Disaster SitRep - A Vertical Search Engine and Information Analysis Tool in Disaster Management Domain

Li Zheng, Chao Shen, Liang Tang, Chunqiu Zeng, Tao Li,
Steve Luis, Shu-Ching Chen, Jainendra K. Navlakha
School of Computing and Information Sciences, Florida International University
{lzhen001, cshen001, ltang002, czeng001, taoli, luiss, chens, navlakha}@cs.fiu.edu

Abstract

With the rise of heterogeneous information delivering platform, the process of collecting, integrating, and analyzing disaster related information from diverse channels becomes more difficult and challenging. Further, information from multiple sources brings up new challenges for information presentation. In this paper, we design and implement a ***Disaster Situation Reporting System (Disaster SitRep)*** *that is essentially a disaster information collecting, integration, and presentation platform to address three critical tasks that can facilitate information acquisition, integration and presentation by utilizing domain knowledge as well as public and private web resources for major disaster recovery planning and management. Our proposed techniques create a disaster domain-specific search engine and a geographical information presentation and navigation platform using advanced data mining and information retrieval techniques for disaster preparedness and recovery that helps impacted communities better understand the current disaster situation. Specifically, hierarchical clustering with constraints are used to automatically update existing disaster concept hierarchy; taxonomy-based focused crawling component is developed to automatically detect, parse and filter those relevant web resources; a domain-oriented skeleton for each type of disasters is used to extract disaster events from disaster documents by defining the set of structural attributes. Furthermore, the platform can perform not only as a domain-specific search engine but also as an information monitoring and analysis tool for decision support during recovery phase of disasters.*

Keywords: *Data Mining, Disaster Information Management, Vertical Search Engine, Concept Hierarchy, Focused Crawler, Disaster Event Extraction*

1. Introduction

Natural or man-made hazardous disasters cause huge impact in business continuity activities. Thin margins and lack of a well-designed and regularly tested disaster plan make companies, particularly small businesses, especially vulnerable [1-2]. Our previous work [4-5] also demonstrates that building robust and intelligent disaster information extraction and analysis platform can help the public and private sectors work together to apply world class computing tools to deliver the right information to the right people at the right time.

Needs for heterogeneous information integration in disaster management domain: People have been firmly convinced that the use of timely, accurate and effective disaster information can significantly facilitate the disaster recovery process. Typical data resources include news/articles/blogs from web, announcements from governments, business reports from company participants, social media snippets and multimedia data like images and videos. However, information management and processing in disaster management are particularly challenging because of miscellaneous information resources that are publicly available and the unique combination of characteristics of those data, including: a great amount of information production and consumption; time sensitivity of the exchanged information; level of trustworthiness of the information sources; lack of common terminology; and heterogeneous formats [3]. However, very few information integration tools have been developed in disaster management tasks.

Growth of vertical search engine in various domains: General-purpose search engines, such as Google, Yahoo, or Bing have shown their efforts to exhaustively grasp all possible information from the giant web. However, the drawback of the general search strategy is the obviously overwhelming ambiguous and irrelevant information when digging in for a specific topic. Vertical search engine, also called domain-specific search engine, has been deemed as a powerful and necessary complementary tool to overcome those shortcomings. A well-established vertical search engine can substantially improve the efficiency of users getting more insights about a certain topic (in both coverage and relevance) and it also can

IEEE IRI 2012, August 8-10, 2012, Las Vegas, Nevada, USA
978-1-4673-2284-3/12/$31.00 ©2012 IEEE

significantly save the cost for web crawlers in terms of time and storage.

Successful vertical search engines: There are quite a few successful vertical search engines currently serving various communities, such as Flight/Travel (SkyScanner), Law/Legal (FindLaw), BioInformatics (BioMed), and Academic Search (RefSeek). These search solutions focus on one area of knowledge creating customized search experiences and utilize existing knowledge from domain expertise. In disaster management domain, information collection and presentation platforms have also been implemented to gather disaster related information. For example, GeoVISTA [6] from Penn State created GeoTwitter [7] component to plot new tweets in real-time to support for situational awareness; OilReport [8] from Colorado University collected tweets related to oil spill and categorize those tweets by types of events. However, there is no previous work that can simultaneously handle information from heterogeneous resources (web, social media and government official reports, etc.) and systematically integrate resources together in disaster management domain.

1.1 Motivation for Integrated Disaster Information Analysis Tools

Our disaster management team at Florida International University has cooperated closely with experts and participants from South Florida Emergency Management and industry partnerships for over four years. We have designed and implemented a web-based prototype of disaster information sharing platform and a disaster situation awareness and community organization application running on iOS-based mobile devices utilizing the data processing power of advanced information technologies for disaster planning and recovery under hurricane scenarios [3-5]. They can largely help people discover, collect, organize, search and disseminate real-time disaster information.

This collaboration provides us with the opportunity to gain insight into the manner that South Florida public and private sector entities manage and exchange information in a disaster situation. The emergency managers and business continuity professionals eagerly desire that a powerful and intelligent data analysis system be specifically designed for disaster management domain and satisfy the information acquisition needs for different types of users including emergency management officers, business continuity participants, and other users without business recovery capabilities. They agreed that such a system can significantly help them facilitate their disaster management and recovery efforts. In order to efficiently and effectively deliver high-quality disaster related information, several interesting yet crucial information management issues have been brought up.

1. **Real-time disaster information.** People always prefer the latest news and situation reports related to their search interests when being affected by disasters. Timeliness is the most important requirement for information collecting and sharing system.
2. **Heterogeneous information resources.** News portal is no longer the only information sources in disaster situation. Some micro blog or social media applications make faster response when emergency happens. Information from various channels can largely accelerate the information discovery process.
3. **Diverse information presentation.** Textual results are no longer the standard information representing approach. News visualization methods focusing on combining different aspects or dimensions become more popular. For disaster events, geo-location is considered as one of the most important features for disaster preparedness and recovery, which can greatly improve information monitoring and organizing capabilities.
4. **Integrated information portal.** Users prefer information portal that offers a multitude of services to meet their information needs.

In summary, user requirements from professionals who have an operational responsibility in disaster situations have been converging gradually to a disaster information integration and analysis platform that is able to assimilate massive information and provide actionable information for decision support. Non-professional users also like a domain specific search tool that provides them with insight of disaster situations.

1.2 Research Challenges and Proposed Solutions

The following three key tasks have been identified to fully utilize the advantages and overcome the shortcomings of traditional general search and information management platform that have never been applied to disaster management domain.

1. Design and develop effective and dynamic concept hierarchy generation and reuse methods in disaster management domain to help the domain experts, the crawler and search engine behave efficiently in situation. Concept hierarchy, as means of formalizing and sharing knowledge, provides domain experts and knowledge engineers support for modeling specific domain of the world and can be applied in various areas to implement intelligent knowledge and information management system. However, building the hierarchy from scratch is a costly process that requires massive human labor, so automatically improving concept hierarchy generation and reuse becomes a challenging but critical task. Combining existing hierarchy with concepts extracted from Semantic Web contents largely helps to

extend and enrich existing structural concepts in a given domain.

2. Design and develop intelligent focused web crawling techniques to manage the data acquiring process and to increase the information coverage and relevance in disaster domain. Heterogeneous data collected from various sources bring difficulties to assimilate information at different levels. The strategies for general-purpose search engine will lead to many irrelevant web pages being indexed and also the seeds set will be expanded unexpectedly. Intelligent crawling strategies are needed to systematically control the crawling process to guarantee the indexed web contents with high quality and relevance. Also the given seeds can be expanded to a certain level and finally converge to a good seeds list. On the other hand, the query results are required to be personalized to remove duplicity and increase diversity.

3. Design and develop data integration techniques for disaster events identification and extraction. In disaster situation, many recovery processes are running in a confused mass. Undergoing activities and important situations are hard to detect from many information channels in unformatted patterns. How to understand the information and organize useful knowledge in a unified manner becomes especially helpful for government officials, disaster management agents, business continuity staff, and even public users suffer from disorders during disaster recovery phases. After getting related information from the web, particular techniques need to be designed to integrate the raw data into certain format that are ready to be used by the search engine and topic visualization modules.

Generally, to accomplish the above three tasks, Disaster SitRep utilizes the latest advances in database, data mining, and information extraction technologies, to create a user friendly, information-rich service web application in disaster management domain. In particular, to address dynamic concept generation task, we apply document hierarchical clustering with constraints to automatically expand and enrich the existing disaster concept hierarchy. To address focused crawling task, we create a focused crawler that is able to classify newly discovered web resources into different disaster concepts and also discover new concepts simultaneously. To address disaster data integration task, we use dictionary-based and rule-based named entity recognitions to extract disaster related events from massive document in heterogeneous formats.

In this paper, we design and implement a **Disaster Situation Reporting System (Disaster SitRep)** system. The rest of the paper is organized as follows. Section 2 presents the overview of Disaster SitRep system; Section 3 describes the disaster taxonomy generation in detail. Existing taxonomy is utilized as partially known concept hierarchy that can be used as constraints to update taxonomy; Section 4 discusses the focused crawler. We propose a taxonomy-based focused crawling component to automatically detect, parse and filter those relevant web resources; New concepts can be extracted from crawled documents to update exiting taxonomy; Section 5 describes the disaster data integration module for assimilating miscellaneous resources; Section 6 describes the system evaluation.

2. Disaster SitRep Overview

Disaster SitRep is an integrated platform specializing in disaster management domain. It provides a collection of disaster related search, integration, and visualization tools to deliver personalized search results based on specific user needs. The goal to design this system is to help the user efficiently identify important information, organize emergent resources, and understand current damage and/or recovery status.

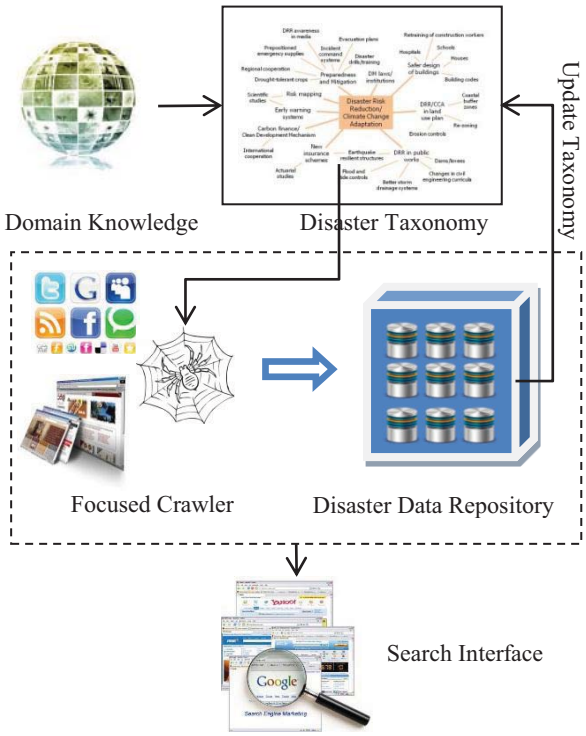


Figure 1. System architecture

Disaster SitRep has several major procedures to combine domain knowledge with disaster information from various resources and provide system users with high-quality disaster information related to their query interest. A hierarchy of disaster concepts is specified at the very beginning to help the focused crawler filter unrelated web pages. The focused crawler fetches web resources including web pages, documents and textual content from major social media. New concepts are generated after certain period of crawling and are used to

update previous concept hierarchy. At this time, heterogeneous data is integrated into a unified data repository for information search and visualization interface. Figure 1 illustrates the system architecture.

2.1 Disaster SitRep System Components

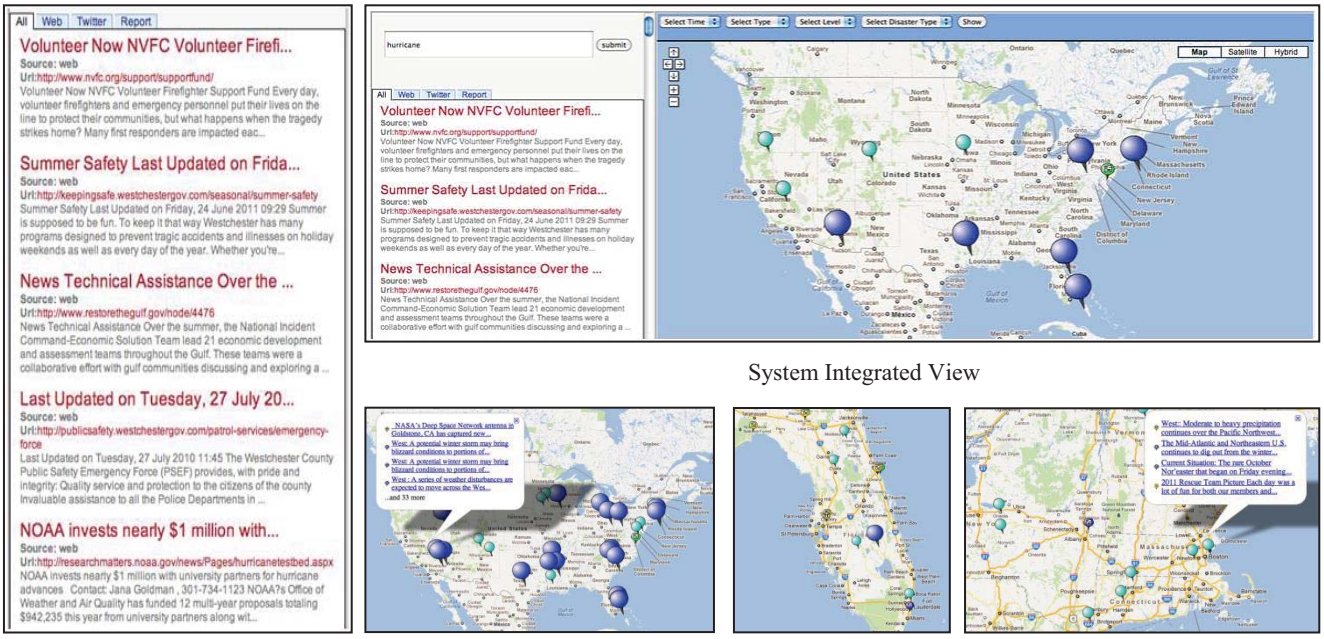
Based on the experience of developing our prototype we have designed two major components that are necessary for disaster information search and visualization:

used to give users a visual summary and corresponding textual summary [19, 20] of the disaster situation for corresponding query conditions.

We illustrate those important components in Figure 2.

2.2 Disaster SitRep System Architecture

Following our previous application framework, Disaster SitRep is designed and implemented to be a lightweight, comprehensive and fully Java implemented Web-based application. Our major information processing



Search Results | System Integrated View | Information Clusters – State Level | City Level -FL | City Level - NY

Figure 2. System components

Search Panel: A search component that supports keyword query provided by users returns a list of most important news crawled from various resources, including web, famous social media, and official government or company announcements and reports submitted to our previous business continuity web portal (www.bizrecovery.org). Results can be displayed based on different resource types or just in an integrated view.

News/Reports Map: Map in our system is used in two modes. Firstly, each of the search results associates with one or more points in map indicating the locations mentioned in the text. This allows users to visually know where those events happened and the geographic distributions of the events associated with a query. Secondly, the map provides a comprehensive view of all disaster information in our repositories. Also, such disaster information can be visually manipulated by utilizing different filters in map module. There are 4 filters including time, resource type, zoom level, and disaster type. Any combination of those filters can be

and representation functionalities are integrated with the following three critical modules: Taxonomy Generation, Focused Crawling and Disaster Event Extraction.

Taxonomy Generation: Based on our cooperation with domain experts, we initialize fundamental disaster taxonomy from disaster expertise. As the system keeps running, more web contents are crawled and extracted from unforeseen sources and new disaster terminologies are dynamically generated and are appended to the existing taxonomy. We propose a semi-supervised hierarchical clustering algorithm to enrich and modified previous taxonomy. Details of taxonomy generation and extension approaches are discussed in Section 3.

Focused Crawling: Our focused crawler is implemented to discover more disaster information by intelligently traversing the web contents based on their relevance to ongoing disasters. Usually, the more a web page is related to a certain topic, the higher probability it contains more resources (including hyperlinks to other web pages or possibly relevant concepts) in the same domain. The

disaster taxonomy in the previous stage can be utilized to classify web pages into various disaster categories. In general, there are two levels of judgments that help scoring the relevance of a page:

- *Web Page Classifier:* The classifier adopts hierarchical classification strategy to automatically categorize a crawled web page into different aspects according to the disaster taxonomy or simply report that current web page is irrelevant to any disaster topic.
- *Queue Prioritizer:* From the categorization results, the focused crawler adjusts the priority of each web page in the queue to guarantee that the most related web resource will be accessed earliest during the crawling process.

Combining these two functionalities, the focused crawling module attempts to assign the most relevant web page with the highest score to make sure such resource can be downloaded earliest. By properly designing those two parts, the crawler can access more related web resources by accessing fewer web pages. Also, as we crawl more disaster related content, it can largely contribute to extend our current taxonomy by including more concepts. Details are discussed in Section 4.

Disaster Event Extraction: Textual documents and situation reports crawled from the websites do not usually provide actionable information immediately, such as time, location, status, etc. The replication of information from various resources also challenges the search engine to provide highly related and diversified content to users. To gain further insight about the disaster event rather than a collection of textual documents, we need a domain-oriented skeleton for each type of disasters. The domain-oriented skeleton is the set of structural attributes that we try to extract from disaster documents. The details will be described in Section 5.

These modules are tightly integrated to provide a cohesive set of services including disaster information searching, querying, and visualization. Furthermore, they constitute a holistic effort on developing a data-driven solution for disaster management and recovery.

3. CONCEPT HIERARCHY GENERATION

Taxonomies or conceptual hierarchies play significantly important role in most knowledge-based information management systems applied in various application domains. They are designed to provide structurally organized terminologies that are formal, application-independent and with common agreement within a community of practice [9, 19]. However, generating taxonomy from the scratch suffers high-cost, low-efficiency problem. Ensembling several existing taxonomies or incrementally integrating new concepts into existing taxonomy becomes effective and well-accepted approach for taxonomy generation and reuse. In our taxonomy generation component, we model this problem as ***document hierarchical clustering with ordered constraints*** in which the constraints are given as a partially known hierarchy, the disaster related concepts extracted from web documents are treated as instances, and our goal is to build a term hierarchy which satisfies the relative hierarchical structure in given partial hierarchy.

3.1 Base Concept Hierarchy Generation

Our initial disaster taxonomy is built manually from the scratch. Based on our long cooperation with Miami-Dade Emergency Operational Center (EOC), we extracted hundreds of frequent terms in its official announcements and situation reports in the past 5 years. We reasonably assume that those terms with high frequency indicate important concepts in disaster domain. Through careful filtering and organizing those terminologies from our staff and developers, our initial disaster taxonomy is obtained and then verified by our domain experts.

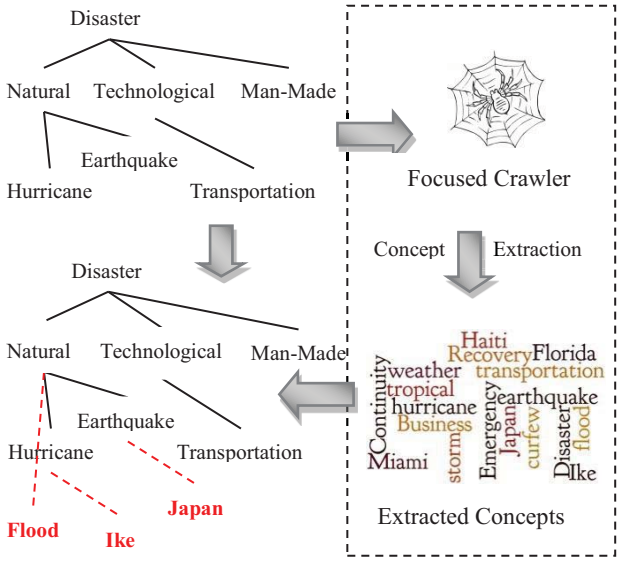


Figure 3. Iterative taxonomy generation

3.2 Iterative Taxonomy Generation

The disaster taxonomy generation process follows an interactive and iterative strategy. The focused crawler utilizes the taxonomy to classify accessed web pages and prioritizes those pages with highest relevance to disaster domain. From the repository of crawled data, high quality data will be analyzed and disaster-related concepts without being mentioned in the existing taxonomy, will

be extracted. Those extracted concepts are considered as highly popular terms that can extend and enrich the existing taxonomy. After integrating those newly-discovered concepts into disaster taxonomy, domain experts can verify the updated knowledge based and provide valuable feedback. Figure 3 shows the typical workflow of iterative taxonomy generation strategy.

3.3 Hierarchical Clustering with Constraints

Our aim is to build a hierarchical structure to model the basic human understanding of the relationships among disaster relevant concepts. A basic taxonomy/concept hierarchy is given at the very beginning of the generation process. In our work, we use agglomerative hierarchical clustering with constraint to algorithmically integrate newly-discovered terms or concepts into the existing ones.

3.3.1 Problem Definition. All concepts in existing taxonomy are denoted as $T = \{t_1, t_2, \cdots, t_n\}$ and the newly-discovered concepts are denoted as $C = \{c_1, c_2, \cdots, c_m\}$. H is the existing concept hierarchy formed by terms from T. Our goal is to generate an updated concept hierarchy H' that is formed by all terms from both T and C. The integration of T and C is non-trivial. There are three important aspects worth mentioning:

1. Each concept in T or C is represented by a set of terms extracted from the web documents repository. So, essentially there is a subset of web documents under each concept.
2. H is essentially a hierarchical clustering on all documents. The hierarchy of the concepts reflects the inclusion or exclusion of documents sets. There is no partial overlap between document sets under different concepts.
3. There is a merging preference/order for each pair of concepts in both H and H' which indicates the level of closeness between two document sets. The new concepts in C should not change the relative merging order of existing concepts in T. The details are given in the following section.

3.3.2 Algorithm and Partial Hierarchy Constraint. The merging preferences mentioned above are modeled as relatively ordered constraints when performing hierarchical clustering on document set. Constraints defined in hierarchical clustering are different from constraints, such as instance-level constraints [10] and prior knowledge [21] in partitional clustering. Several types of constraints that can be applied in hierarchical clustering are defined in the literature [11-13].

In our application, we use Bade's algorithm [11] to refine the given disaster concept hierarchy by considering further extracted concepts. The constraint in [11] is named must-link-before (MLB), shown in Figure 4, which specifies the order in which objects are linked. When applied to concept hierarchy, such order indicates the merge preference between concepts (document sets). Bade's algorithm [11] can utilize the existing concept hierarchy as partially known hierarchy and update it by directly attaching newly-discovery concepts to previous hierarchy. The other two methods do not meet our needs because updated hierarchy requires to be built from the scratch.

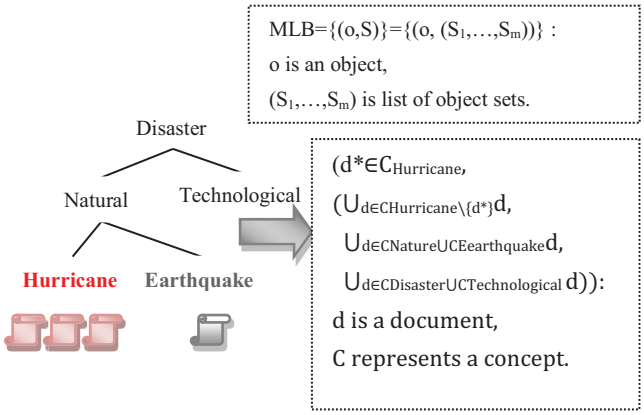


Figure 4. MLB constraints from partial hierarchy

4. FOCUSED CRAWLER

We adopt focused crawling technique to retrieve the disaster aware information in the Web. In addition, contents also come from subscription of some local news feeds and monitoring announcement from government sites. Compared with a standard focused crawler defined in [14, 15], there are some challenges in our problem.

Loose cohesion: Except large disaster events accompanied with intensive reports, most disaster information is scattered in the Web. In news websites, stories about disasters may embed in other types of news. As for government sites like county emergency management homepages, they are more likely to link to websites of the county's other departments than that of another county's emergency management homepage.

Diversity of disaster topic: Disasters we are interested in include many subtopics, from various types of disasters to four different phases of emergency management, it is difficult to evaluate a web page's relevance on a consistent scale among all these subtopics. It is very likely that the crawled data will bias towards some of the subtopics and leave some others uncovered.

To address the above issues, we utilize the concept hierarchy we developed.

4.1 Selection Strategy

Best-first approaches are widely used by focused crawlers, selecting the next page to be crawled from all currently assessed candidate page URLs by their scores as

$$l^* = \text{argmax}_{l \in queue}\, score(l)\ ,$$

where $score(l)$ is calculated based on a classifier indicating whether or not the URL l belongs to the topic. However, the "best" may bias to some of the subtopics of general disaster topic because of the unbalance of these subtopics and a limited initial training dataset. To get a set of web pages with high diversity for a specific disaster, we simultaneously crawl web pages for each disaster concept based on the concept hierarchy. Our selection strategy considers a disaster concept:

$$l_C^* = \text{argmax}_{l \in queue}\, score(l, C),$$

that is, for each disaster concept, select the next page to be crawled from all currently assessed candidate page URLs according to their scores with respect to the concept.

Figure 5. An example page of hurricane Irene.

4.2 Prioritization Based on Concept Relationship

For a web page, instead of classifying it into "Disaster" and "Non-disaster", we assigned to it a concept in our concept hierarchy, such as "weather", "government" and "environment protection". These disaster related concepts increase the coherence of the Web pages of disaster topic, playing a role of bridging between pages of different sites of disaster concepts and pages of different disaster concepts. To calculate the prioritization score of a URL, the concept of the page from which the URL is linked is utilized as follows:

$$score(l, C_d) = P(C_i^* \rightarrow C_d) * P(page_l = C_i^*),$$

where $\mathrm{P}(\mathrm{page_l} = \mathrm{C_i^*})$ is the output of our content classifier indicating the probability the page where the link l is linked from belongs to its optimal concept $\mathrm{C_i^*}$, and $\mathrm{P}(\mathrm{C_i} \rightarrow \mathrm{C_d})$ is the link relationship between concepts, the probability that a page of concept $\mathrm{C_i}$ links to a page of concept $\mathrm{C_d}$. It can be calculated as

$$P(C_i \rightarrow C_d) = \frac{\sum_{p \in C_i} |L_{p,fetched} \cap C_d| + \lambda}{\sum_{p \in C_i} |L_{p,fetched}| + \lambda \sum_{p \in C_i} |L_{p,unfetched}|}\ ,$$

the ratio of the number of links classified as $\mathrm{C_d}$ from pages of $\mathrm{C_i}$ to the number of all fetched links from pages of $\mathrm{C_i}$, with a Dirichlet smoothing using unfetched links. Note that with the process of crawling, $\mathrm{P}(\mathrm{C_i} \rightarrow \mathrm{C_d})$ is being updated, so that the scoring of links is also adaptive with more data crawled.

4.3 Link Prediction

Although a page is disaster relevant, the links of the page may not necessarily lead to other pages of disasters. Figure 5 shows an example page.

To further distinguish the links in a page, a link classifier is trained, using the prediction of the content classifier for crawled pages as training data. The rationale is that many links contain a description of the content of the linked page. Another observation we find is about link structure, that for a pair of link which are in the sibling nodes of the HTML DOM tree, e.g. in a list of the page, they tend to be of a similar topic. We follow the work of [16] and build a link classifier based on Native Bayes. To apply the link prediction:

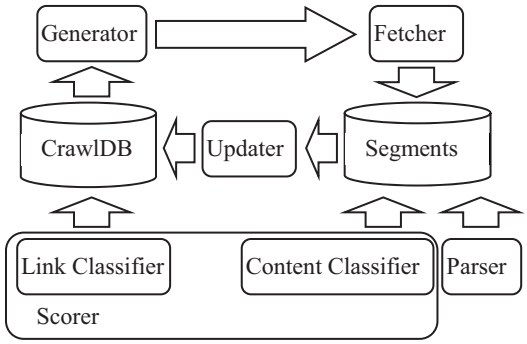


Figure 6. Architecture of the focused crawler.

1. The prioritization score can be extended as:
$score(l, C_d) = P(C_i^* \rightarrow C_d) * P(page_l = C_i^*) * P(C_d|l)$, where $P(C_d|l)$ is the output of the link classifier, probability that link l leads to a page under concept C_d.

2. To reduce the redundancy, we first divide the links into clusters, and constrain the crawler such that links in the same cluster are not fetched at same time. Once a link is fetched, the prediction of links in the same cluster will be updated.

4.4 Architecture of the focused crawler

We build our crawler based on Nutch[17], which is a distributed general crawling tool running on Hadoop[18]

clusters. We customize the scoring module and generator module in Nutch. The current architecture is shown in Figure 6. In each iteration, the Fetcher fetches page content of a list of URLs, and stores them as a segment. The updater updates CrawDB, where the crawled data is associated with a URL. The scoring module assigns a prioritization score to each URL indicating the importance of the URL. The generator module generates a set of URL, covering all disaster concepts in the concept hierarchy. The Fetcher fetches the web page content.

5. DISASTER EVENT EXTRACTION AND INTEGRATION

We hope to gain further insight about the disaster event rather than create a collection of textual documents. For example, the location and the date time of a storm, the status of the electrical power impacted by an earthquake and so on. These are key domain-oriented information of disasters. In our vertical search engine, the rank of the search results is mainly based on this domain-oriented information. On the other hand, when a disaster happens, a huge amount of news, situation reports, and announcements will burst in a very short time. Most of the documents have replicated content. To eliminate replicated content and integrate all related documents, we need a domain-oriented skeleton for each type of disaster. The domain-oriented skeleton is the set of structural attributes that we try to extract from disaster documents.

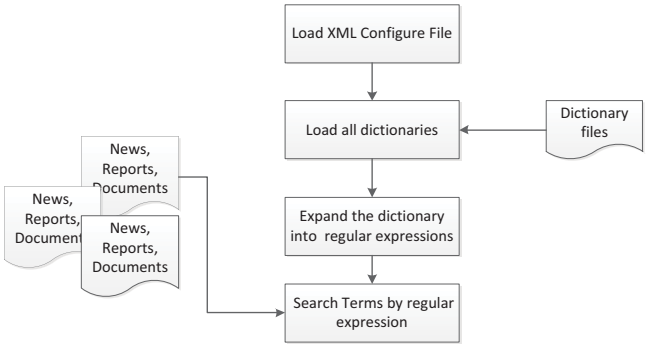


Figure 7. Workflow of event extractor

5.1 Event Extractor

Dictionary-based and Rule-based named entity recognitions are utilized in this part. We do not use probabilistic and training based name entity recognition because that approach requires people to label the web documents word by word, which is time-consuming. Each domain-oriented disaster attribute (e.g. the date time, the location, the power status) is defined as a rule in a XML configuration file. The rule is similar to the regular expression but combines more functionality. It can include a dictionary given by the user and supports approximate word matching by edit distance.

For each concept, we have particular event attributes for extraction. The event attributes are also associated with the extraction rules in the configuration file. Each concept owns the attributes of its ancestors. The work flow for event extractor is shown in Figure 7.

5.2 Event Integration

The disaster event integration is based on the similarity table join. We consider each extracted disaster event as a database record. Each web document corresponds to a database record. The problem of event integration is how to join these database records. In our disaster event integration, we apply the similarity join, which does not require two identical attribute values. It considers the overall similarity of the values of common attributes. If common attribute values have a similarity greater than a threshold, the two records can be joined. If the threshold is 1, the similarity join becomes the traditional equal join.

Let a disaster event consist of n attributes, x and y be two extracted event records. x_i denotes the i-th attribute value of x, i=1,…n. The overall similarity for record joining is defined as follows:

$$\mathrm{sim}_{\mathrm{all}}(x,y) = \frac{\sum_{i=0}^{n} I(i,x)I(i,y)w_i f_i(x_i,y_i)}{\sum_{i=0}^{n} I(i,x)I(i,y)w_i}$$

where I(i,x) is a binary variable that I(i,x)=1 if x has the i-th attribute value, otherwise I(i,x)=0. wi is the weight for the i-th attribute. f_i is the predefined similarity function for the i-th attribute of the event.

Figure 8 shows an example of the joining two records which are extracted from two different web news. W1 is a situation report from an official web site such as FEMA, and W2 is a news report from a local news website. The common attributes of W1 and W2 have identical values, so they are joined to a record W*.

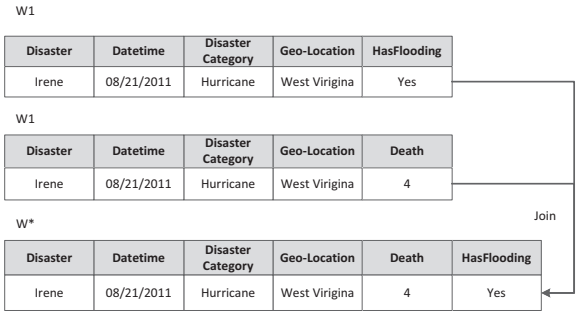


Figure 8. Event record join

Many existing studies on web documents apply textual clustering based methods. Then, the information shown to the users is the representative document of each cluster.

This approach is not quite accurate because it does not consider the content meaning. For instance, it does not distinguish the name of the disaster, the date time of the event, and the geo-locations and other key attributes of disaster reports with other document words. Therefore, our disaster event integration is based the extracted events rather than the original textual documents.

6. DATA COLLECTION AND EVALUATION

To evaluate the algorithms used in our system we use standard performance metrics used in the research literature and carefully compare our algorithms with existing work when applicable.

Our system evaluation process consists of presenting the system to our community of emergency managers, business continuity professionals and other stakeholders for feedback and performing community exercises. The community exercises involve a real time simulation of a disaster event and are integrated into an existing exercise that the community conducts for readiness each year. This evaluation exposes information at different time intervals and asks the community to resolve different scenarios by using the tool developed. The evaluation conducted takes on the form of a "table-top" exercise in which information injects provide details about the current disaster situation and specify potential goals and course of action. In return, the participant uses the system to gather information to best assess the situation and provide details about the actions to be taken. We gather information from the user about what information they found to derive their conclusions or lack thereof. This information allows us to better understand how our techniques overall improve the information effectiveness.

Feedback from our users are overwhelmingly positive and suggest that our system can be used not only to share the valuable actionable information but to pursue more complex tasks like business planning and decision making. There are also many collaborative missions that can be undertaken on our system which allows public and private sector entities to leverage their local capacity to serve the recovery of the community. Our initial work has been recognized by FEMA (Federal Emergency Management Agency) Private Sector Office as a model in assistance of Public-Private Partnerships.

7. ACKNOWLEDGMENTS

This work is supported by NSF grants HRD-0833093 and CNS-1126619, and DHS grants 2009-ST-062-000016, 2010-ST-062-000039, and VACCINE/DHS 4112-35822. We thank Jesse Domack and Jason Clary for their work in the system development and testing.

8. REFERENCES

[1] The Conference Board. Preparing for the worst: A guide to business continuity planning for mid-markets. *Executive Action Series*, February 2006.

[2] R. Berg. Hurricane Ike Tropical Cyclone Report. NHC. Retrieved 2009-09-12.

[3] V. Hristidis, S. Chen, T. Li, S. Luis, and Y. Deng. Survey of data management and analysis in disaster situations. *The Journal of Systems and Software*, 83:1701–1714, 2010.

[4] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S. Chen, and V. Hristidis. Using data mining techniques to address critical information exchange needs in disaster affected public-private networks, *KDD '10*, pages 125–134, 2010.

[5] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, and S. Chen. Applying Data Mining Techniques to Address Disaster Information Management Challenges on Mobile Devices. *In Proceedings of KDD 2011*, pages 283–291.

[6] GeoVista. http://www.geovista.psu.edu/.

[7] A. M. MacEachren, A. C. Robinson, A. Jaiswal, S. Pezanowski, A. Savelyev, J. Blanford. Geo-Twitter Analytics. *Applications in Crisis Management 2011*. Paris, France.

[8] OilReport. http://www.cs.colorado.edu/~starbird/oilreport_map.html.

[9] E. Simperl: Reusing ontologies on the Semantic Web: A feasibility study. *Data Knowl. Eng*. 68(10): 905-925 (2009).

[10] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained K-means Clustering with Background Knowledge. *In Proceedings of ICML 2001*, pages 577-584.

[11] K. Bade and A. Nrnberger. Creating a cluster hierarchy under constraints of a partially known hierarchy. *In SDM 2008*, pages 13–24.

[12] L. Zheng and T. Li. Semi-supervised Hierarchical Clustering. *In ICDM 2011*, pages 982 - 991.

[13] H. Zhao and Z. Qi. Hierarchical agglomerative clustering with ordering constraints. *In WKDD 2010*, pages 195–199.

[14] S. Chakrabarti, M. van den Berg, B. Dom. Focused Crawling: A New Approach to Topic Specific Resource Discovery. *In WWW 1999*.

[15] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the World Wide Web with arbitrary predicates. *In WWW '01*. ACM, 2001.

[16] D. Ahlers and S. Boll. 2009. Adaptive geospatially focused crawling. *In Proceeding of the 18th ACM conference on Information and knowledge management*, pages 445–454.

[17] Nutch. http://nutch.apache.org/

[18] Hadoop. http://hadoop.apache.org/

[19] L. Li, D. Wang, C. Shen, and T. Li. Ontology-Enriched Multi-document Summarization in Disaster Management. *In Proceedings of SIGIR 2010*.

[20] D. Wang, L. Zheng, T. Li, and Y. Deng. Evolutionary Document Summarization for Disaster management. *In Proceedings of SIGIR 2009*.

[21] T. Li, Y. Zhang, and V. Sindhwani. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009), Pages 244-252, 2009.