**Twitter Based Epidemic Trends**

Himanshu Maurya , Mudit Sharma , Arpan Gupta

**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 use of the web to collect information about epidemics has been a target of investigation in the last years.

Twitter functions as microblog, where people post small messages (limit of 140 characters), that can be visualized anyone according to their account settings. The large and still increasing number of users of Twitter makes it a perfect for data crawling purpose.

Data obtained from social network services such as Twitter, can provide real time information from large base of regular users.

**Table of Content**

1. Current Scenario
2. Description
3. Purpose and Scope

**Current Scenario**

Some systems have been developed in the last years to collect disease information from Internet users. Internet monitoring systems (IMS) use obtained from user voluntary reports, such as Gripnet1 or collected automatically, like Google Flu Trends2 and HealthMap3.

Google Flu Trends has been announced as a system capable of predicting influenza epidemics based on search engine query log data. The author 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.

**Description**

* Data Collector

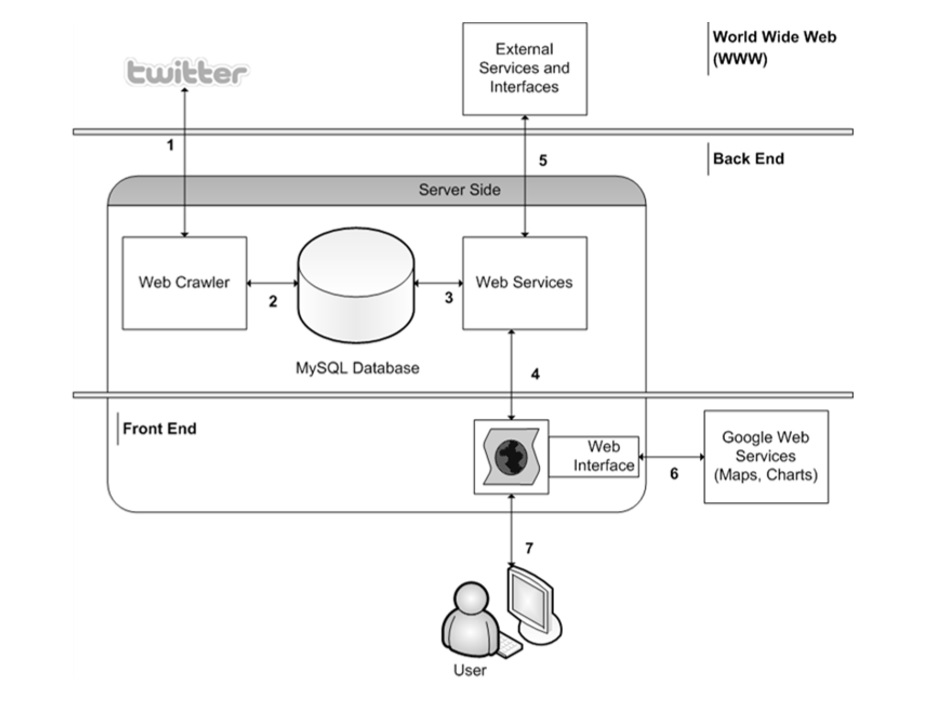
The Data Collector is able to actively collect information about putative infections by automatically retrieving infection alerts from the web. The data retrieved is stored in a local database and made available through web services and a web user interface.

* 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.

* Data Storage

All the information retrieved from the web will be stored in a local database. Data is stored according to the Entity Relationship diagram shown in fig. 2.



**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.

User is requested to Login into the system or is directed to Signup page. Then is directed to a Query page where he can select different queries based on his requirements. The query fired requests the Twitter Crawler API to get data about different disease trends from Twitter. A relationship is generated between Disease – Location, Disease – Symptom, User – Symptom – Disease. Whole data is stored in different attribute of different tables. The frequency of occurrence of disease is stored in an attribute, which is further used for analysis and depiction of data of stored.

* 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>

* 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.

* Demo Mashup Client

We intend to implement dynamical plots, using Google Chart4, which will allow the user to visualize occurrences over time. Another means of visualizing data would be an earth map, using Google Maps5.

**Implementation**

* The Data Collector is being implemented with free open source software.
* The database implemented using MySQL Server 5.0.
* The crawler is implemented using Python.
* Crawler uses Tweepy6 to collect data from Twitter.

**Entity Relationship Model**

