VI
Conclusion

This thesis presented a research about computational support to a practical approach that uses public data available on the Internet for studies about healthcare issues (see Chapter III). The main contribution of this thesis is a process with computational tools to conduct studies based on the proposed approach (see Chapter IV.2). Two research studies on healthcare topics demonstrate the application of the proposed process and tools (see Chapter V). This chapter concludes this thesis according to this organization.

VI.1 An approach to study healthcare on the Internet

The application of communication and information technologies to health is a promising research topic. In this context, the development of new approaches and understandings of the ways in which people join forces, share and identify relevant information can foster improvement in the quality of our health. The proposed approach provides a guide for researchers conducting studies.

The proposed approach is an incremental three-step-in-depth process retrieving and analyzing information from the Internet at different levels. The idea is to start analyzing data from search engine services that summarizes user search trends and information available on the Web. The next step utilizes social networking sites, looking for discussions related to the subject of interest based on keywords identified in the previous step. The final step provides a deeper analysis of a specific online community related to the research questions.

Although available data and tools exist to support the execution of the approach in most studies, better tools and data can improve its application. The evolution of social media requires initiatives of opportunistic and inventive use of data and technology. The proposed process is an initiative as such. The results presented in the research studies demonstrate the feasibility of this process approach application.

A multi-agent system architecture is proposed to support experts in analyzing social media. The system design addresses the process of social media analysis and key-features identified from social media expert desires. The main goal of the system is to empower social media experts in their duty, providing tools to leverage them in the process. The system development is in a seminal stage, which has its basis defined in this thesis. The system growing and evolution follow the framework defined by the architecture. The multi-agent system architecture is flexible in incorporating new kinds of collectors, processors and analysis methods. This enhancement is expected and required in order to catch up with the fast-pace of the Internet technology evolution.

VI.2 Supporting the proposed process

Community Association Map

The Community Association Map aims to reveal users' interests based on their associations within their community. The epistemology considered in the process development is presented in this thesis. The process is divided into three steps. The first is the data gathering step, for collecting users and community membership information. Next, follows the model and measurement step, which contemplates the model creation and the measuring. The final step, the visualization plot, presents an appealing view of the model by revealing the associations for the experts' appreciation. An example of the process application followed by a discussion about the process opportunities and drawbacks are also presented.

The advantage of the process application is to flesh out users' interests based on the processing of all these users' associations to the other communities they belong to. The disadvantage is the lack of guidelines for filtering. 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].

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. Finally, this process could be added as a component on a social media analysis platform, such as part of the multi-agent system proposed in Chapter III.

Supporting analysts in exploring and selecting content from online forums

The content selection problem is a difficult one for social scientists that embrace new ventures in conducting research based on the vast content available in online communities. The problem sets two objective goals: to maximize the number of selected participants and minimize the content volume to be analyzed. These are conflicting goals. The solution to the problem is also driven by the research interests, which are not measurable (so far).

This thesis introduces a process to support researchers in tackling the problem of content selection from online community forums. This process is based on unsupervised machine learning techniques, such as hierarchical clustering. This presents consolidated and structured results to support researchers in selecting interesting content to be analyzed for their research. These results include measurements and a content exploration method. Examples of this problem in real world research studies are also presented with the strategies employed to solve them. The process development was based on acquired experience. The process creation rationale and its description are presented. As an application of the proposed process, a tool based on the process was created to aid researchers to apply it, called TorchSR.

User's evaluations of TorchSR show that the hierarchical clustering created through the proposed method helps to tackle the problem. Although this tool is a prototype, the evaluation through a TTF instrument indicates that the content measurement is presented in the right level of detail. Enhancement of content exploration method and the data manipulation operation features, such as ordering and searching, were desirable features as identified in the evaluations. Users' feedback also reveals how hard and important this problem is for qualitative researchers that use online forum discussions for analysis. So far, TorchSR is the best tool available to support users to tackle this problem.

Without the proposed tool, researchers rely only on superficial metrics about the content of the topics, or they must look at the whole forum content to perform the content reducing task. The proposed tool aims to support researchers in tackling 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 aid in solving the problem, the subjective goal of what is of interest to be analyzed is still a burden upon researchers.

Further research in the measures used to calculate posts' similarities can also provide the user with better results, and improve the process. Another interesting enhancement, is to consider user's feedback of what is "interesting content" in their search, so the recommendation rankings can be improved iteratively. This is a new trend in machine learning research, called active

learning. Analytical support of the proposed process should be useful as a feature of a professional system to support in a full process of qualitative research. Therefore, a natural evolution for TorchSR is to go from a support tool to tackle the content selection problem to become a system to support users in the whole process of qualitative research.

Finally, future research can aim at the broader problem of finding answers to questions of interest from the analysis of community forums. This would be better if it assumes that the discussions in forums are not known entirely and therefore the unsupervised machine learning approaches cannot be applied directly. Adding up, the solution should solve the problem at scale. In this sense, many techniques found in related works on the similar topics should be evaluated, showing the advantages of a better approach by conducting experimental studies.

VI.3 The research studies

Two research studies demonstrate the proposed method and tools. The first research study is based on the research theme of Madeira [Mad11], which used a similar method. Instead of analyzing discussion from specific questions posted in the forum, the conducted analysis showed that it is possible to identify relevant insights about a community by analyzing content previously available.

The study presents a disclosure of the preliminary findings considered original and promising. The word frequency / content analysis approach expressed needs of social support and material assistance that may provide subsidies for further qualitative approach and public health policies aimed to HCV carriers. This leads to the identification of patterns of recurring demands. In this research, identification Patterns of Recurring Demands requires small resources in its development in contrast with important outcomes in terms of depiction of demands from patients with chronic diseases underestimated by other perspectives. The word frequency and content analysis can furnish hypotheses, linking concepts and "bounding ideas" which are essential to the portrayal of collective ideas and social support demands. The present findings describe some evidence of need for social and material support. This may subside public policies aimed at carriers of HCV.

The second research study aimed at a major social problem in Brazil, drug abuse. The broad spectrum in data collection of the proposed method, which take into account data from more data sources (i.e. Internet, Social Networking Sites, Online Community), enabled researchers to draw an overview of the online population regarding the study theme. Although the proposed

method is more suitable as an early stage research, like an exploratory study about a theme, the in-depth data filtering and analysis characteristic can lead to more qualitative knowledge by deeper analysis of the available content. The computational tools employed in this endeavor supported research in the hard task to tame the huge amount of data. However, it is a burden upon the researchers to define filters to find interesting content to their studies. The study results are interesting and have been a subject of 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.

In conclusion, this thesis research involved many partners and collaborators. It was oriented by real-world problems. The work strategy was pragmatical problem solving. This strategy enabled fast practical results. Even though, the diverse interest of the Anamnesis members led it to end its operation in its first year, it is an interesting research activity, but not profitable. The balance between pioneering research in hot new topics and anchoring the findings in traditional research fields of well-established sciences was a great lesson taken from this thesis research. This can be seen as a drawback of pursuing multidisciplinary applied science. The results exposed in this document can — and will — be explored through further research. In time, it is also important to conclude this research cycle.

The proposed process and tools are useful. However, technological evolution must require evolution in this process. An example of this evolution is the recent user migration from Orkut to Facebook in Brazil. Previous phenomenon happened with Myspace and Facebook in USA [Boy07]. This is a natural behavior on Internet. Successful sites with great audience such as Altavista, Geocities, AOL, lost their audience to new ones such as Google and Facebook. Although this research might look like outdated in such short notice because it deals with Orkut instead of Facebook, the foundation in the proposed approach to deal with social networking sites, regarding if it is Orkut or Facebook, is still the best source of social media, especially regarding online communities. Besides that, the proposed method and tools are designed to adapt and evolve with its natural (virtual) environment. This was explored by a software architecture by the means of lousy-coupled (smart) agents. There are many open problems and possible paths to continue this research, as some of them has been addressed in their context along this chapter.