A Web Based Tool For the Detection and Analysis of Avian Influenza Outbreaks From Internet News Sources

Ian Turton and Andrew Murdoch

ABSTRACT: There is an ever increasing amount of publicly available data available relating to disease outbreaks on the Internet. However it is almost impossible for an individual to keep up with all of this data anymore. This paper introduces an automated system that takes in RSS news feeds related to avian influenza and automatically geocodes them. These articles are then stored in a spatially enabled database and served to the Internet using open standard web mapping protocols. A custom client with mapping and time bar facilities is provided to allow non-expert users easy access to the data.

Real-time RSS feeds coupled with a mapping interface allow for up-to-date data that depicts active disease spreading. By utilizing frequently updated official and unofficial sources through the acquisition of RSS feeds, origins and dispersal patterns of diseases such as Avian Influenza can be analyzed and hopefully aid in containing the disease, treating those affected, and improving future prevention methods.

KEYWORDS:web mapping, RSS, news feeds, avian influenza, disease outbreaks, OGC, open source

Introduction

An increasing amount of health related data is now available on the Internet (Woodall, 1997; Heymann and Rodier, 2001; Woodall, 2001; M'ikanatha et al., 2006), while some of this data is unverified (e.g. news reports, blogs) much is produced by medical professionals on the ground (e.g. ProMED, WHO). It is increasingly difficult for interested analysts to keep up with the amount of data available from the many sources but with increased worries about the risk of global pandemics or intentional bio-terrorism incidents it is becoming increasingly important that quick detection and fast response is possible. Using the increasing prevalence of RSS feeds on the web and a simple news aggregator to harvest articles from feeds is an effective way to track diseases such as Avian Influenza. Initially, the experimental system relied heavily on the World Health Organization's Avian Influenza RSS feed. However, no longer is the WHO's feed the only reliable source to report Avian Influenza outbreaks. It is becoming easier to subscribe to RSS web feeds as they are becoming more integrated into previously operating news aggregators such as Google News and Yahoo! and into medical sites, both official and unofficial, that regularly report on actively occurring disease outbreaks

(Woodall, 1997). Many sites either offer several feeds pertaining to specific topics, such as Avian Influenza, or can be queried for feeds pertaining to a given keyword, in essence producing a customized RSS feed tailored to the user's desired topic.

This paper describes an experimental project to provide health analysts with a simple web based tool to allow them to quickly and easily view of events in the current outbreak of Avian Influenza (`http://www.experimental.geovista.psu.edu/andrew/html/avian_influenza_map.html`). The system is constructed using a collection of open source tools that with minor customization could be used to visualize any phenomena that can be accessed using an RSS feed. The system is best characterized as a client server system where the user's web browser provides the execution environment for the client and the server is implemented as a Java middle ware component (GeoServer) over a spatially enabled SQL database (PostGIS).

Previous Work

Mykhalovskiy and Weir (2006) relate the development of the Global Public Health Intelligence Network (GPHIN) and it's role in the detection of the 2002-3 outbreak of SARs. (Freifeld et al., 2007; Brownstein et al., 2008) discuss one of the most well known automated disease mapping systems HealthMap (`http://healthmap.org`). In both cases the system makes use of the growing power of the Internet to collect information on outbreaks of an illness or disease worldwide and provide a summary for experts to access through their desktop computer. In the case of GPHIN the system is restricted to subscribed users and simply scans incoming news items and either posts them to a web site or drops them if they fall below a set threshold. HealthMap makes use of a variety of data sources but applies text processing techniques to them extract information about diseases and locations.

Both of these systems are designed to collect information about all infectious diseases of interest. HealthMap also attempts to extract locational information using a simple dictionary look up technique to assign a text string to one of 2,300 locations known to the system. HealthMap then stores the text, location and disease data to a database for display on a web page. The system is based on the Google Maps API (see figure 1) which is familiar to users but lacks custom-ability.

System Design

The system described in this paper consists of a client server combination. The server is responsible for collecting the data from the news sources on the Internet, extracting the geographic locations mentioned on the text, geocoding these locations and storing the results in a database. The server is also responsible for serving the data to the client for display by the user. The client provides the map interface and a time bar to allow the user to select periods of interest. As the system is constructed using open international standards as defined by the Open Geospatial Consortium (OGC) the data stored on the server is also available to expert users using fully featured desktop GIS tools.

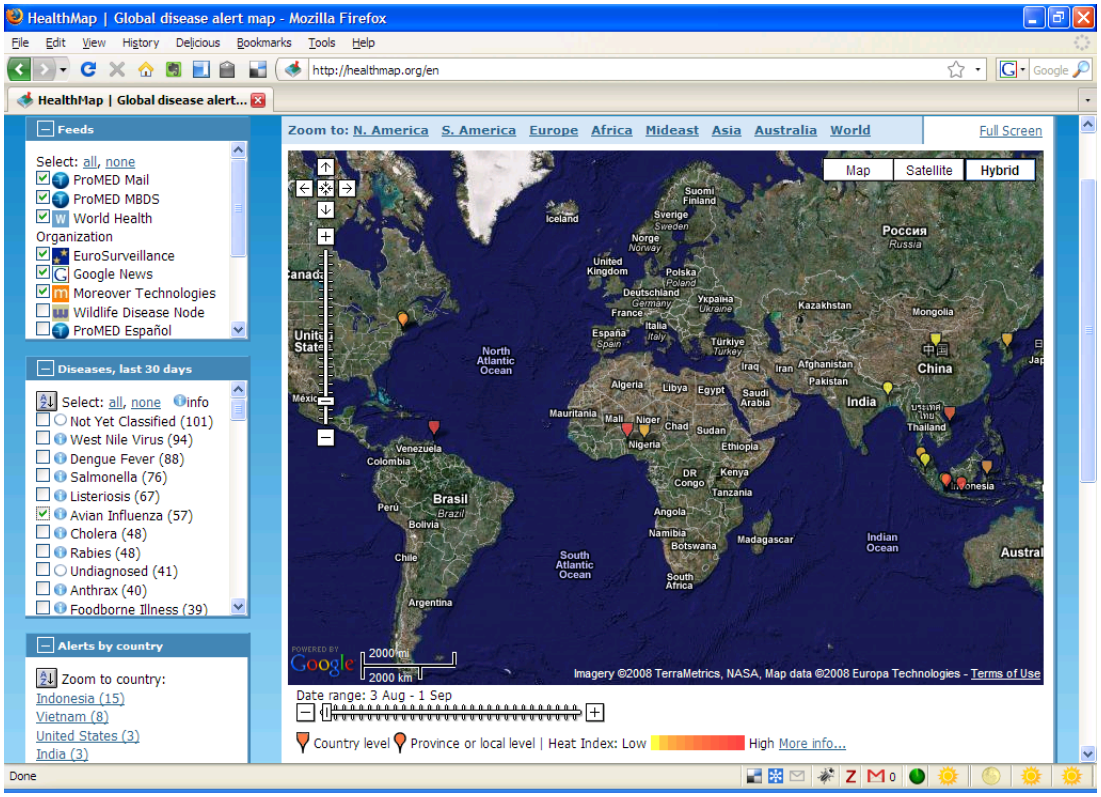

Figure 1: HealthMap showing recent Avian Influenza reports

The server makes use of the open source web mapping server GeoServer serving data from a PostGIS database that stores the georeferenced news items and other background maps. At a set interval each day a java process is spawned by the cron system on the server that collects the latest items from the RSS news feeds and passes them through a custom geocoder based on FactXractor. FactXtractor is an information extraction web service for Named Entity Recognition (NER) and Entity Relation Extraction (RE) developed at PSU and available at `http://julian.mine.nu/snedemo.html`. FactXtractor processes a text document using GATE (Cunningham et al., 2002) and identifies entity relations using Stripped Dependency Tree kernels. The named entities are then processed using a custom geocoding system that makes use of the `www.geonames.org` database to convert raw place names into geographic locations. As each place name is extracted from the document it is returned to the reader which then attempts to disambiguate it and geolocate the resulting place. Disambiguation involves taking a bare place name like 'London' and determining which of the 2683 places called London (`http://ws.geonames.org/search?q=london`) in the GeoNames gazetteer is the correct one. While there is much literature discussing methods and algorithms to carry out this process most assume either that the user has a pretagged corpus to train a machine learning method on (Rauch et al. 2003), or that more information than simply London is available to help disambiguate the place name (Amitay et al. 2004).

Amitay et al. (2004)discuss how place names extracted from text documents can be disambiguated by applying a series of heuristic rules which they use to determine

Given two locations A and B:
Choose A if A is a Political Entity and B is not,
Choose B if B is a Political Entity and A is not,
Choose A if A is a Region and B is not,
Choose B if B is a Region and A is not,
Choose A if A is an Ocean and B is not,
Choose B if B is an Ocean and A is not,
Choose A if A is a Populated Place and B is not,
Choose B if B is a Populated Place and A is not,
Choose A if A's population is greater than B's,
Choose B if B's population is greater than A's,
Choose A if A is an Administrative Area and B is not,
Choose B if B is an Administrative Area and A is not,
Choose A if A is a Water Feature and B is not,
Choose B if B is a Water Feature and A is not,
Choose A.

Figure 2: Disambiguation Algorithm

the geographic focus of a text document (in their case a web page). They define the two types of ambiguity that can occur in this sort of process as geo/non-geo and geo/geo. A geo/non-geo ambiguity is one where a place name is also a person (Washington) or is a common word (turkey), while a geo/geo ambiguity is where a place name occurs for many distinct places in the world, e.g. London, UK and London, Ontario; Springfield – as in "The Simpsons"' – is also a very popular choice, having been used as a place name in at least 25 US states. In their Web-a-where system Amitay et al. (2004) initially built a dictionary of likely geographic and non-geographic words from the geo/non-geo group of words from a corpus of documents and the number times a word was capitalized, implying it was a proper noun. Then they make use of the fact that when several places are mentioned in a document they are most likely to be near each other to resolve geo/geo ambiguities. In the system described in this paper we found that the news items were too short and often about too many places to be able to apply these and other more complex disambiguation algorithms. Therefore a simpler heuristic was applied which was found to work well in most cases (see figure 2). It works with the GeoNames feature codes (`http://www.geonames.org/export/codes.html`) which are unique to each location.

These news items are then stored in the spatially enabled database from which they can be served to the map client. By making use of GeoServer's ability to produce output in multiple formats it is possible to send the data to the client encoded as GeoJASON which can be read by both OpenLayers and the SIMILE Timeline.

The client consists of a time line and a map that the user can use to zoom into any region of the Earth that is of interest to them and limit the markers displayed by selecting a particular time period using the time bar (see figure 3). This functionality is provided using the OpenLayers mapping client (`http://www.openlayers.org`) which provides Open Geospatial Consortium (OGC) standard compatible web map server display as well as the ability to display GeoRSS feeds on the map. The timeline is based

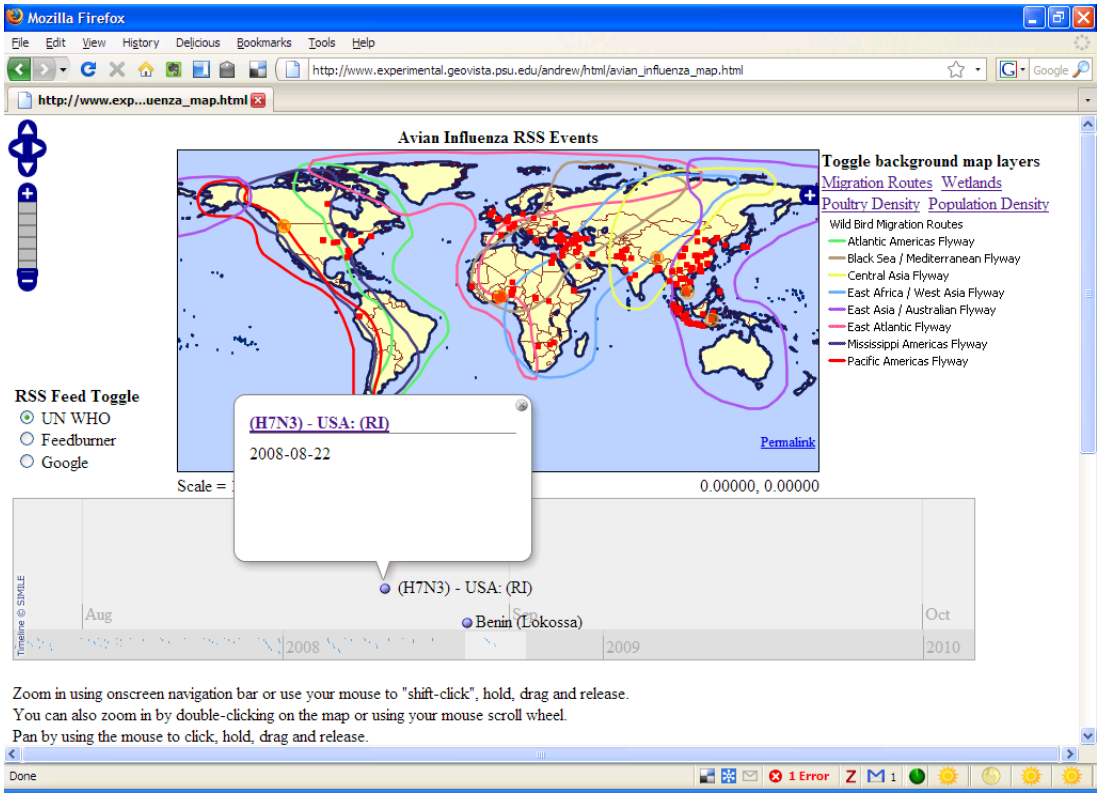

Figure 3: The AvianFlu Viewer

on the MIT SIMILE project's time line tool (`http://simile.mit.edu/timeline/`). A GeoRSS feed is an RSS feed, such as those used for distributing updates to web sites and news, that has been enhanced by the addition of geospatial coordinates (`http://www.georss.org`).

Figure 3 shows the viewer in action, here the user has selected a point in the time bar, which has popped up a "bubble" view containing a link to the original news item and the date that the news item was published. If a summary of the item had been available it would have been shown here too. On the map the red dots show the locations of all places mentioned in the news items since the system was launched in 2007, the points which are related to the selected time period (the lighter area at the bottom of the time bar) are highlighted in orange on the map. A user can also click on a point on the map to see all news items related to that location sorted by publication date.

Data Sources

Currently the system is based on three feeds, one a combination of the WHO avian influenza feed from the Epidemic and Pandemic Alert and Response (EPR) section which contains information about outbreaks of avian influenza from around the world as they are reported to the WHO and an RSS representation of the ProMED email list relating to Avian Influenza. The second feed is a collection of news feeds from news agencies and aggregation services such as Google News, this is more experimental. It

suffers from some problems related to being collected from a wider range of sources such as a tendency to collect the same news item from several sources (usually a news agency story that has been widely published by local newspapers) as well as stories that are about avian influenza but not about actual outbreaks. The third feed is a second experiment which takes results of a search for "avian influenza H5N1" run on the Google News site and returns them as an RSS feed. This shows some promise as it combines the advantages of the feed burner collection but Google attempt to remove the duplicate news items to a certain extent using a pre–processing algorithm.

Future work is planned to make use of classification techniques (Konchady, 2006; Segaran, 2007) to reduce the duplication of articles and to classify the news items in to groups that are of more interest to an analyst.

ProMed is a widely known disease reporting system with an overall error rate of only 2.6% (Woodall, 2001). It has a specific Avian Influenza RSS feed, which makes it ideal for the task as it will not yield reports about other diseases. EuroSurveillance is a scientific journal that reports on medical research and disease outbreaks, but its Avian Influenza feed is fairly inactive and for that reason this feed is not as pertinent as others. The Moreover Technologies news aggregator Avian Flu feed suffers from impurity because many of its articles have to do simply with health-related topics. Duplicate articles are also a problem.

Aside from the feeds used by HealthMap, additional feeds are being experimented with, some of which are more successful than others in generating desired articles. The European based real-time news alert system MedISys (`http://www.medusa.jrc.it`) provides an Avian Flu feed that reports on outbreaks but suffers from advertisements that infiltrate the feed and decrease the feed's precision and therefore desirability. The CDC Flu Updates feed provides articles and podcasts focusing on flu prevention rather than outbreak information and so is of no interest to this project. EpiSPIDER is a surveillance system that supplies many feeds, some of which are Avian Influenza related feeds from askMEDLINE (`http://www.askmedline.nlm.nig.gov`), which compiles sources from medical journals and abstracts. While the Medline feeds provide many articles, they have to do almost exclusively with academic research rather than outbreak information. Other potentially useful sites about Avian Influenza outbreaks include the European Influenza Surveillance Scheme (`http://eiss.org`) and the Global Public Health Intelligence Network (`http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk-eng.php`), but these sites do not produce RSS feeds and therefore are not helpful for the system. In the longer run the authors may investigate building screen scrapers to generate RSS feeds from these sites but they are not included in the present system.

Results

The current system is based on hand constructed scripts that collect news items from carefully selected RSS feeds. There is no attempt to allow the system to detect general disease outbreaks in the way that HealthMap does (though being based on a dictionary makes that claim suspect). But taking this limitation into account the system preforms very well at its aims. The place name recognition and geocoding system (while not yet

formally tested) seems to work very well. This allows the automated collection of a large database of geocoded disease information which is of use to both professional health workers and the general public.

Early user testing has shown that by making use of a "slippy map" interface the system is relatively easy for inexperienced users to navigate around the map, the time bar also is intuitive to users with a simple drag interface.

Future Work

In future work we hope to develop the system to work with other diseases and possibly allow users to construct their own personalized portals by providing their own RSS feeds of interest (or selecting from a menu of preselected feeds). This will allow users to visit the application on a regular basis (or possibly make it a widget on their homepage) so that they can get a regular update on diseases of particular interest to them.

Another interesting and potentially useful feature we plan to add in the near future is to allow users to click on a link to give feedback on the quality of the geocoder, so that if a place name is being miscoded the user can give feed back to the system either directly affecting the weightings used in the algorithm or alerting the programmers that there is a problem that needs attention.

As was discussed in the data sources section some feeds had to be dropped as they contained too many adverts or other irrelevant posts. To deal with this we plan to experiment with a Bayes-enabled feed aggregator in which each article in an initial set of articles is manually given a positive or negative rating based on relevance (Orchard, 2005). For instance, an article reporting on an Avian Influenza outbreak receives a positive rating while an article having to do with the history of Avian Influenza receives a negative rating. Once this initial training is complete, the system becomes automated so that every incoming article gets automatically scored, which aids in determining what articles are relevant. However, due to the desire for additional refinement, Bayesian classification alone is not enough to categorize the articles and analyze them effectively. Again end users could vote on the relevance of the news items on the site to help train the Bayesian classifier in the same way that modern email spam systems learn from their users.

Clustering methods could be further used to categorize the articles and allow for even easier analysis of groups within the data. The non-negative matrix factorization clustering method groups the articles based on common keywords/features, which allows for an easy way to identify groups, the number of which can be adjusted (Segaran, 2007). Once the articles are clustered, it is simple to see which articles are relevant and to study patterns in the data to aid in refining a search or eliminating feeds that produce more irrelevant than relevant articles.

Conclusions

This paper has presented initial results from an experimental system that automatically takes news items from Internet RSS news feeds and geocodes them. The news

items are then presented to the user in a client using a combined map and timeline that allow users to zoom into areas and time periods of interest. The system is constructed using open source components that allow for easy customization, and by utilizing open standards the server can be accessed by other clients with very little effort on the part of expert users who require the ability to carry out more advanced analysis than is possible using the web based client.

References

Amitay, E., Har'el, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA. ACM Press.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Medicine*, 5(7):e151+.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the ACL*.

Freifeld, C. C. C., Mandl, K. D. D., Reis, B. Y. Y., and Brownstein, J. S. S. (2007). Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc*.

Heymann, D. L. and Rodier, G. R. a. (2001). Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. *The Lancet infectious diseases*, 1(5):345–353.

Konchady, M. (2006). *Text Mining Application Programming (Programming Series)*. Charles River Media.

M'ikanatha, N. M., Rohn, D. D., Robertson, C., Tan, C. G., Holmes, J. H., Kunselman, A. R., Polachek, C., and Lautenbach, E. (2006). Use of the internet to enhance infectious disease surveillance and outbreak investigation. *Biosecurity and Bioterrorism: biodefense strategy, practice, and science*, 4(3):293–300.

Mykhalovskiy, E. and Weir, L. (2006). The global public health intelligence network and early warning outbreak detection: a canadian contribution to global public health. *Canadian journal of public health. Revue canadienne de santé publique*, 97(1):42–44.

Orchard, L. M. (2005). *Hacking RSS and Atom*. Wiley Publishing Inc.

Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL 2003 Workshop on Analysis of Geographic References*.

Segaran, T. (2007). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly Media, Inc.

Woodall, J. (1997). Official versus unofficial outbreak reporting through the internet. *International Journal of Medical Informatics*, 47(1-2):31–34.

Woodall, J. P. (2001). Global surveillance of emerging diseases: the promed-mail perspective. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, 17 Suppl:147–154.

Ian Turton, GeoVISTA Center, Pennsylvania State University, University Park, PA 16802, ijt1@psu.edu
Andrew Murdoch, MGIS Program,World Campus, Pennsylvania State University, a_murdoch@hotmail.com