IV Supporting tools to the proposed process

This chapter presents two tools created in order to support the process presented in the previous chapter. The first one generates the Community Association Map, which is based on a process aiming to support experts in the understanding of users' interests. The second tool, TorchSR, supports users to find interesting content in online forums for discussion analysis.

IV.1 Community Association Map

Community Association Map is an analytical tool based on a process aiming to support experts in the understanding of users' interests based on their associations within their communities, for the benefit of the online community content analysis. The process goal is to reveal the interests of users through a map of associated communities, which provide information that helps to characterize the population involved in the community under analysis. The user membership is utilized to establish the relationship among communities. This map, called Community Association Map, shows the interests of members from a specific community in other communities. The research study that motivated the process development is available in the technical report [Car10], and the results presented here has been published in the Proceedings of the 8th International Conference on Web Information Systems and Technologies (WEBIST2012) [CAM12].

The proposed process focuses on user membership within communities. Membership could be seen as the user interest in the community topic. Whenever a user is member of a specific community, one can infer a strong interest of the user to the community topic. Consequently, a map of associated communities reveals the interests of community members to other topics. This map represents a model of community relationships.

Models and measurements help experts to achieve the complete social media analysis, which also considers other elements associated with their study. The process relies on a model of related communities based on users' membership. The user membership of two communities establishes a relationship between these communities. Therefore, a community relates to another if one user is a member of both communities. The purpose of this model is an attempt to reveal users' interests in other topics by associating their communities. A key point for measurement considered in the process development is the use of appealing visualization for its outcome. To achieve this goal, the process relies on a software package for visualizing data and information. A plotted graph reveals the strongest affinities by weight of the connecting lines linking communities. These graphs help experts in discovering trends that will further their understanding in specific social matters.

Although similar approaches are found in literature, such as the feature presented on Orkut system to provide up to nine related communities on a community page [Che09], not all data required by other methods are available to be used. Moreover, experts require more suitable tools for helping them in the analysis of social media, such as the one presented here.

(a) Processing the Community Association Map

The process has three steps: data gathering, model and measurement, and visualization. The first step, data gathering, obtains the social media data from online community sites. Next is the model and measurement step, as discussed in the previous section. The collected data is processed and organized in an appropriate format to plot the data visualization. The visualization step is responsible for displaying data in a more appealing way, easing the work of the experts. Details of the steps to build the Community Association Map (CAM) are presented as follows.

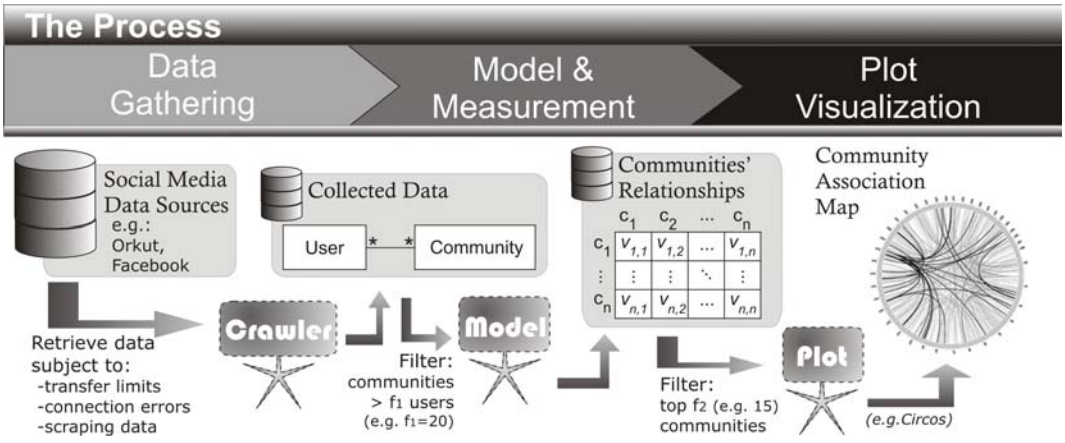

The Process
Data Gathering
Model & Measurement
Plot Visualization
Social Media Data Sources
e.g.: Orkut, Facebook
Collected Data
User
*
*
Community
Communities' Relationships
c_1 c_2 ... c_n
c_1 $v_{1,1}$ $v_{1,2}$ ... $v_{1,n}$
c_n $v_{n,1}$ $v_{n,2}$ ... $v_{n,n}$
Community Association Map
Retrieve data subject to:
-transfer limits
-connection errors
-scraping data
Crawler
Filter: communities > f1 users (e.g. f1=20)
Model
Filter: top f2 (e.g. 15) communities
Plot
(e.g.Circos)

Figure IV.1: The process that generates the Community Association Map.

1) Data gathering

Collect data from an online community site (social media data source). The collection step requires acquiring all user information from an isolated community of interest. Then, for each user, membership information for other communities is obtained. The gathered data set is a collection of users, communities, and the relationships among them.

2) Model and measurement

Create the model of communities' relationships. In a formal notation, if $user_x$ is a member of community cmm_a and cmm_b then cmm_a has a relationship with cmm_b. If $user_y$ is also a member of cmm_a and cmm_b, then the relationship is reinforced (+1). Therefore, applying this rule over a given list of users and their membership information builds the model with the weighted relationship among the communities.

The most important measurement obtained from this model is a list of the top relationships within the communities (strongest ties), which means the communities with more users in common. This measure is exposed through the community association map, which is displayed in the next step, plot.

Depending on the data available, the model size can cause problems in the process application. To overcome eventual processing limitations, there are filters for the communities considered in the model. For instance, one can choose to model only relationships among communities with more than 10 members. Applying this filter, the model will keep its features, because the modeling goal is to identify the most related communities.

Filters can also be applied to change the focus of the measurement, for instance, taking into account only communities of some specific category. An extended discussion about filtering communities by number of members or category is available in [Car10]. An example of filtering by number of members is in Section V.1.

3) Plot the visualization

The final step is to plot the community association map. Given that the map is used to reveal the most interesting associations among communities, it is useful to limit the number of communities presented in the final graph.

The outcome of this process is to support experts in the study of social media. For this purpose, a software package for visualizing is used to display the table containing the relationship among communities. Visualization tools like heat map tables are helpful in this way. The chosen one was Circos, which plots graphs that visualize data in a circular layout – making it ideal for exploring relationships between objects or positions [Krz09].

Figure IV.1 gives a glance at the whole process application. At the top of the image are the three steps of the general process in a sequence. Below the steps, corresponding with each one, there is a diagram with the data flow from the social media source to the final plot of the community association map. In between, there are agents representing the software that manipulates the data, transforming it into suitable formats. Results of the process application are presented in Chapter V.

(b) Discussion

A promising way to study social media is to focus on the analysis of the online communities' discussions. Identifying the users' interests in these online communities is an important part of this study. Specialists can apply the process presented here to reveal these interests and search for textual clues in the discussions to corroborate their findings, or vice-versa. The process generates an element of analysis that shows the interests of users through a map of associated communities as part of a study on social media.

The process application is limited to situations in which users are also members of other communities. Consequently, there must exist many other communities being hosted by the same social media site of the community under analysis, or a way to identify links between users from the communities under analysis to other communities in different social media sites. For instance, the SIOC (Semantically-Interlinked Online Communities), an initiative that aims to create and leverage a layer of semantic data in online communities [Bre06], could be used to map these communities over different sites.

The advantage of the process is to provide information about users' interests of an online community. It fleshes out users' interests based on the processing of all these users' associations to others communities that they belong to. Although it is up to the experts to choose which users and communities are interesting for the analysis, there are several filters like those shown in the previous section. The aforementioned example gives clues of how one can place filters on the data set to reveal specific interests, such as considering only active users that are more relevant to the study. An evaluation of the process results is achieved by comparing them to other elements of analysis (*e.g.* discourse analysis) as proposed by Netnography [Koz09].

According to Cormode et al. [Cor10], the process is characterized as follows. The data collection can be retrieved by connecting to the social media site API, or scraping-based when the API is not available, which was the case in the examples described in this chapter. The sampling methodology is based on an exhaustive crawling within a defined boundary, with a starting point in a community of interest. The measurement efforts focus on showing the weighted relationships among the communities through a map. An evaluation of process results is achieved by comparing them to other elements of analysis (*e.g.* discourse analysis).

A considerable step for advancing this work is to minimize process application restrictions. Instead of relying on explicit user membership information, other ways of community detection like finding interests on exchanged messages could be applied to determine community association data. In the same line of reasoning adding text-mining capabilities could further the extraction of more information from social media.

IV.2 Tackling the content selection problem

Online forums have an increasingly huge amount of content to be analyzed. A technique that helps to perform the content analysis is the Discourse of the Collective Subject, a qualitative technique with roots in the Theory of Social Representations [Lef06]. The aim of this technique is to identify groups, aggregated by central ideas, and describe them through a created discourse based on a patchwork from the collective members' speeches, synthesizing the discourse as one collective subject. However, this technique and similar ones (*i.e.* qualitative analysis) demand increasing efforts proportional to the amount of content for analysis. Therefore, the content selection is a pivotal problem in considering content available in online forums for qualitative analysis.

In order to support users to find relevant content for qualitative analysis in online forums, we propose a process based on an automatic and intelligent organization of the content in topic hierarchies. This process relies on the machinery available Torch [Mar10], tailoring its application in such way to deal with characteristics of the content selection problem. Some initial results of this research was 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 [Car12b]. The details of how this process attempts to tackle this task are presented in this section.

(a) The problem of content selection in online forum

The problem of content selection can be summarized as the task of finding discussions in online community forums in which content looks promising to answer the research questions of a study. This data cut must also consider resource constraints, such as team and time, because qualitative analysis of "big data" usually requires effort beyond the resources available for the study. Therefore, this task can be defined as a problem with two distinct objectives. Firstly, researchers want to maximize the number of discussion participants (*i.e.* users that posted messages). Secondly, they want to minimize the number of topics to be analyzed (*i.e.* relevant content volume).

In the definition of this problem, an online forum is a set of topics with a set of posts (*i.e.* messages). Figure IV.2 outlines the problem's main entities (Forum, Topic, Post) with their relevant attributes (User, Content, Date). This general model of online forum might be mapped from another online forums representations, such as those of major social networking sites like Orkut and Facebook as presented in Figure IV.2.

It is important that the discussion analysis considers the context of the topic, because the analysis of an interesting post requires the comprehension of the whole discussion as presented in other posts of its topic. Therefore, the solution of the problem of content selection for analysis is a set of topics selected, containing post messages that are interesting for a study, from the forum of an online community.

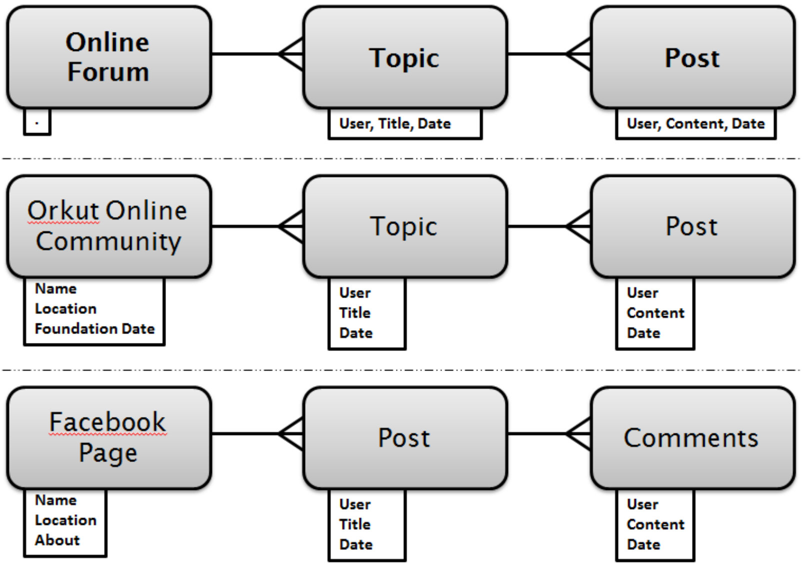


Figure IV.2: The considered entities of an online forum and examples of its equivalents in the social networking sites Orkut and Facebook.

Solving strategies and support tools

The content selection problem is a judgment task that has many solving strategies and solutions too. Two different strategies employed in research cases (see Chapter V) are presented below.

In the first research study about Hepatitis C, the analysts chose only one of all the 1284 topics. The chosen topic was the second most active, with 589 messages (3% of total), 88 participants (19%), and very interesting content for the research objectives — the goal of this topic is to simulate a confession. This was an easy and trivial solution, according to the analysts' evaluation, and a desirable one too, because it had valuable content for the study. The other topics were irrelavent consisting mostly of small talk. Even relevant discussion would be added for analysis, the effort required to go further in this selection sum-up with the already overload effort required from the specialists team to analyze the already selected data was the decision breaker at this point.

The difficult aspects of the problem of content selection showed up in the second research study. In this case, none of the 384 topics is representative enough for the study. It required a set of topics. The broad scope of the research goal to study the motivations about starting and ceasing of drug abuse was the first difficulty faced by the analysts. At that time, their judgment skills were limited due to simple computational support – they relied only on a spreadsheet with the content of 8,655 messages and measures about the topics and authors. In this context, the first selection consisted of 116 (30%)topics with 4,943 (57%) messages. This content selection had almost 3 million characters, whose amount is equivalent to Bible B42 of Guttenberg. This measurement was only possible later on with the development of scripts to measure the selected content (*i.e.* counting characters). Considering time constraints and empowered by the new measures provided by scripts, the analysts refined their selection to 39 (10%) topics (37 were among the first 116 selected) with 129 (30%) participants and 925 (11%) messages, totaling 107.488 (14%) words, or 602,332 (13%) characters. In the analysts' opinion, these measures enabled them to properly tackle the task to select content to be analyzed.

(b) Methodological background of 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. In the first step, a term co-occurrence network is used to identify meaningful relationships among text terms. In the second step, the relationships from the term co-occurrence network are organized in clusters and subclusters by a hierarchical clustering algorithm. This organization summarizes the textual data in distinct themes. The first two steps rely on the machinery already available on Torch [Mar10]. In the third and final step, users can explore the organization and select themes. After selecting a theme of interest, topics and posts, relevant to this theme, are presented to the user. Figure IV.3 shows the whole process at a glance.

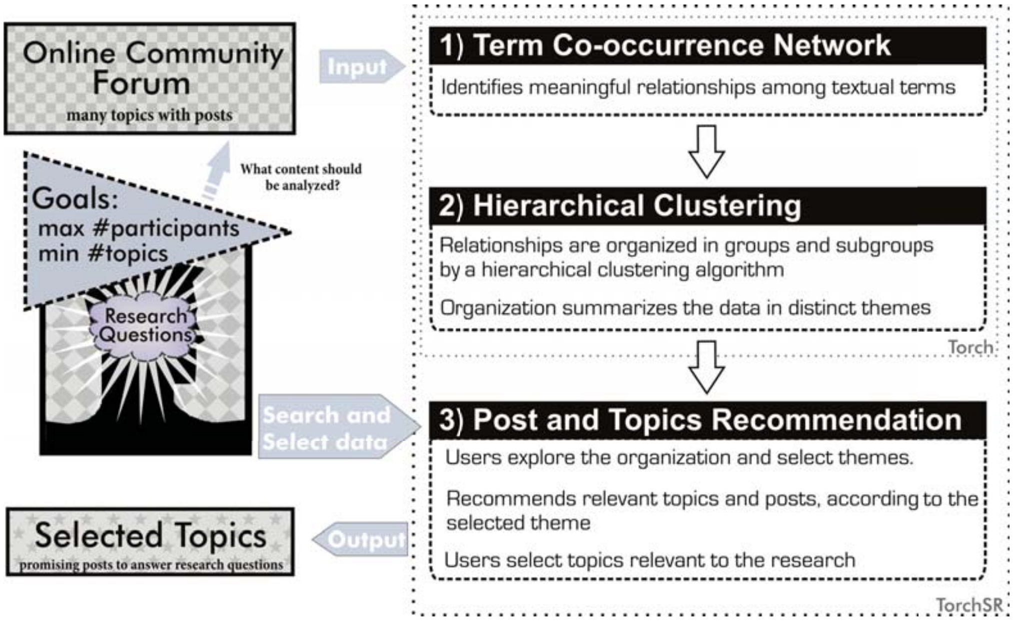

Online Community Forum
many topics with posts
What content should be analyzed?
Goals:
max #participants
min #topics
Research Questions
Input
Search and Select data
Output
Selected Topics
promising posts to answer research questions
1) Term Co-occurrence Network
Identifies meaningful relationships among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

Figure IV.3: The proposed process to tackle the content selection problem.

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 [Agg12]. The vector space model is one of the most common structures for text representation [Agg12]. In this model, each document is represented by a vector of terms $x = \{t_1, t_2, ..., t_m\}$, where each term i has a value t_i associated to its relevance (weight) to the document. In this work, t_i is the frequency value in the document. A document cluster $G = \{x_1, x_2, ..., x_n\}$ also has a representation in the vector space model, which is defined by a centroid $C_G = (\sum_{i=0}^{n} x_i)/|G|$, i.e., the mean vector of all documents from G.

The similarity between two documents (or document clusters) represented in the vector space model is usually calculated by the cosine measure. In this measure, let x_i and x_j be two documents, then the cosine $cos(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}$ 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 $dis(x_i, x_j) = 1 - cos(x_i, x_j)$.

The process execution has as input a text collection $D = \{P_1, P_2, ..., P_n\}$, composed by n posts P, that are retrieved from an online community forum. A post P has a representation in vector space model. The representation of a topic T in the vector space model is obtained by Equation 1, where T_{set} is a set with all posts that belong to the topic T. Following that, each step of the process execution is detailed.

$$T = \frac{1}{|T_{set}|} \sum_{\forall P \in T_{set}} P \qquad (1)$$

Term Co-occurrence Network

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

The **vertexes** 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 is 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** are 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 2:

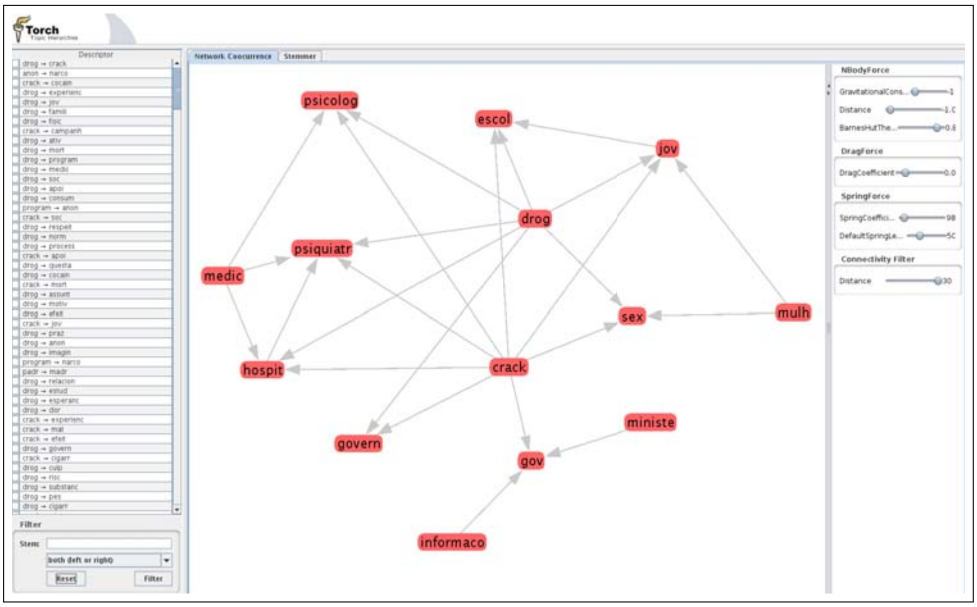
$$w(e) = C(t_i \cap t_j) \qquad (2)$$

in which $C(t_i \cap t_j)$ is the centroid that represents the document 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:

1. Capability to identify meaningful relationships among terms from the online community text, based on the co-occurrence frequency; and

2. Capability to extract posts' subset (represented by centroids), in which the pairs of terms (edges) are used to describe the content of these posts.

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.



Descriptor

drog → crack
anon → narco
crack → cocain
drog → experienc
drog → jov
drog → famili
drog → foic
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 → assum
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 → mal
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.C

BarnesHutThe... 0.E

DragForce

DragCoefficient 0.0

SpringForce

SpringCoeffici... 98

DefaultSpringLe... 5C

Connectivity Filter

Distance 30

Figure IV.4: Example of a term co-ocurrence network.

Hierarchical Clustering

The term co-occurrence network, in general, contains all relationships that have 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 of the 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. Thus, it is possible to use the same cosine similarity measure and traditional hierarchical clustering algorithms, like UPGMA and Bisecting K-means. This is the fundamental technique considered in Torch [Mar10].

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. The user can visualize the information at different levels of granularity allowing them to explore iteratively the textual content of the online community. This thematic organization has an important role for the users, as it allows them to perform an exploratory search. Usually, users have little prior knowledge about the data from online communities, especially at the beginning of an analysis. Nevertheless, users can rely on the available content labeling, *i.e.*, each post cluster has a descriptor set (terms of the co-occurrence network) that contextualize and give meaning to the clusters, as a guide to their search.

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

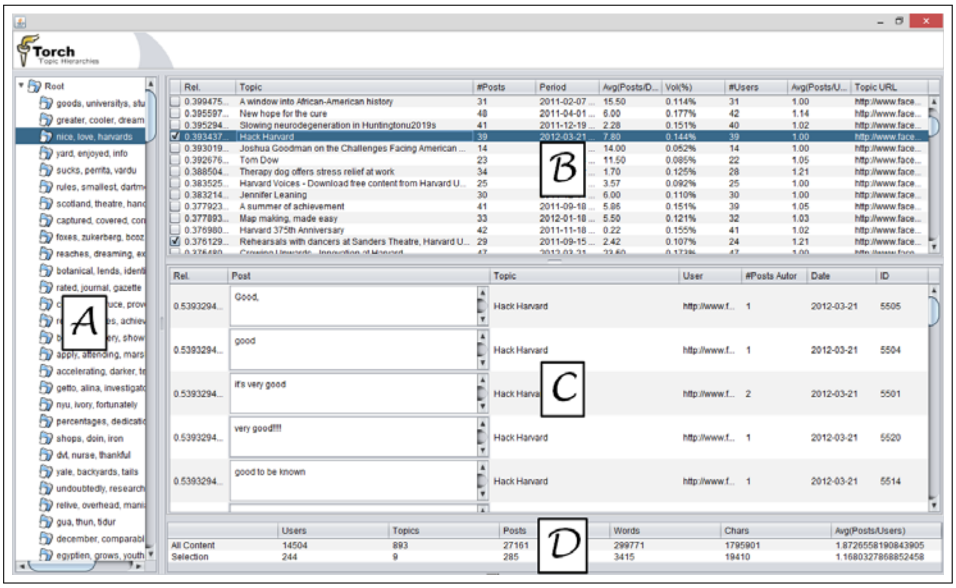
Post and Topics Recommendation

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

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

In our proposed process, the topics and posts with the highest ranking are the best candidates to be selected, meaning they should have the most interesting content for the users. Although the organization and summarizing is an unsupervised process, the users have the important task to choose the themes of interest and, then, to select the most relevant content for their goals. This approach significantly minimizes the amount of textual data required to be analyzed. In the next section, we present our software tool to illustrate how to help users in the problem of content selection from online communities.

(c) TorchSR: tackling the content selection problem



Root
- goods, universitys, stu
- greater, cooler, dream
- nice, love, harvards
- yard, enjoyed, info
- sucks, perrita, vardu
- rules, smallest, dartm
- scotland, theatre, hanc
- captured, covered, con
- 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, tr
- 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...
□	0.376480...	Growing Upwards - Innovation at Harvard	47	2012-03-21 ...	23.60	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 IV.5: Screenshot of the tool in execution.

The software tool developed to support the content selection from online communities is an extension of the Torch - **To**pic Hiera**rch**ies [Mar10]. Since it is an extension created to support in **s**ocial **r**esearch, we call this tool TorchSR. This tool provides techniques for text pre-processing and hierarchical clustering, such as available in Torch. In addition, many functionalities were added to support users to deal with the content selection problem (*i.e.* step 3 of the proposed process). A customized module for posts and topics recommendation was developed with a whole new visual interface to explore the topic clustering results from topics and posts. Users are able to select topics and obtain content assessment through metrics regarding the available content (*i.e.* topics and messages) and the selected one. More details about these functionalities are presented as follows.

The content presentation is based on several attributes of the posts and topics collected from the online forums. These attributes were proposed together with social media experts during a research study development as an extension of the measures defined by Madeira [Mad11]. 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 in the message of the post, the publication date of the message and the post author. Each post belongs to a particular topic of the online community forum and the forum has many topics. Thus, the attributes considered for representation of the topics were the topic title, the period of existence (defined by the publication date of the first and last post) and the number of participants.

After collecting a set of posts and topics, the tool performs the textual data pre-processing. The first step is the stopword removal where pronouns, articles and prepositions are discarded. Then, the terms are simplified by using the Porter Stemming algorithm [Por09, Nog08]. 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 preprocessed texts is the first structure available for the users to explore the textual content of the online communities. Our tool allows the users to analyze significant relationships among terms through an interactive interface. Figure IV.4 illustrates part of a term co-occurrence network extracted from an online community forum about drugs in the Orkut social network. It is noteworthy to say that the important relationships found in the content of the forums are highlighted by the network, for example, the "crack → escol" ("crack in schools") and "drog → jov" ("young people and drugs"). Moreover, users can remove relationships that are not of interest to them.

Figure IV.5 shows the main interface of the tool created to support in 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 user in succinct themes (Figure IV.5A). When the users select a theme, the most relevant topics (Figure IV.5B) and posts (Figure IV.5C) are presented according to the ranking strategy described in Section IV.2(b). The user can select content through check boxes at the left side of the topic or post names. The selection of a post in the lower part (Figure IV.5C) automatically includes all posts of its topic, because the comprehension of the discussion requires all posts of the topic.

The problem solving progression is described by a set of measures (Figure IV.5D) that describe the number of participants, that 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 users 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 social scientists that embrace new ventures in conducting research based on the vast content available on the Internet. It has two objective goals, maximize the number of selected participants and minimize the content volume to be analyzed. These are conflicting goals. The solution is mainly driven by the research interests, which are not measurable (so far). Without the proposed tool, the researchers rely only on the general measures about the content of the topics, or they must scrutinize manually 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 measures and models help in solving the problem, the subjective goal to identify relevant content to be analyzed is still a burden for researchers.

(d) Evaluation of TorchSR

In order to evaluate if the proposed tool, TorchSR, accomplishes its goal to support and help analysts solving the content selection problem, a twofold evaluation was conducted in an exploratory case study. An evaluation instrument based on the task-technology fit was created to verify if TorchSR (*i.e.* technology) supports and improves the resolution of the content selection problem (*i.e.* task). Furthermore, another evaluation verifies if the proposed process really improves a simpler tool as used in the exploratory research to support users to undertake the task.

Task-technology fit theory

User evaluations of information systems are valuable resources to assess if these systems accomplish their purposes. Especially in the case of the evaluation of a tool created to support users in solving a problem, an evaluation would test if it helps users instead of misleading them, that is, if it tackles the problem of content selection. The Task-Technology Fit (TTF) instruments might be used to assess systems like TorchSR [Bre12, Goo95]. These instruments are conceptually based on the task-technology fit theory. In this theory, the correspondence between information system functionality and task requirements leads to positive user evaluations, and positive performance impacts.

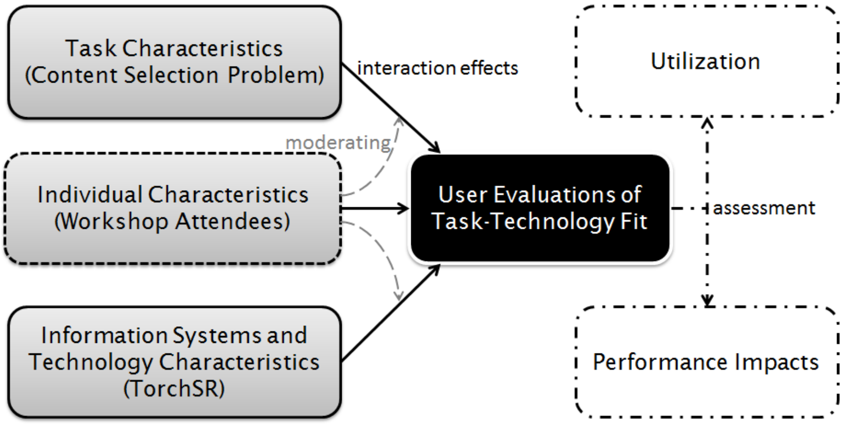


Figure IV.6: The basic model of the task-technology fit with the interaction effects, its moderating factor, and assessments.

Although there are many user evaluation methods that would be used in this evaluation, a special characteristic of the TTF regarding other user satisfaction evaluations is that user's "satisfaction" (fit between personal needs and benefits of using a system) would be most appropriately measured by assessing how a user feels about his/her system [Goo95]. On the other hand, task-technology fit would be most appropriately measured by assessing the users' beliefs of how satisfactorily systems meet task needs, regardless of how the user might feel about the systems. Further, performance benefits are better assessed from TTF beliefs than "user feelings" towards the systems. The basic model of the task-technology fit is presented in Figure IV.6, with its interaction effect being moderated by the individual user's characteristics. The result of the TTF evaluation is an assessment of the system (*i.e.* TorchSR) utilization and performance impact (*i.e.* positive or negative).

The seven dimensions considered in the TTF instrument to evaluate the TorchSR are based on the work from Goodhue [Goo95], here called referential TTF (RTTF). Consequently, the questionnaire is an adaptation of the RTTF considering this specific task and tool under evaluation. Nine dimensions of the sixteen proposed in the RTTF were discarded because they does not fit in this the context. These changes are part of experiment design in order to be useful and make sense in specific contexts. For example, from the original 16 dimensions proposed in the RTTF, only 12 were considered as valid and useful for the final instrument defined to evaluate enterprise information systems. For instance, the dimensions The Right Data and The Right Level of Detail in the RTTF faced user-understanding problems and due validity issue. Though, The Right Data dimension was removed from the final questionnaire in the RTTF. Bearing in mind the task definition and the tool implementation in the context of this research, which provides all original content and is an evaluation prototype, the nine dimensions The Right Data, Accuracy and Compatibility, Flexibility, Assistance, System Reliability, Currency, Training, and Authorization are not used in this evaluation. The remaining seven dimensions were used in this evaluation and are presented below with the statements used to evaluate each of them. The TTF questionnaire is composed by these statements presented in random order and 7-point scale from *Yes* to *No*.

- D1 - The right level of detail (maintaining the right data at the right level or levels of details)

  Sufficiently detailed data is available to accomplish the task.

  The system provides data at an appropriate level of detail for my purpose.

- D2 - Locatability (ease of determining what data is available and where)

  It is easy to locate interesting messages on a particular topic, even if that message was not seen before.

  It is easy to find out what messages are in the forum on a given subject.

- D3 - Accessibility (ease of access to desired data)

  I can get data quickly and easily when I need it.

  It is easy to get access to data that I need.

- D4 - Meaning (ease of determining what a data element in the screen means)

  In the system, the exact meaning of data elements is either obvious, or easy to find out.

The exact definition of data fields relating to my tasks is easy to find out.

- D5 - Ease of Use (ease of doing what I want to do using the system for accessing and analyzing data)

  It is easy to learn how to use the system.

  The system is convenient and easy to use.

- D6 - Presentation

  The data that I need is displayed in a readable and understandable form.

  The data is presented in a readable and useful format.

- D7 - Confusion

  There are so many different accesses to data, each with slightly different behavior, that it is hard to understand which one to use in a given situation.

  The data is stored in so many different places and in so many forms; it is hard to know how to use it effectively.

Exploratory case study

Besides the evaluation of TorchSR by an instrument based on the task-technology fit (TTF) as shown before, another evaluation is proposed with the following goal: assess if the hierarchical clustering supports users to tackle the task. In order to achieve this goal, we organized workshops of 4 hours in which attendees received training in qualitative analysis research and could practice the task of content selection from an online forum as follows.

For evaluation purposes, we considered a tool that provides measures of the content and selection. This tool is similar to TorchSR, but without the hierarchical clustering component (Figure IV.5A). This is an improved tool, which reflects the same functionalities available for the analysts in the second research study and was evaluated by them as OK (see Section IV.2(a)). TorchSR, which was presented in the previous section, is the tool being evaluated in the experiment.

Two data sets were considered for the evaluation. The first data set is the same used in the research study about drug abuse(see Section V.2), an online forum of Orkut Community “Crack, Nem Pensar - AJUDA”. Considering all content since its creation in July 16 of 2004 to September 2011, this data set has 434 participants, 384 topics and 8,655 messages, representing a total of 76,646 words, or 4,515,087 characters. Because of the research goals, this data set is in Portuguese. The second data set considered for the evaluation is the Facebook Page of Harvard (`http://www.facebook.com/Harvard`). Using the Facebook API, 27,161[1] 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 research question proposed for the evaluation is to analyze why Harvard is a good University; why students want to study there; and what are the main issues regarding studying there.

The evaluation process consists of four parts, which is illustrated in Figure IV.7. The first part, introduction, starts with an explanation of 90 minutes about the qualitative research field, with many examples of studies, and presenting the content selection problem. In this part, the experimental procedure is also explained. A pre-interview intends to verify if the attendee understands and is capable of tackling the task. For the data set assignment, half of the attendees that claim a better knowledge of English is assigned to the Harvard Page data set and the other half goes on with the data set in Portuguese. The second and the third parts are the interaction of attendees with one of the tools, which is randomly chosen - one tool per part. The attendees have around 40 minutes of interaction with each tool to solve the content selection problem, but they were able to stop the practice at any time. At the end of each interaction, attendees have to fill in the TTF evaluation questionnaire. The last part is a post interview in which subjects answer which tool is their favorite, why and what features they would like to be available.

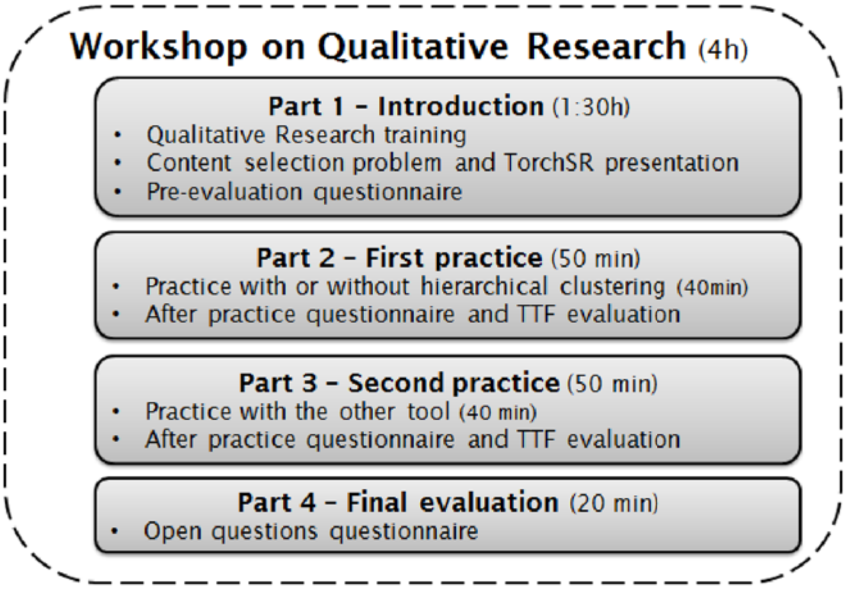
The multiple choice questions considered in the evaluation have answers ranging from Yes(1) to No(7). Some of them asked for explanations of the choice, such as: "The task of selecting content from online forums for analysis was understood? (If not, why?)." The open questions of the last part are presented below. The last question of this questionnaire was meant to provide a practical conclusion for the workshop practice.

Part 4 - Final evaluation

1. What is your assessment of the effort needed for a research to carry out the proposed task? Explain.

2. What are the main difficulties that may be encountered when performing these tasks?

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



Workshop on Qualitative Research (4h)

Part 1 – Introduction (1:30h)
- Qualitative Research training
- Content selection problem and TorchSR presentation
- Pre-evaluation questionnaire

Part 2 – First practice (50 min)
- Practice with or without hierarchical clustering (40min)
- After practice questionnaire and TTF evaluation

Part 3 – Second practice (50 min)
- Practice with the other tool (40 min)
- After practice questionnaire and TTF evaluation

Part 4 – Final evaluation (20 min)
- Open questions questionnaire

Figure IV.7: The conducted exploratory study case.

3. Which tool has helped you accomplish better selection of content? Explain.

4. What other data would you like to be available in the tools? Explain.

5. What other features would you like to be available in the tools? Explain.

6. Give an (approximate) answer to the research questions based on your best analysis.

The selected people for the evaluation were potential end-user analysts. An analyst is anyone with a college degree in any field, who understands the problem, and feels capable of solving it. The workshop attendees were considered in this user's evaluation. The workshop was realized twice in November of 2012 having in total eight software engineering researchers from the Software Engineering Laboratory of PUC-Rio as attendees interested in learning more about qualitative research. They were six men and two women, two having PhD in Computer Science (CS), five PhD Students in CS and one Master Student also in Computer Science. They had no prior knowledge of qualitative research or were familiar with the subject of the topics of the experimental data sets. However, their computer science background grants them the skills that facilitate learning new computation tools and allow a better assessment of software systems.

Results and Discussion

In the first part of the evaluation process, the question regarding whether the task of selecting content from online forums for analysis was understood and had positive results. It scored 1.5 in median, with the arithmetic average (avg) of 1.5 and standard deviation (sd) of 0.7. The next question about their evaluation of someone with their training skills being able to accomplish the task also had positive results with 2 in median (avg 2, sd 0.7). It means that they think they had the required skills to perform the task. An attendee pointed out that their background in computer science did not help in performing the task. Regarding language skills, one attendee has Spanish as her native language and the others had Portuguese. The Spanish speaker and another three most skilled in English had been assigned with the Harvard Dataset and the others had the Drug Abuse Dataset. Two attendees started with the simple tool (without the clustering hierarchies) for the first part of the practice and the other six attendees started with TorchSR.

After the first practice (second part), the answer to the question whether the initial explanation of the task of content selection was enough for understanding had positive answers with 2 in median (avg 2.1, sd 0.8). Attendees complained about the few examples used to explain the task and how to solve it using the tool. This lack of explanation was intentional because it was part of the evaluation to assess the ease of use. This considers if it is easy to learn how to use the tool. The following question verified whether attendees felt able to evaluate the task performance, particularly, about the process of content selection from the forum supported by the tool, after the practice. Despite the positive answers with median of 2 (avg 2.5, sd 2.1), three of the eight attendees answered a full Yes-1, but one attendee answered a full No-7, and the other attendees were in between. The attendee with full No-7 answer changed his opinion in the following experimental part and explained that he had gotten used to the tool. In future evaluations, a flexible amount of time might be considered in order to mitigate this problem. However, it must be considered that it will affect the workshop logistic. For a better assessment of one tool, a flexible amount of time seemed to be the best option. For an experimental practice as considered in this evaluation, the given time of 40 minutes for practice is enough for the tool assessment according to the attendees. The following question asked whether someone with education and training like theirs (considering the workshop) is able to accomplish the task. The answer was also positive with the median making 2 (avg 1.8, sd 0.9). The answers to this question are very similar to the previous one, but the attendee that answered a full No-7 in the previous question answered 4. He explained that

Table IV.1: The TTF user evaluations: scale from 1(Yes) to 7(No) – 1 is best for all, but D7-Confusion.

Dimension	**Median**	Average	S.D.
D1 - The right level of detail	2	2	1.2
D2 - Locatability	3	3.3	1.7
D3 - Accessibility	3	3.2	2.1
D4 - Meaning	3	3.3	1.7
D5 - Ease of Use	2	2.3	1.4
D6 - Presentation	3	2.9	1.3
D7 - Confusion	5	4.8	2.1

operational training of the tool is required because it was difficult to use. It also means that a better training in the tool can minimize this, because the next evaluation after part 3 was better. This can lead to another experimental design and introduce bias in the user evaluations about the ease of use. The final question of this part was whether attendees understanding of the task had changed after practice, having positive answers of 1 in median (avg 1.3, sd 0.4). This result is expected because of the short explanation of the task; it means that the attendees learned more about performing the task. The explanations were around the realization of how hard the task is and how they underestimated the efforts required to finish it. Although they thought it was possible for them to complete the task, it required long hours and specific knowledge of the content used for the study.

After the second practice (third part of the experimental evaluation), the question whether the practice time is enough for an evaluation had a median 2 (avg 1.6, sd 0.4). The following question about their understanding of the task after practice also got a positive answer with median 2 (avg 2.3, sd 2.1). Many attendees said that they learned more about the task completion and how hard it was to finish. Two of them said that they already knew how to solve from what they have learned in the previous part.

The TTF evaluation provided an assessment from the end-users point of view of TorchSR utilized to solve the content selection problem. Table IV.1 presents consolidated results for the seven evaluated dimensions of this assessment. The amount of evaluations is small, therefore further statistical analysis is not suitable, but we can grasp some peculiarities from the assessment and try to explain them. This was also possible due to the analysis of the answers from the last part of the evaluation.

The first TTF dimension evaluated, regarding the right level of detail (D1), that is, maintaining the right data at the right level or levels of details, achieved a positive result of 2 in median. It seems that attendees interpretation of what they meant by purpose and task is a little different, because the assessment of the data required to accomplish the task had better evaluation with 1.5 in median (avg 1.8, sd 1.0) than the evaluation regarding their purposes with 2 in median (avg 2.3, sd 1.3).

The result of the second TTF dimension, regarding the locatability (D2), that is the ease of determining what data is available and where, is slightly towards a positive evaluation. However, this dimension evaluates the main activity in solving the content selection problem – finding relevant content for analysis. We believe that these assessments of this dimension, and of the next, regarding accessibility (ease of access to desired data), were mostly influenced by this factor. In the evaluation of the fourth part this was clarified by some of the attendees. In the third dimension (D3-Accessibility), a performance evaluation stood out because the assessment considering how quickly users can get the data they need had 4 in median (avg 3.6, sd 2). This is worse than the evaluation without considering the speed to get the data that scored 2 in median (avg 2.8, sd 2).

The fourth TTF dimension, regarding meaning (D4), that is, the ease of determining what a data element in the screen means, had positive evaluation, despite the lack of tool tips and the quick and short explanation about the tool and the task. Most of the attendees evaluated TorchSR in the first practice, but they said that after the second practice the tool, when they had more time to use it, the data made more sense to them. There are many different content descriptors and making sense of them is difficult. It is worthwhile to remember that the evaluators are post-grad level professionals in computer science, consequently a professional tool for a broader public requires enhancement in explaining all data elements presented to the users. This is somehow related to the presentation dimension (D6) that had similar assessment.

The fifth TTF dimension, regarding the ease of use (D5), that is, ease of doing what I want to do using the system for accessing and analyzing data, had positive results. It shows that the tool appears to be usable even with short explanation. The attendees were able to learn and use its features only by interacting with it. However, the last TTF dimension, regarding confusion (D7), might imply something slightly different, because the 5 in median shows users faced confusion in doing what they wanted. When considering the answers of the last questionnaire, it seems that this confusion is more related to the task difficulty than the tool operation.

The final part of the evaluation process was helpful to provide textual explanations to the TTF evaluation. All attendees said that TorchSR was their favorite tool and the hierarchical clustering helps them to find relevant content.

The attendees though they could finish it if given a proper amount of time. The effort to process the huge amount of content, especially the waste of time with useless content, was pointed out as their main concern about the task. The most desirable features were data filtering, sorting and searching.

The experiment is a task performing simulation. Despite limited realism, the experimental assessment tries to grasp users' evaluation of TorchSR. As long as a user understands the problem and has a solution strategy based on the provided tool, their assessment of which tool helps better based on the task completion should be similar. It is expected that the evaluation of the meaning and user confusion improves, as the users get used to the tool. For a more accurate assessment of this process, real situation tool utilization is more appropriate, such as in a study case. However, such evaluation requires industry level resources, beyond the scope of the academic environment. The workshop assessment provides an evaluation in a reasonable length of practice. In real projects, the task completion requires days, if not weeks for a comprehensive selection, of full-time work.

It is noteworthy to say that the development is grounded in experience from previous research studies. Early prototypes of computational support to analysts were based on spreadsheet systems and simple scripts in Lua (*i.e.* programs) to provide customized measurements. A tool that offers these features helps analysts even more by providing a better work environment. Instead of using a set of scattered tools, analysts can rely on an integrated support system to help them in selecting content. Therefore, in the professional analysts' opinion, scripts to provide the customized measurements enabled them to properly tackle the task of content selection. However, these opinions were from few analysts and did not consider a tool with advanced capabilities as in TorchSR. Though, this experimental evaluation investigated an assessment of a tool featuring all those advanced capabilities, especially regarding the hierarchical clustering. The result indicates that, regardless of required enhancements like data ordering and filtering, the process for hierarchical clustering built into TorchSR helps the users in tackling the task of content selection for analysis from online forums.