Focus on Media-based Biosurveillance JAMIA

Model Formulation ■

HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports

Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, John S. Brownstein

Abstract **Objective:** Unstructured electronic information sources, such as news reports, are proving to be valuable inputs for public health surveillance. However, staying abreast of current disease outbreaks requires scouring a continually growing number of disparate news sources and alert services, resulting in information overload. Our objective is to address this challenge through the HealthMap.org Web application, an automated system for querying, filtering, integrating and visualizing unstructured reports on disease outbreaks.

Design: This report describes the design principles, software architecture and implementation of HealthMap and discusses key challenges and future plans.

Measurements: We describe the process by which HealthMap collects and integrates outbreak data from a variety of sources, including news media (e.g., Google News), expert-curated accounts (e.g., ProMED Mail), and validated official alerts. Through the use of text processing algorithms, the system classifies alerts by location and disease and then overlays them on an interactive geographic map. We measure the accuracy of the classification algorithms based on the level of human curation necessary to correct misclassifications, and examine geographic coverage.

Results: As part of the evaluation of the system, we analyzed 778 reports with HealthMap, representing 87 disease categories and 89 countries. The automated classifier performed with 84% accuracy, demonstrating significant usefulness in managing the large volume of information processed by the system. Accuracy for ProMED alerts is 91% compared to Google News reports at 81%, as ProMED messages follow a more regular structure.

Conclusion: HealthMap is a useful free and open resource employing text-processing algorithms to identify important disease outbreak information through a user-friendly interface.

■ **J Am Med Inform Assoc.** 2008;15:150–157. DOI 10.1197/jamia.M2544.

Introduction

Internet-based resources, such as online newspapers, blogs, and discussion forums, have increased in number, volume, and coverage, and show potential as useful data sources for disease surveillance and early outbreak detection—currently, nearly all major outbreaks investigated by the World Health Organization are first identified through these informal online sources.[1,2] However, electronic sources of infectious disease news are not well organized or integrated. Reading and assimilating a broad range and large number of reports as they appear on a daily basis has already become increasingly burdensome.[3,4]

The HealthMap project has begun to address this challenge through automated querying, filtering, integration, and visualization of Web-based reports on infectious disease outbreaks, to facilitate knowledge management and early detection.[5,8] A freely available Web site operating since September 2006, HealthMap.org integrates data from a variety of electronic sources, including news through the Google News aggregator, expert-curated accounts such as ProMED Mail, and validated official alerts such as World Health Organization announcements. Through the use of automated text processing algorithms, the system classifies alerts by location and disease and then overlays them on an interactive geographic map. It currently processes an average of 30 disease alerts per day; with the default 30-day time window, the system typically displays approximately 1,000 alerts at any particular time. The filtering and visualization features of HealthMap thus serve to bring structure to an otherwise overwhelming amount of

Affiliations of the authors: Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology (CCF, KDM, BYR, JSB), Boston, MA; Division of Emergency Medicine, Children's Hospital Boston (KDM, BYR, JSB), Boston, MA; Department of Pediatrics, Harvard Medical School (KDM, BYR, JSB), Boston, MA.

This work was supported by R21LM009263-01, 1 R01 LM007677, and N01-LM-3-3515 from the National Library of Medicine, National Institutes of Health, and the Canadian Institutes of Health Research.

Correspondence: Clark C. Freifeld, Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, 300 Longwood Ave., Boston, MA 02115; e-mail: <clark.freifeld@childrens.harvard.edu>.

Received for review: 06/29/07; accepted for publication: 11/29/07

information, enabling the user to quickly and easily see those elements pertinent to her area of interest.

Background

HealthMap is part of a new generation of health surveillance systems that help supplement existing public health systems by focusing on event-based monitoring of infectious diseases by leveraging Internet news and other electronic media. One of the earliest systems to harness some of these resources is the Global Public Health Intelligence Network (GPHIN).[9,10] GPHIN has shown that extensive monitoring and analysis of news media around the world can effectively aid in early detection of emerging disease threats. Most notably, GPHIN was able to identify the 2002–2003 outbreak of Severe Acute Respiratory Syndrome (SARS) well in advance of official reporting.[10,11] On an ongoing basis, GPHIN also provides a large fraction of initial outbreak reports directly to the WHO for investigation.[1,2] Another successful online disease alerting service is the ProMED Mail email announcement list, with 38,000 subscribers and a panel of expert moderators.[12–14] Other systems include MedISys,[15] Argus,[16] and EpiSPIDER,[17] all of which also leverage informal electronic datasets for disease outbreak information.

While projects such as GPHIN and ProMED serve public health authorities, infectious disease Web sites that serve the general public are also gaining in popularity and helping to increase awareness of public health issues, especially for international travelers. One such site, FluWikie.com, which reports on avian influenza and other topics relating to pandemic influenza, is heavily trafficked and was cited along with similar sites by the CDC as "critical to CDC's ability to prepare for and respond to an influenza pandemic."[18]

In addition to existing online public health resources, recent years have seen the rise of "Web 2.0" technologies[19] including the proliferation of Really Simple Syndication (RSS)[20] and Asynchronous JavaScript and XML (AJAX).[21,22] These tools create new opportunities for interactive software such as HealthMap. On the backend, RSS is a first step towards the goal of a "semantic Web,"[23] allowing for greater possibilities in extracting structure algorithmically from a variety of disparate data sources. On the frontend, the Google Maps public API allows the Web developer to create mapping applications using a powerful and well-known user interface. Finally, rich JavaScript and asynchronous HTTP requests, the AJAX building blocks, enable us to create responsive, highly customizable Web user interfaces that begin to approach the desktop software experience.

The power of HealthMap as a disease surveillance tool lies in its potential to bring together automated processing of a broad range of Internet data sources and rich, accessible visualization tools for lay and public health users alike. In this report, we describe the software architecture and implementation, as well as challenges and future plans.

Formulation Process

The principal objective of HealthMap is to provide access to the greatest amount of potentially useful health information across the widest range of geography and pathogens, without overwhelming the user with excess information or obscuring important and urgent elements. To accomplish this goal, the system must be able to correctly classify reports, provide flexible and useful visualization output, and be responsive under heavy usage load.

Classification

The system is only useful to the extent that it can correctly identify the primary locations, diseases and other outbreak-related factors of a large percentage of alerts, based on words, phrases and other available contextual information for each report.

In addition to the "correctness" of classification, the system must also take end-user objectives into account. For example, if a single alert contains references to fifty different places, the strictly correct classification would generate markers in all fifty locations. However, this alert, likely a summary of known ongoing activity, would then overload the map view with less important information and provide little benefit to the user. Another condition where optimum classification is difficult is in the case of multiple country involvement in a single outbreak. For instance, Switzerland may send disease specialists to help combat a dengue fever outbreak in Paraguay. In this case the primary locations of the alert are Switzerland and Paraguay, but if the system presents alert classifications in such a way as to imply that an outbreak of dengue fever is occurring in Switzerland, the user will be justifiably confused. The classifier must thus be designed to integrate its output with the user display.

Visualization

With respect to visualization, a key objective of the system is to maximize flexibility in two key areas: in the user interface and in the collection of the underlying data. Specifically, HealthMap is designed to organize data across different dimensions (such as date, location and disease) and allow users to customize the view according to the geographic location, disease, and type of outbreak. However, the system must balance flexibility with simplicity; in certain cases, it should impose assumptions in organizing the data, so as not to overwhelm the user with customization controls. In general, the visualization interface should be intuitive and easy to use for the novice user—who may be a novice with respect to both software interfaces and infectious disease epidemiology—as well as allow the advanced user sophisticated and flexible customization of the display.

Behind the user interface, as the system collects reports, the goal is to allow the underlying data to shape the view as much as possible. Avian influenza, for example, is currently a topic of significant public health concern and extensive media coverage. However, the system should not place *a priori* emphasis on any given disease; instead it should adapt its mode of display to infectious disease threats as they emerge. The next global threat may come from an unexpected source, or the focus of public health and media attention may shift.

Accordingly, while HealthMap focuses primarily on human disease surveillance, one of our design objectives is comprehensive coverage of disease activity, encompassing animal and plant diseases, as well as some insect pests and other invasive species. This disease coverage is important as many infectious diseases of public health concern are zoonotic, naturally circulating among wildlife reservoir hosts before emerging in the human population.[24–26]

Along the same lines, the system should, where possible, avoid biases towards specific geographic areas. The next

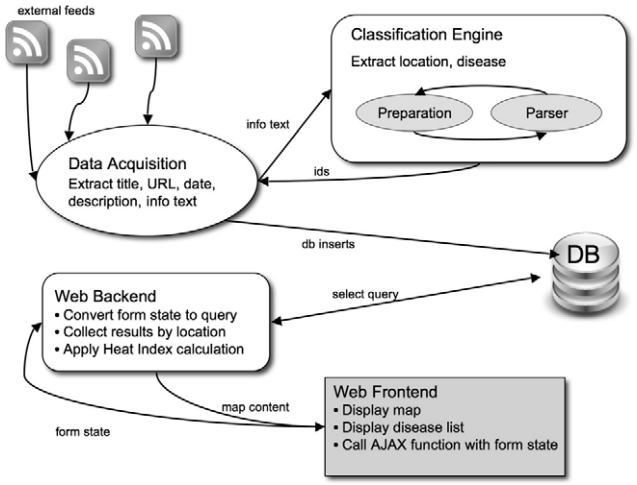

Figure 1. HealthMap System Architecture.

noteworthy outbreak may as easily come from a major urban center in North America as a rural village in Africa.

Performance and Scalability

As the system scales to include more sources and more dimensions of classification, it must be capable of rapidly processing a large number of reports. And as the user interface is enhanced to provide more sophisticated data visualization and customization, it must be able to accommodate a large number of simultaneous users and still be responsive. This scalability is critical, as the Web site could receive a burst of traffic in the event of a broadly publicized disease outbreak.

Model Description

The HealthMap system consists of five modules: the Data Acquisition Engine, Classification Engine, Database, Web Backend, and Web Frontend. As illustrated in Figure 1, the system gathers alerts, classifies them by location and disease, stores them in a database, and then displays them to the user.

Data Acquisition Engine

As the system loads raw data from the Web, it converts each disease outbreak report into a standard "alert" format, containing four fields: *headline, date, description,* and *info text*. The *headline* is the alert headline, *date* is the date of issue of the alert, and the *description* is a brief summary of the alert, generally the first few sentences of the article. The *info text* is the text that will be fed into the parsing engine for the initial classification pass. In general, this initial text consists of the alert headline, stripped of elements that may trigger a false positive. For example, with Google News, the system removes the name of the originating publication from the headline.

The standardization of the alerts, when not already available from the RSS structure, is accomplished through the use of basic assumptions about the HTML and text formatting of the input for each feed. The drawback to making these assumptions is that the data source may change its format without warning, creating unexpected results in the data acquisition and requiring rapid adaptation of the system, though this has not yet proven to be a problem.

Classification Engine

The classification engine determines the primary locations and diseases associated with each alert. It is comprised of two modules: the Preparation Module, which takes the raw input from the source, segments it and prepares it for input to the parser, and the Parser Module, which takes text input and produces disease and location codes as output.

Preparation Module: Tiered Approach

While many alerts contain references to multiple locations or multiple diseases, the aim of the classifier is to identify the *primary* locations and diseases for each alert. To this end, the input is processed in stages: if the classifier is unable to identify location and disease from the initial input provided by the feed, namely the modified headline, it can request additional text from the feed. For example, in the case of the Google News aggregator, the system examines the headline, then the description, which generally consists of the first one or two sentences of the article, followed by the article's body text, and finally, the name of the online news source. Frequently, a publication originating in one area will refer to events occurring in another area, making the publication name and location an unreliable source for the location of the alert. However, articles that don't refer to a well-known location, such as "Suburban school closed after flu outbreak," generally refer to a location near the publication headquarters. By processing the input in stages, the classifier avoids the incorrect classification of the first case while capturing the true location in the second case.

The extraction component of the preparation module processes the full HTML body of the article itself. Clearly, the article text contains the best indicators as to the locations and diseases of the event in question. However, blindly feeding the full article into the parser, while increasing sensitivity, would also significantly increase the false positive rate, especially due to JavaScript code, CSS and hyperlinks mixed with the body text, any of which may contain text elements that would trigger an incorrect match. The extractor must also contend with the wide variety of HTML formats of different news sources, including potentially malformed HTML code. By means of a collection of regular expressions and cautious assumptions about the input, the system confronts some of these challenges.

Parser Module

The Parser Module uses a word-level N-gram approach to match input against a dictionary of known patterns (an N-gram, as applied in the HealthMap software, is an N-word text extract, generally 1 to 10 words in length). After the initial data acquisition, the parser receives the input text, strips it of non-alphanumeric characters and splits it into word tokens. It then converts all capital letters to lowercase, except for those tokens that are two characters or fewer in length. The parser then compares the input to its dictionary of place and disease patterns, mapping text patterns to the database IDs of all locations and diseases known to the system. As part of the ongoing development of HealthMap, the dictionary is updated daily to improve the accuracy of the system; at the time of this writing it consists of over 2,300 location and 1,100 disease patterns.

Because the dictionary patterns are stored in memory as a tree, where each node is a hash table that maps single tokens

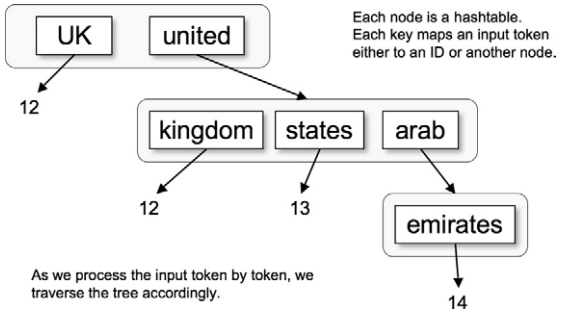

Figure 2. Lookup Tree.

to either subnodes or IDs (leaves), the system can look up each input token in constant time (see Figure 2). Thus, the classification time is linear on the number of input tokens, i.e., the length of the input. In the case where a word may have multiple spellings, for example the American "diarrhea" and the British "diarrhoea," we simply stock the dictionary with multiple patterns. With the addition of patterns to the dictionary, memory consumption increases, but lookup time does not increase substantively.

The disadvantage of this approach is that because the input is hashed, each token must match exactly, making it difficult to accommodate fuzzy matching, wildcards, or regular expression approaches. Further, if we change the input processing, for example, to retain more of the input data, such as capitalization and punctuation, we must update the entire pattern dictionary. A further disadvantage of the dictionary approach is that the system can only identify locations and diseases already known and stocked in the database. Moreover, a key step in enhancing the parser resolution consists of augmenting the database by capturing correct locations and disease names, often involving careful manual data entry. As national borders shift and names of places change, albeit infrequently, the system must be manually updated to reflect new geography. For example, we have already been affected by this issue, as we needed to update the parsing system to reflect the designation of Serbia and Montenegro as separate nations on June 5, 2006.

A key advantage to the pattern dictionary approach is that it is relatively easily translated to other languages: we can simply employ a different dictionary within the existing architecture. A language expert is needed to perform the initial translation, refine the pattern library, help with capitalization and punctuations subtleties, and provide other adaptations, but the basic approach can be re-applied without major changes to the system. Further, the language expert need have only very minimal technical knowledge with respect to natural language syntax or software development to contribute to the dictionary. With the help of collaborators at the Naval Medical Research Center Detachment in Peru, we have already successfully adapted the classification engine to accommodate Spanish-language input, albeit with a smaller pattern dictionary.

Container Relationships

A key component of the location classifier is its use of relationships among geographical entities. Our goal is to identify the most specific primary location or locations for a given alert. In many cases, we are presented with input such as "UK (England)," or "Boston, MA." In these cases, each input contains two distinct patterns that are coded as separate locations in the dictionary. However, Boston is contained by Massachusetts and England is contained by the United Kingdom. In order to correctly process this type of input, after it has identified a list of locations, the classification engine executes a secondary step, eliminating apparently redundant locations based on container relationships. In the given example, the system will initially identify both Boston and Massachusetts as locations for the alert, and then eliminate Massachusetts, as it is considered to be redundant with Boston.

We also apply container relationships to disease matching, as the input can contain analogous cases. For instance, avian influenza is a type of influenza and Norwalk-like viruses cause gastroenteritis—if the system identifies both Norwalk-like virus and gastroenteritis in an alert, it thus eliminates gastroenteritis as a redundant, less specific disease category. One key difference in the case of disease taxonomy is that unlike a location, a disease can "belong" to more than one container disease: *E. coli* is more specific than food poisoning, while norovirus, cholera and *E. coli* can each cause diarrhea (or gastroenteritis). If no disease category can be identified from the text, we designate the alert as Not Yet Classified. Such alerts may be non-disease-related news items that have slipped through the filter, but they may be important if they indicate initial investigation of an unknown disease or a rare condition that is not yet represented in the HealthMap database.

Database

Once the alerts are classified by location and disease, the system stores them in a MySQL database. The database is designed according to standard relational database normalization principles. The primary tables store alerts, diseases and locations, while linking tables map alerts to their respective categories as identified by the classification engine. This standardized data model allows the HealthMap software flexibility to perform a variety of queries and display different views of the data. While the database is designed primarily to support features of the Web application, the data as they are stored are readily accessible for retrospective epidemiological studies, public health risk mapping and other research applications.

Output Renderer

The initial Web page is loaded by the user's browser from a server-side cache which is updated every hour, following the capture and classification of new alert data. If the user adjusts the viewing parameters, he will trigger an AJAX request to the server. The request indicates the current state of the page controls, and from it the server generates a database query. The database then returns the alerts that match these parameters.

From these query results, the system then tallies the number of alerts, diseases, and feeds for each day at each location. To this tally it applies an algorithm, based on an exponentially weighted average, to determine a "heat" rating for each location. In order to give particular emphasis to more recent alerts, through qualitative assessments, we have currently set the decay parameter "alpha" of the exponential weighting to 0.17. (A greater alpha value means the weighting will

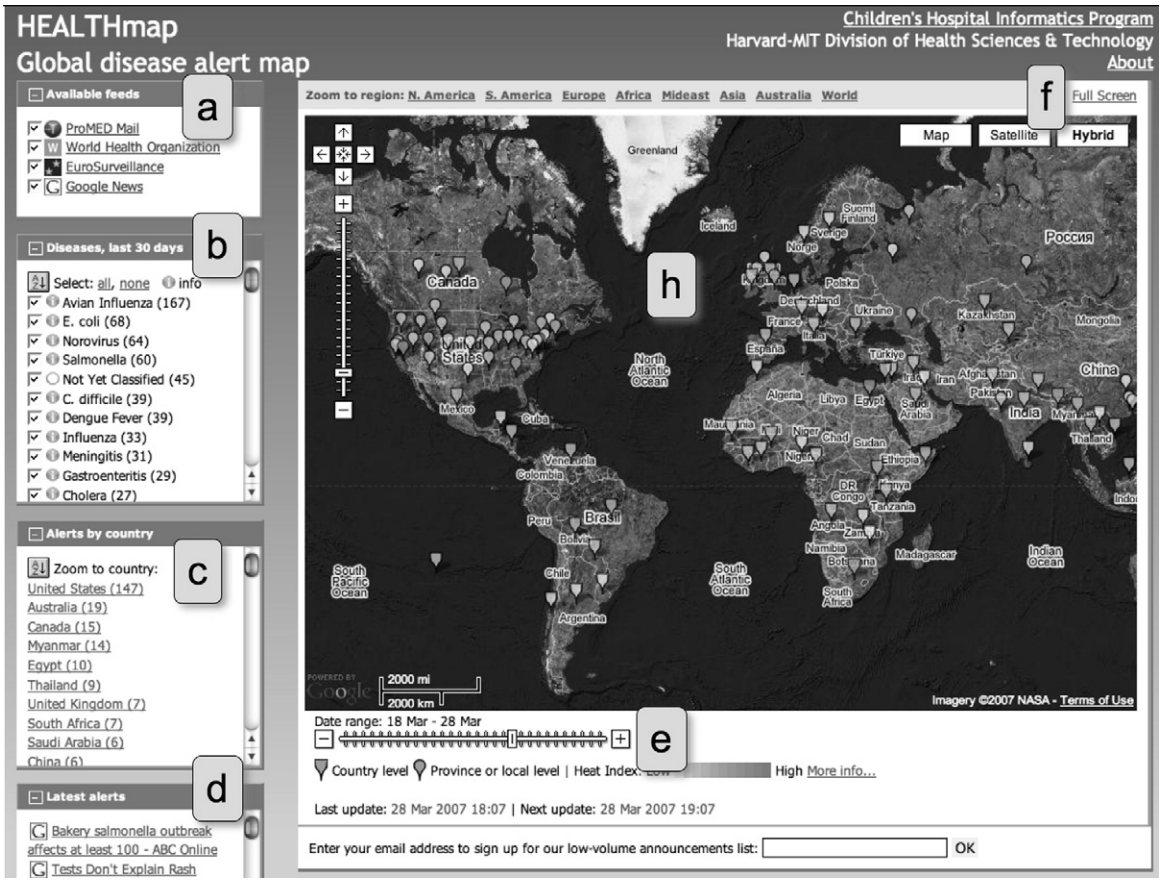

Figure 3. User Interface.

decay more rapidly as we progress into the past.) Locations that have a greater number of feeds and diseases associated with them are also given increased weighting. Our qualitative justification for this boost is that if multiple sources have corroborated an outbreak it deserves more emphasis, and if the same source is reporting the same disease, it deserves less emphasis.

After the computation is complete, the system normalizes the heat scores across the set of markers and assigns each marker an integer value from 0 to 10. Because it computes the Heat scores for the currently requested marker set, the user can, for example, choose a particular disease category and quickly see where the hotspots are for that disease, in addition to the default view indicating general levels of outbreak activity.

User Interface

Figure 3 shows the HealthMap main page, featuring a variety of information boxes and user controls. The "Available feeds" box (Figure 3a) allows the user to select which sources to display on the map by means of the checkboxes along the left-hand side. Below the feeds menu, the "Diseases, last 30 days" box serves both to display the currently active diseases as well as to allow the user to select which diseases to display (Figure 3b). The "i" button brings up a menu with links to further information about the particular disease from the Wikipedia, WHO, CDC, PubMed, and Google Trends Web sites. In the next section, the "Alerts by country" box indicates the number of alerts active in each country for the currently selected parameters (Figure 3c). Clicking on a country name zooms the map view to that country for easy viewing of alerts in that location. The "Latest alerts" box displays the most recent alerts in reverse chronological order (Figure 3d). An icon next to each headline indicates the alert source.

Moving across to the map display window, the date slider at the bottom allows the user to control the date range of displayed alerts (Figure 3e). The end date is fixed as the current date, but the user can set the start date to any point in the previous thirty days. "Full Screen" mode expands the map to cover the full browser window, allowing for richer display and navigation (Figure 3f). It also allows for "situation room" use, allowing the user to display the map on a non-interactive screen and monitor ongoing alert activity. On the map itself, the color of a marker indicates the Heat Index value for the location, with the deeper red color indicating more intense recent activity as contrasted with the paler yellow color.

Validation

Example Report Illustrating Classifier Operation

To illustrate the functioning of the system, we examine a sample report and how it is processed by the HealthMap classification engine. A local newspaper report concerning an outbreak of shigellosis at a school in Wisconsin enters the system via the Google News aggregator. The system begins by examining the article headline:

Elementary School Deals with Outbreak of Bacteria

As there are no known patterns found for either location or disease, the classifier then progresses to the article "description," an extract provided by Google News:

Elementary School Deals with Outbreak of Bacteria 58 minutes Smith A bacterial outbreak at a Fond du Lac school is prompting the district to alert parents and do some extra cleaning in hopes of stopping the . . .

While there is an indication of the location provided in this extract, "Fond du Lac" is currently not included in the dictionary, and therefore not recognized. Still lacking both location and disease information, the classifier examines the article body text, as prepared by the parsing engine from the original HTML:

WEB SEARCH BY A bacterial outbreak at a Fond du Lac school is prompting the district to alert parents and do some extra cleaning in hopes of stopping the bacteria from spreading. State health officials say there were 14 confirmed cases of ***shigellosis****, a bacterial infection, in Fond du Lac County in the past three months. Five confirmed cases prompted Roberts Elementary School in Fond du Lac to notify parents. "We want to get the information out to parents: Here it is and here are steps you can take," Marian Sheridan, the Fond du Lac school health and safety coordinator said. The concern is that this infection is fast-spreading. Although the* ***Wisconsin*** *health department says 300 to 400 cases are reported each year, the uncomfortable abdominal cramps, fever, and* ***diarrhea*** *are symptoms no one wants running rampant through schools. "I think we re getting the message out early enough, and I think that s one of the benefits of working with school districts staff to get the word out so we can contain it before it s widespread," Joyce Mann of the Department of Health and Family Services said. "Parents are used to the school sending them health notices, and it s never to alarm but it s rather to inform," Sheridan said. "Normally what we do is go in with a ten-percent bleach solution and everything gets wiped down—telephones, door knobs, desk chairs, desktops, the bathrooms are thoroughly gone through," building and*

As indicated in bold, the classifier now matches three different patterns in the text. The first identifies the disease category as Shigellosis; the second places the report in Wisconsin. The third match corresponds to the Diarrhea disease category, but based on the container relationships described above, the system correctly identifies Diarrhea as redundant with Shigellosis, and eliminates the former. At this point, the classifier has completed its work, and proceeds to the next report. Had it not identified both disease and location from the body text, it would have further examined the name of the publication as provided by Google News:

WBAY, WI

Upon processing of this text, it would also have identified the location based on the abbreviation WI, which is listed in the dictionary as a synonym of Wisconsin. However, in this particular case, the publication information is ignored as the classifier has already achieved matches using other components of the report.

Classifier Performance

Because the classification engine places alerts into many hundreds of different location and disease categories (currently over 700 total), as well as combinations of multiple categories of each type, it is not possible to apply traditional binary classification metrics such as precision and recall to measure its performance. However, because we curate all reports on a daily basis to correct misclassifications, we can examine various aspects of performance based on the changes performed.

Table 1 ■ Location and Disease Classifier Performance over the One Month Period from 10 October 2007 to 9 November 2007

Source	Total	Edited	Location	Disease	Accuracy
All	778	123 (16%)	87 (11%)	47 (6%)	84%
ProMED only	207	19 (9%)	14 (7%)	5 (2%)	91%
Google News only	547	104 (19%)	73 (13%)	42 (8%)	81%

At the most basic level, the accuracy of the classifier can be measured by the percentage of reports entering the system that need not have their disease or location classifications corrected in any way. At a more detailed level, we can examine the number of alerts requiring a correction of disease classification as compared with the number requiring a location correction. Table 1 provides a full breakdown of the classifier performance both by source and by disease and location. As shown, the overall accuracy of the system is 84%, thus correctly classifying 655 out of 778 reports over the one-month period from October 10 to November 9. As one might expect, performance on ProMED alerts, at 91%, is substantially better than on Google News reports (81%), as ProMED messages represent data curated specifically for disease outbreak reporting and follow a more regular structure.

There are, however, important limitations to this performance analysis. In particular, in some cases, the correction of the classification serves merely to shift between related categories, such as reclassifying Gastroenteritis as Norovirus, or UK as England. In other cases, the correction is more drastic, such as correcting Influenza to Equine Influenza, or Washington, DC to Washington State. Clearly the change is more significant in the latter cases, but we don't capture this distinction in the current analysis. As it is difficult to capture rigorously, for the moment we take the most conservative view in computing accuracy. As part of our ongoing research, we are developing more fine-grained metrics.

Discussion

As HealthMap is still in the early stages of development, a number of important enhancements are either currently under development or in the planning stage. The primary design goal of HealthMap is to provide broad coverage of ongoing outbreaks without overwhelming the user. In the pursuit of improved coverage, we are exploring the use of other sources, including additional news aggregators—such as Yahoo news, Factiva, and LexisNexis—blogs, and veterinary news sources such as the World Organization for Animal Health (OIE). In pursuit of improved filtering, we are developing natural language processing techniques for additional automated data categorization, such as clustering similar reports, identifying specific outbreak pertinence, distinguishing discrete outbreaks from endemic activity, and identifying reports indicating the absence of disease or the end of a previously identified outbreak.

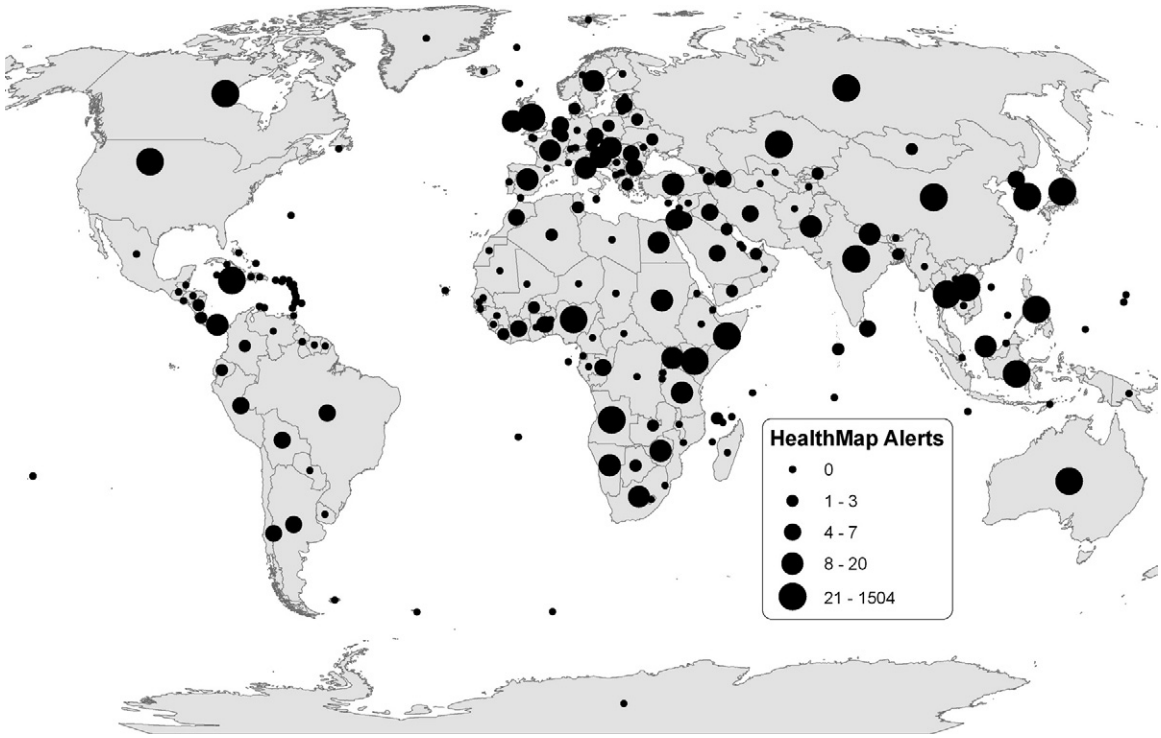

Figure 4. Geographic coverage of the HealthMap system.

As part of our own evaluation, as mentioned in the Formulation Process, an important goal of the system is to cover as broad a range of geography and disease as possible, without bias toward particular regions or pathogens. While the internal architecture of the system itself largely meets these goals (particularly as we add more geographical subdivisions around the world), the alert data we process and display leaves much to be improved. Because we currently rely heavily on the US edition of Google News for reports, the system is biased toward the US and Canada as well as other English-speaking countries around the world, as shown in Figure 4. To address this problem, we have developed a Spanish-language version of the system and are currently expanding to other languages and data sources as resources permit. However, given the uneven distribution of media and reporting resources around the world, we will continue to face this issue for the foreseeable future.

In addition to adding new capability, we are also working on improving the accuracy of the existing classifier, both by expanding the pattern dictionary and by improving the preparation module. We will add more locations and diseases, including administrative divisions for countries such as Indonesia, Brazil, Sudan and Mexico as well as major cities worldwide. For diseases, we will be adding more disease categories and refining our disease taxonomy, as well as tagging diseases with category metadata to allow for improved searching. We will also explore more advanced techniques such as fuzzy matching and Bayesian machine learning for improving the resolution and accuracy of our automated classification algorithms, as well as categorizing alerts by relevancy, clustering similar alerts, and extracting other useful attributes.[27–29] On the human side, taking inspiration from the highly successful Wikipedia model,[30] we plan to work with networks of experts to evaluate community collaboration as a mechanism for alert acquisition and classification.

As we expand functionality, performance will naturally become an increasing concern. We have a few optimizations in progress, such as moving to memory-based caching, more intelligent, "lazy" loading of the pattern dictionary, and better optimized database queries. We are also exploring ways to better employ client-side caching without overloading the browser.

On the frontend, we have plans to improve the user experience with added features and improved customization. Examples include keyword searching, RSS output, saved preferences, endemic background disease rates, notification messaging via email, and temporal visualization. (Notably, the EpiSPIDER system has already taken steps in this area, incorporating a timeline view of ProMED reports.[17]) We also plan to conduct a usability observation study, to gather feedback from our target demographic on priority features as well as how best to improve the HealthMap user interface.

Along with user-level evaluation, we are also working to develop more rigorous evaluation metrics for the integrated system, including its ability to cover a broad range of geography and pathogens, limit noise, detect outbreaks early, and accurately characterize alerts in each dimension of classification.

HealthMap is part of a new generation of disease surveillance systems that process unstructured and unclassified data sources. Comprehensive evaluation of these types of systems and data sources is also an important area and part of our ongoing and future research and collaboration with other disease tracking systems such as GPHIN, EpiSPIDER, MedISys, and Argus, would enable an in-depth comparison.

With that said, there are a few broad comparisons we can draw between systems. One key area is accessibility: HealthMap is freely available to the public, whereas some systems are currently closed systems, requiring either paid subscription or approved access. Another key area is in the use of automation. While we certainly perform manual curation in maintaining HealthMap, our goal is to maximize automation in order to leverage the human contribution. The value of a full-time staff of language and domain experts to read and analyze reports around the clock should also be addressed as part of a broader research initiative.[6,7]

Conclusion

The promise of HealthMap lies in its ability to extract useful, customizable messaging and views from a mass of unstructured data. While the site has already generated significant interest as a publicly available surveillance tool, many improvements remain to be made for it to be a truly useful resource for both public health professionals and the general public. In particular, adding more languages and expanding our usage of general data sources such as newspapers and blogs will increase coverage and further demonstrate the value of the visualization and filtering features. Moreover, only as time progresses, as more people use the system, and further significant outbreaks unfold in the global disease ecosystem, will we know the true potential of the software, and how best to improve it.

References ■

1. Grein TW, Kamara KB, Rodier G, Plant AJ, Bovier P, Ryan MJ, et al. Rumors of disease in the global village: outbreak verification. Emerg Infect Dis. 2000 Mar-Apr;6(2):97–102.
2. Heymann DL, Rodier GR. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect Dis. 2001 Dec;1(5):345–53.
3. Hiltz SR, Murray T. Structuring computer-mediated communication systems to avoid information overload. Communications of the ACM. 1985;28(7):680–9.
4. Berghel H. Cyberspace 2000: Dealing with information overload Communications of the ACM. 1997;40(2):19–24.
5. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. HealthMap: Internet-based emerging infectious disease intelligence. In: Institute of Medicine, editor. Infectious Disease Surveillance and Detection: Assessing the Challenges—Finding Solutions. Washington, DC.; 2007. 183–204.
6. Holden C. Netwatch: Diseases on the move. Science. 2006 December 1st.;314(5804):1363d.
7. Captain S. Get your daily plague forecast. Wired News. Available at: http://www.wired.com/science/discoveries/news/2006/10/71961. Accessed Apr 4, 2007.
8. Larkin M. Technology and public health: Healthmap tracks global diseases. Lancet Infect Dis. 2007 February;7:91.
9. Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health. 2006 Jan-Feb;97(1):42–4.
10. Mawudeku A, Blench M. Global Public Health Intelligence Network (GPHIN). 7th Conference of the Association for Machine Translation in the Americas 2006. Available at: www.mt-archive.info/MTS-2005-Mawudeku.pdf. Accessed Apr 26, 2007.
11. Eysenbach G. SARS and population health technology. J Med Internet Res. 2003 Apr-Jun;5(2):e14.
12. Morse SS, Rosenberg BH, Woodall J. ProMED global monitoring of emerging diseases: design for a demonstration program. Health Policy. 1996 Dec;38(3):135–53.
13. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. 2004 Jul 15;39(2):227–32.
14. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. Arch Med Res. 2005 Nov-Dec;36(6):724–30.
15. Health Threats Unit at Directorate General Health and Consumer Affairs of the European Commission. MedISys (Medical Intelligence System). Available at: http://medusa.jrc.it/. Accessed Apr 4, 2007.
16. Wilson J. Argus: Use of Indications and Warnings for Global Tactical Detection and Tracking of Biological Events." Georgetown Hosts 3rd Annual Conference on Infectious Disease; 2007; Washington, DC; 2007.
17. Tolentino H. Scanning the Emerging Infectious Diseases Horizon—Visualizing ProMED Emails Using EpiSPIDER. International Society for Disease Surveillance Annual Conference; 2006; Baltimore, MD; 2006.
18. Bernhardt JM. Centers for Disease Control and Prevention: Director's Blog. Health Marketing Musings 2006. Available at: http://www.cdc.gov/healthmarketing/blog_101106.htm. Accessed Apr 17, 2007.
19. O'Reilly T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software 2007. Available at: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. Accessed Apr 17, 2007.
20. Cayzer S. Semantic blogging and decentralized knowledge management. Communications of the ACM. 2004;47(12):47–52.
21. Garrett JJ. Ajax: A New Approach to Web Applications. 2005. Available at: http://www.adaptivepath.com/publications/essays/archives/000385.php. Accessed May 10, 2007.
22. Paulson LD. Building Rich Web Applications with Ajax. Computer. 2005;38(10):14–7.
23. Berners-Lee T, Hendelr J, Lassila O. The semantic Web. Scientific American. 2001;284(5):28–37.
24. Gratz NG. Emerging and resurging vector-borne diseases. Annu Rev Entomol. 1999;44:51–75.
25. Dobson A, Foufopoulos J. Emerging infectious pathogens of wildlife. Philos Trans R Soc Lond B Biol Sci. 2001 Jul 29; 356(1411):1001–12.
26. Brownstein JS, Holford TR, Fish D. Enhancing National West Nile Virus Surveillance. Emerg Infect Dis. 2004; In press.
27. Zheng W, Milios E, Watters C. Filtering for medical news items using a machine learning approach. Proc AMIA Symp. 2002: 949–53.
28. Chen H. Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. J Am Soc Inform Sci 1999;46(3):194–216.
29. Ribeiro-Neto B, Laender AHF, deLima LRS. An Experimental Study in Automatically Categorizing Medical Documents. J Am Soc Inform Sci 2001;52(5):391–401.
30. Giles J. Internet encyclopaedias go head to head. Nature. 2005 Dec 15;438(7070):900–1.