II
Related works

From a computer science point of view, the interest of this thesis research is about computational processes, methods and tools to support the conduction of social studies. In the context of this thesis research, the utmost goal would be a system with an input box for the research interest and a button one must hit in order to obtain a processed report. However, computers cannot yet produce such reports. As an example to illustrate the complexity involving the generation of these reports, the results produced by a typical study are considered for a master thesis in public health. To this day, the research and development attained by computer science provide enough support to cover all phases of the process, which involves many disciplines of this science. This chapter presents related works that was used as inspiration to this research rooted in computer science, social science and public health.

II.1 Research on social media

Some experts in the scientific community claim that a new scientific field is arising: the coming of age of computational social science [Laz09]. Research over user-generated content available on the Internet used to give special attention to social network analysis. In this field, the state of the art in computer science is around the creation of analytical models to perform or support users in specific tasks. In the literature, there are studies that attempt to characterize the users' population of important sites, such as Facebook [Naz08], Twitter [Hub08] and YouTube [Che08]. The mainstream theory used to analyze social networks is based on complex networks studies [Cos07]. Other studies aim to understand the users on social networks and perform sentiment analysis, so one can identify special users on these networks and predict the network evolution and its implications [Ben09, Wil09, Ahm10, Big10]. The diversity of available data and research possibilities can be explored through different methods and techniques, Netnography [Koz09] being an inspiring one.

Following ethnographic principles, Netnography observation happens in users' natural habitat [Koz09]. Online communities are the natural (virtual)

PUC-Rio - Certificação Digital Nº 0921317/CA

habitat for people gathering to discuss their interests. The definition and limits of these virtual communities varies among the scientific disciplines and applications [Pre05]. It has interesting features like the possibility of anonymous interaction, which is a way that people have found for open talk about their life without embarrassment. For instance, in online forums individuals feel comfortable talking about their health conditions that would cause social stigma and embarrassment in face-to-face contacts. Furthermore, besides exchanging information, people support others in the community, creating strong ties in these relationships.

The next section describes the realm of the Web as a data source for studies in public health, showing some applications, and presenting studies analyzing available data. In terms of healthcare investigation, the world of online communities is an outstanding place to find information about healthcare issues. Accordingly, a discussion on online community is presented based on studies highlighting its relevance to the study of healthcare issues.

II.2 Research on healthcare issues based on social media

The Internet usage leaves records that can be used to investigate health care issues [Eys09]. The Internet is a medium used by a significant fraction of the population of Brazil. According to comScore estimates from 2010 on Brazilian online audience [Cscr11], there are around 77 million Internet users in Brazil, representing access from home, work and public locations (lan house, school, etc). This represents almost 40% of the Brazilian population [IBGE11] and the numbers are growing fast. Social network sites reach around 85% of Internet users in Brazil, according to those same estimates. Especially in the case of socially stigmatized illness, users go to the Internet to look for information and to share their experiences [Ber05]. On online forums, individuals feel comfortable talking about health conditions that would cause social stigma and embarrassment in face-to-face conversations. Anonymous interaction is a way that people can talk openly about their lives without embarrassment. Furthermore, besides exchanging information, members start to support others in the community, binding strong relationship ties.

According to the Brazilian Internet Steering Committee [CG10], 87% of users use the Internet to search for information and utilize online services. From these users, 35% use the Internet to search for information related to healthcare or health services. In developed countries like the USA and the UK, the majority of users perform searches on healthcare topics. According

PUC-Rio - Certificação Digital Nº 0921317/CA

to research of the Centers for Disease Control and Prevention [NCHS10] from January through June 2009, 51% of American adults aged 18-64 had used the Internet to look up health information during the prior 12 months. In the UK figures seem to be even higher [Tel10]. A Porter Novelli EuroPNStyles survey showed that 65% of those questioned chose to search on the Internet when they wanted to know the answer to a medical query, compared to 43% who asked their doctor, while only 27% look for information via television shows. A mere 14% of interviewed people rely on government health information services.

The analysis of search, communication and publication behavior on the Internet can reveal interesting patterns about public health [Eys09]. An example of this data application is the Google Flu Trends service[1]. The Google Flu Trends is a project that aims to detect and anticipate flu epidemics based on the analysis of search terms used on Google. Due to the population's increasing habit to search on the Web about symptoms and medication, even before looking for a doctor, Google can perform this real time Flu tracking system. According to Carneiro and Mylonakis [Carn09], "Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems."

Another remarkable research on disease surveillance systems on the Internet is the Dengue surveillance system[2], that shows the evolution of dengue situations reported on Twitter. Dengue is a mosquito-borne infectious disease and a leading cause of illness and death in tropical and subtropical regions, like Brazil. The system was built based on an active surveillance methodology. Gomide *et al.* [Gom11] found a high correlation between the number of cases reported by official statistics and the number of tweets posted over the same period.

PUC-Rio - Certificação Digital Nº 0921317/CA

II.3 Research focusing online communities

Online communities are the most natural virtual habitat for people gathering to discuss their interests. The definition and delimitation of these virtual communities varies among the scientific disciplines and applications [Pre05]. The broadest definitions see online community as any group of people on Internet. A stricter definition is that an online community must be lively, have a minimum number of members, policies, and purposes, and occur on the Internet. In this work, the investigation goes from a broad view of online communities towards a special online community, where answers to the research questions of interest are more likely to arise.

[1] `http://www.google.org/flutrends`
[2] `http://www.observatorio.inweb.org.br/dengue`

Lasker et al. [Las05] 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, this work considers a broad definition of online forum [Pre05] as any online system where people can discuss things - they share content. Whitehead [Whi07] provides an integrated review of the literature on methodological and ethical issues in Internet-mediated research in the field of health.

A more recent phenomenon on the Internet is the emergence of social networking sites [Boy07], which have effectively served to recruit many users to the Web. They are a fertile environment for the establishment of online communities. In the context of health, these systems have been explored in many studies. Gold *et al.* [Gol11] realized a systematic examination of the use of online social networking sites for sexual health promotion; they found 178 sexual health promotion activities (meeting their inclusion criteria) and only one of these activities was identified in the literature. The authors comment, "SNSs are being used for sexual health promotion, although the extent to which they are utilised varies greatly, and the vast majority of activities are unreported in the scientific literature."

Using Facebook, *Greene et al.* [Gre11] 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 [Ben11] 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 (236, 38.1%), product or service promotion related to fundraising or awareness (61, 9%), or patient/caregiver support (46, 7%).

Discussions in online forums (*i.e.* content created by users) are detailed and contextualized in the community. If it is publicly available, it can be retrieved in non-intrusive ways, enabling an opportune, effective and efficient processing [Koz09]. In this way, users are not summoned to participate in a reactive fashion (*e.g.* online surveys) allowing the analysis of freely constructed manifestations. New research venues are being created such as the workshop on Words and Networks: Language Use in Socio-Technical Networks (WON 2012) at the ACM Web Science Conference [WON12]. This workshop presents

PUC-Rio - Certificação Digital Nº 0921317/CA

many examples of related research on analysis of online forums.

Madeira [Mad11] investigated in Orkut's online communities about transformations of the power relationship between physician and patient for her PhD thesis. This investigation mainly considered discussions fostered in online forums and analysis of content through discourse analysis using the Discourse of the Collective Subject (DSC) technique [Lef06]. Although there are many discussion analysis methods, based on Madeira's work the DSC showed to be a powerful technique available to organize the freely constructed manifestations on online forums and present meaningful results. However, it is noteworthy that this methodology discussion is more pertinent to social science. The purpose of this thesis research is to support the application of some discourse analysis technique.

II.4 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 this organization, thus allowing users to explore the collection interactively through topics that indicate the contents of the available documents [Rez11].

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-kmeans (divisive) achieve the best results in textual data [Zha05]. 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, construction of digital libraries of patents [Kan07], search engines [Carp09], webmining [Liu11], and, more recently, analysis of online communities and social networks [Kad12].

In particular, text clustering for online forums analysis has recently gained the attention of the research community because of the need to automat-

PUC-Rio - Certificação Digital Nº 0921317/CA

ically organize the huge volume of texts published daily. Most of the existing works found in the literature investigate extraction of user profiles [Shm10], event detection from social data [Zha07], and recommendation tasks [Dav12].

The tool Torch – **To**pic Hiera**rch**ies [Mar10] – was considered as a research starting point. This choice is based on the literature revision presented in [Rez11]. This tool implements various clustering algorithms, aiming to build topic hierarchies from growing text collections. Since the interpretation of the results of clustering is a difficult task for users, this tool tries to present the results with understandable cluster labels, such as simple descriptions like terms (words) that indicate the contents of the documents in the clusters. It also supports topics overlapping, which is a particular feature of textual collections is that documents can belong to more than one topic. Thus, topics overlapping are a desirable effect, because it allows maintaining the multi-topic property of the texts. However, this tool as proposed in [Mar10] is not suitable to perform meaningful analysis of online forums, because it does not consider the hierarchy of topics and messages. The research challenge explored in this thesis was to extend this tool development in order to properly process online forum content and support analysts in solving the content selection problem.

PUC-Rio - Certificação Digital Nº 0921317/CA