Discovery, Analysis, and Characteristics of Event Impacts

Robert M. Patton
Applied Software Engineering Research
Oak Ridge National Laboratory
Oak Ridge, TN, U.S.A.
pattonrm@ornl.gov

Jim N. Treadwell
Applied Software Engineering Research
Oak Ridge National Laboratory
Oak Ridge, TN, U.S.A.
treadwelljn@ornl.gov

Ryan A. Kerekes
Applied Software Engineering Research
Oak Ridge National Laboratory
Oak Ridge, TN, U.S.A
kerekesra@ornl.gov

Thomas E. Potok
Applied Software Engineering Research
Oak Ridge National Laboratory
Oak Ridge, TN, U.S.A.
potokte@ornl.gov

Abstract – Assessing and monitoring events and their impacts continues to present multiple challenges. From financial markets to disaster management to epidemiology, the importance of understanding the impacts that events create cannot be understated. This work describes an approach that incorporates information from multiple sources and then analyzes the overall information flow to identify temporal patterns related to event impacts of interest. This approach is then applied to the analysis of news reports from multiple news sources pertaining to several different natural disasters. Results show that the approach provides insight into the impacts created by these natural disasters as well as indications to the severity of these impacts. In addition, results show that specific types of disaster consistently produce similar impacts when each time they occur.

Keywords: event detection, text analysis, multi-source & multi-resolution information fusion, and event correlation

1 Introduction

As global connectivity continues to grow, the worldwide relevance of formerly local events has become much more profound. Events that decades ago would be isolated news stories, now have the potential to create a cascading effect resulting in a global reaction. The development and expansion of the Internet as a news and information source provides much of the fuel for this cascading effect. News of events can now travel the globe at nearly the same rate at which the event itself unfolds. A recent example is that of the controversial Danish cartoons published in September 2005 [1]. Immediately, these cartoons created a strong reaction from the local Muslim community in Denmark. Within several months, this event led to a global reaction resulting in approximately 139 people dead, extensive property damage, and various bans on imports and exports between countries [1].

Identifying such events before they create a global reaction presents many difficulties, but is of significant importance in order to prepare a proper response. There are several major challenges to this. First, information concerning these events is often obscured, initially, within a massive amount of data. Consequently, this information is not likely to become "visible" until these events have created a global reaction. Second, since events may now have a global reach, information concerning these events may occur at any time within a 24-hour period, and must often be analyzed and responded to within a very short time frame. This creates a dynamic environment in which to track the latest "hot issues". Such dynamic environments pose significant challenges due to their unpredictability. Another challenge is that automation of event extraction from electronic sources remains very difficult [2][3]. Many of the difficulties are directly related to natural language processing. Unfortunately, thorough and accurate event extraction from massive data remains a laborious, manual process and is sometimes even impossible for humans to perform. Finally, another challenge is that events have a temporally based reaction characteristic. Some events cause an initial global reaction immediately. Some events cause only a local reaction that does not propagate. Unfortunately, there remains no simplistic or clearly defined approach for defining and measuring a reaction to an event.

Given these challenges and the significance of a solution to the problem, the grand vision is to automatically detect events that may cause a global reaction before that reaction occurs. In an attempt to move towards this vision, this paper describes a novel approach to discovering and analyzing evidence of the impact of events on different aspects of society. This work narrowly focuses on resolving a specific problem: identifying and characterizing the impacts of natural disasters as they relate to physical infrastructure and social behavior using information obtained via news reports. The work reported here is a preliminary investigation into solving this problem.

Section 2 discusses some background information and related works that provide the context under which this new event impact analysis approach was developed. Section 3 describes the design and process for this approach. Section 4 describes the application of this analysis approach to natural disasters that occurred in 2007. Sections 5 and 6 provide observations and summary, respectively.

2 Related Works

Investigating the impacts of natural disasters on society is not a new research area. There are many works relating to the identification of different impacts as well as different studies that have been performed [4][5][6]. Many of these works and studies, however, have identified impacts that are difficult to measure quantitatively either before, during, or after the event. In addition, some works have relied on remote sensing techniques for assessing and monitoring disaster situations [7][8][9]. For this work, the goal is to develop an approach that would provide a more quantitative view of the event impact as well as be founded on field reporting about the event. Inspiration for this approach came from work performed in sunspot detection and indirect detection of extrasolar planets.

In the field of sunspot detection [10], if a sunspot (i.e., an active region on the surface of the sun) is facing the earth, it can cause an effect on the earth's atmosphere. Such active regions on the sun emit different characteristics than other non-active regions of the sun. In the work of [10], the authors determined that there is a particular type of radiation that can be measured such that sunspots can be detected before they begin facing the earth. This method of indirect detection provides an early warning of increased solar activity that may affect the earth.

Another analogous work is that of indirectly detecting extrasolar planets [11]. In their work, the authors describe several approaches for detecting extrasolar planets based on indirect measurements of the environment in which the planet exists. According to their work, the very existence of a planet will produce measurable effects on the stars that they orbit. Depending on the effects, it may even be possible to determine the size of the planet and distance from the star that it orbits.

Others have also investigated the impact of events from news reports, but in different ways. In the work of [12], information from blogs was analyzed with respect to the actual sales of a book. The authors discovered that there was a direct relationship between blog chatter prior to a book release and the actual volume of book sales after the book release. The more people talked about a book on their blogs prior to its release, then the higher the volume of sales for the book after its release. The authors show clearly that blog chatter is a good indicator of the impact that a book release will create on sales volume. In the work of [13][14], the relationship between financial news and stock prices was investigated. The foundation of these works is the premise that financial news can have either a positive, negative, or natural impact on the price of a stock, and that the time lag between the news and the stock price was minimal, if any. The authors observed that, in fact, some of the financial news could provide indications as to the direction of the stock price (up, down, unchanged). However, their results are based on market simulations using real data. Regardless, their work provides evidence that events and event impacts can be monitored and gauged using news reports.

In the field of natural language processing, there is substantial work in automated event detection, including related works [15][16][17]. Generally, these works focus on using clustering or categorization techniques for event detection. Unfortunately, for massive data sets, clustering techniques become computationally intensive and categorization techniques become plagued by the challenge of defining appropriate categories [18]. In addition, these works do not focus on the event impact, but rather just the event detection and tracking of the event. Unlike these approaches, our work does not depend on clustering or strict categorization techniques.

In the field of time-series analysis, there is also work involving the temporal aspects of events and their impacts. In the work of [19], a time-series is decomposed into three sub-series referred to as seasonal, trend, and noise. The goal is to find complete or partial periodic patterns in a time-series with trends. Their work provides an approach to observe both short and long-term periodic patterns. They demonstrate their approach on atmospheric CO_2 levels as well as stock prices. In the work of [20], the problem of identifying temporal patterns involving multiple time granularities is investigated. The authors developed an event structure that relates different events on various time scales to each other. The authors then look for repetitive patterns of these event structures as they relate to stock prices.

3 Approach

In this work, the primary focus is to detect, analyze, and characterize the impact of events as observed in the news media. Much like the work of [12][14], the relationship between event impacts and the news media is investigated. Unlike these works, however, this work also incorporates the concept of decomposing a time-series into sub-series as discussed in [19]. Figure 1 shows the conceptual view of our approach.

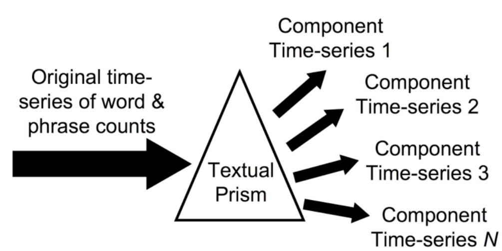

Figure 1. Concept of approach

The original data begins as a time-series of all of the word and phrase counts that are observed in the news media. Next, a "textual prism" is created that is comprised of a set of taxonomies that describe the words and phrases related to the topics of interest by the user. Each taxonomy is defined by the user and is applied to the original time-series in order to produce a component time-series. This component time-series shows how the words and phrases of the taxonomy change over time.

Once the component time-series are created, wavelet-based noise filtering is then performed on each component series [21][22]. Using a wavelet-based approach allows a multi-resolution analysis of the signal, whereby small fluctuations in the signal due to noise can be filtered out while the larger, more meaningful spikes are retained. In this approach, a Haar wavelet basis is used for the analysis due to its efficient computational structure. The Haar wavelet coefficients of each component time-series are computed, and any coefficients whose square falls below a certain threshold are set to zero. The resulting sets of coefficients are transformed back to the time domain to yield a filtered version of the time-series. The next section will describe the application of this approach to real data.

4 Application

To demonstrate this approach, this work focuses on natural disaster events. In terms of events and event impacts, natural disasters are fairly simple and well defined. In addition, they generally have a short duration (usually measured in hours or days, but rarely weeks) and rarely have any preceding or subsequent events. Consequently, impacts of the event should be more easily identifiable. Future work will focus on expanding this approach to address more challenging events such as man-made events. However, the current approach should be applicable to other domains such as finance and medical.

When natural disasters occur, they often make significant impacts on society. As discussed in section 2, there have been investigations into the various impacts caused by natural disasters. For this work, the research performed by [23] provides the foundation for the impacts to be investigated. This work was chosen for its simplicity and extensibility. In [23], impacts can be observed in 4 areas: Technical, Organizational, Social, and Economic. Technical refers to the infrastructure of a society (roads, bridges, power grid, water systems, etc). Organizational refers to areas such as service crews that maintain or respond to the Technical aspects. Social refers to areas related to housing, shelters, provisions for human needs, etc. Economic refers to the impacts on the economy. This work focuses on the Technical and Social impacts of a natural disaster.

To begin, our data source consists of time-series data of word and phrase counts from closed captioning feeds of news sources such ABC, CBS, NBC, CNN, and Fox News. Every word and phrase observed from these sources is counted each day. Each particular word or phrase has an associated vector of how many times that word or phrase was used across all of the information sources over a specific time period.

Next, a set of taxonomies were developed for both the Technical and Social categories as shown in Table 1. These taxonomies were developed manually by analyzing news reports from the time period of August 25, 2005 to September 5, 2005. This is the time period when Hurricane Katrina made landfall in the Gulf Coast of the United States [24]. This natural disaster created significant and widespread damage and resulted in extensive impacts on U.S. society. The news reports during this time period were clustered. The clusters were then analyzed to determine the most popular words and phrases that were used to describe specific conditions. These words and phrases were then categorized according to the framework defined in [23] as shown in Table 1.

As a specific example, the Shelters taxonomy consists of word and phrases such as: taking cover, taking shelter, seek refuge, shelters. The Movement taxonomy consists of words and phrases such as: mandatory evacuation, dawn curfew, flights canceled, etc.

Table 1. Natural Disaster Impact Taxonomies

Technical	**Social**
Power	Water (Health)
Communications	Movement
Roads	Medical
Lights	Health Hazards
Water (Systems)	Crime
Outage	Shelters

After creating these taxonomies, they are then applied to the original time-series as a form of “textual prism” to create component time-series as described in Figure 1. In this case, the original time-series consists of news reports from the time period of January 1, 2007 to December 31, 2007. News reports consisted of closed caption data received from multiple sources such as ABC, CBS, NBC, CNN, and Fox News.

For brevity, only the results from January 2007, February 2007, and October 2007 are discussed here. The following figures show the time-series results from applying the taxonomies to the original time-series. Figures 2 – 5 show the Social and Technical components for the month of January 2007. Figures 6 – 9 show the same components for the month of February 2007. Figures 10 – 13 show the same components for the month of October 2007.

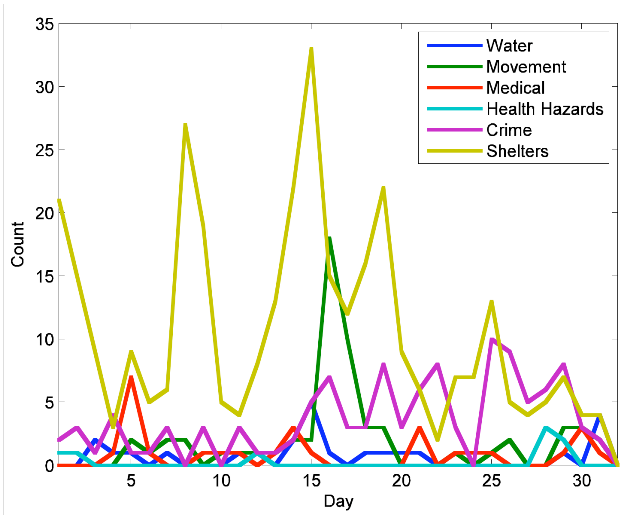

Figure 2. January 2007 Social Components

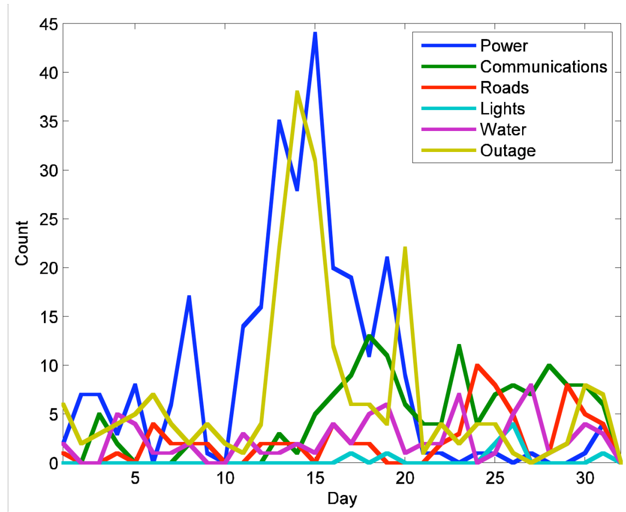

Figure 3. January 2007 Technical Components

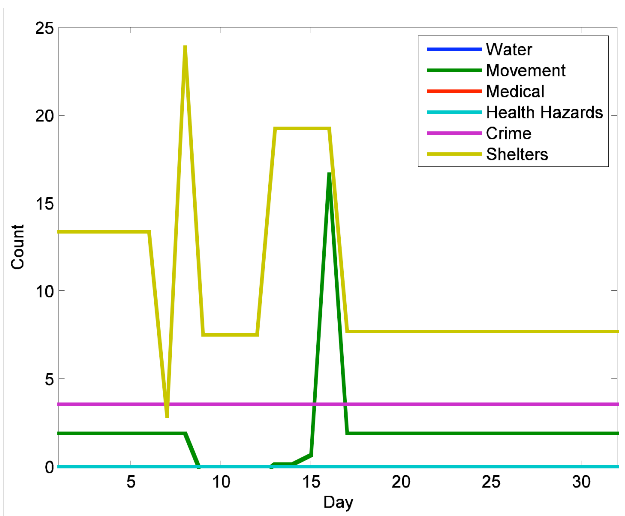

Figure 4. January 2007 Social Components Filtered

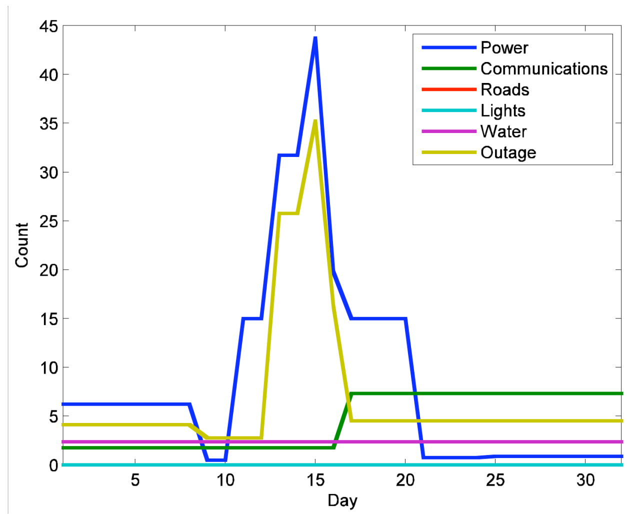

Figure 5. January 2007 Technical Components Filtered

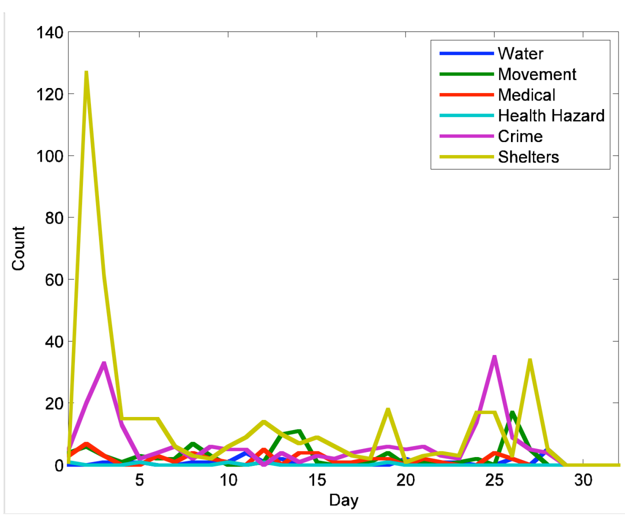

Figure 6. February 2007 Social Components

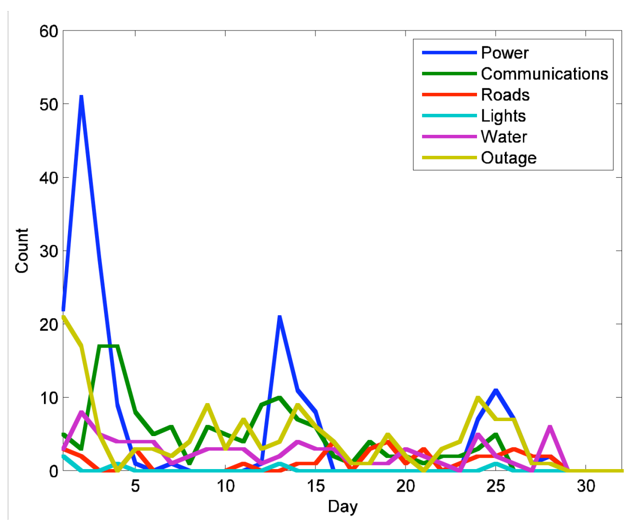

Figure 7. February 2007 Technical Components

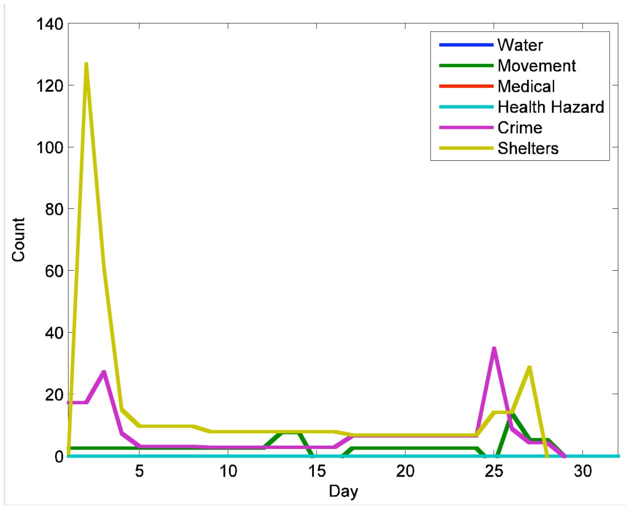

Figure 8. February 2007 Social Components Filtered

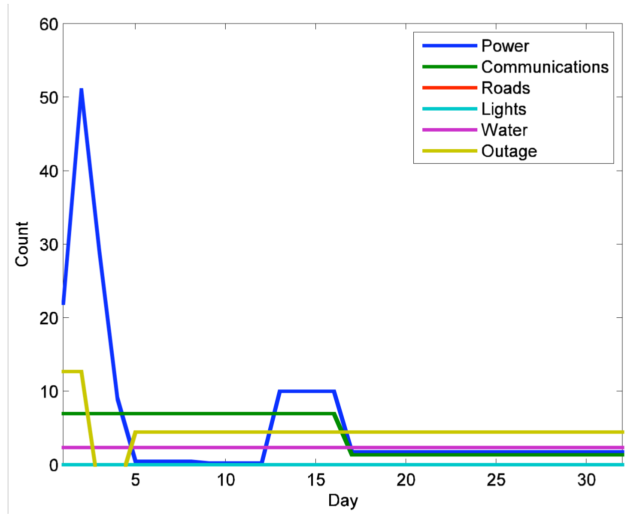

Figure 9. February 2007 Technical Components Filtered

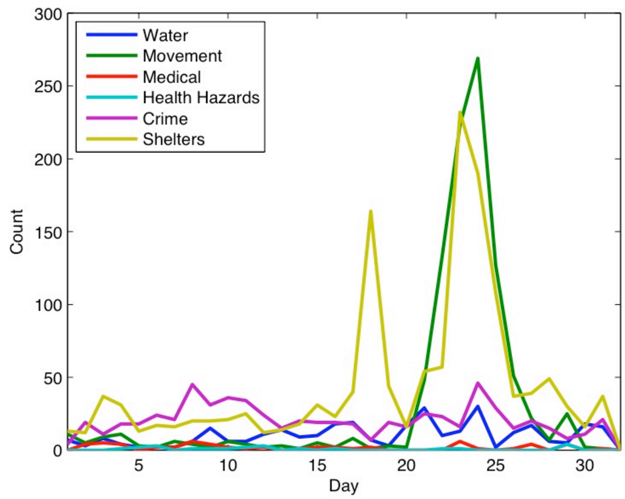

Figure 10. October 2007 Social Components

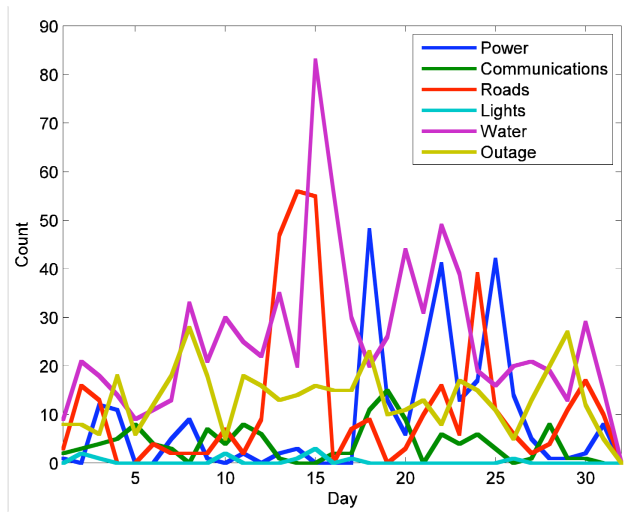

Figure 11. October 2007 Technical Components

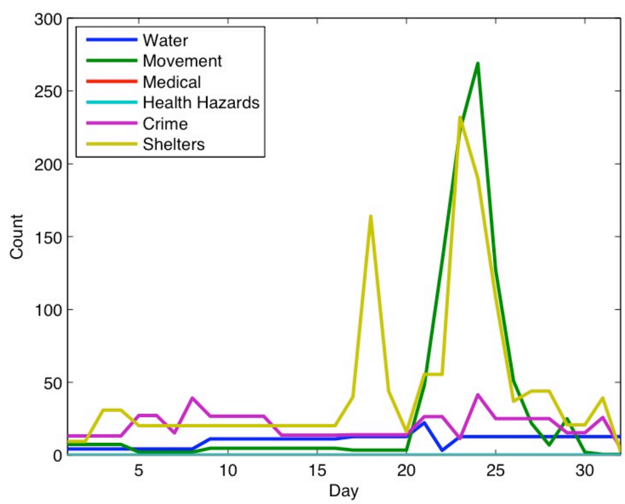

Figure 12. October 2007 Social Components Filtered

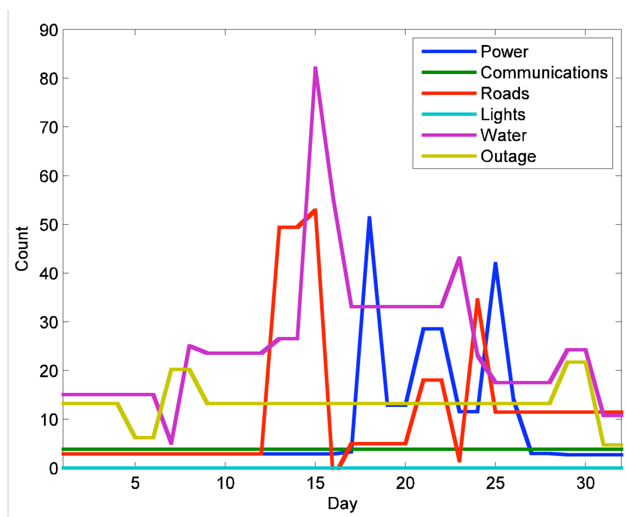

Figure 13. October 2007 Technical Components Filtered

Observations

In 2007, several natural disasters occurred in North America. Table 2 shows a sample of these disasters. In January, several tornadoes and an ice storm occurred [25][26]. In February, several tornadoes and a winter storm occurred [25][27][28]. Starting June, a heat wave began in the Western United States [29][30]. Finally, in October, several tornadoes and several wildfires occurred [31][32].

Table 2. Selected Natural Disasters in 2007

Event Type	Begin	Peak	End	Areas Impacted
Tornadoes	01/04	N/A	01/07	LA, MS, AL, GA
Ice Storm	01/11	01/14	01/24	Central & Eastern U.S.
Tornadoes	02/02	N/A	02/02	FL
Winter Storm	02/12	02/14	02/20	Midwest & Eastern U.S.
Tornadoes	02/23	N/A	02/24	KS, LA, MS, AR
Heat Wave & Drought	June	July & Aug.	Ongoing	Western & Eastern U.S.
Tornadoes	10/17	N/A	10/19	Midwest & Southern U.S.
Wildfires	10/20	10/21-10/25	11/09	CA

With the exception of the tornadoes in February, most of these disasters impacted very large regions of the U.S. Therefore, the data shown in Figures 2 – 13 provide a national level view of the impacts of these events. Future work will seek to expand this view to regional and local views.

In Figures 2 – 5, significant spikes occur in the Shelters and Power components on Jan 7. These spikes are directly attributed to the tornadoes that occurred on those same days. In addition, there are significant spikes that occur in the Shelters, Movement, Power, and Outage components on Jan 12 – Jan 20 with the highest peaks occurring on Jan 14. These spikes are directly attributed to the ice storm.

Tornadoes are difficult to predict, and are much more destructive on homes than an ice storm. Consequently, the Shelters component for the tornadoes has much higher amplitude than the Power component. In contrast, the ice storm is a more predictable event and has a more destructive impact on infrastructure such as the power grid and roads than on homes. Therefore, the Power and Outage components created much higher amplitudes than Shelters and Movement. In addition, since the ice storm is more easily predicted than tornadoes, the Movement component created much higher amplitude for the ice storm than for the tornadoes.

These patterns in the Social and Technical components for the tornadoes and ice storm in January are also evident in the month of February. In Figures 6 – 9, significant spikes occur in the Shelters and Power components on Feb 2. These spikes are directly attributed to the tornadoes that occurred on those same days. In addition, there are significant spikes that occur in the Shelters, Movement, Power, and Outage components on Feb 12 – Feb 16. These spikes are directly attributed to the winter storm. Finally, significant spikes occur in the Shelters and Power components on Feb 23 – 27. These spikes are directly attributed to the tornadoes that occurred on those same days.

Like January, the tornadoes in February created maximum amplitudes in Shelters followed by Power. On the other hand, the winter storm created significant amplitudes in Power, Outage, Shelters, and Movement. These patterns were similar to the ones observed for the ice storm in January.

Figures 10 – 13 show the component time-series for the month of October. For this month, there are several interesting events and patterns. First, there were tornadoes that occurred on Oct 17 – Oct 19. Like the tornadoes in January and February, these tornadoes also produced significant spikes in the Shelters and Power over the same period as when the tornado events occurred. Between the three months, another interesting observation is that the Shelters spike for tornadoes is, on average, at least twice as large as the Power spike. Further analysis is necessary to determine the consistency of this pattern.

In addition to the tornadoes in October, there were also wildfires. The impacts of the wildfires can be seen from Oct 20 – Oct 27. These wildfires created one of the largest evacuations in U.S. history [32] with nearly 1 million people forced to evacuate. Consequently, the Movement and Shelters component show tremendous spikes during that period. Between the three months, these spikes are easily the largest in amplitude. In addition, the wildfires also created spikes in the Power and Roads component time-series. This is a result of the wildfires threatening the power grid and also shutting down several major highways.

Finally, one of the most complex events that occurred in 2007 was a heat wave and drought. These two events are complex for a variety of reasons. First, there is no clear start and end dates that can be associated with them. They are more indicative of "conditions" rather than a particular event. According to [29][30], the heat wave started approximately in June 2007 in the Western U.S and moved eastward in July and August 2007. For the drought, there is no indication of when the conditions began. In addition to this, these events are also complex in that their impacts are not as clearly evident as those for the other disasters. However, they did create impacts either in different ways or at different times. According to

[32], one of the causes for the wildfires was the extreme dry condition that was exasperated by the heat wave. In fact, 2007 was the driest year on record for California. Consequently, while there were no immediate impacts observed by the heat wave, it appears to have manifested the impacts in the form of wildfires. Furthermore, the heat wave, which primarily occurred in the summer of 2007, created more severe drought conditions in Southeast U.S. However, it was not until October 2007 when these drought conditions began to make significant impacts on the water supplies. In Figures 11 – 13, both the Social and Technical Water component time-series are significantly elevated in comparison to January and February. The Social Water component reflects the need for drinking water and water usage as it relates to human needs. The Technical Water component reflects the water supplies, reservoirs, sewage systems, and other infrastructure related to water. The elevated levels of these components are directly attributed to problems with the water and sewage systems and water reservoirs reaching record lows in the month of October 2007 [33]. Numerous communities were severely impacted and many were forced to implement emergency water management practices [33].

Table 3 summarizes the impact characteristics that were observed for the various natural disasters shown in Table 2. Using the approach described here, these impacts are identifiable and consistent. Future work will explore additional natural disasters as well as additional impacts.

Table 3. Dominant Impacts from Natural Disasters

Winter & Ice Storms	**Tornadoes**	**Wildfires**	**Heat wave & Drought**
Power	Shelter	Movement	Water (Tech)
Outage	Power	Shelter	Water (Social)
Shelter	Outage	Power	
Movement	Medical	Roads	

5 Summary

This work focuses on identifying and characterizing the impacts of natural disasters as they relate to physical infrastructure and social behavior using information obtained via news reports. In this work, an approach is proposed that enables event impacts to be discovered and characterized. Results indicate that events such as tornadoes, winter storms, and wildfires create consistent, measurable and specific impacts on the physical infrastructure and social behavior as reported by the news media. This approach leverages the news media to further enhance situational awareness of natural disasters.

While this approach provides a fresh perspective on events and their impacts, future work will explore additional natural disasters, additional types of impacts, and more challenging events such as man-made events.

References

[1] Wikipedia: Danish cartoons, Current February 2008, http://en.wikipedia.org/wiki/Jyllands-Posten_Muhammad_cartoons_controversy

[2] Huttunen, S., Yangarber, R., and Grishman, R. *Complexity of Event Structure in IE Scenarios*, in Proc of the 19th International Conf. on Computational Linguistics, August 2002.

[3] Turmo, J., Ageno, A., and Catala, N. *Adaptive Information Extraction*, ACM Computing Surveys, Vol. 38, No. 2, July 2006.

[4] Alexander, D. *The Study of Natural Disasters, 1977-97: Some Reflections on a Changing Field of Knowledge*, Disasters, Vol. 21, No. 4, pages 284–304, December 1997.

[5] Little, R.G., Toward more robust infrastructure: observations on improving the resilience and reliability of critical systems, Proc. of the 36th Annual Hawaii International Conference on System Sciences, January 2003.

[6] Kreimer, A., Arnold, M., and Carlin, A., *Building Safer Cities: The Future of Disaster Risk*, World Bank Publications, 2003.

[7] Llinas, J., *Information fusion for natural and man-made disasters*, Proc. of the 5th International Conference on Information Fusion, July 2002.

[8] Chi, T., et. al., *Research on information system for natural disaster monitoring and assessment*, Proc of the 2003 IEEE International Geoscience and Remote Sensing Symposium, July 2003.

[9] Hussain, M., et. al., *Emerging geo-information technologies (GIT) for natural disaster management in Pakistan: an overview*, Proc. of the 2nd International Conference on Recent Advances in Space Technologies, June 2005.

[10] Bertaux, J.L., et. al. *Monitoring solar activity on the far side of the Sun from sky reflected Lyman alpha radiation*, Geophysical Research Letters, Vol. 27, No. 9, pages 1331-1334, May 2000.

[11] Marcy, G.W. and Butler, R. P., *Detection of Extrasolar Giant Planets*, Annual Review of Astronomy and Astrophysics Vol 36: 57-97, September 1998.

[12] Daniel Gruhl, et. al., *The predictive power of online chatter*, Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, 2005, pp. 78-87.

[13] Gabriel Pui Cheong Fung, et. al., *Stock prediction: Integrating text mining approach using real-time news*, Proc. of the 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003, pp. 395-403.

[14] Gabriel Pui Cheong Fung, et. al., *The predicting power of textual information on financial markets*, IEEE Intelligent Informatics Bulletin, Vol. 5, No. 1, June 2005.

[15] Yang, Y., Pierce, T., and Carbonell, J. *A Study on Retrospective and On-Line Event Detection*, in Proc. of the 21st annual international ACM SIGIR, August 1998.

[16] Allan, J., Papka, R., and Lavrenko, V. *On-line New Event Detection and Tracking*, in Proc. of the 21st annual international ACM SIGIR, August 1998.

[17] Yang, Y., Ault, T., Pierce, T., and Lattimer, C. *Improving text categorization methods for event tracking*, in Proc. of the 23rd annual international ACM SIGIR, July 2000.

[18] Reed, J., et al. *TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams* in Proc. of the 5th International Conference on Machine Learning and Applications (ICMLA'06). Orlando, FL., 2006.

[19] Jeffrey Xu Yu, et. al., *Patterns discovery based on time-series decomposition*, Proc. of 5th Pacific Asia Knowledge Discovery and Data Mining, 2001.

[20] Bettini, C., et. al., *Mining temporal relationships with multiple granularities in time sequences*, Data Engineering, Vol. 21, No. 1, March 1998.

[21] Percival, D. B., and Walden, A. T., *Wavelet Methods for Time Series Analysis*, Cambridge University Press, 2000

[22] Vetterli, M., and Kovacevic, J., *Wavelets and Subband Coding.* Englewood Cliffs, NJ: Prentice Hall, 1995.

[23] Bruneau, M., et. al. *A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities*, Earthquake Spectra, Vol. 19, No. 4, pages 733-752, November 2003.

[24] Wikipedia: Hurricane Katrina, Current February 2008, http://en.wikipedia.org/wiki/Hurricane_Katrina

[25] Wikipedia: Tornadoes of 2007, Current February 2008, http://en.wikipedia.org/wiki/Tornadoes_of_2007

[26] Wikipedia: January 2007 North American ice storm, Current February 2008, http://en.wikipedia.org/wiki/January_2007_North_American_ice_storm

[27] Wikipedia: February 2007 Central Florida tornadoes, Current February 2008, http://en.wikipedia.org/wiki/2007_Central_Florida_tornadoes

[28] Wikipedia: February 2007 North America Winter Storm, Current February 2008, http://en.wikipedia.org/wiki/February_2007_North_America_Winter_Storm

[29] Wikipedia: 2007 Western North American heat wave, Current February 2008, http://en.wikipedia.org/wiki/2007_Western_North_American_heat_wave

[30] Wikipedia: 2007 North American South and Eastern heat wave, Current February 2008, http://en.wikipedia.org/wiki/2007_North_America_South_and_Eastern_heatwave

[31] Wikipedia: Mid-October 2007 Tornado Outbreak, Current February 2008, http://en.wikipedia.org/wiki/Mid-October_2007_Tornado_Outbreak

[32] Wikipedia: October 2007 California wildfires, Current February 2008, http://en.wikipedia.org/wiki/California_wildfires_of_October_2007

[33] NOAA U.S. Drought Watch, Current February 2008, http://www.ncdc.noaa.gov/oa/climate/research/2007/oct/us-drought.html