Using Data Mining Techniques to Address Critical Information Exchange Needs in Disaster Affected Public-Private Networks

Li Zheng, Chao Shen, Liang Tang, Tao Li, Steve Luis, Shu-Ching Chen, Vagelis Hristidis

School of Computer Science

Florida International University

11200 S.W. 8th Street, Miami, Florida, 33199, U.S.A.

{lzhen001, cshen001, ltang002, taoli, luiss, chens, vagelis}@cs.fiu.edu

ABSTRACT

Crisis Management and Disaster Recovery have gained immense importance in the wake of recent man and nature inflicted calamities. A critical problem in a crisis situation is how to efficiently discover, collect, organize, search and disseminate real-time disaster information. In this paper, we address several key problems which inhibit better information sharing and collaboration between both private and public sector participants for disaster management and recovery. We design and implement a web based prototype of a Business Continuity Information Network (BCIN) system utilizing the latest advances in data mining technologies to create a user-friendly, Internet-based, information-rich service and acting as a vital part of a company's business continuity process. Specifically, information extraction is used to integrate the input data from different sources; the content recommendation engine and the report summarization module provide users personalized and brief views of the disaster information; the community generation module develops spatial clustering techniques to help users build dynamic community in disasters. Currently, BCIN has been exercised at Miami-Dade County Emergency Management.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; H.3.3 [**Information Search and Retrieval**]: Clustering; H.3.5 [**Online Information Services**]: Web-based services; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Management

Keywords

Data Mining, Critical Information, Disaster Management, Information Extraction, Dynamic Dashboard, Multi-Document Summarization, Spatial Clustering

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

It is well known that hurricanes, earthquakes, and other natural disasters cause immense physical destruction, loss of life and property around the world. Annually, the gulf coast states are especially susceptible to hurricane impact. In 2004-2005, Florida was struck eight times by hurricanes that caused enormous economic damage. An untold amount of time and effort was also spent on repeated preparation and recovery efforts causing fatigue and anxiety in the affected communities.

To better understand hurricane behavior and impact, a significant amount of scientific study is being conducted to improve the predictive ability of hurricane relevant atmospheric and hydrological models, or to develop technologies that improve the quality of building materials and building codes. However, despite the lessons learned in a succession of storms such as Andrew, Katrina, and Wilma, there is very little knowledge of how to collect, manage, find, and present disaster information in the context of disaster management phases: Preparation, Response, Recovery, and Mitigation.

In recent years, Federal Emergency Management Agency has recognized the importance of the private sector as a partner in addressing regional disasters. The State of Florida Division of Emergency Management has created a Business and Industry Emergency Support Function that is designed to facilitate logistical and relief missions in affected areas. Four counties, Palm Beach, Broward, Miami-Dade and Monroe, which constitute the Southeastern population of South Florida and include over 200,000 business interests, are developing Business Recovery Programs to help facilitate quicker business community recovery through information sharing and collaboration.

For over three years, a multi-disciplinary team of researchers at Florida International University with South Florida Emergency Management and industry partners, such as Wal*Mart, Office Depot, Wachovia, T-Mobile, Ryder Systems, IBM, and others have investigated the way South Florida public and private sector entities manage and exchange information in a disaster situation. According to our preliminary work on our partner network in disaster management situations, the efficiency of sharing and management of information plays an important role in the business recovery [4]. For instance, it is critical that companies receive information about their facilities, supply chain, and city infrastructure. They seek this information from media outlets like television/radio newscasts, employee reports, and speak with other companies they have a relationship with. With so many

sources of information at granularities, with different levels of redundancy and accuracy, possibly generated by different varieties of reports (structured and unstructured), it is difficult for companies to quickly assimilate such data and understand their situation.

Furthermore, we have learned that large-scale regional disaster causes a disruption in the normal information flow and channels, and the relationships between information producer and consumer. Effective communication is critical in a crisis situation. What is not very well known is how to effectively discover, collect, organize, search and disseminate real-time disaster information.

In this context, we design and implement a Web-based prototype implementation of a Business Continuity Information Network (BCIN) that is able to link participating companies into a community network, provide businesses with effective and timely disaster recovery information, and facilitate collaboration and information exchange with other businesses and government agencies. The project website is http://www.bizrecovery.org.

Based on observations we have made during our preliminary research, we have identified several key problems that inhibit better information sharing and collaboration among both private and public sector participants for disaster management and recovery. In this paper, we will focus on these problems.

1. How can the system quickly capture the status report information? Participants will communicate status reports through many channels, including direct emails, mailing lists, web pages, press releases, conference calls. It is desirable to capture such status information the minute it is available and prevent redundant reporting. To facilitate the reuse of such materials, users can upload status information in the form of unstructured documents such as plain text, Adobe PDFs, and Microsoft DOC. It is thus necessary to identify the useful information in the documents.

2. How can the system effectively understand the situation from a large collection of reports? In larger organizations, or in cases where there is a large accumulation of companies in an area, like a corporate park, reports about a particular area can be redundant. It is important to build a summarized view to understand the situation users are interested from these reports.

3. How can we automatically capture users' interests and effectively deliver the relevant information to the users? The status reports are collected through many different channels and are concerned about different categories. During disaster preparation and recovery, users typically don't have the time and patience to go through the system to find the information they want.

4. How can we take advantage of the community information for disaster recovery? Participating companies and organizations interact in different communities, such as being members of the same industry sector, or using the same shipping company. Identifying how participants interact with these communities in a disaster situation is very important since it may reveal information that would be helpful in a recovery scenario.

To pursue solutions for these challenging problems, BCIN utilizes the latest advances in database, data mining, and information extraction technologies to create a user-friendly, Internet-based, information-rich service and acts as a vital part of a company's business continuity process. BCIN is a platform of information sharing, integration, extraction, and processing for disaster management and recovery. It is also a data mining solution for disaster management and recovery that is able to process and analyze the data from diverse and heterogeneous information sources of different types (categorical events and continuous data) with different formats (structured and unstructured: database records, document news, reports).

In particular, to address Problem 1, we apply information extraction to automatically extract the status information from documents. To address Problems 2 and 3, we create a user interface capability called the Dynamic Dashboard to improve information quality at matching the user's interests, and use document summarization techniques to give users a quick access to multiple reports. To address Problem 4, we adopt spatial clustering techniques to track assets like facilities, or equipment, which are important to participants. The geo-location of such participants can be organized into dynamic communities, and these communities can be informed about events or activities relevant to their spatial footprints.

Our analysis shows that, if our system helps 5% of the companies in the South Florida area to speed up their hurricane recovery by 1 week, it will prevent $219,300,000 of non-property economic losses which would result from that week's closure.

The rest of the paper is organized as follows. Section 2 presents an overview of the BCIN system. Sections 3-6 describe BCIN's four key modules (information extraction module, dynamic dashboard, summarization module, space clustering module) in detail, respectively. Section 7 describes the system development and operation, and finally Section 8 presents the evaluation exercises BCIN has participated and concludes the paper.

2. BCIN OVERVIEW

While there is no shortage of disaster planning or recovery toolkits, disaster information or news portals, emergency telecomm services, etc., there is no single technology, solution, or service that answers pressing business questions like, "What's the damage to my company's facility and surrounding area? Are utility, transportation, and fuel services available?" We have developed our prototype BCIN system to help answer those questions.

2.1 Concepts in BCIN

Based on the experience of developing our prototype we have established a series of abstractions necessary for modeling information management in our application. These concepts are listed below:

Entity: An entity in BCIN represents anything of interest in which we can inquire status during a disaster, including companies (like Wal*Mart and T-Mobile), assets of companies (like a retail outlet), resources the companies are providing, facilities (like bus and seaports), public service (like power and water), and even orders (like curfew). Most entities are associated with geographic information and can be annotated by a point or a range on the map. The sharing of entity status information is one of the most important motivations' of our approach.

Report: Users share their information by reports, which are mainly about the status of the entities that the users are related with. Reports from County Emergency Management Offices through an Emergency Operation Center, activated when there is a pending storm threat, are referred to as EOC reports. And reports from companies are called company reports and messages, the

difference of which is that company reports are accessible to users via group and role based access controls while messages are only received by specified target users.

Community: Community is defined as a set of entities that have a specific relationship between each other. There are two kinds of communities: static and dynamic. A static community is a type of entity set whose relationships are pre-existing in the system (before a storm impact), whose relations among entities of the same set typically do not change. For example, a static community is the list of companies that participate in a type of industry. A dynamic community is a type of entity set in which existing relations among entities may break, new relations may establish and new entities may occur as the situation changes. Dynamic communities are designed to help users discover potential collaborations under disaster recovery scenarios. For example, in the post-disaster phase, hundreds of companies open or close every week, with many reports and messages submitted every day. It is difficult for a user to browse all the different types of reports gathered in the system, and thus providing reports related to the user's interest is highly desirable. In other words, the reports producers and consumers are forming a community during a certain period of time.

Dashboard: The dashboard of our system is a quick view of the current entity information for users. When a disaster event occurs, users are pressed for time and dealing with personal and business crisis situations, requiring immediate status information. Users will not have the patience to click the different sections in the system and find the information they seek. Therefore, the aim of the dashboard is to predict the most needed information by different users and to display that information directly. In our BCIN project, we have developed 4 different dashboards, each of them uniquely addressing information based on feedback obtained from our business users:

Situation Dashboard: The screenshot of the situation dashboard is shown in Figure. 1. In the table of this figure, each column represents a facility and each row represents a jurisdiction. The status of each facility at each jurisdiction is shown as a sign with a color. For example, a green cycle indicates the opening status and a red cycle indicates the closed status.

Thread Dashboard: The threat dashboard is designed to display the latest official advisories from National Hurricane Center (NHC). For example, Figure 2 shows a short description about a fictitious hurricane called "Tony" and is published on Thu Nov 13, 2007 at 7:00 AM. In this example, the storm named "Tony", whose center is 720 Miles away from Fort Lauderdale is expected to impact the city with tropical storm force winds by Thursday, Nov. 15, 2007 at 1:00 AM.

Event Dashboard: The event dashboard is a calendar that displays recovery event activity reports on a grid for each day in a month. Figure 3 is a screenshot of this dashboard.

Company Dashboard: The company dashboard displays a list of companies' news and status. The collection of companies is composed of two categories. The first category is determined by the selections provided by the user. Users can change their selections as needed to reflect their interests in following different companies displayed in the dashboard. The second category is the recommended companies by the system that the user may be interested in. Figure 4 shows an example of the Wal-Mart's status in Miami. Each record in that table shows a status message of a specific Wal-Mart asset located in Miami.

2.2 Architecture of BCIN

Figure 5 shows the system architecture of BCIN. The system allows company users to submit reports related to their own business, and government users to make announcements on the public issues. To collect more information during the disaster, BCIN can monitor the news published on the websites and takes the news as its input. Like traditional information systems, these reports and news, and the status information of entities they contain can be retrieved and accessed by queries. For example, reports can be viewed according to alert categories or geo-locations, and resources can be viewed according to status or usages. Furthermore, BCIN not only displays users-submitted information but also conducts necessary and meaningful data processing work. BCIN makes recommendations based on the current focus and dynamically adapts based on the users' interests. BCIN summarizes reports and news to provide users with brief and content-oriented stories, preventing users from being troubled when searching in huge amount of information. By introducing the concept of Community, BCIN offers users a hierarchical view of important reports or events around them.

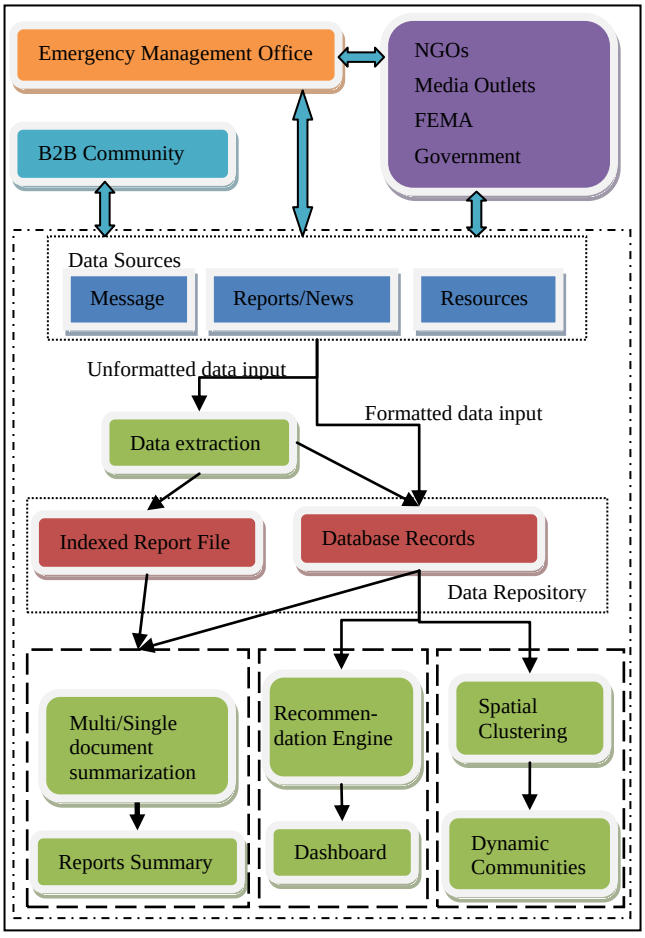

Figure 5. BCIN information processing.

In this paper, we discuss the following four main information processing and representation components: Information

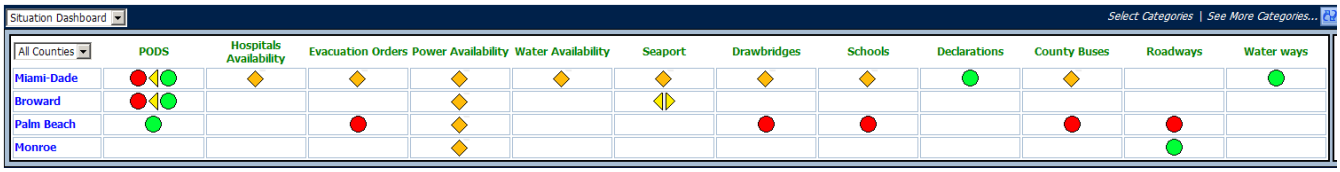

Figure 1. Situation dashboard.

Figure 2. Threat dashboard.

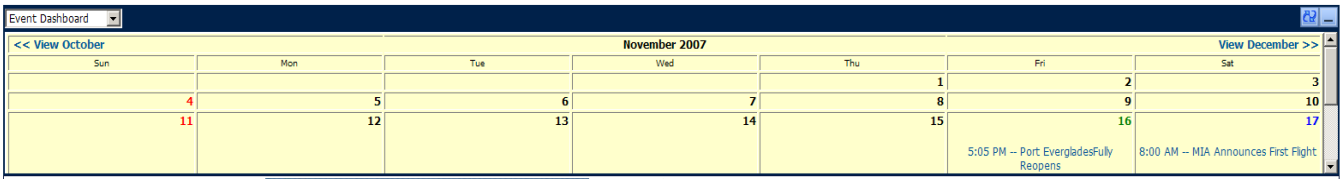

Figure 3. Event dashboard.

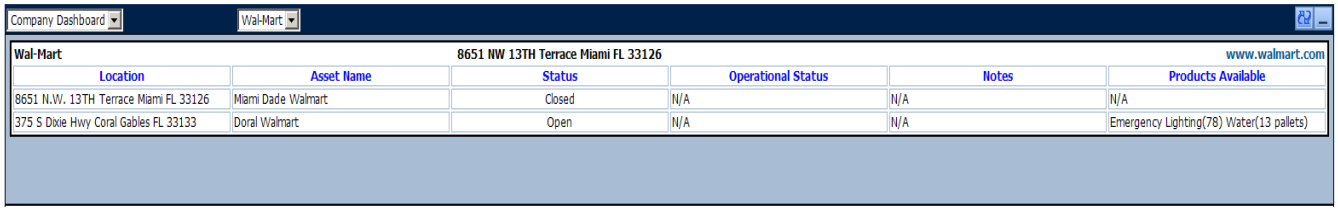

Figure 4. Company dashboard.

Extraction, Dynamic Dashboard, Report Summarization, and Dynamic Community Generation. These four components utilize and develop data mining and machine learning techniques and apply them to disaster management and recovery.

Information Extraction: As a data pre-processing component, we adopt sequence tagging and classification methods to extract the structured information from text to integrate different input without a unified format. The detailed approaches used for information extraction are presented in Section 3.

Dynamic Dashboard: In order to improve the relevance of information to match the user's interests we have created a user interface capability called the Dynamic Dashboard. The dynamic dashboard analyzes user interactions such as what kinds of reports the participant has submitted and viewed and automatically recommends similar information to display on the dashboard. The dynamic dashboard provides with the users a convenient and fast approach to obtain the disaster information that they probably want during the emergent time. The dynamic dashboard's content is personalized with the concerns of different users. The main contributions of the dynamic dashboard lie in two parts. 1) It automatically removes the redundant companies' reports, news and other information by clustering methods. 2) It ranks the information by both the relevance to the current user and the importance of information. The details are discussed in Section 4.

Report Summarization: the BCIN system provides users a report summary which is generated from multiple reports to show the updated changes about the process of the disaster. In the summarization process, structured information extracted from text and stored in the database is used to generate the summary to reflect the latest and changed status of an entity. Details of summarization approaches are discussed in Section 5.

Community Generation: Participating companies and organizations interact in different communities, such as being members of the same industry sector, or using the same shipping company. Identifying how participants interact with these communities in a disaster situation is very important since it may reveal information that would be helpful in a recovery scenario. Using spatial relationship techniques we can track assets like facilities, or equipment, which are important to the participants. The geolocation of such participants can be organized into dynamic communities, and these communities can be informed about events or activities relevant to their spatial footprints. By generating dynamic communities, users can directly select those events happening around them and make more efficient and accurate decision. We adapt spatial clustering algorithms in an interactive way to provide users a multilevel view of related communities. The details are described in Section 6.

These different components are tightly integrated to provide a cohesive set of services and constitute a holistic effort on developing a data-driven solution for disaster management and recovery.

3. STRUCTURED INFORMATION EXTRACTION FROM REPORTS

The BCIN system is an information sharing system for companies and government agencies. To provide a user-friendly interface to all these users, we do not request a unified format for them to submit the reports. Instead, we use information extraction methods to integrate reports from different sources. For example, Table 1 shows an example of EOC reports.

Table 1. An example of EOC report.

Time: October 21, 2005 12:30 p.m.
Miami-Dade Emergency Operations Center is currently activated at a level II and officials and emergency managers are carefully monitoring Hurricane Wilma. Residents are urged to finalize their personal hurricane preparations. On Monday, October 24, Miami-Dade County offices, public schools, and courts will be closed. Currently, transit bus and rail service continues, including Metrobus, Metrorail and Metromover. Miami International Airport is open. However, if you have travel plans please check with your airline for flight information. Tomorrow afternoon, the American Red Cross will open hurricane evacuation centers for residents who do not feel safe in their homes or live in low-lying areas.

In the BCIN system, the key information is "What was/is/will be the status of Facilities/Services/… at the time of …". From the EOC reports, we need to extract such information in the form of a triple: (entity, time, status), which reveals the status information of the entity at a certain time. In EOC reports, the entity may be a facility or public service like "Miami International Airport", "schools", "bus", and an order like "curfew". If the entity is referred to an order, the triple means whether the order is in effect or not at that specific time. We extract these triples through two steps: first, we extract entities and time expressions, then, we classify a pair of (service, time) to a proper category, "no relation"/ "open" / "close" / "unclear". We assume that the information of one event will not span on different sentences, so we process every sentence individually to extract an event. To extract those triples, both entity and relation extraction will be performed.

3.1 Entity Extraction

For each report, sentence segmentation is conducted first, and each sentence is POS-tagged. To extract entities and time expressions, we manually label some news and train a linear chain conditional random fields (CRF) model to tag all words of sentences, using "BIO" annotation [7,8]. A word tagged as [TYPE-B]/[TYPE-I] means it is the beginning/continuing word of the phrase of the TYPE, and the ones tagged as O means it is not in any phrase. Here TYPE can be E or T, referring to the entity and time expression. Using CRF, given the sentence X, the probability of its tags Y is as follows:

$$p(Y|X) = \frac{1}{Z_X}\exp\left(\sum_{i,k}\lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l}\mu_l g_l(y_{i-1}, y_i, X)\right),$$

where Z_X is the normalization constant that makes the probability of all state sequences sum to one; $f_k(y_{i-1}, y_i, X)$ is an arbitrary feature function over the entire observation sequence and the states at positions i and $i\text{-}1$ while $g_l(y_{i-1}, y_i, X)$ is a feature function of the states at position i and the observation sequence; λ_k and μ_l are the weights learned for the feature functions f_k and g_l, reflecting the confidence of feature functions by maximum likelihood procedure. The most probable labels can be obtained as

$$Y^* = argmax_Y P(Y|X)$$

by Viterbi-like dynamic programming algorithm[7]. We use for features the local lexicons and POS tags, and plus the dictionary composed of the existent entity names in the database.

Table 2 shows the result of the entity extraction result of the report in Table 1.

Table 2. Entity extraction result of the report in Table 1.

Miami-Dade Emergency Operations Center is currently activated at a level II and officials and emergency managers are carefully monitoring Hurricane Wilma. Residents are urged to finalize their personal hurricane preparations. On <T>Monday, October 24</T>, <E>Miami-Dade County offices</E>, <E>public schools</E>, and <E>courts</E> will be closed. <T>Currently</T>, <E>transit bus</E> and <E>rail service</E> continues, including <E>Metrobus</E>, <E>Metrorail</E> and <E>Metromover</E>. <E>Miami International Airport</E> is open. However, if you have travel plans please check with your airline for flight information. <T>Tomorrow afternoon</T>, the American Red Cross will open <E>hurricane evacuation centers</E> for residents who do not feel safe in their homes or live in low-lying areas.

Table 3. Features used to classify whether the entity e is associated with the time expression t.

DistanceBetween (e, t)
WordBetween(e,t)
TenseOf Sentence(e,t)
NegativeVerbsInSentence(e,t)
PositiveVerbsInsentence(e,t)
ContainDate(t)
PrepositionBefore(t)
FromDocument(t)

3.2 Relation Extraction

If a sentence contains an entity but no time expression, the time associated with the report will attached to the end of the sentence. To generate the triple by connecting the entity with the time expression with a proper status label, we train a multi-category SVM [9] to classify each pair of (entity, time) to a proper category, "no relation"/ "open"/ "close"/ "unclear". Table 3 shows the features we used for classification. Among them,

TenseOfSentence(s,t), NegativeVerbsInSentence(s,t) and PositiveVerbsInsentence(s,t) are extracted by heuristic rules to indicate the tense of the sentence, the verbs with and without negative modifier semantically in the sentence, respectively. Note that FromDocument(t) indicates whether the time is the time associated with document or not.

Finally, we extract those pairs of entity and time expression in the "open" or "close" categories to form the triple. Meanwhile the time expressions are formatted into an absolute form of expression from relative time expression such as "next Monday", "this afternoon" and etc. using the time of report as a benchmark. The structured information extracted from the report in Table 1 is shown in Table 4.

Table 4. Information extracted from the EOC report shown in Table 2.

Service	Time	Status
Miami-Dade County offices	October 24, 2005	close
public schools	October 24, 2005	close
courts	October 24, 2005	close
transit bus	October 22, 2005 6:30 p.m.	open
Rail service	October 22, 2005 6:30 p.m.	open
…		
Miami International Airport	October 22, 2005 6:30 p.m.	open
hurricane evacuation centers	October 23, 2005 afternoon	open

4. DYNAMIC DASHBOARDS

4.1 The Challenges for Dashboards

The dashboards provide condensed views for users to quickly explore the recent news and reports rapidly. It cannot display all the information in such a small area. Usually, when a disaster happens, the system will receive a huge amount of information from official government, companies and media, which is about almost everything. So it is necessary for the system to select a small portion of entities that a user really cares about to display in the dashboards.

Another problem in practice is the information sent from company users may have a lot of redundancies. For instance, when a hurricane arrives South Florida, almost all the company users in that area will report the same hurricane information: "*The storm has arrived South Florida*". Thus, the same information may be reported hundreds of times by different users. Therefore, the system has to identify which information is redundant and the redundant information should not appear in the dashboards.

We address these problems by introducing the dynamic dashboard supported by the content recommendation engine in our system. The engine's main task is to extract the most important, relevant and non-redundant information about entities from news and reports.

4.2 The Content Recommendation Engine

Figure 6 shows the data flow related to the content recommendation engine. In the system, there are four main data sources: EOC reports, news, company reports, and company messages. Since reports and news may contain information about multiple entities, in content recommendation engine, each of reports or news is divided into several documents, and each document consists of a sentence containing entity status information plus a context window (one previous and next sentence).

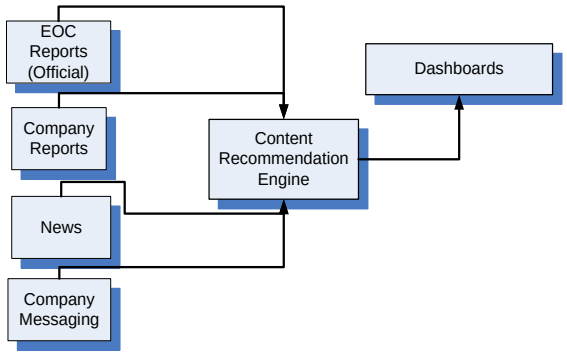

Figure 6. The Content Recommendation.

The content recommendation consists of two steps. The first step is text clustering, which is to cluster the same description of entities into one cluster. The second step is ranking the text by the relevance and presenting the top *k* items to the dashboards.

Note that the content recommendation engine is based on the text data while situation dashboard, thread dashboard and company dashboard display the structure information. The results of the information extraction of the text data are used to transform the results of recommendation engine to the formats that those dashboards need.

The four dashboards are denoted as Db_S(Situation Dashboard), Db_T(Threat Dashboard), Db_E(Event Dashboard),Db_C(Company Dashboard). The maximum numbers of items allowed to show in the dashboards Db_S, Db_T, Db_E, Db_C are denoted as $size_S$, $size_T$, $size_E$, $size_C$ respectively.

Table 5. The data sources for different dashboard.

Dashboard	EOC Reports	Company Reports	News	Company Messaging
Db_S	√		√	
Db_T	√			
Db_F	√	√	√	√
Db_C		√		√

The content recommendation engine recommends information from different data sources to the four dashboards. Table 5 shows the relationship between the data sources and the four dashboards.

Since the dashboards show the latest information, we use the last 48 hours records and news as the input of the engine.

For any user *u* in the system, the set of information submitted by *u* is denoted by *I*(*u*) and the set of reports/news of which the details are viewed by *u* is denoted by *J(u). u*'s profile in the system is composed of *I(u)* and *J(u)*.

4.2.1 Document Clustering

Before performing clustering, we first transform the text data (report, news and so on) to the vectors. We adopt the tf-idf transformation here [1]. Suppose the set of documents is $D=\{d_1,d_2,\dots d_N\}$, which is the data source of one dashboard shown in Table 1, and a set of words $W=\{w_1,w_2, \dots , w_M\}$. $n_{i,j}$ is the number of occurrences of the considered word w_i in document d_j, i=1,2…,M, j=1,2,…,N. $|d: w_i \in d|$ means the number of documents which contain word w_i, $d \in D$, and $w_i \in W$. We have

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} ,$$

$$\mathrm{idf}_i = \log \frac{|D|}{|d: w_i \in d|} ,$$

$$\mathrm{tf-idf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i .$$

The similarity between two documents can be calculated by the cosine similarity [2].

Suppose $t(d_i)$ and $t(d_j)$ are the transformed vectors from two document d_i and d_j, $d_i \in D$, $d_j \in D$. The similarity sim(d_i,d_j) of d_i and d_j is defined below:

$$\mathrm{sim}(d_i, d_j) = \frac{t(d_i) \cdot t(d_j)}{|t(d_i)| \cdot |t(d_j)|} .$$

Function *t*(*d*) is the tf-idf transformation from document *d* to the vector.

We apply the *k*-medoids [13] algorithm to cluster the documents. Note *k* is a user-defined parameter, which is determined by the managers of the system. It is also relevant to the number of items allowed to be displayed on the dashboards.

After clustering, each cluster contains the duplicated information about an entity and one document can be selected from a cluster to show the status of the entity. But before that, we have to decide which cluster and which document should be selected.

4.2.2 Content Ranking

For a specific user *u*, there are three priorities of the information. The three priorities from highest to lowest order are of EOC reports, Company Partner's information, that is the messages received, and other users' information, that is the company reports. In the system, the three priorities are denoted by user-defined parameters pr_1, pr_2 and pr_3 respectively, $pr_1 > pr_2 > pr_3 > 0$. For a given document $d_i \in D$, we use $pr(d_i)$ to indicate the priority of this document, $pr(d_i) \in \{ pr_1, pr_2, pr_3\}$.

Suppose the current user in the system is *u*, we can obtain the *u*'s feature f_u by the users profile as:

$$f_u = \alpha \frac{\sum_{u \in I(u)} t(u)}{|\sum_{u \in I(u)} t(u)|} + (1-\alpha) \frac{\sum_{u \in J(u)} t(u)}{|\sum_{u \in J(u)} t(u)|}$$

The parameter α is used to tune the importance weights of the reports submitted and viewed as the profile.

The Importance Score of each document $d_i \in D$ is calculated as follow:

$$score(d_i) = \mathrm{sim}(f_u, t(d_i)) \cdot pr(d_i)$$

For each dashboard, we use a top-*K* query to greedily search the *K* highest scores' documents from its corresponding data sources, where $K \in \{\mathrm{size}_S , \mathrm{size}_T, \mathrm{size}_E, \mathrm{size}_C\}$ and no two documents are selected from the same cluster. The set of *K* highest scores' documents is just the result of the content recommendation engine. Normally, since the EOC official reports has the highest priority, even some of them are not very relevant to the current user, information from these reports still is likely to appear on the Event Dashboard, where information can be from arbitrary sources.

5. REPORTS SUMMARIZATION

Besides providing the dashboards for a quick view of the current information, BCIN also allows users to search reports with keywords or query forms like traditional information systems. To provide a belief view of the latest information the user is interested, we lean on multi-document summarization to generate a summary of the returned reports, focusing on the latest status of the entities involved.

Using the last 48 hours as the criteria, we divide the returned documents into two sets, $\mathrm{D}^1 = \{\mathrm{d}_1^1, \mathrm{d}_2^1, \cdots, \mathrm{d}_{\mathrm{N}^1}^1\}$, which contains the documents submitted within the last 48 hours and $\mathrm{D}^2 = \{\mathrm{d}_1^2, \mathrm{d}_2^2, \cdots, \mathrm{d}_{\mathrm{N}^2}^2\}$, which containing the remaining documents. Since in Section 3, we extract structured information from sentences, sentences in $\mathrm{D}^1 \cup \mathrm{D}^2$ can also divided into two sets, $\mathrm{S}^1 = \{\mathrm{s}_1^1, \mathrm{s}_2^1, \cdots, \mathrm{s}_{\mathrm{M}^1}^1\}$ containing M^1 sentences having the structured information, and $\mathrm{S}^2 = \{s_1^2, s_2^2, \cdots, \mathrm{s}_{\mathrm{M}^2}^2\}$ containing the rest M^2 sentences. Our task is to extract sentences from $\mathrm{D}^1 \cap \mathrm{S}^1$ and $\mathrm{D}^1 \cap \mathrm{S}^2$, and the extracted sentences should be discriminative from D^2. Users can tune the parameter to control how many sentences are from S^1 and S^2. To extract the sentence set *V* from $\mathrm{D}^1 \cap \mathrm{S}^2$, the BCIN system adopts the sentence selection method proposed by Wang et al.[10]. We take the sentences as documents' features and different document sets as the labels, and then use the Minimum Redundancy and Maximum Relevance (mRMR) feature selection method to select the sentences which correlate strongly with the labels while keeping the redundancy of the selected sentences minimum. Formally, the sentence is selected according to the following formula [11]:

$$\max_{s_j \in S - V_{k-1}} [I(s_j, l) - \frac{1}{k-1} \sum_{s_i \in V_{k-1}} I(s_j, s_i)],$$

where V_k is the subset of S composed of *k* sentences extracted, *l* is the document label indicating which document set D^1 or $\mathrm{D}^2$is the document from, and $I(a,b)$ is the mutual information of two attributes a and b in all documents.

To extract sentences from $\mathrm{D}^1 \cap \mathrm{S}^2$, we make use of the extracted structured information. Since triples (entity, time, status) in each document have been extracted, instead of taking the sentences as features of documents, we take the extracted entities as the features and the status as the value. Using the following formula

$$\max_{e \in E - E_{k-1}} [I(e_j, l) - \frac{1}{k-1} \sum_{e_i \in E_{k-1}} I(e_j, e_i)],$$

where E is the set all entities extracted from D[1] and E_k is the subset of E composed of k entities selected, we can select a set of entities. Then for each selected entity, the latest sentence containing the entity will be selected to form the summary.

6. COMMUNITY GENERATION

There are two characteristics in disaster recovery scenarios which motivate us to consider geo-location information in the BCIN system. The first characteristic is that any event extracted from a report is associated with a/several location(s) indicating the place(s) where the announced event takes place. The second characteristic is that spatially co-located entities are more likely sharing similar disaster damage situation.

These two characteristics motivate the concept of community as we mentioned in Section 2.1. A community is a certain geographical region in which entities tend to share more recovery status or interests in common. So geographically identifying those communities is one of the most important issues to help companies understand what the current disaster situations are or their interested resources nearby.

The BCIN system addresses community generation by adapting existing spatial clustering algorithms. In practice, we provide an interactive spatial clustering interface for users to access multi-

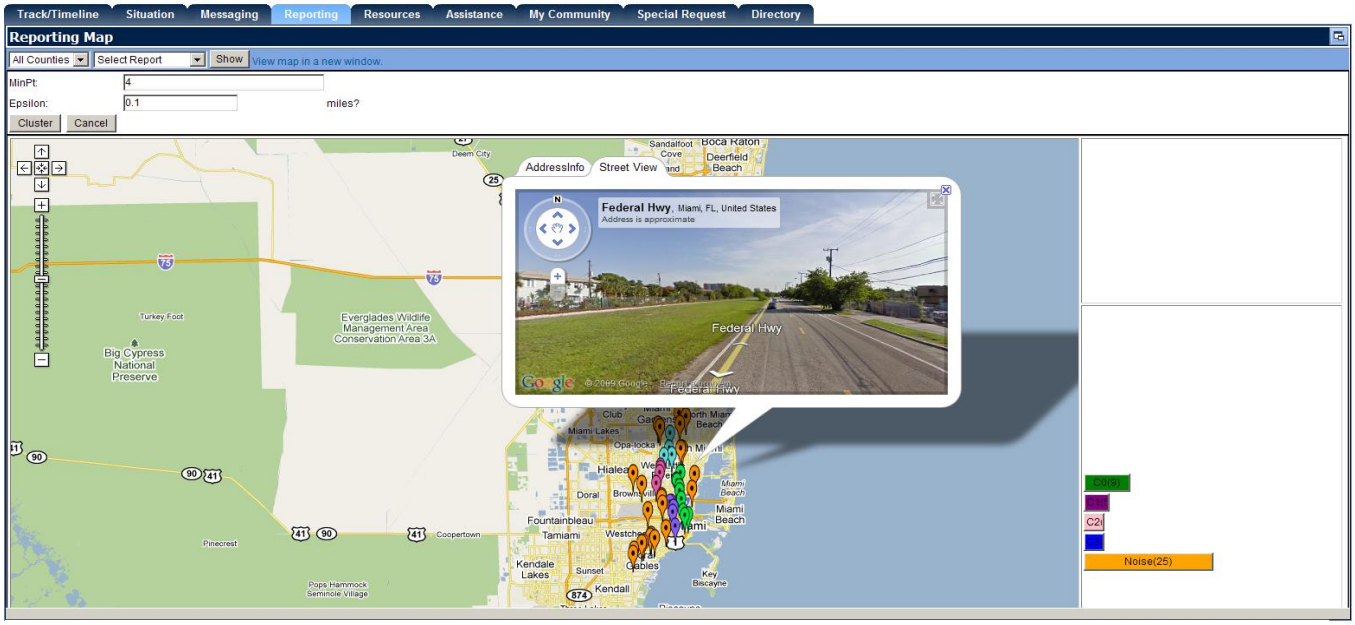

Figure 8. Interactive community display.

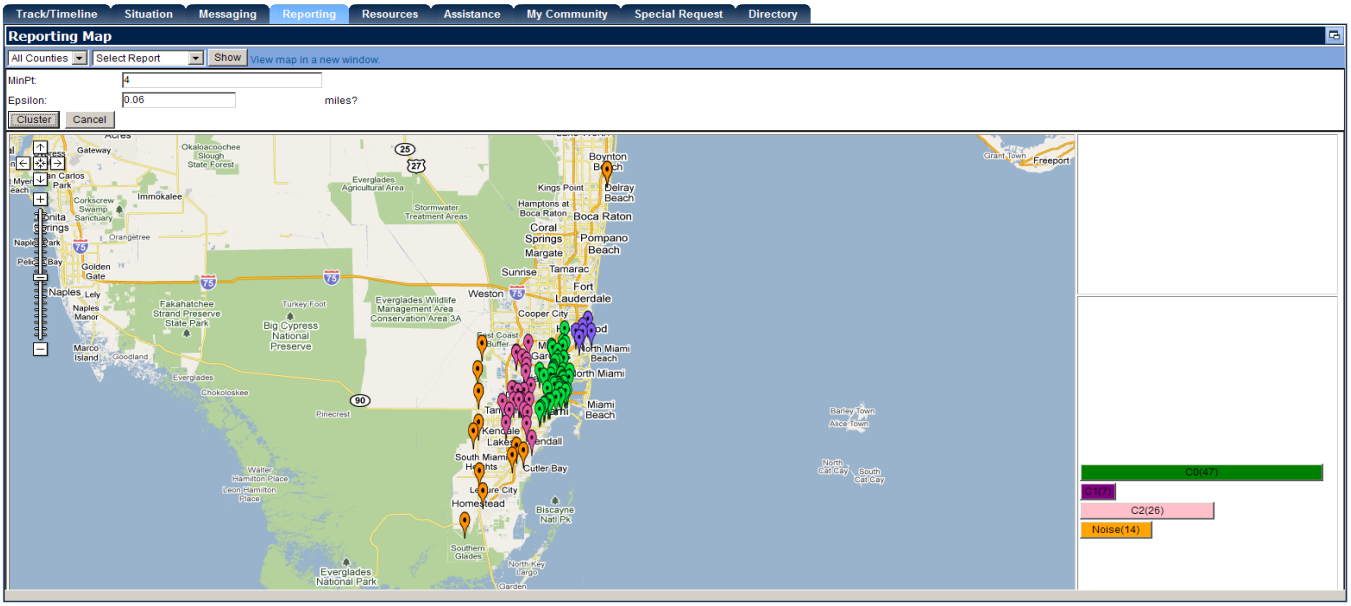

Figure 7. Dynamic community generation result.

level communities in a top down manner and consider physical or non-physical obstacles when generating spatial clusters to form more practical communities.

6.1 Spatial Clustering

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set [2,13]. Many spatial clustering techniques [12,14,15] have been developed for identifying clusters with arbitrary shapes of various densities and with different physical constraints.

In practice, communities formed by geographically related entities can be of various shapes. So we extend DBScan [14], a well-known density-based clustering algorithm, which is capable of identifying arbitrary shape of clusters, to generate dynamic communities.

6.1.1 Spatial clustering with constraints

We consider the method of spatial clustering with constraints. Generally, there are three types of constraints [13]: 1). Constraints on individual objects: Such constraints are non-spatial instance-level constraints that can be preprocessed before performing clustering algorithms. 2). Constraints as clustering parameters: Such constraints are usually confined to the algorithm itself. Usually, user-specified parameters are given through empirical studies. 3). Constraints as physical obstacles: Such constraints are tightly intertwined with clustering process. It is clear that physical obstacles are such constraints which prevent two geographically close entities from being clustered together. In real case, the bridge, highway and rivers are of this type.

In our BCIN system, we focus on object constraints and physical constraints.

Object constraints: We have two ways to obtain object constraints: 1) users submit formatted reports through report interface. Those reports are immediately recorded in the database; 2) our system extracts entity status from reports. For example, Table 4 can be used as object constraints.

Obstacle constraints: Polygon is a typical structure in spatial analysis to model objects. Obstacles modeled by a polygon can be represented as a set of line segments after performing polygon reduction [15].

Figure 7 shows the communities generated by clustering all open facilities and companies in Miami with the constraint: "I75 closed".

6.1.2 Interactive Spatial Clusters

In order to deal with imbalanced size of clusters, we provide users with an interactive mechanism to track the sub-community information within a large size community. By using this mechanism, users can obtain clusters with different granularities and more meaningful results. Figure 8 shows the interactive clustering results within the largest cluster in Figure 7.

7. SYSTEM DEVELOPMENT AND OPERATION

Through a series of interviews with public and private sector partners we identified the specific information both side could share and needed as part of their preparedness and recovery processes. The system then functionally established four key capabilities: Messaging, Reporting, Resources, Situational Browsing, so we can do things like alert a user via messages that a particular resource has been reported available at a local business. The proposed enhancements to the base system we have discussed in this paper provide new ways to connect reports, with resources, and the people/communities that need it.

FIU has spent over $600K in the development of the application and has received over $400K in sponsored research or industry donation. The system is monitored 24/7 via scripts that verify application, database, web server, and hardware availability. The system is managed in a revision control system and is run through a test suite that validates key functionality such as report submission, field validation, and role based access control. Over 100 companies (local and national) and government agencies in the south Florida area are utilizing the system，working closely with County emergency managers to collaborate on their mutual interest of disaster preparedness, response and recovery. The private sector benefits by receiving timely, accurate, information which impacts business operations and has the ability to report in situational information regarding disaster impact and infrastructure needs which are a priority for their business resumption. The public sector benefits by helping the business community receive and better understand disaster related information and can use disaster related situational reports from private sector to make better assessment of disaster impact.

Before the deployment of the BCIN, in a disaster situation, simultaneous reports from thousands of participants would overwhelm participants, making it very difficult to assess the status without dedicating a significant amount of time by all parties to process this potentially huge volume of information.

Using the proposed information extraction and report summarization techniques, the flooding status of an important commerce area such as Dadeland can be determined even if there are 1,000 companies providing status information. For instance, if these Dadeland based companies each logs into the system and enters a Flood report, or uploads a relevant document that contains relevant information, in an unstructured format, such as flooding area, depth, and public safety issues (nearby canals, down powerlines). The exercise has shown that the proposed techniques are able to identify critical common features of the flooding and summarize these, providing situational reporting in the Dynamic Dashboard. Further, if many of these companies are displaced by the damage, Dynamic Community Generation can inform community members about logistical concerns or assistance opportunities available.

8. EVALUATION AND CONCLUSION

Up to now, BCIN has been exercised at Miami-Dade County Emergency Management for the hurricane disaster management and recovery for three times. Miami-Dade, Florida is a very concentrated urban area (4th largest in the US), with tens of thousands of commercial concerns in a 25 square mile area. Miami-Dade County Emergency Management is interested in assisting this large, diverse business ecosystem to prepare and recovery quickly from hurricane impact.

Table 6 is a belief description of the exercises BCIN has taken part in. In a regional disaster such as a hurricane, business continuity professionals are under extreme pressure to execute their continuity of operation plans because many of the usual sources of information and services about the community and supply chain are completely disconnected, sporadic, redundant,

Table 6. Three exercises BCIN has participated.

Date	Description of the Exercise
Jun/01/2009	In Florida Dept. of Emergency Management's Statewide Hurricane Exercises, BCIN was utilized in a scenario where Miami-Dade County Emergency Management Business Recovery Desk facilitated the logistics to deploy portable ATMs at Shelters and PODs in Miami-Dade County.
Jun/29/2009	In Miami-Dade UASI exercise, BCIN was the tool responsible for communicate and collaborating with several companies that participated in the event as observers.
Aug/20/2009	In a full scale company training of BCIN where about 30 companies participated, companies were given injects to provide information to resolve different information requests.

and many times lack actionable value. The BCIN system focuses user input and collaboration around actionable information that both public and private sector can use. According the user surveys after these exercises, BCIN provided simple and effective interfaces for companies to communicate with each other and get information on critical needs, and scalable information sharing and recovery collaboration tools between these participants.

Feedback from our users suggest that our system can be used not only to share the valuable actionable information such as when are the schools going to close, what bridges have failed, etc, but to pursued more complex problems such as how county emergency management can collaborate with local banks to provide portable ATMs at shelters, whether we can provide a supply waypoint at local hospitals where nurses serving hospice patients can receive medicine, whether we can have a wireless provider setup cells-on-wheels (COWs) to provide local businesses at an area mall with communications. There are many such collaborative missions that can be undertaken on our system which allows public private sector entities to leverage their local capacity to serve the recovery of the community.

9. ACKNOWLEDGMENTS

This work is supported by NSF grant HRD-0833093 and DHS grant 2009-ST-062-000016. We thank Jason Allen and Mark Oleson for their work in the system development and testing.

10. REFERENCES

[1] G. Salton, and M. J. McGill. 1983. Introduction to modern information retrieval. McGraw-Hill.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, 2005. Introduction to Data Mining, Addison-Wesley.

[3] R.E. Anderson, Social impacts of computing: Codes of professional ethics. Social Science Computing Review, 2:453-469, 1992.

[4] K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. Li. Towards a business continuity information network for rapid disaster recovery. Intenational Digital Government Research Conference. 2008: 107-116.

[5] W.E. Mackay. Ethics, lies and videotape. In Proc. of CHI. 1995.

[6] M. Schwartz. Task force on bias-free language. Guidelines for Bias-Free Writing. Indiana University Press, Bloomington IN, 1995.

[7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of ICML. 2001.

[8] F. Sha, and F. Pereira. Shallow parsing with conditional random fields. In Proc. of HLT-NAACL. 2003.

[9] C.W. Hsu, and C.J. Lin. A comparison of methods for multiclass support vector machines. IEEE Transaction on Neural Networks, 13(2):415-425, 2002.

[10] D. Wang, L. Zheng, T. Li, and Y. Deng. Evolutionary document summarization for disaster management. In Proc. of SIGIR. 2009.

[11] F.L. Han, C. Peng, and C. Ding. Feature selection based onmutual information: criteria of max-dependency, max-relevance,and min-redundancy. IEEE Transaction on Pattern Analysis and Machine Intelligence, 27:1226–1238, 2005.

[12] O.R. Aaiane, A. Foss, C.-H. Lee, and W. Wang. On data clustering analysis: Scalability, constraints and validation. In Proc. of PAKDD 2002.

[13] J. Han, and M. Kamber. 2006. Data Mining Concepts and Techniques 2nd. Morgan Kaufmann.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discoverying clusters in large databases with noise. In Proc. of KDD. 1996.

[15] C.-H. Lee. Density-based clustering of spatial data in the presence of physical constraints. Master's thesis, University of Alberta, Edmonton, AB, Canada, July 2002.