computed and ordered by their relevance to this theme. The cosine similarity measure defines the relevance criteria, using the proximity value between the centroid of the theme and the vector representation of the post or topic.

In our proposed process, the topics and posts with the highest ranking are the best candidates to be selected, meaning they should possibly bear the most interesting content for the analysts. Although the organization and summarizing is an unsupervised process, only analysts know what themes are of their interest. So, the analyst must select the most relevant content to their study goals. This is what we call the problem of content selection from online forums.

The problem of content selection can be summarized as the task to find discussion in online forums in which content looks promising to answer study research questions. This problem has two distinct objectives, the first being to maximize the number of discussion participants (*i.e.* online forum users) and the second being to minimize the number of topics to be analyzed (*i.e.* content volume). It is important to say that the discussion analysis must be performed considering the context of the topic, because the analysis of one interesting post requires the comprehension of the whole discussion as presented in other posts of the topic. Therefore, the solution of the problem of content selection for analysis is a set of topics selected from the online forum.

The proposed approach aims to significantly minimize the amount of textual data required for analysis. In the next section, we shall present our software tool to illustrate how to support analysts in exploring and selecting content selection from internet forums.

4. A Tool to Support Solving the Content Selection Problem

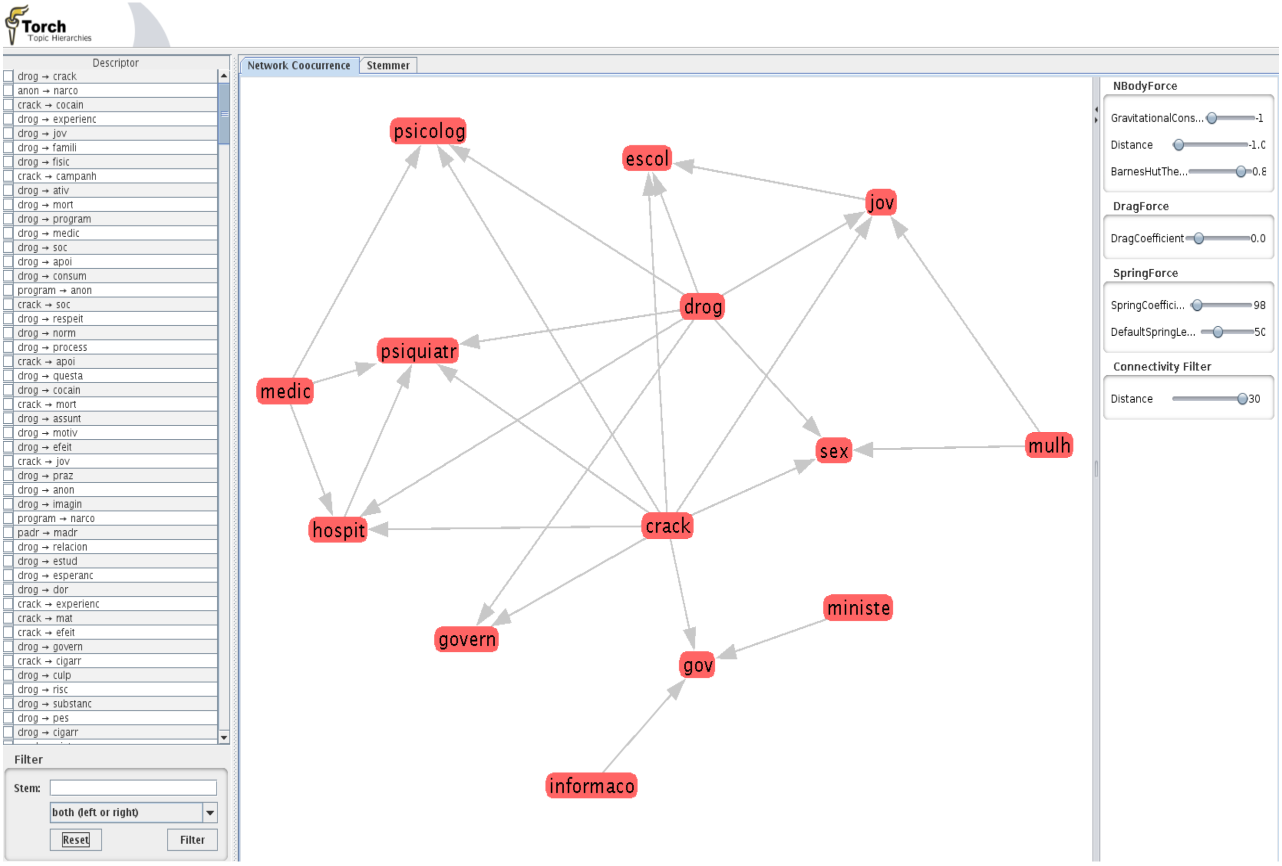
The software tool developed to support in exploring and selecting content from online forums is an extension of the Torch—Topic Hierarchies [21]. This tool provides techniques for text preprocessing and hierarchical clustering algorithms. In addition, we developed a module for recommendation of posts and topics and a GUI to explore the clustering results from topics and posts.

The posts and topics collected from the online forums have several attributes. We use a set of attributes that is common in many social networks to allow a wider application of the tool. The attributes considered for each post were the text message of the post, the publication date of the message and the post author. Each post belongs to a particular topic of the online forum and the forum has many topics. Thus, the attributes considered for representation of the topics were the topic title, its period of existence (defined by the publication date of the first and last post) and the number of participants in the topic.

After collecting a set of posts and topics, the tool performs the textual data pre-processing. The first step is the stop words removal, in which pronouns, articles and prepositions are discarded. Then, the terms are simplified by using the Porter Stemming algorithm [22,23]. Thus, morphological variations of a term are reduced to its radical. Finally, a feature selection technique based on document frequency obtains a reduced and representative subset of terms.

The term co-occurrence network obtained from pre-processed texts is the first structure available to analysts for exploring the textual content of the online forums. Our tool allows the analysts to analyze significant relationships among terms via an interactive interface. Figure 2 illustrates part of a term co-occurrence network extracted from an online forum about drugs in the Orkut® social network. This example context is explained in Section 5. It is noteworthy to say that the important relationships found in the forums' content are highlighted by the network, for example, the "crack → escol" ("crack → schools") and "drog → jov" ("drugs → youth"). Additionally, analysts can remove relationships that are not of interest to them.

Figure 3 shows the main interface of the tool created to support solving the content selection problem. The term co-occurrence network was summarized with hierarchical clustering. Thus, the various topics discussed in the forums are presented to the analysts in succinct themes (Figure 3(a)). When an analyst selects a theme, the most relevant topics (Figure 3(b)) and posts (Figure 3(c)) are presented according to the ranking strategy described in Section 3.3. The analyst can select content through check boxes at the right side of the topic or post names. The selection of a post in the lower part (Figure 3(c)) automatically includes all posts of its topic, because the comprehension of the discussion requires all posts of the topic.



Descriptor
□ drog → crack
□ anon → narco
□ crack → cocain
□ drog → experienc
□ drog → jov
□ drog → famili
□ drog → fisic
□ crack → campanh
□ drog → ativ
□ drog → mort
□ drog → program
□ drog → medic
□ drog → soc
□ drog → apoi
□ drog → consum
□ program → anon
□ crack → soc
□ drog → respeit
□ drog → norm
□ drog → process
□ crack → apoi
□ drog → questa
□ drog → cocain
□ crack → mort
□ drog → assunt
□ drog → motiv
□ drog → efeit
□ crack → jov
□ drog → praz
□ drog → anon
□ drog → imagin
□ program → narco
□ padr → madr
□ drog → relacion
□ drog → estud
□ drog → esperanc
□ drog → dor
□ crack → experienc
□ crack → mat
□ crack → efeit
□ drog → govern
□ crack → cigarr
□ drog → culp
□ drog → risc
□ drog → substanc
□ drog → pes
□ drog → cigarr

Filter

Stem:

both (left or right)

Reset Filter

NBodyForce

GravitationalCons... -1

Distance -1.0

BarnesHutThe... 0.8

DragForce

DragCoefficient 0.0

SpringForce

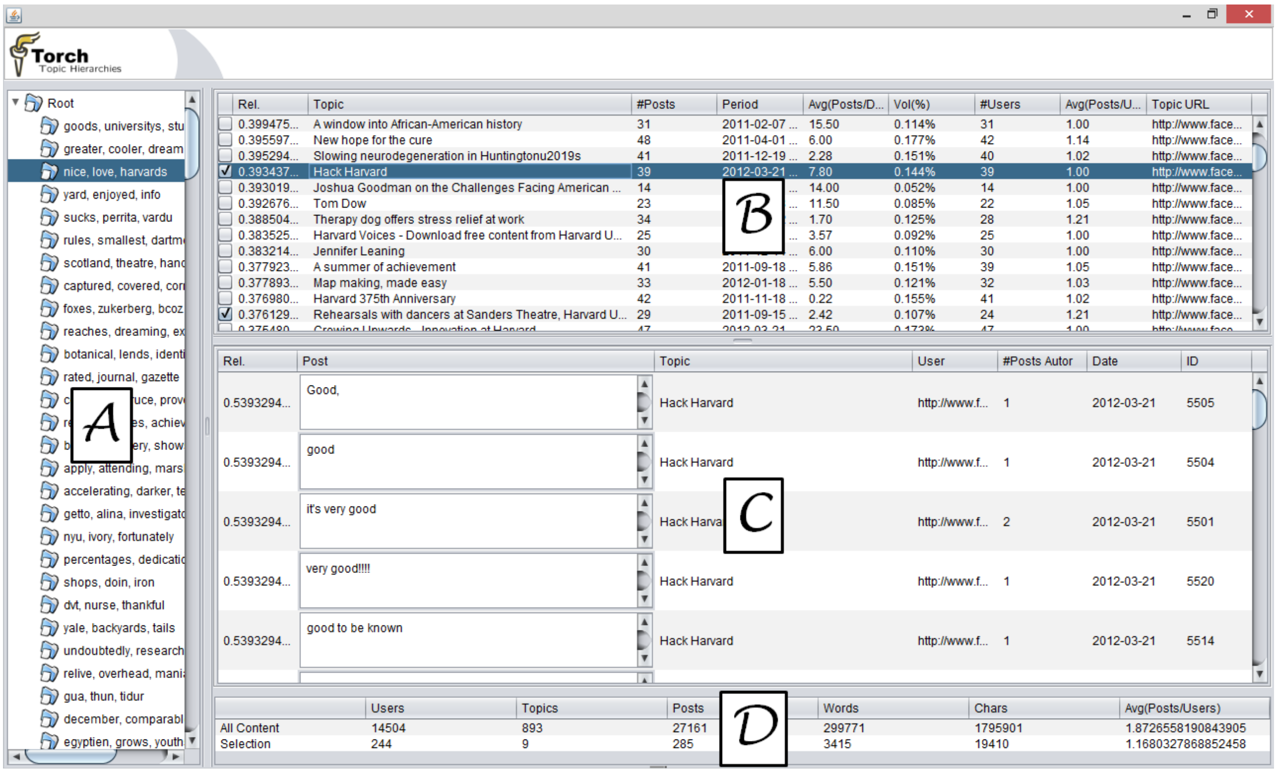
SpringCoeffici... 98

DefaultSpringLe... 50

Connectivity Filter

Distance 30

Figure 2. **Example of term co-occurrence network.**



Root
- goods, universitys, stu
- greater, cooler, dream
- nice, love, harvards
- yard, enjoyed, info
- sucks, perrita, vardu
- rules, smallest, dartm
- scotland, theatre, hanc
- captured, covered, cor
- foxes, zukerberg, bcoz
- reaches, dreaming, ex
- botanical, lends, identi
- rated, journal, gazette
- c uce, prov
- r es, achiev
- b ery, show
- apply, attending, mars
- accelerating, darker, te
- getto, alina, investigato
- nyu, ivory, fortunately
- percentages, dedicatio
- shops, doin, iron
- dvt, nurse, thankful
- yale, backyards, tails
- undoubtedly, research
- relive, overhead, mani
- gua, thun, tidur
- december, comparabl
- egyptien, grows, youth

	Rel.	Topic	#Posts	Period	Avg(Posts/D...	Vol(%)	#Users	Avg(Posts/U...	Topic URL
□	0.399475...	A window into African-American history	31	2011-02-07 ...	15.50	0.114%	31	1.00	http://www.face...
□	0.395597...	New hope for the cure	48	2011-04-01 ...	6.00	0.177%	42	1.14	http://www.face...
□	0.395294...	Slowing neurodegeneration in Huntingtonu2019s	41	2011-12-19 ...	2.28	0.151%	40	1.02	http://www.face...
☑	0.393437...	Hack Harvard	39	2012-03-21 ...	7.80	0.144%	39	1.00	http://www.face...
□	0.393019...	Joshua Goodman on the Challenges Facing American ...	14	...	14.00	0.052%	14	1.00	http://www.face...
□	0.392676...	Tom Dow	23	...	11.50	0.085%	22	1.05	http://www.face...
□	0.388504...	Therapy dog offers stress relief at work	34	...	1.70	0.125%	28	1.21	http://www.face...
□	0.383525...	Harvard Voices - Download free content from Harvard U...	25	...	3.57	0.092%	25	1.00	http://www.face...
□	0.383214...	Jennifer Leaning	30	...	6.00	0.110%	30	1.00	http://www.face...
□	0.377923...	A summer of achievement	41	2011-09-18 ...	5.86	0.151%	39	1.05	http://www.face...
□	0.377893...	Map making, made easy	33	2012-01-18 ...	5.50	0.121%	32	1.03	http://www.face...
□	0.376980...	Harvard 375th Anniversary	42	2011-11-18 ...	0.22	0.155%	41	1.02	http://www.face...
☑	0.376129...	Rehearsals with dancers at Sanders Theatre, Harvard U...	29	2011-09-15 ...	2.42	0.107%	24	1.21	http://www.face...
□	0.375480	Growing Upwards - Innovation at Harvard	47	2012-03-21	23.50	0.173%	47	1.00	http://www.face

Rel.	Post	Topic	User	#Posts Autor	Date	ID
0.5393294...	Good,	Hack Harvard	http://www.f...	1	2012-03-21	5505
0.5393294...	good	Hack Harvard	http://www.f...	1	2012-03-21	5504
0.5393294...	it's very good	Hack Harva	http://www.f...	2	2012-03-21	5501
0.5393294...	very good!!!!	Hack Harvard	http://www.f...	1	2012-03-21	5520
0.5393294...	good to be known	Hack Harvard	http://www.f...	1	2012-03-21	5514

	Users	Topics	Posts	Words	Chars	Avg(Posts/Users)
All Content	14504	893	27161	299771	1795901	1.8726558190843905
Selection	244	9	285	3415	19410	1.1680327868852458

Figure 3. Screenshot of the TorchSR in execution.

The content showed by the tool in **Figure 3** is from the Harvard[1] Facebook Page. Using the Facebook API, 27,161[2] comments (*i.e.* messages) from 893 posts (*i.e.* topics) were retrieved in June 2012. This is a famous page in Facebook, with more than 1,800,000 "likes". It is also the community in which Facebook® was created.

The problem solving progression is described by a set of measures (**Figure 3(d)**) that describe the number of participants, which should be maximized, the content volume (topics, posts, words, characters), that should be minimized, and a relationship of both measures (median of posts per participant). The first line shows the measurement considering the whole forum content and the second has the selection measurement. It is up to the analysts to decide when the current solution (*i.e.* selection) is satisfactory, and these measures help them to make this decision.

The content selection problem is a difficult problem for analysts who embrace new ventures in conducting research based on the vast content available on the Internet. The two objective goals are to maximize the number of selected participants and minimize the content volume to be analyzed. These are conflicting goals. The problem's solution is also driven by the research interests, which are not computationally measurable (so far). Without the tool, researchers rely only on the general metrics about the topics' content, or they must look at the whole forum content to perform the content reducing task. The proposed process aims to support the researchers to tackle this problem in a smarter way, leveraging them with the best machine learning techniques available so far. Although the content mining and description through metrics and models help in exploring and selecting content, the subjective goal of what is of interest to be analyzed is still a burden upon the analysts.

5. Example of a Study That Would Benefit from the Proposed Process

We present a study about motivations for drug abuse to start and cease, specifically with regard to the drug crack cocaine in Brazil. This is a case study that could benefit from the process proposed in this paper. As a result of the community content analysis, the report compiled answers the following questions: 1) what are the factors leading to crack use; 2) what are the optimal turning points to start a treatment; 3) what are abstinence maintenance factors; 4) what favors the restart of drug abuse; 5) what criticism exists for official health treatment; and 6) which kind of help are the codependents looking for.

Since the major source of social media in Brazil at the

[1] http://www.facebook.com/Harvard

[2] Despite no error reported by the Facebook Open Graph API, only part of the 52,986 comments informed were retrieved through this interface. A double check through web interface faced the same issue. Therefore, it was impossible to retrieve all messages from the Facebook Page of Harvard.

time of the study was Google's Orkut® [24], this social networking site was used as the investigation platform in this study. A search for the term "crack" in the Orkut system, considering filters for location (Brazil) and language (Portuguese), gave 995 results in September of 2011. The next step was to select communities where content would have contextualized discourses about people's experience related to drug use. Categorizing the 995 communities, it was possible to identify 278 (28%) communities related to selection objective, 360 (36%) communities that did not seem to be directly related to the selection objective, and 357 (36%) communities that were not using the term "crack" in reference to a drug (e.g. instead referred to programs and password cracking). Narrowing the research, the 278 identified online communities were filtered down to 13 communities considering others criteria such as: 1) possessing more than 300 members; 2) having been in existence for more than 6 months; 3) having exhibited recent activity; and 4) having publically available content. The last step of this stage was to choose one community for evaluation. The community "Crack, Nem Pensar—AJUDA"[3] was selected for in-depth analysis because, out of the 13 remaining communities, it is the oldest and has the most members (11,102). In a quick evaluation of its content, the community presented an intense conversation among its members, which in later analysis showed a median of 3.3 messages per day since its creation in July of 2004.

The community analysis focused on participating members who had engaged in conversation in the community forum. It is important to make this distinction, as all members have the potential to follow the discussions, but most choose not to participate (*i.e.* lurkers). This analysis is based on the postings of the participants in the forum. From the participant data available, there were 57% men and 43% women identified. At the time of the study, September 2011, the community forum had 434 participants, 384 topics and 8655 messages, representing a total of 76,646 words, or 4,515,087[4] characters. The content analysis was conducted by applying the Discourse of the Collective Subject technique [4]. Considering the great volume of data and efforts required for content analysis, a data cut was performed to focus the investigation on a suitable content analysis to study the objectives. This data cut task would benefit from the **proposed process** in this paper. From the original dataset, 39 (10% of the total) topics were selected, with 129 (30%) participants and 925 (11%) messages, totaling 107,488 (14%) words, or 602,332 (13%) characters.

The study results have been subject to discussion in a seminar organized by the Sírio-Libanês Hospital in January of 2012 in São Paulo (Brazil), with attendees from the Brazilian government, health organizations and general public. The online forum analysis mainly "*identified that the speech of dependents and codependents* (*family and friends of the dependents*) *are mingled and complete each other, therefore both require care and attention*" [3]. The reality of these people is transcript through discourse syntheses that answer to the study research questions. A compilation of the results has been submitted to a journal and is still under evaluation to the time of this article written. This example gives an indication of the hard task researchers might face when performing content analyses of online forums and the valuable outcomes that can be achieved from its analysis.

6. Conclusions

This paper introduces a process proposed to support analysts in exploring and selecting content from online forums. This process is based on unsupervised machine learning techniques like hierarchical clustering and term co-occurrence network. Consequently, analysts can explore the online forum through consolidated and structured content. This supports them in selecting interesting content to be analyzed for their research. The process creation rationale and its description are presented. As an application of the proposed process, a tool based on that process, called TorchSR, was created to aid researchers to apply it. This includes content measurements and exploration methods. An example of a real world study that could benefit from the proposed process contextualizes the process application.

The created tool is already a useful prototype. Further research in the measures used to calculate the similarities between posts can also provide better results to the user, and improve the process. Another interesting enhancement is to consider the user's feedback of what is "interesting content" to their search, so the recommendation rankings can be improved iteratively. This can be achieved by predicting the odds that a topic will be selected by the analysts, based on forum content and user interaction.

Acknowledgements

The authors would like to thank Rodrigo Pazzini for the help with his expert social media skills. This work was sponsored by CNPq (Brazilian Council for Research and Development), process 142620/2009-2, and FAPESP (State of Sao Paulo Research Foundation), process 2010/20564-8 and 2011/19850-9.

REFERENCES

[1] J. Preece and D. Maloney-Krichmar, "Online Communities: Design, Theory, and Practice," *Journal of Computer-*

[3] http://www.orkut.com/Main#Community?cmm=175318

[4] B-42 Gutenberg's Bible has around 3 million characters.

Mediated Communication, Vol. 10, No. 4, 2005, Article 1. http://jcmc.indiana.edu/vol10/issue4/preece.html

[2] R. V. Kozinets, "Netnography: Doing Ethnographic Research Online," Sage Publications Ltd., London, 2010.

[3] D. Carvalho, W. Madeira, M. Okamura, C. Lucena and S. Zanetta, "A Practical Approach to Exploit Public Data Available on the Internet to Study Healthcare Issues," *Proceedings of the XII Workshop on Medical Informatics*, *XXXII Congress of the Brazilian Computer Society Computer Society*, Curitiba, 2012.

[4] F. Lefevre and A. M. C. Lefevre, "The Collective Subject That Speaks," *Interface-Comunicação*, *Saúde*, *Educação*, Vol. 10, No. 20, 2006, pp. 517-524. http://dx.doi.org/10.1590/S1414-32832006000200017

[5] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy and M. Van Alstyne, "Computational Social Science," *Science*, Vol. 323, No. 5915, 2009, pp. 721-723. http://dx.doi.org/10.1126/science.1167742

[6] J. N. Lasker, E. D. Sogolow and R. R. Sharim, "The Role of an Online Community for People with a Rare Disease: Content Analysis of Messages Posted on a Primary Biliary Cirrhosis Mailing List," *Journal of Medical Internet Research*, Vol. 7, No. 1, 2005, p. e10. http://dx.doi.org/10.2196/jmir.7.1.e10

[7] L. C. Whitehead, "Methodological and Ethical Issues in Internet-Mediated Research in the Field of Health: An Integrated Review of the Literature," *Social Science & Medicine*, Vol. 57, No. 4, 2007, pp. 782-791. http://dx.doi.org/10.1016/j.socscimed.2007.03.005

[8] J. A. Greene, N. K. Choudhry, E. Kilabuk and W. H. Shrank, "Online Social Networking by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook," *Journal of General Internal Medicine*, Vol. 26, No. 3, 2011, pp. 287-292. http://dx.doi.org/10.1007/s11606-010-1526-3

[9] M. Berger, T. H. Wagner and L. C. Baker, "Internet Use and Stigmatized Illness," *Social Science & Medicine*, Vol. 61, No. 8, 2005, pp. 1821-1827. http://dx.doi.org/10.1016/j.socscimed.2005.03.025

[10] W. Madeira, "Transforming Needed: Changes in Power Relationships Established between Doctor and Patient," Ph.D. Thesis, Universidade de São Paulo, São Paulo, 2011.

[11] Y. Zhao, G. Karypis and U. Fayyad, "Hierarchical Clustering Algorithms for Document Datasets," *Data Mining and Knowledge Discovery*, Vol. 10, No. 2, 2005, pp. 141-168. http://dx.doi.org/10.1007/s10618-005-0361-3

[12] I.-S. Kang, S.-H. Na, J. Kim and J.-H. Lee, "Cluster-Based Patent Retrieval," *Information Processing & Management*, Vol. 43, No. 5, 2007, pp. 1173-1182. http://dx.doi.org/10.1016/j.ipm.2006.11.006

[13] C. Carpineto, S. Osiński, G. Romano and D. Weiss, "A Survey of Web Clustering Engines," *ACM Computing Surveys*, Vol. 41, No. 3, 2009, p. 17. http://dx.doi.org/10.1145/1541880.1541884

[14] B. Liu, "Unsupervised Learning," *Web Data Mining*: *Exploring Hyperlinks*, *Contents*, *and Usage Data*, 2nd Edition, Springer Berlin Heidelberg, 2011, pp. 133-166. http://dx.doi.org/10.1007/978-3-642-19460-3_4

[15] C. Kadushin, "Understanding Social Networks: Theories, Concepts, and Findings," Oxford University Press, Oxford, 2012.

[16] M. Shmueli-Scheuer, H. Roitman, D. Carmel, Y. Mass and D. Konopnicki, "Extracting User Profiles from Large Scale Data," *Proceedings of the* 2010 *Workshop on Massive Data Analytics on the Cloud*, Article no. 4, ACM, New York, 2010.

[17] Q. Zhao, P. Mitra and B. Chen, "Temporal and Information Flow Based Event Detection from Social Text Streams," *Proceedings of the* 22*nd National Conference on Artificial Intelligence*, *AAAI*'07, Vol. 2, AAAI Press, Vancouver, 2007, pp. 1501-1506.

[18] E. Davoodi, M. Afsharchi and K. Kianmehr, "A Social Networkbased Approach to Expert Recommendation System," *Proceedings of the* 7*th International Conference on Hybrid Artificial Intelligent Systems* (*HAIS*'12), Vol. Part I, Springer Berlin Heidelberg, 2012, pp. 91-102.

[19] D. B. F. Carvalho, R. M. Marcacini, C. J. P. Lucena and S. O. Rezende, "Towards a Process to Support Solving the Content Selection Problem from Online Community Forums," *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*, *XXXII Congress of the Brazilian Computer Society Computer Society*, Curitiba, 2012.

[20] C. C. Aggarwal and C. Zhai, "A Survey of Text Clustering Algorithms," In: C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*, Springer, US, 2012, pp. 77-128. http://dx.doi.org/10.1007/978-1-4614-3223-4_4

[21] M. Marcacini and S. O. Rezende, "Torch: A Tool for Building Topic Hierarchies from Growing Text Collection," *Proceeding of WTA*'2010: *IX Workshop on Tools and Applications*, 8th Brazilian Symposium on Multimedia and the Web (Webmedia), Belo Horizonte, 2010, pp. 133-135.

[22] M. Porter, "The Porter Stemming Algorithm," 2009. http://www.tartarus.org/martin/PorterStemmer

[23] B. M. Nogueira, M. F. Moura, M. S. Conrado, R. G. Rossi, R. M. Marcacini and S. O. Rezende, "Winning Some of the Document Preprocessing Challenges in a Text Mining Process," *Proceedings of IV Workshop on Algorithms and Data Mining Applications*, *XXIV Brazilian Symposium on Database*, Porto Alegre, 2008, pp. 10-18.

[24] A. Banks, "State of the Internet in Brazil," 2011. http://www.comscore.com/Insights/Presentations_and_Whitepapers/2011/State_of_the_Internet_in_Brazil