Automated Social Network Epidemic Data Collector

Luis F Lopes, João M Zamite, Bruno C Tavares, Francisco M Couto, Fabrício Silva, and Mário J Silva

LaSIGE, Universidade de Lisboa
epiwork@di.fc.ul.pt

Abstract. Recent epidemiological surveillance projects began collecting data from the Internet to identify inflectional diseases propagation. These systems collect data from pre-selected data sources somehow related to the subject. However, other sources, like web social networks, may present early evidences of an infection event. With the increasing popularity of social networks, where people post personal data interactively, we may hint the outbreak of an epidemic event from these data. This paper presents the architecture of a system capable of collecting epidemiological data from social network services. A preliminary prototype collects data from Twitter into a local database, where it can be analyzed and made available to other applications. Initial evaluation results show that relevant epidemiological data can be found in this source.

Key words: Epidemiological surveillance, Data Retrieval, Social networks, Twitter

1 Introduction

Epidemiological surveillance systems are essential for the study and control of diseases. Data recovered by these systems should be as extensive as possible in order to obtain enough information to understand the propagation of diseases, assessing their impact in public health through epidemiological prediction tools.

In recent years there has been an increase of quantitative social, demographic and behavioral data available, which can be used by statistical and mathematical models to improve the traditional disease surveillance systems, providing faster and better geo-referenced outbreak detection capacities.

The epidemiological surveillance is traditionally done by governmental and international health organizations, such as the WHO (World Health Organization), and is mostly based on cases reported by health services.

Although official statistics should be accurate and based on reliable data, new technologies can be implemented to collect epidemiological data that could be regarded in the future as a valuable complement to national reporting systems.

The use of the web to collect information about epidemics has been a target of investigation in the last years. State-of-the-art systems can collect data from

different kinds of sources, such as search-engine log data, news sites and user-provided input. However, social networking sites are potential targets for the retrieval of epidemic information, given their role as forums where people meet and post information about themselves. Text messages exchanged on the web can be used to identify cases of diseases or at least to provide an idea of the spread of a disease in a community.

Twitter[1] functions as a microblog, where people post small messages with a limit of 140 characters that can either be visualized by anyone or by specific friends, according to the account settings. The large and still increasing number of users of Twitter makes it the perfect case case study for our approach.

Data obtained from social network services such as Twitter, can provide real time information from a large base of regular users. Furthermore, since this information is available instantaneously, it can be used for early detection comparatively to information collected from official sources as suggested by Google Trends observations [3].

This paper presents the Data Collector that is being developed as a module of the Epiwork's Epidemic Marketplace[2], an epidemiological data management platform. The objective of this module is to explore the use of web based social network services for the collection of epidemiological related data. This module collects data from Twitter, searching for disease keywords in the published messages. Besides the text of the message, the module retrieves information about the author, location and posting date. In this work we present some statistics related to the data collector and analyze the data obtained for H1N1 in some European countries and discuss how to improve the system.

2 Related Work

Some systems have been developed in the last years to collect disease information from Internet users. Internet monitoring systems (IMS) use data obtained from user voluntary reports, such as Gripenet [6] or collected automatically, like Google Flu Trends [3] and HealthMap [1].

An IMS is an information collection system that depends on the active participation of registered users, which receive weekly newsletters about the flu and are invited to fill a questionnaire about the flu symptoms (or their absence in the previous week). Gripenet was developed after the model of Holland's Influenzanet [5]. It has been implemented until now in 6 countries, besides Portugal and Holland: Belgium, Italy[3], Brasil[4], Mexico[5], United Kingdom[6] and Australia[7].

[1] http://www.twitter.com/
[2] http://epiwork.di.fc.ul.pt/
[3] http://www.influweb.it/
[4] http://www.gripenet.com.br/
[5] http://reporta.c3.org.mx/
[6] http://www.flusurvey.org.uk/
[7] http://www.flutracking.net/

More should be expected to appear in the future with the objective of having real time flu monitorization.

Healthmap[1] is a website that displays in the world map information about new cases of diseases, especially infectious diseases, collected from several sources. These include news sources, such as Internet news sites, the ProMED-mail newsletter[4], Eurosurveillance and WHO. The degree of reliability of these sources is variable, ranging from media news to validated official alerts. The data displayed is organized by disease using an automated text processing system and then displayed on the Earth map. The original information is available by a hyperlink to the original source.

Google Flu Trends has been recently announced as a system capable of predicting influenza epidemics based on search engine query log data [3]. It was initially deployed for the United States of America, where about 90 million users are believed to search online for information about specific diseases. It has now been applied in Australia New Zealand and experimentally to Mexico. The authors of this work state that certain queries can be used as a surveillance tool to predict the number of influenza cases and that these predictions can be produced one to two weeks before the official data.

In recent years Web based social network services, where people share their interests and activities, have become a popular trend. Moreover the types of information shared in these Social Networks are varied, from link sharing to short text messages. This variety allowed for News Services to gather and publish information in these networks.

3 Data Collector

The Data Collector is able to actively collect information about putative infections by automatically retrieving infection alerts from the web. The data retrieved is stored in a local database and made available through web services and a web user interface.

3.1 Architecture

The data collector (fig. 1) is composed by a backend and a frontend. The backend contains the web crawler which retrieves messages from Twitter, a relational database used to store the collected data and web services which can be used to access this data. The frontend consists of a web interface that will provide a dynamic graphic environment for the user to explore the data.

3.2 Data Storage

All the information retrieved from the web will be stored in a local database. Data is stored according to the UML class diagram shown in fig. 2. The two main classes are Disease and Location, which contain the diseases and the geographical locations to monitor. Each disease will have an official name and a type;

alternative names are instances of the Disease Alias class. In the same way, each location will have an official name and a type; alternative names are instances of the Location Alias class. The type of a location can be a city, country, etc. Each location will be geo-referenced by the latitude and longitude of its centre and a radius limiting the area to monitor. For example, cities will tend to have smaller radius than countries.

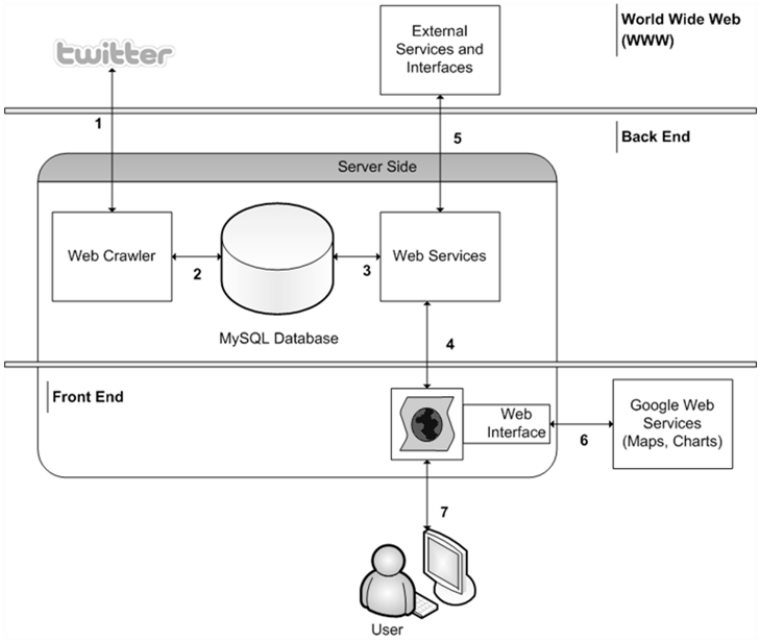

Fig. 1. Architecture of the Data Collector. 1 - The Web Crawler requests XML messages from twitter containing a disease and a location; 2 - The Web Crawler stores detected occurrences in a relational database; 3 - Web Services query the database for information; 4 - The Interface uses the Web Services to present data using 6 - other available Web Services to 7 - display the information to user in an interactive manner. 5 - Other external parties can use the web services for their own purposes.

A relationship between a location and a disease is stored when an alert is detected for the given disease at the given location. The Occurrence will store the date, author, source, evidence, type and score. The author, the source and evidence are information that may be retrieved or received attached to the occurrence alert. The attribute score will store a confidence score for each detected occurrence allowing the weighting according to specific attributes, once mathematical models are developed.

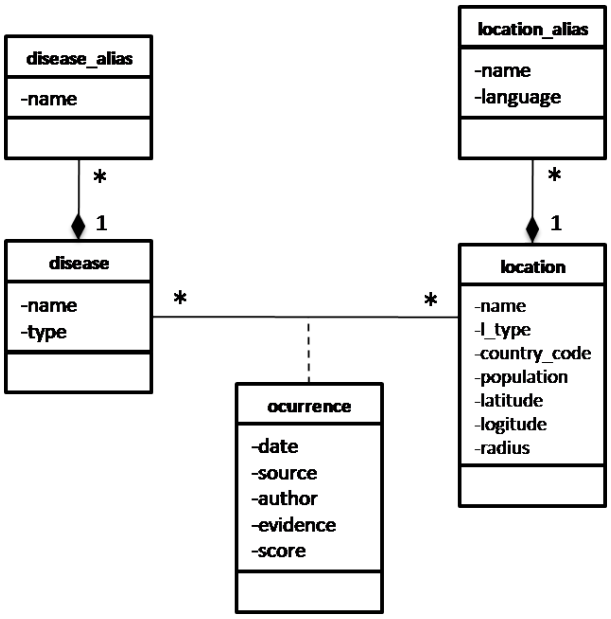

Fig. 2. UML class diagram of the relational database.

3.3 Web Crawler

This module actively collects information about putative infections by automatically crawling the social web. The current crawler accesses the Twitter using the web services documented in the Twitter Search API. To detect disease referencing tweets it queries for the name of a disease and of a location. For example, using: http://search.twitter.com/search.atom?lang=enq=h1n1+italy (see fig. 3).

3.4 Application Programming Interface (API)

The Data Collector provides a RESTful API which specifies the methods and parameters through which other users and applications can access the data contained in the database.

There are three main divisions in the API which specify methods to retrieve locations, diseases and occurrences in the database, displayed in HTML tables.

Retrieving information on diseases is done through the "diseases" method[8] which returns the names and all the aliases of each disease in our database. Requesting information on locations is done through the "locations" method[9] which returns all the information on the locations in our database.

[8] http://epiwork.di.fc.ul.pt/collector/diseases
[9] http://epiwork.di.fc.ul.pt/collector/locations

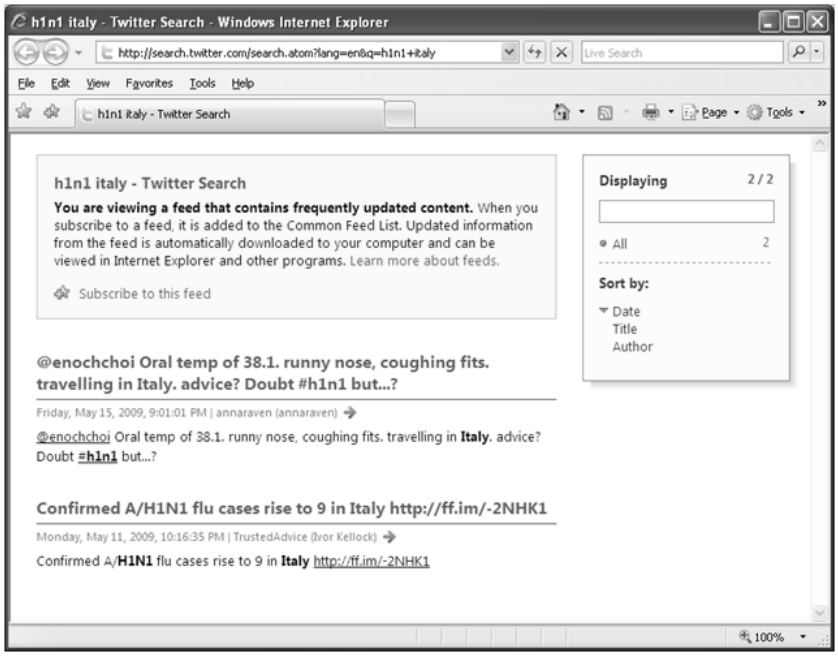

Fig. 3. Output of Twitter API when searching for h1n1 and Italy.

The "occurrence" method[10] returns the main name of the detected diseases, the main name of the location where they were detected as well as the source, author, evidence, score and date. This method can be filtered by location and by disease names using the "location=" and the "disease=" parameters. It can also be further filtered by date, specifying the "before=" and "after=" parameters in the format: "yyyyMMDDhhmmss". The "occurrence" method can also be used with the parameter 'format=tsv' which results in the method returning a dataset in tab-separated values (TSV) format. An example of the usage of this web service would be:

http://epiwork.di.fc.ul.pt/collector/occurrences?disease=h1n1&location= portugal&before=20090518000000&after=20090515000000&format=tsv

This request would produce a TSV containing the attributes of occurrences of H1N1 in Portugal, in a specific time interval.

3.5 Demo Mashup Client

The web interface (see fig. 4) will use AJAX technology to produce dynamic trends graphs and geographical maps that change in real time. It will permit the user to select and visualize data within a specific timeframe (i.e. within a week,

[10] http://epiwork.di.fc.ul.pt/collector/occurrences

month, year, etc), allowing for easier visualization of trends that occur in that window of time. The user interface is currently under development.

For this purpose, we intend to implement dynamical plots, using Google Chart[11], which will allow the user to visualize occurrences over time. Another means of visualizing data would be an earth map, using Google Maps[12]. The map will display markers color-coded according to the number of occurrences which will give the user immediate insight into the disease distribution. The displayed information could be interactively filtered by disease, location and source.

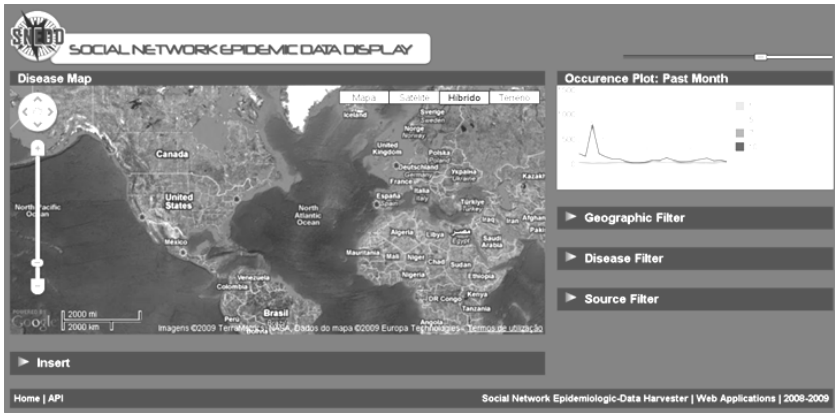

Fig. 4. Snapshot of the mashup webpage.

4 Implementation

The Data Collector is being implemented with free open source software. The collected data is currently being stored in a relational database implemented with MySQL Community Server 5.0. The crawler was written in PHP. It queries Twitter using its Search API for messages containing location names and disease names. It does so for all the possible combinations of locations and diseases. Location information about the countries and their capitals was retrieved from the geonames[13] webservice.Using the searchJSON method we retrieved the countryCodes and geonameId of all the countries of the world and their capitals[14][15]. The country area was retrieved using the CountryInfoCSV method[16]

[11] http://code.google.com/intl/pt-PT/apis/chart/
[12] http://code.google.com/intl/pt-PT/apis/maps/
[13] http://www.geonames.org
[14] http://ws.geonames.org/searchJSON?featureCode=PCLI
[15] http://ws.geonames.org/searchJSON?featureCode=PPLC&country=<countryCode>
[16] http://ws.geonames.org/countryInfoCSV?country=<countryCode>

which was used for calculating the radius of interest for each country. Using the geonameId the remaining information, including all the aliases was retrieved using the getJSON method[17]. Diseases were retrieved manually from the United States CDC website[18] and Health Protection Agency website[19] as well as other sources, including Healthmap. We are currently tracking 89 diseases as well as all the countries and their capitals (17165 names being used). The process of querying Twitter for all this information currently takes from 2 to 3 days, which we consider adequate to the system's needs considering the large number of location names being used. A new querying process is started each week, and since twitter stores messages for over 20 days there are no gaps or discrepancies on the messages obtainable each day.

5 Results

The crawler has been collecting information from tweets that contain a disease and a location in the message body. This type of search is still very simple and is probably missing lots of tweets. However, even so the system is collecting on average 3200 messages every day, from which 700 pertain to H1N1. Using the Data Collector web services, we produced datasets containing the data collected for the disease term H1N1 for several countries: Portugal, Spain, France, United Kingdom (UK), Holland and Italy. From the datasets the number of daily messages was analyzed (see fig. 5).

We observed that France and the UK are, among the countries analyzed, the ones where there were more references to H1N1 (or the alias swine flu). Even using a naïve search method it is possible to obtain a large number of messages. It is also possible to observe an increase in the number of messages being collected, which is correlated to the increase of the number of cases.

6 Conclusions

This work shows that it is possible to collect large amounts of epidemiological data from twitter, a system with a large base of users and daily messages.

Despite the large number of messages already collected by this system, it can be improved to collect even more information in a more accurate way. Since the first implementation the number of diseases and locations has already been largely increased. The system is now looking for a large number of locations and its aliases, based on data retrieved from Geonames. Although the number of diseases being currently tracked is already very high (89 diseases), the number of aliases used is still low. In the future this will be improved by implementing the UMLS ontology [2], which will increase disease coverage.

[17] http://ws.geonames.org/getJSON?geonameId=<geonameId>
[18] http://www.cdc.gov/ncphi/disss/nndss/phs/infdis2007.htm
[19] http://www.hpa.org.uk/

These messages that have been collected had the country in the text body while in the future it will recover messages that were posted in those countries but do not necessarily have the country name in the message.

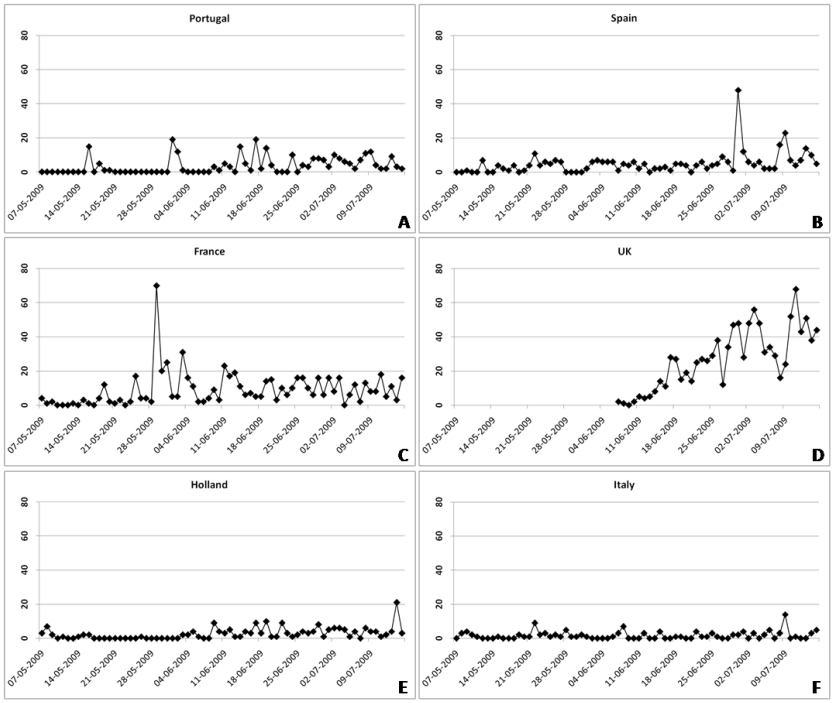

Fig. 5. Number of daily messages containing the word H1N1, from the 7th of May until the 15th of July of 2009, in six countries: A- Portugal, B- Spain, C- France, D- United Kingdom (only data from the 7th of June to 15th of July), E- Holland and F- Italy.

The hypothesis we intend to test is that the number of messages posted on Twitter related to a specific disease should be related to the number of disease cases in the population. Although there are several detected cases where collected messages do not reflect specific cases, we suppose that the amount of data is so large that those cases will not influence the results significantly,as observed by Google Trends [3].

The presented system will surely be more efficient in some countries, where the number of people using Twitter or other web based social network services is comparatively high. However this effect may be corrected through normalization or other bias elimination methods.

In the future the use of better text mining techniques will enable filtering the messages containing non relevant messages, improving the data accuracy.

Further statistical analysis and data collection is also required to validate the usage of data collected by this system in epidemiological propagation predictions.

7 Acknowledgements

The authors want to thank the European Commission for the financial support of the Epiwork project under the Seventh Framework Programme (Grant # 231807), the Epiwork project partners and FCT (Portuguese research funding agency) for its LASIGE Multi-annual support.

References

1. Brownstein JS, Freifeld CC. (2007). HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill, 12(11): E071129.5.
2. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(Database issue):D267-270.
3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. (2008). Detecting influenza epidemics using search engine query data. Nature, 457, 1012-1014.
4. Madoff, L. C. (2004). ProMED - mail: An Early Warning System for Emerging Diseases. Clinical Infectious Diseases, 39:227-232.
5. Marquet RL, Bartelds AI, van Noort SP, Koppeschaar CE, Paget J, Schellevis FG, van der Zee J. (2006). Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season. BMC Public Health, 6:242.
6. van Noort SP, Muehlen M, Rebelo de Andrade H, Koppeschaar C, Lima Loureno JM, Gomes MG. (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. Euro Surveill, 12(7): E5-6.