UNCOVERING SOCIAL MEDIA DATA FOR PUBLIC HEALTH SURVEILLANCE

Ming Yang, School of Management, Harbin Institute of Technology, Harbin, China, yangming.hit@gmail.com
Yi-jun Li, School of Management, Harbin Institute of Technology, Harbin, China, liyijun@hit.edu.cn
Melody Kiang, Department of Information Systems, College of Business Administration, California State University, Long Beach, United States, mkiang@csulb.edu

Abstract

Social media is playing an increasingly important role as the sources of health related information. The goal of this study is to investigate the extent social media appear in search engine results in the context of health-related information search. We simulate an information seeker’s use of a search engine for health consultation using a set of pre-defined keywords in combination with 5 types of complaints. The results showed that social media constitute a significant part of the search results, indicating that search engines likely direct information seekers to social media sites. This study confirms the growing importance of social media in health communication. It also provides evidence regarding opportunities and challenges faced by health professionals and general public.

Keywords: Social media analytics, Disease surveillance, Health informatics, Public health service
1 INTRODUCTION

With the advent of Web 2.0, the past few years have witnessed the rapid rise of social media, such as Web forums, blogs, wikis and media-sharing websites. Social media websites harness the “wisdom of crowds”, giving Internet users convenient and instant access to information and the communities (Eysenbach 2009). These new tools and novel information sources are also becoming ubiquitous due to the increased access to the Internet and mobile communication, such that participants in social media systems can obtain public health information more quickly and directly than at any time in history (McNab 2009). Moreover, members of the general public can readily discuss their health conditions, disease questions, anxiety, and illness treatments interactively in social media systems. These user-generated contents are rapidly emerging as tremendous assets for syndromic surveillance, which is concerned with the continuous monitoring of public health-related sources and early detection of adverse disease events (Yan & Zeng 2008).

Syndromic surveillance systems aim to provide effective prevention, detection, and management of infectious disease outbreaks, whether naturally-occurring or caused by bioterrorism (Zhang et al. 2009). Data sources used in developing a syndromic surveillance system are expected to provide timely re-diagnosis health indicators (Yan & Zeng 2008). Hence, there is an increasing need to make full use of the critical health information in social media. However, finding the best data source from social media for syndromic surveillance is difficult, due to the amount and diversity of information available in social media. Methods that are capable of discovering relevant data sources from social media for syndromic surveillance are desired. They are helpful for Centers for Disease Control and Prevention (CDC) to detect pandemic disease outbreaks, and provide early warning to the management of public health emergencies.

In order to find relevant data sources, search engines have become powerful interfaces that serve as the “gateway” to health-related information as well as important communication channels. However, search engines often overwhelm users by producing laundry lists of irrelevant results and creating information overload problems. Finding a usable amount of relevant, accurate, and trustworthy information is not always easy, and without expert subject knowledge, the information seeker may be misled or misinformed (Abbott 2010). Hence, much time and effort can be wasted in following up worthless, out-of-date, and inaccurate information.

In this paper, we address the aforementioned problems by proposing and implementing a semi-automated method for collecting and analyzing health-related information in social media. Leveraging human preciseness and machine efficiency, the method consists of various steps including collecting, filtering, analyzing, and visualizing health-related information in social media. This research investigates the extent the method can assist CDC analysts in collecting and analyzing social media data for syndromic surveillance. This study is helpful for documenting the current trends on the Internet and providing useful insights for public health service.

2 BACKGROUND

In health communication, timely, accessible and credible health information is critical for improving public health outcomes, whether to help people take action during an outbreak or to prevent illness (Rimal & Lapinski 2009). Among healthcare consumers, computer-mediated communication (CMC) for health information is increasingly perceived to be essential to health and well-being (Seckin 2010). Social media, a great CMC form, is radically transforming the way people communicate around world.
and presumably emerging as an important component in health communication (Kreps & Neuhauser 2010). For instance, countless blogs on health topics written by specialist and non-specialists alike are read, commented on and shared globally. Social networks websites such as Facebook and Myspace are also utilized by hundreds of millions of people to communicate about a vast range of topics, including health (Boulos et al. 2010). Social messaging service such as Twitter enables users to follow public health events instantly from their desk or mobile devices. The World Health Organization used Twitter during the influenza A (H1N1) pandemic and, at that time of writing, had more than 11,700 followers (McNab 2009).

Recently, a growing body of research has been done trying to analyze health related content in the social media. For instance, Denecke and Nejdl (2009) compared the contents of medical Question & Answer Portals, medical weblogs, medical reviews, and Wikis. The results showed that there are substantial differences in the contents of various health related social media. Boulos et al. (2010) evaluated a number of emerging technologies and tools that can be used to exploit these inherent social Web features in real or near real-time and harness them for public health, environmental and national security surveillance purposes. Based on text mining and social network analysis, Corley et al. (2010) proposed an approach that can identify online communities for targeted public health communications to assure wide dissemination of pertinent information. Syndromic surveillance requires high-quality data to construct an accurate prediction model. However, previous studies have placed limited emphasis on how to collect and select relevant data from social media. Although public health professionals are well aware of web-based analytical tools such as Google (Ginsberg et al. 2009; Seifter et al. 2010), and analytical tips from the social media community (Denecke & Nejdl 2009; Boulos et al. 2010; Collier et al. 2008), finding the best knowledge source for a specific information need is difficult due to the fact that relevant information can be either hidden in Web pages or encapsulated in social media (Denecke & Nejdl 2009). At the same time, it is needed to reshape the guidelines for the collection, dissemination, and intervention of the Global Health Surveillance in accordance with the newly emerged and re-emerged infectious diseases, new cycles of pandemics, and the threats of bioterrorism (Castillo-Salgado 2010).

The deficiency of existing techniques for social media analytics in syndromic surveillance, coupled with the challenges associated with building effective techniques warrant the use of new guidelines to collect and analyze health related information on social media to improve public health outcomes.

3 METHODS

To uncovering the social media data for syndromic surveillance, we propose a semi-automated method that integrates various information collection and analysis techniques and human domain knowledge. Figure 1 shows the framework of the proposed method aiming to assist human investigators to obtain health informatics on social media using collection, filtering, and analysis techniques.
3.1 Information Sources

The Information has been collected from social media websites, which include weblogs, microblogs, wiki, online forums, Web communities, and social networking websites. Denecke and Nejdl (2009) found substantial differences in the contents of various health-related social media. They showed that weblogs and answer portals mainly handle diseases symptoms and medications while wiki and encyclopedia provide more information on anatomy and procedures.

3.2 Collection Methods

Collection methods make automatic searching, browsing, and harvesting of information from identified sources possible. Meta-searching uses related keywords as input to query multiple search engines from which investigators or automated programs can collect top-ranked results and filter out duplicates to obtain pertinent URLs of Web pages. With careful formulation of search terms and appropriate linguistic knowledge, they can obtain highly relevant results (Chen et al. 2008). However, some Web pages are encapsulated in social media, and they cannot be reached by search engines. We employ domain spidering, which starts with a set of relevant seed URLs and relies on an automated Web page collection program to harvest Web pages linked to the seed URLs. In order to search Web pages which have hyperlinks pointing to a target Web domain or page, we utilize back-link search supported by search engines such as Google (www.google.com) and Baidu (www.baidu.com). Moreover, members on social media websites often share the same interests, therefore we use group/personal profile search and applied to major social networking such as Facebook and Myspace.

3.3 Filtering Methods

Filtering involves sifting through collected information and removing irrelevant results, and it requires domain and linguistic knowledge to perform this task (Chen et al. 2008). Domain knowledge refers to knowledge about public health, syndromic surveillance, as well as homeland security and emergency management. Linguistic knowledge deals with terms, textual and multimedia clues in the native
language. Filtering can be automatic or manual, depending on requirements for efficiency of process and precision of the results. Typically, manual filtering achieves high precision, but it is less efficient and relies on domain experts who have had years of experience in the field. Automatic filtering is very efficient as it often uses computers and machine learning to process large amount of data but the results are less precise. Investigators can obtain high-quality data for analysis from filtered repositories.

3.4 Analysis Methods

Analysis methods provide insight into data and helps investigators identify trends and verify conjectures. Several tools support these analytical tasks. Indexing relates textual terms to individual Web pages, thereby supports precise searching of the pages. Extraction identifies meaningful entities such as complaints, frequently mentioned symptoms, and suspicious syndrome. Classification finds common properties among entities and assigns them to predefined categories to help investigators monitor the trends of public health communications. Visualization presents voluminous data in a format perceivable by human eyes, so investigators can picture the relationship between syndromes and diseases.

4 CHINESE HEALTH RELATED INFORMATION: A CASE STUDY

To demonstrate the effectiveness of our method, we apply the proposed method to infectious disease problems in China for collecting and analyzing the data from social media. The prevention and management of pandemic disease outbreaks have becoming a growing concern of the Chinese public health. Since a significant portion of Chinese population has difficulty in obtaining public health service, the problems caused by pandemic disease outbreak can be exacerbated. Recent outbreaks in China include: Server Acute Respiratory Distress Syndrome (SARS) epidemic in 2002-2003, human bird flu in 2005-2006, foot-and-mouth disease in 2008, and pandemic influenza A (H1N1) in 2009. These disease outbreaks have weakened public confidence in the government’s ability to respond to emergencies, undermined the nation’s social order, catalyzed regional instability, and caused adverse economic impact, including trade restrictions. We describe the steps of applying the proposed methodology as follows.

4.1 Data Collection

To collect data, we created a health consultation scenario by mimicking individual’s use of a search engine when searching for health related information. The idea was to examine several aspects of social media as represented by a search engine based on certain queries. These aspects include: (1) the proportion of social media among all the search results retrieved by the search engine; and (2) the way the search engine presented social media websites across different search result pages.

In this study, we focus the information search context to five infectious diseases. The five complaints, AIDS, hepatitis B, tuberculosis, influenza, and foot-and-mouth disease, were selected to represent the major threats to China’s public health service in recent years. They were selected based on number of physician visits, infected population sizes, as well as to reflect geographic diversity data published by Chinese Center for Disease Control and Prevention. This selection was deemed appropriate given the exploratory nature of the study.

We defined 8 keywords (in Chinese) in combination with the five complaints to form queries for the search engine. These keywords, including “symptoms”, “transmission”, “prevention”, “cure”, “diagnosis”, “vaccine”, “latent period”, and “medical inspection”, represent the top level health-related terms that will likely be used by individuals when they are looking for health information about a specific infectious disease. The selection of these keywords was based on the
heuristics of the search engine.

In this study, we choose Baidu the state-of-the-art Chinese search engine due to its high popularity among Internet users and dominant place in the online search market in China. In addition, Baidu offers a number of services to locate information using Chinese-language search terms, such as, search by Chinese phonetics, advanced search, snapshots, news, knows (provide users with a query-based searchable community to share knowledge and experience), postbar, images, and video information. As such, Baidu was considered the best “candidate” when assessing aspects related to online health communication in Chinese.

The 8 keywords in combination with the five complaints, as mentioned above, resulted in 40 queries, were entered in Baidu to obtain the search results. According to Rains and Karmikel (Rains & Karmikel, 2009), the majority of search engine users only review search results appeared in the first three pages. In this research, search results on the first 10 pages were retrieved so as to provide a more comprehensive representation of social in the context of health-related information search. First, we sent out a query in Chinese, e.g., “乙肝传播途径 (the transmission of hepatitis B)”, to Baidu. By specifying the number of pages (10 in this case), the URLs associated with each of the search results were spidered through parsing the contents of the Web pages. Fig. 2 shows a typical example of Baidu search result. In this case, the URL associated with the first line, i.e., “乙肝传播途径 百度知道 (the transmission of hepatitis B, Baidu knows)”, was extracted and saved into a database along with the complaint and search keywords that were used to generate the search results. This process was repeated until all keywords and complaints were used. Baidu provides 10 search results on a single search result page by default. However, the search engine also provides additional results on the first result page for certain queries. We then used the back-link search function of Baidu to obtain several hundred URLs that point to the additional results on the first result page. Table 1 presents the total of 6,559 results extracted for the five complaints.

<table>
<thead>
<tr>
<th>Complaints</th>
<th>Number of Search Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS</td>
<td>1150</td>
</tr>
<tr>
<td>Hepatitis B</td>
<td>1139</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>1147</td>
</tr>
<tr>
<td>Influenza</td>
<td>2110</td>
</tr>
<tr>
<td>Foot-and-mouth disease</td>
<td>1013</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6559</strong></td>
</tr>
</tbody>
</table>

Table 1. Search results retrieved from Baidu

4.2 Filtering and Coding

We conducted filtering and coding to pre-process the collected data. First, we manually filtered out unrelated sites, such as news or governmental websites that report public health events. We retained social media websites containing health related contents.

Two domain experts were employed to categorize the websites in the search results in a two-step
First, we employed two native Chinese speakers to categorize the 6,559 search results into social media and non-social media websites. The coders visited the Web pages by following the URLs and indicated whether the website belongs to social media. Inter-coder reliability was checked using Kappa statistics, which was 0.89, indicating there was a high level of agreement between the two coders. In total, 725 search results were identified as social media websites, which constituted approximately 11% of all the search results extracted.

The second step involved coding the identified websites into different types of social media. Based on the previous research in social media for health communication, we classified social media into four categories. They include: Web forums, Blogs, social networking sites, and Wikis. After the coding was completed, inter-coder reliability was once again checked. The kappa statistics was 0.77, which was a bit lower than the results of the first round of coding. Considering that it involved more categories, the results of the second round of coding still indicate reasonably good inter-coder reliability. The differences in coding were resolved by discussing with the coders until an agreement was reached for each case.

4.3 Analysis and Discussion

Among the total 6,559 search results, there were 725 (approximately 11%) identified as search results collected from social media websites. Given the otherwise rather fragmented nature of the search results, this suggests that social media, indeed, represent a substantial part of the online health service. We performed descriptive analysis, classification, and visualization on the 725 search results collected using an exhaustive breadth-first search spidering program.

We first plotted the distribution of unique domain names among the social media results using an Ogive chart in Figure 3. In total, there were 175 unique domain names found among the 725 search results. In this graph, the top 20 unique domain names with the highest frequencies represent approximately 50% of all the social media websites; and the top 67 unique domain names represent 80% of the social media websites. As such, there seems to be a “core” and a “long tail” in this distribution. That is, there are a relatively small number of websites forming the core of the social media portion of the domain, while a significant number of websites only occurred once or twice represent the long tail of the domain. The top 10 social media websites are shown in Table 2. As can be seen, these websites represent the most “popular” (as determined by Baidu) social media websites containing health-related contents, which can be used as major data sources for public health surveillance. It is interesting to note that the top 10 websites contain various types of social media including Web forums, Blogs, wikis, and media sharing websites. However, social networking
websites are not well represented among the most prominently discussed websites.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Frequency</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>baike.baidu.com</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>zhidao.baidu.com</td>
<td>14%</td>
<td>29%</td>
</tr>
<tr>
<td>120ask.com</td>
<td>13%</td>
<td>42%</td>
</tr>
<tr>
<td>tieba.baidu.com</td>
<td>12%</td>
<td>54%</td>
</tr>
<tr>
<td>ask.39.net</td>
<td>10%</td>
<td>64%</td>
</tr>
<tr>
<td>ks.cn.yahoo.com</td>
<td>9%</td>
<td>73%</td>
</tr>
<tr>
<td>bbs.39.net</td>
<td>9%</td>
<td>82%</td>
</tr>
<tr>
<td>club.health.sohu.com</td>
<td>8%</td>
<td>90%</td>
</tr>
<tr>
<td>wenwen.soso.com</td>
<td>5%</td>
<td>95%</td>
</tr>
<tr>
<td>healthbbs.net</td>
<td>3%</td>
<td>98%</td>
</tr>
</tbody>
</table>

Table 2. Top 10 unique domain names among social media search results

In the second analysis task, we examined whether there is any relationship between specific search queries and social media representation. Figure 4 shows the distribution of social media websites among the selected complaints. As can be seen, the number of social media websites for most of the complaints remains very similar except for foot-and-mouth disease. While this is a small sample, it seems that the number of social media sites for the major pandemic diseases are relatively stable.

Figure 4. Distribution of social media by complaints

Figure 5. Correspondence between keywords and types of complaints
To reveal the relationship between keywords and types of complaints, we used a radar diagram to perform the visualization task. This was intended to answer the question whether certain complaints would more likely to co-appear with certain types of keywords in the search results. Figure 5 shows the co-occurrence frequencies of complaint types and keywords. The eight sides of the radar plot represent the eight keywords of a certain complaint. Each of these eight dimensions represents a normalized scale between 0 and 1, showing the frequency of occurrence on the keywords. The co-occurrence frequency of complaint c on keyword k was calculated by the following formula:

$$\text{Co-occurrence Frequency} (c, k) = \frac{\sum_{i}^{n} \sum_{j}^{m} w_{i,j}}{m \times n}$$

where

$$w_{i,j} = \begin{cases} 1 & \text{keyword i occurs in complaint j search results} \\ 0 & \text{otherwise} \end{cases}$$

n = total number of search results containing the specified keyword k;

m = total number of search result belonging to the specified complaint c.

The radar diagram indicates that the people’s concern for various aspects of the complaint were fairly different across different type of diseases. For example, AIDS is highly associated with the keywords “symptoms”, “transmission”, and “medical inspection”. Influenza is highly associated with the keyword “vaccine”. We found that social media is frequently utilized to share knowledge of disease symptoms, transmission, and prevention. Social media include a variety of websites that allow patients to share their experiences in different ways, ranging from posting their stories, comments, to even pictures and video clips. Moreover, social media frequently appear on the first few search result pages in Baidu. This suggests that these social media sites are substantial in terms of their size, the up-to-date nature and the relevance of their contents, and the level of connectivity with other sites on the Internet.

5 CONCLUSION AND FUTURE DIRECTIONS

Collecting and analyzing public health information on social media has challenged both researchers and practitioners due to the abundance of Web information. It has also made it difficult to obtain a comprehensive trend in syndromic surveillance. In this study, we propose a method to address these problems. Using advanced Web mining, content analysis, visualization techniques, and human domain knowledge, the method exploited various information sources to identify and analyze 175 social media websites. Information visualization tool was used to help identify invaluable data sources for public health surveillance. Our evaluation showed that the method yielded promising results that would be useful to assist investigation of public health and have potential to guide policy-making and intelligence research.

Given its exploratory nature, this study has several limitations. In addition to the lack of comprehensiveness due to the limited number of keywords and complaints selected, this research employed a cross sectional study in which the data reflected only a snapshot of the social media represented through a popular search engine at one specific point in time and for a specific group of complaints. We plan to extend our study in two main directions to address the limitations of current study. First, we will conduct longitudinal studies capturing the role of social media over time in the dynamic environment of syndromic surveillance. Second, we plan to develop scalable techniques to collect such volatile yet valuable contents to visualize large volume of public health data and extract meaningful health informatics from social media websites. These efforts will help investigators trace and prevent pandemic disease outbreak.
References


