Social Web Information Monitoring for Health

Alessio Signorini
alessio-signorini@uiowa.edu

Department of Computer Science
University of Iowa

June 7, 2009

Abstract

The growth of social networks and blogging services radically changed the way users interact with the Internet. The so called "Web 1.0", seen by many as a giant, free library, is progressively being replaced by the more interactive "Web 2.0" where every user is at the same time a producer and a consumer. The results of the studies presented in this report demonstrated a high level of correlation between social activities performed online and public perception of current topics, which could be very useful to supplement the current data collection processes especially in the health context.

1. Introduction

1.1. Spread of the Internet

According to a 2004 study of the U.S. Department of Commerce, the number of households with a computer increased from 36% to 62% between 1997 and 2003, with 88% of these using the computers to connect to the Internet [2]. The same study reveals that the number of broadband connections doubled between 2001 and 2003, increasing from 9.1% to 19.9%.

A recent study by Nielsen Online reported that, as of November 2008, more than 81% of households have a computer and 92% of these use the computer to access the Internet [7]. These data are confirmed by another Nielsen Home Technology study which showed that more than half (57%) of American homes have access to a high-speed Internet connection.

Similar statistics can be found in the updated reports of Internet World Stats, which show 73% Internet penetration among the population of the US. While the world average is much lower (24%), Oceania, Australia and Europe closely follow North Americans with 55% penetration among their respective populations [10]. It is interesting to note how in Europe the northern states (e.g., Iceland, Norway, Finland and Netherlands) lead the chart with an Internet penetration of nearly 90%.

A Nielsen report on US Internet Usage shows daily usage by the average American of about 60 sessions online per month, viewing an average of 2,400 pages [8]. The same study reveals how the average male spends more than 70 hours online per month. The time spent online seems to grow with the age of the person, going from 25 hours per month among age 12-24 to more than 85 hours per month among older people. Clearly, the increased popularity of computers with high speed connections has changed the lives and behavior of millions of people. According to a Mediamark Research Survey done in fall 2008, many tasks that were once done manually are now typically completed online [6] (see Table 1).

Activity	% of American
Read News Online	46.00%
Pay Bills Online	39.60%
Personal Shopping	37.20%
Shared Photos	25.40%
Searched for Recipes	24.80%
Arranged a Travel	20.50%
Obtained Medical Advices	19.90%
Looked for Movies Showtimes	19.70%
Searched for Employment	15.30%
Traded Stocks	13.20%
Listened to the Radio	13.10%

Table 1: Percentage of Americans which performed the activity online in the last 30 days

1.2. Social Aspects of the Web

In addition to the more mundate, information gathering, tasks just mentioned, users seem to like sharing their knowledge and opinions on various topics. The most well-known collaborative effort is Wikipedia[1] a free encyclopedia created, edited and updated by users around the world. Since its creation in 2001, Wikipedia has attracted more than 75,000 editors who have created more than 10 million documents. Although the English version is the largest (2.8M documents), many of the articles are available in 260 languages and have attracted more than 684 million users globally since 2001.

Together with Wikipedia, other social networking sites found fertile ground in the Internet during recent years. A PEW report from January 2009 reveals that more than 35% of Internet users have a profile on a social networking site, a percentage which increases to 75% among 18-24 year old users [19]. Small university projects like Facebook[2] have became hugely popular, gathering more than 250 million active users of all ages. Users connect (or re-connect) with friends, partners and colleagues, sharing photos, videos, and other personal information.

This compelling desire to share fueled the need for real-time thought-sharing sites like Twitter[3], where anybody with a cell phone or computer can share (in no more than 140 characters) what they are doing or thinking right now. Founded in May 2006, the service attracts nearly 9 million people each month, who generate more than 70 million page views [11]. Similarly, according to MRI's survey, lots of Internet users also share their thoughts in their personal blogs [6]. About 11% of the users interviewed admitted to visiting blog sites, and more than 4% had one of their own.

Is it interesting to note how all of these sites imply graphs of users: for example, two Wikipedia users could be connected if both of them have modified the same document, while Facebook and other social networking sites allow users to explicitly declare their friendship. Users of Twitter can decide to "follow" each other, while bloggers usually refer to their friends' blogs in their own text.

1.3. Accessing Health Information

The growth of the Internet has also increased the amount of information accessible to the general public on any topic. The Web is seen by many as a giant (free) library where anything can be found. In fact, many schools and teachers have had to introduce strict no-Internet-references policies in their classes, forcing the students to find "real" sources as references for their assignments. The role of electronic encyclopedias, epitomized for many years by the Encyclopedia Britannica and Microsoft Encarta, has now been replaced with Internet-based crowd-sourced publications like Wikipedia. Many paper-only publications suffered a similar fate. Scientific journals and conference proceedings are often no longer offered on paper but rather distributed on some sort of electronic medium, such as DVDs or memory cards. All this content is also made available on a website, where it is easily found and indexed by the major search engines.

[1] http://www.wikipedia.org
[2] http://www.facebook.com
[3] http://www.twitter.com

For any type of content, accessibility and searchability are very important properties in today's connected world, where geographical locations and borders are less of an impediment. Thanks to the Internet, researchers in Italy can easily share their data with groups based in Tokyo, or compare their results with the early outcomes of similar studies done in Canada, all practically in real time. Experiments of this kind are already a reality: in November of 2000, a monkey at Duke University in North Carolina was connected through the Internet to a robotic arm in Massachusetts Institute of Technology's (MIT) Touch Lab, more than 600 miles away [1]. Maps of planets and stars, weather forecast and history, geological graphs and photos that were once available only to a restricted circle of scientists and graduate students, can also be found by anybody in just a few minutes.

Among the most popular sites are health-care related websites. Questions that were once answered by consulting the Medical Encyclopedia are now looked up online. Even small laboratory studies, which a few years ago were at best published in low-circulation venues, receive a lot of attention thanks to references from passionate bloggers who elevate these studies into the general media.

Other websites cover a wide range of possible medical-related necessities. Fitness and weight loss are among the most popular, with sites like Self[4], Men's Health[5] and Weight Watcher[6] leading the category with the highest number of visitors. Another popular category are the disease-centered websites, where any user can try to auto-diagnose by selecting the symptoms experienced and letting the site suggest possible causes. Among the most popular sites in this category we can find WebMD[7], Mayo Clinic[8] and Yahoo!Health[9]. Finally, support group websites (for addictions, substance abuse, or rare diseases) are also among the most visited. These are generally non-profit sites that aim to connect people in similar situations, to exchange information, help users deal with their shared problems and speed recovery.

1.4. Log Mining

Many of the tasks once completed manually can now also be accomplished digitally by simply connecting to some website. In a few cases, the physical form has been completely supplanted and its digital counterpart is now the standard. Shopping, paying bills, forwarding the mail, getting insurance quotes, reading newspapers, booking a vacation, checking the weather, watching tv, and banking are good examples of activities which are progressively moving to the web. According to a ComScore study, more than 44% of U.S. citizens use online banking, as do about 67% of Canadian citizens [4]. Not surprisingly, among users in the 25-55 age group, the penetration is even higher (about 75%). Users spend on average about 46 minutes per month on banking sites. Thanks to this increase of online activity and the spread of digital transactions (e.g., credit/debit cards payments) lots of data are now available in some digital form.

At the same time, computing power and disk space are becoming much cheaper: "cloud computing" (that is, out-sourcing your data and processing requirements to a farm of computers owned by somebody else

[4] http://www.self.com
[5] http://www.menshealth.com
[6] http://www.weightwatcher.com
[7] http://www.webmd.com
[8] http://www.mayoclinic.com
[9] http://health.yahoo.com

and paying some sort of rent) has become a viable option. These farms, built out of thousands of low-end machines, are usually situated in remote locations, where real estate and electricity are less expensive. Although each single unit is often not powerful enough, the parallelization software which runs them is capable of subdividing each job into multiple tasks which are then performed simultaneously on different machines; the output is recombined to provide an overall solution. The most common parallelization paradigm is probably Map-Reduce, invented and evangelized by Google's engineeers [15].

Many companies offer digital storage and cloud-computing solutions: Amazon was one of the first to introduce its Amazon Web Services[10] (2006), Google followed shortly with its Google AppEngine[11] (2008), and Microsoft is launching its own solution named Azure[12]. The availability of these solutions has revolutionized the Internet start-ups market. While in the past big investments were necessary to build scalable infrastructure, today it is sufficient to know some Python or Java and take advantage of cloud computing to launch a new service. For example, Twitter is entirely based on Amazon Web Services.

Cheaper disks and processing power also make it more convenient for companies to move paper-dominated businesses into the digital age, where every document or piece of information is stored somewhere in electronic form. This digital migration tremendously increases the amount of data available for reference and study purposes, which allows companies to reduce risks and expenses by optimizing their products and production lines.

Most of our everyday transactions are stored in digital form: parking tickets, bills, medical records, calendars and appointments, real estate prices, credit card transactions, phone calls, repair services, movie rentals, purchases, and so on. All of these data are collected together, usually in a relational database, and studied by marketing analysts. Although sometimes businesses ask our permission to collect these data, in most cases it is done without our knowledge or any explicit interaction on our part. In some cases, customers implicitly help to collect and aggregate these data in exchange for a small discount on their purchases.

Shop fidelity cards are good examples: they allow the business to track the purchases of each customer which are then used to optimize stocked products reducing costs and increasing sales. In exchange, the customer receives a small discount on their everyday bill and thus keep using the card and returning to the same business. During the application process for the card, usually the customer is invited to provide some personal information such as their birthday, address and family composition. When the card is used, the company is able to track and digitally store the items bought by the customer, the day, time and date of the sale, and the total amount spent. Looking at the data it is possible, for example, to infer that most families do the bulk of their grocery shopping during the weekend and thus the store might want to make sure to stock up on family products on those day. On the other hand, sales of perishable items like fish might reach their lows on those days (since Sunday fish is generally perceived as "less fresh") and thus the company might want to avoid introducing a new Sunday delivery.

Using the same data, stores (especially if they have a built-in pharmacy) might also deduce health-related information about each of its customers. For example, by monitoring purchases the store might detect

[10] http://aws.amazon.com
[11] http://code.google.com/appengine/
[12] http://www.microsoft.com/azure/

when somebody is sick (e.g., they bought a new medicine), what illness they have (e.g., they bought cough medicine), and possibly also when they start to get better (e.g., they did not come in for a refill, or they skipped last weekend's shopping but returned the following week).

This process is very powerful and can be extended to acquire almost any sort of information: there is a new baby in the family (e.g., they started buying baby food), their kids now go to school (e.g., they bought school supplies right before the semester starts), somebody wants to lose weight (e.g., they started buying Weight-Watchers products) or have already lost weight (e.g., they started buying smaller clothes), they own a PlayStation (e.g., they started buying video games), and so on.

Analyzing massive amounts of data in search of useful (and possibly unknown) patterns requires advanced computer science algorithms and methods. These techniques belong to the "data mining" field, defined by the Encyclopedia Britannica as:

> Type of database analysis that attempts to discover commercially useful patterns or relationships in a group of data. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neuralnetwork techniques. A major goal of is to discover previously unknown relationships among the data, especially when the data come from different databases. Businesses can use these new relationships to develop new advertising campaigns or make predictions about how well a product will sell. – *Concise Encyclopedia Britannica, "data mining"*

Data mining techniques have become more useful since massive computing power and digital data collection become available. Knowledge Discovery in Databases (KDD) is the name coined by Gregory Piatetsky-Shapiro in 1989, who outlined the foundations of data mining in a fall 1996 paper [13]. Broadly speaking, data mining generally involves four types of tasks:

- **Classification** Often referred to as "supervised learning", these methods are used to partition data into predefined groups. A common example is the spam detection filter of an email client, which attempts to classify incoming email as legitimate or spam. Well-known algorithms include nearest neighbor, naive Bayes classifier and neural networks.
- **Clustering** Similar to classification, but with no predefined groups, these methods are also named "unsupervised learning". The algorithm will use the given "similarity function" to evaluate the differences among data points and group similar ones together.
- **Regression** These methods attempt to find a function which models the data with minimum error. While classification and clustering partition the given data, these methods try to infer the function which allows a correct partitioning. Genetic programming techniques are often used for this task.
- **Rule Learning** Searches for relationships among variables. Supermarkets often use association rule learning to determine which products are frequently bought together and then utilize this information for marketing purposes. These kind of analysis are often referred to as "market basket analysis".

The application of data mining techniques is not limited to marketing purposes. Those methods are widely used across many other fields. One of the early adopters of data mining techniques is the insurance business. Generally speaking, the purpose of these companies is to cover your expenses in case of accident, in exchange for a fixed monthly premium: auto and health insurance companies represent a big sector of the US economy. Those companies make a profit if the premium paid by their customers generates higher earnings than the expenses necessary to cover the generated claims. Accident risk levels vary for different classes of individuals, as well as the maximum premium that those individuals are willing (or able) to pay. For these reasons, data mining techniques acquired an extraordinarily important role allowing insurers to uncover patterns, trends and similarities in customers claims data. For example, they found that young 25 year old males driving red sports cars are the most prone to accidents.

Credit cards issuers are also "big customers" of data mining technology. These enterprises allow their customers to buy costly items and pay them off with monthly payments. Although already common, the recent increase of Internet shopping helped credit cards usage to grow exponentially, quickly becoming the standard way to pay on the web. Unfortunately, it is sufficient to know the card number and some personal information to fraudulently purchase items online. For this reason, data mining techniques became widely used to detect illegal card activities. Purchases made at uncommon times or on questionable foreign sites are usually a good indicator of potentially fraudulent activities.

Online shopping activities have increased at a steady pace since 2005, generating about 10 billion dollars of purchases per month in 2008 [5]. The broad use of web search engines and better targeting of online advertising allow users to easily find what they are looking for. The techniques applied by marketers in grocery shops for many years have been ported and perfected for online business. Browser cookies replaced fidelity cards allowing companies to track user's activities and purchases on their sites (and often across partner sites) providing a massive amount of information on which to apply data mining techniques. Amazon[13], the colossal online book store, was one of the first online companies to introduce and exploit data mining techniques on their site, recommending books which a user might like based on purchase histories.

A similar strategy has been pursued by Google, both in its search engine and for its online advertising business. Every query is saved together with user location (mapped from user IP address), time and date, and the results selected. This massive amount of data (Google receives more than 8.5 billion searches per month [9]) is later mined for similarities among searches and click patterns, and allows Google to, for example, improve local search results (e.g., if everybody looking for "pizza" in the user's neighborhood clicks on a specific link it probably is a good result for that location). At the same time, advertisements shown during search with relative clicks and conversion rates (i.e., the percentage of people who make a purchase following the ads) are also carefully collected and mined in search for patterns and new marketing ideas.

[13] http://www.amazon.com

1.5. Privacy Concerns

While the broad availability of customer data and the recent improvements in data mining techniques please marketers and companies, they raise many privacy concerns among users and customers. The idea that so much data has been collected about one's activities and that all these data sources could potentially be liked together to produce an accurate and complete picture of each user can definitively raise some concerns.

In her 1998 report, Ann Cavoukian, Commissioner for the Ontario Information and Privacy Committee, claimed that data mining "may be the most fundamental challenge that privacy advocates will face in the next decade" [13]. In her report, she recommends that, at the moment of purchase, customers be given a choice among 3 levels of opt-out policies:

1. Do not allow any data mining of user's data;
2. Allow data mining only for internal use only; and
3. Allow data mining for both internal and external uses.

Privacy concerns are even more pressing when dealing with medical data, since a data leak or massive data aggregation could influence an individual's insurance status. Hundreds of papers and books have been published in recent years just on this topic, with the aim of exposing the flaws of the system and increasing the confidence in data mining techniques with solutions that allow anonymous aggregation of the data while preserving its important properties. Most of the solutions proposed, as for example the one published by Segre & al., take advantage of cryptographic algorithms to scramble identifying fields while still allowing statically useful data analysis [22].

2. Information Monitoring on the Social Web

In this work we focus on the health sciences, collecting, studying and validating available data as an additional signal to monitor and better manage diseases outbreaks.

2.1. Current Surveillance System

According to the Center for Diseases Control and Prevention[14] (CDC), cases of a disease or other condition of interest are primarily identified within the health care system. Once identified, cases are typically reported to a local health department, often using paper-based data collection forms. At the local health department, forms may be entered into a computerized electronic data management system and transmitted to the State, or they may be copied, filed at the local level and then sent directly to the State where they are manually entered into the State health department's electronic system. Some of these data may then be aggregated at Federal level. These reporting processes are generally the same, regardless of the disease or condition that is being reported.

There are a number of problems that can arise during the reporting process. These problems, in turn, often place a large burden on medical care staff who have responsibility for disease reporting. For example, cases may be reported from a variety of sources from within the health care setting (such as clinical laboratories and private physicians) whose staff are already overworked. Nevertheless, it is often left up to health care providers to determine if a case meets public health surveillance case definitions and to figure out how to fill out the wide variety of forms produced by CDC and health departments. They may also spend significant time tracking down patient records in response to requests for more information from the health department.

To reduce the burden imposed on medical care staff, minimize human error, and facilitate the transmission of these important medical data, the CDC designed and introduced the National Electronic Disease Surveillance System (NEDSS). NEDSS facilitates the collection of case report forms from providers in two important ways.

First, standards are being developed to assure uniform data collection practices across the nation. The public health data model and common data standards will recommend, for example, a minimum set of demographic data that should be collected as part of routine surveillance. In addition, guidelines will provide a consistent method for coding data on the data collection forms. It is expected that standardizing data collection forms should ease the burden on physicians and their staff by providing a more uniform data entry portal for all reportable conditions via secure web-based systems or, for larger organizations via electronic data exchange that is automatic and imposes minimal burden on health-care reporters.

Second, NEDSS will include recommended standards that can be used for the automatic electronic reporting of surveillance data. Specifically, NEDSS will recommend a standard data architecture and electronic data interchange format to allow computer systems to automatically generate electronic case reports that can be sent to local or State health departments. These types of standards would both ease the burden on large organizations that already have computerized data systems (such as regional laboratories,

[14] http://www.cdc.gov

hospitals, managed care organizations) and would ensure that all cases that are in the providers data systems are reported to public health officers.

2.2. Possible Applications of Social Web Activity Monitoring

Although NEDSS will surely improve the effectiveness of US health care surveillance systems, it still relies on a small number of humans (e.g., doctors or nurses) to manually report cases of diseases or conditions they encounter. Whether submitted using paper forms or electronically, the system relies heavily on each medical office's efforts to find the time to promptly transcribe and report their cases. Doctor's offices are notoriously under-staffed, especially when economic conditions are poor. Moreover, many people do not consult a doctor for what they perceive to be common or minor health problems. In fact, a Consumer Health-Care Product Association[15] survey report that nearly 80% of Americans relied on over-the-counter medications to treat a personal condition and that 73% would rather treat themselves at home than see a doctor [24]. For these reasons, it is very likely that many potentially interesting diseases and conditions will remain unreported and thus undetected.

The spread of the reach of the Internet and the increase of social web activity could represent a good supplement to official data. On a daily basis, millions of social network status updates, blog posts and search queries travel through the network. In these messages, people express their feelings, look for solutions to their problems, or seek suggestions from peers. Monitoring and analyzing these data could provide hints on the perception and mood of the public with respect to certain health subjects, as well as clues to new and potentially unreported outbreaks.

2.3. Query Log Analysis

Until just a few years ago, many groups and individuals published lists of their favorite web pages focused on specific topics. The linked structure of the Internet allowed users to start from these "hubs" and follow the links to discover new, interesting, content. In exchange, the user might create and publish their own favorite list, and the cycle would repeat. Unfortunately, the rate at which new pages are created and old ones disappear made the task of creating and maintaining such lists manually very time consuming. Moreover, with the recent introduction of social network profiles, blogs and dynamic-content websites, it is sometimes impossible to provide a direct link to dynamically generated resources.

Theese issues make the use of web search engines a necessity. Every day people rely more and more on the results provided by search engines to accomplish many tasks, even not strictly related with the web. For example, almost all the current search engines allow users to discover the current time in various cities of the world (e.g., search for "time in Rome, Italy" on Ask.com) as well as movie theater listings (e.g., search for "80302 movies" on Google) or the correct spelling of a word (e.g., search for "analizing" on Yahoo!). As the reach of the Internet grew, people also started using web search engines as substitutes for their medical encyclopedias to find updated information on health questions. The creation and diffusion

[15] http://www.chpa-info.org

of health-related websites (e.g., health.com, webmd.com, mayoclinic.com) encouraged an increase in this behavior.

All the queries submitted to a search engine by its users are aggregated and saved for later analysis in databases which are commonly referred to as "query logs". Over the past few years, query log analysis generated many interesting studies in a broad range of fields. Google Flu[16] is the best-known query log analysis effort. In their paper the authors analyzed hundreds of billions of queries contained in 5 years of Google query logs [16]. The query logs were anonymized, but information about the location of the users (obtained through geo-location of the source IP address) was retained to provide localized statistics. Flu-related queries were automatically identified by an automated classification system developed at Google and their daily count was normalized by the total number of queries performed on their system on each particular location. The results obtained during their experiments were validated against official CDC data on Influenza-Like Illness (ILI) doctor visits.

During their experiments, the authors identified 45 search queries which are significantly more useful in predicting the number and location of ILI-visits as depicted by CDC data. These queries were then used to to develop a linear model using weekly ILI percentages between 2003 and 2007. The model was able to obtain a good fit with CDC-reported ILI percentages with a mean correlation of 0.90. The model was also validated against an previously untested data from 2007 through 2008 and showed a mean correlation of 0.97. Data from the state of Utah allowed the authors to test the model on a more local scale, obtaining a mean correlation of 0.90.

The findings of this study confirmed the results of an earlier study conducted by Polgreen & al. using Yahoo! search queries [21]. In their study, the authors studied the correlation between the percentage of ILI-related queries and official CDC data, developing a linear model which allows to predict influenza outbreaks 1-3 weeks in advance. A similar model was also developed to predict an increase in mortality attributable to pneumonia and influenza up to 5 weeks in advance.

In both experiments, the queries used were identified by the presence of a few specific influenza-related terms. Although user queries are usually very short, it could be interesting to apply some more advanced classification methods to the query logs and extract a bigger collection of health-related queries. The use of a larger dataset could perhaps improve the precision of these methods or increase the lead time of outbreak predictions.

2.4. Blog Posts Analysis

Another recent Internet trend is the increased popularity of personal blogs. While company blogs are usually used to advertise new services or products, personal blogs can be thought as the modern version of old-fashioned "secret diaries". According to recent estimates, about 900,000 blog entries are published every day [23]. A recent (March 2008) comScore[17] study reported that globally a total of 346 million of users read blogs.

[16] http://www.google.com/flutrends
[17] http://www.comscore.com

In their blog posts, people express personal feelings and opinions about life, products, recent news or events. Since many users threat their blogs as a personal diary, the language adopted and the entities cited can often allow the identification of many personal details. For example, it is not uncommon to find posts titled "my 30th birthday", which allow analysts to determine the age of the writer with high precision. Some posts may describe an evening out, mentioning identifiable landmarks (e.g., "we got a cab to lower Manhattan"), places (e.g., "Time's Square was packed") or venues (e.g., "we had dinner at the Four Seasons"). Other posts offer clues about the gender of the writer, for example, comments about a new pair of shoes, relationship problems or a new dress might suggest a female writer, while opinions on the current situation of the stock market or the weekend's sport results, increase the probability of facing a male blogger.

While such details might help to identify the location, gender and age of the writer, the complexity of the language used in the posts makes it difficult to automatically identify the mood and attitude of the writer (e.g., happy, confused, frustrated) as well as the category of the post (e.g., sports, politics, history). Although difficult to achieve, automatic categorization of blog posts could be very useful in many occasions, as for example while trying to summarize the opinion of the public about certain products or topics.

There have already been many attempts to classify blog post. In 2005, Gilad Mishne published a paper describing the early outcomes of his experiments leading to the development of MoodView[18] [20]. In his work, Gilad obtained about 850,000 mood-annotated blog posts from LiveJournal[19] and tried to identify discriminative features (and their weights) in the post's text for each different mood. Unfortunately, the precision achieved by the method tested is barely above (67%) the baseline (50%, random guess) and more work is clearly necessary to make it usable.

Similar work has been been published by Paula Chesley et al. in 2006 [14]. In their work the authors simplified the approach taken by Mishne and tried to classify the posts into just 3 main classes: objective, positive or negative. The classification method was based mainly on the identification of the polarity of adjectives and verbs which they obtained from Wikitionary and the weight of each term was computed using Support Vector Machine (SVM) classification. The final accuracy of the method was close to 90% both for verb and adjectives.

Automatic classification of blog posts could be really useful in identifying the perception of the general public of some products or topics. In the health context, it could be useful to identify moods and opinions about certain diseases or vaccines which might permit public health officials to better address problems and concerns.

2.5. Social Status Update Analysis

In the last few years the increase in popularity of blogs coupled with the exponential diffusion of cell phones and PDAs created a new kind of services commonly referred to as "micro-blogs". These services remove the technological barrier (usually a computer and Internet access) imposed by a blog, allowing users to update their status or submit a new post from nearly anywhere using their phones. In many cases these services

[18] http://www.moodview.com
[19] http://www.livejournal.com/

also allow users to supplement their messages with pictures or videos, which an increasingly number of mobile devices now support.

The most popular micro-blogging service is Twitter, which boasts more than 30 million users world-wide and receives more than 10 million updates per day. Another very popular service in this category is the Indian SMSGupShup4[20] which recently announced the registration of its 20 million-th user. In the last few months, Facebook[21], the popular social networking service, has also introduced some micro-blogging features which allow its users to update their status using their mobile phones. Although fairly new, micro-blogging is increasing its popularity among users of all ages which use it to exchange ideas and share opinions about products or events. As with traditional blogs, many users of micro-blogs use them to keep a live journal of daily life. They often start publishing status updates as soon as they wake up in the morning and then detail places and feelings until late at night.

Analyzing and classifying this stream of real-time information could be very useful for early detection of diseases outbreaks as well as to measure the public perception of certain products or topics.

2.6. Proxy Log Analysis

While a large number of people use search engines, blogs, and social networks, the traffic generated by these services represents only a fraction of total Internet traffic. Moreover, while users might not feel comfortable sharing their symptoms in a blog post or their Facebook profile, they may well visit a health-related website seeking suggestions or remedies for their symptoms.

Companies and universities (but in general every big institution) generally route the traffic generated by the internal network through firewalls and proxies. These systems are put in place to safeguard internal data and forbid certain operations (e.g., downloading illegal software), and usually log on disk all the URLs visited by the users, together with the originating IP address and time of the day, for debugging purposes. With access to these logs one could not only see the query traffic of many search engines (since the query is embedded in the final URL, e.g., `http://www.google.com/search?q=<query>`) but also the requested pages related to certain topics (e.g., http://www.webmd.com/cold-and-flu/swine-flu/). These data could be of help identifying symptoms of outbreaks or other health-related concern.

[20] http://www.smsgupshup.com
[21] http://www.facebook.com

3. Current Results

In this section we present some preliminary results.

3.1. Monitoring the Swine-Flu Outbreak

Novel influenza A (H1N1) is a new flu virus of swine origin that was first detected at the beginning of April 2009 in some regions of Mexico. This mutation of the virus, capable of infecting humans, spread from person-to-person sparking outbreaks of illness all over the United States. An increasing number of cases have been reported internationally as well. The CDC issued the first outbreak report on April 23rd, 2009 after which human cases of H1N1 infection were identified in San Diego County and Imperial County, California as well as in San Antonio, Texas. Media outlets all over the world depicted this pandemic as disastrous, forecasting thousands of deaths and hospitalizations.

On April 26th, the CDC published some general prevention guidelines (e.g., cover your nose and mouth when sneezing, wash your hands often, ...) while announcing that face masks had been distributed in community settings where spread of influenza was detected. In the same update, the CDC announced that the virus seemed to be susceptible to common antivirals such as Tamiflu and Relenza. While the number of new cases identified increased only by a few dozen per day, the inscrease in number of articles and news reports published was hundreds of times higher, making the spreading of the influenza a common topic of discussion. Fearing an epidemic many prepared for the worst and stockpiled in food, water and medical supplies. Travel to and from Mexico, but also anywhere within the U.S. was curtailed, and in many airports passengers and workers started wearing surgical masks at all times.

Fueled by the desire to monitor and estimate the response to the situation, on April 28th we started collecting the related posts on Twitter. Using their search API[22] we retrieved all the H1N1-related tweets published between April 28th and May 15th. Each entry was timestamped and contained various information about its publisher (for example, geo-location). Swine-related entries have been identified by searching through Twitter's public stream for tweets matching specific keywords, as for example:

- swine AND (flu OR influenza)
- H1N1
- (face OR surgical) AND (mask OR masks)
- relenza OR zanamivir
- tamiflu OR oseltamivir
- (hand AND (wash OR washing)) OR handwashing

[22] Application Programming Interface

At the same time we created a client-side JavaScript application for live monitoring of H1N1-related tweets published in the United States. This interface continuously updates a Google map with the last 500 most recent matching tweets, yielding a real-time view of flu-related public sentiment. Users can read any tweet by placing the cursor over its corresponding colored dot on the map (see Figure 1).

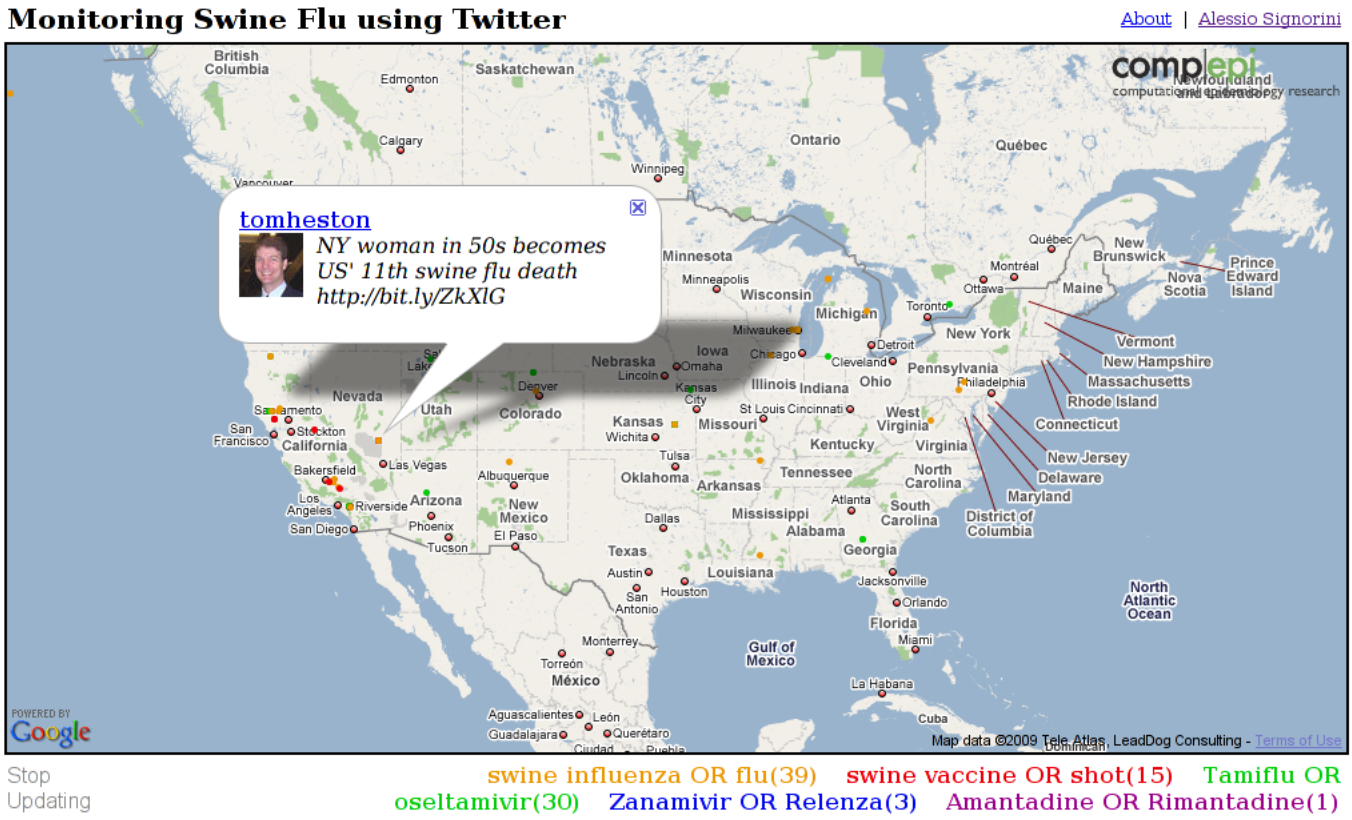

Figure 1: Client-Side JavaScript for live Monitoring of H1N1-related Tweets

During the period in question we collected a total of 592,543 H1N1-related tweets. After removing irrelevant entries from the data, we aggregated the tweets into categories (e.g., antiviral, handwashing, masks) and compared their temporal distributions with the increase of cases identified, and the public announcements of the CDC and other health organizations.

Looking at the graph it is interesting to notice how the majority of the tweets were published before May 7th, when the number of cases detected was still well under one thousand. The volume of conversations on H1N1-related topics does not seem to grow proportionally with the number of cases detected, which might suggest a high correlation between the tweet stream and the general perception of the outbreak.

Moreover, the peaks of the various categories immediately follow public announcements of health organizations. For example, tweets containing references to antiviral drugs peaked on April 29th, the same day in which the World Health Organization (WHO) raised the pandemic warning level to 5, but

fell as soon as official reports indicated that most U.S. cases were relatively mild and did not require hospitalization. Nevertheless, as a reaction to the warning, numerous media agencies republished the safety guidelines issued by the CDC just a few days earlier, which probably generated the peak of tweets with hand-sanitizing references on the following day.

In addition, many health organizations recommended canceling all unnecessary travel and wearing surgical masks (as a precaution) while in crowded public spaces (such as planes or airports). It is interesting to note how the number of tweets corresponding to these topics peaked on the announcement's following day (May 2nd, 2009).

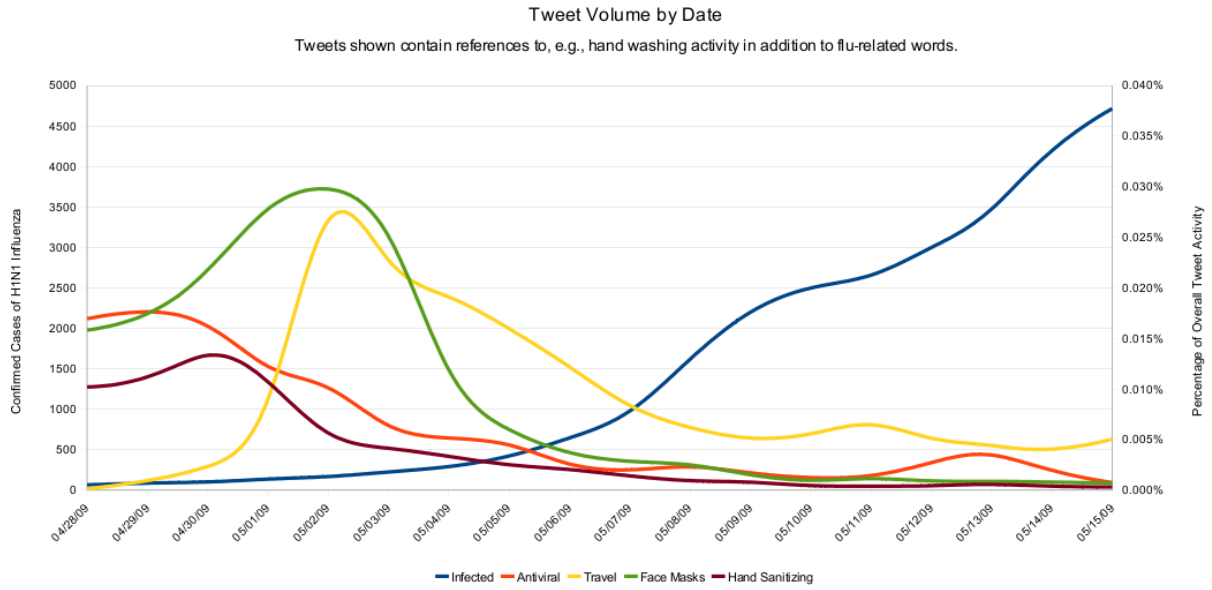

Figure 2: H1N1-related Tweets between April 28th and May 15th, 2009

This evidence suggests that it might be possible to use tweet analysis as an inexpensive way to determine not only the levels of anxiety and concern but also to gauge response to news and official public health messaging.

3.2. Predicting the American Idol 2009 Winner

American Idol is a reality-show competition to find new solo musical talent. It debuted on June 11, 2002 and has since become one of the most popular shows on American television. The program is a spinoff from Pop Idol, a reality program created by British entertainment executive Simon Fuller and first aired in 2001 in the United Kingdom. The program conducts a series of nation-wide auditions looking for the best singers. The American public decide the outcome of the final stages through telephone voting. The judges (usually record producers, singers or music executives) on the show give critiques to the contestants after each performance. On American television, the show is usually aired on two consecutive weekdays: on the first evening each contestant performs one or more songs, and on the following night the outcome

of the phone voting is announced and one of the contestants is sent home.

Given the popularity of the show and the fact that its target audience is compatible with the power users of Twitter, we decided to perform some analysis on the AI-related tweet stream. We used the Twitter search API to retrieve tweets which contained the name of each of the last 5 contestants of this season's American Idol (Adam Lambert, Danny Gokey, Matt Giraud, Allison Iraheta and Kris Allen) between April 28th and May 20th. Since the number of fans of each contestant might vary, we decided to normalize the data of each contestant by its average number of daily tweets for the period in exam.

Interestingly, we observed that even on such a small time scale the sequence of peaks of the candidates' tweets closely track the order in which they sing during the show. Figure 3 depicts the distribution of the tweets for each contestant during the night of May 5th, 2009. Analyzing the sequence of the peaks it is possible to reconstruct the order in which the contestants sang: Kris Allen, Adam Lambert, Danny Gokey and Allison Iraheta (see Figure 3).

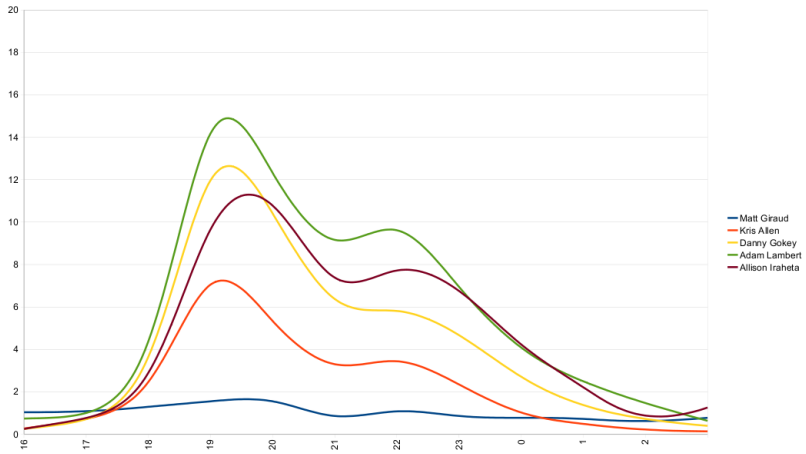

Figure 3: Tweets for each contestant during performances night, May 5th, 2009

On May 20th, 2009 (the morning before the finale of the current season), Lara Hejtmane published on the well-known blog Mashable an article/study in which she tried to apply Google Flu's prediction model to guess the outcome of the popular TV show [18]. In her study, Hejtmane observed how the distribution of queries based on finalists' names closely matched their final order in the show for each season. Figure 4 shows search query trends for American Idol 7. Analyzing the most recent query trends the author announced that Adam Lambert should win season 8 of American Idol. This curious use of search query trends allowed the article to gain substantial popularity on the web and the prediction made by Hejtmane was endorsed by many other bloggers.

Inspired by this article, we studied the number of tweets published for each finalist during the 2-night finale. Our analysis discovered that the total number of tweets published for each contestant did not offer any particular clue on who might be public's favorite. For this reason, we decided to limit our analysis only to positive tweets (e.g., containing words like "love", "best", "win").

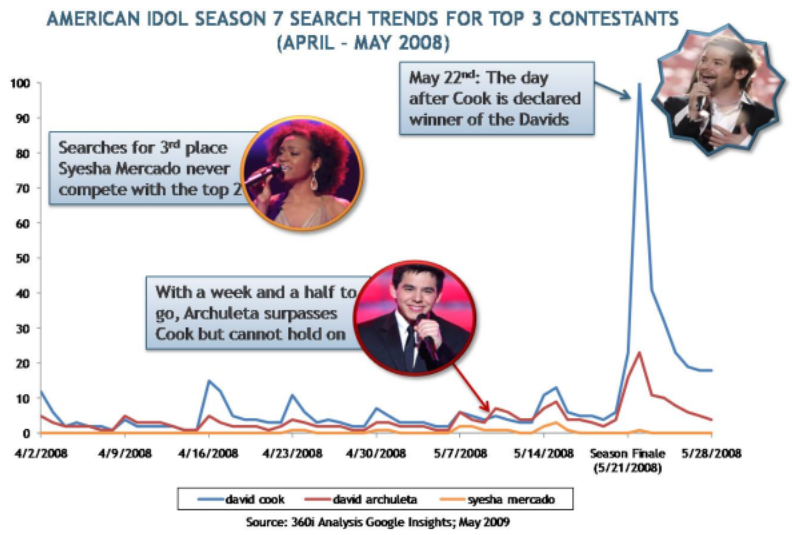

Figure 4: American Idol 2009, Search Trends for Top-3 Contestant

Figure 5 shows the relative number (or "frequency") of positive tweets obtained by each contestant during their final performance. Observing this graph, Kris Allen's performance seems to have received a higher public appreciation with respect to that of his rival Adam Lambert. Confident in our data, we published a blog post announcing our prediction for the winner of this season of American Idol [17]. Although our post did not receive the same attention as that of Mashable, our prediction was the more accurate: Kris Allen won American Idol season 8.

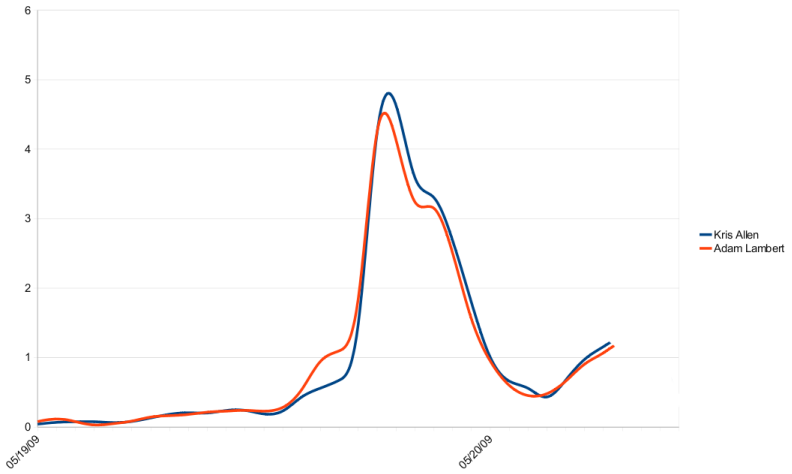

Figure 5: Positive tweets after American Idol 2009 performances

The results obtained in this other (more frivolous) experiment seem to confirm an high level of correlation between the tweets published regarding certain topics and corresponding public opinion.

4. Future Work

The results obtained in our early experiments is very encouraging. Social web activities seems to be highly correlated with public perceptions of certain topics, and the aggregation of information implicitely released in public conversations (e.g., a tweet or a blog post) could be very effective in a public health context. The following is a brief outline of the future works planned.

4.1. Twitter

In our early studies we retrieved relevant tweets matching specific keywords with their content. While simple to implement, this approach tends to collect a lot of noise (especially for broad terms like "flu") and may fail to identify other possibly interesting conversation, (for example, when synonyms are used). Although due to their length tweets do not offer much context to work with, the precision of the detection could probably be increased with simple grammatical analysis coupled with other well-known classification techniques.

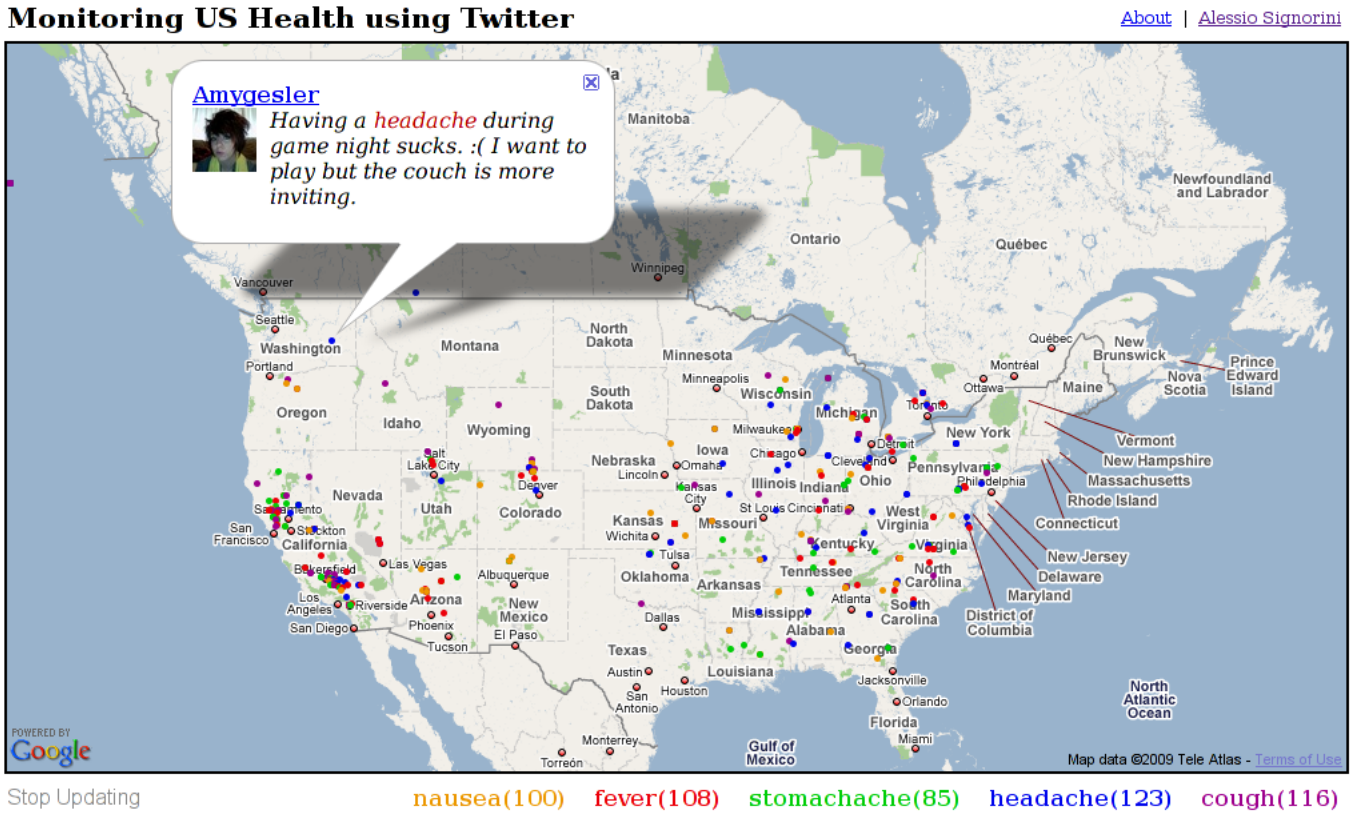

Figure 6: Health-related Tweets Live Map

A similar approach could also be used to identify the mood and the tone of the messages exchanged, which could be very useful to measure public perception with respect to certain topics. The Twitter Search API already attempts to detect the "mood" of each message looking for the presence of e-smiles (e.g., ":-]"). Unfortunately, according to our analysis more than 92% of the tweets do not include these features, or the publisher adopted other ways (e.g., LOL) to express their feelings. Implementing lexical analysis could produce interesting results.

Finally, we intend to study the geo-location data that are associated with tweets, for example, extending our Twitter monitoring system to capture more generic flu-related symptoms and comparing the data collected with official CDC reports. A new client-side JavaScript map interface (Figure 6), capable of capturing tweets correlated to generic health topics, has already been released. Many Twitter users seem to enjoy sharing with the world their arrival by plane in some new location. Every day thousands of messages containing "landed in ..." are published on the service. Utilizing tweets' implicit geolocation, or the home location specified by the user in their profile, it might be possible to create an approximate map of flight traffic.

Matching the data collected with official CDC reports of disease outbreaks could lead to the discovery of new interesting patterns which can then be used by medical authorities to supplement the information collected through traditional channels. Although widely used in our early experiment, Twitter conversations are not the only interesting source of social activity. Social networks (e.g., Facebook or MySpace), blog posts and news releases are other important sources of data which we will include in our experiments.

4.2. Wikipedia

Wikipedia Wikipedia is one of the principal sources of information on the Internet and its pages often appear among the top 3 URLs in search engines results. This free encyclopedia is maintained by thousands of passionate volunteers all over the world who constantly create, update and perfect its articles. In the past few years, Wikipedia has been very fast in reacting to new trends and topics. Deaths of celebrities and major political events were often captured on its pages only few minutes after the corresponding event.

During the recent swine flu outbreak, information about recent events was published in the "Swine Influenza" article on April 24th, just minutes after the first CDC public announcement, and a dedicated article was created on the following day. The "2009 Swine Flu Outbreak" article on Wikipedia received 1.5M visits during its first 5 days, with a peak of 417,200 on April 29th. Figure 7 shows the distribution of page views for the month of May 2009. This suggests that monitoring pages visits, creations and updates could offer an accurate picture of the most interesting current topics as perceived by the general public.

4.3. Blog Posts

While in the past years there have been many attempts to classify the mood of blog posts, to the best of our knowledge nobody has examined health-related posts for mood and topic. Well-known classification techniques might be useful for this task and already showed good precision in early tests. Using the location declared by the user in their profile, or alternatively trying to guess it through clues found in the posts,

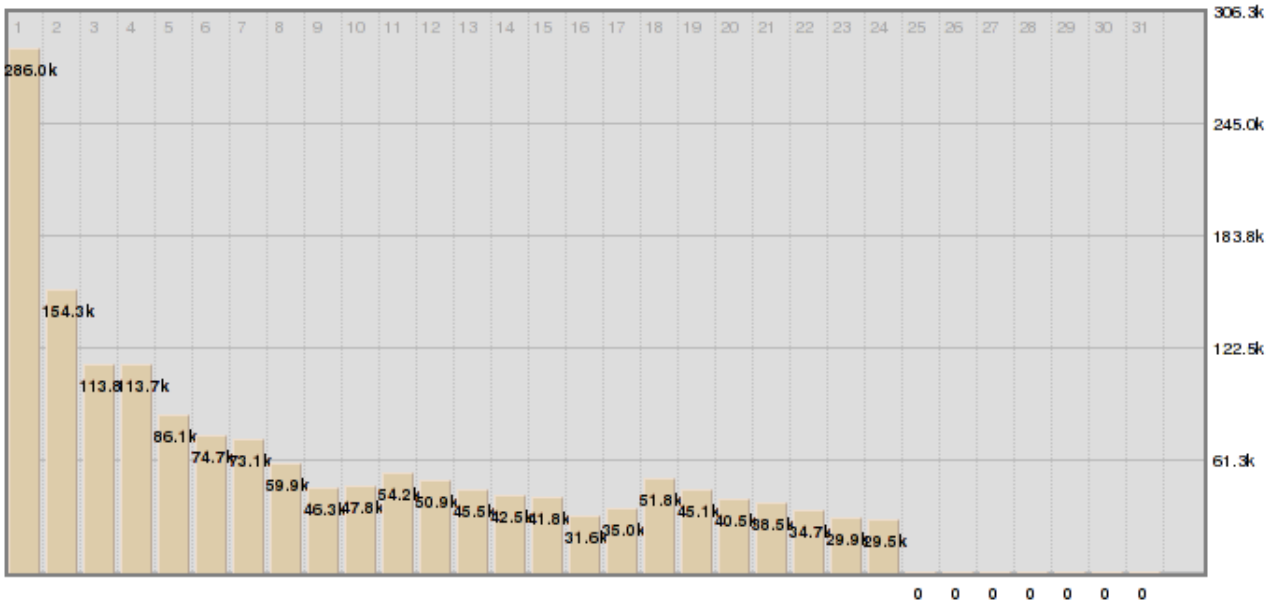

Figure 7: Page visits of Wikipedia article "2009 Swine Flu Outbreak", May 2009

it might also be possible to plot the data on a map to be compared with official local reports collected through the traditional channels. Finally, it might be interesting to compare the data collected with the number of articles in main-stream news channels, to identify the degree of correlation and the delay (if any) in public response.

4.4. Browsing History and Search Queries

The majority of Internet users probably do not own a blog nor use Twitter. However, conducting searches and browsing pages could still provide lots of useful information about their perception of current events. For example, it is possible that health-related websites like WebMD and MayoClinic receive an higher-than-usual amount of traffic when a pandemic warning is in effect. People, fearing to have contracted the illness, probably visit such sites to compare their symptoms with the ones reported in their databases.

Since most of the search engines display the query performed in the URL of the results page, users' browsing history could also be used to extract a large amount of query terms across all the different search engines. While previous research (e.g., Google Flu Trends) focus on a very specific set of queries, health-related queries represent more than 7% of daily search engine traffic and offer a very interesting untapped source of information.

The analysis of traffic logs might also support the study of correlation between user's behavior and medical data on a variety of topics. For example, given that Utah is one of the states with higher incidence of sunburn, it would be interesting to observe if correlated pages receive more visits from Utah than the rest of the country, or if there is a difference in the number of sunburn-related search queries [3]. Similarly, it would be interesting to analyze search queries and Internet traffic from users in Louisiana and Delaware,

which, according to recent (2005) statistics published by the CDC and the National Cancer Institute (NCI), have the highest incidence of cancer [12].

5. Conclusion

The growth of social networks and blogging services radically changed the way users interact with the Internet. The so called "Web 1.0", seen by many as a giant, free library, is progressively being replaced by the more interactive "Web 2.0" where every user is at the same time a producer and a consumer. The results of the studies presented in this report demonstrated a high level of correlation between social activities performed online and public perception of current topics, which could be very useful to supplement the current data collection processes especially in the health context.

Our future studies will extend and expand the experiments presented in this paper. We will analyze many sources of social activity data (e.g., Twitter, Wikipedia, Blogs, ...) in search of trends and patterns which can be directly correlated to the public sentiment. Well-known data mining techniques (e.g., classification, clustering and entity extraction) will be used to extract and identify the features and characteristics of each trend. Our experiments will be correlated to current events (e.g., seasonal flu trends) and validated against official health data (e.g., CDC flu reports) so that any findings can be directly applied by health authorities.

References

[1] *Monkey controls robotic arm using brain signals sent over Internet.* MIT News Office, December 2000.

[2] A nation online: Broadband age. Technical report, US Department of Commerce, September 2004.

[3] *Morbidity and Mortality Weekly Report*, volume 56, pages 524–529. Center for Disease Control and Prevention, June 2007.

[4] *Canada Leads World in Online Banking Usage.* ComScore, July 2008.

[5] *Holiday E-Commerce Spending Accelerates in Most Recent Week as this Year's Compressed Shopping Season Increases Urgency to Spend.* ComScore, December 2008.

[6] Internet activities, us adults with access to the internet. Technical report, Mediamark Research & Intelligence, November 2008.

[7] An overview of home internet access in the us. Technical report, Nielsen Company, December 2008.

[8] Topline u.s. data for march 2008. Technical report, Nielsen Online, March 2008.

[9] *comScore Releases February 2009 U.S. Search Engine Rankings.* ComScore, February 2009.

[10] Internet usage statistics. Technical report, Internet World Stats, March 2009.

[11] *Twitter.com Site Analytics.* Compete, March 2009.

[12] United states cancer statistics: 1999-2005. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, 2009.

[13] A. Cavoukian. Data mining: Staking a claim on your privacy. Technical report, Information and Privacy Commissioner / Ontario, January 1998.

[14] P. Chesley, B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 27–29, 2006.

[15] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Sixth Symposium on Operating System Design and Implementation.* OSDI, December 2004.

[16] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, February 2009.

[17] C. Hagen. *American Idol Winner: Sorry, Mashable... You Predicted the Loser.* OneRiot, May 2009.

[18] L. Hejtmane. *American Idol Winner: Can Google Predict the Results?* Mashable, May 2009.

[19] A. Lenhart. Adults and social network websites. Technical report, Pew Internet, January 2009.

[20] G. Mishne. Experiments with mood classification in blog posts. In *Style2005 - 1st Workshop on Stylistic Analysis of Text for Information Access*, 2005.

[21] P. Polgreen, Y. Chen, D. Pennock, and F. Nelson. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, December 2008.

[22] A. M. Segre, A. Wildenberg, V. J. Vieland, and Y. Zhang. Privacy-preserving data set union. In *Privacy in Statistical Databases*, pages 266–276, 2006.

[23] A. Singer. 49 amazing social media, web 2.0 and internet stats. Technical report, TheFutureBuzz, January 2009.

[24] R. S. Worldwide. Americans take healthcare into their own hands. Technical report, Consumer Healthcare Products Association, January 2001.