Final Grant Report to VT CCSR

Social Media for Cities, Counties and Communities

PIs: Andrea Kavanaugh, Virginia Tech and Apostol Nastev, IBM
Co-PIs: Edward A. Fox, Steven Sheetz, Donald Shoemaker, Virginia Tech, and Lexing Xie, IBM
Graduate Researchers: Seungwon Yang, Lin Tze Li, Venkat Srinivasan

March 11, 2011

1 Overview

Social media (i.e., Twitter, Facebook, Flickr, YouTube) and other tools and services with user-generated content have made a staggering amount of information (and misinformation) available. Some government officials seek to leverage these resources to improve services and communication with citizens, especially during crises and emergencies. Yet, the sheer volume of social data streams generates substantial noise that must be filtered. Potential exists to rapidly identify issues of concern for emergency management by detecting meaningful patterns or trends in the stream of messages and information flow. Similarly, monitoring these patterns and themes over time could provide officials with insights into the perceptions and mood of the community that cannot be collected through traditional methods (e.g., phone or mail surveys) due to their substantive costs, especially in light of reduced and shrinking budgets of governments at all levels. We conducted a pilot study in 2010 with government officials in Arlington, Virginia (and to a lesser extent representatives of groups from Alexandria and Fairfax, Virginia) with a view to contributing to a general understanding of the use of social media by government officials as well as community organizations, businesses and the public. We were especially interested in gaining greater insight into social media use in crisis situations (whether severe or fairly routine crises, such as traffic or weather disruptions). Our findings from the pilot study fall into three main areas:

1) local government agencies use social media without knowing its costs and benefits, or who their actual audience is, who in their organization should monitor communications, how and when they should be responding, and what effect their social media communications have on the public;

2) new tools are needed to help government and citizens make sense of the overwhelming amount of data that is being generated, to model the flow of information, and to identify patterns over time; and

3) digital libraries are needed to archive and curate generated content, especially for crisis and social convergence situations, but also for analyses that cover longer time frames.

Our work also builds on a larger set of studies carried out by a Virginia Tech team funded by NSF (IIS-0916733) with the overarching goal of building a Crisis, Tragedy and Recovery Network (CTRnet) (http://www.ctrnet.net).

2. Social Media Use by Government, Local Organizations and the Public

Our six-month planning project (July-December 2010) funded by the Virginia Tech Center for Community Security and Resilience (CCSR) was an exploratory study of how social media data analysis can be applied in Arlington and environs to improve services and communication with citizens (Figure 1). Based on interests and needs demonstrated in a CCSR workshop with

officials from Arlington County and the National Capital Region (NCR) (the area around Washington, D.C.) and from review of prior research, we were selected to conduct a planning study that explored social media applications to improve community resilience in times of crises, as well as a timely and complementary open source of information for improving city, county, and community services.

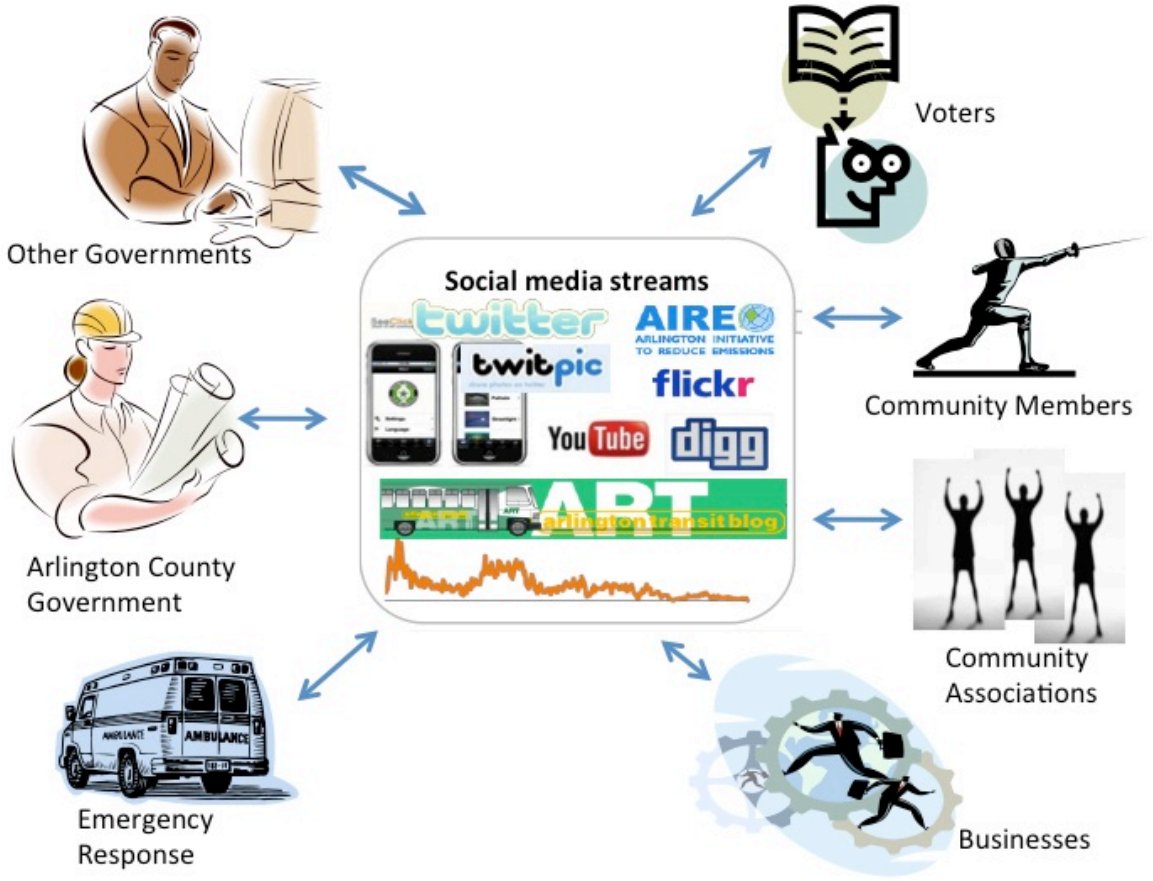

Figure 1: Social media streams to improve services and communication with citizens

Target information sources included official county publicity portals, blogs, news, community forums, as well as relevant postings by the public on social media sites such as Twitter, Facebook, YouTube, and Flickr. Applications of such analyses include monitoring public opinion before and after large public events, monitoring planned or unplanned activities, identifying and categorizing important community issues over time and location, enhancing community recovery in response to crises or tragedies, and monitoring and tracking the development of long-running themes in civil life.

3. Methods

We used a combination of qualitative and quantitative techniques to assess current use of social media, user interests and needs and potential areas of application for government agencies, private and non-profit organizations and the general public. We surveyed application domains (e.g., education, transportation), collected and analyzed area-specific social media (SM) sources, conducted focus group interviews with 24 county officials (specifically, personnel from

emergency management services, the police department, and volunteer leadership office), and identified a simple taxonomy of social media issues in community contexts.

Survey of application domains: we identified examples of social media systems and usage in six different sectors of urban life. The purpose is to enumerate examples of social media in action, examine their goals, means and status, so as to extrapolate and derive trends and patterns for future development. We grouped the examples into six broad sectors based on the respective civil functions.

Stakeholder and user study: We conducted three focus group interview sessions with Arlington County officials and citizens to identify their needs and interests regarding social media monitoring and emergency response opportunities and issues. In an initial meeting with area representatives, we met with a individuals from business, no-profit groups (including homeowner associations).

Acquisition of social media sources: We used monitoring tools to capture relevant sources based on characteristics such as: textual descriptions, timeliness, the ability to reflect current events quickly, and multimodality -- providing the ability to illustrate events visually, with scenes, objects, and actions (and locations in the case of some geo-tagged or other location aware applications). We focused on topics and content related to city life and civil services.

Data analysis. We analyzed social media for main purposes: to determine the adequacy of available data around the different facets of civil life, and to extrapolate on the potential insights and value for users. We used analysis in connection with the interviews, in order to bootstrap the goals and functionality of a more complete system. We used existing analysis tools and algorithms from the IBM and Virginia Tech teams, including content summarization and visualization, ontology design, visual classification, spatial and/or temporal trending and profiling, and preliminary network and topic analysis.

4. Results

4.1 Survey of application domains: social media use in city life

In this section we survey examples of social media systems and usage in six different sectors of urban life. The purpose is to enumerate examples of social media in action, examine their goals, means and status, so as to extrapolate and derive trends and patterns for future development. We group the examples into six broad sectors based on the respective civil functions. The examples below are not meant to be exhaustive, given the large amount of ever-growing social media applications online.

Public safety

Public safety involves the prevention of and protection from events that could endanger the safety of the general public from significant danger, injury/harm, or damage, such as crimes or disasters (natural or man-made) [27]. The main public safety concerns among law enforcement, city officials and citizens include condition monitoring, emergency alertness, general awareness, rapid information dissemination, and so on.

For abrupt natural disasters such as earthquake the timely sensing and reporting is a primary

concern. Sakaki et al have shown [25] that with proper information extraction and statistical estimation, aggregated information from twitter messages can detect earthquake in near real-time, and that the spatial-temporal traces of tweets faithfully reflect the trajectory of hurricanes. The Quake-Catcher Network led by Cochran and Lawrence [21] employs existing networked laptops and desktops to form a dense, distributed computing seismic network. Software developed by IBM [26], on the other hand, utilizes data center machines as earthquake sensors. These distributed sensing approaches can be very effective where expensive seismic sensing equipment is not available.

Social media can also be information aggregators for emergency managers and public safety workers. Catawba County in North Carolina used twitter hashtags to monitor the progress for Hurricane Earl in September 2010 [24]. The role of social media there was to serve as a central information hub and reduce strain on 911 centers having to monitor many difference sources. Twitter and other social media sites can also be a good platform for local governments to collaborate with and inform each other. A recent survey by the American Red Cross [22] presented rich statistics on the usage and preferences of social media in emergency contexts. Among those surveyed, the ratio of people getting emergency-related information from social media was still relatively low (16%). However, the ratio that reported they will sign up for notifications (50%) and those who would post emergency information (72%) are much higher. Furthermore, 69% respondents agreed that response agencies should monitor social media and respond to information posted, and three-fourths expected that help would arrive within an hour. These observations indicate that the adoption of social media in emergencies should be on the rise despite the current usage levels.

Transportation

Urban and suburban transportation is one of the key factors influencing citizens' quality of life. It involves the movement of people and goods from one location to another via a number of means such as air, rail, road, water. Further, it has a number of aspects such as infrastructure, vehicles, and operations. Social media provide an easy-to-use means to gather and disseminate information about various aspects of transportation. It is particularly well-suited for distributed coordination, such as traffic conditions and mapping routes [14], finding gas prices [9], ride sharing [18, 8], broadcasting transit information [2], campaigning for environmentally friendly lifestyle and transportation [4], and spotting parking violations [6]. While the effectiveness of these social media varies by geography and depends largely on the size of the active online local community, the applications have clear advantages in de-centralized coordination conditions, such as ride sharing, as well as distributed information gathering, such as price aggregation.

Energy and utilities

Energy and utilities are often taken for granted, but are also a focus for discussions around the environment, conservation, and global warming. Being a crucial civil sector with a long history and bound by a variety of regulations, there has been some hesitation to use social media in utilities. Nonetheless, there are initial steps of adoption [7, 5] as many utilities started participating by twitter accounts and Facebook pages; scenarios include advertising green energy sources, joining consumer conversations, and improving customer service. It was also argued [11] that there is an urgent need for utility firms to tap into and maintain a presence with social media, in part because of risk mitigation (i.e., word gets around fast) and more diverse ways of monitoring discussions.

Healthcare

Social media adoption varies widely among the difference parties in the healthcare sector. Patients, physicians and the general public embrace the new means of sharing information, compare experiences, and build online communities that are otherwise impossible due to widely dispersed geography. PatientsLikeMe.com [10] is one example of an innovative online service that enables patients who suffer from life-changing diseases to converse with one another, allowing them to share data on improving their outcomes, empathize with each other and to learn any techniques or medication other sufferers are trying in order to improve their health. Forward-looking government agencies use social media platforms for public communication and education, in order to promote its image and increase transparency; the statement made by Singapore Ministry of Health [13] is a good example. Aggregate information in healthcare is often of greater value than the mere collection of individual pieces. An example of such indirect web 2.0 applications is Google Flu Trends, where aggregating web queries relating flu reveals a very telling picture of seasonal disease worldwide.

Education

The web has been an indispensable resource for learning in all stages of education. Traditional classroom education and continuing education both make use of social media. Wikipedia and other wikis are living examples of massive-scale distributed knowledge creation. Many examples and discussions [17, 16, 23, 19] have used wikis, including wikipedia and online sources for creating teaching topics, find references, and create missing references. Wikis as a living lab for large-scale collaboration ”for fun” also provide extra motivation for students. Moreover, social media can be used to both teach appropriate research and writing standards, and can also expose students early to network safety and privacy topics.

For continuing education and self-learning, wikis, forums, and question-answer sites plays an even more important role. Social media sites here can be generic [1] or specialized [15], and they are often paired with reputation-based reward system that both selects good quality answers and motivates participants, creating a win-win operation model.

Development

Urban development and maintenance of a city’s infrastructure is vital yet often unnoticed part in ensuring the smooth function of many aspects of our urban life. While the management and execution of these functions are specialized agencies, the surveillance and reporting aspects can be open to wider participation. Recently such applications have emerged as both independent online services [12] and as part of the government’s official mobile applications [3]. They are used as an additional channel for reporting infrastructure and environment problems. Such applications certainly encourage citizen participation, however, the final verdict is still out on their efficacy. This is because the outcome is limited by the liability of the execution agencies, and ultimately depends upon the responsiveness of government entities in charge.

Summary

This survey intends to cover the broad aspects of using social media in urban life, rather than being exhaustive. From the examples, we can see that there are several patterns where social media applications can be effective. (1) Distributed, asynchronous collaboration stimulated by common interest: The Internet is a great platform for bringing together birds-of-a-feather for

reference authoring, question answering, or mutual support. (2) The whole is greater than the sum of its parts: trends, predictions, and alerts can be derived from hundreds of thousands of individual contributions. (3) Improving communication: by increasing openness and encouraging citizen participation, such applications benefit government's public image if executed well. A number of these scenarios may benefit from further automation and content collection, organization and analysis tools.

4.2 Focus Group Interviews

We conducted three focus group sessions lasting two hours each with three different types of groups: Emergency Management Services employees of Arlington County (recruited by the EMS office), the Arlington County Politic Department, and a mixed group of volunteer organization representatives associated with the overarching 'Arlington Volunteers' office of Arlington County. Each two hour interview session consisted of two steps, beginning with the participants engaged in electronic brainstorming to generate a substantive number of ideas quickly, followed by their identifying categories that grouped the ideas by similarity.

Using group support software, the participants anonymously generate and enter ideas, beliefs, issues, or concepts, in the form of short sentences or phrases that they feel are important to the situation. We developed and used a set of framing questions (shown in Figure 2) to cue participants to begin entering ideas. Participants' responses to these questions are visible on each participant's computer screen as the responses are generated, allowing ideas generated by one person to be expanded by others or to cue others to generate related ideas. As a single group interview participants aided by a facilitator discuss, create and name the meaning of response items or categories in order to organize their ideas by similarity.

- What are the missions and objectives of your organization?
- What are you trying to accomplish using social media?
  - Do you feel you are currently accomplishing this goal effectively with social media? (if yes, why?)
  - If not, what do you need [to know? – to do? --in order] to use social media more effectively?
- What concerns do you have about using social media?
- What difficulties do you have about using social media?
- What information would you like to have about how your organization uses social media?
- What information would you like to have about how social media is being used in your community?
- Is there anything else you would like to know about social media that would be helpful?

Figure 2: Framing questions used in previous focus groups

The categories generated by the focus groups are merged based on the ideas that participants group together by similarity into the taxonomy started in the previous groups. Participants in our three focus group sessions identified shows the 23 categories of ideas (shown in Figure 3). The categories consist of factors related to the organization and factors related to the information exchanged between the organization and the community.

Organization Factors

Organization factors (Figure 3) include policies, legal issues, costs, and training. The organization requires that polices be adopted to provide the environment needed for employees to achieve. Management buy-in is essential if benefits are to be realized and costs are to be

controlled. To utilize SM effectively the activities and roles implemented are institutionalized through Human Resources (HR) developing job descriptions and ensuring related types of communication are managed effectively. There are attempts to control information and to communicate the government's opinions and actions that the public should pursue. In addition organizations desire to define the types of information to be shared and the manner in which it is shared. The participants perceive the substantive legal issues related to maintaining government transparency, often through the Freedom of Information Act (FOIA), as important considerations of using SM. For example, are tweets by a government employee public record, most likely; what about tweets related to their non-work life, perhaps not. Costs are always important to organizations, and government budgets have been squeezed due to reduced receipts resulting from slowing economic activity and increased use of government services. Yet the participants perceive that the potential exists for achieving efficiencies using SM and the potential return on investments should be evaluated. Complicating this calculation is the value placed on reaching previously uninvolved constituents and the most interested participants. One of the costs of adopting SM is the training of the employees that will conduct the activities. In addition, the public must be educated to understand how the government will interact with them and what expectations for interaction are appropriate.

Information Factors

Information factors include issues related to the quality and quantity of information generated through SM. They also include the tone of and types of communications in which government desires to participate, including outreach, feedback, and two-way communications. Additional types of information that can be obtained from some SM channels, e.g., detecting the locale of emerging events, are of substantial interests for emergency management and policing functions. Finally, the security of technology used to provide SM capabilities and new tools needed to meet legal obligations for saving public records comprise a set of technology issues that contribute to the information factors. Lastly which existing SM tools should be utilized remains a substantial question across the focus groups.

Together the factors identified by the participants describe a broad range of interests and concerns of the Arlington County government in relation to their use of SM. Each of these categories also contains a set of ideas from the electronic brainstorming that further clarify the intentions of the participants about the meaning of the categories.

- Information Factors
  - Communications
    - Community Outreach (emergency, crime/traffic alerts, 24/7 level of service, recruitment)
    - Feedback (from community to organization, social trends, locale, fast spreading ideas)
    - Population Reached (misses traditional/older population or can't afford technology)
    - One Way vs Two Way (pushing out vs creating dialogue, effort/costs different)
    - Tone (Government presents just the facts, not stories, not press release, listen then educate)
  - Information
    - Quality of Content (accuracy, facts of situation, un-vetted information, misinformation)
    - Quantity of messages (how to be heard, from 1 to 10 to 1000s, overwhelming, loss of control)
    - Personal Level (information overload, ability to write complete thoughts, nuances of face-to-face lost)
  - Technology
    - Security (network exposed to world)
    - Technology and Equipment (cost of technology and maintenance, cost savings, training)
    - Social Media (SM) Outlets (knowing audience/expertise, users expect transparency, so many outlets)
    - Public Record/FOIA (are SM public record, tools needed to save, outdated polices)

- Organization Factors
  - Policy
    - Management Buy-In (unknown expectations, under valued, need to set culture)
    - Control Issues (how much to control, what we can control, telling how/what to think/do)
    - Human Resource (HR) Components (job descriptions, evaluation, expertise, dialogue, positive and negative)
    - SM Communications Policy (what not to do/say, right people to make SOP, moving target)
    - Professional Level (privacy concerns, devices owned by county, investigative purposes)
  - Legal Issues
    - Data Maintenance (FOIA data maintenance and related costs)
    - Owing Vs Using Someone Else (official outlet versus imposter, use in investigations.
    - Public Record/FOIA (are SM public record, tools needed to save, outdated polices)
  - Costs
    - Resource Issues (SM adds to previously full time job, other duties, limit 24/7 expectation)
    - ROI/cost to value (how to measure value, who are we reaching, enough received messages)
  - Training
    - Education (tools to manage, learning from each other, train constituents where to go)
    - Training (best practices for dividing duties, case studies, understanding management's concerns)
    - Other (educate nonusers, establish boundaries)

Figure 3: Simple taxonomy of categories identified by focus group participants

Some of the Arlington County interview participants said that they need social media aggregation tools. In general, dashboard services that accept search keywords and phrases help monitor information from multiple social media, such as trackur (http://www.trackur.com/social-media-monitoring) and Netvibes (http://www.netvibes.com/). But these tools are designed to support businesses not government or citizens, so they are not optimal for civic needs. We will modify existing tools, as required. Having geo-mapping features would be very useful for the needs of cities and communities, which are not currently enabled in dashboard tools. We will select from emerging applications that allow citizens to contribute geo-tagged photos and video to a community database. For example, MIT's Mobile Media Experience Laboratory has developed a place-based application called Locast for this purpose (http://www.locast.mit.edu). The video analytic software IBM has developed will help to organize and cluster images of similar content or location. This would make it easier for users to find content of interest and to contribute to ongoing information exchange regarding a particular issue or place.

Feedback from participants in our focus groups in Arlington also indicated that recent or projected budget cuts could erode 15 years of community outreach; the County wants to understand how to use technology to maintain and sustain established communications with citizens. We will intervene at this critical time. The neighborhood/civic associations are key, but not all neighborhoods have associations. These residential neighborhoods (usually with a lower socio-economic population) fall through the cracks. Social media may be particularly helpful for such households and neighborhoods, especially in combination with cell phones.

4.3 Social Media Data Analysis

In order to study the pattern of communication and the information communicated using social media, we collected publicly available data from Twitter, Facebook and YouTube related to Arlington County and environs. We identified 34 civic organizations, some of which are government agencies, in the NCR that were tweeting; we collected and analyzed their tweets for 30 days between September and October 2010.

Twitter Analysis

We analyzed the tweets as well as the biographical information posted as profiles of the organizations' followers using Natural Language Toolkit, tag clouds, and graphs. Figure 4 shows the number of followers for the 34 civic organizations. When we look into the number of followers of these followers, however, we see the extensibility of the communication chain radiating out beyond the organization originating tweets (Figure 5). Further analyses show us which words are used most commonly in the tweets or bios during this period. The predominance of various words (most common words appear larger in a tag cloud) provides a quick overview of what is being said or characterized (in the case of followers' bios).

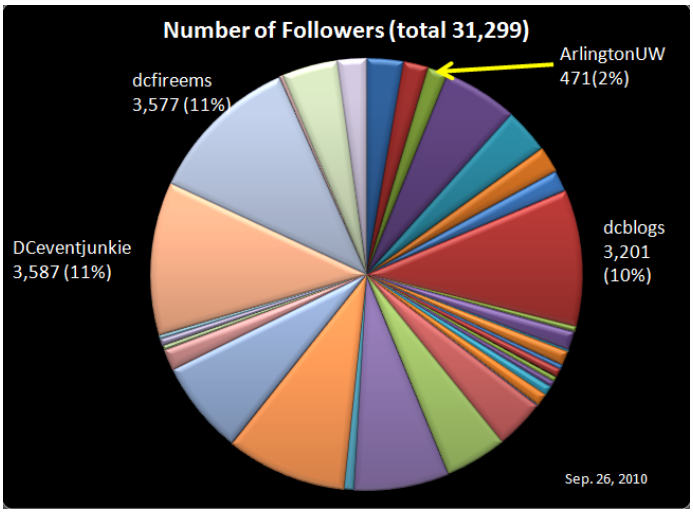

Figure 4. Number of follwers for 34 NCR civic organizations

For the 34 civic organizations that were tweeting during the September – October 2010 period, we see there are a total of about 31,000 'direct' followers (i.e., people who subscribe to the RSS feed that carries each organization's twitter posts). What is interesting to note is that the 'direct' followers are themselves being 'followed' by other people – what we refer to as 'followers of followers' (Figure 5). The number of followers of followers for these same organizations is over 67 million.

For an organization such as Arlington Unwired (Arlington UW), that disseminates announcements about local events, shown with an arrow in Figure 4, there were 471 followers on the date we captured these data (September 26, 2010). We can see from the analysis of the number of Arlington UW followers' followers (Figure 5) there are over 8 million followers. This is not to say that a tweet from Arlington UW will go beyond the 471 direct followers; however, if there is a crisis in the Arlington area (such as a major catastrophe or extreme violence) it is very likely that the *indirect* followers will retweet (forward along the same twitter post) regarding such a catastrophe to their own set of followers (i.e., over 8 million followers). In this way, we can see the potential reach of a critical piece of information being disseminated throughout a community way beyond the direct Twitter followers to a larger population of followers' followers.

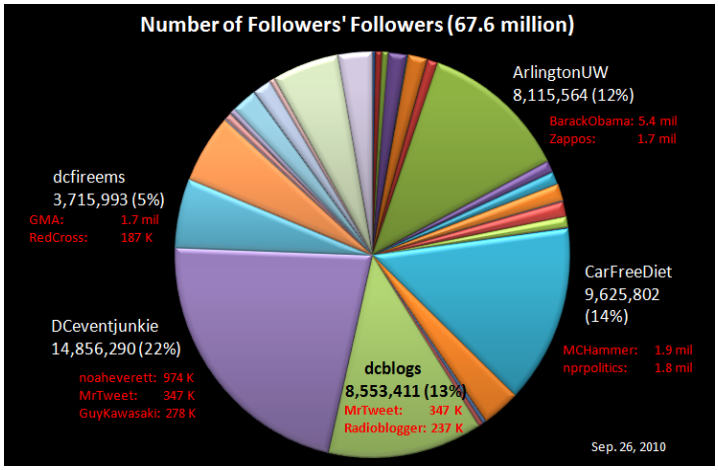

Figure 5. Followers of Organizations' Followers

It is also important to note that among the followers of Arlington UW is 'Barak Obama' – and the number of followers of 'Barak Obama' is over 5 million.

In order to get a sense of who are the followers of these 34 civic organizations in the National Capitol Region, we collected the publicly available biographical profile information that followers list on their own twitter accounts. Figure 6 shows in a tag cloud the predominant descriptors that followers' use in their profiles.

Figure 6. Profile biographies of Arlington UW followers

This kind of analysis and visualization provides quick overview of the type of individuals and their interests who are following a given organization. Followers of tweeting organizations or individuals are likely to be tweeting themselves. We analyzed and visualized in a tag cloud the 20 recent tweets of the followers of Arlington Unwired (UW) at the time of the data capture (September 26, 2010) shown in Figure 7.

By looking at the recent tweets of followers, we see a kind of 'mood' and 'buzz' among users. The large 'RT' stands for re-tweet, meaning that this is the most common term appearing in the twitter posts for these users. The organization knows from this analysis that many of their posts are going well beyond their immediate (direct) followers.

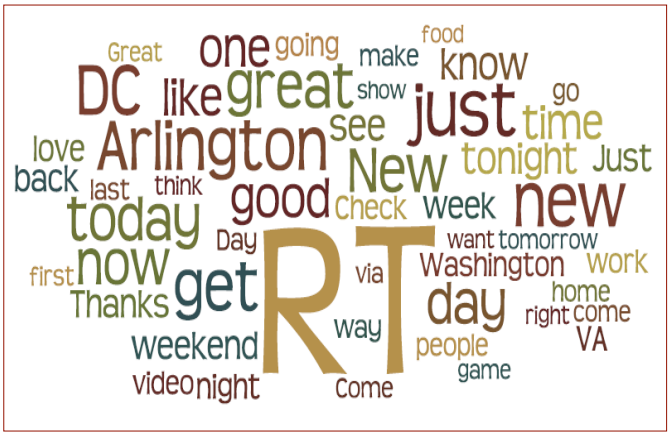

Figure 7. Twenty recent tweets by followers of Arlington UW

The purpose of these analytical and visualization tools, as noted earlier, is to allow government and citizens to see quickly and easily the big picture of the information and communication flows that interest them.

Analysis of Facebook Comments

Arlington County government has maintained a Facebook page since early 2010 (http://www.facebook.com/ArlingtonVA). The page had roughly 4500 fans at the end of September 2010 (by February 2011, there were close to 4900 fans). We analyzed a two-month period (August -September, 2010) of posts by the County and responses (comments) from the public by conducting a simple content analysis by topic. There were a total of 112 posts; the top 10 most frequent topics are shown in Figure 8.

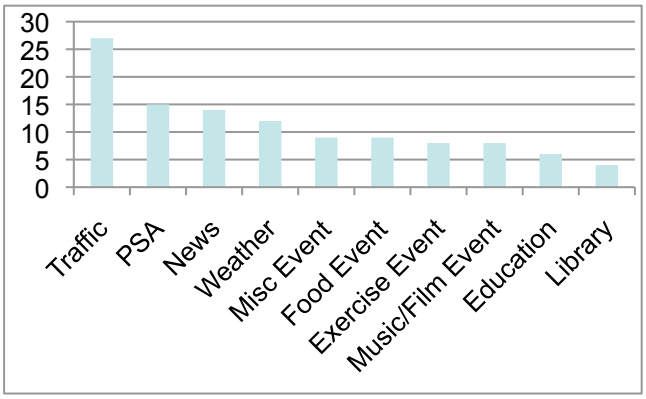

Figure 8. Top Ten Topics in Arlington County Facebook

The most common posts by the County on the Facebook page were about traffic (e.g., conditions, closures, metro outages), followed by public service announcements (PSA). News (shout-outs, updates, and other County announcements) and weather related posts (National Weather Service and Arlington Weather Service advisories) were followed by various events (good, biking, walking, music or film) in terms of frequency of posts. There were only a few posts related to education (Arlington County School District) and library services (e.g., closures, speakers, special activities) during this two month period.

There were a total of 824 public comments to the County posts during this two-month period. Half of the comments pertained to about a fifth (19%) of the County posts (the top 21 posts by the County). Figure 9 shows the distribution of the bulk of the comments on the same top 10 County posts seen in Figure 8.

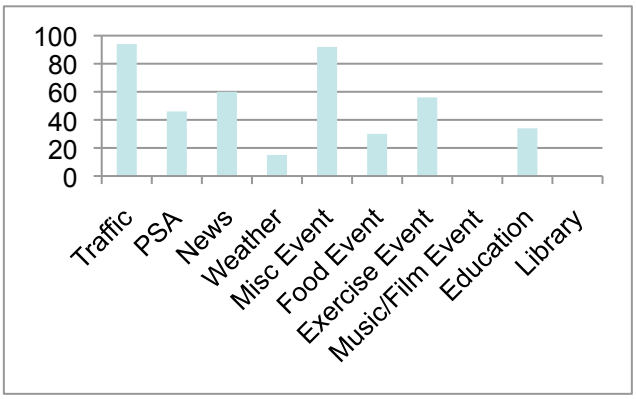

Figure 9. Public Comments by Arlington Facebook Topic

The comments are predominantly related to traffic and miscellaneous events (that is, events that do not fall into the other 'event' categories shown, such as food, exercise, music and film). Exercise events (bikes, walks) and News announcements generated the next most frequent number of comments from the public. Almost all the comments were highly consistent with the social media policy of the County (e.g., no profanity or off topic comments) and were overwhelmingly positive in tone, including many "Likes" hits.

Lastly, we collected videos in YouTube pertaining to Arlington, Virginia and conducted a tag analysis of the video collection using image software developed by IBM. We performed a search using Perl script and the phrase 'Arlington County;' this produced about 1800 videos from YouTube. We then developed two types of tag clouds generated using video titles and video tags (see Figure 10).

Figure 10. Tag Cloud of Arlington YouTube Videos

The tag cloud as a visualization quickly and easily represents the frequency with which different terms appear in the search thereby providing a snapshot of what is in the large dispersed collection. The more frequently a term appears in the image collection, the larger it appears in the tag cloud. The cloud visualization also provides an indication of the importance of various

civic issues to members of the community. The recurring civic themes revealed in the video analysis can be further explicated in the six categories shown in Table 1.

Table 1. Tag Cloud Categories for Arlington Videos

Law enforcement	Police, cops, officer, courthouse, robbery, accident, ACPD, surveillance
Transportation	Metro, street, boulevard, highway accident, parking, transit
Social issues	Environment, diversity, community, city, neighborhood, accountability
Economic development	Growth, sustainability, development, bank, private, local
Political	Government, elections, agencies, department
Communication	Media, ABC, NBC, CBS, television, news, network, bilingual, NoVAPJ, Spanish

The further clustering of video tags and video titles as shown in Table 1 allows government and other users to make sense more easily of the interests and needs of the community as expressed in the YouTube collection at any given point.

4.4 Usage ontology from knowledge sources

Our goal with the CTRnet digital library (DL) is to collect, organize and serve resources that can cover the disaster and emergency management domain comprehensively. Without knowing which concepts are important and how they are related with one another, we may not see the big picture. This might lead the CTRnet DL to collect and provide resources that are unbalanced in covering the domain. For this reason, we are developing a CTR ontology with in-depth coverage.

Currently the resources being archived (with support from our main project partner, the Internet Archive) are mostly news web pages about events. Through expansion to include more information -- such as emergency responses for specific types of events, emergency management plans, and crisis informatics content -- people in emergency response teams and researchers in crisis informatics would benefit, too.

The CTR domain is broad. Therefore, having solid domain knowledge will help to collect and organize resources. It also will help visitors navigating through information in the CTRnet digital library. Developing a CTR ontology will allow us to assemble a key form of domain knowledge. To confirm accuracy across the broad CTR area, and to make the effort feasible and scalable, it makes sense to create the ontology through a semi-automatic methodology involving human as well as computational efforts (e.g., natural language processing). To prove that the ontology is useful in organizing resources, an ontology-based classifier will be implemented and then compared with both probabilistic and non-probabilistic classifiers.

We have bootstrapped the ontology development process by merging the classification variables from four databases designed to record the existence of disasters. We selected this method due to the commitment in resources required to develop a database and supporting information system, assuming that those capable of designing and developing such systems would uncover the essential elements of disasters in their requirements analyses and design their systems to support those elements. By evaluating multiple databases we hope to identify elements that are common as well as elements more idiosyncratic to particular activities. The databases we evaluated included:

1) Richmond Disaster database (http://learning.richmond.edu/disaster/index.cfm),

2) EM-DAT database (http://www.emdat.be/),
3) Canadian disaster database (http://www.publicsafety.gc.ca/res/em/cdd/search-en.asp),
4) DesInventar tool for defining disaster databases (http://www.desinventar.org/).

The resulting draft ontology current consists of 185 elements. Our process for merging the disaster classifications from the four databases was straightforward. We started with the disaster classification form in the Richmond database, and then merged into it the EM-Dat database classification elements. The process went smoothly with the Richmond database providing more leaves and the EM-Dat elements comprising higher level elements. Next we merged the Canadian disaster database into the ontology, assigning each of its elements to an existing element in the draft ontology or creating a new element in the ontology. The overall hierarchy of the draft ontology was stable through this process with most of the Canadian disaster database mapping to leaf elements in the ontology. Finally we merged the DesInventar disaster classification into the merged ontology from the other three databases.

A large majority of elements matched the draft ontology. However, the concept of the "cause" of the disaster was not included in any of the other databases. We have decided to exclude this element from the merged ontology due to the lack of consensus across databases. However, DesInventar includes an extensive set of slots to be filled for every disaster and we have adopted them for integration with the ontology. Figure 11 shows a small subset of the elements identified. Panel a (in larger font) shows the highest levels of the draft ontology. There is substantive consensus across all the databases we evaluated for a "manmade" versus "natural" type of disaster at the highest level of describing disasters. Panels a, c, and d show how the ontology expands as elements at higher and middle levels are selected. In this example Manmade Disasters are expanded to show those that are Conflict Based, and one of its elements, i.e., Terrorism, is finally expanded to show leaves of the tree. Similar expansions are included for the higher elements presented in Panel a.

We are currently mapping the attributes or slots captured by the databases, e.g., the number of people impacted or the number of homes destroyed, for a disaster into this draft ontology. Initial efforts show that associating a disaster with the appropriate elements and obtaining values for their attributes will capture a comprehensive representation of and record of facts related to disasters. One attribute related issue remaining to be resolved is geographic location. While it may seem that all disaster database systems would simply adopt a standard GPS representation that could be translated to other coordinate systems as desired, this is not the case. Although several databases included such attributes, e.g., longitude and latitude, they all include other, mostly unique, geographic representations including states, provinces, counties, boroughs, etc., apparently intended to allow for the maximum flexibility of definition. This is also complicated by the nature of some disasters that have impacts across multiple countries, states within countries, and cities.

a. Highest level of ontology

- Disaster
  - Manmade
    - Crowd Event
    - Conflict Based
    - Human Systems Failure
  - Natural
    - Biological
    - Climatological
    - Geophysical
    - Meterorological
  - Unclassified
    - Compound Event

b. Manmade -> Human Systems Failure elements.

- Human Systems Failure
  - Fire
    - Tire
    - Tunnel
    - Underground
    - Urban
  - Industrial
  - Public Health
  - Transport
    - Air
    - Rail
    - Road
    - Water

c. Manmade -> Conflict Based elements.

- Conflict Based
  - Attack
  - Civil War
  - Cultural Disappearance
  - Ethnic Violence
  - Genocide
  - Illegal Immigration
  - Mass Murder
  - Piracy and Maritime Terrorism
  - Prison Incident
  - Refugees
  - Riot
  - School Shooting
  - Terrorism
  - War

d. Conflict Based -> Terrorism elements.

- Terrorism
  - Aircraft Attack
  - Assassination
  - Attempted
  - Biological
  - Chemical
  - Cyber
  - Hijacking
  - Hoax
  - Kidnapping
  - Nuclear
  - Occupation
  - Property Destruction
  - Robbery
  - Rocket Attack
  - Shooting
  - State Conventional
  - Threat

Figure 11: Selected elements from ontology merged from four disaster database projects

5. Directions for full investigation of social media use by government and citizens

Our pilot study was intended to advance technologies and systems for social media analysis relating to both routine day-to-day civil life and critical incidents or emergencies. Specifically, we have begun to:

1) leverage and further refine tools for collecting and correlating large amounts of public social media data relevant to Arlington County, VA and environs,

2) archive and curate collected social media data over a period of time into a digital library, including social media for crisis conditions, and

3) identify, research and implement applications of multimedia analytics and text mining for government services and communication.

In a fuller investigation, we would be able to develop and leverage technology more fully to help government manage information and facilitate interaction in meaningful ways in order to achieve broader public participation than is possible through normal channels (e.g., public commenting at county board meetings). Deep analysis of social media streams can provide access to segments of the community that have not participated even in traditional *online* ways (e.g., web browsing and email). Further, mining a diverse real-time feed of social streams related to real-world events could enable officials to make better sense of the vast amount of information that is generated by local organizations and the public. In so doing, government should be able to act more effectively on matters both routine (e.g., ongoing issues of public concern) and critical (e.g., major weather or traffic disruption, public safety or rapid response).

The results of social media data analysis have the potential to treat questions or issues more fully than the gather-and-report style of journalism involving traditional sources. When, where and how are local events or issues unfolding? What are the different views on a given event or concern? Which social media should government use to communicate most effectively with a diverse public? How should messages be formed and framed across social media to be effective? To what extent can messages in social networks be used to explain how influential messages form and spread? Who are the influential users in an online or local community? Is civic information, disseminated through social media as opposed to through the Web or email, more likely to reach some traditionally underrepresented groups, such as those with lower socio-economic status (SES) or younger voters? What role do social media play in the general mix of information sources for citizens to communicate about civic life, with each other and with government? Does social media use affect civic participation differently than traditional online channels?

In December 2010, we submitted a proposal to the National Science Foundation's Program on Human-Centered Computing to conduct a full investigation of social media use and impacts, especially related to crisis situations, in Arlington and the National Capitol Region. Our stated research objectives in the proposed work are to study the use and impact of social media and to identify and develop methods to effectively meet a variety of local government and community needs. To achieve these goals we proposed to study comprehensively the use and impact of social media in Arlington, Virginia and environs. Specifically, we will crawl, collect, aggregate, and archive relevant social media data; administer a stratified random sample household survey; conduct focus group interviews with key stakeholders, community leaders and the public; and develop tools to analyze and render data more usable and meaningful for local governments and citizens.

In this larger investigation based on the pilot study, we seek to address a combination of technical and social science challenges through this research; on the technical side, these include: 1) recognizing relevant information accurately and in a timely manner, especially short content from micro-blogging sites (e.g., Twitter); the limited information in a tweet (i.e., less than 140 characters) makes it difficult to identify its meaning and context which may lead to incorrect classification and misleading analysis of tweet data;

2) alerting government officials to the analyzed information from multiple social media sources in real-time; due to the massive volume of the social media data stream, it is a challenge to quickly analyze the collected information from different sources and to make a decision based on the analysis; and

3) visualizing the current and past status of incoming information and the analysis of it; simple yet informative visualization design is essential in making-sense of the data presented. We support the sense-making process by incorporating interaction methods with the visualization to deal with the large amount of data. By mining content and services covering multiple media types (i.e., text, audio, image, video) we can develop tools to recognize events and alert government, citizens, and community groups to see quickly the 'big picture' through visualizations of social media activity and content and changes in both over time.

On the social science side, we build on social network analysis and social and political participation research to understand the use and impact of social media (i.e., civic awareness, collective efficacy, engagement). We seek to contribute to crisis informatics research and an understanding of the role of social media in crisis situations, including more mundane crises, such as weather or traffic problems, and in social convergence situations, such as crowds, rallies and other large gatherings.

Regarding the ontology of crisis, tragedy and recovery, once we have developed the final ontology and confirmed its usefulness, we can apply it in various ways. For example, we will classify resources to suit the differing needs of a variety of stakeholder groups, using the CTR ontology. We will use concepts and their relations in the ontology to monitor social media. For example, we will be notified when a disaster event happens and people begin to communicate about it on the social network; that will guide our re-tweeting to suitable community and neighborhood groups. Semantic browsing and query expansion will be provided, too. They should lead to increased user satisfaction with the information retrieval capabilities of CTRnet. To further ensure scalability, and to move toward sustainability, the CTR ontology will be moved to a public space to ensure that interested members of the community will continuously update it for broader use.

In addition to extending our work with ontologies, we will undertake four other types of digital library (DL) related research:

1) While there has been considerable attention given to handling social media services, there has been little focus on integrating these with digital libraries archiving/preservation, and a combination of stream and collection oriented text mining and visualization. We will leverage our formal analysis of DLs and our system-building experience to devise a new theory-based integration of collections and communities, with high quality.

2) While news services at various levels (local, regional, national, international) commonly report on crises and tragedies at those or higher levels, rarely is there aggregation of that information and comparative analysis across events that can lead to generalities and identification of patterns in how information was disseminated and caught public attention, (re) shared, commented upon, and how it affected people over time and space. Understanding and

capturing different patterns enable better understanding, planning for, and posting alerts for events. We will distinguish between expected behavior and what requires the attention of decision makers, citizens, and other authorities. Special patterns will trigger human alerts as well as cause components of our specialized system to act, e.g., to start collecting more information, or to summarize trends to be presented as evidence for a crisis response.

3) Moreover, since social media include text, images, and video, we will research, for each media type and for each multimedia combination, suitable mechanisms for dealing with them (organize, classify, retrieve, present) and making sense of them to aid diverse audiences.

4) Finally, since different people make use of different social media, only by collecting information from all social media can we obtain a clear picture of the activities of the many groups and organizations that operate in our nation's metropolitan areas.

References

[1] Answers.com. http://www.answers.com/.
[2] Arlington transit blog. http://www.arlingtontransit.com/pages/news-events/arlington-transit-blog/.
[3] Boston citizens connect. http://www.cityofboston.gov/doit/apps/citizensconnect.asp.
[4] Commuterpageblog. http://commuter.typepad.com.
[5] Creating, executing and benefiting from social media. http://esnews.wapa.gov/wordpress/?p=539.
[6] Csi: Arlington parking. http://arlingtonparking.blogspot.com/.
[7] Do utilities need social media? http://www.elp.com/index/display/article-display/7864762152/articles/electric-light-power/volume-88/issue-1/sections/do-utilities_need.html.
[8] eRideShare Arlington. http://www.erideshare.com/carpool.php?city=Arlington&dstate=VA.
[9] Gasbuddy: Find local gas prices. http://gasbuddy.com.
[10] Patients like me. http://www.patientslikeme.com.
[11] The power of Twitter and Facebook for utilities. http://www.intelligentutility.com/article/09/12/power-twitter-and-facebook-utilities.
[12] See click fix. http://www.seeclickfix.com.
[13] Singapore health ministry on social media. http://www.futuregov.asia/articles/2010/mar/29/singapore-health-ministrys-successful-foray-social.
[14] Smarter transportation: 10 social media tools to navigate your city. http://mashable.com/2009/09/29/social-media-transportation.
[15] Stackoverflow. http://www.stackoverflow.com/.
[16] Teaching with wikipedia and other wikimedia foundation wikis. http://www.wikisym.org/ws2010/Teaching+with+Wikipedia+and+other+Wikimedia+Foundation+wikis.
[17] Wikipedia in the writing classroom. http://www.suite101.com/content/wikipedia-a164366.
[18] Zimride. http://www.zimride.com/.
[19] N. Augar, R. Raitman, and W. Zhou. Teaching and learning online with wikis. In *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference*, pages 95–104. Citeseer, 2004.
[20] M. Chui, M. Lffler, and R. Roberts. The internet of things. *McKinsey Quarterly*, 2 2010.
[21] E. Cochran, J. Lawrence, C. Christensen, and R. Jakka. The Quake-Catcher Network:

Citizen Science Expanding Seismic Horizons. *Seismological Research Letters*, 80(1): 26, 2009.
[22] T. A. R. Cross. Social media in disasters and emergencies, Aug. 2010. http://www.redcross.org/www-files/Documents/pdf/other/SocialMediaSlideDeck.pdf.
[23] P. Konieczny. Wikis and Wikipedia as a teaching tool. *International Journal of Instructional Technology & Distance Learning*, 4(1): 15–34, 2007.
[24] A. Opsahl. County experiments with monitoring social media in emergencies, 9 2010. http://www.emergencymgmt.com/disaster/NC-County-Monitoring-Social-Media-Emergency.html.
[25] T. Sakaki, M. Okazaki, and Y.Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In*WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, New York, NY, USA, April 2010. ACM.
[26] J. D. Sutter. Your computer as an earthquake sensor, 9 2010. http://articles.cnn.com/2010-09-17/tech/ibm.earthquakes_1_sensors-computer-earthquake?_s=PM:TECH.
[27] WikiPedia. Public safety, 12 2010. http://en.wikipedia.org/wiki/Public_safety.