Journal of the Association for Information Systems
JAIS

Research Article

An Information Diffusion-Based Recommendation Framework for Micro-Blogging*

Jiesi Cheng
University of Arizona
chengj@email.arizona.edu

Aaron Sun
University of Arizona
asun@email.arizona.edu

Daning Hu
University of Zurich
hdaning@gmail.com

Daniel Zeng
University of Arizona
zeng@email.arizona.edu

Abstract

Micro-blogging is increasingly evolving from a daily chatting tool into a critical platform for individuals and organizations to seek and share real-time news updates during emergencies. However, seeking and extracting useful information from micro-blogging sites poses significant challenges due to the volume of the traffic and the presence of a large body of irrelevant personal messages and spam. In this paper, we propose a novel recommendation framework to overcome this problem. By analyzing information diffusion patterns among a large set of micro-blogs that play the role of emergency news providers, our approach selects a small subset as recommended emergency news feeds for regular users. We evaluate our diffusion-based recommendation framework on Twitter during the early outbreak of H1N1 Flu. The evaluation results show that our method results in more balanced and comprehensive recommendations compared to benchmark approaches.

Keywords: *Micro-blogging, Recommender System, Information Diffusion.*

* Kalle Lyytinen was the accepting senior editor. This article was submitted on 20th May 2010 and went through three revisions.

Volume 12, Issue 7, pp. 463-486, July 2011

1. Introduction

Micro-blogging – a new paradigm of Web-based and mobile application – is experiencing rapid growth and gaining explosive popularity worldwide. Compared to traditional blogging, micro-blogging allows a more instant and flexible form of communication. Micro-blogging sites typically restrict the length of posted messages. These messages can be published and received via a wide variety of means, including the Web, text messaging, instant messaging, and other third-party applications. Such a flexible and broad-based architecture significantly lowers the threshold for participation and encourages users' frequent updates. Consequently, the public has widely adopted micro-blogging to share/seek real-time information, especially during emergency events. For example, in the early stages of the recent H1N1 Flu (Swine Flu) outbreak, the volume of H1N1 Flu-related messages on Twitter – one of the most popular micro-blogging sites – increased 1,500 times over four days (Apr 24 ~ 27, 2009), and accounted for nearly 2 percent of all Twitter traffic in that time period (Nielsen Online, 2009). Meanwhile, a large number of people turned to Twitter searching for the latest updates on the outbreak, causing the keyword "Swine Flu" to be listed as the "top trending topic" on Twitter Search consistently.

On the other hand, the exponentially expanded micro-blogging community results in a tremendously large and constantly updated information stream repository, making it increasingly difficult for users to find content of interest. During emergencies, seeking newsworthy and timely information can be difficult due to the explosion in the volume of micro-blog postings. There is a need for users to be able to find relevant and timely information efficiently, such as through. a search function. The current real-time search functions available in Twitter and some major search engines, such as Google and Bing, allow users to input a query, and then return the latest updates, which contain the search keywords. The most advanced real-time search is able to return updates posted seconds before the search is performed (Singhal, 2009). Nevertheless, the typical search results simply rank updates in reverse chronological order. Therefore, the quality of the search results fluctuates with the timing of the search action.

The approach proposed in this study is to leverage the social network feature in micro-blog communities. In Twitter, if a user u follows user v, all v's updates will be displayed on u's home page. In other words, with user u's subscription of user v's micro-blog, all v's updates are instantly "pushed" to u. The "feed subscription" allows users to receive the latest updates instantly. Our observation is that there exist numerous micro-bloggers who play the role of "news reporters" during emergency events by posting instant news stories on their micro-blogs. We empirically observe that these reporters operate in social settings: they re-broadcast and refer to news stories from one another, maintaining strong interlinking to facilitate rapid diffusion of news stories. Our intuition is, if we could understand how these reporters capture news stories during their diffusion processes, we could effectively measure the importance of each reporter from various diffusion perspectives (e.g., the number of diffusions captured and/or the average time needed for the capture) and make recommendations.

Therefore, in this paper, we formulate the task of navigating micro-bloggers to their desired information as a recommendation problem. As such, instead of letting users actively perform searches, we aim to identify a small number of quality "news reporters" and recommend them to information seekers as emergency news feeds. Such a task is distinctly different from standard content-based and link-based recommendation investigated in the blogging domain with the primary task of "finding blog articles of interest that are not viewed yet" for users. (1) User interest in the blogging context could be extracted from user history. However, in many cases, especially in emergency contexts, information seekers in micro-blogging communities are not necessarily the contributors of the discussions on the emergency events. (2) Information seekers expect to find timely information from micro-blogging communities, especially regarding emergency events. With consideration of the unique challenges, we propose a novel information diffusion-based framework to deal with the unique characteristics and requirements of micro-blogging recommendation. Specifically, we develop diffusion-based metrics to evaluate micro-bloggers, formulate the recommendation problem into a multi-objective optimization problem, and subsequently propose a diffusion-based candidate selection algorithm to recommend quality micro-bloggers during time-critical events. The purpose of the recommendation is not to let information seekers read the

retrospective updates; instead, we aim to enable the users to receive future relevant tweets posted by the recommended micro-bloggers immediately after the updates are posted.

The rest of this paper is organized as follows. We begin by reviewing major micro-blogging applications and relevant recommendation techniques in a blogging context in Section 2. In Section 3, we propose a diffusion-based micro-blogging recommendation framework that utilizes information diffusion patterns. We then present an empirical study to illustrate the potential usefulness and practical value of this diffusion-based recommendation method in Section 4. Finally, we discuss contributions and future directions in Section 5.

2. Literature Review

2.1. Micro-Blogging

Since its launch in 2006, Twitter has become the largest and most well-known micro-blogging platform. As such, Twitter is an ideal candidate site for our study. Twitter allows users to send text-based posts (tweets) that are up to to 140-characters to a network of followers via a variety of means. By default, tweets are public so that users can follow and read each other's posts without permission. Early studies in this area have focused on understanding the prevalent usage and structural patterns of micro-blogging. Java, Song, Finin, and Tseng (2007) studied the topological and geographical properties of Twitter's social network and summarized different user intentions for using Twitter, such as daily chatting and information sharing. Also focusing on the social networking aspects, Krishnamurthy, Gill, and Arlitt (2008) characterized distinct classes of Twitter users and their behaviors, including “broadcasters” (e.g., online radio stations and media outlets), “acquaintances” (users who exhibit reciprocal relationships), and “miscreants” (e.g, spammers).

Recent studies have shifted the attention to some novel micro-blogging applications. For instance, Jansen, Zhang, Sobel, and Chowdury (2009) studied Twitter as a platform for online word-of-mouth branding. They analyzed more than 10,000 micro-blog posts containing branding information and claimed that micro-blogging could play an important role in designing marketing strategies and campaigns. Ehrlich and Shami (2010) and Zhang, Qu, Cody, and Wu (2010) discussed the adoption and use of micro-blogging in the workplace – enterprise micro-blogging. By analyzing users’ posting activities and reading behaviors, they found that enterprise micro-blogging could facilitate conversation and mutual assistance. Such user-to-user exchanges and collaborations via Twitter were also identified in a public setting (Honeycutt & Herring, 2009) in which the authors explored the potential to use Twitter as a collaboration tool.

The rich textual data that are freely available from Twitter also attract interest from the text mining community. O'Connor Balasubramanyan, Balasubramanyan, and Smith (2010) applied the sentiment analysis technique to extract public opinions and attitudes from a large body of tweets. They compared the results with opinions derived from standard polling and survey data, which highlighted the promise of using Twitter as a substitute or supplement for traditional polling. Jansen et al. (2009) used similar, but simpler, techniques to understand user opinion fluctuations toward a particular brand.

Another important application of micro-blogging that is of our interest is its widespread adoption and use during mass crises and emergency events. Though traditional official and media communication channels remain in place, Web-based social media, such as online forums, blogs, and micro-blogs have emerged as alternative forms of rapid dissemination of information (Brownstein, Freifeld, & Madoff, 2009). Apart from the H1N1 F example presented above, micro-blogging has been widely used for status updates and live news reports on occasions of emergency such as during the Southern California wildfires in 2007 (Sutton, Palen, & Shlovski, 2008), the Mumbai terrorist attack in 2008 (Caulfield & Karmali, 2008), the H1N1 Flu outbreak in 2009 (Ostrow, 2009), the Icelandic volcano eruption in 2010 (Nigam, 2010), and others. Such emergency usages of micro-blogging have received increasing attention from academic researchers.

Hughes, Starbird, and Palen (Hughes & Palen, 2009; Starbird & Palen, 2010) were among the first researchers to study this phenomenon. They observed Twitter usage patterns surrounding emergency events and compared those with regular use patterns. They noted that information propagation was more likely to happen in emergency situations than in regular situations. Hughes and Palen (2009) and Starbird and Palen (2010) took advantage of the popularity of Twitter and monitored incoming tweets to detect crisis events such as earthquakes and epidemic outbreaks. These applications clearly indicate micro-blogging's role transition from a daily chatting tool into a valuable information sharing platform during emergencies. However, as mentioned earlier, the explosion in the volume of messages can pose a significant challenge for finding noteworthy information in a timely manner. The occurrence of an emergency compounds this problem when a considerable number of unplanned messages arrive in a short period of time. As such, we propose using a recommender system to alleviate this problem of information overload. In the next subsection, we discuss our research motivations, starting with a review of relevant literature on blog recommendation.

2.2. Blog Recommendation

To our knowledge, this paper presents the first study on micro-blog recommendation, and there has been limited published work in this area. The closest related work to ours is the blog recommender system that has been extensively studied in the literature. In this subsection, we review previous studies related to blog recommendation services only. For a comprehensive review of the recommender system, especially its application in the e-commerce domain, interested readers can refer to Herlocker, Konstan, Terveen, and Riedl (2004) and Schafer, Konstan, and Riedi (1999).

There exist two major types of blog recommendation techniques: content-based and link-based recommendation (Abbassi & Mirrokni, 2007). The core of the content-based recommendation is suggestion of an item (e.g., a blog article or a blogger) to the reader based upon the degree of match between the content description of the item and user interest. A majority of blog recommendation methods can be grouped into this category. The simplest approach is to pre-label blogs to facilitate understanding and categorization. For example, Technorati (www.technorati.com) fetches blog posts that are associated with user-defined tags. Articles under the same tag-based category are then presented to interested readers. Another common approach is to represent a blog article as a term-frequency vector, and to use a scoring system to calculate the distance between this article and user interest, which is also represented as a vector.

Arguello, Elsas, Callan, and Carbonell (2008), developed different document representation models for recommending blogs in response to a user query. Li, Yan, Fan, Liu, Yan, and Chen (2009) developed an incremental vector-space clustering method to identify new topics from the incoming stream of blog articles. The article that best represented a given topic was then selected and recommended to the reader. Note that for blogs annotated by descriptive tags, the contents can be directly characterized using tag vectors (Hayes, Avesani, & Veeramachaneni, 2007).

Blog articles can also be transformed into a tree-like or graph-like hierarchical ontology. Ontology is defined as a formal specification of a shared conceptualization consisting of entities, attributes, and relationships. Nakatsuji, Miyoshi, and Otsuka (2006) used Web Ontology Language to extract user-interest ontology from blog articles. Then they applied an ontology-based similarity measurement to cluster bloggers whose interests were alike. Readers could then find like-minded bloggers through the personalized recommendations that were made. El-Arini, Veda, Shahaf, and Guestrin (2009) carried out a similar study to achieve a different recommendation objective. The authors characterized the blog postings by various semantic features, such as name entities, topics identified from the corpus, and their high-level relations. A set of blogs was then selected and recommended that covered the most features (best coverage).

At the other end of the spectrum, link-based recommendation is implemented on the basis of explicit or implicit network structures extracted from the blogging community. Explicit connections among blogs include hyperlinks from one blog to another and mentions/comments about other blogs in blog entries. Implicit connections, on the contrary, do not physically exist. Instead, they are inferred

artificial links that supplement the explicit connections. Such connections are usually weighted, with the weight indicating the likelihood for two blogs being connected in a certain manner. In the literature, various structural features can be used to estimate the common factors among individual bloggers or blog articles. Abbassi and Mirrokni (2007) measured the similarity among blogs using the eigenvalues of the adjacency matrix of an explicit blog graph. They later applied a spectral clustering method to partition relevant blogs and recommend.

Kritikopoulos, Sideri, and Varlamis (2006) developed a modified version of PageRank to rank nodes on the blog graph by their ranking scores. To address the sparsity problem of the hyperlink-based graph, they created a denser graph by incorporating artificial weighted links that denoted the similarity among bloggers into the original graph. Chau and Xu (2007) proposed a semi-automated approach that consists of a set of Web mining and network analysis techniques to identify influential opinion leaders in hate groups.

Although the link structure is generally useful for blog recommendations, we found that only a limited number of studies are purely link-based. More frequently, structural features are used in combination with content features to improve the outcomes, which leads to a hybrid form of recommendation. Hsu, King, Paradesi, Pydimarri, and Weninger (2006) proposed such a hybrid approach by using both the link structure and interests declared by bloggers. Another hybrid method reported in Li and Chen (2009) considered both link and content information. They also took into account a third dimension – the trust among bloggers – to enhance the reliability of the recommender system.

As we have discussed before, micro-blogging differs significantly from regular blogging in its extensive use during emergency situations. As such, our research faces distinctive challenges. In a blogging context, the primary task of recommendation is to “find blog articles of interest that are not viewed yet” for users. However, this consideration is not applicable in a micro-blogging context, especially during emergencies, because the lightweight design of micro-blogging tends to generate an overwhelming volume of messages that are inefficient to process (Kristina, 2009). In addition, a substantial number of these messages are related to personal conversations that have little value to the general public, such as one's own fears about the H1N1 Flu epidemic. Another noteworthy phenomenon in micro-blogging is that there exist numerous “news reporters” who regularly post the latest news stories on their micro-blogs, mostly on a voluntary basis. These reporters provide important information filtering and amplification services and can be effectively leveraged for recommendation. We believe that it is more convenient to recommend these reporters to ordinary users as live news feeds, rather than to recommend individual postings. In the next section, we will present our diffusion-based recommendation framework.

3. An Information Diffusion Based Recommendation Framework

3.1. Information Diffusion and Diffusion-Based Valuation Criteria

Information diffusion through online social networks has recently become an active research topic. In blogging communities, the propagation of information from one blog to the next is frequently observed, as a result of low-cost information sharing and publishing. Gruhl et al. (2004) examined the information propagation pattern from 11,000 blog sites at two different levels: individual-level diffusion among blog entries and community-level diffusion among blogspaces. They then adapted a cascade model to characterize individual behaviors in different stages of diffusion. Adar and Adamic (2005) analyzed the internal link structure of blogspace to track the flow of information among blog entries. In particular, the authors addressed the problem of “infection inference,” focusing not only on utilizing the explicit link structure, but also inferring implicit routes of infection/diffusion. In the end, they built a diffusion tree to visualize the likely routes of transmission for a specific diffusion. Such diffusion patterns are also prevalent among micro-blogs. Lerman and Ghosh (2010) studied the diffusion of news stories on Twitter and compared it with the diffusion on Digg in terms of rate and scope. It was noticed that the diffusion on Twitter generally maintained a consistent rate and penetrated farther than on Digg.

Yang and Counts (2010) studied Twitter users' interaction behaviors. One important finding of the study was that the majority of the interactions were one-way rather than reciprocal. This finding was also supported in a recent study (Starbird & Palen, 2010), which claimed that news stories were more likely to be distributed from media outlets and traditional service organizations, and then spread into the population.

Our observations on Twitter confirmed this uni-directional information flow. We noticed that during the outbreak of H1N1 Flu, first-hand news stories normally originated from a limited number of professional news agencies (e.g., BBC and Reuters) and public health organizations (e.g., CDC Emergency), though exceptions exist. These stories then propagated across the community through the process of reposting (retweeting) or commenting. Figure 1 illustrates such a news story diffusion example corresponding to the tweets collected during the early outbreak of H1N1 Flu (listed in Table 1.) Clear story diffusion can be traced from the source "CDCEmergency" to other Flu news reporters (as indicated by their account IDs) within 36 hours after the story was first posted by the source.

During a news story's diffusion process, any micro-blog that posts the story is called a participant who "captures" the story. If we want to recommend no more than k micro-blogs as emergency news feeds (k is an exogenous parameter that is reasonably small to avoid information overload), these diffusion patterns could be helpful because they will reveal how each micro-blog participated in past diffusion processes. Intuitively, a micro-blog is more likely to be favored and recommended during emergencies if it captures news stories of interest more accurately and rapidly. More specifically, we use the following measures to quantify various aspects of this valuation process.

(1) Story Coverage (SC). Multiple news stories regarding one broad topic can simultaneously spread. For example, when the H1N1 Flu outbreak occurred, CDCEmergency reminded people that they would not get infected from eating pork, while ForbesNews was concerned about the Flu's impact on financial markets. The SC metric measures the set of stories of interest captured by micro-blogs under study, and stronger recommendation is given to micro-bloggers that have wider coverage, other conditions being identical.

(2) Reading Effort (RE). Certain micro-blogs can be crowded with messages. Too many messages compromise readability and raise the cognitive effort to filter out irrelevant content. The RE measurement indicates the set of messages one has to read after subscribing to the selected micro-blogs.

(3) Delay Time (DT). The postings of one news story s on micro-blogs are time-stamped. The delay time of s equals the time passed from the first appearance of s in the community until s is captured by one of the selected micro-blogs, usually measured in hours. In our application setting, a delayed capture of s is certainly undesirable.

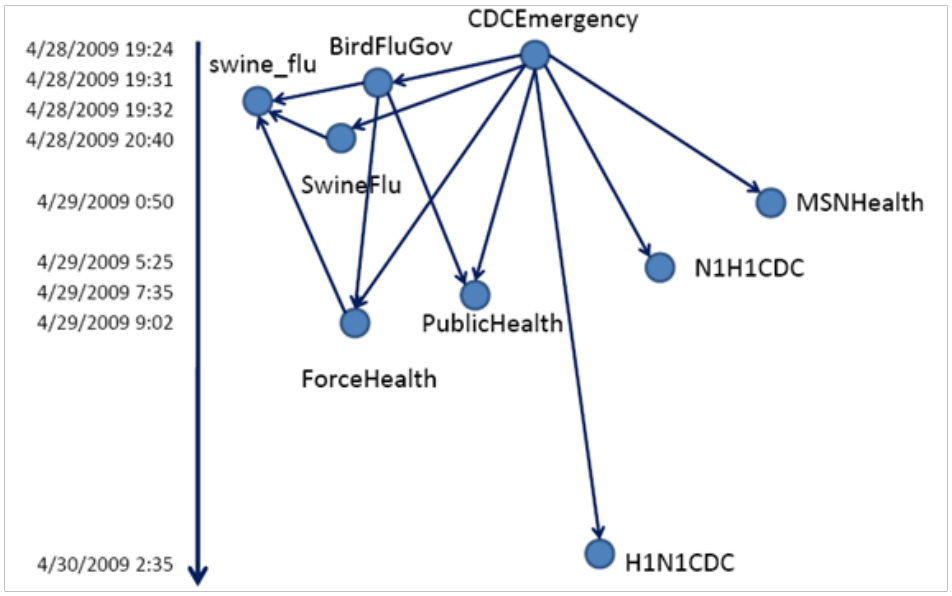

Figure 1. A Story Diffusion Example

Table 1. Story Diffusion among Micro-blogs

Timestamp	Micro-blog	Tweet Content
4/28/2009 19:24	CDCEmergency	CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/28/2009 19:31	BirdFluGov	RT @CDCemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/28/2009 19:32	swine_flu	RT @CDCemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/28/2009 20:40	SwineFlu	RT CDCemergency: CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu http://ow.ly/4kKe
4/29/2009 0:50	MSNHealth	RT @cdcemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/29/2009 5:25	N1H1CDC	CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/29/2009 7:35	PublicHealth	@cdcemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/29/2009 9:02	ForceHealth	RT @CDCemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/29/2009 14:44	kcnews	RT @CDCemergency CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1 #swineflu
4/30/2009 2:35	H1N1CDC	CDC reminds you that you can NOT get swine flu from eating pork. http://bit.ly/16YpY1

Obviously there are other criteria that can also evaluate the quality of a recommendation, but in this study we focus exclusively on these three basic measures. The above discussions are illustrated in Figure 2 where rounded rectangles represent micro-blogs, and solid/open shapes represent news stories relevant/irrelevant to the user interest. Directed links indicate the flow of a news story, and each link is associated with a numeric label representing the units of time spent. We define a diffusion path as the route through which a news story flows from its source to other micro-blogs. In Figure 2,

three diffusion paths co-exist with each corresponding to a news story. In terms of recommending micro-blogs, many combinations turn out to have the largest coverage, such as the sets $\{m1, m2, m3\}$, $\{m3, m5\}$, and $\{m7\}$. However, when considering other measurements, each combination has its own advantages and limitations. $\{m1, m2, m3\}$ is relatively weak in RE but competitive in DT; $\{m3, m5\}$ and $\{m7\}$ have a better RE score but spend longer DT in capturing each story. Such examples show that it is possible to customize the combination of micro-blogs and achieve various recommendation objectives.

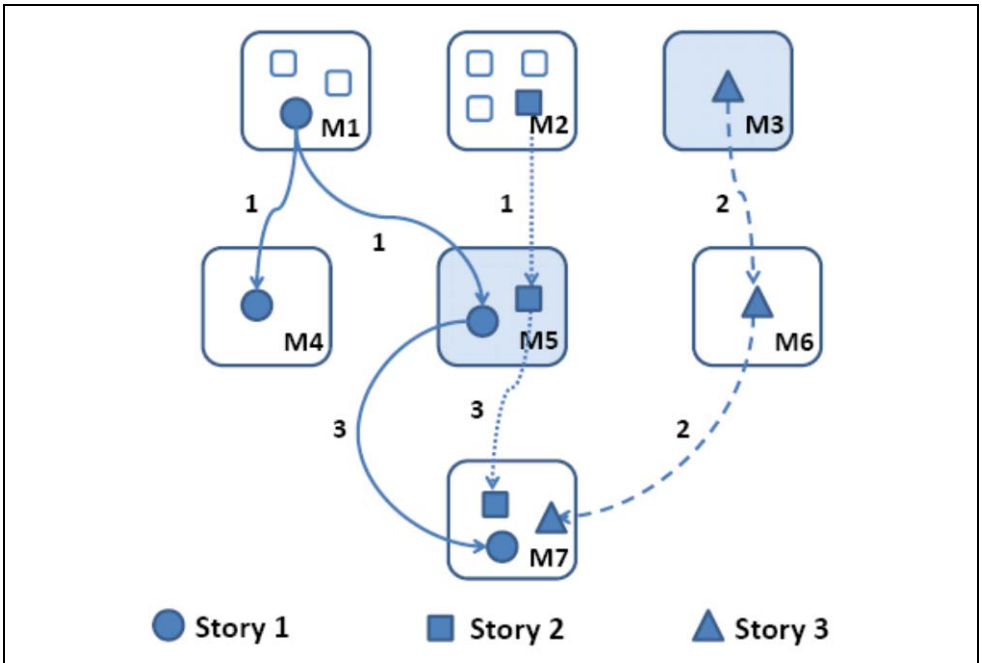

Figure 2. An Example of Diffusion Path

3.2. The Recommendation Framework

Inspired by the work of Krause, Leskovec, Guestrin, VanBriesen, and Faloutsos (2008) on outbreak detection in water distribution networks, we have developed a diffusion-based recommendation approach. In general, our approach is to design a set function b, and associate each recommendation set I with a real number $\mathrm{b}(\mathrm{I})$, the benefit score. Given user preferences, we intend to maximize this benefit score and recommend the user an optimized result set. To this end, we have quantitatively assessed each candidate micro-blog m from the above three diffusion-related aspects. Suppose that we can identify a set of major news stories $\mathbb{S}$ during a certain period and subsequently reconstruct their diffusion paths, we denote the raw observations for micro-blog m as $\mathrm{SC}(\{\mathrm{m}\})$, $\mathrm{RE}(\{\mathrm{m}\})$, and $\mathrm{DT}(\{\mathrm{m}\})$. As described in the previous section, $\mathrm{SC}(\{\mathrm{m}\})$ is a subset of $\mathbb{S}$ whose member stories are captured by $\{\mathrm{m}\}$; $\mathrm{RE}(\{\mathrm{m}\})$ is the set of posts one has to receive and read from $\{\mathrm{m}\}$ in order to discover $\mathrm{SC}(\{\mathrm{m}\})$; and $\mathrm{DT}(\{\mathrm{m}\})$ indicates a set of delay time corresponding to each story s in $\mathbb{S}$. In case story s is not captured, we let $\mathrm{DT}_s(\{\mathrm{m}\}) = \infty$. Based on these raw observations, we assign a score vector $\overline{b(\{m\})} = \left(\mathrm{b}_{\mathrm{SC}}(\{\mathrm{m}\}), \mathrm{b}_{\mathrm{RE}}(\{\mathrm{m}\}), \mathrm{b}_{\mathrm{DT}}(\{\mathrm{m}\})\right)$ to micro- blog m, which quantifies the benefit incurred by following m based on the diffusion histories of $\mathbb{S}$. The components of the score vector are set functions $\mathrm{b}_x(\cdot), X \in \{\mathrm{SC}, \mathrm{RE}, \mathrm{DT}\}$ that can transform the raw observation sets into real numbers. More concretely, $\mathrm{b}_{\mathrm{SC}}(\{\mathrm{m}\})$ equals the number of stories covered by $\{\mathrm{m}\}$, denoted by $|\mathrm{SC}(\{\mathrm{m}\})|$. This set function takes a simplifying assumption that all diffusion stories are of equal importance. In practice, we have also developed an improved set function by assigning an importance weight $\mathrm{w}(s)$ to story s. The importance weight $\mathrm{w}(s)$ could be defined as the number of micro-blogs in the community who capture s, which can be estimated from the constructed diffusion paths. Based on this

weighted function, $b_{SC}(\{m\})$ equals $\sum s \in SC(\{m\})\, w(s)$, and a micro-blog gains a higher benefit score for capturing more "important" stories. For the second component of the score vector, we define $b_{RE}(\{m\}) = -|RE(\{m\})|$, where $|RE(\{m\})|$ is the size of set $RE(\{m\})$. Note that RE takes a negative value such that smaller $|RE(\{m\})|$ can lead to a higher benefit. The last component of the score vector transforms the $DT(\{m\})$ set as follows: for any $s \in S$, we assume that the score of DT drops exponentially with increased delay time, determined by $b_{DT_s}(\{m\}) = e^{-\beta \cdot DT_s(\{m\})}$ where β is a positive scalar. The cumulative score for the entire story set S is then defined as the sum of individual scores, $b_{DT}(\{m\}) = \sum s \in S\, e^{-\beta \cdot DT_s(\{m\})}$. Note that if none of the stories is captured by $\{m\}$, we consider no benefits can be obtained from following m so m is removed from the candidate set.

The score vector can be also associated with multiple micro-blogs to measure their aggregated benefits. Given a set of micro-blogs M, $SC(M)$ and $RE(M)$ can be expressed as $\cup_{m \in M} SC(\{m\})$ and $\cup_{m \in M} RE(\{m\})$ respectively. For any $s \in S$, the delay time of s by subscribing M is the minimum delay time of subscribing $m \in M$, denoted as $\min_{m \in M}(DT_s(\{m\}))$. $DT(M)$ can now be readily written as $\cup_{s \in S}\left(\min_{m \in M}(DT_s(\{m\}))\right)$. Finally, a score vector $\overrightarrow{b(M)} = (b_{SC}(M), b_{RE}(M), b_{DT}(M))$ can be built based on these raw measurements.

3.3. Multi-criterion Optimization

We now formulate the micro-blog recommendation problem in an optimization framework. Given the diffusion story set s and the micro-blog set M, we aim to identify a subset of, at most, k micro-blogs $I \subseteq M$ and $|I| \leq k$ that optimizes the benefit score of $\overrightarrow{b(I)}$. In this optimization problem, we intend to simultaneously optimize multiple objectives. However, the three objectives might be in conflict, and the situation can arise that two recommendations I and J are incomparable, e.g. $b_{SC}(I) > b_{SC}(J)$, but $b_{DT}(I) < b_{DT}(J)$. In this case, we use the Pareto-optimality (Boyd & Vandenberghe, 2004). A recommended set I is called Pareto-optimal if there does not exist another recommendation J such that $b_X(J) \geq b_X(I)$ for all measurements $X \in \{SC, RE, DT\}$, and $b_Y(J) > b_Y(I)$ for at least one measurement $Y \in \{SC, RE, DT\}$. One common approach for finding such Pareto-optimal sets is scalarization (Boyd & Vandenberghe, 2004; Hayes et al., 2007). By choosing weights λ_{SC}, λ_{RE} and λ_{DT}, we can optimize an objective function $b(M) = \sum_{X \in \{SC,RE,DT\}} \lambda_X \lambda_X(M)$ as an alternative to $\overrightarrow{b(M)}$. Any solution that optimizes $b(M)$ is guaranteed to be Pareto-optimal to $\overrightarrow{b(M)}$, and by adjusting exogenous parameter λ_X, our recommendation can take full consideration of all three aspects but also allow varying degree of emphasis depending on user preferences.[1] In practice, some users might expect to receive as many timely posts as possible to get informed on everything new about the topic of interest without caring too much about being flooded with many posts, while others expect to read the minimum number of posts to capture the overall story of the events of interest. These different preferences could be operationalized as optimization problems with different weights λ_X on the three objectives. In the next subsection, we prove that this scalarized objective function is submodular. In general, submodular function optimization is NP-hard (Khuller, Moss, & Naor, 1999). We then propose a greedy algorithm as an effective heuristic solution.

3.4. A Heuristic Greedy Algorithm

Consider an arbitrary function f that maps subsets of a finite ground set U to real numbers. We call f submodular if it shows a "diminishing returns" property: "The marginal gain from adding an element to a set A is at least as high as the marginal gain from adding the same element to a superset of A (Kempe, Kleinberg, & Tardos, 2003)." Formally, this submodular property can be expressed as: $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$ for all elements v and all pairs of sets $A \subseteq B$. For submodular objective functions, a greedy algorithm is frequently used for obtaining a bounded approximation guarantee (Nemhauser, Wolsey, & Fisher, 1978). The optimization objective in our recommendation problem satisfies the submodular property as well, which can be proved as follows.

[1] In practice, we apply a Z-transformation to score vectors so that the mean and variance (mean = 0 and standard deviation = 1) of scores are equivalent across different categories.

Theorem 1: the scalarized objective function $b(I)$ *is submodular*.
Theorem 1.1: the set function $b_{SC}(\cdot)$ *is submodular*.
Theorem 1.2: the set function $b_{RE}(\cdot)$ *is submodular*.
Theorem 1.3: the set function $b_{DT}(\cdot)$ *is submodular*.

Proof. See the appendix.

Intuitively, in the micro-blogging context, the submodular property can be understood as: reading a micro-blog after we have only read a couple of blogs provides more information (and other benefits) than reading it after we have read many micro-blogs.

The submodular property can be utilized to develop a greedy hill-climbing algorithm that approximates the optimum of the problem to within a factor of $(1 - 1/e)$ (where e is the base of the natural logarithm) (Nemhauser et al., 1978). We have followed this approach and developed a greedy algorithm, illustrated in Algorithm 1. This algorithm starts with the empty recommendation set, and repeatedly adds a micro-blog to the set that maximizes the benefit score. The algorithm stops once k micro-blogs are selected or the incremental benefit is less than a predefined small value ε. In the next Section, we evaluate this diffusion-based recommendation method using a recent emergency case on Twitter.

Function: Greedy ($S, M, k, \varepsilon, \lambda_{SC}, \lambda_{RE}, \lambda_{DT}$)
$I \leftarrow \emptyset$, $\Delta \leftarrow +\infty$, $B \leftarrow 0$;
While $\exists i \in M \backslash I : |I \cup \{i\}| \leq k$ and $\Delta > \varepsilon$ do
    $i^* \leftarrow \mathbf{argmax} \sum_{X \in \{SC, RE, DT\}} \lambda_X b_X(I \cup \{i\})$
    $I \leftarrow I \cup \{i^*\}$
    $\Delta \leftarrow \sum_{X \in \{SC, RE, DT\}} \lambda_X b_X(I) - B$
    $B \leftarrow \sum_{X \in \{SC, RE, DT\}} \lambda_X b_X(I)$
Return I

Algorithm 1. A Greedy Algorithm

4. An Empirical Study

4.1 The H1N1 Flu Dataset

We collected data from Twitter.com using its API from May 10 to May 16, 2009 during the early outbreak of H1N1 flu. We used keywords “swine flu” and “h1n1” to search Twitter every 15 minutes throughout the week. Each time Twitter search provided a maximum of 1,500 real-time messages ranked by their published time, and we identified 1,034 unique accounts who had mentioned either keyword more than five times during that week. We then continued to retrieve all of each user's available tweets (up to 3,200 historical tweets.) In our data set, for a majority of users, 3,200 tweets were more than adequate to cover their two-month histories, which means that we were able to collect these users' near-complete tweets since the outbreak of H1N1 Flu (late April, 2009.) In the end, we collected a total of 1,308,800 tweets from the 1,034 candidate accounts, among which, 35,091 tweets contained the keywords “swine flu,” or “flu,” or “h1n1.” We refer to these tweets as H1N1-related tweets hereafter.[2]

[2] The set of documents (tweets) that are relevant to the topic of interest is an essential input for constructing diffusions. Therefore, the task of document retrieval on the topic of interest is an important step of preprocessing in our recommendation framework. In many cases, different terms are used to describe the same topic. In our particular case study, we chose to use “swine flu,” “flu,” and “H1N1” as the keywords for collecting the basic pool of tweets, which involved a manual selection based on domain

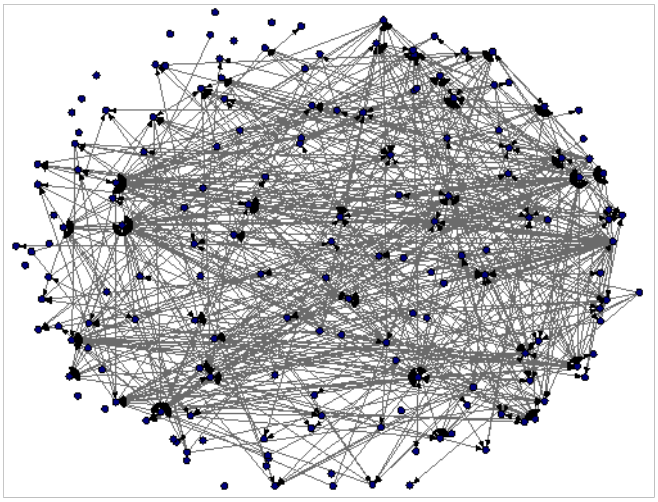

Figure 3. Twitter Network

As we mentioned earlier, each Twitter user u can maintain a list of friends and followers. For these 1,034 accounts, we also included their friend and follower connections in our dataset. We used a directed graph (Figure 3) to show the relationships among these candidate accounts. The graph contains 1,034 nodes and 6,876 directed links. (If v is a friend of u, equivalently, u is a follower of v. We establish a linkage from v to u to indicate the direction of information flow.) By the time we collected data, most of these accounts had a large community of followers. Although many of the followers were not included in our dataset, the resulting subgraph was still well connected, which has enabled the formation of diffusion.

4.2 Research Design

Our detailed experimental design is illustrated in Figure 4. We first divided all H1N1-related tweets into two groups by their published time. We placed tweets time-stamped from Apr 26 to May 2, 2009 (week 1) in the first group, which we used as training data to feed the recommendation algorithm. We used tweets posted in the two weeks immediately after that (from May 3 to May 16, 2009) as testing data to evaluate the performance of the recommended micro-bloggers. With the above setting, we intended to simulate the recommendation made for the trending topic "swine flu" on May 3th, and then to evaluate the quality of the recommended micro-bloggers via their tweets in the subsequent two weeks.

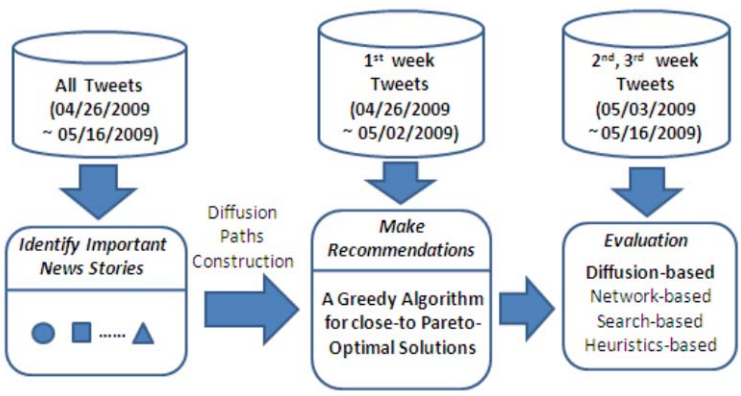

Figure 4. Research Design

knowledge. In practice, the selection of relevant keywords for the same general topic could be automated using classic approaches from information retrieval literature such as Latent Dirichlet Allocation (LDA) (Blei, Ng., & Jordan, 2003).

For each group of tweets, we conducted a term frequency analysis after tokenizing, stemming, and removing stopwords. We used the common English stopword list provided by Natural Language Toolkit (NLTK) (Loper & Bird, 2002). We also filtered out additional common words that were not helpful for distinguishing stories such as "flu" and "h1n1." Term frequencies for tweets in Group 1 are visualized in a word cloud (Figure 5). In this figure, the size of each word is arranged proportionally according to its term frequency.[3] For simplicity, we only considered frequent terms that appeared five times or more.

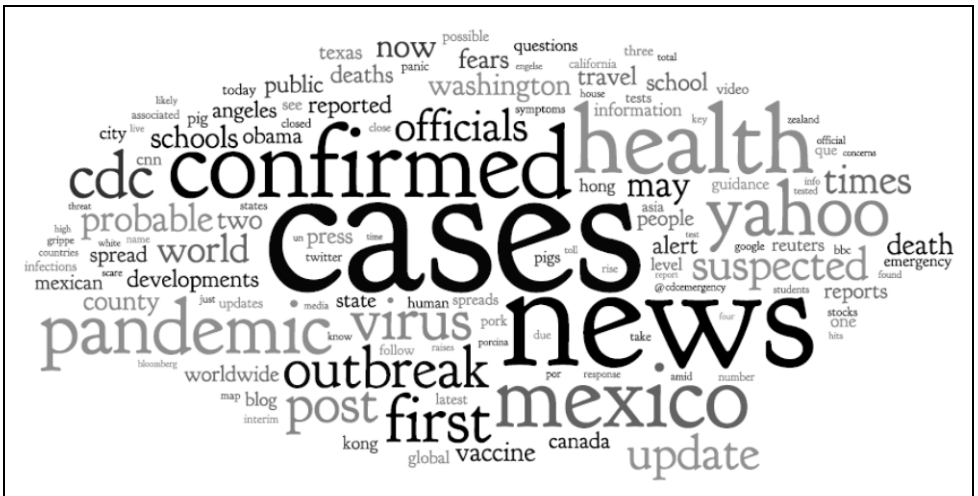

Figure 5. Word Cloud for Tweets from Apr 26, 2009 to May 3, 2009

Each H1N1-related tweet was classified into one of the two groups and represented as a term frequency vector. We then implemented top-down hierarchical clustering techniques for story detection within each group. We first applied the K-means clustering algorithm provided by NLTK to partition each group of vectors into k segments. However, the K-means clustering did not generate stable outputs. In practice, we found that it would return dissimilar clustering results at each run with a given k. In addition, it was difficult to accurately estimate the number of clusters for different inputs. As a result, we set k to be a small number (e.g., $k = 20$), so that each generated cluster contained tweets that represented a broad topic, and tweets about the same story were less likely to be separated into different clusters. These clusters were used as intermediate results, and we applied extended Jaccard similarity to further subdivide each cluster into smaller ones if possible. In the clustering literature, Jaccard similarity and its extensions are appropriate for dealing with our high-dimensional, sparse term frequency data. We, therefore, calculated the extended Jaccard similarity between every two vectors $\overrightarrow{tv_1}$ and $\overrightarrow{tv_2}$, within each cluster, denoted by $J(\overrightarrow{tv_1}, \overrightarrow{tv_2}) = \frac{\overrightarrow{tv_1} \cdot (\overrightarrow{tv_2})^T}{|\overrightarrow{tv_1}|^2 + |\overrightarrow{tv_2}|^2 - \overrightarrow{tv_1} \cdot (\overrightarrow{tv_2})^T}$ (Strehl, Ghosh, & Mooney, 2000). These vectors were then be visualized as a complete graph, with an artificial link connecting two vectors as nodes, and the similarity score being the weight of the link. We continued to remove those links whose weights were less than a given threshold, and extracted all connected components (subgraphs) of the reduced graph. Each connected component was viewed as a finer-level cluster. In practice, we chose the threshold value to be between 0.5 and 0.7. The clustering results were quite satisfying, owing greatly to the fact that tweets are short in length. As a tweet only contains at most 140 characters, it is typical that each tweet centers on a single news story. Additionally, users frequently repost stories with only minor modifications, which means a set of shared terms tend to be used by different users for describing one common story.

We subsequently selected those "important" stories in each group that had been posted by three or more distinct accounts. There were 59 and 287 important stories identified in group 1 and 2, respectively. Each story was stored with a rich set of metadata: users who posted this story, friend/follower relationships among these users, and time of postings. In the next step, we used these

[3] Word cloud layout algorithm by Jonathan Feinberg, Wordle.net and IBM Research, http://www.wordle.net/credits.

metadata to construct diffusion paths for each story. Various methods have been developed to identify information diffusion paths from the history of user interactions (Adar & Adamic, 2005; Gruhl et al., 2004). We adopted the following rules to approximate a diffusion path for each news story s. (1) Explicit referring: In Twitter, a tweet can include the use of "RT" and/or "@username" to indicate that this tweet is a repost of one of the username's earlier tweets. As such, if user u posted story s after user v did, and u referred v's username explicitly, s was then assumed to flow from v to u. (2) Implicit referring: According to (Adar & Adamic, 2005), if user u followed user v, and u frequently posted the same stories after user v did, then we assume diffusions from v to u. Such inferred diffusion routes can be identified by running the association rule-mining algorithm across all tweet clusters. (3) Unknown referring: When neither condition above is satisfied, we assume that u received the story from a dummy "Real World" node (Gruhl et al., 2004). As a result, for both groups 1 and 2, we have obtained the corresponding diffusion paths for identified stories, and the benefit scores of SC, RE, and DT for each participant micro-blog were calculated for the tasks of recommendation and evaluation. Table 2 summarizes the major data features.

Table 2. Grouping and Clustering Data

	Group 1	Group 2
Time Period	4/26/2009 ~ 5/2/2009	5/3/2009 ~ 5/16/2009
Total Number of Tweets	13,416	21,679
Term Frequency Vectors Dimension	2538-dimensional	3612-dimensional
Number of Stories Identified	59	287
Size of the Largest Cluster	11	17
Average Size of the Clusters	6.17	6.84

4.3 Evaluation Results and Discussions

Given the recommendation objectives and the quantitative measurements of each micro-blog, we selected a close-to Pareto-optimal set of $k = 3$ micro-blogs using the Greedy Algorithm proposed in Section 3.4. To demonstrate the effectiveness of the algorithm, we evaluated the algorithm against four representative user preferences.

(a) $\lambda_{SC} = 1, \lambda_{RE} = 1, \lambda_{DT} = 1$. In this setting, we calculated the total benefit score by placing equal emphasis on the three aspects. In other words, we intended to recommend micro-blogs that capture as many important news stories as possible at a relatively early time, and moreover, accompanied by as few irrelevant or spam tweets as possible.

(b) $\lambda_{SC} = 1, \lambda_{RE} = 1, \lambda_{DT} = 0$. This setting only took SC and RE into consideration. Namely, late capture was acceptable.

(c) $\lambda_{SC} = 1, \lambda_{RE} = 0, \lambda_{DT} = 0$. In this setting, we exclusively focused on the coverage of important stories.

(d) $\lambda_{SC} = 1, \lambda_{RE} = 0, \lambda_{DT} = 1$. In this setting, the reading effort could be compromised, and the aspects of SC and DT were equally weighted.

Meanwhile, we also selected another six recommended sets using benchmark methods. The performance evaluations for all candidate sets using the news stories identified in Group 2 are listed in Table 3 and Figure 6.

The candidate micro-blogs selected by our diffusion-based recommendation algorithm varied with user preferences, whereas in all four settings, they obtained higher benefit scores than those obtained from using benchmark methods. “YourDNAknows,” “swineflualerts,” and “SwineFluPanic” were selected for setting (a). These three micro-blogs achieved a balanced performance in all three measures. In setting (b), by ignoring delay time, we were able to achieve even higher story coverage and lower reading effort, but the average time of capturing stories was delayed by more than two hours. When we exclusively considered the story coverage, micro-blogs selected in setting (c) captured the highest number of stories. Incidentally, the results of setting (d) were identical to those obtained in (a).

The performances of benchmark methods are also illustrated in Table 3 and Figure 6. As the first benchmark, we used the friend/follower graph (Figure 3) and made recommendations by the top three Authority/Hub scores generated by Hyperlink-Induced Topic Search (HITS) algorithm (Kleinberg, 1999). These “authority” nodes delivered only moderate performances partly because their interests are narrowly specialized. For example, “CDCEmergency” is the official Twitter account for Centers for Disease Control, which only posted two original tweets, on average, per day without retweeting or commenting; “swineflu_help” is an account registered in Mexico, which primarily updated Mexico-related Flu stories. On the other hand, these Authority accounts typically post many original tweets, thus, the delay time is relatively small. The performances of Hub nodes were no better than those of Authority nodes. In Twitter, only tweets posted by a user’s direct friends are pushed to the user’s homepage. Hence, tweets posted by Hub nodes’ friends do not automatically cascade to Hub nodes’ followers unless the Hub nodes re-tweet the updates. The empirical results show that although these Hubs were structurally important, they only played the role of good listeners rather than influential opinion leaders. The total number of posted tweets by the top three Hub nodes in the two-week time period was as low as two per day on average, despite the fact that these top nodes followed many Authorities.

We next used Google Site Search and Twitter “Find People” to select top three ranked results using the query “swine flu.” Both search engines performed reasonably well in terms of SC/RE. Although the performance of each individual micro-blog selected by Google Site Search was satisfactory, the aggregated coverage was low due to content overlap, while our algorithms tended to avoid such an overlapping in order to maximize the coverage. In addition, recommended accounts from Twitter “Find People” had relatively large delays for not considering the temporal factor.

Last, we made recommendations by using two simple heuristics. First of all, the top three users with the largest number of followers were “nytimes” (NYTimes.com), “sanjayguptaCNN” (CNN Chief Medical Correspondent), and “bbcbreaking” (Breaking news alerts from the BBC.) These accounts represented traditional mass media outlets, whose number of followers could range from hundreds of thousands to millions. However, they typically published news stories covering a wide range of topics and underperformed in a specific topic category such as H1N1 Flu. Another set of users was selected based on its total number of tweets posted in the first week. This set performed surprisingly well due to the highest volume of tweets, but this advantage was offset by the low RE scores.

A closer observation reveals that the micro-bloggers selected by the proposed diffusion-based recommendation framework are not the most “popular,” “well-known,” “authority,” and “productive” accounts. In contrast, they are the ones who have narrower focus, who follow many “popular,” “well-known,” “authority,” and “productive” accounts, and who actively and frequently retweet interesting updates originally posted by their friends. Empirical results have demonstrated that these selected micro-bloggers play the role of efficient information aggregator and disseminator, and our proposed diffusion-based recommendation framework is able to identify these high quality micro-bloggers and recommend them to information seekers in the micro-blogging community.

Table 3. Recommendation Performance

Methods		Story Coverage (SC)		Reading Effort (RE)		Delay Time (DT)		Top 3 Candidates Selected
		SC	Weighted SC (%)	Total RE	SC/RE	Average (hrs)	Median (hrs)	
Diffusion-Based	$\lambda_{SC}=1\ \lambda_{RE}=1\ \lambda_{DT}=1$	172	62.22%	785	21.91%	1.89	0.24	'YourDNAknows', 'swineflualerts', 'SwineFluPanic'
	$\lambda_{SC}=1\ \lambda_{RE}=1\ \lambda_{DT}=0$	201	72.96%	369	54.47%	4.08	1.27	'YourDNAknows', 'swineflulatest', 'H1N1CDC'
	$\lambda_{SC}=1\ \lambda_{RE}=0\ \lambda_{DT}=0$	248	87.58%	901	27.52%	3.38	0.84	'YourDNAknows', 'News_SwineFlu', 'SwineFluPanic'
	$\lambda_{SC}=1\ \lambda_{RE}=0\ \lambda_{DT}=1$	172	62.22%	785	21.91%	1.89	0.24	'YourDNAknows', 'swineflualerts', 'SwineFluPanic'
Network-Based	**HITS (Authority)**	65	21.20%	351	18.52%	3.27	0.77	'CDCEmergency', 'h1n1info', 'swineflu_help'
	HITS (Hub)	26	10.70%	91	28.57%	8.68	2.35	'swinevirus', 'h1n1info', 'SWINE_FLU_INFO 1'
Search-Based	**Google Search**	106	41.60%	564	18.79%	2.82	0.62	'stopswineflu', 'swineflualerts', 'swineflu_news_'
	Twitter "Find People"	53	22.80%	331	16.01%	5.08	1.60	'swineflubrk', 'SwineFluTicker', 'DrSwineFlu'
Heuristics-Based	**# of Followers**	6	2.40%	303	1.98%	1.35	0.15	'nytimes', 'sanjayguptaCNN', 'bbcbreaking'
	# of Tweets	144	56.40%	1084	13.28%	4.33	1.18	'h1n1swineflu', 'swineflu2', 'swineflualerts'

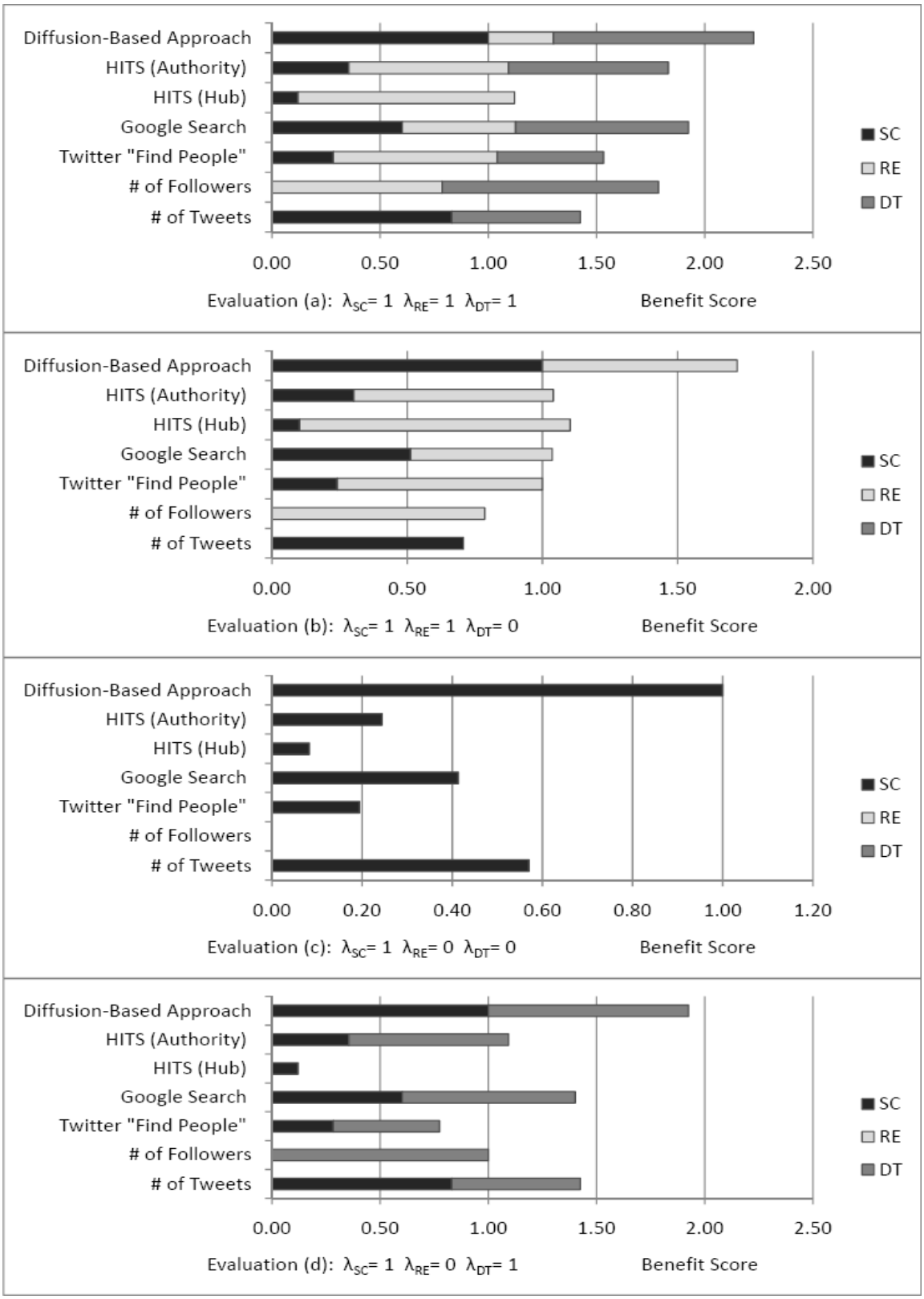

Figure 6. Evaluation Results – Benefit Score

Scores were normalized by max-min transformation.

In addition to benefit scores, we also evaluated the results of the proposed diffusion-based recommendation and the benchmark methods using another set of evaluation metrics. In information retrieval studies, recall and precision are two widely used performance measures (Makhoul, Kubala, Schwartz, & Weischedel, 1999). Recall is computed as the fraction of retrieved and relevant instances among all relevant instances, while precision is the fraction of retrieved and relevant instances among those the algorithm retrieves. In our study, the two metrics can be operationalized with the valuation metrics SC and RE.

$$\text{Recall} = \frac{\text{retrieved and relevant}}{\text{relevant}} = \frac{\text{SC}}{\text{Total \# of relevant stories}}$$

$$\text{Precision} = \frac{\text{retrieved and relevant}}{\text{retrieved}} = \frac{\text{SC}}{\text{RE}}$$

The harmonic mean of recall and precision is often used to integrate the two measures, which is referred as the F-measure.

$$F = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$$

We then adopted the spirit of F-measure to evaluate the micro-bloggers selected by the proposed diffusion-based recommendation framework and the benchmark methods. In addition to recall and precision, the temporal factor is another critical aspect for micro-bloggers' performance evaluation. Hence, we extended the definition of F-measure to integrate the three diffusion-based valuation criteria to evaluate selected micro-bloggers' performance in the emergency context. We define the Extended F-measure, denoted by *F**, as the weighted harmonic mean of *Recall*, *Precision*, and *Average Delay Time.*[4]

$$F^* = \frac{\lambda_{SC} + \lambda_{RE} + \lambda_{DT}}{\frac{\lambda_{SC}}{\text{recall}} + \frac{\lambda_{RE}}{\text{precision}} + \frac{\lambda_{DT}}{\text{AvgDelayTime}}}$$

The evaluation results are displayed in Table 4. Though the diffusion-based approach does not result in the highest score for all the three dimensions (*recall*, *precision*, and *AvgDelayTime*), it outperforms benchmark methods on *F** in all four parameter settings.

Intuitively, one would consider following those popular, official accounts when seeking information from the micro-blogging community in emergency contexts. However, our evaluation has shown that there exist topic-specialized and timely-updated micro-blogs that are worth following. The comparison of our proposed approach with other benchmark methods demonstrated that, especially in a time-critical context, the proposed diffusion-based recommendation framework and the proposed algorithm can provide useful recommendations for finding out these less popular but high quality micro-blogs.

[4] Average Delay Time (DT) is normalized by max-min transformation to be scaled into [0,1].

Table 4. Evaluation Results -- Extended F-measure (F*)

Methods	Performance Measure	Diffusion-Based Approach	Network-Based		Search-Based		Heuristics-Based	
			HITS (Authority)	HITS (Hub)	Google Search	Twitter "Find People"	# of Followers	# of Tweets
$\lambda_{SC}=1$ $\lambda_{RE}=1$ $\lambda_{DT}=1$	Recall	**0.599**	0.226	0.091	0.369	0.185	0.021	0.502
	Precision	0.219	0.185	**0.286**	0.188	0.160	0.020	0.133
	AvgDelayTime	0.926	0.738	0.000	0.799	0.491	**1.000**	0.594
	F*	**0.410**	0.269	0.000	0.323	0.219	0.030	0.268
$\lambda_{SC}=1$ $\lambda_{RE}=1$ $\lambda_{DT}=0$	Recall	**0.700**	0.226	0.091	0.369	0.185	0.021	0.502
	Precision	**0.545**	0.185	0.286	0.188	0.160	0.020	0.133
	AvgDelayTime	0.628	0.738	0.000	0.799	0.491	**1.000**	0.594
	F*	**0.613**	0.204	0.138	0.249	0.172	0.020	0.210
$\lambda_{SC}=1$ $\lambda_{RE}=0$ $\lambda_{DT}=0$	Recall	**0.864**	0.226	0.091	0.369	0.185	0.021	0.502
	Precision	0.275	0.185	**0.286**	0.188	0.160	0.020	0.133
	AvgDelayTime	0.723	0.738	0.000	0.799	0.491	**1.000**	0.594
	F*	**0.864**	0.226	0.091	0.369	0.185	0.021	0.502
$\lambda_{SC}=1$ $\lambda_{RE}=0$ $\lambda_{DT}=1$	Recall	**0.599**	0.226	0.091	0.369	0.185	0.021	0.502
	Precision	0.219	0.185	**0.286**	0.188	0.160	0.020	0.133
	AvgDelayTime	0.926	0.738	0.000	0.799	0.491	**1.000**	0.594
	F*	**0.728**	0.347	0.000	0.505	0.268	0.041	0.544

4.4 Time complexity and Scalability

The diffusion-based recommendation framework is composed of two major computational steps, (1) diffusion detection, and (2) recommendation. The diffusion detection algorithm has a worst-case time complexity of $O(T^2)$, where T is the total number of tweets to process. The recommendation algorithm is a Greedy Algorithm. The worst-case time complexity of the Greedy Algorithm is $O(kns_{\max}M)$, where k is the number of micro-bloggers to recommend, n the number of keywords (combination of keywords) of trending topics, $s_{\max}$ the maximum number of story diffusions detected for each keyword (combination of keywords), and M the size of the candidate micro-bloggers set. Since both k and $s_{\max}$ are usually small, the Greedy Algorithm is fairly efficient even for processing a large number of topics and/or a large network. Hence, the scalability of the recommendation is largely determined by the diffusion detection step. In practice, the diffusion detection process can be carried out offline. When dealing with a huge number of tweets, the computation could be handled by distributed systems to increase efficiency. Moreover, the latest tweets could be processed incrementally to eliminate unnecessary computations.

5. Concluding Remarks and Future Directions

5.1 Contribution

The advance of Internet technologies has provided a wide availability of data sources for us to monitor and capture emerging trends and patterns (Brownstein et al., 2009). As such, Web browsing and searching has become an important means for people to explore and discover such time-critical knowledge (Ginsberg et al., 2009). Meanwhile, Web-based social media, such as online discussion forums, blogs, and micro-blogs have emerged as alternative forms of rapid dissemination of

information. Recent studies on micro-blogging have focused on the role transition of micro-blogging from a social communication tool into an important platform for sharing/seeking up-to-the-second information during emergencies. In this study, we proposed a novel diffusion-based micro-blogging recommendation framework, aiming to recommend micro-bloggers during time-critical events. We developed a set of measures assessing the value of micro-bloggers from a diffusion standpoint and formulated the recommendation problem into a multi-objective optimization problem. We then proved the submodular property of the proposed recommendation objectives to solve this optimization problem, and we adopted a scalarization approach to reduce the dimension of the objectives. The solution to the alternative objective function is guaranteed to be Pareto-optimal of the original problem. We further developed a heuristic greedy algorithm that exploited submodularity to find near-optimal node selections. Though the evaluation metrics in this study are geared toward solving the recommendation problem in emergency response contexts, the underlying idea – evaluating micro-bloggers via information diffusion-based metrics and formulating the recommendation task into a multi-objective optimization problem – could be applied to recommendations in a broader context.

We extensively evaluated our diffusion-based recommendation framework and the proposed algorithm using Twitter data collected during the early outbreak of H1N1 Flu. The empirical results showed that our method outperformed other benchmark approaches, and could achieve a more balanced and comprehensive recommendation. Moreover, the evaluations under different parameter settings demonstrated that our recommendation framework can accommodate diverse user preferences and provide customized results.

The practical contribution of our study lies in a more user-friendly design of the micro-blogging platform to facilitate information seeking during emergency events. The proposed recommendation framework could be implemented as a function on Twitter. On Twitter, a snapshot of the most mentioned topics (trending topics) in the micro-blogging community is displayed in the sidebar of a user's homepage. The snapshot is generated based on Twitter's proprietary algorithm (Cheong & Lee, 2009), and could be modified with user preferences on trending topics of the minute, day, week, etc.

In practice, our proposed micro-blog recommendation could work as follows: when clicking on a topic of interest from the trending topic, a user navigates to a list of recommended micro-bloggers. By following the recommended micro-bloggers the user subscribes to all their future updates. The idea of taking trending topics as the set of keywords for the recommendation has the following advantages compared to letting users input search terms freely: (1) Users do not need to know the topic of interest prior to receiving the recommendation. This is particularly important for bursty events. It is not practical to expect users to input search terms for new events they have never heard of. Hence leveraging the list of trending topics to navigate users for the recommendation is a more natural design. (2) The diffusion detection process with the list of terms in the trending topics is independent from users' requests. Hence, it could be carried out offline. (3) With the trending topics identified from a tweets timeline, the system is prevented from dealing with potential ambiguity introduced by a synonym or typo in user input. (4) Though the proposed implementation only provides recommendations for trending topics, it is reasonable to expect it could handle a majority of users' interests due to the power-law distribution of user interest frequently observed in online communities. Despite the advantages of implementing the recommendation framework as navigation from Twitter trending keywords, there are still other ways to apply the proposed recommendation framework, such as instant search applications for content from micro-blogging sites. Besides being implemented as a Twitter application, it could also be implemented as a third-party Web service that uses Twitter API for retrospective tweets and social network structure.

It was demonstrated that the recommendations made by the proposed diffusion-based framework are capable of pushing timely and relevant updates to the subscribers without flooding them with too many posts. In addition, more comprehensive recommendation could be provided with users' input on the perceived importance of different dimensions including "story coverage," "reading effort," and "time delay." From an interface design perspective, the users' preferences may be adjusted by a slider control.

5.2 Future Work

We conclude this paper by discussing our ongoing and future work. The application of the proposed diffusion-based micro-blogging recommendation framework is not limited to the emergency context. The approach can also be applied to other time-critical tasks. For instance, there exist a number of Twitter accounts that consistently post real-time financial news, such as "FTfinancenews," "ftnewsblog," "fnnews," and others. Likewise, recommendations can be made by measuring these accounts from a diffusion perspective. Beyond the recommendation, the diffusion-based framework could provide extended services such as monitoring and surveillance.

Identifying stories from tweets is a critical stage in our recommendation framework. One limitation of the current study is that we detect stories solely based on the vector-space-model. This method fails to consider the semantics (meaning) of the terms, such that two tweets describing the same story could be assigned to different clusters if they are paraphrased using a different set of terms. Existing studies on semantic text clustering could be helpful for addressing this issue. For example, it has been reported that clustering with semantic features outperformed the term frequency-based method (Choudhary & Bhattacharyya, 2002; Hotho, Staab, & Stumme, 2003). A possible the future direction of study would be to integrate semantic text clustering approaches into the task of story identification such that more accurate diffusion patterns could be inferred to support the recommendation task.

Another future direction of study could be to develop new valuation criteria for more comprehensive recommendations. For example, users might be only interested in news stories occurring in locations geographically close to them. In this case, geographical proximity can be developed to measure the distance between the recommended micro-bloggers and the reader. At the time we collected data for this study, we could only extract self-proclaimed location information from public user profiles. Such data can be incomplete or inaccurate. In March 2010, Twitter launched its new Geolocation feature which enabled users to track the latitude and longitude of tweets as long as this feature is activated. Should the tweet-level location data be available for analysis, future research could incorporate this location-related metric for more comprehensive recommendations.

Last but not least, to better evaluate the perceived usefulness of the proposed recommendation framework, future research could conduct a user study to capture user experiences using information seeking in micro-blogging communities with diffusion-based recommendations.

Acknowledgements

The research reported in this paper was partially supported by the Chinese Academy of Sciences through grants #2F07C01 and #KGCX2-YW-122-05; the National Natural Science Foundation of China grants #71025001, #90924302, #91024030, #70890084, #60875049, #60921061; and the Ministry of Health grants #2009ZX10004-315 and #2008ZX10005-013.

References

Abbassi, Z., & Mirrokni, V. S. (2007). A recommender system based on local random walks and spectral methods. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis,* 102-108.

Adar, E., & Adamic, L.A. (2005). Tracking information epidemics in Blogspace. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 207-214.

Arguello, J., Elsas, J.L., Callan, J., & Carbonell, J.G. (2008). Document representation and query expansion models for blog recommendation. *Proceedings of the 2nd International. Conference on Weblogs and Social Media (ICWSM)*.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Boyd, S., & Vandenberghe, L. (2004). Convex Optimization. *Cambridge University Press*.

Brownstein, J.S., Freifeld, C.C., & Madoff, L.C. (2009). Influenza A (H1N1) Virus, 2009 - Online Monitoring. *The New England Journal of Medicine,* 360, 2156.

Caulfield, B., & Karmali, N. (2008). Mumbai: Twitter's moment. *Forbes.com*.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: a study of online hate groups. *International Journal of Human-Computer Studies,* 65(1), 57-70.

Cheong, M., & Lee, V. (2009). Integrating Web-based intelligence retrieval and decision-making from the Twitter trends knowledge base. *Proceedings of the 2nd ACM workshop on Social Web Search and Mining*, 1-8.

Choudhary, B., & Bhattacharyya, P. (2002). Text clustering using semantics. *Proceedings of the 11th International World Wide Web Conference (WWW).*

Ehrlich, K., & Shami, N.S. (2010). Microblogging Inside and Outside the Workplace. *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

El-Arini, K., Veda, G., Shahaf, D., & Guestrin, C. (2009). Turning down the noise in the blogosphere. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), ACM*, 289-298.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M. S., & Brilliant, L.(2009). Detecting influenza epidemics using search engine query data. *Nature,* 457, 1012-1015.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through Blogspace. *Proceedings of the 13th International World Wide Web (WWW) Conference*, 491-501.

Hayes, C., Avesani, P., & Veeramachaneni, S. (2007). An analysis of the use of tags in a blog recommender system. *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI'07)*, 2772-2777.

Herlocker, J.L., Konstan , J.A., Terveen, L.G., & Riedl, J.T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS),* 22(1), 5-53.

Honeycutt, C., & Herring, S.C. (2009). Beyond Microblogging: conversation and collaboration via Twitter. *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS).*

Hotho, A., Staab, S., & Stumme, G. (2003). Text clustering based on background knowledge. *Technical report, University of Karlsruhe, Institute AIFB*. Report No. 425.

Hsu, W.H., King, A.L., Paradesi, M.S.R., Pydimarri, T., & Weninger, T. (2006). Collaborative and structural recommendation of friends using Weblog-based social network analysis. *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 55-60.

Hughes, A.L., & Palen, L. (2009). twitter adoption and use in mass convergence and emergency events. *Proceedings of the 6th International ISCRAM Conference*.

Jansen, B.J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Micro-blogging as online word of mouth branding. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems (CHI),* 3859-3864.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding Microblogging usage and communities. *Proceedings of the 9th WEBKDD and 1st SNA-KDD Workshop on Web mining and social analysis.*

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Khuller, S., Moss, A., & Naor, J. (1999). The budgeted maximum coverage problem. *Information Processing Letters,* 70(1), 39 – 45.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment.*Journal of ACM,* 46(5), 604-632.

Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., & Faloutsos, C. (2008). efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6), 516-526.

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about Twitter. *Proceedings of the 1st workshop on online social networks (WOSN'08)*.

Kristina, M. (2009). Bursts of Information: Microblogging. *The Reference Librarian,* 50(2), 212-214.

Kritikopoulos, A., Sideri, M., & Varlamis, I. (2006). BlogRank: ranking weblogs based on connectivity and similarity features. *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications, ACM Press*.

Lerman, K., & Ghosh, R. (2010). Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

Li, X., Yan, J., Fan, W., Liu, N., Yan, S., & Chen, Z. (2009). An online Blog reading system by topic clustering and personalized ranking. *ACM Transactions on Internet Technology (TOIT)*.9(3).

Li, Y.-M., & Chen, C.-W.(2009). A synthetical approach for blog recommendation: combining trust, social relation, and semantic analysis. *Expert Systems with Applications: An International Journal,* 36(3), 6536-6547.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching NLP and Computational Linguistics*, Philadelphia, PA, 63-70.

Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999). Performance measures for information extraction. *Proceedings of DARPA Broadcast News Workshop*, 1999.

Nakatsuji, M., Miyoshi, Y., & Otsuka, Y. (2006). Innovation detection based on user-interest ontology of blog community. *Lecture Notes in Computer Science,* 4273, 515-528.

Nemhauser, G., Wolsey, L., & Fisher, M. (2009). An Analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1), 265-294.

Nielsen Online. (2009). Swine Flu news and concern dominates online buzz. http://blog.nielsen.com/nielsenwire/online_mobile/swine-flu-news-and-concern-dominates-online-buzz/

Nigam, S. (2010). How Social Media Helped Travelers During the Iceland Volcano Eruption. http://mashable.com/2010/04/22/social-media-iceland-volcano/

O' Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From tweets to polls: linking text sentiment to public opinion time series. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

Ostrow, A. (2009). Swine Flu Hysteria: 10,000 Tweets Per Hour. http://mashable.com/2009/04/27/swine-flu-twitter/.

Schafer, J.B., Konstan, J., & Riedi, J. (1999). Recommender systems in e-commerce. *Proceedings of the 1st ACM conference on Electronic commerce*, Denver, Colorado, United States 158-166.

Singhal, A. (2010). Relevance meets the real-time Web. *The Official Google Blog*. http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html

Starbird, K., & Palen, L. (2010). Pass it on?: retweeting in mass emergency. *Proceedings of the 7th International ISCRAM Conference*.

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on Web-page clustering. *Proceedings of the AAAI Workshop on AI for Web Search (AAAI)*, 58-64.

Sutton, J., Palen, L., & Shlovski, I. (2008). Back-channels on the front lines: emerging use of social media in the 2007 southern California wildfires. *Proceedings of the 2008 ISCRAM Conference*, *Washington, DC.*

Yang, J., & Counts, S. (2010). Interaction network properties of Twitter. *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.

Zhang, J., Qu, Y., Cody, J., & Wu, Y. (2010). A case study of micro-blogging in the enterprise: use, value, and related issues. *Proceedings of the 28th international conference on Human factors in computing systems* (CHI), ACM, New York, NY, USA, 123-132.

Appendix. Proofs of Theoretical Results

Theorem 1: the scalarized objective function $b(I)$ is *submodular*.

Theorem 1.1: the set function $b_{SC}(\cdot)$ is *submodular*.

Proof: Let M_1 and M_2 be two subsets of micro-blog set M, and $M_1 \subseteq M_2 \subseteq M$, such that $SC(M_1) \subseteq SC(M_2)$. Let m be an arbitrary micro-blog that does not belong to M_1 or M_2, $\exists m \in M, m \notin M_1$ and $m \notin M_2$. The value of $b_{SC}(M_1 \cup \{m\}) - b_{SC}(M_1)$ is the number of posts that are in $SC(\{m\})$ but not in $SC(M_1)$, i.e. $b_{SC}(M_1 \cup \{m\}) - b_{SC}(M_1) = |SC(\{m\})| - |SC(M_1) \cap SC(\{m\})|$. This number is at least as large as the number of posts that are in $SC(\{m\})$ but not in $SC(M_2)$:

$$\because M_1 \subseteq M_2 \therefore |SC(M_2) \cap SC(\{m\})| \geq |SC(M_1) \cap SC(\{m\})|.$$

$$\begin{aligned}\therefore b_{SC}(M_1 \cup \{m\}) - b_{SC}(M_1) &= |SC(\{m\})| - |SC(M_1) \cap SC(\{m\})| \\ &\geq |SC(\{m\})| - |SC(M_2) \cap SC(\{m\})| = b_{SC}(M_2 \cup \{m\}) - b_{SC}(M_2)\end{aligned}$$

Theorem 1.2: the set function $b_{RE}(\cdot)$ is *submodular*.

Proof: Let M_1 and M_2 be two subsets of all micro-blogs and $M_1 \subseteq M_2 \subseteq M$, such that $RE(M_1) \subseteq RE(M_2)$. Let m be an arbitrary micro-blog that does not belong to M_1or M_2. The value of $b_{RE}(M_1 \cup \{m\}) - b_{RE}(M_1) = -|RE(\{m\})| = b_{RE}(M_2 \cup \{m\}) - b_{RE}(M_2)$. Therefore, the set function of RE is modular, which also satisfies the submodular property.

Theorem 1.3: the set function $b_{DT}(\cdot)$ is *submodular*.

Proof: Let M_1 and M_2 be two sets of micro-blogs and $M_1 \subseteq M_2 \subseteq M$, such that for any story $s \in S, DT_s(M_1) \geq DT_s(M_2)$ (i.e. M_2 never takes a longer time to capture s than M_1.) Let m be an arbitrary micro-blog that does not belong to M_1 or M_2. For any story $s \in S$, there exist four possibilities:

(1) $\{m\}$ does not capture s. In this case, $DT_s(\{m\}) = \infty$;

(2) $\{m\}$ captures s, but not earlier than either M_1 or M_2, $DT_s(\{m\}) \geq DT_s(M_1)$ and $DT_s(\{m\}) \geq DT_s(M_2)$

Under these two conditions, $DT_s(M_1 \cup \{m\}) = DT_s(M_1)$ and $DT_s(M_2 \cup \{m\}) = DT_s(M_2)$, therefore, $b_{DT,s}(M_1 \cup \{m\}) - b_{DT,s}(M_1) = b_{DT,s}(M_2 \cup \{m\}) - b_{DT,s}(M_2) = 0$.

(3) $\{m\}$ captures s, but not later than either M_1 or M_2, $DT_s(\{m\}) \leq DT_s(M_1)$ and $DT_s(\{m\}) \leq DT_s(M_2)$;

Under this condition, $DT_s(M_1 \cup \{m\}) = DT_s(M_2 \cup \{m\}) = DT_s(\{m\})$, and, therefore, $b_{DT,s}(M_1 \cup \{m\}) - b_{DT,s}(M_1) \geq b_{DT,s}(M_2 \cup \{m\}) - b_{DT,s}(M_2)$.

(4) $\{m\}$ captures s earlier than M_1 but later than M_2, $DT_s(\{m\}) < DT_s(M_1)$ and $DT_s(\{m\}) > DT_s(M_2)$.

Under this condition, $DT_s(M_1 \cup \{m\}) = DT_s(\{m\}) < DT_s(M_1)$ and $DT_s(M_2 \cup \{m\}) = DT_s(M_2)$, hence $b_{DT,s}(M_1 \cup \{m\}) - b_{DT,s}(M_1) > 0 = b_{DT,s}(M_2 \cup \{m\}) - b_{DT,s}(M_2)$.

$\because b_{DT}(M_1 \cup \{m\}) - b_{DT}(M_1) = \sum_{s \in S} b_{DT,s}(M_1 \cup \{m\}) - b_{DT,s}(M_1)$
and $b_{DT}(M_2 \cup \{m\}) - b_{DT}(M_2) = \sum_{s \in S} b_{DT,s}(M_2 \cup \{m\}) - b_{DT,s}(M_2)$

According to (1)~(4), $b_{DT}(M_1 \cup \{m\}) - b_{DT}(M_1) \geq b_{DT}(M_2 \cup \{m\}) - b_{DT}(M_2)$.

As proved, three individual objectives satisfy the submodularity property. Since submodularity is close under nonnegative linear combinations, the new scalarized objective is also submodular.

About the Authors

Jiesi CHENG is currently a Ph.D. student in Management Information Systems in the University of Arizona, Tucson, AZ. She received her B.E. degree in Electrical Engineering from Beijing University of Aeronautics and Astronautics, Beijing, China. Her research interests include social network analysis, text mining, and viral marketing.

Aaron SUN received the Ph.D. degree in Management Information Systems from the University of Arizona, Tucson, AZ. He received his B.S. degree in Computer Science from Najing University of Aeronautics and Astronautics, China. His research interests include quantitative modeling, data mining and social network analysis.

Daning HU received his Ph.D. degree in Management Information Systems from Eller College of Management, the University of Arizona, Tucson, Arizona, USA. He is currently an assistant professor in the Department of Informatics, University of Zurich. His research interests include network modeling and analysis for financial markets, business intelligence, and social computing. His work has been published or accepted by Decision Support Systems, Journal of the American Society for Information Science and Technology, and Information Systems Frontier.

Daniel ZENG received the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University and the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China. He is a Research Professor at the Institute of Automation in the Chinese Academy of Sciences and a Professor and Eller Fellow in the Department of Management Information Systems at the University of Arizona. His research interests include intelligence and security informatics, infectious disease informatics, social computing, recommender systems, software agents, and applied operations research and game theory. He has published one monograph and more than 180 peer-reviewed articles. He has also co-edited 15 books and proceedings, and chaired many conferences including the IEEE International Conference on Intelligence and Security Informatics (ISI), the Biosurveillance and Biosecurity Workshop (BioSecure), and the International Workshop on Social Computing (SOCO). He serves on editorial boards of 15 IT journals. He is also active in information systems and public health informatics professional activities and is Vice President for Publications for the IEEE Intelligent Transportation Systems Society and Chair of INFORMS College on Artificial Intelligence. His research has been mainly funded by the U.S. NSF, the NNSF of China, the Chinese Academy of Sciences, U.S. DHS, and MOST and MOH of China.