I
Introduction

Use of the Internet as a communication medium has set a new paradigm for information sharing in modern society. The virtual world has some interesting capabilities, for example, persistent footprints left by human interactions on the web such as surfing, posts on social networking sites, and discussions archived on online communities. These capabilities have brought about a great debate regarding privacy issues and ethical standards concerning the tracking of users' behavior on the Internet, as well as the use of such data, especially for scientific research [Whi07, Koz09, Eys09]. In fact, this new realm opens great opportunities to establish new paradigms in many fields of academic research. Computer science is the main instrument enabling this new kind of science [Hey09]. For instance, a group of researchers recently coined the term *computational social science* [Laz09] to show and discuss new possibilities for social science regarding the computational processing of huge amounts of data, typically obtained from the Internet, on a scale impossible to conceive with traditional research methods. In the field of health, reliable sources of social media are being used for many studies (*e.g.* [Eys09, Carn09]). However, enabling sophisticated use of social media requires innovative computing research [Shn11]. This thesis research explores a novel approach to perform studies in healthcare that consider data from social media, specifically presenting a process and tools to analyze data especially from online communities in social networking sites.

I.1 Background and motivation

Inspired by new perspectives for societal studies based on Internet content, the research for this thesis began with exploratory research on opportunistic use of social media. This research has started as an opportunistic joint venture between business and academia. Although it was not a formal research project at the beginning, it happened in the context of the flourishing Web Science research agenda, especially in the context of the INCT Web Science Brazil. The motivation came from the practical opportunities in social

media research and business that have been explored in recent years. In this multidisciplinary environment, many experts claim that a research agenda requires and establishes new science fields, such as Computer Social Science and Web Science [Laz09]. Nevertheless, computer science is the foundation science in this research, because the Internet is a technological artifact built on its principles. Therefore, I believed since the beginning this research would lead me to interesting results in computer science. Laureate computer science researchers also share the point of view that "realizing the value of social media requires innovative computing research" [Shn11]. The employed strategy was that the results would come from applied research on real-world problems found in relevant applications.

In 2010 I established a joint venture with an important marketing agency in Brazil, Mapa Digital. At the early stage of the research we were looking for opportunities to exploit the biggest sources of social media available on the (Brazilian) Internet. The regional constraint to work with Brazilian social media is consequent with our practical approach to bring value to the business partners. The team at the time was a social media expert, Rodrigo Pazzini, and me. Rodrigo has a background in Social Sciences and Design and is well versed with computers. At the time, he was working at Mapa Digital in Belo Horizonte (Minas Gerais, Brazil), a company specialized in digital marketing in Brazil. He is also a friend from childhood. The very first activities were aimed at supporting social media analysts with new processes and tools (*i.e.* scripts) to help them study business branding on the Internet [Aki08, Koz09]. Since the major source of social media in Brazil at that time was the social networking site Google's Orkut [Boy07, Cscr11], we focused on the data publicly available on this platform. Online communities on Orkut are the main place where people get together and discuss their interests. Early findings of this exploratory research are available in a technical report [Car10].

Based on the social media analysts' experience, we focused our research on the field of public health. In addition to reduced costs compared to conventional field research, the analysis of online communities brings an opportunity to capture written speech published in moments of acute need for support. It is known that the segmentation of the Internet into virtual communities seems to transcend their merely informative context [Fer11], and to acquire a unifying force aimed at overcoming great obstacles. Besides posting messages on topics for mutual enlightenment and social support, participants in online forums feel more comfortable talking about personal difficulties and living conditions, compared to the discomfort of a conventional medical consultation. In general, stigmatized diseases or health conditions

encourage individuals to take advantage of the internet as their main source of information and environment for sharing experiences [Ber05]. This choice was also reinforced by evidences found in many other research cases [Whi07, Eys09, Carn09, Mad11].

Starting research from real world problems brought many opportunities and made it easy to show the relevance and application of the results. As a multidisciplinary field, web science research requires a skillful and intrepid team formed by members from complementary backgrounds. Besides computer science, the research of healthcare issues through social media analysis also required experts in social and health sciences. It is important to acknowledge that this thesis research is the result of teamwork. This team is presented later in this section citing some of the members' contributions. I try to point out my personal commitment to this research.

Since the beginning, this research applied pragmatic decision making. It means that our focus was on obtaining interesting results based on existing technology. This strategy led us to fast results. That is a positive outcome. However, this strategy helped only a little on my research in computer science, my main concern in this endeavor. It is noteworthy to mention that the Community Association Map (Section IV.1) was created at this stage, and further elaborated on with the help of Prof. Hugo Fuks as part of my Qualification Exam. My role in this partnership was to provide support to Rodrigo with general consulting about technology and tailored computational tools.

After this warm-up, it was time for a pilot project in order to test the ideas' feasibility. At this time we brought another social media expert to the team, journalist Sergio Rosa. Rodrigo and Sergio were co-workers at Mapa Digital where they worked together for three years. Mapa Digital was ceasing activities at the time and they were looking for a start-up business in social media analysis. The story of how this happened is too complex and goes beyond the purpose of this discussion, but at the end we had a start-up formed. I tried to persuade them to analyze social media related to the stock market, where I had more business partners to work with. However, they had more altruistic feelings to do something to help people in general. They also knew from previous works that the field of health was a promising one to work in. We called the start-up Anamneses, just to recall the medical procedure and reinsure the analysis purpose of the business. Subsequently, the team needed some specialists in social media analysis applied to public health.

From an interview on a regional TV channel in Minas Gerais, Sergio took notice of scientific research on public health conducted in online communities.

He called Wilma Madeira, recently graduated with a doctorate in public health at USP (São Paulo), and invited her to be part of our team. Her answer was positive. She has a background in social science and informatics, besides the doctorate. With this team, we set up a pilot study of three months about hepatitis C. Hepatitis C was one disease studied by Wilma in her thesis [Mad11]. The study about hepatitis C was a success. Even though it was really hard to perform, especially for the analysts, I managed to conclude the project almost on time, the arbitrary chosen three months. The achieved results were reasonable enough to demonstrate the feasibility of our proposed approach. The analysis was carried out only by Rodrigo and Sergio. I supported them with general consulting, tools, reviewing (QA) and managing the project. Wilma was too busy at the time and could not participate in this study. She was the CTO of the Instituto de Responsabilidade Social Sírio Libanês (São Paulo), besides being involved with teaching as well. At the end, she helped the Anamnesis team with our next project.

Considering the doctoral thesis of Wilma Madeira [Mad11] as a starting point, the next step was to study an important healthcare issue through an opportunistic use of services and data publicly available on the web. The case considered in this study was about hepatitis C and its related themes. This case study was also a subject of Madeira's doctoral thesis, but instead of asking questions in online forums, we analyzed data previously available there. It is believed that the support from virtual social networks dedicated to the problem represents an important resource for hepatitis C patients who encounter obstacles in adapting to everyday difficulties arising from their condition. The themes of interest that were found are related to a combination of problems, treatment side effects, prospects for recurrence, and need for radical lifestyle changes. This brings challenges that carriers could not tolerate without relying on social support from spouses, relatives, friends, and, of course, other carriers of hepatitis C virus (CHPV). As a result, we identified patterns of recurring demands by analyzing the data available in virtual communities dedicated to CHPV in Brazil. The computational support to enable the study completion was developed according to the needs of the experts.

Before starting the second study, the hepatitis C study follow up had another relevant outcome to this thesis research. Confident of our good results, we tried to make contact with many practitioners, businesses, and researchers. We received positive feedback from specialists, *e.g.* Carlos Varaldo, hepatitis C expert and founder of hepato.com, and some media coverage, such as a publication from the Instituto Ciência Hoje [Mor11]. From these contacts, I met the medical doctor Prof. Paulo Vasconcellos. He is a researcher at Instituto

Oswaldo Cruz (IOC — Fiocruz) in Rio de Janeiro and a professor at Escola Nacional de Saúde Pública (ENSP). We found him by searching scientific literature on health specialists that have worked with information technology. I am still working with him on the results of the hepatitis C study. At this time, we received a good first round feedback soliciting minor reviews. This publication sets ground for the discussion analysis publications. Besides this, we are looking for further research in health science, applying innovative information and communication technologies (ICT). This partnership is basically the same as the first — the specialist takes care of the application, and I take care of support and underlying technologies.

The hepatitis C research study also served as a portfolio to foster a new venture partnership in further research. The next research study was a challenge proposed by the medical sanitarist Dr. Sergio Zanetta, director of philanthropy of Hospital Sírio Libanês and head of the Instituto de Responsabilidade Social Sírio Libanês (IRSSL). The study theme was drugs, with special interest in crack cocaine. This study was developed as a partnership between IRSSL and Anamnesis. The Anamnesis team was composed of Rodrigo, Sergio and me. From the IRSSL, the team was Wilma Madeira, the nurse and data analyst Mirna Okamura, and Dr. Sergio Zanetta. Considering that this study was based on the first one, it took advantage of our already acquired know-how and expertise, so I expected that it would be easier for the analysts. In practice, it was more difficult than the first. The first study relied on the academic background provided by Wilma's thesis. However, for this study the background had to be built from scratch; we had no specialist in the study of drugs in our team. Once again, the analysts did brilliant work in four months and delivered the study. The head of the Research and Education Institute of Sírio Libanês Hospital, Dr. Roberto Padilha, was impressed by the results delivered in such short time. His comment was shared during the presentation and discussion of the results in a Seminar organized by IRSSL in January of 2012 in São Paulo [HSL12]. Dr. Zanetta was so confident of our promising approach and results that he invited the Brazilian Minister of Health, Alexandre Padilha. The minister did not go because of a mission overseas, but he sent his secretary that deals with the drug problem. At the national level, the drug problem budget is around R$ 4 billion, requiring joint activities of many public health segments and even security, police and military. The seminar [IRSSL2012] also had attendance of authorities from São Paulo state and city governments, mental health specialists, and civil society. Representing Anamnesis, I presented and discussed the results in this seminar. Apart from this, my involvement was basically the same as the first study.

My last personal challenge was to make sense of the developed work in a way to show innovative and scientific contribution to computer science. Hence, I could conclude this thesis research. Since our work method in the two studies was pragmatical problem solving in conducting the research studies, I thought more innovation in computer science was required to justify a computer science thesis. Based on my field experience helping analysts, I decided to investigate the problem for which our approach would potentially differentiate from others and computational methods which could make a big difference. This was found to be the content selection problem.

From these two research studies, a research approach [Car12] was proposed. The critical stage of this approach is the deep analysis of an online community, explaining the community reality as related to the research questions of interest. This analysis might employ traditional social science techniques and take advantage of new computational processes too, such as the Community Association Map, developed along this thesis research. In order to process all available data, computational tools (*i.e.* scripts that help to collect, organize, and process the available data) support experts in the task of making sense of all information available and report it as findings of the study. We understand data here as any piece of information in raw format and information as data in a context, and as an utmost goal experts try to consolidate knowledge from the findings of these studies in a report. In order to achieve this goal, the analysis should consider a comprehensive study of the discussion presented in the community forum. A valuable technique that can be used to describe the community population is the so-called Discourse of the Collective Subject [Lef05, Lef06, Mad11], a qualitative technique with roots in the Theory of Social Representations. Since the amount of data is huge and most of the (qualitative) research techniques are very time consuming, computational tools to support experts who select data through manual scrutiny is very desirable in this scenario. Additional data available in social networking sites can also be automatically analyzed, generating big data aggregates as results.

In a qualitative analysis, the content selection problem could be explained as a data cut phase. Even though it is basic procedure in qualitative research that faces large volume of data to be analyzed, this problem has some peculiarities worthy of being studied. In the context of our approach, which is beyond traditional social science research, analysts have to cut data from online forums. Moreover, analysts face a huge amount of text (*i.e.* Big Data) that has a predefined structure (*i.e.* organized as topics with messages). Following research studied this problem from a computer science point of view, which involves natural language processing by (unsupervised) machine

learning techniques. This is a hot research topic with results being used in practice by many applications. The processing of professionally written text is considered a hard but feasible task. However, automated processing of online forums is a very hard task. The low quality text — *i.e.* not written in proper language, with colloquialisms, and typos — is enough to justify the task with this classification. The challenge was to research and develop new approaches that would benefit in content exploration and selection.

In the scientific literature I found Torch [Mar10], a tool for organizing documents in hierarchies of themes. This seemed to be a promising representation in order to improve exploring the online forum by other representation of its topics. I could have tried to redo their work, but I think a partnership can foster better achievements (*i.e.* solutions). Besides that, it is more productive. Consequently, I contacted the authors, Prof. Solange Rezende from the Instituto de Ciências Matemáticas e de Computação of the Universidade de São Paulo (São Carlos, São Paulo) and her student Ricardo Marcacini. In this partnership I took care of the domain application, definitions and validation, and Ricardo supported me in the process design and implementing the tools. This is basically the opposite set of roles than my previous partnerships.

In the end, I took care of the assessment of the proposed process to support analysts to tackle the content selection problem. It is worth mentioning that the proposed process is an evolution of previous solution, which is better by definition (*i.e.* conception). Nevertheless, the tool is still a prototype. Accordingly, it is not suitable enough for user interface evaluations. The purpose of the experimental evaluation was to uphold that the principles employed in the proposed process are valid and provide insights of weakness to guide further enhancing. At this point, I acknowledge the help and support of the Software Engineering Laboratory of PUC-Rio in the realization of this evaluation, under the leadership of Prof. Carlos Lucena and mentoring of Prof. Hugo Fuks.

I.2 Methodology

This thesis research followed a pragmatic approach of problem solving, whose focus was to exploit social media to perform social studies, especially related to healthcare issues. The initial research was an exploratory study about how to exploit social media to study relevant topics about healthcare issues. This was achieved by seeking answers to the following questions:

– Which data sources should be used in these studies?

– How to support analysts to deal with the great volume of data and conduct the studies?

– According to the study goals, how to provide meaningful analytical results based on the available data?

A practical approach to conduct such studies was proposed as a guideline based on this initial research. A process compliant with this practical approach was defined to cover all phases required to conduct a study. However, its coverage is too broad for an in depth-research for the expected time span of this thesis research. Consequently, the following research focused in the designing of tools to support specific tasks of this process.

Finally, two research studies on relevant healthcare issues were conducted to show the application of the proposed process and tools. The first study investigated hepatitis C and related theme, which was one of the studies conducted by Madeira in her doctoral thesis research [Mad11]. The other research study explored the realm of drug abuse, especially the motivations for start and cease of use of crack cocaine in Brazil.

I.3 Thesis statement

This research thesis aimed to exploit valuable social media, especially the ones available in online communities of social network sites, to perform social studies about healthcare. A process based on a practical approach was defined to conduct such studies. This approach was designed with three stages to guide analysts to achieve their study objectives. It starts with a content analysis of users' online search, which aims to survey how users are looking for information on Web. In the second stage, the analysts searches for a specific online community in social networking sites, where users discuss about the research theme among themselves, until finding one that looks promising with regards to the study goals. The last stage is an in-depth analysis of the online community. This is a systematic use of free data available on the Web.

A system to support analysts in this proposed process was defined. This development follows software engineering principles, especially the ones designed for multi-agent systems. Many computational tools were created, tailored for specific tasks and needs of the considered approach. Two tools that stand out are presented because of their utility and the complexity of the process in which their build was based on. For the benefit of online community analysis, a tool was created based on a process to aid in the analysis of the inter-community relationships through the Community Association Map. The goal is to reveal the interests of users through a map of associated communities.

Users' membership is utilized to establish the relationship among communities. This map shows the interests of members in other communities. The other tool regards the task to find discussion in online community forums in which content looks promising to answer study research questions. Since amount of discussions available is too vast, the data selection from online forums to be manually analyzed by (qualitative) research techniques (*e.g.* content and discourse analysis) is an important problem to be overcome by analysts. This problem, called the content selection problem, was tackled with a process designed based on unsupervised machine learning.

The results of significant studies about healthcare issues were obtained by the application of the proposed approach. One study identified demand patterns of carriers with hepatitis C. The other studied the motivations for drug abuse start and cease, specifically with regard to the drug crack cocaine in Brazil. These are examples of how to provide meaningful analytical results from the analysis of social media, especially considering online communities.

Lastly, this research tries to bring together different disciplines, as is required to conduct such studies, and to step into the arena of computational social science [Laz09]. It presents results obtained by mixing classic techniques from social science with innovative approaches from computer science. The work as a whole provides insightful perspective for applied research in this new field.

I.4 Thesis organization

This is organized as follows. Chapter II presents related works rooted in computer science, social science and public health that inspired and served as start point to this research. The practical approach derived from the studies carried out in exploratory research, with a description of the computational system to support this approach, is presented in Chapter III. Chapter IV shows two developed tools to support a process based on the proposed approach. The Community Association Map is a tool developed to help in online communities analysis, and TorchSR is proposed to support analysts in solving the problem of content selection from online community forum. It also contains an evaluation showing a tool, which was built based on the above process, improves solving the task of content selection for analysis. Chapter V presents the main results of the two research studies from the exploratory research, about hepatitis C and drug abuse. Lastly, Chapter VI concludes this thesis.