

# Content Collection and Analysis in the Domain of Epidemiology

Roman YANGARBER <sup>a,1</sup>, Peter von ETTER <sup>a</sup>, and Ralf STEINBERGER <sup>b</sup>

<sup>a</sup>*Department of Computer Science, University of Helsinki, Finland*

<sup>b</sup>*European Commission—Joint Research Centre, Ispra, Italy*

**Abstract.** We describe a system that tracks the spread of epidemics by automatically extracting content from the Web. The system continuously monitors a large set of news sources, extracts information from new articles, and accumulates the extracted facts in a database in real time. The system provides functionality for visualizing results, as well as alerting capability. We present the current state of the system and discuss challenges and further enhancements.

**Keywords.** Data analysis-extraction tools, Event-based systems, Knowledge management, Linguistics

## 1. Introduction

We address the task of global epidemic surveillance by means of gathering and analyzing textual content from Web-based sources. National and supra-national (e.g., European-level) Health Authorities require timely information about threats posed to the public by emerging infectious diseases and epidemics. The authorities invest substantial manual labor in searching news media for mentions of new or known threats. Electronic sources include newspaper sites and commercial news aggregators (Factiva, Lexis-Nexis). Search is typically performed by queries—Boolean combinations of keywords—that filter potentially relevant news items out of large collections. Relevant items are stored in a database, along with analysis and notes on actions taken in response to the threat.

The number of on-line news sources is steadily increasing. Further, it can no longer be argued that one needs to concentrate only on “local” events and that events in remote parts of the world are not relevant to local threat assessment—infectious agents do not respect national borders and can be transported across the globe in a matter of hours. Boolean keyword queries alone are not suitable for handling the growing volume and complexity, since queries that reach any considerable coverage are well-known to cause the problem of over-generation (low precision). Due to these factors, sophisticated text-analytic techniques are necessary to identify potentially important items. We describe an on-going distributed effort to combine an information retrieval (IR) system, *MedISys*, with an information extraction (IE) system, *PULS*. *MedISys* tracks web-based media and provides early-warning alerts based on sudden increases in reports about Public Health-related issues. *PULS* analyzes documents flagged as relevant by *MedISys*, and extracts from them structured *facts*, or metadata, about outbreaks of communicable disease.

---

<sup>1</sup>Corresponding author; e-mail: [firstname.lastname@cs.helsinki.fi](mailto:firstname.lastname@cs.helsinki.fi)

### 1.1. Related work

Several systems exist for surveillance of infectious disease outbreaks and gathering information about them. Some of these systems rely heavily on human/specialist analysis, others rely more on automatic processing; we are primarily interested in the latter.

Global Health Monitor [1] follows about 1500 RSS news feeds hourly and matches the words found in the new articles against a taxonomy of about 4300 entities: infectious diseases, country, province, and city names. Disease names are organised in an ontology with properties for synonyms, associated symptoms and syndromes, and hosts. Only those disease-location pairs are retained that frequently appear in an independent reference corpus. The system visualises the successful matches on a geographic map.

HealthMap [2] monitors articles from Google News and emails from the collaborative portal ProMED-Mail,<sup>2</sup> and extracts infectious diseases and locations. The results are passed to a human moderator for *manual* inspection, stored in a database and visually presented on a map. Diseases and locations are identified if words in the text match entities in the HealthMap taxonomy, which contains about 2300 location names and 1100 disease names. HealthMap identifies 20–30 outbreaks per day. Among the languages covered by ProMED, HealthMap currently covers English, Spanish and Russian, (in [2], only English-language processing is described).

The system we present in this paper covers a large number of sources, a wide range of languages (currently, over 40) and health-related topics (nuclear, chemical, and radiological incidents, bio-terrorism). Particular emphasis is placed on aggregating information collected from a variety of sources and across time, and using the aggregation to provide additional features and functionality for the users—for example, urgent warnings about unexpected spikes in levels of incidence in a given area.

## 2. MedISys: Information Retrieval

The *Medical Information System*, MedISys, gathers reports concerning Public Health in over 40 languages from thousands of Internet sources world-wide, classifies them according to several hundred categories, detects trends across categories *and* languages, and provides early-warning alerts to users. The MedISys site—[medusa.jrc.it/](http://medusa.jrc.it/)—maintains quantitative summaries of the latest epidemics reports. MedISys also tracks reports on toxins, bioterrorism, bacteria (e.g., anthrax), viral hemorrhagic fevers, viruses, medicines, water contaminations, animal diseases, etc. MedISys is part of the *Europe Media Monitor* (EMM) product family, [3], developed at the EC's *Joint Research Centre* (JRC).

MedISys currently collects an average of 40,000 news articles per day from about 1400 news portals around the world, from commercial news providers, and from about 150 specialised Public Health sites. The sources were selected to achieve a balanced geographic coverage. For each country, MedISys tracks mentions of disease names, and sends out alerts when it detects a sudden spike in incidence of a given country-disease pair, by comparing the statistics for the last 24 hours against a two-week rolling average.

The objective of MedISys is to save users' time and to give them access to more news reports in different languages. Early-warning statistics are calculated using infor-

---

<sup>2</sup>[www.promedmail.org](http://www.promedmail.org)

Viewing 604 events in 353447 documents

	Published	Source	Disease	Country	Begin	End	Total	†	Descriptor	✓
				china	2008					
[543] +	2008.05.08	xinhuanet_en	Hand , Foot , And Mo...	China	2008.05.01	2008.05.01	180		180 cases	
[543] +	2008.05.06	xinhuanet_en	Hand , Foot , And Mo...	China	2008.05.01	2008.05.01	42		42 sporadic cases	
[450] +	2008.05.02	angolapress_en	Enterovirus	China	2008.05.01	2008.05.01	1	†	A child	
[450] +	2008.05.02	bbc	Enterovirus	China	2008.05.01	2008.05.01	1	†	A child	
[450] +	2008.05.02	xinhuanet_en	Enterovirus	China	2008.05.01	2008.05.01	2 946		2,946 children	
[450] +	2008.05.02	reuters	Enterovirus	China	2008.05.01	2008.05.01	1	†	one more fatality	
[450] +	2008.05.02	thestar	Enterovirus	China	2008.05.01	2008.05.01	1	†	one more fatality	
[543] +	2008.05.07	voanews	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	--		people	
[450] +	2008.05.07	voanews	Enterovirus	China	2008.04.30	2008.04.30	2		Both children	
[543] +	2008.05.07	xinhuanet_en	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	28	†	28 deaths	
[543] +	2008.05.07	xinhuanet_en	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	28	†	28 lives	
[543] +	2008.05.07	AfghanistanSun	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	15 000		15,000 children	
[543] +	2008.05.07	AfghanistanSun	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	2 000		young children	
[543] +	2008.05.07	AfghanistanSun	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	28	†	--	
[543] +	2008.05.07	thestar	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	more		more cases	
[543] +	2008.05.07	news_yahoo	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	28	†	--	
[543] +	2008.05.07	news_yahoo	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	2 000		young children	
[543] +	2008.05.07	news_yahoo	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	more		more cases	
[543] +	2008.05.07	haveeru	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	--	†	children	
[543] +	2008.05.07	news_yahoo	Hand , Foot , And Mo...	China	2008.04.30	2008.04.30	15		15K children	

<< 1 2 3 ... 8 9 10 11 12 13 14 15 16 17 18 ... 29 30 31 >>

**Figure 1.** A table of extracted incidents, selected by the query in the top-most row: events in China, starting in 2008. (604 records matched these criteria. This is the 13th page-full of rows, from a total of 31 pages, with 20 rows per page.) The table shows that *hand, foot and mouth disease* (HFMD), caused by an enterovirus, is dominating the news. The left-most column shows, for each record, how many records in the database are linked with it, belong in its thread. (Note: at present, the system treats HFMD and enterovirus as two unrelated threads. Ideally, the system should recognize that HFMD and enterovirus records are related.)

mation aggregated from news articles across *all* languages. MedISys is therefore able to alert users to events that may not yet be available in their country or language. The development of MedISys is sponsored by the European Commission's (EC) Directorate General *Health and Consumer Protection* (DG-SANCO) to support national and international Public Health institutions in their work on monitoring health-related issues.

### 3. PULS: Extracting Facts

MedISys is a powerful tool for finding and categorising documents mentioning infectious diseases from a very large number of Web sources. For some users it is important to perform deeper analysis of the documents: is there an epidemic event occurring in the mentioned location(s), and if so, what is its extent and what threats does it pose.

This analysis is known as fact extraction, or information extraction (IE). After MedISys identifies documents where the *alerts* (i.e., disease names) fire, IE delivers more detailed information about the specific incidents reported in the documents. IE can help boost precision, since keyword-based queries may trigger on documents which are off-topic but happen to mention the alerts in non-relevant contexts. Pattern matching in IE assure that the keywords appear in relevant contexts only. IE focuses on *specific* scenarios involving diseases—outbreaks and epidemics, vaccination campaigns, etc.—as opposed to monitoring documents that mention diseases in a broader context.

PULS, the *Pattern-based Understanding and Learning System*, is developed at the University of Helsinki to extract factual information from plain text. PULS has been adapted to analyse texts in the epidemiological domain.<sup>3</sup> For each document found by MedISys, PULS extracts a set of *incidents* reported in the text. An incident is a structured representation of an event<sup>4</sup> involving some communicable disease, reported in a text news article. An incident is described by a set of fields, or attributes: location and country of the incident, name of the disease, the date of the incident, and information about the victims—their type (people, animals, etc.), their number, whether they survived or died, etc. The incident may cover a single occurrence or larger time interval, as in “dozens of chickens have died on the farm since the outbreak began last month.” The system also identifies events in which the disease is *unknown*, or undiagnosed so far, which are especially important for surveillance. A snapshot of the database is shown in Figure 1. These records were returned in response to a query, specified by constraints on some of the attributes. The user may enter the constraints—the country (*China*), starting date (*2008*), etc.—into the text boxes below the column/attribute names.

For detailed information about the design of PULS, please see, e.g., [4,5,6]. The system uses extensive domain-independent and domain-specific *knowledge bases*. An example of domain-independent knowledge is the location ontology containing names of countries, states, provinces, cities, etc. An example of a domain-specific knowledge base is the medical ontology, containing names of diseases, viruses, drugs, etc., organized in a conceptual hierarchy. PULS operates by pattern matching; the system has a large set of domain-specific patterns, which map from surface-syntactic representation of the facts in the sentence to the semantic representation in the database records. Populating the knowledge bases for a new domain is a major bottleneck in IE development. PULS employs weakly-supervised learning to reduce the amount of manual labour, by bootstrapping the knowledge bases from large un-annotated document collections, [7,8].

#### 4. Aggregation of Facts

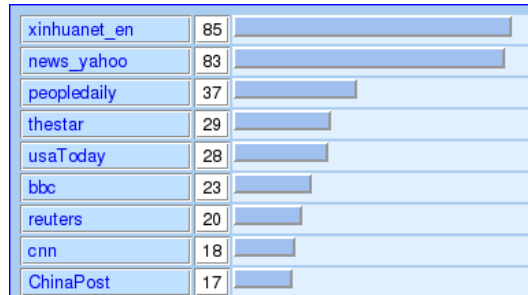
Achieving good accuracy in filtering and categorisation is clearly a crucial goal for MedISys. Another important problem the system must address is multiple reporting. Because MedISys tracks a large number of independent sources, the same event may be found multiple times, resulting in undesirable redundancy. MedISys and PULS aggregate individual reports into larger units to reduce the overload on the user.

In MedISys, similar news reports arriving close to each other (within 8 hours) are clustered together. If reporting continues steadily, articles from different days may be grouped into the same cluster. Article similarity measure is based on a vector-space representation of the first 200 word tokens. That is, not only multiple reports of the same story may be grouped together, but also similar reports about different cases for the same disease outbreak. This method also allows users to discard entire groups of *non-relevant* articles (e.g., discussions about vaccination campaigns) at once.

---

<sup>3</sup><http://doremi.cs.helsinki.fi/jrc>

<sup>4</sup>NB: The term *event* has different connotations in the medical and the computational communities. In the medical context, an event denotes the *entire course* of an epidemic episode, from its inception to completion. In IE literature, event denotes a single “factoid”, that may belong to a group of factoids, which together cover the course of an epidemic. In this paper, the latter meaning is intended.



**Figure 2.** The top of a histogram showing the coverage of events in Figure 1—the enterovirus epidemic in China—by different sources (obtained from the table, by clicking the *[chart]* button, in the upper-right corner.)

PULS aggregates individual *facts* into groups. In a typical IE system, documents are processed separately and independently: facts found in one document do not interact with information found in other documents. In PULS, after facts are extracted from each document locally, the system globally unifies them into *threads*. A thread is a chain of related incidents. At present, we group incidents by simple heuristics: they must share the disease name and country, and occur “close” together in time. Closeness is specified by a fixed time window.<sup>5</sup> Thus, each thread is a kind of a “bin” for related incidents, and provides an added level of abstraction over the individual incidents.

Note that each extracted record in the database may be equivalently viewed as *meta-data*, annotating the news item from which the event originated. Various views may be used to present these data to the user. Especially important are views that *aggregate* information according to the user’s criteria. PULS provides aggregate views in the form of geographic maps, and bar-charts or histograms, as shown in Figure 2.

Having a mass of news text automatically annotated with metadata enables us to apply a range of analytic methods, which cannot be done with plain text. The user may issue complex queries, aggregating over a choice of dimensions. For example, if an analyst needs to rank sources according to thoroughness or completeness of coverage (say, for a given epidemic), s/he can construct queries and histograms, as in Figure 2; by fixing the country parameter in the query, the histogram is made to reflect the coverage of the epidemic in the given country across all sources. More sophisticated methods, such as *link analysis*, may be applied to the metadata, e.g., to rank sources according to *importance*, or reliability (similarly to how search engines rank importance of Web pages).

## 5. Integration and Future Work

In conclusion, we describe the on-going integration effort between MedISys and PULS, and argue that—even at this early stage—the whole is greater than the sum of its parts.

An RSS tunnel has been set up between MedISys (running at the Joint Research Centre, in Italy) and PULS (at the University of Helsinki, Finland). Every 10 minutes, MedISys sends a batch of documents newly discovered from the Web, that trigger MedISys alerts, through the tunnel to PULS.<sup>6</sup> This is done in addition to the normal

<sup>5</sup>The time window is currently fixed at 15 days; it could be made more sensitive, e.g., dependent on the disease type, since for some diseases, outbreaks evolve more slowly and sparsely than for others.

<sup>6</sup>PULS currently processes only English-language documents; more languages to be covered in the future.

processing on the MedISys side, that is, tracking running averages for alerts, etc.

On the PULS side, the IE system analyses the received documents, and returns facts extracted from them back through the tunnel—in structured form (also at 10 minute intervals). The communication is asynchronous, with both sites operating in real-time.

The integration of MedISys and PULS offers users complementary functionality through a unified user interface. In the field of communicable disease outbreaks, which is covered by both systems, the combination of IR and IE yields additional advantages. The IE processing in PULS, which is computationally heavier, needs to be applied only to a small subset of the document collection, pre-filtered by MedISys. Secondly, the medical event extraction patterns act as a filter to focus only on disease outbreak reports. While MedISys captures all news articles mentioning diseases, for users interested in monitoring disease outbreaks, PULS's event extraction reduces the number of reports by about two thirds. (Results from a preliminary performance evaluation are found in [4].)

The current status of integration can be improved further: at present, the systems don't make full use of each other's information aggregation methods. The categorisation of news items by MedISys can help the analysis performed by PULS, and is yet to be exploited. The taxonomies used by the systems are overlapping, but have not yet been fully integrated. These and other issues are to be tackled in future work.

## Acknowledgements

We thank the anonymous reviewers for their comments. We are grateful to the members of JRC's Web Mining and Intelligence team who have contributed to the development of MedISys, especially the group leader, Erik van der Goot. Work on PULS is supported in part by the Academy of Finland grant 118653, National Center of Excellence "Algodan" (*Algorithmic Data Analysis*).

## References

- [1] S. Doan, Q. Hung-Ngo, A. Kawazoe, and N. Collier, "Global Health Monitor—a web-based system for detecting and mapping infectious diseases," in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.
- [2] C. Freifeld, K. Mandl, B. Reis, and J. Brownstein, "HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports," *J Am Med Inform Assoc*, vol. 15, pp. 150–157, 2008.
- [3] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby, "Europe Media Monitor—system description," EUR, Tech. Rep. 22173 EN, 2005.
- [4] R. Steinberger, F. Fuart, E. van der Goot, C. Best, P. von Etter, and R. Yangarber, "Text mining from the web for medical intelligence," in *Mining Massive Data Sets for Security*, D. Perrotta, J. Piskorski, F. Soulié-Fogelman, and R. Steinberger, Eds. Amsterdam, the Netherlands: OIS Press, 2008.
- [5] R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen, "Extracting information about outbreaks of infectious epidemics," in *Proc. HLT-EMNLP 2005*, Vancouver, Canada, 2005.
- [6] R. Grishman, S. Huttunen, and R. Yangarber, "Information extraction for enhanced access to disease outbreak reports," *J. of Biomed. Informatics*, vol. 35, no. 4, 2003.
- [7] R. Yangarber, "Counter-training in discovery of semantic patterns," in *Proc. ACL-2003*, Sapporo, Japan, 2003.
- [8] W. Lin, R. Yangarber, and R. Grishman, "Bootstrapped learning of semantic classes from positive and negative examples," in *Proc. ICML Workshop: Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.