EVENT RECOGNITION IN EPIZOOTIC DOMAINS

by

SWATHI BUJURU

B.E., Osmania University, 2007

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Computing and Information Sciences
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Approved by:

Major Professor
William H. Hsu

Abstract

In addition to named entities such as persons, locations, organizations, and quantities which convey factual information, there are other entities and attributes that relate identifiable objects in the text and can provide valuable additional information. In the field of epizootics, these include specific properties of diseases such as their name, location, species affected, and current confirmation status. These are important for compiling the spatial and temporal statistics and other information needed to track diseases, leading to applications such as detection and prevention of bioterrorism.

Toward this objective, we present a system (*Rule Based Event Extraction System in Epizootic Domains*) that can be used for extracting the infectious disease outbreaks from the unstructured data automatically by using the concept of pattern matching. In addition to extracting events, the components of this system can help provide structured and summarized data that can be used to differentiate confirmed events from suspected events, answer questions regarding when and where the disease was prevalent develop a model for predicting future disease outbreaks, and support visualization using interfaces such as *Google Maps*. While developing this system, we consider the research issues that include document relevance classification, entity extraction, recognizing the outbreak events in the disease domain and to support the visualization for events. We present a sentence-based event extraction approach for extracting the outbreak events from epizootic domain that has tasks such as extracting the events such as the disease name, location, species, confirmation status, and date; classifying the events into two categories of confirmation status- confirmed or suspected. The present approach shows how confirmation status is important in extracting the disease based events from unstructured data and a pyramid approach using reference summaries is used for evaluating the extracted events.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I offer my sincere gratitude first to my advisor, Dr. William Hsu, for his valuable suggestions and guidance to my research work. He has been a constant source of inspiration for me throughout my master's work. It has been an intellectually stimulating and rewarding experience for me to work with him. I truly feel privileged to have joined his research group. I also thank the members of my graduate committee, Dr. Doina Caragea and Dr. Gurdip Singh, for all their advice and encouragement.

I also thank the members of the KDD (Knowledge Discovery in Databases) group at Kansas State University specially Svitlana Volkova who helped me to complete this study successfully. This research would not have been completed without their support.

I would like to thank my family members and friends for supporting and encouraging me to pursue this degree. Their encouragement was a key factor in successful completion of my degree.

1 Introduction and Problem Formulation

1.1. Introduction

The epizootics have a direct impact on animal lives and an indirect impact on world's economy and trade. To predict and control severe epizootic outbreaks, there is a high requirement to discover locations which are disease prone and take precautionary actions against the spread of epizootics. Useful information about the epizootics outbreak could be available through many resources such as new papers, online archives, emails, and blogs, several health and agricultural organizations and so on. Information in most of these resources would be present in some sort of unstructured format while some may contain in structured format. For an instance, data from health and agricultural organizations would be in a structured format while data from blogs, other online archives would be in an unstructured format. Thus there lies a need for the available raw data (from blogs, online archives) to be summarized in terms of space and time using machine learning and data mining techniques for effective usage of raw data. It would take many years to manually analyze and summarize the data. To evade this manual effort, a system, which can automate the summarization process, is highly necessary. To support this need, an event extractor is developed which would summarize the data in terms of events, thereby making raw data in unstructured format more structured and organized.

High voluminous amount of information about disease outbreaks is available in this world in the form of unstructured data which mostly exists in textual format. Unstructured data could be scattered in many places and could exist in any form such as emails, literature papers, research papers, news articles, and blog posts. It can also exist in any human readable and spoken language. To take an example, news archives from epizootic domain contains factual information about the disease such as the location that is been affected, the time when the disease was reported, the number of victims affected, the status of the disease and many more. This kind of news archival information can be useful for analysts, or to government agencies that monitor infectious disease outbreaks. Thus, structuring of data from the unstructured format of the news archives could be beneficial in a lot of ways. The available information could be queried and referred more easily if unstructured data is represented in a structured format.

1.1.1 Animal Disease Impacts

Epizootics have a great impact on humans, economical trade and economical crisis and the following sections focus more on the details of the negative impacts of outbreak diseases.

1.1.1.1 Impact on Humans

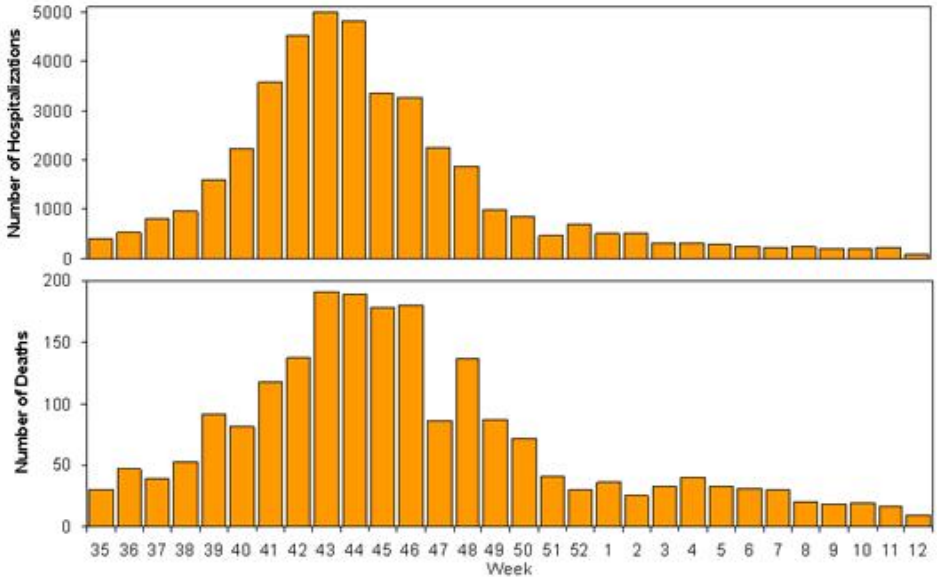


Figure 1-1: *H1N1 Hospitalizations and deaths reported to AHDRA, National Summary, during Aug '09 - March '10 [38]*

For centuries, humans have been affected with diseases that originate from animals. Many of the disease causing agents, responsible for human epidemics, have their origins from animals including diseases such as tuberculosis, influenza, bubonic plague, food-borne illness, and AIDS. This episode of animal borne diseases also referred as zoonotic diseases or zoonoses has been continuing for years and is spread throughout the global system. Scientists expect the rise in zoonotic diseases episodes to continue [25]. One cannot predict or expect when or where the new zoonotic pathogen will emerge or what its ultimate impact might be. Hence investigation at the first sign of emergence of a new zoonotic disease is rather important. The following Figure 1-1 gives an example of how a disease (H1N1) impacts human beings.

1.1.1.2 Impact on Economic Trade

Animal disease outbreaks have a great impact on economic trade throughout the world. Some of the outbreaks which had a big impact previously are listed: Bovine Spongiform Encephalopathy (BSE) found in cattle in the European Union (EU), Canada, and United States

(US); swine fever in the EU; and Foot-and-Mouth Disease (FMD) found in cattle in Taiwan, Japan, Korea and other parts of the world and many more. When such an outbreak happens, trade bans are usually imposed on exports coming from counties infected with such kind of disease outbreaks. The following Figure 1-2 gives an example of how port exports are affected through the disease impacts.

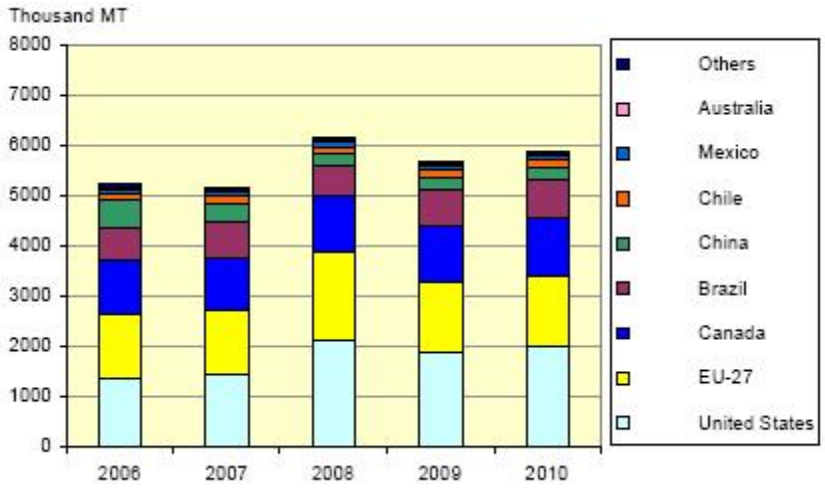


Figure 1-2: *World Pork Exports by country [27]*

1.1.1.3 Impact on Economic Crisis

Impacts after disease outbreaks are costly when evaluating prevention and mitigation measures after post disease outbreaks. Immediate impacts after a disease outbreak include a reduction in the productive capacity of the animal products industry and a subsequent reduction in the supply of meat products. Also, disease outbreaks may reduce the demand for meat and meat products. Allied agribusinesses bear an initial loss in the supply of meat products, and later, increased costs when locating and certifying safe food supplies [32].

This introductory chapter provides the details of the impact of epizootics on society, rationale behind the project, problem statement and objective for event extractor being developed.

Related work that includes discussion of the types of animal disease surveillance systems both automated and manual; the differences between the automated surveillance systems and the proposed system are explained in Chapter 2. The extractors used for tagging the entities within the raw text, the details of the confirmation status extractor & its importance and the different gazetteers developed for confirmation status extractor are discussed in Chapter 3. Different

stages in the overall event extraction system, the event extraction methodology, visualization module specifics are given in Chapter 4. In Chapter 5, the discussion on the example of the data set used for processing is given, followed by the discussion of the different experimental set ups using the confirmation status gazetteers and the details on the evaluation metric used are given. The results from the different confirmation status gazetteers and the screen shots of the visualization maps produced are presented in Chapter 6. Chapter 7 describes limitations of the system and also some future work proposals.

1.2 Problem Formulation

From the information given in the previous sub-section, it is clear that a process is required which would extract structured information from the available unstructured data format. This whole process of extracting structured data could be referred as Information Extraction. Through information extraction systems, one can recognize entities such as person names, locations, organizations, and disease names and also relationships between entities in natural language text. Once the entities and relations are identified, the structured data can be stored in database for quick references and querying. Also we can answer a lot of questions by querying at the derived database instead of going through the complete documents.

The classifiers that can tag disease names, species, dates, location in the unstructured data were developed in the Knowledge Discovery in Databases (KDD) laboratory. All these extractors can be used for event extraction process. My main goal, for this project is to develop an effective event extraction system, is split into three sub goals. The first sub-goal of this project is to extend the attributes <*disease name, date, location, species*> of events by adding a new attribute called "*confirmation status*" that conveys the severity of the disease and can classify the event related sentences. Second goal is to develop the event extractor that can summarize the eventual information automatically from the unstructured data in the form of <*disease name, date, location, species, confirmation status*>. And finally the last goal is to plot the events on *Google Maps* and to be able to cluster the events on the map.

1.2.1 Challenges in unlocking unstructured data

As mentioned previously, information could be either structured or unstructured. Structured data is the one that has a data model and is easy to understand by human/system. Relational databases and spreadsheets are examples of structured data. Unstructured data is the

one that doesn't have some specific model to follow, making it a challenge for human/computer to follow. Examples of unstructured data include emails, blogs, news, literature and so on.

Most of the health organizations gather information from user opinions, polls, literature, survey feedback, reviews, and preserve it in the form of blogs, forums, portals and many other online sources, either as text, audio, images or other forms. This way the information is available in an unstructured format which makes it a difficult to analyze or use it.

"We are drowning in information but are starving for knowledge," says Mani Shabrang, technical leader in research and development at Dow Chemical Co.'s business intelligence (BI) center in Midland, Mich. He also adds "Information is only useful when it can be located and synthesized into knowledge."

By using new text mining tools, unstructured data can be unlocked for extracting meaningful relationships and summarizing the knowledge. With such kind of automatic tools, one can easily convert the raw text into knowledge for easy understanding.

Several challenges are to be faced while text mining the unstructured data. Due to the difficulty level involved in identifying the relevant information over irrelevant ones, many stringent rules have to be framed for classifying the relevant information. Framing the rules alone is not sufficient for effective mining, cases of disambiguation, problem of aggregation, spurious results, finding the specific time mentioned should also be taken into consideration.

Problem of disambiguation can be observed in locations, abbreviations and in many scenarios. The below examples highlight some scenario of disambiguation.

Example 1: Cricket fever in India. - Need to disambiguate cricket as a game rather than a new virus;

Example 2: "BSA" AND "bovine serum albumin" NOT "body surface area".

Another challenge which has to be dealt with is that of aggregation of data. After retrieving the knowledge from the unstructured data, it has to be aggregated/ summarized for further analyzing part. But, there could be a case where different figures could be present or conflicting information might have been given in different sources.

As an example of multiple reports of a single outbreak, DEFRA (Department for Environment, Food and Rural Affairs) [17] reports: "100 cases reported on 21st Feb, 2010", while a regional site reports: "90 cases infected on 21st Feb, 2010".

In these kinds of scenarios where spurious results are caused by redundant reports, it is better to rely on most reliable and trusty organizations and to get the average number of reported cases from most of the reliable organizations.

The most important piece of information from any text is temporal information for maintaining statistical figures. In most of the reports, the case or issue will have temporal information such as “reported today”, “suspected 20 as on Monday”, *etc.*, there should be some other means to refer back to the text and answer the questions, “What is `today’?” or “What date was `Monday’?” One of the solutions for this problem could be to trace the reported date of the article itself rather than the case.

1.2.2 Problem Statement and Objective

We aim to develop a system [40], as shown in Figure 1-3, which extracts related and necessary attributes from unstructured data and convert them into structured tuples, representing events, which would answer questions pertaining to diseases in context. The main objective of the proposed system is to frame the stringent rules for effective event extraction from unstructured data. Another main objective of the system should be to minimize the recognition error rate, when extracted events are compared using an evaluation metric and can be plotted in a visualization map.

The events would be extracted from raw data and outputted as tuples in the below format.
<*Disease-Name, Location, Species, Date, Confirmation-Status*>

- ✓ *Disease-Name*: It would give the name of the specific disease in context.
- ✓ *Location*: Place, where the disease outbreak was noticed.
- ✓ *Species*: Specimens that have been affected with the disease.
- ✓ *Date*: Specify the date or period of disease outbreak
- ✓ *Confirmation-Status*: Identify the severity of the disease outbreak.

Our desired event extractor should be able to recognize the events from the outbreak documents that have past information or latest information and following are examples for past/latest outbreak news.

- Example 1: "*Rift Valley Fever was first identified in 1930, after an epidemic among sheep in Kenya"* is an example for outbreak event that occurs in the past.
- Example 2: "*NIGERIA - The country has reported a new outbreak of* Foot-and-Mouth *disease at Chanchaga in Niger State on Aug 27, 2010.*" is an example of the latest outbreak event.

The proposed system should be able to differentiate the non-event related sentences from event related sentences and following are the examples in both of the categories as stated.

- Example 3: "*Rift Valley Fever killed 124 people in the United Kingdom between September 2000 and May 2001, while 760 of the 884 infected people recovered*." is an example of an event related sentence.
- Example 4: "*Rift Valley fever (RVF) is an acute, fever-causing viral disease that affects domestic animals (such as cattle, buffalo, sheep, goats, and camels) and humans*" is an example of non event related sentence.

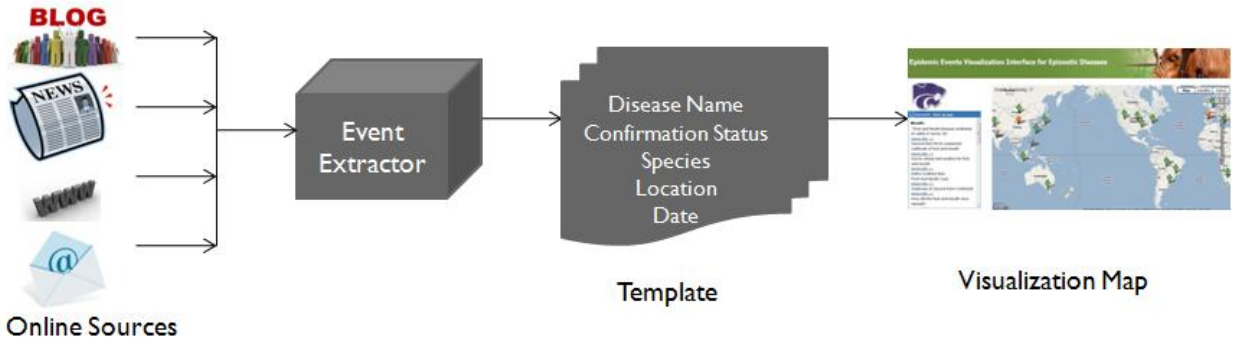


Figure 1-3*: Entire System Components*

2 Related Work

Now that we have seen the different disease impacts which are possible, we would look into the previous and present systems/ works which keep track of disease outbreaks.

2.1 Animal Disease Surveillance Systems

There are many surveillance systems that maintain the outbreak information about animal diseases and also provide web interfaces for visualization. These systems include both automated and manual surveillance systems. Most of these systems are from health organizations, state and federal government agencies. The primary goal of these systems is to maintain the database of animal disease outbreaks and to support the epidemic surveillance system. The main resources of information for these systems include news, blog, emails, literature and many more. In the next section we will look into the different kinds of surveillance systems, both manual and automated.

2.1.1 Manual Surveillance Systems

The following surveillance system databases are updated manually by gathering outbreak information from different sources.

2.1.1.1 WAHID (World Animal Health Information Database) interface & Handistatus II

WAHID is one of the biggest database sources for animal disease outbreaks. It is an interface supported by World Organization for Animal Health (OIE), an intergovernmental organization responsible for improving animal health worldwide. This system maintains database of animal disease outbreak from 2005 and thereafter [23].The main interface is shown below in Figure 2.1

.

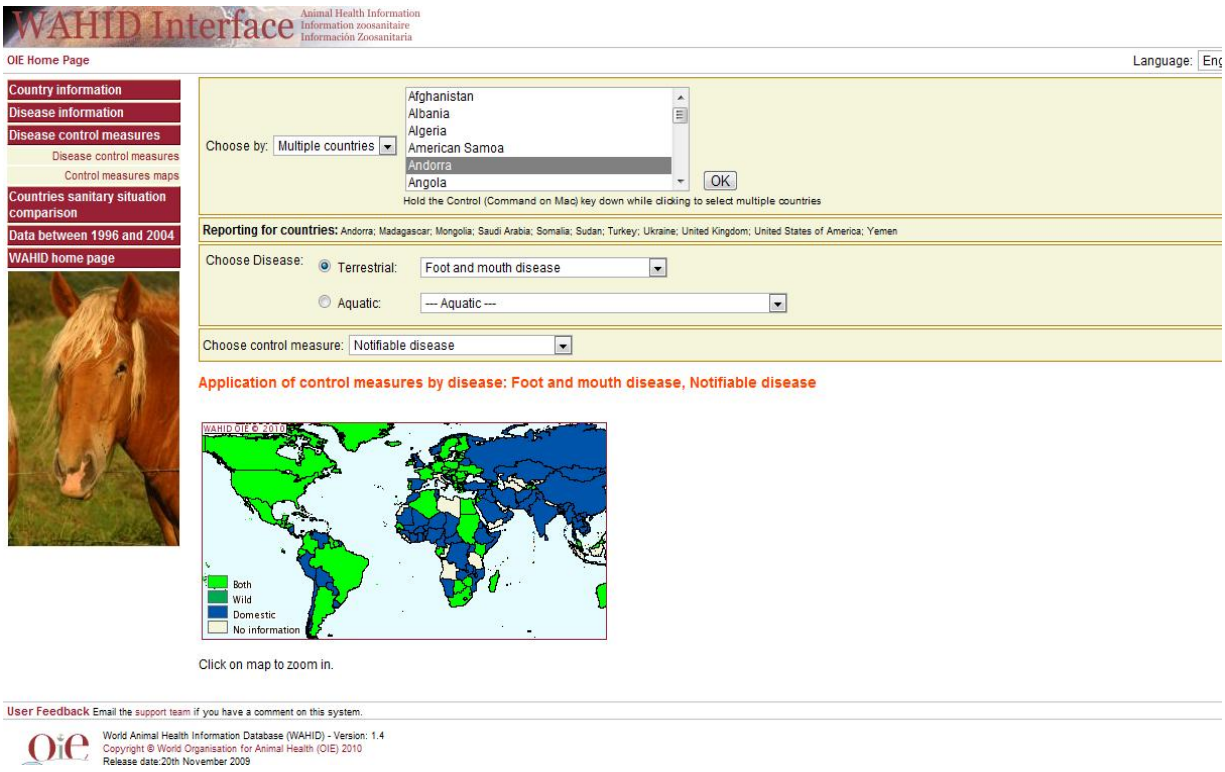


Figure 2-1: *World Animal Health Information Database (WAHID) Interface [23]*

The main drawback of this system is that it does not provide information on outbreaks before 2005, instead it is separately maintained in different web interface, Handistatus II [1] which makes harder to analyze the data.

2.1.1.2 WHO -The Global Health Atlas

WHO -World Health Organization (Global Health Atlas) [29] launched the first global online atlas of infectious diseases which is build over the features of *HealthMapper*. Over 300 indicators and 20 infectious diseases are maintained in the database. This atlas is bringing together the statistical data of animal diseases for analysis and comparisons at country, regional and global levels. Maps are used to display data on the prevalence of individual diseases and so on and following is the Figure 2-2 showing screen shot of WHO's visualization map.

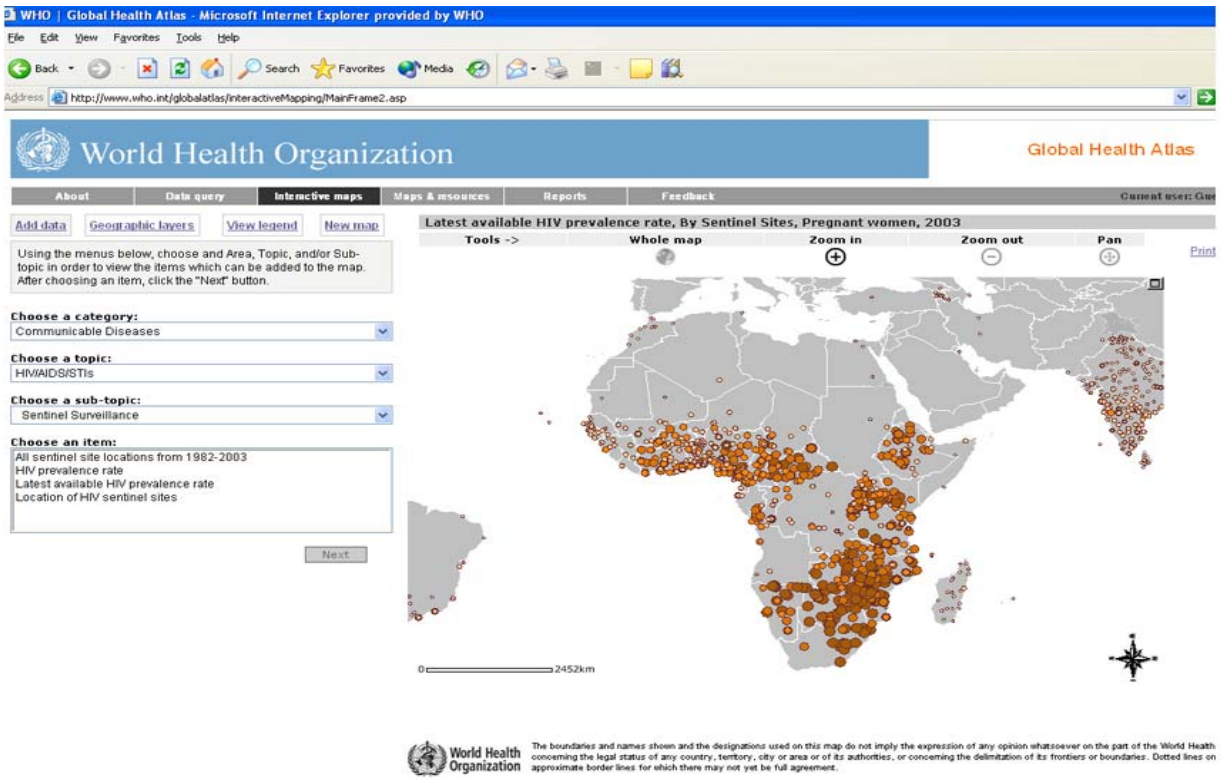


Figure 2-2: *WHO- Global Health Atlas showing data on Interactive Map [29]*

2.1.1.3 EMPRES- Global Animal Disease Information System

EMPRES-I (EMPRES - Global Animal Health Information System) [2] website is a global animal health information system of FAO's emergency prevention programme for Transboundary Animal Diseases (EMPRES), that focus on the users need of easily finding and collecting in one place all the information available for animal health and transboundary animal diseases. It enables user to access and retrieve animal disease outbreaks information throughout the world according to the search criteria (disease, date, species, location *etc*.) and also allowing user to save the data into pdf or excel files for further analysis. It also provides maps that allows user to select outbreaks/cases from the database and represent them graphically as charts (by time or by location) or geographically on a map that can be tailored by adding optional layers, such as livestock population, biophysical layers, socioeconomics, animal health, trade and so on. These layers are created and maintained by the Global Livestock Production and Health Atlas (GLiPHA) [3] which is a user-friendly, highly interactive electronic atlas using the Key Indicator Data System (KIDS) [4] and the following Figure 2-3 gives the screen shot of EMPRES.

Figure 2-3: *EMPRES- Global Animal Disease Information System web interface [2]*.

2.1.1.4 National Notifiable Diseases Surveillance System-CDC

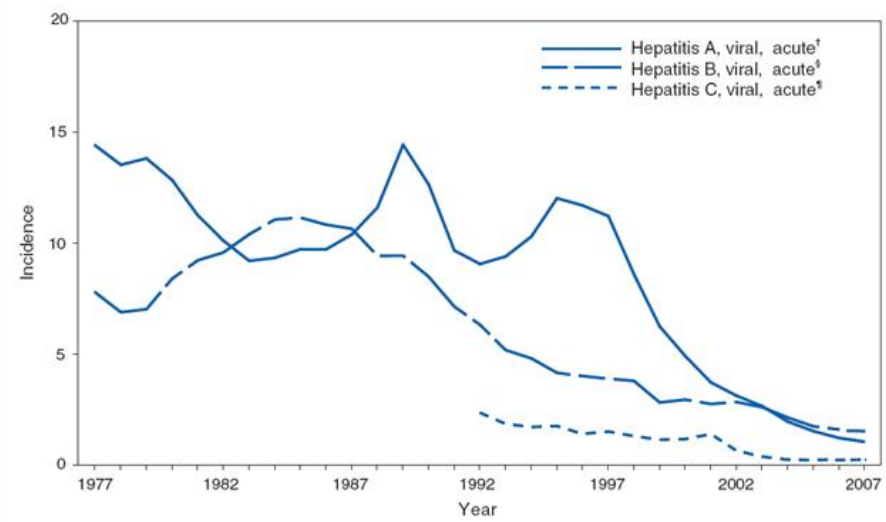


Figure 2-4: *Hepatitis, Viral Disease Incidence by year, US 1977-2007 [31]*

CDC - Centers for Disease Control and Prevention has taken the responsibility of gathering the data concerning nationally notifiable diseases such as cholera, smallpox, plaque and many more, when US marine hospital service authorized it to do so. CDC has been preparing

the annual summary reports, in the form of graphs and maps, for all the notifiable diseases. Each year, the reports are been updated for each noatifiable disease [26]. A sample graph depicting a snapshot for the Hepatitis, Viral disease Incidence, by year — United States, 1977-2007 is shown in Figure 2-4.

2.1.1.5 USGS-Global Wildlife Disease News Map

USGS-United States Geological Survey [5] works on reporting the wildlife mortality events that provide information on locations, species and causes of death. This information can be used for framing the disease prevention and mitigation strategies and also to balance the interconnections between the human and animals. Global Wildlife Disease News Map [30] depicts the events from the news from Wildlife Disease News Digest [6] which have got a geographical reference within the last 45 days. The map is been updated periodically with news over multiple times within a week and it maintains filters such as wildlife health topic, wildlife/human topic, domestic animal/wildlife topic, disease type, species, country, and date.

2.1.1.6 Center for Food Security and Public Health (CFSPH)

Center for Food Security and Public Health (CFSPH) at the Iowa State University College of Veterinary Medicine, is working to increase the bioterrorism and foreign animal diseases among veterinarians, general public, and farmers. CFSPH maintains fact sheets, power point presentations and handouts on important emerging diseases in USA [18].

2.1.1.7 FMD BioPortal

The FMD BioPortal [22] was developed by a joint effort of the Institute for Animal Health (the FMD World Reference Laboratory) at Pirbright, England, the Artificial Intelligence Laboratory at the University of Arizona and the FMD Laboratory at the University of California, Davis. Its primary objective was to develop a web based system using the FMD based data, used at the Pirbright Lab and make it available to the general public, applying the data basic search and analytic tools with both graphical and tabular representation of the data. It also provides an option of downloading selected records. The data includes the outbreaks of FMD, submitted to the laboratory of Pirbright since 1957. The FMD Bio Portal also includes the FMD News which is a near real time web search to identify and capture FMD- related news items appearing worldwide.

2.1.1.8 BioPortal

It is a powerful bio surveillance tool which can be used for various data analysis applications. It is a collaborative effort between the International BioComputing Corporation (IBC), a University of Arizona spin-off company, and the Artificial Intelligence lab at the University of Arizona lab. The main goal of this portal is for information sharing, data analysis and visualization of disease outbreak and to improve the health awareness of infectious diseases among public [20].

2.1.1.9 DEFRA: Department for Environmental Food and Rural Affairs

The Department for Environmental Food and Rural Affairs [17] is a government department in United Kingdom that protects the environment for the future generations and making the environment more sustainable and thereby improving the quality of life for well being. It maintains the database for all the notifiable diseases which includes data sheet, and summary profile [28].

The first three systems, WAHID, EMPRES, and WHO Global Atlas, are being maintained at international level and National Notifiable Diseases Surveillance System-CDC, USGS-Global Wildlife Disease News Map, Center for Food Security and Public Health (CFSPH), FMD BioPortal, DEFRA are maintained at regional level.

2.1.2 Automated Surveillance Systems

News about disease outbreak could be in any form *i.e.*, email, literature, research, news, blog information and in any human speaking language. It would be costly and time consuming if we adapt manual methods for maintaining animal disease monitor systems. This problem calls for the presence of Automated Systems to maintain the disease outbreak database. They are less time consuming and less expensive compared to the manual systems. Developing an Automatic Surveillance System involves several technical challenges such as accurately labeling linguistic markers for semantic roles and correctly interpreting the geo-temporal dynamics of disease/virus spread. Other challenges include coping with a very large volume of data, the need to interpret described properties of diseases as quickly as possible, recognizing each named entity and containing phrase in many languages, and finally inferring inherent context from natural language text, such as for word sense disambiguation. These challenges characterize the difficulty of the task.

There are many web surveillance projects that require lesser degree of automation for analyzing disease epidemics and epizootics. Some of them include *BioCaster [28]*, *MedISys [36]*, *GPHIN*, *Argus*, *ProMed-Mail [16]*, *EpiSPIDER [8]* and *HealthMap* .

2.1.2.1 BioCaster

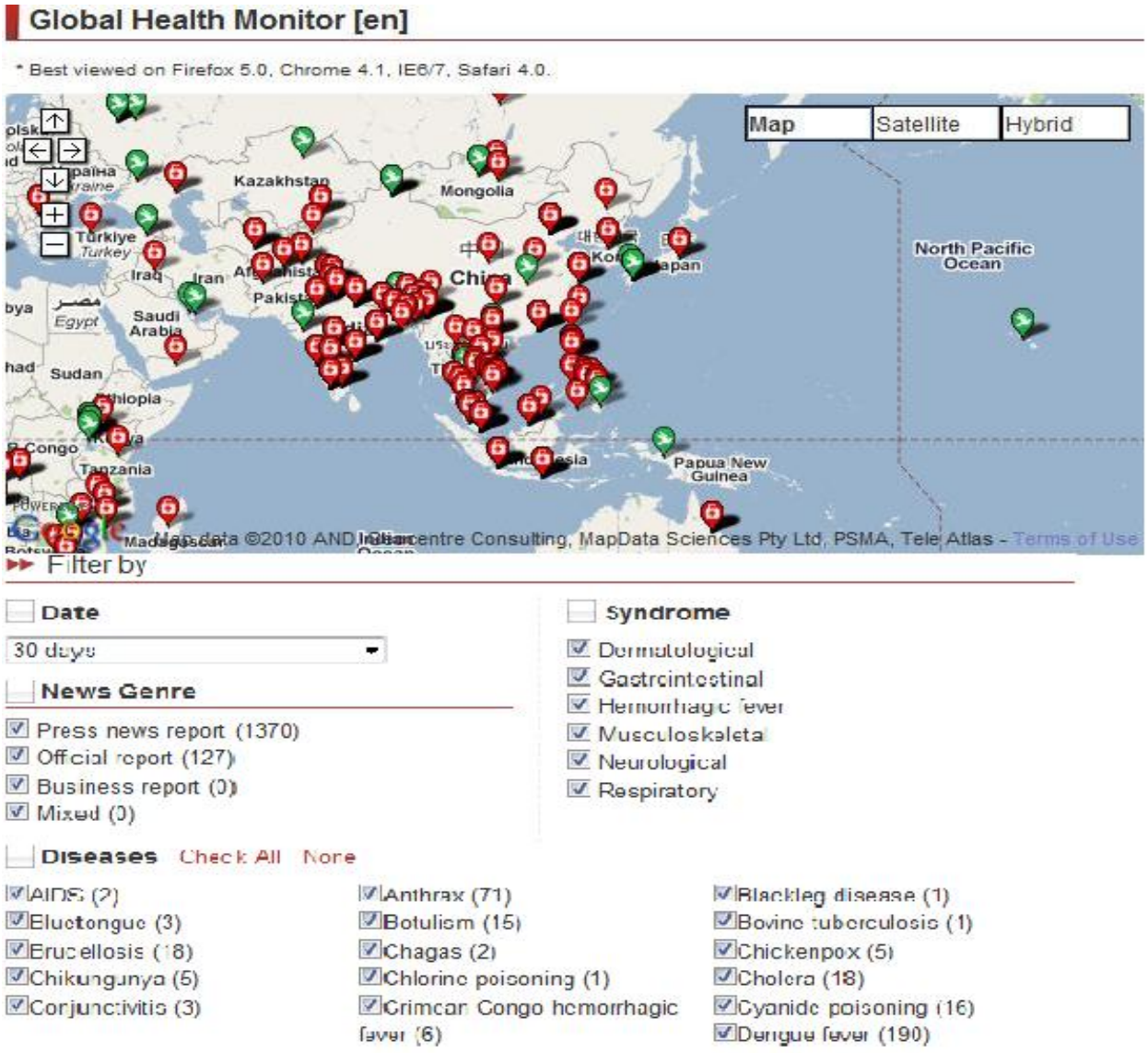


Figure 2-5: *Global Health Monitor: A web interface for BioCaster [28]*

BioCaster [28] is a non-governmental web surveillance system that uses ontology based approach, which has information regarding diseases and their pathogens and geographical location names including latitude and longitudinal values in eight different languages, as its text

mining technique. *BioCaster* aims in providing an early warning monitoring system in reporting epidemic and environmental diseases (human, plant, and animals). In order to achieve its goal, it aggregates the outbreak news from various resources and processes them to analyze the unusual patterns. It is trying to extend and gather other sources of textual information in its growth. It continuously analyzes information at every hour reported over 1700 RSS feeds and classifies them into topical relevant and non-relevant and plot the events on the *Google Maps*. It gets updated every 30 minutes. The Figure 2-5 shows us the web interface used by *BioCaster*, known as Global Health Monitor.

2.1.2.2 MedISys

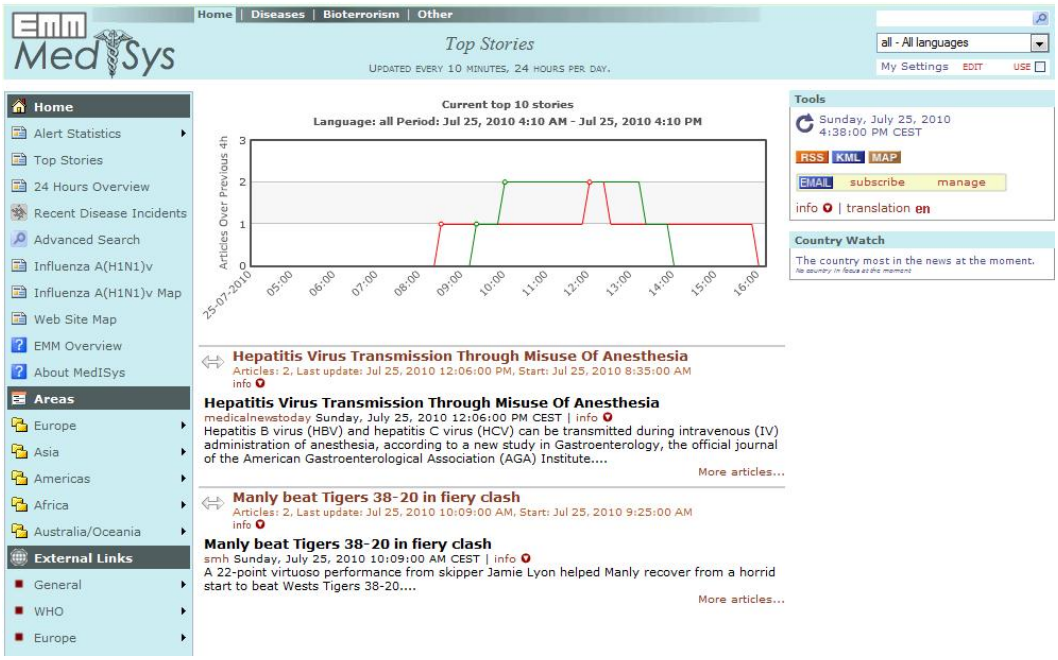


Figure 2-6: *Alert Statistics on H1N1 for 24 hours (July 25th) on MediSys* ***[36]***

MedISys (*MedISys*) [36], the Medical Information System, is an automatic web surveillance tool that gathers health information from various sources and categorizes them into various disease sub types in 25 different languages. *MedISys* processes over 20000 articles per day from over 4000 sites of approximately 1600 news sources (news and medical sites) in 45 languages. It dynamically updates its disease statistics every 10 minutes. *MedISys* operates 24/7 and 365 days per year. After aggregating the information, *MedISys* alerts the users regarding any emerging diseases. *MedISys* sends the documents to *PULS* (*Pattern-based Understanding and*

Learning System) [7] and *PULS* returns events and updates its database. Figure 2-6 and Figure 2-7 shows the web interfaces for *MedISys* and *PULS* respectively.

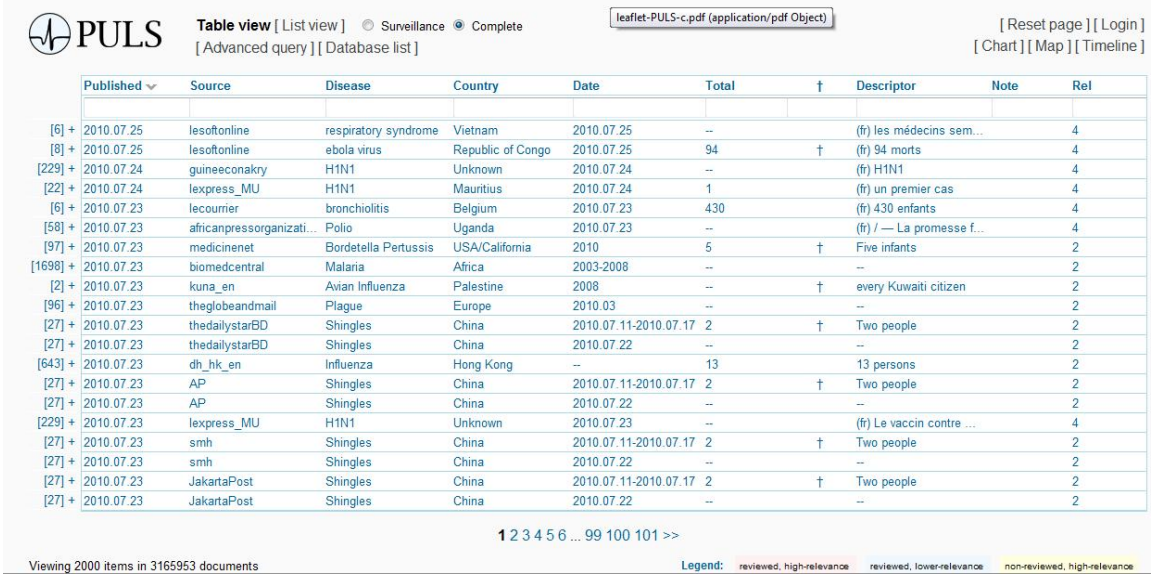


Figure 2-7: *Events reported on PULS* **[7]**

2.1.2.3 HealthMap

HealthMap [21] is a web-based system which collects data from various data sources and presents a unified and comprehensive view of the current global states of infectious diseases to human and animal health. This web site uses information from a variety of electronic sources: online news wires, Really Simple Syndication (RSS) feeds, expert-curated accounts (*ProMED-Mail*), and validated official alerts (WHO). Once this information is collected, the data now is summarized by source, type of disease and geographical location and is displayed over an interactive map to the user of the system. The specialty of this system is that it generates "meta-alerts" color codes depending upon the reliability and amount of data. After categorizing based on location and disease, tagging of data is done. The data is basically tagged as the following: breaking news, warning, follow-up, background- context, and not disease related. Currently it processes 133.5 disease alerts per day on an average, with approximately 50% categorized as breaking news. As of 20 November 2007, *HealthMap* had processed over 35,749 alerts across 171 disease categories and 202 countries or semi-autonomous/overseas territories since it was launched. Among these, the major alerts came from news media (92.8%), followed

by *ProMED* reports (6.5%). A sample map depicting the news alerts received by *HealthMap* is shown in Figure 2-8.

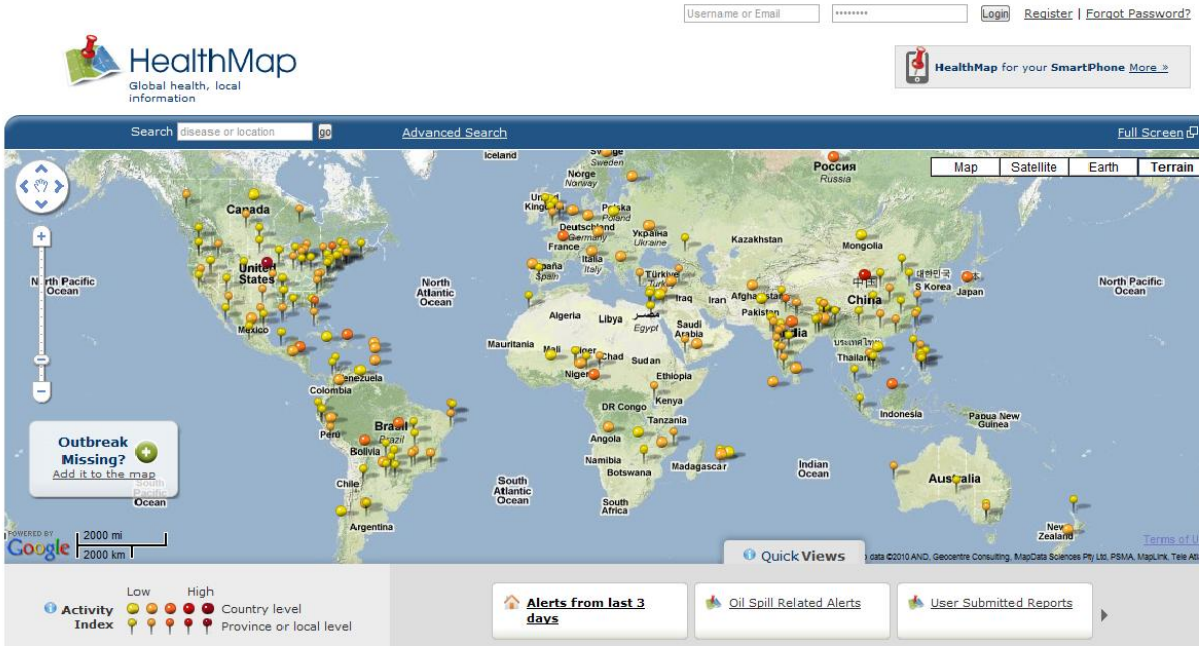


Figure 2-8: *HealthMap Visualization Support* ***[21]***

2.1.2.4 EpiSPIDER Project

The *EpiSPIDER* [8] project was designed to serve as a visualization support for Program for Monitoring Emerging Diseases – ProMED [16] mail reports, a global electronic system for reporting emerging infectious diseases. ProMED has a global mailing list with more than 35,000 subscribers to which it provides an integrated and summarized reports on outbreaks of diseases of infectious or toxic etiology that affect plants, animals and humans. *EpiSPIDER* connects to news portals and uses natural language processing for transforming unstructured data to structured data.

2.2 Differences among the Automated Surveillance Systems and the Proposed System

Following Table 2-1 describes the differences between the *BioCaster*, *MedISys*, *HealthMap* and RuleBasedExtractor [41].

	BioCaster	***MedISys***	***HealthMap***	**Rule-Based Extractor**
Started In	2006	April-07	September-06	June-2010
Sources	1700 RSS Feeds	20000 articles per day from over 4000 sites of approximately 1600 news sources	Google and Yahoo news media (92.8%) and *ProMED-Mail*s (6.5%).	Any epizootic homogenous focused domain set
Taxonomy	50 infectious diseases and locations (243 countries and 4,025 sub-countries)	4300(human and animal diseases) 70,000 locations	2300 location names and 1100 disease names	more than 1000 animal diseases and million locations from NGA GEOnet Names Database
Outbreaks /day	25-30 locations on 40 diseases /day	50k news reports/day	20-30 outbreaks /day	Depends on the size of collection
Kind	Automatic	Automatic	Manually verification of reports	Automatic
Languages	8 languages	45 languages	7 languages	one language

Output	disease/location pairs	Disease name, time, place , type of victims (human/animal), status(dead/hospitalized)	disease and location	Disease name, time, location, species and Confirmation Status
Visualization Map	Yes	Yes	Yes	Yes
Cluster Visualization	No	No	No	Yes

Table 2-1: *Differences among Automated Surveillance Systems and the Proposed System*

3 Information Extraction System Components

This chapter gives the details of the event extraction tools used in the project. In section 3.1, we will look into the generic NER – Named Entity Recognition and we will follow up with all the system components used in the project in section 3.2.

3.1 Named Entity Recognition (NER)

Named entity recognition (NER) is a key part of many information extraction systems. It involves identification of proper names in texts, and classification into a set of predefined categories of interest. The categories could be person, location, organization, date/time expressions, quantities *etc.* In the present system; the categories disease names, location, species, confirmation status, date extractors are been taken into account. NER is just recognizing the entities but not the events. Implementing a practical NER system is not easy, as it may have to handle many ambiguous senses. For example: John Smith (company *vs.* person), Washington (person *vs.* location), 1945 (date *vs.* time).

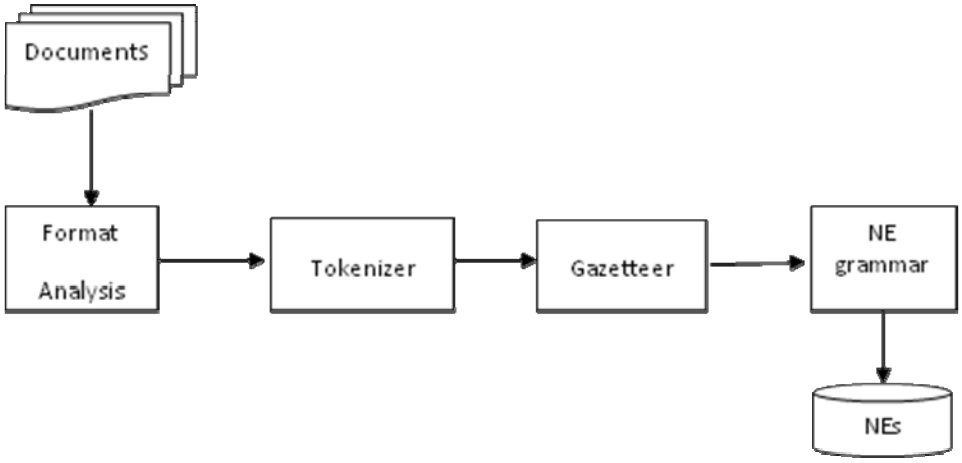


Figure 3-1: *NER - Named Entity Recognizer*

The generic architecture of any NER system is shown in Figure 3-1 and will have some set of documents formatted into the required form and the document is chunked into tokens using tokenizer. Tokenizer is to segment the text into tokens such as words, sentences, *etc*. And each token is tested for occurrence of any named entity using gazetteers. Gazetteers are the

dictionaries for named entities NE (places, organizations, *etc*.). If the extracted entity satisfies NE grammar, then it is extracted as NE otherwise it is skipped.

3.2 System Components (Extractor Tools)

3.2.1 Disease Extractor

This named entity extractor extracts diseases from any epizootic related document. Each document is split into tokens and each token is tested against the disease gazetteers. Almost 457 diseases with their synonyms, abbreviations and various disease causing viruses are maintained in disease gazetteers.

Examples for disease names: FMD, Foot-and-Mouth Disease, Orthopoxvirus, RVF *etc*.

3.2.2 Location Extractor

A Location Extractor has been developed for extracting the location names from raw text. It uses Stanford NER [9] and NGA GEOnet Names Server (GNS) [14] for location disambiguation and for retrieving latitude and longitude. GNS is the official repository of standard spellings of all foreign place names, sanctioned by the United States Board on geographic names. The database also maintains variant spellings for cross reference purpose.

3.2.3 Species Extractor

The Species Extractor recognizes the species from any epizootic domain document. Every document is chunked into tokens and each token is tested by using pattern matching against the species gazetteers. Species gazetteer used in this system is a stemmed dictionary of animal names from Wikipedia.

3.2.4 Date/Time Extractor

A Date Extractor, adapted from KSNES (***KDD- Service based Numerical Entity Searcher***), is a system that should be able to identify numerical and temporal information from raw text [33]. It extracts dates using a set of regular expressions. It is able to recognize the dates irrespective of the format used.

3.2.5 Confirmation Extractor

The main purpose of confirmation status extractor is to differentiate between outbreak related and non-outbreak related sentences and to generate the templates as shown in Figure 3-2. Typically, a sentence that is outbreak related will have a disease name and supporting verb phrase for reporting an event. Without this verb phrase, every possible combination *i.e.*, <disease name, location >, <disease name, species> *etc.*, will be an event. For example, a sentence such as "FMD Disease is mostly seen in pigs" is a factual sentence, but not an outbreak sentence. So, Confirmation Extractor provides a means to find the difference between the outbreak and non outbreak related sentences.

Confirmation status extractor does the pattern matching against the verb in the raw text using its own gazetteer which has separate lists of words for two distinct groups, confirmed and suspected within confirmation status. The sentences that have words such as "outbreak", "strikes" *etc.*, would fall in confirmed events category and the sentences that have got terms such as "catch", "threat" *etc.*, would come in suspected events category. These extractor gazetteers have few basic restricted words (*e.g.*. kill, threat, catch) in each category confirmed and suspected that are not extended using any lexical database of English. The initial gazetteer is extended using lexical databases such as *GoogleSets* [10] and *WordNet* [39] by adding various related synonyms for the initial verbs framed in initial set.

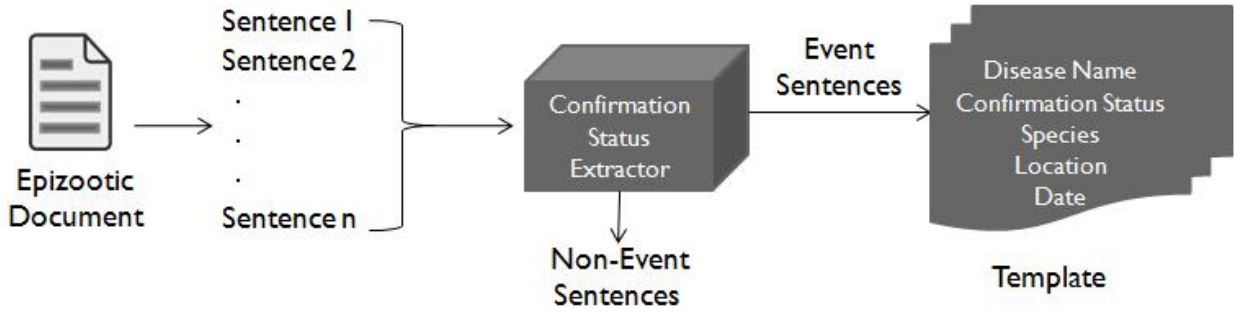


Figure 3-2: *Confirmation Status Extractor*

Following is the Table 3-1 that gives statistics about the number of verbs and verb phrases in each category (confirmed and suspected) from three different gazetteers.

Status	Confirmed	Suspected
Initial Set	8	8
WordNet	57	47
GoogleSets	68	57

Table 3-1: *Statistical data from three different gazetteers of confirmation status*

For example, "OUTBREAK of the highly contagious foot-and-mouth disease is discovered among dairy cows in east China near Shanghai on August 14th, 2002" is classified as a confirmed event due to occurrence of the confirmed word "Outbreak" and "FMD threat in Taiwan" is classified as suspected because of the verb threat that is from suspected list.

3.2.5.1 Gazetteer Construction

Gazetteer construction is the most important step in the development of confirmation extractor. While developing this dictionary, several different kinds of animal disease domain articles, news, blogs, were referred and came up with the initial list of words that could fit in the different categories of the confirmation status – confirmed and suspected. The words chosen were verbs and noun phrases. After the initial set is developed, gazetteer is extended by adding the synonyms and related meanings using WordNet and GoogleSets and following Table 3-2 lists all the verbs/ noun phrases in the confirmation status gazetteer.

Initial	**Confirmed**	**confirmed, infected, strike, outbreak, tested, positive, detected, diagnosed, diseased**
	Suspected	spread, catch, threat, danger, suspect, risk of infection, subject, warning
Google Sets	Confirmed	confirm, open, close, select, search, review, buy, alert, prompt, reserve, set timeout, quote, clear timeout, fetch, write, set interval, prepare, delete, print, describe, execute, scroll to, scroll by, move to, add, clear, save, back, infect, the exchange of, strike, ball, strike looking , fouled off the pitch, outbreak, tested positive, tested negative, those affected, at risk, detected, request, reset, not enabled, not detected, diagnosed, facilitated, represented, assessed, clarified, collected, advocated,

		assisted, guided, supported, demonstrated, referred, familiarized, educated, provided, arranged, ensured, diseased, remedy, truth, given, ill, dead, morbid
	Suspected	spread, catch, try, finally, throw, threat, risk, hazard, danger, warning, caution, risk, hazard, peril, note, jeopardy, para, flammable, threat, poison, endanger, corrosive, attention, notice, endangerment, chance, menace, imperil, tip, hint, error, important, section, hazardous, safety, imperilment, jeopardize, threaten, combustible, signs, explosive, gamble, pitfall, fatal, emergency, death, caustic, toxic, harmful, chapter, warned, mobile, phone, blocked, suspect, risk of infection, susceptibility to infection
WordNet	Confirmed	confirmed, infected, strikes, claims, outbreak, outbreaks, tested positive, detected, diagnosed, corroborate, sick, infested, disease-ridden, plague-ridden, affected, influenced, morbidity, morbid, reported, emitting, emitted, virus production, recovered, removed, disposed, euthanasia administered, culled, IP cull (Infected Premise culls), DC cull (dangerous/ direct contact) cull, died, dead, cleaned, death, mortality, CP cull (contiguous premise), buried, end, eradicated, eradicate, electrocution, cull, withdrawn, isolate, isolated, retrieve, recover, ured, slaughter
	Suspected	spread, catching, threat, danger, risk of infection, warning, predict, Alert, strike again, scares, Re-emergence, surmise, expect, expected, believed, believe, venture, be taken in, fall for, give in, impressionable, influenced, liable, movable, predisposed, prone, receptive, responsive, sensile, sensitive, subject, susceptive, susceptible, vulnerable, head count, herd count, agricultural census, density, suspected, population, healthy, risk, exposed

Table 3-2: *Verbs and Noun Phrases from different confirmation status gazetteers*

4 Methodology

This chapter explains the methodology used for the proposed system. The explanation regarding the different stages in the system is given in section 4.1 and the details about each stage (web crawling, topic classification, searching, event extraction and visualization) are provided in the sub sections.

4.1 Stages of the System

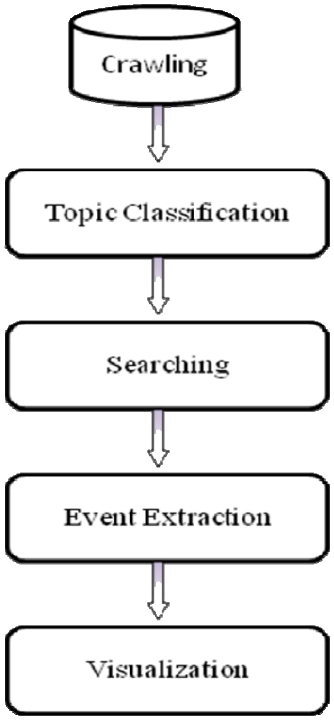


Figure 4-1: *Stages of the Event Extractor System.*

The proposed system shown in Figure 4-1 includes five main stages: web crawling, topic classification, searching, event extraction, and visualization. It is a time consuming and tedious process to extract the entities and the relevant relationships from the unstructured data. For extracting entities, a system first needs good collection of relevant domain set, else entire pre-processing tasks before entity extraction may go waste. The first three stages of event extractor

system - crawling, topic classification, and searching was dealt by Svitlana Volkova, member of Knowledge Discovery in Databases (KDD) group at Kansas State University.

4.1.1 Web Crawler

If one searches for any topic on search engines such as Google, Yahoo, one may end up with million pages. It is a fact that only few documents contain relevant information. It is utmost important to filter out useless documents because not every document that has matched topic of interest can be used for entity extraction. To zoom in on the promising documents, web crawling is done using Heritrix [10], an internet archive's open-source, extensible, web-scale, and archival-quality web crawler. Starting with a set of user-provided seed tuples (ProMED-Mail, DEFRA, OIE, CDC *etc*.) for the target relation, Heritrix retrieves a small set of documents likely to be useful for the further extraction process.

4.1.2 Topic Classification

This stage helps to identify the documents with disease-related topics and retains the relevant ones for further processing. This stage uses Naïve Bayes as the classification algorithm from MALLET [35] for classifying the documents as relevant and non-relevant. MALLET is a Java based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. Mallet provides tools for document classification which can convert textual documents to features using algorithms such as Naïve Bayes, Maximum Entropy, and decision trees, and it can evaluate the classifier performance using several commonly used metrics.

4.1.3 Searching

After classifying the crawled data on the basis of its relevance, the narrowed data collection is passed on to searching module for further processing. The powerful search tool Lucene [19] can perform complicated searches on several gigabytes of document data after indexing that data. Lucene has a powerful feature for processing the data that makes use of metadata in addition to document data. This data is typically about title and author information of the documents being indexed. Lucene uses its sophisticated ranking model while searching for best matched documents for the given query. A ranking model uses factors such as the frequency of a particular query term with individual documents and the frequency of the term in the total

population of documents while calculating the rank of the query term. For the present system, the documents are searched using queries that have animal diseases name and/or location. In this way, one can improve to get the focused crawled set.

4.1.4 Event Extraction

The narrowed domain set is then given for the information extraction module for event extraction. This module extracts disease name, location, species, date and the confirmation status from the raw text. More details on this module are given later in Section 4.2. Not every sentence contains all the entities but it would be a combination of these.

Example: *WHO officials said 160 cases of RVF were reported in Yemen in June, 2001*.

Event extracted is {*RVF, Confirmed, Yemen, 2001*}

4.1.5 Visualization

The event extraction task is a challenging issue because it deals with extracting the meaningful entities. So, it is very helpful and meaningful for plotting the events on *Google Maps*. More details regarding the visualization will be presented in the section 4.3.

4.2 Event Extraction Methodology

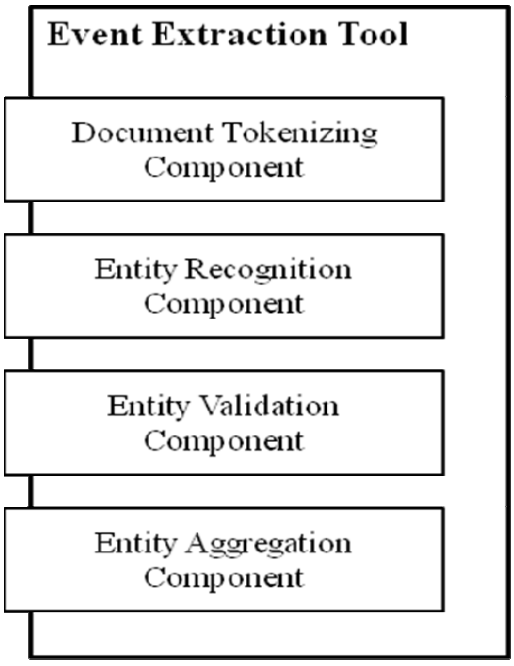


Figure 4-2: *Components in Event Extraction Process*

Each document chosen for event extraction is epizootic related, that is been crawled and classified as relevant from the stages of crawling, topic classification and searching. Raw text from document is now processed using entity extractor tool and the atomic entities such as disease-name, location, confirmation status, date and species classified into different categories as shown below.

- Disease names – {*"FMD", "RVF," arenavirus"}*
- Species with quantities – {"*pigs", "100 horses", "armadillo*", "*5 million sheep*"}
- Locations – {"*Netherlands", "UK*"}
- Date – {"*Today*", "*Sept 22, 1985*", "*Monday*"}
- Confirmation Status – {"*Outbreak", "hits", "catch*"}

After extracting the individual entities, event extractor tries to find the relationship among the entities for tracing out the meaning attributes from the raw text.

Components in the event extractor system are shown in Figure 4-2. In the first stage of processing, each document is chunked into sentences using tokenizer component. Each sentence that is chunked is given to the event extraction tools - Disease Extractor, Location Extractor, Species Extractor, Confirmation Extractor and Date Extractor for extracting disease names, location, species, confirmation status and date entities from each sentence. The following step is important in differentiating the event related sentences from the non event related sentences. For "e.g" "FMD Disease is mostly seen in pigs" a factual informational sentence can be easily filtered from event related sentences using this step. After the entity set is extracted from the sentence, it is evaluated against the rules that are framed for the system which are explained in 4.2.3. If the extracted entity set contains a verb that could be categorized into one of groups of confirmation extractor along with disease-name and any of the <species, date, location>, then the entity set is treated as the valid one or else the entity set is ignored. If the entity set is evaluated as a valid one, related entities sets from the same sentence are aggregated. Further detailed operations are shown in the architecture.

4.2.1 System Architecture

Following Figure 4-3 explains how a document is parsed using event extractor tool and how each event extractor component is associated with each processing step is shown in the figure.

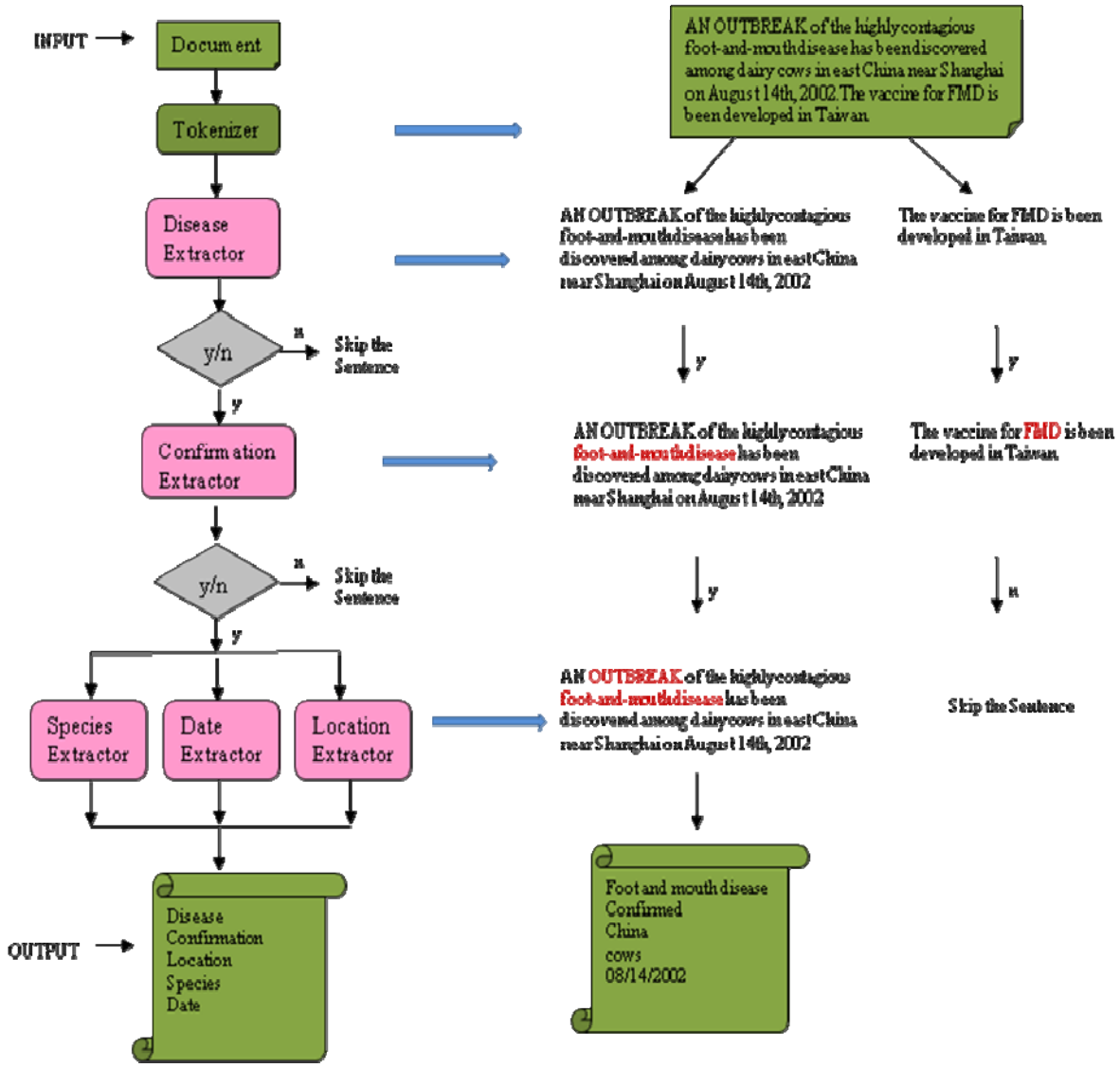


Figure 4-3: *Event Extraction Architecture*

i. Any web document from disease domain is taken as an input for event extraction.

ii. The whole document is chopped into sentences (tokens) using tokenizer module.

iii. Each token is processed iteratively with the next following steps.

iv. Each token is given to the disease extractor module to tag a disease in case of its presence. If disease is tagged within the token, it is sent to the confirmation status extractor module or else the token is skipped.

v. Confirmation status extractor will tag confirmation status in case of its presence or else the token is skipped off.

vi. If both of the mandatory terms *i.e.*, <*disease-name, confirmation status*> are tagged, then only the token is processed with the next steps for non mandatory terms.

vii. Non mandatory terms < *species, location, date*> are tagged using species, location and date extractor.

viii. The output file (template) is written only when at least one non mandatory term and all mandatory terms are found in the token (sentence).

4.2.2 Event Extraction Algorithm

Following is the algorithm developed for the event extraction tool. In this algorithm, while extracting the entity set <*disease, date, location, species, confirmation status*>, if we find that date is not stated within the sentence, article reporting date is extracted from the document for maintaining temporal information.

Input: Set of web documents D
Output: Set of extracted events $e_k \in$ E for each document $d_j \in$ D

```
EVENT EXTRACTOR()
foreach document d_j ∈ D do
    S = TokenizeToSentences(d_j);
    foreach sentence s_i ∈ S in d_j do
        status = ExtractConfirmationStatus(s_i);
        if status ≠ ∅
            disease = ExtractDiseaseEntities(s_i);
            date = ExtractDateEntities(s_i);
            if date = ∅
                date = ReportedDate(d_j);
            location = ExtractLocationEntities(s_i);
            species = ExtractSpeciesEntites(s_i);
        else
            skip sentence s_i;
        if (status ≠ ∅ AND (disease ≠ ∅ OR location ≠ ∅ OR species ≠ ∅))
            e_i = GenerateTuples (disease; date; location; species; status) ;
    foreach entity e_i ∈ E in d_j
        A_k = AggregateTuples(e_i);
```

4.2.3 Features and Rules

Complete feature set is *{Disease Name, Confirmation Status, Location, Species, and Date}*

- Disease Name would give a picture of disease name that is in context.
- Confirmation Status would tell about whether the disease is confirmed or not.
- Place, where the disease has spread to can be seen through Location.
- Specimens that are affected with the disease can be seen through Species.
- When question about the disease can be answered through Date.

The disease name and confirmation status attributes are mandatory among all attributes. Rests of the attributes <location, species, and date>are non-mandatory ones. An event will have all mandatory attributes and at least one of the non-mandatory attributes.

Examples of Feature set:

- <disease name, confirmation status, location> Outbreak [verb] of FMD[dis] in USA[loc].
- <disease name, confirmation status, date> Outbreak[verb] of FMD[dis] in Nov, 2008[date]
- <disease name, confirmation status, species> Cows [species] are tested positive [verb] for RVF [dis]
- <disease name, confirmation status,loc1,loc2,…> RVF [dis] broke out[verb] in China[loc], India[loc].

4.3 Visualization Map

For supporting visualization map, *Google Maps* API (*Google Maps* API) [11] is used for plotting the extracted events on the visualization map. Different technologies used for developing the visualization module are as follows: PHP for developing the visualization module, JSON (JSON-JavaScript Object Notation) [13] for representing the events, and JQUERY, MYSQL for maintaining the database of events. We first introduce these different technologies and then look at the web application developed.

Google Maps API: *Google Maps* has a wide array of APIs that let you embed the robust functionality and everyday usefulness of maps into an application, and overlay data on top of them. Google Map API is one of those intelligent bits of Google technology that can help to add any relevant data that is useful and to customize the look and feel of the map to fit any personalized style. ClusterMarker [42] is a marker manager for use with maps powered by the *Google Maps* API and it can be used for clustering the points plotted. ClusterMarker detects

any group(s) of two or more markers whose icons visually intersect when displayed and the markers that are close in latitudinal and longitudinal values fall in the same group. Each group of intersecting markers is then replaced with a single cluster marker. The cluster marker, when clicked, simply centers and zoom the map in on the markers whose icons intersect. MapIconMarker [43] is a freely available JavaScript API used in conjunction with the *Google Maps* to display icons on the map. We can use this API to produce markers of different size and color and many other properties. The dynamic markers which represent different events from our data are generated using this API.

For plotting the events on to the map, the events are converted into JSON stands for JavaScript Object Notation. It is a lightweight text-based open standard designed for human-readable data interchange. It is derived from the JavaScript programming language for representing simple data structures and associative arrays, called objects.

4.3.1 Visualization Module

The web application was developed using PHP and MYSQL database and deployed on a XAMP web server. The communication between the database and the application is done in PHP. Using the structured data, obtained from the rule based extractor and stored in the database, the points are plotted on a map depicting the locations of the disease spread. Thus the structured data is visualized geographically on Google Map. The map depicts the places where the disease outbreak was reported by using markers. A marker at a location depicts a single event processed. Also a set of markers are grouped as a cluster by using Cluster Marker of *Google Maps*.

Now in the main page, the JSON result set is decoded and assigned as markers. Using the values of latitude and longitude information obtained for a specific event, the marker for that event is represented on the map. The markers in the Map are plotted using two different colors, based on the type of event, either suspected or confirmed. The color convention used is Orange for suspected events and Gray for confirmed events.

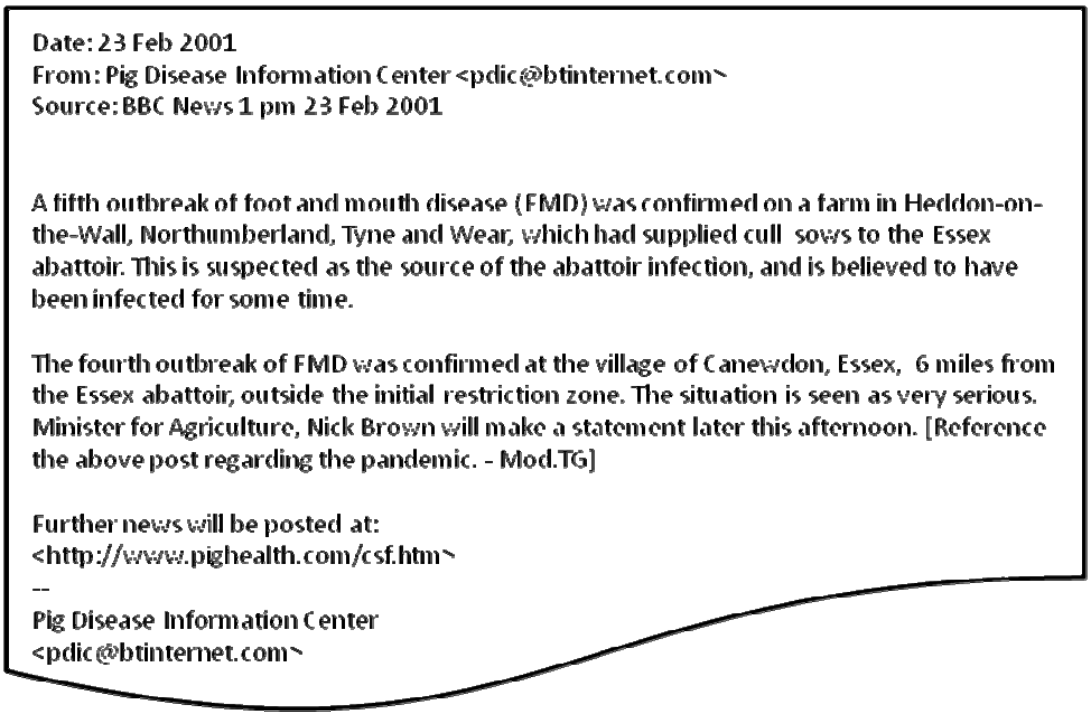
Clusters on the map are depicted using a green arrow and once we click on them, we can further see inner clusters or the markers part of that cluster.

5 Experiment Setup and Evaluation Measures

In this chapter, we will discuss in detail about the dataset used, experiments conducted for extracting the results and the different confirmation status gazetteers which have been developed. Some experiments were conducted to evaluate the efficiency of the extractor in extracting the events from animal disease domain. In Section 5.1, we discuss about the data set been used. In Section 5.2, we explore about the different confirmation status gazetteers that have been developed for the event extractor followed by the evaluation metrics in Section 5.3.

5.1 Data Set

To perform the evaluation, we used the data set of 250 documents related to two diseases: FMD (Foot-and-Mouth Disease) and RVF (Rift Valley Fever). These documents have been collected from the health organizations, animal disease monitor systems such as *ProMED-Mail*, *PULS* Data, news archives, blogs and many more. An example of a document from the dataset is shown below in Figure 5-1.


Date: 23 Feb 2001
From: Pig Disease Information Center <pdic@btinternet.com>
Source: BBC News 1 pm 23 Feb 2001

A fifth outbreak of foot and mouth disease (FMD) was confirmed on a farm in Heddon-on-the-Wall, Northumberland, Tyne and Wear, which had supplied cull sows to the Essex abattoir. This is suspected as the source of the abattoir infection, and is believed to have been infected for some time.

The fourth outbreak of FMD was confirmed at the village of Canewdon, Essex, 6 miles from the Essex abattoir, outside the initial restriction zone. The situation is seen as very serious. Minister for Agriculture, Nick Brown will make a statement later this afternoon. [Reference the above post regarding the pandemic. - Mod.TG]

Further news will be posted at:
<http://www.pighealth.com/csf.htm>
--
Pig Disease Information Center
<pdic@btinternet.com>

Figure 5-1: *Example of a document from Data Set*

As shown from the Figure 5-1**,** almost all the documents of the data set have a "reported date" mentioned before detailing about the facts of the disease spread. We can use the reported date as an approximate date closest to the effective outbreak period in the case of date missing from the eventual information. This approximation is necessary as temporal information is important while analyzing the outbreak information. To explain this scenario, consider the following example. In the above figure 8, 23 Feb, 2001 can be taken as the outbreak date as the actual mention of outbreak date is missing from the facts from the document. Thus the dataset was used taking these approximations into considerations.

5.2 Different Confirmation Status Gazetteers

In this section, we evaluate the event extraction methodology by executing the extractor using three different confirmation status gazetteers in stemmed and non stemmed format. As discussed earlier in Section 3.6.1, the three gazetteers- Initial, Google Sets and WordNet are stemmed using Porter Stemmer [34]. If the stemmed versions are used as gazetteers, we need not maintain all the occurrences of a verb in the dictionaries while if we use non stemmed versions, we have to populate the gazetteer with each and every occurrence of the verb and this could make the working dictionary verbose.

For example, for the word – '*Suspected',* we need to maintain words such as 'Susceptible', 'Suspected', 'Suspecting' and all other possible words for non stemmed versions while we just need to maintain 'Suspect' for stemmed versions. One may even end up not populating all possible words for a verb in the case of non stemmed.

5.3 Evaluation Metrics

The result of the event extractor is in the form of the template, providing information regarding disease spread that can be compared to a summary of the document. Summary evaluation is a challenging task. To evaluate any summary, it has to be compared with a model summary that has covered all important facts from the original document. There would be different opinions for different people about the information that needs to summarize. Thus, it is very difficult to get such kind of model summaries for comparison. In this scenario, we would be interested in collecting as many summaries as possible and collect all common facts of a

document from them. These summaries are collected from different people in order to avoid monotonic opinions about the information that needs to be summarized.

There comes the question of validating a system generated summary with the summaries that have been collected. If the evaluations of the summaries with the model summaries is done manually (human resource), it is a laborious and are not reliable too. For this purpose we would use an automatic evaluation method (PYRAMID) [37] which approximates human judgments.

We would be creating two referential summaries to compare with the system generated summary templates. A set of summaries was written by Karthik Tadepalli and another set by me. The evaluation procedure would take these reference summaries as models and generated templates as inputs. Mostly closed terms (clauses) among summaries would be treated as SCU's. Based on these SCU's, we would be calculating pyramid score for a peer summary based on the closeness of the peer summary with the optimal summary. Now, we would look in detail about the actual implementation and steps that goes into the calculation of the pyramid scores.

5.3.1 Pyramid Method

The pyramid method is based on the fact that summaries from different persons always share overlapping content and this approach is framed to calculate the relatedness of the human written summaries to the model summary. The pyramid model is a manual methodology used for the evaluation of the summaries, and its main purpose was to address the problem of summarizing, that of different humans choosing different content when summarizing documents. The pyramid method addresses this problem by using multiple human summaries to create a template that could be compared against any system generated summary. It also makes note of the frequency of information from the human summaries in order to assign importance to different facts. The approach involves two phases of manual annotation: pyramid construction and annotation of unseen summaries against the pyramid to determine which Summarization Content Units in the pyramid have been expressed in the peer summary. The pyramid evaluation has characteristics of both a precision measure (as the score is a function of the size of the summary) and of a recall measure (as the score is also a function of the weights of the optimal SCUs) [24].

The *DucView Tool* [15] is an annotation tool used for creating pyramids from model summaries, the annotation of peer summaries against an existing pyramid and also for computing pyramid scores.

5.3.1.1 Rationale for Pyramid Approach

The main reason for using a pyramid approach for evaluation for this system, rather than any other evaluation method, is that pyramid end results can stand on their own without additional dependencies on subjective ground truth. The proposed system event tuples are being augmented with new attributes (confirmation status) that previously have not been extracted from text corpora before.

Moreover, the pyramid results can be considered as a measure to validate the event-higher the pyramid score, better the chance that event includes most of the vital information. When we adopt for comparative study between the present system and any other similar system, one may get some spurious results because the system that is used as ground truth may not be perfect. The main goal of the present system is summarization which cannot be compared perfectly with any third party's generated summary. Hence, we rely on the manual summaries which are perfect that are ground truths used in pyramid approach.

5.3.1.2 SCU's Construction

This new approach, the pyramid method, developed by Ani Nenkova and Rebecca Passonneau (Columbia University) [37], is based on Summarization Content Units (SCUS):

- As different people include different information when making a summary, SCUs annotation highlights what people agree on.
- SCUs are sub-sentential content units, not bigger than a clause, taken from a corpus of manually-made summaries.
- An SCU consists of a label -a concise sentence that states the meaning of the Contributing Units - CU, and many contributors -snippets of text from summaries which are semantically related to the label.

Each SCU has a weight corresponding to the number of summaries it appears in.

Following is the detailed discussion of construction of the Summarization Content Units from the peer summaries developed for one of the document. The Contributing Units that are

common to both summaries are colored in a same color and the Table 5-1 gives the "SCU Label: Contributing Units" pairs from the peer summaries and the contributing units for each pair.

Summary A

Foot-and-Mouth disease outbreaks in China, where the infection spread leading to cows and sheep being slaughtered. FMD was also reported in Argentina.

Summary B

FMD hit China in two provinces, which led to a slaughter of 89 cows and 110 sheep. Argentina has also been hit by FMD.

SCU Label: Contributing Units	Summary A	Summary B
Disease Name: FMD	Foot-and-Mouth disease	FMD
Location: China	China	China in two provinces
Confirmation Status: Confirmed	Outbreaks	hit
Species: Cows and Sheep	cows and sheep	89 cows and 110 sheep

Table 5-1: *SCUs Annotations showing the contributing units from two summaries*

5.3.1.3 Pyramid Construction

For constructing a pyramid, annotators identify contributing units from a pool of summaries written by different persons who have read the same set of documents. For this purpose, two sets of summaries for 200 documents are written by two different annotators and the pyramids are created using DucView Tool. Following Figure 5-2 is the snapshot of the Pyramid Construction using DucView Tool.

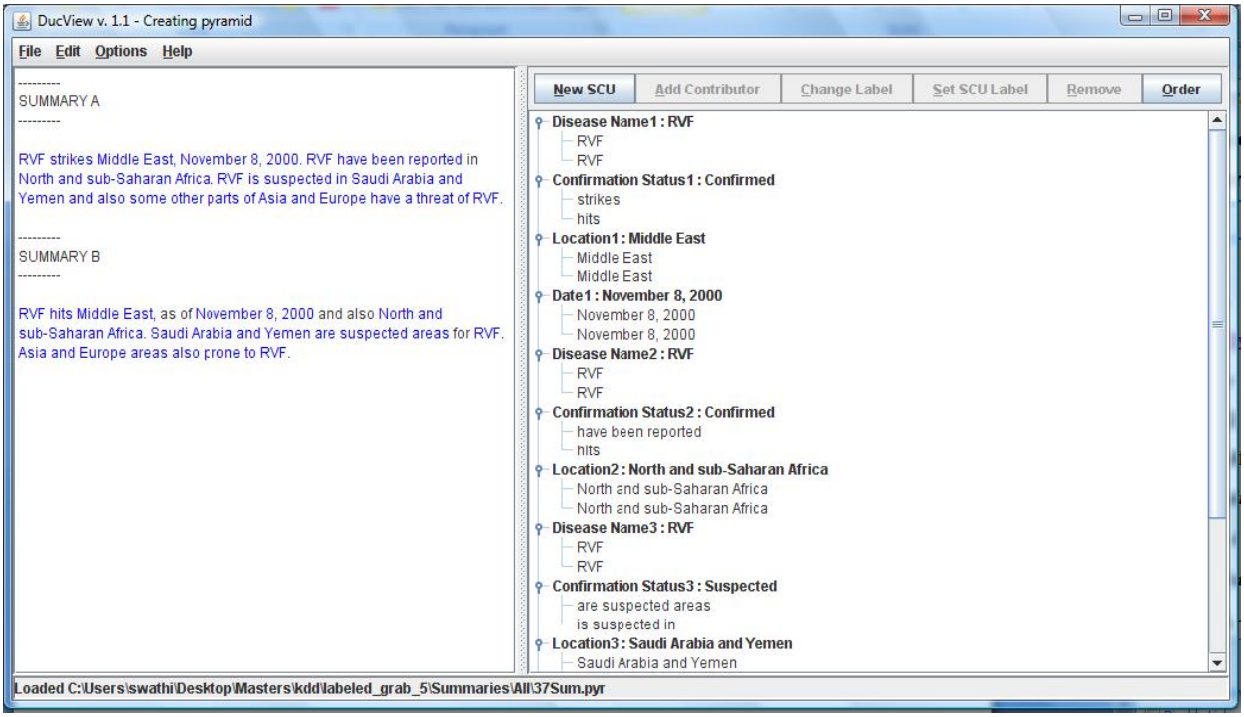


Figure 5-2: *Pyramid Construction*

After deciding on the SCUs (SCU Label: Contributing Units), the matching contributing units from the summaries are added to the corresponding SCUs. This way we can match the CUs with the SCUs.

For Example, the contributing units, *have been reported* and *hits* from Summary A and Summary B respectively in the figure, are added to the matching SCU, *Confirmation Status2: Confirmed*.

The number of tiers in the pyramid is equal to the number of summaries that are been considered. A pyramid of order n will have weights from n to n-1 for each tier. The SCUs that has got the same weight will go to the same tier. Given a pyramid of order n, we can predict the optimal summary content, which should contain all the SCUs from the top tier, and, if length permits, SCUs of the next tiers & so on.

5.3.1.4 Peer Annotation

After identifying all the SCUs from the peer summaries, a pyramid is constructed and this pyramid could be used as model summary, which have covered all important facts that a good

summary should have, to compare against a peer summary for measuring the proximity. Given a model pyramid for a document set, each peer summary is annotated against the corresponding pyramid. Peer annotation is easier than pyramid construction because the set of SCUs is already constructed. Annotators select words in the peer summary that express the same information expressed in an SCU of a pyramid constructed. Also if the peer summary has got information that is repeated more than once, the annotator can reselect the same SCU. If the annotator does not find the matching SCU for the information that is relevant in the peer summary, he will add the information to the non matching SCU list. This way all peer summaries are annotated using its corresponding pyramid. Figure 5-3, shows how the peer summary is annotated using its Pyramid.

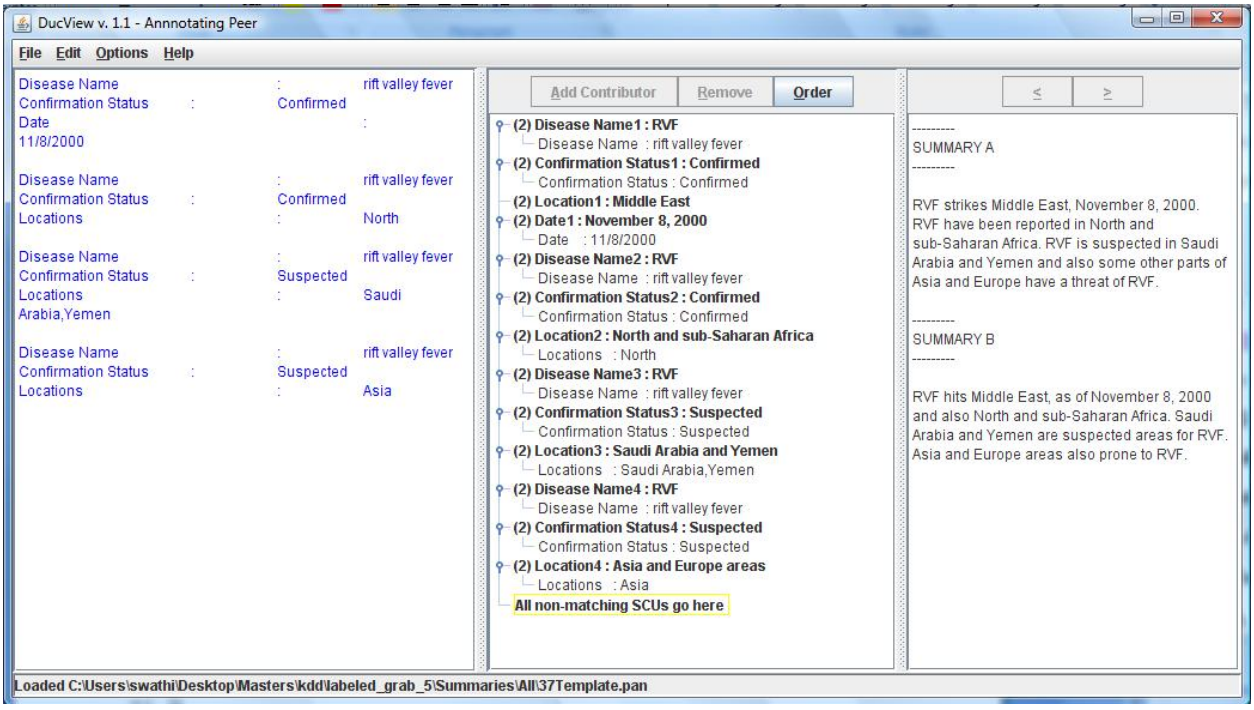


Figure 5-3: *Peer Annotation*

5.3.1.5 Pyramid Score Computation

Two scores are computed from the peer annotations namely OBS and MAX. OBS is the sum of the weights of the SCUs found in the peer. MAX is the sum of the weights of SCUs found in pyramid. If the number of SCUs of a given weight i that occur in a summary is O_i , the sum of the weights of all the SCUS in a summary is given by

$$OBS = \sum_{i=1}^{n} i \times O_i$$

The number of SCUs used in computing MAX_o is same the number used to compute OBS. If we designate the pyramid tiers by their weight (T_i), MAX_o is given by

$$MAX_o = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$$

$$\text{where} \quad j = \max i (\sum_{t=i}^{n} |Tt| \geq X)$$

The pyramid score is given by OBS/MAX_o, indicates the ratio between the sum of the weights of its SCUs in peer summary (OBS) and sum of the weights of an optimal summary (MAX_o). It ranges from 0 to 1, with higher scores indicating that relatively more of the content is as highly weighted as possible.

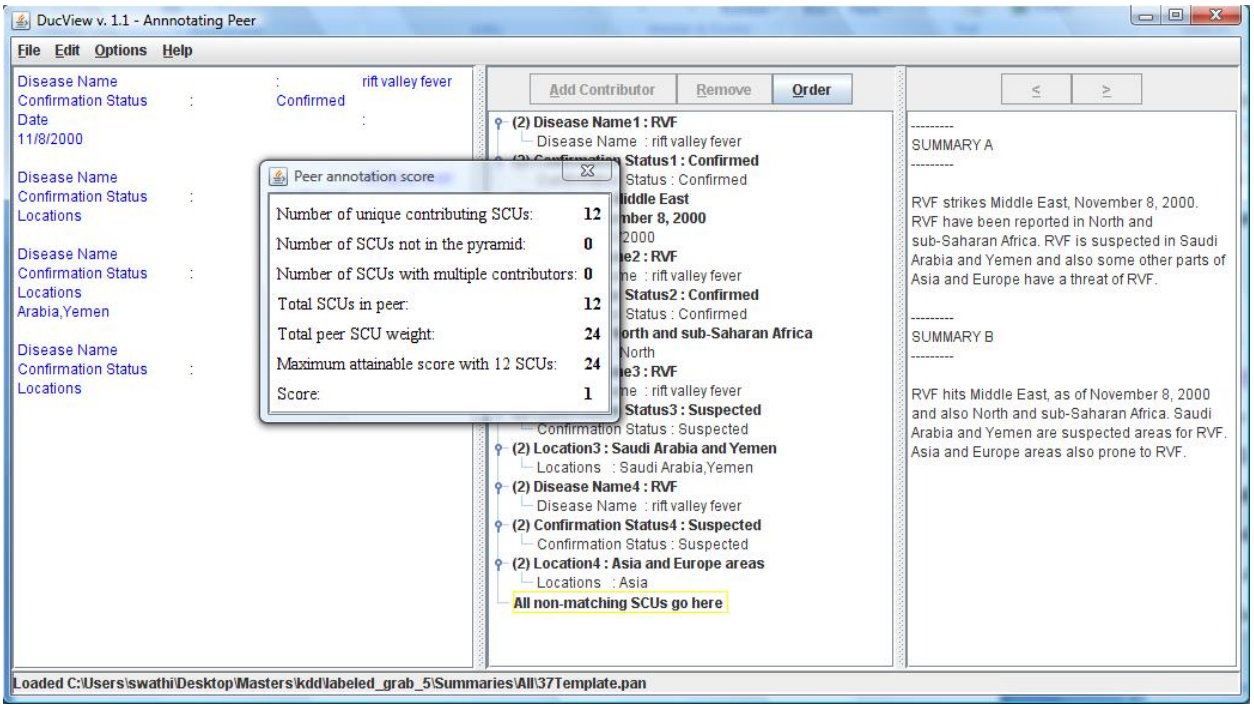


Figure 5-4: *Pyramid Score*

Figure 5-4 presents the screen shot of the peer annotation score from the DucView tool. In addition to the pyramid score, it also reports the information regarding the number of unique contributing SCUs, number of SCUs not in the pyramid, number of SCUs with multiple contributors, total SCUs in peer, total peer SCU weight and maximum attainable score with the total number of SCUs.

6 Results and Evaluations

Having seen the evaluation metrics developed and used in event extraction, in this chapter, we will see the results of the experiments conducted and also the evaluation of the results obtained. In section 5.1, we would evaluate the event extractor results with three different types of confirmation status gazetteers. In section 5.2, we will look into the screen shots of visualization of the events on the map.

6.1 Event Extractor Behavior with different Confirmation Status Gazetteers

6.1.1 Pyramid Score Ranges

Pyramid scores resulting from events with three different gazetteers - the initial set, *GoogleSets*, and *WordNet* - are broken down in three different ranges 0.0 - 0.4, 0.4 - 0.7, 0.7 - 1. The following experiment, finding the trend in percentage of loss in events as a function of pyramid score, is conducted to finding the splitting range within pyramid scores (0 to 1) with respect to the similar behavior observed.The experimental results are shown in Figure 6-1, Figure 6-2, Figure 6-3 for *WordNet*, *GoogleSets*, and the initial set, respectively.

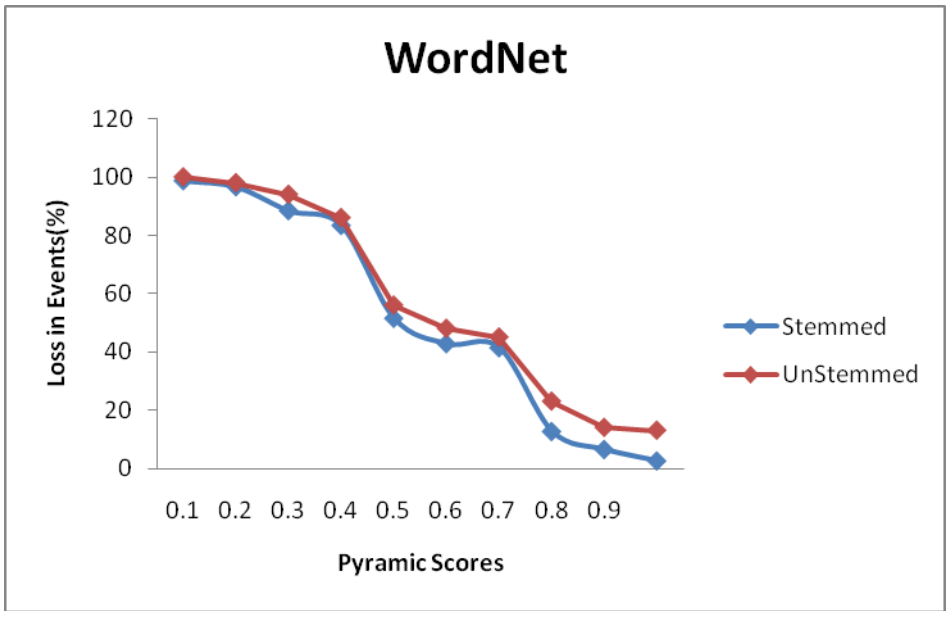


Figure 6-1: *Loss in Events (%) vs. Pyramid Scores: WordNet Data.*

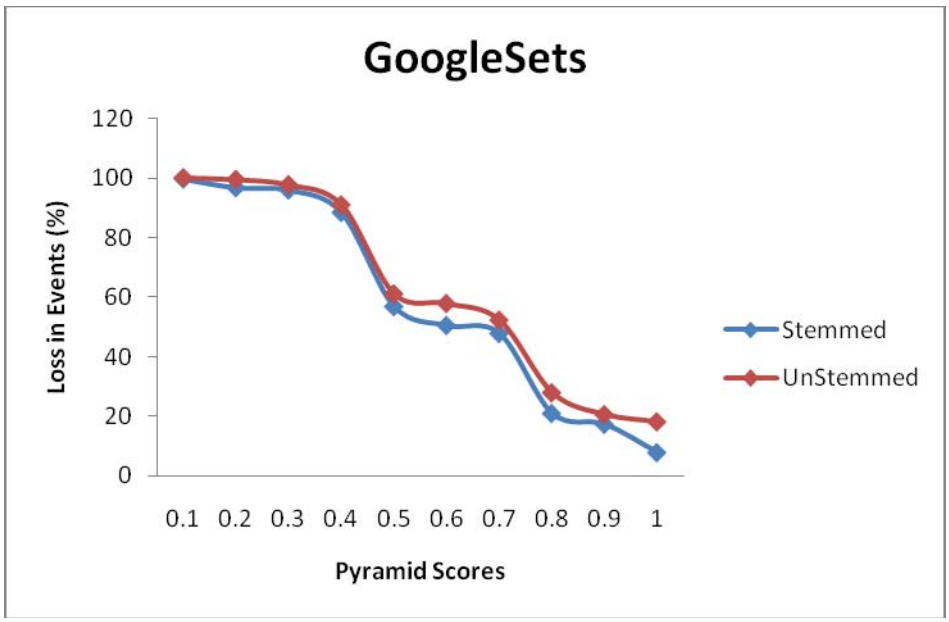


Figure 6-2: *Loss in Events (%) vs. Pyramid Scores: GoogleSets Data.*

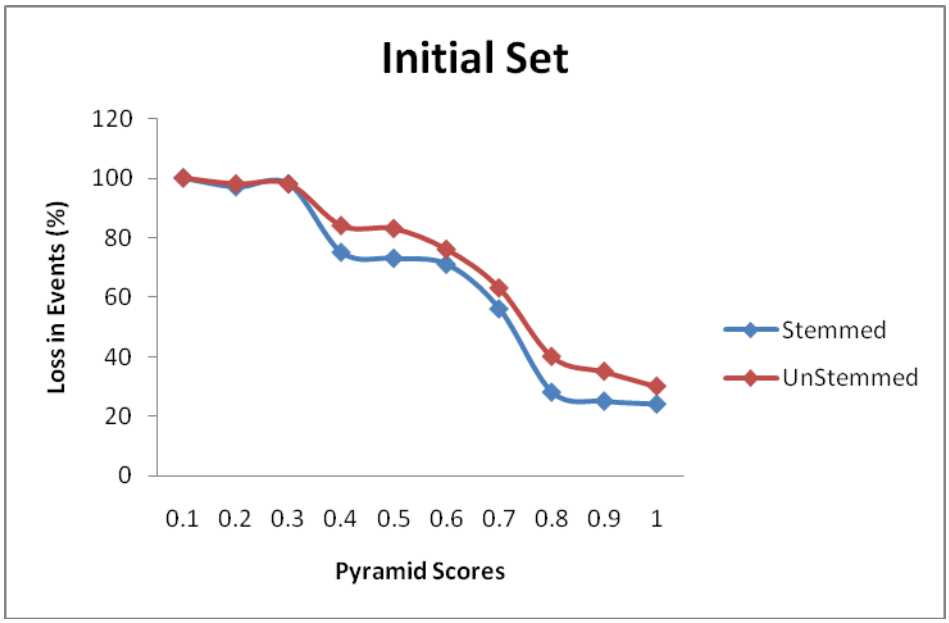


Figure 6-3: *Loss in Events (%) vs. Pyramid Scores: Initial Set.*

From the following experiments conducted against the three confirmation gazetteers, similar pattern behavior is observed within the pyramid range of 0.0 - 0.4, 0.4 - 0.7, 0.7 - 1 as there are major drop offs observed after 0.4 and 0.7.

6.1.2 Experimental Results

To evaluate the event extractor, it is run using three different gazetteers – Initial gazetteers, Google Sets gazetteers, and the WordNet gazetteers.

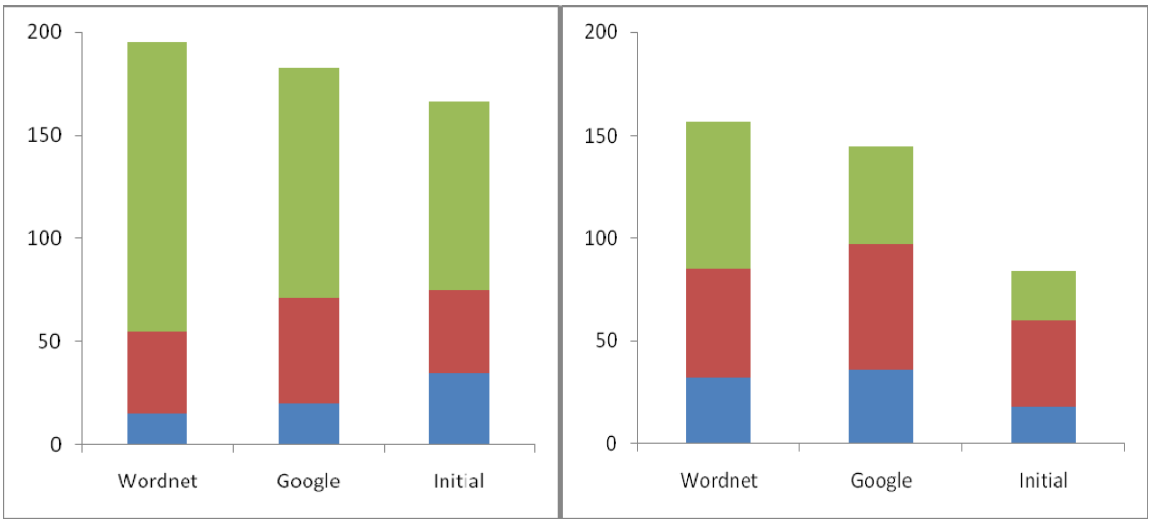


Figure 6-4: *Pyramid Scores with Stemmed and Non-Stemmed Gazetteers for Initial, GoogleSets and WordNet.*

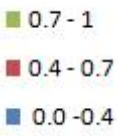
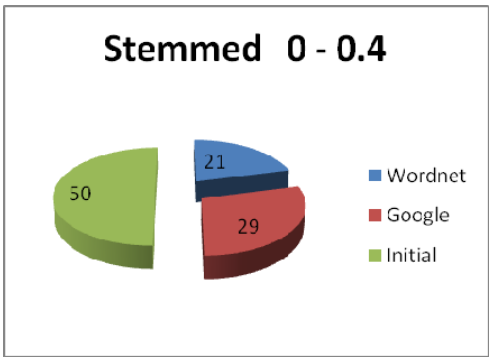


Figure 6-5: *Pyramid Ranges Distribution*

Graphs showing pyramid scores using three gazetteers – initial set, GoogleSets, WordNet are shown in the figures Figure 6-4. The pyramid scores ranging 0.0 – 0.4 is colored in blue, 0.4 – 0.7 is colored in red and 0.7 – 1 is colored in green. Figure 6-4 depicts the pyramid scores with stemmed and non-stemmed gazetteers for Initial, GoogleSets and WordNet. From the results, we can clearly say that stemmed results are much better compared to that of non stemmed results. Higher the pyramid score, higher the relevance of the event extracted. From Figure 6-4, we can clearly say that WordNet gazetteers are much better than GoogleSets and initial gazetteers by looking at the pyramid scores once again.

Following are the pie-diagrams that give the picture of how the pyramid scores of 0-0.4, 0.4-0.7 and 0.7-1 are organized in stemmed and non stemmed gazetteers.



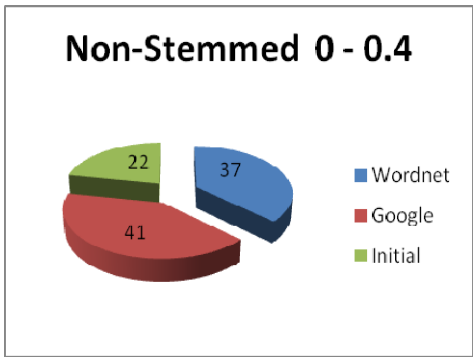
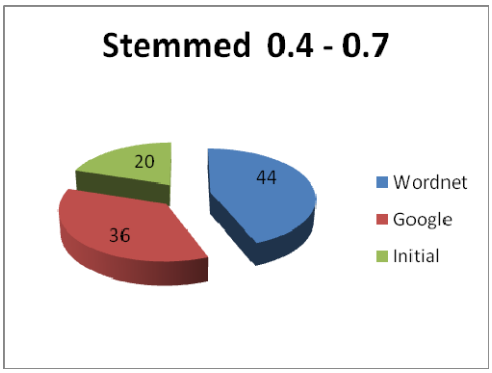


Figure 6-6: *Pie Diagrams for Stemmed Gazetteers and Non Stemmed Gazetteers with pyramid score range (0 – 0.4), showing the highest percentage of coverage in stemmed initial gazetteer.*



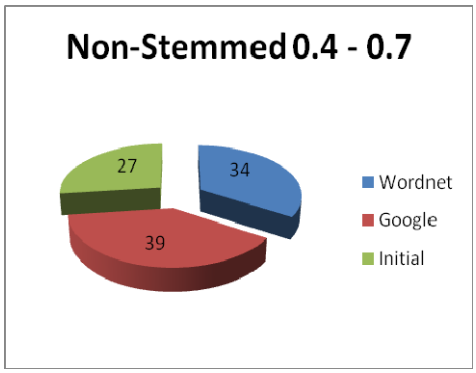
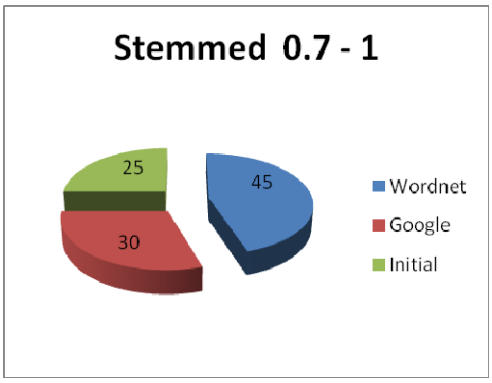


Figure 6-7: *Pie Diagrams for Stemmed Gazetteers and Non Stemmed Gazetteers with pyramid score range (0.4 – 0.7), showing the highest percentage of coverage in stemmed GoogleSets gazetteer.*



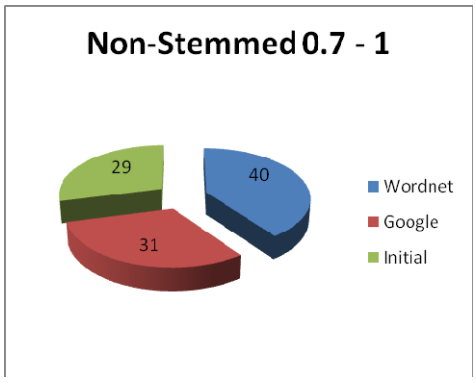


Figure 6-8: *Pie Diagrams for Stemmed Gazetteers and Non Stemmed Gazetteers with pyramid score range (0.7 – 1), showing the highest percentage of coverage in stemmed WordNet gazetteer.*

From all these figures Figure 6-6, Figure 6-7, Figure 6-8, we can conclude that stemmed gazetteers are comparatively better than non stemmed gazetteers. Higher the pyramid scores, higher is the efficiency in the event extraction. Hence, from the results one can say WordNet gazetteer is better in extracting the events having pyramid scores 0.7-1.

6.2 Visualization Map

In this section we will see the various features that are available in visualization map. With the set of events from the database, using the longitude and latitude values for those events, the events are plotted on a map.

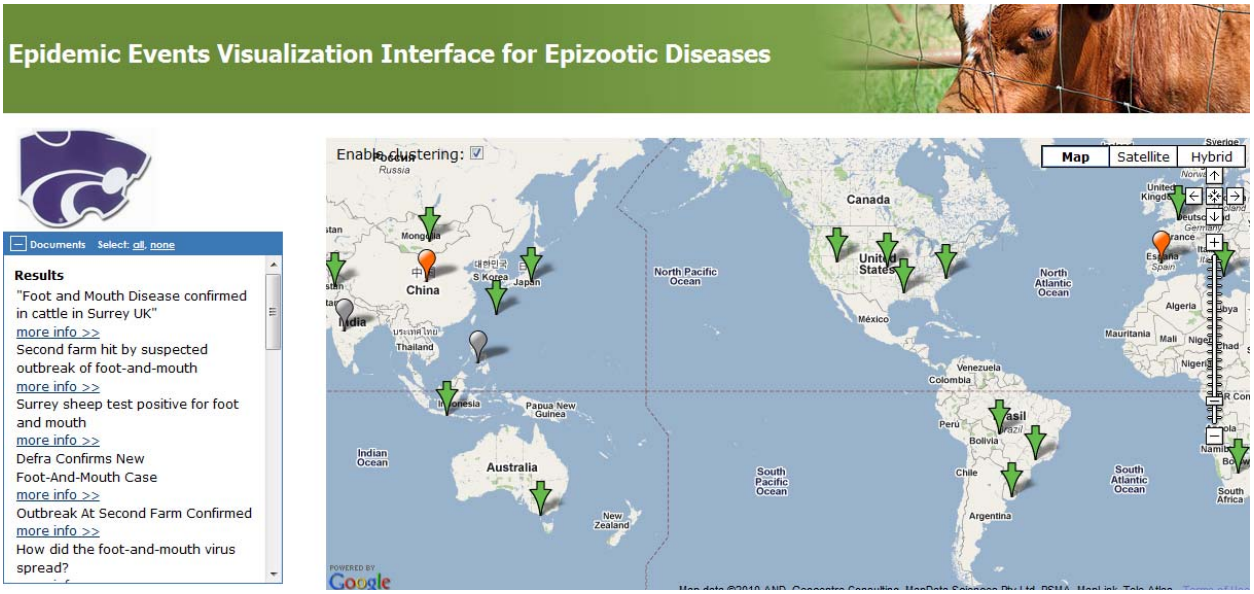


Figure 6-9: *Visualization Map*

Figure 6-9 shows the visualization map that shows the cluster markers (green anchored symbols) for the events extracted from the rule based extractor. Results on the left side show the description of the events extracted. And the events that are not clustered are shown using a bubble marker. There is a check box that can enable and disable the clustering for the events plotted.

Figure 6-10: *Disable Clustering on Visualization Map*

Figure 6-10 shows the visualization map after the clustering is disabled. And it shows only markers and it is able to differentiate the confirmed and suspected events in different colors such as confirmed events are represented in grey color and suspected events in orange color.

Figure 6-11: *Map showing the number of markers in a cluster*

Figure 6-11 shows the snapshot of the map that states the number of markers that are grouped within a single cluster. One can click on the cluster marker to see in the markers that are clustered. The Figure 6-12 below shows the number of markers that are residing inside the cluster. This way one can manage the huge number of markers to be displayed on the map. Map takes more time for populating the huge number of markers rather than the cluster hiding the markers.

Figure 6-12: *Map showing the markers within the cluster*

7 Summary

The system that has been developed previously [40] is very similar to the present one, but differs in a few key features. In the present system, the species extractor is able to retrieve the quantities along with the specimens, for example “1000 pigs” or “145 goats”, which is different from the previous system. The confirmation status extractor gazetteers word lists were reviewed with new words for the present system. The rule of keeping track of published date of the article has been taken into consideration for filling out the default dates within the tuple set in case of the exact date is missing in case of the present system. The visualization module is developed for the present system.

We have provided event extraction results in the form tuples, in animal epizootic domain, which demonstrates the efficiency in ontology based approach and sentence based event extraction approach. There were many challenges faced during the development of this system such as the problem of extracting meaningful entities from the unstructured data, problem in mining the epizootic domain related articles while web crawling.

Confirmation status – a verb that describes the severity of the disease is helpful while filtering out the non-event-related sentences. A rule of entity set- “{confirmation status verb, disease name, {[species, date, location]}}”, of which {confirmation status verb, disease name} are non optional entities and [species, date, location] are optional entities, requiring either one of them was formulated. This rule was applied for the sentence level event extraction,

Extracted tuples are evaluated using pyramid approach and the *DUCView* tool is used for generating the pyramid scores that can be used for measuring the accuracy of the proposed system. To increase the efficiency of the extractor, confirmation status gazetteers are extended using high structured lexical database using WordNet and GoogleSets.

The comparative study shows that WordNet results are relatively better than GoogleSets. When compared with other mining systems such as *BioCaster*, ProMedMail, *HealthMap*, *MedISys* our present system can provide the visualization support for past outbreaks, can also show clusters on visualization map. The developed system is totally automated and it can extract more descriptive information such as confirmation status, species, date along with disease-name and location.

7.1 Contributions

The first three steps involved in the event extractor system *i.e.,* web crawling, topic classification and searching were mainly dealt with by Svitlana Volkova and the rest of the steps *i.e.,* event extractor and visualization module were my individual contribution.

The contributions made in the development of the system through my research work were mainly for three subtasks – development of confirmation extractor, development of efficient event extractor and finally visualization system.

Development of confirmation extractor involves developing the gazetteers by gathering the verbs and noun phrases inherited from the documents set in epizootic domain and categorizing them under two groups of confirmation status extractor confirmed/suspected. Gazetteer construction involves reviewing the massive collection of epizootic documents. Event extractor involves writing many stringent rules for recognizing the relationship among the individual entities recognized within the unstructured data. Finally, development of visualization module that takes extracted events as inputs and displays them on *Google Maps* and visualization map must be able to show clusters on the visualization map.

7.2 Limitations and Future Work

- Our system is able to process only the documents that are in English. Our future work includes processing any document irrespective of its language.
- Our system does not resolve the problem of co-reference resolution. Co-reference resolution is a task of determining which noun phrases in a text or dialogue refer to the same real-world entity that has been previously introduced in the text.
- As part of future work we intend to apply a deeper syntactic analysis of the sentence and part-of-speech tagging in addition to the list of verbs that we used.

References

[1] Office International des Épizooties (OIE). (2010). *Handistatus II Prototype.* November 17, 2010. Retrieved from http://www.oie.int/hs2/report.asp?lang=en

[2] Food and Agriculture Organization of the United Nations (FAO). (2010). *EMPRES - Global Animal Health Information System.* Retrieved from http://empres-i.fao.org/empres-i/home

[3] Food and Agriculture Organization of the United Nations (FAO). (2010). *Global Livestock Production and Health Atlas*. Retrieved from: http://kids.fao.org/glipha/

[4] Food and Agriculture Organization of the United Nations (FAO). (2010). *Key Indicator Data System.* Retrieved from: http://kids.fao.org

[5] United States Geological Survey (USGS). (2010). *USGS Home Page.* Retrieved from: http://www.usgs.gov

[6] The National Biological Information Infrastructure (NBII) Wildlife Disease Information Node (WDIN). (2010). *Wildlife Disease News Digest.* Retrieved from: http://wdin.blogspot.com

[7] Du, M., von Etter, P., Huttunen, S.. Vihavainen, A., & Yangarber, R. (2010). *Pattern-based Understanding and Learning System (PULS).* Retrieved from: http://*PULS*.cs.helsinki.fi/medical/

[8] Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., Eysenbach, G., & Brownstein, J. S. (2010). *EpiSPIDER*. Retrieved from: http://www.*EpiSPIDER*.org/index.php

[9] Manning, C., & Jurafsky, D. (2010). *The Stanford Natural Language Processing Group.* Retrieved from: http://nlp.stanford.edu/index.shtml

[10] Google, Inc. (2010). *Google Sets*. Retrieved from: http://labs.google.com/sets

[11] Internet Archive & Nordic. (2010). *Heritrix Web Crawler*. Retrieved from: http://crawler.archive.org

[12] Google, Inc. (2010). *Google Maps Application Programmer Interface.* Retrieved from: http://code.google.com/apis/maps/index.html

[13] JavaScript Object Notation (JSON) Working Group. (2010). *JavaScript Object Notation (JSON) Home Page.* Retrieved from: http://www.json.org

[14] National Geospatial Intelligence Agency (NGA). (2010). *NGA GEOnet Names Server (GNS).* Retrieved from: http://earth-info.nga.mil/gns/html/

[15] Passoneau, R. (2006). *Pyramid Annotation Guide: DUC 2006.* Retrieved from: http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html#start

[16] International Society for Infectious Diseases. (2010). *ProMED-Mail*. Retrieved from: http://www.promedmail.org/pls/apex/f?p=2400:1000

[17] United Kingdom Department for Environment, Food and Rural Affairs (DEFRA). (2010). *2001 Foot and Mouth Disease data.* Retrieved from: http://ww2.defra.gov.uk

[18] Iowa State University. (2004). *Center for Food Security and Public Home Page.* Retrieved from: http://www.cfsph.iastate.edu/

[19] The Apache Foundation (2005). *Apache Lucene: a high-performance, full-featured text search engine library*. Retrieved from: http://lucene.apache.org

[20] BioComputing Corporation. (2006). *BioPortal System.* Retrieved from: http://biocomputingcorp.com/bphome.html

[21] Brownstein, J. & Freifeld, C. (2010). *HealthMap – Global Health, Local Information.* Retrieved from: http://HealthMap.org/en/

[22] University of California – Davis. (2010). *Foot-and-Mouth Disease (FMD) BioPortal.* Retrieved from: http://fmdbioportal.ucdavis.edu/

[23] Office International des Épizooties (OIE). (2009). *World Animal Health Information Database (WAHID) Interface.* Release date: November 20, 2009. Retrieved from: http://www.oie.int/wahis/public.php?page=home

[24] Harnly, A., Nenkova, A., Passoneau, R., & Rambow, O. (2005). Automation of summary evaluation by the pyramid method. *In Recent Advances in Natural Language Processing (RANLP).* Borovets, Bulgaria, September, 2005.

[25] Benjamin, G. (2003). *Animal-Borne epidemics out of control: threatening the nation's health.* Trust for America's Health.

[26] Centers for Disease Control and Prevention (CDC). (2010). *Centers for Disease Control and Prevention (CDC) home page*. Retrieved from: http://www.cdc.gov/

[27] TheBeefSite.com. (2010). *CME: Regionalisation Important for Animal Trade.* Retrieved from: http://www.dailylivestockreport.com/documents/dlr%204-14-2010.pdf

[28] Collier, N. D. (2008). *BioCaster*: Detecting public health rumors with a Web-based text mining system. (pp. Bioinformatics, 24(24):2940-2941). Oxford University Press.

[29] World Health Organization. (2010). *Global Health Atlas*. Retrieved from: http://apps.who.int/globalatlas/

[30] The National Biological Information Infrastructure (NBII) Wildlife Disease Information Node (WDIN). (2010). *Global Wildlife Disease News Map version 2.0*. Retrieved from: http://wildlifedisease.nbii.gov/wdinNewsDigestMap.jsp

[31] Centers for Disease Control and Prevention (CDC). (2007). *Graphs and Maps from the Summary of Notifiable Diseases - United States 2007*. Retrieved from: http://www.cdc.gov/ncphi/disss/nndss/annsum/2007/07graphs.htm

[32] Pritchett, J., Thilmany, D., & Johnson, K. (2005). Animal Disease Economic Impacts: A Survey of Literature and Typology of Research Approaches. *International Food and Agribusiness Management Association.*

[33] Karumuri, S. (2009). *KDD- Service Based Numerical Entity Search (KSNES).* Master of Software Engineering Project, Kansas State University. Retrieved from: http://people.cis.ksu.edu/~sowji/100jiMSE/

[34] Porter, M. F. (2001). *The Porter Stemming Algorithm*. Retrieved from: http://snowball.tartarus.org

[35] McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from: http://mallet.cs.umass.edu

[36] MedISys. (n.d.). *Europr Media Monitor.* Retrieved from: http://medusa.jrc.it/medisys/homeedition/all/home.html

[37] Nenkova, A. P. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing.*

[38] Phelps, R. (2010). Weekly Confirmed Hospitalization and Death Report, August 30, 2009 – March 27, 2010. Retrieved from: http://emssolutionsinc.wordpress.com/2010/04/05/

[39] Princeton University. (2010). *About WordNet.* Retrieved from: http://wordnet.princeton.edu/

[40] Volkova, S., Caragea, D., Hsu, W., & Bujuru, S. (2010). Animal disease event recognition and classification. *Proceedings of the First International Workshop on Web Science and Information Exchange in the Medical Web (MedEx 2010)*. Raleigh, NC, USA. April 30, 2010.

[41] Yangarber, R., von Etter, P., & Steinberger, R. (2008). Content Collection and Analysis in the Domain of Epidemiology. *In Proceedings of the 1st international MIE'2008*

workshop on describing medical web resources (DRMed), held at the 21st International Congress of the European Federation for Medical Informatics. Göteborg, Sweden.

[42] Pearman, M. (2010). *ClusterMarker. Google Maps API Projects.* Retrieved from: http://googlemapsapi.martinpearman.co.uk/home.php

[43] Fox, P. (2008). *MapIconMaker.* Retrieved from: http://code.google.com/p/gmaps-utility-library/source/browse/trunk/mapiconmaker/1.1/src/mapiconmaker.js?r=124