Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Decision Support Systems 47 (2009) 508–517

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Automatic online news monitoring and classification for syndromic surveillance

Yulei Zhang [a,*], Yan Dang [a], Hsinchun Chen [a], Mark Thurmond [b], Cathy Larson [a]

[a] Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA
[b] FMD Lab, Center for Animal Disease Modeling and Surveillance (CADMS), University of California, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 8 September 2008
Received in revised form 6 February 2009
Accepted 25 April 2009
Available online 3 May 2009

Keywords:
News classification
News monitoring
Feature selection
Syndromic surveillance

ABSTRACT

Syndromic surveillance can play an important role in protecting the public's health against infectious diseases. Infectious disease outbreaks can have a devastating effect on society as well as the economy, and global awareness is therefore critical to protecting against major outbreaks. By monitoring online news sources and developing an accurate news classification system for syndromic surveillance, public health personnel can be apprised of outbreaks and potential outbreak situations. In this study, we have developed a framework for automatic online news monitoring and classification for syndromic surveillance. The framework is unique and none of the techniques adopted in this study have been previously used in the context of syndromic surveillance on infectious diseases. In recent classification experiments, we compared the performance of different feature subsets on different machine learning algorithms. The results showed that the combined feature subsets including Bag of Words, Noun Phrases, and Named Entities features outperformed the Bag of Words feature subsets. Furthermore, feature selection improved the performance of feature subsets in online news classification. The highest classification performance was achieved when using SVM upon the selected combination feature subset.

1. Introduction

Syndromic surveillance is concerned with the continuous monitoring of public health-related information sources and early detection of adverse disease events [48]. Syndromic surveillance systems aim to provide effective prevention, detection, and management of infectious disease outbreaks, whether naturally-occurring or caused by bioterrorism attacks. The Centers for Disease Control and Prevention (CDC) defines syndromic surveillance systems as those that "collect and analyze morbidity, mortality, and other relevant data and facilitate the timely dissemination of results to appropriate decision makers" [5].

Increasing globalization, combined with accelerating population mobility and more frequent travel, has made the prevention and management of infectious disease outbreaks a growing concern in public health [19]. Outbreaks of various diseases routinely threaten the public health of the world's populations. Recent outbreaks include: the anthrax attacks of 2001, the outbreaks of severe acute respiratory syndrome (SARS) in Asia in 2003, and the continuous avian flu outbreaks in recent years. These outbreaks have highlighted the need of syndromic surveillance systems which can detect and monitor an outbreak and minimize associated morbidity and mortality [5,9].

These days, important public health related news is increasingly available on the World Wide Web in electronic form, and has been shown to be a useful data source for syndromic surveillance [14]. However, the volume of news is very large and there is a question on how to most effectively use this kind of information for syndromic surveillance to accurately detect the signals indicative of disease outbreaks.

Syndromic surveillance systems that include a classification component can facilitate follow-up analysis and outbreak detection [48]. However, to our knowledge, there is currently no automatic online news monitoring and classification system specifically designed for specific infectious diseases. It is also not clear what kind of document representation approach and machine learning algorithm perform best on online news classification for syndromic surveillance. This study is aimed at designing and examining automatic online news monitoring and classification methods for syndromic surveillance and global situational awareness.

The remainder of this article is organized as follows. Section 2 provides an overview of literature concerning syndromic surveillance, news-based syndromic surveillance systems for infectious diseases, online data acquisition, text document representation, and feature selection used in text classification. In Section 3, we describe our research questions. In Section 4, we outline our architecture for automatic news monitoring and classification, after which we present our experiments and the results on foot-and-mouth disease (FMD) related online news in Section 5. Finally, we describe our conclusions and future directions in Section 6.

* Corresponding author.
E-mail address: ylzhang@email.arizona.edu (Y. Zhang).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.04.016

2. Literature review

2.1. Syndromic surveillance

Early knowledge of a disease outbreak plays an important role in improving response effectiveness [35]. Syndromic surveillance, as a public health surveillance approach, is employed to systematically collect, analyze, and interpret "syndrome"-specific data for early detection of public health aberrations [48]. Syndromic surveillance is concerned with continuous monitoring of public health-related information sources and early detection of adverse disease events. It has attracted significant attention in recent years [48].

Collecting data is a critical early step when developing a syndromic surveillance system. Data sources used in syndromic surveillance systems are expected to provide timely pre-diagnosis health indicators and are typically electronically stored and transmitted [48]. Different types of data used for syndromic surveillance typically include: emergency department (ED) visit chief complaints, ambulatory visit records, hospital admissions, over-the-counter (OTC) drug sales from pharmacy stores, triage nurse calls, 911 calls, work or school absenteeism data, veterinary health records, laboratory test orders, and health department requests for influenza testing [25]. Table 1 lists the major data sources used for syndromic surveillance [48].

As shown in Table 1, public news reports and bulletins, published on the Internet to the public, are among the major data sources used for syndromic surveillance. Nowadays, Internet-based resources such as discussion sites and online news centers have become invaluable sources for a new wave of syndromic surveillance systems [7]. The World Health Organization (WHO) relies on those informal sources for about 65% of their outbreak investigations [18]. Currently, nearly all major outbreaks investigated by the WHO are first identified through online sources [14]. The earliest descriptions of the SARS outbreak in Guangdong Province, south China, came from informal online reports [7]. In this study, we chose the online public sources, specifically the online news reports and bulletins, as syndromic surveillance news sources.

2.2. News-based syndromic surveillance systems for infectious diseases

ProMED-mail, Argus, MiTAP and HealthMap are major news-based syndromic surveillance systems for infectious diseases. Table 2 lists the data sources and domain for each system.

The Program for Monitoring Emerging Diseases (also known as ProMED-mail, http://www.isid.org/) is an electronic outbreak reporting system that monitors infectious diseases globally. It is one of the largest publicly available emerging disease and outbreak reporting systems in the world. Originally founded in 1994, and a program of the International Society for Infectious Diseases (ISID, http://www.isid.org/) since 1999, ProMED-mail is distributed without a fee to more than 40,000 e-mail subscribers in over 165 countries, and the site gets roughly 10,000 hits a day [31]. ProMED-mail provides up-to-date information on human diseases, zoonotic diseases and diseases that affect sources of human nutrition (both plants and livestock animals) [26]. Each day, ProMED-mail's editor, assisted by five associates and about 25 scientific experts, cull through the dozens of e-mailed reports of mysterious outbreaks sent in from experts and amateur disease watchers throughout the world. The ProMED-mail team gathers and posts, from newspapers, health department alerts, government reports and other sources, the information for threats to public health that official syndromic surveillance systems may not yet be circulating [31]. In 2003, ProMED-mail was the first to report the disease that turned out to be SARS. In 2006, ProMED-mail's prompt reporting on a cattle die-off in northeastern Kenya that turned out to be the first outbreak of Rift Valley fever in nearly a decade enabled officials to contain the virus [31].

However, one problem for ProMED-mail is that in an attempt to display only the most relevant information, all submissions have to be manually processed by a group of experts and volunteers. This approach is limited by the number of staff available to process the volume of reports submitted throughout the world in multiple languages [44].

In contrast, Argus (http://biodefense.georgetown.edu/projects/argus.aspx) is an automatic disease surveillance system. It is one of the ongoing projects of the Division of Integrated Biodefense (DIB) at Georgetown University. The purpose of Project Argus is to create and implement a global biological event detection and tracking capability that provides early warning alerts [44]. The Argus analytic team consists of multilingual analysts that utilize state of the art online media processing software designed in collaboration with the MITRE Corporation [44]. Argus currently covers 34 languages, and manages between 2200 to 3300 active case files with update report threading for approximately 175 countries and over 130 disease entities [44]. The Argus team has developed a social disruption model to enable rapid detection and assessment of biological threats that may require swift intervention by the international public health community [46]. Social disruption is a deviation from a routine daily activity that can be tracked and used in lieu of direct reporting of diseases [45]. With over 200 social disruption parameters, the model was created by conducting more than 60 in-depth retrospective case studies for socially disruptive biological events affecting animals and humans, and has been operationally validated against over 20,000 prospective biological events detected and tracked to-date [45].

Table 1
Major data sources used for syndromic surveillance [48].

Data source	Description
Chief complaints from ED visits or ambulatory visits	Patient-reported signs and symptoms of their illnesses
School or work absenteeism	Data collected from school or workplace
Hospital admission	Data that is recorded when hospitalization takes place
Triage nurse calls, 911 calls	Symptoms or signs recorded during patient calls when consulting health care nurses
ICD-9	Preliminary diagnoses for billing
ICD-9-CM	Allow assignment of codes to diagnoses and procedures; often used for third-party insurance reimbursement purpose
Laboratory test orders	Orders for laboratory tests
Laboratory test results	Results of laboratory tests
Public sources	News reports or bulletin notification

Table 2
Major news-based syndromic surveillance systems.

System	Data sources	Domain
ProMED-mail	Media reports, health department alerts, government reports, local observers, and other sources	Human diseases, zoonotic diseases and diseases that affect sources of human nutrition (both plants and livestock animals)
Argus (DIB)	Active case files, event reports, articles, and etc.	Over 130 infectious diseases
MiTAP	Multiple information sources (epidemiological reports, newswire feeds, emails, online news, transcribed audios) in multiple languages (English, Chinese, French, German, Italian, Portuguese, Russian, and Spanish)	Infectious diseases
HealthMap	A variety of electronic sources: online news wires, Really Simple Syndication (RSS) feeds, expert-curated accounts (such as ProMED-mail), and validated official alerts (such as WHO)	About 90 infectious diseases

Table 3
Studies on adding intelligence (heuristics) into the crawling strategies.

Study	Intelligence (heuristics)
Chen, Chung, Ramsey, & Yang [11]	Best first search, and genetic algorithm
Chakrabarti, Berg, & Dom [8]	Naïve Bayesian
Rennie & Mccallum [36]	Reinforcement learning
Menczer & Belew [30]; Pant & Srinivasan [34]	Evolutionary algorithm, and neural network
Chau & Chen [10]	Neural network, traditional graph search, and PageRank algorithm
Johnson, Tsioutsiouliklis, & Giles [21]	SVM with linear kernel
Pant & Srinivasan [34]	Compared various machine learning algorithms

Another well known syndromic surveillance system is MiTAP, which was developed by the MITRE Corporation as an experimental prototype using human language technologies to monitor infectious disease outbreaks. MiTAP aims at providing timely, multilingual, global information access to analysts, medical experts and individuals involved in humanitarian assistance and relief work [12]. The system collects, annotates and categorizes documents from multiple open news sources, including foreign sources both in English and native languages that MiTAP translates into English before processing [12]. MiTAP architecture has three phrases: (1) information capture, (2) information processing, and (3) user interface [12]. The information capture process supports Web sources, electronic mailing lists, newsgroups, news feeds, and audio/video data. The information processing is carried out by the Alembic natural language analyzer [2,3] using machine-learned rules and WebSumm [27] to generate a summary for each document. MiTAP consists of a number of different user interfaces to the processed data. The core MiTAP system provides two methods of access: via newsgroups and a web-based search facility. The user-based design approach and the integration of human language components have made the MiTAP system a success [12].

Healthmap (http://www.healthmap.org/en) is also a major news-based syndromic surveillance system. It is an Internet-based alert and mapping surveillance system which integrates outbreak data from a variety of electronic sources: online news wires, Really Simple Syndication (RSS) feeds, expert-curated accounts (such as ProMED-mail), and validated official alerts (such as WHO) [6]. Data are acquired automatically every hour and characterized via text mining to determine the disease category and location of the outbreak [6]. Currently alerts are geocoded to the country scale with province-, state-, or city-level resolution for select countries. Once processed, the outbreak data is visualized on an interactive world map for user-friendly access to the original reports. Users can choose among various information sources. In addition, users can look at all diseases, or the ones they are most interested in. The alerts then appear as flags on the map. By clicking on each of them, users can get the headline and a link to the full story.

However, all the systems shown in Table 2 consider a large number of general infectious diseases, and thus none of them specifically focus on or provide a large portion of information on a particular infectious disease. They do not provide the functionality of classifying the outbreak news of a particular infectious disease into different topic categories either. In this study, we proposed a general framework for building a domain specific news monitoring system for syndromic surveillance, and a classification component to automatically classify the related online news into different topic categories.

2.3. Online data acquisition

With the rapid growth of the Internet, users are often faced with information overload and find it difficult to search for relevant and useful information on the Web. To alleviate the problem, Web crawler programs are often used. These programs exploit the graph structure of the Web by starting at a seed page and then following the hyperlinks within it to attend to other pages. This process repeats with the new pages offering more hyperlinks to follow, until a sufficient number of pages are fetched or a certain higher level objective is reached [34]. To search for important information in a specific domain, vertical search engines are often used by keeping indexes only in that domain [10]. The Web crawler programs used in vertical search engines are selective about the pages fetched and ensure as best as possible that these are relevant to the initiating topics. In Web crawler programs, adding intelligence (heuristics) into crawling strategies can help improve the performance of collecting relevant and important Web pages [34]. Table 3 shows various studies on different intelligence (heuristics) including: best first search, genetic algorithm, reinforcement learning, evolutionary algorithm, Neural Networks, Naïve Bayesian, SVM, PageRank algorithm, etc.

2.4. Text document representation

In order to reduce the complexity of text documents (collected by the crawling programs) and make them easier to handle, full text documents have to be transformed to document vectors which describe the contents. A simple way to transform a text document into a feature vector is by using a Bag of Words representation, where each feature is a single token [23]. In the Bag of Words representation, the semantically empty stop-words need to be removed and the remaining terms are used as the textual representation. Such subset of terms building upon Bag of Words can get the main concepts of an article [32]. The Bag of Words representation has been widely used because of its simple nature and ability to produce a suitable representation of the text [38]. However, sometimes the word-based representation of content is imprecise. It suffers from noise issues associated with seldom-used terms as well as problems of scalability where immense computational power is required for large datasets [37]. An improved representational system which addresses a majority of these shortcomings is Noun Phrases. This representation retains only the nouns and noun phrases within a document and has been found to adequately represent the important article concepts [42,43]. As a consequence, this technique uses fewer terms and can handle article scaling better than Bag of Words [37]. A third representational technique is Named Entities, an extension of Noun Phrases. It functions by selecting the proper nouns of an article that fall within well-defined categories. This process uses a semantic lexical hierarchy [41] as well as a syntactic/semantic tagging process [28] to assign candidate terms to categories. Selected categorical definitions are prescribed by the Message Understanding Conference (MUC-7) Information Retrieval task, and they encompass the entities of date, location, money, organization, percentage, person, and time. This method allows for better generalization of previously unseen terms [37].

2.5. Feature selection used in text classification

No matter which representation approach is used, a typical real world textual dataset usually has a large number of features. However, not all the features are necessary to learning the concept of interest. Many of them may be noisy or redundant and feeding all these features into a model often results in over fitting and poor predictions [29]. Therefore feature selection that aims at identifying a minimal-sized subset of features relevant to the target concept can be applied [13]. The objective of feature selection is threefold: improving the prediction accuracy, providing faster and more cost-effective prediction, and providing a better understanding of the underlying process that generated the data [22,43]. A feature selection method generates different candidates from the feature space and assesses them based on some evaluation criterion to find the best feature subset [22]. Each feature selection method contains two parts: the evaluation criterion

and the generation procedure of candidates. The evaluation criterion is used to assess the goodness of features or feature subsets, and the generation procedure determines how to explore different candidates to find the optimal ones. Graph-based search algorithms are often used to find the optimal features [22].

Text classification assigns category information to documents which are characterized by a set of features. The dominant approach in text classification is based on machine learning techniques. It is an inductive process that automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents [40]. The advantages of the machine learning approach include effectiveness, considerable savings of expert labor, and straightforward portability to different domains [40]. A number of studies have documented the relative merits of text classification algorithms. Well known algorithms with good performance reported in the literature include: K-nearest neighbour (KNN), learn Bayesian net (LBN), Naïve Bayesian (NB), support vector machine (SVM), etc. Using feature selection in text classification can help to select a condensed subset of more relevant features from the initial set to classify future documents [24]. In general, feature selection improves the performance of text classification by offering more concise and precise feature representations of documents [39].

2.6. Research gaps

From prior research, we identified several research gaps. While there are some well known automatic news-based syndromic surveillance systems for general infectious diseases, none has been specifically designed for a particular infectious disease. For classifying online news articles, it is not clear what kind of document representation approach works best, and which machine learning algorithm yields better classification results for syndromic surveillance. Finally, in building an automatic online news classification component for syndromic surveillance, it is yet to be known how feature selection improves performance.

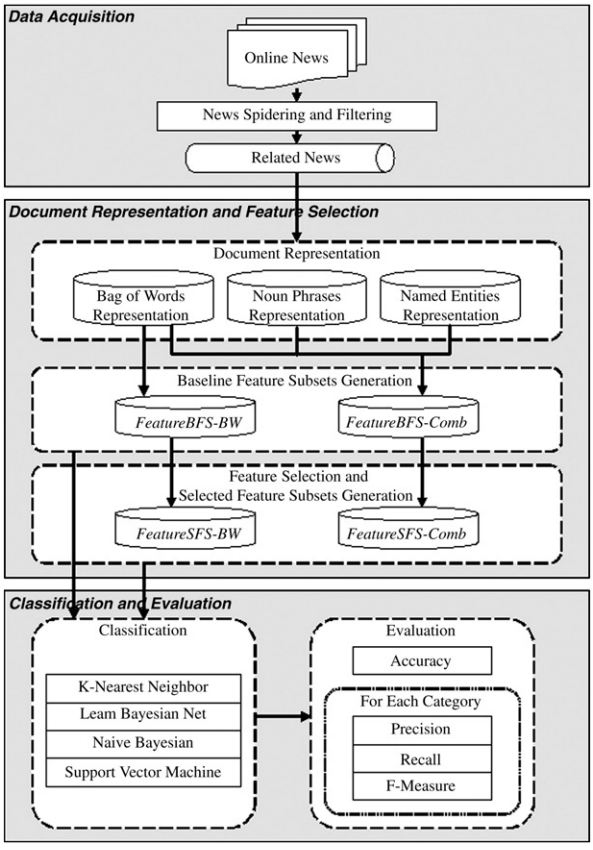

Fig. 1. System architecture of automatic online news monitoring and classification.

Table 4
FMD keywords used by the FMD Lab at UC-Davis.

Language	FMD keywords
English	Foot and mouth disease/hoof and mouth disease
Spanish	Fiebre Aftosa
Portuguese	Febre Aftosa
French	Fièvre aphteuse

3. Research questions

From the research gaps identified above, we formulated the following research questions:

1. How can we monitor online news reports and bulletin notification for syndromic surveillance?
2. Can the features generated by combining the Bag of Words, Noun Phrases and Named Entities approaches outperform those only from the baseline Bag of Words representation in domain specific online news classification for syndromic surveillance?
3. Can feature selection help improve the performance of domain specific online news classification for syndromic surveillance?
4. Which machine learning algorithm performs better for domain specific online news classification for syndromic surveillance?

4. System design and architecture: News monitoring and classification

To answer the research questions, we developed the system architecture as shown in Fig. 1, which contains three components: (1) data acquisition, (2) document representation and feature selection, and (3) classification and evaluation. These components are described in the following sub-sections.

4.1. Data acquisition

In this component, Web crawler programs are developed to spider news articles from the Internet.

The crawler programs are set up to monitor important infectious disease news sources on the Internet. In this study, we use two sets of important FMD news sources containing 27 and 100 news websites respectively identified by the domain experts in the FMD Lab at UC-Davis. Detailed information about these news websites are described in Section 5.1. After collecting the news as HTML pages, we use keyword filtering to filter out unrelated news, thus only keeping FMD related news. The keywords used in this study are listed in Table 4 (see Section 5 for detailed description). We then store all the FMD related news into a local database. This process follows how the domain experts in the FMD Lab at UC-Davis collect and monitor FMD news from the Internet. However, the difference is that here we use automatic ways instead of the manual collecting and filtering done by the FMD Lab.

In this study, since our focus is not to examine Web crawler programs, we do not add heuristics to our crawler programs. In the future, we could add and compare different heuristics to see whether they can improve the performance of the Web crawler programs.

4.2. Document representation and feature selection

In document representation and feature selection component, news documents are transformed from full text versions to document

vectors. We use three widely used document representation methods (Bag of Words, Noun Phrases and Named Entities) to build the document vectors. Each document vector is a set of features. Two feature subsets are generated as baseline feature subsets (BFS). One (denoted as FeatureBFS–BW) is created by using the Bag of Words representation, and the other (denoted as FeatureBFS–Comb) is built by combining the Bag of Words, Noun Phrases, and Named Entities representations. Only features appearing more than five times are taken into account. By conducting feature selection respectively upon FeatureBFS–BW and FeatureBFS–Comb, we build another two feature subsets (denoted as FeatureSFS–BW and FeatureSFS–Comb). The performances of the four different feature subsets are compared in the news classification component as described in Section 4.3.

To conduct feature selection, we choose Correlation-based Feature Selection (CFS) as the evaluation criterion, and Best First Search as the generation procedure. CFS is a widely used feature evaluation criterion. The advantage of CFS is that it evaluates the group of attributes together rather than individually [15–17]. CFS uses a subset evaluation heuristic which assigns high scores to subsets containing attributes that are highly correlated with the class and have low inter-correlation with each other. Hall and Holmes [17] compared the performance of six feature evaluation methods including Correlation-based Feature Selection (CFS), Information Gain Attribute Ranking (IG), Relief (RLF), Principal Components (PC), Consistency-Based Subset Evaluation (CNS), and Wrapper Subset Evaluation (WRP) in the classification on 15 standard machine learning data sets from the University of California, Irvine (UCI) collection. The study showed that CFS performed consistently well on different data sets and suggested that CFS is a good overall performer. In addition, CFS chose fewer features and therefore performed faster. Since we did not intend to compare different feature evaluation methods, we chose CFS for our feature evaluation.

Best First Search is often used as a feature generation procedure. It searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. It does not just terminate when the performance starts to drop but keeps a list of all attribute subsets evaluated, sorted in order of the performance measure, so that it can revisit an earlier configuration [47]. With a good evaluation criterion, Best First Search can drastically reduce the amount of searching needed, thus performing very quickly.

4.3. Classification and evaluation

In the classification and evaluation component, machine learning techniques are used to conduct the online news classification task, and the performances are evaluated based on standard machine learning evaluation metrics: accuracy, precision, recall, and *F*-measure.

Due to their good performances as reported in the literature, we chose four widely used classification algorithms to perform news classification: K-nearest neighbour (KNN), learn Bayes net (LBN), Naïve Bayesian (NB), and support vector machine (SVM). First introduced into text classification by Joachims [20], SVM achieves the best performance for various text classification tasks [1, 33,49]. In this study, we compared SVM with the other three algorithms to see whether it can achieve relatively higher performance on online news classification for syndromic surveillance.

To conduct evaluation, accuracy measures the overall correctness of classification:

$$\text{accuracy} = \frac{\text{number of all correctly classified news items}}{\text{total number of news items}}$$

Precision, recall, and *F*-measure evaluate the correctness for each class. Specifically, precision indicates the correctness of classification, and recall indicates the completeness of classification. *F*-measure is the harmonic mean of precision and recall.

$$\text{precision(i)} = \frac{\text{number of correctly classified news items for class i}}{\text{total number of news items classified as class i}}$$

$$\text{recall(i)} = \frac{\text{number of correctly classified news items for class i}}{\text{total number of news items in class i}}$$

$$\text{F-measure(i)} = \frac{2 \times \text{precision(i)} \times \text{recall(i)}}{\text{precision(i)} + \text{recall(i)}}$$

5. Experimental study: FMD news monitoring and classification

In this study, we chose foot-and-mouth disease (FMD) related online news to conduct our experiments. FMD is one of the most devastating diseases of farm animals. It occurs throughout the world and is a significant hazard to agriculture. The 2001 epidemic in the UK led to the loss of six million livestock [4]. The total costs arising from this outbreak have been put at no less than 9 billion (http://www.fmd.brass.cf.ac.uk/). Global situational awareness can play a critical role in warning of potential and imminent outbreaks.

The FMD Lab at UC-Davis (http://fmd.ucdavis.edu/) has developed models and systems for global FMD surveillance, including the FMD BioPortal Web-based system developed jointly with the Artificial Intelligence (AI) Lab at the University of Arizona. They have also been gathering and processing FMD-related news from the FMD World Reference Laboratory, the OIE, the FAO, etc., and classifying the news into different categories. The news is collected daily, weekly or biweekly, according to the update frequencies of different websites. They also search Google and Yahoo everyday for FMD-related news in four languages by using the keywords shown in Table 4. They perform all news gathering and classification work manually. This manual work is time consuming and labor intensive, and can also lead to information loss. However, to our knowledge, there is currently no automatic news monitoring and classification system specifically for FMD.

5.1. Automatic online FMD news monitoring

To monitor the FMD related news from the Internet, we developed Web crawler programs for two sets of important FMD news sources identified by the domain experts from the FMD Lab at UC-Davis. The first set, as shown in Table 5, contains 27 online news sources that fall into four categories: news websites, government websites, international organizations, and research labs. Although all of them provide FMD outbreak related news, the news websites and government websites mainly focus on the social and economic consequences of an outbreak and general information, and the international organizations and research labs emphasize epidemiological reports and analysis.

Table 5
Important online FMD news sources identified by the FMD Lab.

News website	Government website	International organization	Research lab
All Africa	European Commission for Agriculture	New OIE	WRL
PigSite	FGI ARRIAH	Old OIE	FMD Lab in UC Davis
BBC FMD News	Federation of American Scientists	FAO	DEFRA
The New Vision	EUFMD	SEAFMD	
Bloomberg News	Agriculture, Canada	ProMED	
Times Online	Argentina-SENASA	OIE-JP	
Agrolink Noticias	Peru-SENASA		
World Farming News	SESA		
Arabic News	European Commission for Agriculture		
9 sites	9 sites	6 sites	3 sites

The second set of important FMD news sources are identified from a collection of 2832 pieces of important FMD news manually gathered from the Internet by domain experts from the FMD Lab at UC-Davis. These news items, reported between October 6, 2004 and January 20, 2007, come from 878 different websites including the 27 sources listed in Table 5. As shown in Fig. 2, only four out of the 878 websites have more than 50 news items each in the collection. They are: (1) Agrolink (www.agrolink.com.br) with 191 news items, (2) Allafrica (http://allafrica.com) with 144 news items, (3) OIE (www.oie.int) with 141 new items, and (4) Cattlenetwork (www.cattlenetwork.com) with 95 news items. 564 other websites have only one news item each. Among all the 878 websites, there are 100 websites, each of which has more than 5 news items in the collection. These 100 websites form the second set of sources used in FMD news monitoring.

Once online news are collected as HTML pages from the two sets of important news sources using the Web crawler programs, we use keyword filtering to identify all the FMD related news. The keywords used are listed in Table 4. Till now, we have gathered more than 180,000 FMD related news documents from the 27 sources in the first set and more than 650,000 FMD related news documents from the 100 sources in the second set.

5.2. Testbed for automatic FMD news classification

In order to examine the news classification component, we conducted experiments on a testbed containing 1674 pieces of FMD related news. They were culled from the collection containing 2,832 important FMD news articles gathered by the FMD Lab from the Internet. We use them as the testbed for our news classification component because these 1674 news articles have category information assigned by domain experts and therefore can be used to conduct the proposed evaluation metrics. The category assigned to each news item by domain experts is used as the golden standard for evaluation.

There are three categories assigned to the test data: (1) FMD outbreak related news, (2) FMD control program related news, and (3) FMD social, economic and general information. The classifiers we built are based on these categories.

5.3. Hypotheses

In this section, we list the three sets of hypotheses we tested in this study. All the hypotheses are obtained based on our literature review and the research gaps and questions we identified. These hypotheses are all focused on the automatic online news classification component.

For domain specific online news document representation, we aimed to compare the combined feature subsets with the Bag of Words feature subsets. As discussed in the literature review, since Noun Phrases and Named Entities features can capture more syntactic and semantic information with less noise than the basic Bag of Words features, we posit that using all three types of features can achieve better performance than using only the Bag of Words features in domain specific online news classification. The specific hypotheses tested are as follows:

H1. The combination of Bag of Words, Noun Phrases and Named Entities features outperform the baseline Bag of Words features in domain specific online news classification (regardless of whether or not feature selection is conducted).

H1a. FeatureBFS–Comb>FeatureBFS–BW.

H1b. FeatureSFS–Comb>FeatureSFS–BW.

For domain specific online news feature selection, we aimed to compare the feature subsets created by conducting feature selection with the baseline feature subsets without conducting feature selection. As discussed in the literature review, feature selection can often improve the text classification performance. Therefore, we posit that conducting feature selection can improve the performance of domain specific online news classification. The specific hypotheses tested are as follows:

H2. The selected features developed by conducting feature selection outperform the baseline features in domain specific online news classification (regardless of whether or not the three representation approaches are combined).

H2a. FeatureSFS–BW>FeatureBFS–BW.

H2b. FeatureSFS–Comb>FeatureBFS–Comb.

For domain specific online news classification, we aimed to compare support vector machine (SVM) with the other three widely used machine learning algorithms. SVM is often reported to have the best performance in text classification, although K-nearest neighbour (KNN), learn Bayesian net (LBN) and Naïve Bayesian (NB) also perform well in general. Therefore, we posit that SVM outperforms the other three algorithms in domain specific online news classification. The specific hypotheses tested are as follows:

H3. Support vector machine (SVM) outperforms K-nearest neighbour (KNN), learn Bayesian net (LBN) and Naïve Bayesian (NB) for the selected features developed by conducting feature selection in domain specific online news classification.

H3a. SVM>KNN on FeatureSFS–BW.

H3b. SVM>KNN on FeatureSFS–Comb.

H3c. SVM>LBN on FeatureSFS–BW.

H3d. SVM>LBN on FeatureSFS–Comb.

H3e. SVM>NB on FeatureSFS–BW.

H3f. SVM>NB on FeatureSFS–Comb.

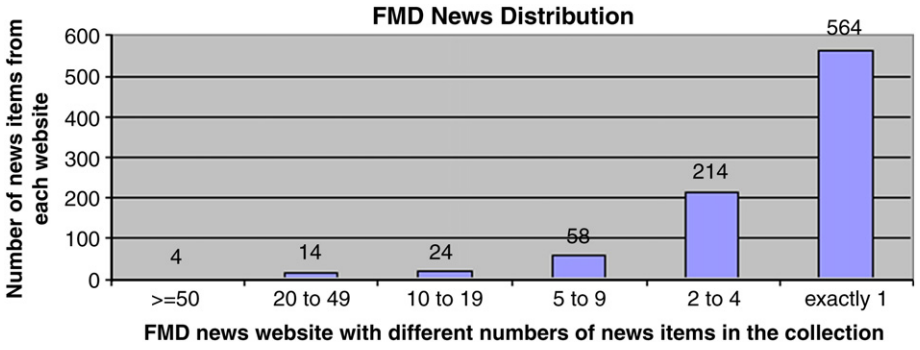

Fig. 2. The news distribution on different websites.

5.4. Experiment results

As mentioned before, our experiments are based on the automatic online news classification component. We tested the classification performance of different feature subsets on different machine learning algorithms as summarized in Table 6. The four classification algorithms we used, including KNN ($k=10$), LBN, NB, and SVM are from the Weka Data Mining Package [147]. For each classifier, we used 90% of the news articles in the testbed for training, and predicted the class labels of the remaining 10% news articles as testing. We repeated this process one hundred times by randomly splitting the testbed for statistical analysis. For each feature subset with each classification algorithm, we report in Table 6 the accuracy, the average precision, the average recall, and the average *F*-measure values on the three FMD news categories. The values in bold fonts denote the best performances.

In our experiments, the classifier using SVM upon feature subset FeatureSFS–Comb achieved the best performance. It achieved the highest accuracy of 77.04%, the highest average precision of 77.10%, the highest average recall of 77.05%, and the highest average *F*-measure of 77.08%. We can also see from Table 6 that in most cases, the two combination feature subsets worked better than the two Bag of Words feature subsets. Specifically, in most cases, FeatureBFS–Comb had higher evaluation values than FeatureBFS–BW, and FeatureSFS–Comb had higher evaluation values than FeatureSFS-BW. By conducting feature selection, the average accuracy of Bag of Words features increased from 70.66% to 73.27%, and the average accuracy of the combination features increased from 72.13% to 74.96%.

The errors in classification were mainly caused by such news that has information related to more than one category. For example, some news articles mentioned both FMD outbreaks which belong to category 1, and plans on how to control the outbreaks which belong to category 2. Some other news articles first talked different control programs on FMD which belong to category 2, and then compared them in terms of their social and economic influences which belong to category 3. In this case, to decide which category a news article belongs to, the domain experts from the FMD Lab at UC-Davis read the content to see which topic the large portion of the article is about. However, for automatic classifiers, the results sometimes were inconsistent with the golden standard provided by domain experts.

Furthermore, in order to test our hypotheses, we conducted pair wise single-sided *t* tests on accuracy and average *F*-measure. The *p* values and results for the tests of our hypotheses are presented in Table 7, where *p* values with * and ** indicate significant differences at the levels of $\alpha=0.05$ and 0.01, respectively. The underlined *p* values indicate that the results contradict the hypotheses. Overall, most of our hypotheses were supported by our experiments.

For Hypothesis 1a, FeatureBFS–Comb significantly outperformed FeatureBFS–BW on accuracy for each of the four classification algorithms. Although Hypothesis 1a on the average *F*-measure was not confirmed for KNN ($p=0.0550>0.05$), it was confirmed for the other three algorithms with *p* values less than or equal to 0.001. For Hypothesis 1b, FeatureSFS–Comb significantly outperformed FeatureSFS-BW with most *p* values less than or equal to 0.001, except for the accuracy of KNN ($p=0.1260>0.05$). Overall, the combination features achieved better performance than Bag of Words features, because the combination features capture not only the common words but also some important noun phrases and named entities.

For Hypothesis 2a, FeatureSFS–BW significantly outperformed FeatureBFS–BW with most *p* values less than 0.0001, except for the average *F*-measure of SVM ($p=0.4968>0.05$). For Hypothesis 2b, six out of the eight *p* values were less than 0.0001, although it was not confirmed on the average *F*-measure for SVM ($p=0.1499>0.05$). Overall, the feature subsets conducted feature selection outperformed the baseline feature subsets.

For Hypothesis 3, all six sub hypotheses were confirmed for accuracy, and four of them were confirmed on the average *F*-measure. Overall, SVM was the best performer with selected feature subsets.

Table 6
Performance measures of the online news classification component.

Feature subset	Classification algorithm	Accuracy	Average precision	Average recall	Average *F*-measure
FeatureBFS–BW	KNN	63.89%	70.68%	63.89%	67.11%
	LBN	71.56%	72.83%	62.64%	67.35%
	NB	74.96%	74.64%	63.78%	68.78%
	SVM	72.22%	72.35%	72.19%	72.27%
	Average	70.66%			
FeatureBFS–Comb	KNN	64.43%	54.70%	64.45%	59.18%
	LBN	73.94%	74.44%	73.95%	74.19%
	NB	75.85%	75.45%	75.87%	75.66%
	SVM	74.30%	74.60%	74.26%	74.43%
	Average	72.13%			
FeatureSFS–BW	KNN	72.22%	58.05%	72.21%	64.36%
	LBN	73.17%	71.77%	73.18%	72.47%
	NB	72.58%	71.44%	72.61%	72.02%
	SVM	75.13%	74.59%	75.13%	74.86%
	Average	73.27%			
FeatureSFS–Comb	KNN	71.56%	74.60%	71.57%	73.05%
	LBN	76.15%	75.19%	76.14%	75.66%
	NB	75.07%	74.24%	75.06%	74.65%
	SVM	**77.04%**	**77.10%**	**77.05%**	**77.08%**
	Average	74.96%			

Table 7
Results of hypothesis testing.

No.	Hypothesis	*p* value on accuracy	Result	*p* value on average *F*-measure	Result
H1a	FeatureBFS–Comb> FeatureBFS–BW				
	KNN	0.0008**	Confirmed	0.0550	Not confirmed
	LBN	<0.0001**	Confirmed	<0.0001**	Confirmed
	NB	<0.0001**	Confirmed	0.0010**	Confirmed
	SVM	<0.0001**	Confirmed	0.0005**	Confirmed
H1b	FeatureSFS–Comb> FeatureSFS-BW				
	KNN	0.1260	Not confirmed	0.0010**	Confirmed
	LBN	<0.0001**	Confirmed	<0.0001**	Confirmed
	NB	<0.0001**	Confirmed	<0.0001**	Confirmed
	SVM	<0.0001**	Confirmed	<0.0001**	Confirmed
H2a	FeatureSFS–BW> FeatureBFS–BW				
	KNN	<0.0001**	Confirmed	<0.0001**	Confirmed
	LBN	<0.0001**	Confirmed	<0.0001**	Confirmed
	NB	<0.0001**	Confirmed	<0.0001**	Confirmed
	SVM	<0.0001**	Confirmed	<0.4968	Not confirmed
H2b	FeatureSFS–Comb> FeatureBFS–Comb				
	KNN	<0.0001**	Confirmed	<0.0001**	Confirmed
	LBN	<0.0001**	Confirmed	<0.0001**	Confirmed
	NB	<0.0001**	Confirmed	0.0384*	Confirmed
	SVM	<0.0001**	Confirmed	0.1499	Not confirmed
H3a	SVM>KNN on FeatureSFS–BW	<0.0001**	Confirmed	<0.0001**	Confirmed
H3b	SVM>KNN on FeatureSFS–Comb	<0.0001**	Confirmed	<0.0001**	Confirmed
H3c	SVM>LBN on FeatureSFS–BW	<0.0001**	Confirmed	0.4577	Not confirmed
H3d	SVM>LBN on FeatureSFS–Comb	0.0041**	Confirmed	0.2919	Not confirmed
H3e	SVM>NB on FeatureSFS–BW	<0.0001**	Confirmed	<0.0001**	Confirmed
H3f	SVM>NB on FeatureSFS–Comb	<0.0001**	Confirmed	0.0002**	Confirmed

Note. Significance levels * $\alpha=0.05$ and ** $\alpha=0.01$.

Table 8
The 56 features in FeatureSFS–Comb.

No.	Feature	No.	Feature	No.	Feature
1	**Animal health service**	20	**Mad cow disease**	39	Such as
2	Beef	21	**Mass vaccination**	40	Susceptible
3	Board	22	**Mass vaccination campaign**	41	**Susceptible animals**
4	Campaign	23	Measures	42	The disease
5	Cattle	24	Meat	43	The embargo
6	Company	25	Ministry	44	**The FMD**
7	Confirmed	26	Mouth	45	The herd
8	DES	27	**Negative Results**	46	The outbreak
9	Detected	28	Origin	47	**The outbreak of foot and mouth disease**
10	Director	29	Outbreak	48	Threatening
11	Disinfection	30	Prices	49	To control
12	Emerging	31	Quarantine	50	US
13	FMD	32	Received	51	Vaccinate
14	Herds	33	Reported	52	Vaccinated
15	Imports	34	Results	53	**Vaccination campaign**
16	Infected	35	Sacrificed	54	**Vaccine doses**
17	Information	36	Samples	55	Venezuela
18	Isolated	37	Serotype	56	Village
19	**Livestock production**	38	Stricken		

The feature subsets FeatureBFS–BW, FeatureBFS–Comb, FeatureSFS–BW, and FeatureSFS–Comb contain 1473 features, 2130 features, 48 features and 56 features respectively. As shown above, FeatureSFS–Comb performed best in FMD news classification. The feature subset size was reduced significantly by using CFS + Best First Search described before. With fewer features, the performance speed of FMD news classification increased dramatically.

5.5. Discussions

From our experiment results, we can see that, overall, the combination of Bag of Words, Noun Phrases and Named Entities features outperform the baseline Bag of Words features in online FMD news classification; feature selection can improve the classification performance; and SVM achieves overall better performance than the other commonly used classification algorithms. The best performance was achieved using SVM upon the selected feature subset of the combination features, i.e. FeatureSFS–Comb.

In Table 8, we list the 56 features in FeatureSFS–Comb in alphabetical order. The bold fonts indicate some important Noun Phrases and Named Entities features that did not appear in FeatureSFS–BW (48 Bag of Words features). We believe it is because of those features that FeatureSFS–Comb achieved overall higher performance than FeatureSFS–BW in our experiments. As shown in Table 8, those important and semantically meaningful features include: animal health service, livestock production, mad cow disease, mass vaccination, mass vaccination campaign, negative results, susceptible animals, the embargo, the FMD, the herd, the outbreak, the outbreak of foot and mouth disease, to control, vaccination campaign and vaccine doses.

6. Conclusions and future directions

Increasing globalization, population mobility and travel frequency have made the prevention and management of infectious disease outbreaks a growing concern in public health. Syndromic surveillance can protect the public's health against infectious diseases by providing effective prevention, detection, and management of infectious disease outbreaks. Among the major data sources used for syndromic surveillance, online news is an important one. Some well known automatic news-based syndromic surveillance systems exist for general infectious diseases. However, to our best knowledge, none of them specifically focus on or provide a large portion of information on a particular infectious disease.

In this study, we have described a general framework for building an infectious disease specific news monitoring and classification system for syndromic surveillance. Our experimental study is based on FMD news. For FMD news monitoring, we have set up Web crawler programs to collect news articles from important online FMD news sources. We then use keyword filtering to identify FMD related news. For FMD news classification, we compared the performance of different feature subsets on different machine learning algorithms. The results showed that the combined feature subsets including Bag of Words, Noun Phrases, and Named Entities features outperformed the Bag of Words feature subsets. Furthermore, feature selection improved the performance of feature subsets in FMD news classification. The highest classification performance was achieved when using SVM upon the selected combination feature subset (FeatureSFS–Comb).

This study has made several contributions. First, we propose a general framework for building an automatic, infectious disease specific news monitoring and classification system for syndromic surveillance, which is becoming increasingly critical for protecting the public's health but has received little investigative attention. The well known existing news-based syndromic surveillance systems are based on general infectious diseases instead of a specific disease. Our approach focusing on automatic monitoring and classifying online news for a specific disease provides an informative point of departure for continued research. With such a framework, we can automatically gather a particular infectious disease related news from the Web and organize the information into different categories, thus providing accurate and timely information to health providers. Therefore, we believe such a unique framework is important and beneficial to researchers interested in syndromic surveillance, especially on infectious diseases. Second, the techniques used to develop the framework are based on the algorithms which have consistently reported the best performances. In order to identify what kind of document representation approach works best for classifying online news articles, which machine learning algorithm yields better classification results, and how feature selection improves the classification performance for syndromic surveillance, we examine different high-performance algorithms. None of the techniques have been previously used in the context of syndromic surveillance on infectious diseases. In addition, we demonstrate the viability of using our proposed framework to automatically monitor and classify FMD related online news. The automatic approach is much more advantageous than the labor-intensive and time-consuming way of monitoring and classifying FMD related online news manually done in the FMD Lab at UC-Davis; therefore, alleviates the burden caused by the manual work of the FMD Lab. Besides FMD, this general framework can be applied to automatically monitor and classify online news related to other particular diseases. Thus, the current study provides important guidance and implications for future infectious disease specific classification component development for syndromic surveillance systems.

This study has some limitations that can also be explored further. For the news monitoring component, we first used Web crawler programs to collect all the news Web pages from the online sources and then used keyword filtering to identify the FMD related news from the large collection. However, adding intelligence (heuristics) into crawling strategies could be more effective and efficient than keyword filtering in collecting relevant and important FMD news Web pages. To evaluate our news classification component, we used a set of FMD related online news provided by the FMD Lab as our testbed instead of using the news gathered by our monitoring component, because this set of FMD news has the category information assigned by domain experts; thus, can be used to calculate the evaluation metrics. However, these news items may be cleaner with fewer noises than the news gathered by the automatic monitoring component. Thus, this may lead to better classification results in our experiment. In the future, we will use the news gathered by the

automatic news monitoring component to test our classification component. In addition, we only focus on English news sources in this study. However, due to the globalization and mobility nature of infectious diseases, some important and timely news, especially the outbreak news may be reported in other languages. Future research could extend to incorporate multilingual processing component to deal with important news sources in other languages. However, even in its current state, the empirical evidence clearly indicates that our proposed framework for building an automatic, infectious disease specific news monitoring and classification system for syndromic surveillance represents a solid foundation upon which to build more advanced automatic, infectious disease specific syndromic surveillance systems.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation (NSF ITR 0428241 "BioPortal: A National Center of Excellence for Infectious Disease Informatics").

References

[1] A. Abbasi, H. Chen, Identification and comparison of extremist-group web forum messages using authorship analysis, IEEE Intelligent Systems 20 (5) (2005) 67–75.
[2] J. Aberdeen, J. Burger, D. Day, L. Hirschman, D. Palmer, P. Robinson, M. Vilain, MITRE: description of the Alembic system as used in MET, Proceedings of the TIPSTER 24-Month Workshop, 1996.
[3] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, M. Vilain, MITRE: description of the Alembic system as used for MUC-6, Proceedings of the Sixth Message Understanding Conference (MUC-6), 1995.
[4] P.R. Bessell, M.J. Keeling, M.J. Tildesley, M.E.J. Woolhouse, T8.6: Future risks of foot and mouth disease spread in the UK, infectious diseases: preparing for the future, Office of Sscience and Innovation, 2006.
[5] D.M. Bravata, K.M. McDonald, W.M. Smith, C. Rydzak, H. Szeto, D.L. Buckeridge, C. Haberland, D.K. Owens, Systematic review: surveillance systems for early detection of bioterrorism-related diseases, Annals of Internal Medicine 140 (11) (2004) 910–922.
[6] J. Brownstein, C. Freifeld, HealthMap: the development of automated real-time internet surveillance for epidemic intelligence, Eurosurveillance 12 (48) (2007).
[7] J.S. Brownstein, C.C. Freifeld, B.Y. Reis, K.D. Mandl, Evaluation of online media reports for global infectious disease intelligence, Advances in Disease Surveillance 4 (3) (2007).
[8] S. Chakrabarti, M.V.D. Berg, B. Dom, Focused crawling: a new approach to topic-specific web resource discovery, Proceedings of the 8th International World Wide Web Conference, 1999.
[9] W. Chang, D. Zeng, H. Chen, A stack-based prospective spatio–temporal data analysis approach, Decision Support Systems (2008), doi:10.1016/j.dss.2007.12.008.
[10] M. Chau, H. Chen, Comparison of three vertical search spiders, Computer 36 (5) (2003) 56–62.
[11] H. Chen, Y. Chung, M. Ramsey, C. Yang, A smart itsy bitsy spider for the web, Journal of the American Society for Information Science 49 (7) (1998) 604–618.
[12] L. Damianos, S. Wohlever, R. Kozierok, J. Ponte, MiTAP: a case study of integrated knowledge discovery tools, Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), 2003.
[13] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (3) (1997) 131–156.
[14] C.C. Freifeld, K.D. Mandl, B.Y. Reis, J.S. Brownstein, HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports, Journal of the American Medical Informatics Association 15 (2008) 150–157.
[15] M. Hall, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the 17th International Conference on Machine Learning (ICML 2000), 2000.
[16] M.A. Hall, Correlation-based feature selection for machine learning, Deptartment of Computer Science, Univiversity of Waikato, 1998.
[17] M.A. Hall, G. Holmes, Benchmarking attribute selection techniques for data mining, IEEE Transactions on Knowledge and Data Engineering 15 (6) (2003) 1437–1447.
[18] D.L. Heymann, G.R. Rodier, WHO operational support team to the global outbreak alert and response network. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases, Lancet Infectious Diseases 1 (2001) 345–353.
[19] P.J.-H. Hu, D. Zeng, H. Chen, C. Larson, W. Chang, C. Tseng, J. Ma, System for infectious disease information sharing and analysis: design and evaluation, IEEE Transactions on information technology in biomedicine 11 (4) (2007) 483–492.
[20] T. Joachims, Text categorization with support vector machines: learning with many relevant features, proceedings of 10th European Conference on Machine Learning (ECML 98), 1998, pp. 137–142.
[21] J. Johnson, K. Tsioutsiouliklis, C.L. Giles, Evolving strategies for focused web crawling, Proceedings of the 20th International Conference on Machine Learning (ICML 2003), Washington DC, 2003.
[22] J. Li, H. Su, H. Chen, B.W. Futscher, Optimal search-based gene subset selection for gene array cancer classification, IEEE Transactions on information technology in biomedicine 11 (4) (2007).
[23] C. Liao, S. Alpha, P. Dixon, Feature preparation in text categorization, in: e.S.J.S.a.G.J.W.a.M. Hegland (Ed.), ADM03 workshop (Australian Data Mining Workshop), Canberra, Australia, 2003, pp. 143–162.
[24] F.G.A. López, M.G.A. Torres, B.M. Batista, J.A.M. Pérez, J.M. Moreno-Vega, Solving feature subset selection problem by a parallel scatter search, European Journal of Operational Research 169 (2) (2006) 477–489.
[25] H. Ma, H. Rolka, K. Mandl, D. Buckeridge, A. Fleischauer, J. Pavlin, Implementation of laboratory order data in biosense early event detection and situation awareness system, MMWR (CDC) 54 (Suppl) (2005) 27–30.
[26] L. Madoff, Cooperation between animal and human health sectors is key to the detection, surveillance, and control of emerging disease, Eurosurveillance 11 (51) (2006).
[27] I. Mani, E. Bloedorn, Summarizing similarities and differences among related documents, Information Retrieval 1 (1) (1999) 35–67.
[28] D.M. McDonald, H. Chen, R.P. Schumaker, Transforming open-source documents to terror networks: the Arizona TerrorNet, American Association for Artificial Intelligence Conference Spring Symposia, Stanford, CA, 2005.
[29] R. Meiri, J. Zahavi, Using simulated annealing to optimize the feature selection problem in marketing applications, European Journal of Operational Research 171 (3) (2006) 842–858.
[30] F. Menczer, R.K. Belew, Adaptive retrieval agents: internalizing local context and scaling up to the web, Machine Learning 39 (2000) 203–242.
[31] J. Miller, Website for the germ-obsessed: to its devotees, ProMED, which tracks disease outbreaks worldwide, is a must-read, Los Angeles Times, 2007.
[32] D. Moldovan, M. Pasca, S. Harabagiu, M. Surdeanu, Performance issues and error analysis in an open-domain question answering system, ACM Transactions on Information Systems 21 (2) (2003) 133–154.
[33] B. Pang, L. Lee, S. Vaithyanathain, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.
[34] G. Pant, P. Srinivasan, Learning to crawl: comparing classification schemes, ACM Transactions on Information Systems 23 (4) (2005) 430–462.
[35] R.W. Pinner, C.A. Rebmann, A. Schuchat, J.M. Hughes, Disease surveillance and the academic, clinical, and public health communities, Emerging Infectious Diseases 9 (7) (2003) 781–787.
[36] J. Rennie, A.K. Mccallum, Using reinforcement learning to spider the web efficiently, Proceedings of the 16th International Conference on Machine Learning, Morgan Kaufmann, San Francisco,CA, 1999, pp. 335–343.
[37] R. Schumaker, H. Chen, Evaluating a news-aware quantitative trader: the effect of momemtum and contrarian stock selection strategies, Journal of the American Society for Information Science and technology 59 (2) (2008) 247–255.
[38] R. Schumaker, H. Chen, Textual analysis of stock market prediction using financial news articles, Americas Conference on Information Systems (AMCIS 2006), Acapulco, Mexico, 2006.
[39] S. Scott, S. Matwin, Feature engineering for text classification, Proceedings of ICML-99, 16th International Conference on Machine Learning, 1999, pp. 379–388.
[40] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1) (2002) 1–47.
[41] S. Sekine, C. Nobata, Definition, dictionaries and tagger for extended named entity hierarchy, Forth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.
[42] K.M. Tolle, H. Chen, Comparing noun phrasing techniques for use with medical digital library tools, Journal of the American Society for Information Science 51 (4) (2000) 352–370.
[43] C.-P. Wei, Y.-H. Lee, Event detection from online news documents for supporting environmental scanning, Decision Support Systems 36 (4) (2004) 385–401.
[44] J.M. Wilson, Argus: a global detection and tracking system for biological events, Advances in Disease Surveillance 4 (21) (2007).
[45] J.M. Wilson, Indications and warnings to dectect and track biological events, Presented at the Sixth Annual International Society for Disease Surveillance Conference Indianapolis, IN, 2007.
[46] J.M. Wilson, M.G. Polyak, J.W. Blake, J. Collmann, A heuristic indication and warning staging model for detection and assessment of biological events, Journal of the American Medical Informatics Association 15 (2) (2008) 158–171.
[47] I.H. Witten, E. Frank, Data mining, practical machine learning tools and techniques, 2nd ed.Morgan Kaufmann, San Francisco, California, 2005.
[48] P. Yan, H. Chen, D. Zeng, Syndromic surveillance systems: public health and biodefense, Annual Review of Information Sciences and Technology 42 (2008) 425–495.
[49] R. Zheng, J. Li, Z. Huang, H. Chen, A framework for authorship analysis of online messages: writing-style features and techniques, Journal of the American Society for Information Science and technology 57 (3) (2006) 378–393.

Yulei Zhang received the BS degree in Computer Science and MS degree in Bioinformatics from Shanghai Jiao Tong University. He is currently working toward the PhD degree in information systems and is also a research associate in the Artificial Intelligence Lab, University of Arizona. His research interests include text mining, data mining and bioinformatics.

Yan Dang received the BS and MS degrees in Computer Science from Shanghai Jiao Tong University. She is currently working toward the PhD degree in information systems and is also a research associate in the Artificial Intelligence Lab, University of Arizona. Her research interests include text mining and human computer interaction.

Dr. Hsinchun Chen received the BS degree from the National Chiao-Tung University, Hsinchu, Taiwan, the MBA degree from SUNY Buffalo, and the PhD degree in information systems from New York University. He is a professor of information systems and the director of the Artificial Intelligence Lab, University of Arizona. He has authored/edited 18 books, 17 book chapters, and more than 150 SCI journal articles covering digital library, intelligence analysis, biomedical informatics, data/text/web mining, knowledge management, and web computing. He serves on 10 editorial boards and has been an advisor for major US National Science Foundation, US Department of Justice, US National Library of Medicine, US Department of Defense, US Department of Homeland Security, and other international research programs. He received the IEEE Computer Society 2006 Technical Achievement Award. He is a fellow of the IEEE and the AAAS.

Dr. Mark Thurmond has research interests related to the epidemiology of infectious diseases of cattle and application of epidemiologic principles to prevention, control, and eradication of diseases and infections that affect animal health and productivity. Diseases of special interest include bovine viral diarrhea, neosporosis, diseases of the mammary gland, abortion, and foreign animal diseases, such as foot-and-mouth disease. He received his DVM from the University of California/Davis in 1972, and the MPVM degreed from UC Davis in 1975. He received a PhD from the University of Florida in 1982.

Cathy Larson is Associate Director of the Artificial Intelligence Lab and an Associate Research Scientist in the MIS Department, University of Arizona. She received her B.A. from the University of Illinois at Urbana-Champaign, with majors in Spanish and Anthropology and a minor in Portuguese. Graduation was followed by a stint in the Peace Corps, studying agricultural extension strategies in Costa Rica then serving in Ecuador. She received her M.S. in Library and Information Science from UIUC. Following the award of her M.S., she went to the University of Iowa Library, to serve first as Media Bibliographer (1987), then as Head of the Library's first Preservation Department (1990), with responsibilities in managing grant-funded microfilming projects. A desire to move to the Southwest and an interest in team-based organizations led to her appointment in 1994 as the Fine Arts/Humanities Team Leader at the University of Arizona Library. In 1998, she became the UA Library's first Data Services Librarian, with responsibilities in planning and implementing data services programs and digital library projects.