USING MICROBLOGS FOR CROWDSOURCING AND PUBLIC OPINION MINING

A Thesis Presented

by

CÜNEYT GÜRCAN AKÇORA

Submitted to the Graduate School of the
University At Buffalo, The State University of New York in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

August 2010

Computer Science and Engineering

ABSTRACT

USING MICROBLOGS FOR CROWDSOURCING AND PUBLIC OPINION MINING

AUGUST 2010

CÜNEYT GÜRCAN AKÇORA

B.Sc., KARADENIZ TECHNICAL UNIVERSITY, TRABZON, TURKEY

MS, UNIVERSITY AT BUFFALO, THE STATE UNIVERSITY OF NEW YORK

Directed by: Professor Murat Demirbas

With the advance of social networking sites, citizen publishing has taken the lead in providing a large part of information we daily consume. In this study, we present data mining and crowdsourcing strategies that can efficiently mine, classify and use this noisy information. In the RainRadar project, active mining that requires user contribution upon querying is presented for weather forecasts. We crowdsource the weather forecast for cities in USA and Canada, and show the results on a web site. In the Upinion project, we use Twitter user posts and employ sentiment analysis to detect changes in public opinion. We created a system that can map emotions to eight classes, and built a corpus to populate those classes.

ACKNOWLEDGMENTS

I thank the Fulbright Commission for the scholarship that allowed me to fulfill my dreams. My heartfelt thanks go to Professor Murat Demirbas for advising me throughout my studies.

I am also grateful to work with a few of the Turkey's smartest people in the Ubiquitous Computing Lab; Murat Ali Bayir, Onur Soysal and Yavuz Selim Yilmaz.

I thank Hakan Ferhatosmanoglu for his help in the projects.

Anneme, Cemo'ya. Babama, biliyordur.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

INTRODUCTION

Internet had its early days with static pages, but it did not stop there. We have seen the rise of interactive pages that users could specify their needs for information, and soon after, social networking sites emerged to enable users create their own content to serve other Internet users. Facebook and Twitter have been the flagships of social networks that serve user created contents.

While the information we daily consume have started to come from those social networks, the noisy nature of this information became an obstacle to a better understanding, and a need to classify and mine this information grew. Big companies like Google, Microsoft and Yahoo has taken a step forward and started using social networks to improve their key technologies. All these search engines added *recency information* of user posts to their search modules. These modules mine social networks in real time to find newly emerging topics, events and web sites to serve their users in a timely fashion. By using the recency information mined from social networks, these companies stay ahead of the curve.

In this study, we take the lead in creating efficient data mining and crowdsourcing techniques to fully exploit the recency information of the user created content. We categorize our data mining efforts as active and passive data mining, where the active mining requires querying users and the passive case exploits the content that is already posted by the user. Crowdsourcing is used to implement active data mining for weather forecasts.

In the RainRadar project, we ambitiously set out to integrate mobile phones, sensors and web interfaces to query users about a simple, yet elusive question; where is it raining? In this work, we created a **bio** code that can seamlessly integrate all those devices, and employ user querying techniques to **actively mine** weather data with very fine granularity. An accompanying web site serves the users with an interactive map to show the weather conditions in various cities. In this pioneering work, we found that users reply to our queries in a very timely fashion, and the participation rate is very satisfactory.

To show the benefits of passive data mining where we can tap into the vast social network posts that are publicly available, we created the Upinion project. The project can keep track of public opinion about a topic by simply monitoring posts and mapping them to different emotion classes. The project accepts topic subscriptions, and has an interactive interface to show users posts from each emotion class about a topic.

In our work, we showed opportunities that crowdsourcing and data mining can bring when carefully utilized. Our work resulted in two web sites to convey information about our techniques. We expect crowdsourcing and active user participation to play a bigger role on Internet, and this study contributes to implementing these technologies in real life.

CHAPTER 1
EVOLUTION OF MICROBLOGGING AND TWITTER

1.1 Microblogging roots

Everyone will be tuned in to everything that's happening all the time! No one will be left out. We'll be all normal! [15]. These lines were published in a cartoon by Robert Crumb in 1960's, but he was wrong to visualize wires attached to our heads. Current day microblogging has its popularity thanks to Twitter.com, but roots go deeper than that.

We are delving into the world of 140 characters, microblogging. This fascinating world is constantly changing and evolving into something more complex. Nobody can see where it will take us yet, but for many, Twitter is only the tip of an iceberg, a gateway to something unknown as DOS was to the Windows. While the world is still contemplating its possible future, many researchers took a hands-on approach, writing influential papers on Twitter.

More than just mentioning those papers, we are giving the basics of microblogging, explaining Twitter.com to details, saving you from the trouble of connecting thousands of dots on Internet with this report.

It was both breathtaking and intimidating to read and record all articles and papers on Twitter and microblogging. It was breathtaking, because of the influence of microblogging in many fields ranging from news reporting to alert systems and education. It is intimidating as well, because almost all sources are volatile. For the report, we had to mention many online blogs, newspaper articles but they might disappear or be moved on Web. We do not see this as a shortcoming in the report, because the very name of Twitter comes from chirping sounds of birds. It is only fitting that our sources, just like those chirping sounds, would be a part of our world for a short time and disappear, leaving us with new notions, like ambient intimacy[55].

The real time text messaging began with Instant Relay Chat (IRC) in 1988. IRC was mainly designed for group communications, but also served as a basis for one-to-one communication. IRC was a great success and had a very similar period to Twitter's famous Iran election coverage when it was used to break a media blackout by Soviet Union in 1991. Popular characters that are used in microblogging, such as # and @, have their roots from IRC.

Another push towards a microblogging community came with mobile phones. With mass production of mobile phones starting from 1990's, mobility became possible and text messaging without a computer gained popularity. Soon after, companies started to push for new applications that utilized mobility. Mobile applications for instant messaging or news reports were developed. The main issue that surrounded mobility was to enable people to share information and collaborate.

Recently, researchers have been analyzing the user motivation to use microblogging and searching sites. A study by Broder shows three basic kinds of search queries: navigational, informational and transactional [10]. Navigational queries are performed to reach a particular site the user is looking for. Informational queries are used to find resources on the Web. Transactional queries are used to locate shopping or download sites where further interactions would take place.

In one of the first studies about microblogging, Java et al. aim to find the reason why users post microblogs on Twitter [37]. The paper lists microblogging users' intentions as *daily chatter, conversations, sharing information and reporting news.* Yet, the major reason for microblogging comes from a need to reach out to new, interesting and even expert people that we do not have a chance to meet in real life.

Researchers [55] call this ambient intimacy, and the notion addresses a new trend in public. So far for human relations, an approximation of 150 people is given as a limit to have stable social relations with. This number is called Dunbar's number, and it is named after the British anthropologist Robin Dunbar, who theorized this limit in relation to the neocortical processing capacity of human brain.

With microblogging, this number can be obsolete. We need stable social relations with people to exchange information, emotions and mundane things. What if there is a way to receive information and all the other things without getting to know the other person? Microblogging brings a *personal API* to us, and we can use it to explore minds from everywhere. The relation is there of course, but it is more of an *Intermittent, one-way, more of a crush than love; the feeling of a shared relationship without it being shared, a benevolent form of stalking* [18].

Microblogging offers a way to get past Durban's number. With the advance of search engines, people can find resources ranked in an order that is based on a majority usage rule, yet sometimes an expert's view, or an influential person's ideas might be worth more than all other people's. However, we do not have the required time, energy and mobility to find these people. Microblogging sites eventually address this desire by simply giving users a chance to explore the world's people.

Apple fans no longer need to be friend with Guy Kawasaki to learn what he thinks, likes and reads. Like many other celebrities, politicians and sportsmen, Guy Kawasaki has a personal API on twitter. Even for our relation with the people around us, microblogging gives us a peek at their lives, it increases our interaction level with them. When carefully observed, the future of microblogging can be seen in the Japanese mobile phone culture, *keitai*.

The Japanese text messaging, with a 1000 character limit that is bigger than usual, provided the Japanese youth with a rich character set that can visualize moods, actions, and opinions. In Japan, mobile phone usage has gone to the point of being banned altogether on campuses and public places. Twitter on the other hand, provides us with a smaller text space than usual, but the space can include links to other texts, pictures and videos.

What keitai enabled with characters are also enabled with links on twitter, and furthermore, unlike text messaging, microblogging posts can be seen by more than one people, reacted upon, favorited and re-posted in public. Whether microblogging will be banned in public yet remains to be seen, but we are not that far away from the activity level of keitai culture.

1.2 A Brief History

"When you can measure what you are speaking about, and express it in numbers, you know something about it", Lord Kelvin.

A web 2.0 project, Facebook.com established a status update field in June 2006, but it was Twitter that took status sharing between people to our mobile phones four months later. First named as *Status*, then as *Twttr*, Twitter has been evolving in its creator Jack Dorsey's mind since 2000. Dorsey wrote an application in 2000 that checked an email address for updates and notified his friends.

Dorsey had limited resources until 2005-2006 when SMS took off in USA. In its essence, Twitter is a virtual implementation that mimics how people move in a city. As Dorsey explains, "*Twitter has conceptual roots in the world of vehicle dispatch – where cars and bikes zooming around town must constantly squawk to each other about where they are and what they're up to*" [58].

What followed Dorsey's interest in observing movements was a Web 2.0 microblogging site called Twitter. Twitter has gone beyond status sharing and now become more of an information sharing and news reporting medium. Twitter started its journey in 2006, but its fame started to spread after the South by Southwest festival in

2007. In the event, company set up user accounts for the participants, and used big screens streaming tweets from them in the conference simultaneously.

The effect of the conference was huge for Twitter. According to a report by HubSpot in 2008, despite being functional since 2006, Twitter had its 70% of users joined in 2008, and an estimated 5-10 thousand new accounts were opened per day [33]. In 2008, 35% of users had ten or less followers and 9% of users did not follow anyone at all. 80% of users had a bio specified on their profile. In 2008, Twitter had around 4-5 million users and Twitter had grown 600% in 2008, making it a top 1000 website in web traffic.

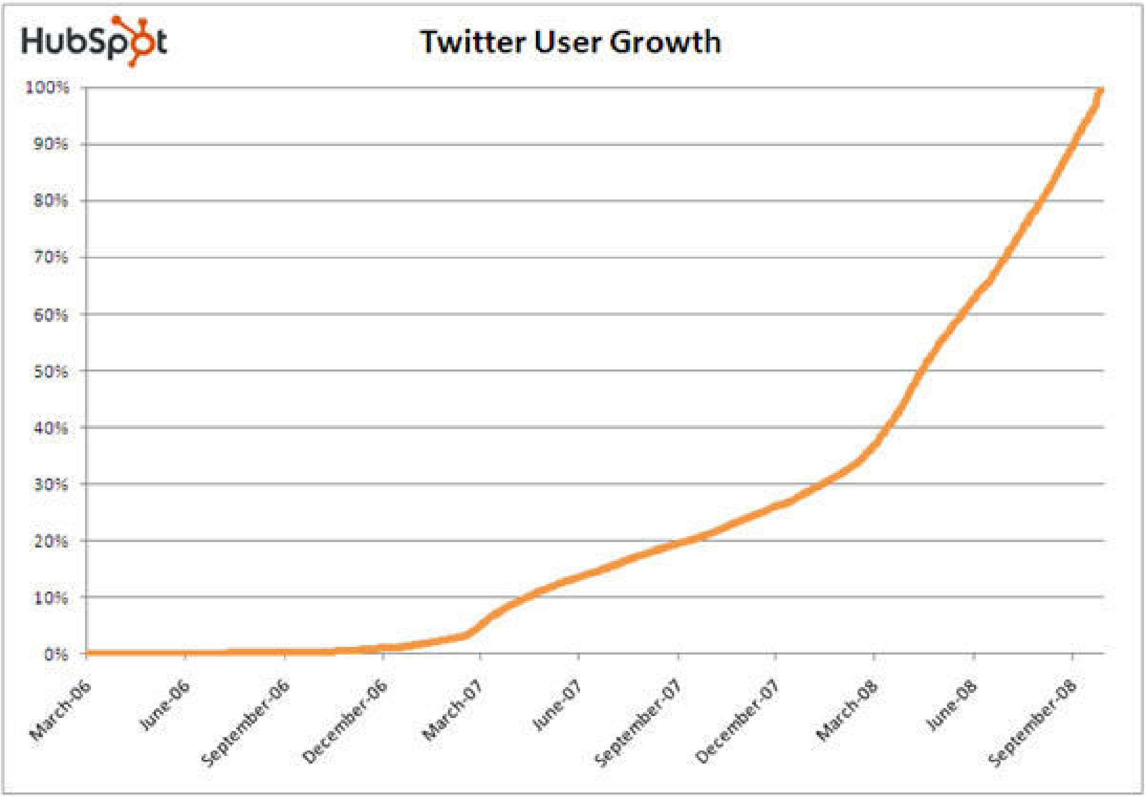


Figure 1.1. Twitter early growth

The trend of growth for Twitter has continued since then. In 2009, HubSpot published a new report and the report gives an astonishing 18.0000% growth rate of users (Figure 1.1), compared to 600% of 2008 [34]. The number of users who are not following anyone jumped to 55.50%, and the number of users who has a bio declined to 24.14%. According to this new report, 79.79% failed to provide a homepage URL, 68.68% have not specified a location and yet more interestingly, 54.88% have never tweeted. Of all users, 52.71% have no followers. By some definitions in the report, 9.06% of all users are inactive.

A discouraging statistics for Twitter is about the inactive user rate, and by correlation, the user retention rate. A report by Nielsenwire showed the retention rate of Twitter to be 40% and well below those of other popular web sites like Myspace and Facebook [48]. Soon, a big amount of criticism of the report was received from the Twitter community for the report's lack of taking Twitter applications into account. Nielsenwire responded by publishing another report that also looked into thirty other applications and web sites that feed into Twitter [47]. The dismaying results of the new report show the same 40% user retention rate for Twitter. The report concludes that if Twitter wants to continue its popularity, it should establish a higher user royalty.

When questioning the accuracy of retention rate in the Nielsenwire report, a point can be made by arguing the metric that is used. Java et al. [37] uses a metric for Twitter that is based on a study by Kolari et al. [39]. Definition of activity is given as "A user is considered active during a week if he or she has posted at least one post during that week“. This metric only takes posting action into account, yet, in a microblogging case study Bohringer et al. found that even the users that do not contribute greatly to the discussion(i.e. do not post microblogs) can be aware of all interactions (of their friends' and some other in Twitter) within the microblogging site [8].

In this case, if Nielsenwire used user posting as the metric, the accuracy of the results can be very doubtful. A better metric would be a combination of user posts and user logins, but this metric can only be evaluated accurately by Twitter itself, as third parties cannot record all user logins.

Another point involves the very nature of the Twitter activity. Twitter has faced larger traffic during big events like Mumbai bombings in India or US elections. Such sporadic events boost Twitter traffic not only by exciting frequent "posters" to post more, but also by drawing "inactive" users back to the site to observe the event and contribute to it. In this sense, a user cannot be regarded lost, as is the case with other web sites, because her inactivity can end in the face of an event that draws her attention. Hence, account removals due to "inactivity" should not be performed in Twitter.

In contrast to other networking sites, being able to use Twitter without establishing a strong friend set first remains a true asset for Twitter, but this asset must be regarded as a double edged sword.

From the user loyalty point of view, The 2009 HubSpot report has some encouraging points. The sense of community is increasing as people gets acquainted with the service and starts looking around to bond some new relations. 1.44% of all tweets are retweets, 37.95% of all tweets contain an "@" symbol (mentions), 33.44% of all tweets start with an "@" symbol (replies). According to the report, week days have more tweets, and Thursday has the most traffic.

Another report by Pear Analytics contributes in some other dimensions [1]. Currently, Twitter does not disclose its user numbers, and all number estimations are made using the third party application results or other web sites' tools. The report gives the total number of Twitter users in USA as 27 million. %55 of these users are male, 48% is between 18 and 34 ages, and %1 of users contribute 35% of total visits. According to the report, only 27% are regular users.

The report confirms HubSpot's results about the Thursday traffic, but classifies this traffic as consisting of 42.50% "pointless babble". In general, the report classifies all traffic as 3.60% news, 3.75% spam, 5.85% self promotion, 40.55% pointless babble, 37.55% conversational, 8.70% pass along value.

The study also reports that 5% of all users generate 75% of all tweets. This is in accordance with a report from Harvard Business Publishing's blog [53]. The blog shows Pareto principle on Twitter. The Pareto principle (also known as the 80-20 rule) states that, for many events, roughly 80% of the effects come from 20% of the causes [77].

According to the Harvard blog, the top 10% of prolific Twitter users account for over 90% of tweets. On other networks, 10% of users account for 30% of all production. The blog's headline, "Men follow men and nobody tweets“ marks a big difference of Twitter from other networks. On other networks, most activity is focused around women, and men follow or request friendship with women even if they do not know them, and women follow women. Twitter, in this aspect, follows a very different pattern because men follow women with 35% and men with 65%.

A recent hot topic about Twitter has been the teenage usage of Twitter. A then 15 year old Morgan Stanley intern, Mathew Robson, wrote a very interesting report on media, Internet and Twitter [56]. The report is hailed as one of the best reports from Morgan Stanley in recent years. Main difficulties, he wrote, evolve around the privacy issues, because unlike other networks, anyone can follow anyone on Twitter. Teens see Facebook as a better Twitter, as Facebook has status sharing too. Twitter always has an edge with its sms service, but this advantage is negated by the fact that cost of text messaging to Twitter is an issue for teens.

Twitter's user profile also drives teens away from the site. Teens do not have their friends on Twitter, and no one is looking at their profiles, so microblogging a post that nobody will read is not an alluring activity. As a blogger writes, "After all, who wants to be the only person they know on Twitter? It ain't fun".

Twitter has seen a rapid growth in the western sphere, and cities like London, New York and San Francisco generate the largest traffic on the site. Top 100 cities list is dominated by US, with the first non-USA city being Toronto, Canada. Tehran, Iran is 18th on the list, making it the first non-western city. As we recall the usage of Twitter in election protests in Iran, this does not come as a surprise. Tokyo, Japan, once in top 10 in 2008, now enters the list as the 21st city. Tokyo is important, because Twitter made its first effort to open up to the world by creating a Japanese version of Twitter. The site has been a success, not only in its traffic, but also as the first monetizing opportunity for Twitter with its ads sections.

1.3 Twitter: Beneath the Hood

Twitter's success can be attributed to two main factors; elegance in design, and simplicity in adding your own improvements to Twitter. Elegance is due to the character limit. Twitter names microblog posts from users as tweets. Each tweet has a 140 character limit and this feature was inherited from text messaging. The original 160 character SMS limit was reorganized into 20 character username and 140 character post fields.

A comparison between blogging and microblogging gives us a good understanding of the reason behind Twitter's popularity. Blogging requires good writing skills, and large content to fill pages. Using Twitter does not require sound grammar knowledge or long thoughts on a topic and everyone finds this encouraging to post small messages. These small posts are often criticized for their meaninglessness.

A study by Pear Analytics shows that 40% Twitter posts are "pointless babbles". Some bloggers have even suggested that Twitter is not microblogging because "the idea that someone can send a 140 character twitpitch or let the world know where about in some city street they are is considered to be blogging is stupid and devalues the hard work that most bloggers do everyday". This is another way of saying that

Twitter is indeed connecting people. The very basic question that Twitter asked was not “what do you think or know” but was “what are you doing“. User posts to this question cannot be expected to be literature pieces.

Recently, the question that Twitter asked on the main page was replaced with “Share and discover what’s happening right now, anywhere in the world“. This change sparked some interest about future plans of Twitter. New sentence implies a shift to being a real time information source. However, as finding useful information became very difficult because of “babbles” and spamming bots, Twitter is coming up with a lists feature. With this feature, users will be able to group users in lists and make these lists private or public. In the public case, other users will be able to add all of the users on list with a click. Simplicity is due to an early decision by Twitter to share posts with third party applications. Starting with Twitterrific in January 2007, many applications have been created.

Twitter applications range from finding similar users to finding answers to your questions and even writing book reviews. Many Twitter applications were written by Twitter users who saw a need for a specific use, and developed an application. As the creator of now defunct Quotably explains; “I was intrigued by just what conversations were shifting into this new space and frustrated by how difficult they were to follow. Threading seemed like to obvious solution and I set out to scratch the itch”. Each application grew in some communities and contributed to Twitter’s success by attracting more users. This trend has been growing so much that, according to the 2009 HubSpot report, only 48.1% of all users used web to access Twitter. The report is from February 2009, and in the face of hundreds of new applications, we can only expect to see web

1.4 Twitter Architecture

"Architecture originates with a disappointment in the provisional", anonymous. Twitter API is based on Representational State Transfer (REST) architecture. The architecture itself can be seen as a philosophy, not as a blueprint. Guiding principles of a REST interface are identification of resources that are kept on the web site and manipulation of resources through representations for users. In this way, clients, if permitted, can modify or delete data on servers. REST has been applied to describe the desired Web architecture, helped to identify existing problems, to compare alternative solutions, and to ensure that protocol extensions would not violate the core constraints that make the Web successful.

From Twitter perspective, the REST architecture means that Twitter will be able to work with web syndication formats. Twitter works with two of these formats. Really Simple Syndication (RSS) and Atom Syndication Format (Atom). Visitors can subscribe to syndication service (feeds) and receives an update every time web administrator changes the page. In this sense each Twitter user has a subscription to the users he follows. Twitter's open API decision lets third party applications read user's feeds, and desktop and mobile phone applications on Twitter use this functionality [32].

For users who followed many people, finding tweets was a major issue. The search engine, Summize.com was acquired in 2008 to solve this problem. Before the acquisition, Summize.com was very popular among Twitter users, and finally the service caught Twitter's attention. After a 15$ million deal, Summize.com workers joined the ranks of Twitter, and the service currently serves as "Twitter search“.

Twitter uses OAuth and Basic Authentication for the authentication. OAuth is an open protocol to allow secure API authorization in a simple and standard method from desktop and web applications. What this protocol implements is of vital importance to Twitter, because OAuth enables third party applications that are abundant on

Twitter to connect to user accounts without a username and password authentication. Any application that needs to connect to a user account can be directed to a page on Twitter, and users can click on tabs to allow the application. This way, the application cannot learn your password, and "if you ever suspect an application to be doing something it shouldn't with your Twitter account, you can simply turn off their connection without having to change your password". Although OAuth proposes a secure way to handle applications, some major security flaws prevented a full dependence on OAuth. In April 2009, after a major flaw was discovered in OAuth protocol, Twitter, as well as Yahoo, pulled their support on OAuth. Current day applications on Twitter can use either OAuth or the Basic Auth that asks users their passwords. As Twitter wiki explains "We (Twitter) would like to deprecate Basic Auth at some point to prevent security issues but no date has been set for that. We will not set a date for deprecation until several outstanding issues have been resolved".

Another security issue was because of the OAuth protocol. Despite those security issues, the biggest threat to Twitter's success lies in scalability issues. Twitter is a famous Ruby-on-Rails deployment. Ruby-on-Rails is an open source web platform for the Ruby programming language. The platform was created as a protest against Java platforms, and Twitter's scalability has often turned into a hot debate between these two platforms. To understand the debate and the issues that can lead to scalability issues, the design philosophy of Ruby-on-Rails (RoR) is worth mentioning. Ruby on Rails is intended to emphasize Convention over Configuration (CoC), and the rapid development principle of Don't Repeat Yourself (DRY). "Convention over Configuration" means a developer only needs to specify unconventional aspects of the application. Conventional aspects are not coded. This leads to less code and less repetition.

“Don’t repeat yourself” means that information is located in a single, unambiguous place. For example, using the ActiveRecord module of Rails, the developer does not need to specify database column names in class definitions. Instead, Ruby on Rails can retrieve this information from the database based on the class name.

With the instant fame after 2007, and increase in the visitor number, Twitter had to think about scalability issues. Former Chief Architect Blaine Cook famously said scaling Rails was easy in April 2007, but the problems continued and the famous “fail whale” image that greets users whenever the site is down became a popular icon in society. Twitter still uses Ruby-on-Rail, but started using SCALA for heavy and asynchronous jobs in April 2009.

1.5 Application Domains

1.5.1 Marketing

To see how Twitter cannot be ignored in marketing, let’s start with exploring how it effects businesses worldwide. Java [37] et al. reports that 13% of all posts include links. In order to overcome the character limit, Twitter had been using URL shortening application TinyUrl but switched to bit.ly. Bit.ly is a portfolio company of Betaworks, and Betaworks became an investor of Twitter after Twitter acquired another Betaworks company, Summize.com. So Twitter’s decision did not come as a big surprise. The drop in the visitor count of tinyUrl.com shows that, Twitter’s impact on the third party applications is not negligible.

Twitter is becoming the next-big-thing on web, and we can compare it to early giants like Google, Youtube and Facebook. What we cannot compare with is the marketing plans of those companies with the marketing plan of Twitter. Twitter critics often point out that Twitter does not have a sound marketing plan, and cannot get profitable. This reminds us the times when Google had no idea how to make money from youtube.

Twitter hype or Twitter's optimistic atmosphere, whatever you may call it, may solve the marketing problem. Human participation on such a marketing plan would provide some useful insights. There are already some signs; David Wilson, a horror and sci-fi author had recently asked his fans about a Twitter promotion and there is already another novel, “Sum” by David Eagleman's , that had a 6000% sales spark thanks to Twitter [63].

This sales increase is described by word of mouth marketing. Word of mouth effect is defined as passing information from people to people, and what can be more successful at it than Twitter? Twitter seems like an ideal salesperson which gives you short information of product, but does not bore you with details if you do not like the product. This aspect of social networks and word of mouth (WOM) marketing has already attracted some interest [36]. A study shows that WOM marketing has 20 times higher elasticity than traditional marketing. Here elasticity is described as the change in sales resulting from each dollar spent [66].

Companies that use Twitter as a market place include Dell, Delta airlines, Buy.com and many others. Recently many blogs have been created to teach companies on how to use Twitter as marketing platform. Those blogs see Twitter as *instant messaging amplified* . Marketing advantages on Twitter are listed as *Rapidly disseminating timely information to groups that are in disparate locations or from live events, as an extension of PR efforts, Personal branding, Enhancing a blog or Web site by making it more real-time and more interactive, As a direct marketing promotional tool* [43].

Brand recognition is another hot issue for companies that has yet to gain a loyal customer base. Twitter can be used to create a brand awareness in public through advertising microposts.

As a direct marketing tool, Dell has used Twitter successfully, and generated $1million in revenue by posting on Twitter [19]. For large corporations, using Twitter to get fast feedback on services and products has been very popular, and many instant posts about product reviews by customers such as *Really not cool Apple! My wife was waiting for this day for the ipod touch with mic and camera.* can be found on Twitter [20].

1.5.2 News

Twitter is also regarded the fastest way to reach to breaking news. Users' collaboration has given Twitter a clear edge over news centers like cnn.com. Recently news centers have set up Twitter accounts and encouraged users to interact with these accounts in order to receive any breaking news in real time. CNN had to buy "cnnbrk" account with its 930 thousand users from the previous owner, and this marked CNN's dedication to Twitter [61]. CNN maintains 45 official Twitter accounts, with @cnnbrk having more than 2,7 million followers and 5 million followers in total. CNN's Twitter account has been a hot topic because Ashton Kutcher had challenged CNN to Twitter popularity contest. Eventually Ashton Kutcher reached 1 million mark before CNN [13].

Personal branding is widely used on Twitter, and Kutcher case was a prime example of this. For celebrity news, Twitter has been a bridge between celebrities and fans. Not only Ashton Kutcher, but many celebrities, including Britney Spears, Oprah Winfrey and even Barack Obama have started to use Twitter, and especially Oprah Winfrey's Twitter debut marked a point in Twitter history as the traffic on the site jumped 43% [35].

During the election protests in Iran, Twitter played a greater role than news centers, and attracted more attention. In recent Mumbai attacks in India, just minutes after the attacks, Twitter was the major source until news sites caught up with updates [64]. Twitter's success to create a bridge between the world and the people who want to reach this world is not only thanks to microblogging posts. As well as posts, information flow to Twitter consists of pictures, links and videos.

Demonstration pictures from Iran such as [68] and the first picture from US Airways plane in the Hudson river [69] increased Twitter's popularity in the public. The Economist called Twitter a winner in this information race [26]. During Iran elections, this information stream also caught the attention of US and Iran governments. US government reportedly warned site owners not to undergo a maintenance for it would break the news stream from Iranian users [12]. Even after banning foreign journalists from covering rallies, Iran could not stop information flow and finally shut down access to Twitter.

As Twitter was used by more people, its popularity increased. The trend of using Twitter for fast information revealed itself when an American student was arrested in Egypt for taking photos. The student was able to post a tweet saying "Arrested". His friends reported his arrest to authorities, and American embassy in Egypt secured his release [27]. Such sporadic news trends helped Twitter gain a prominent role in news reporting .

As Twitter's role in news reporting increased, so did the possibility of its misuse. A tweet post about controversial California gay marriages issue caused a big buzz in Twitter community. The post informed that California Supreme Court had overturned Proposition 8, the voter-approved ban on gay marriage [16]. This false news post created an awareness about news reporting on Twitter. Users are eager to follow breaking news on Twitter, yet they question reliability of the news until it appears on big news centers.

Recently some big media outlets are incorporating Twitter based news into their sites. Users are given an option to turn off automatic one minute updates on a web page that gets content from Twitter. The content includes pictures, videos and posts which are approved for their credibility before being added to the page [28]. Such applications should be carefully observed, because they have the potential to overthrow current day static paged news reporting techniques.

1.5.3 Education

Sharing information in microblogging and emergence of vastly popular web sites pushed mobile education into microblogging domain. A paper by Ebner et al. examines the user behavior and asks the question "Microblogging-more than fun?" [25]. As the paper explains, distant education, or E-learn platforms had previously been defined with rigid terms. Accessibility to learning material was only possible on some defined computers and domains. The next push towards E-learn comes with mobility.

Mobility, in this sense, *should not be restricted to accessibility from home, university or a defined place* [25]. Microblogging gives the users tools necessary to coordinate and simulate a learning experience. As successful websites emerge, users and teachers will have a robust infrastructure to enhance their education experience. Ebner describes the success of weblogs in education based on three factors: usability, collaboration and personality. Usability makes it easy to blog, collaboration makes it fun and personality brings the dedication of user. Microblogging nicely fits into this scheme too. Microblogging enables a real-time interaction between users, and it has been used to simulate a class atmosphere between students that use different applications. According to Ebner et al., *Microblogging does not bring the potential to write articles, but it can be used effectively to connect each other and to inform about interesting things about e-learning.*

In a case study, Holotescu et al. [31] have set up Cirip.ro, a microblogging site, to simulate a class. They found microblogging to *be an effective tool for professional development and for collaboration with students and also provide valuable interactions in educational context.*

1.5.4 Crowdsourcing

Twitter has a huge potential for crowdsourcing. Crowdsourcing is "a neologism for the act of taking tasks traditionally performed by an employee or contractor, and outsourcing it to an undefined, generally large group of people or community in the form of an open call" [76]. Optimism about the future of crowdsourcing runs very high, even to the point that, Laura Fitton, a social media blogger says "I outsource my entire life. I can solve any problem on Twitter in six minutes" [65].

With crowdsourcing on Twitter, users are able to find resources very fast. Examples of crowdsourcing range from reporter's asking users about story ideas [73], minutiae of tax form fillings [70] or many other things [7]. Some applications [72] have already started utilizing crowdsourcing, and it seems to be a dynamic field. Another kind of usage is scholarly crowdsourcing. Although not fully utilized yet, some experiments show signs of future use [6].

Recently, Google joined this wave by buying a Twitter crowdsourcing tool, Aardvark[72].

1.5.5 Enterprise Microblogging

For an enterprise, that has many offices in different regions, microblogging can be used to heighten awareness in the enterprise. Gutwin et al. distinguishes four types of awareness; informal, social, group-structural and workspace [29].

Current day technologies, such as mails or bulletin boards are used to heighten group-structural and workspace awareness, but interaction is very limited in these forms and they fail to heighten the other types of awareness. As microblogging enables user interaction, it can be very effective to heighten all these forms of awareness. On the other hand, using a public microblogging site like Twitter poses great problems, because enterprise communication can reveal important strategies and long term goals, also Twitter is not very reliable and can cause disruptions in data communication when it is down.

In a case study, Bohringer et al. developed and tested a microblogging software without promotion among employees [8]. The results show that, entreblogging can create a *single point of truth* where all employees can contribute to a discussion and learn the goals ultimately. The study found that even the employees who do not contribute greatly to the discussion can be aware of all interactions within the microblogging site. This creates an awareness about the goals of the enterprise.

1.5.6 Alert Systems

The biggest disadvantage of current day alert systems is that, public cannot be alerted if public are not reading their mails or text messages. In the case of Fort Collins emergency system facing a tornado, residents could not get any warning emails or text messages even if they read their messages, because the system failed even before the tornado was going to hit the city [54].

Another issue for cities is the cost that is associated with sending mails or sms messages. Twitter brings up a full-fledged system that can connect residents of a city with virtually no cost. It also increases the abilities of an alert system by inputting more user generated data. Some cities already opted for Twitter to alert their residents [42]. The Virginia Tech incident in 2007 highlighted the security issues on university campuses. In the incident, the university tried every channel it could to alert students. With mobile Twitter applications that students can connect to an alert account, Twitter offers a unique service with again virtually no cost. Pacific university has already implemented a Twitter based alert system for its students, and the trend is likely to grow [22].

As with all good things, an alert system on Twitter has some drawbacks. The most notable one is the fact that Twitter is serving on an “as is” basis. Virtually it cannot guarantee a fully functioning service. It has maintenance periods which mean a big gap of time in the face of an emergency, and worse than that, it can get overloaded pretty often. A fail-whale image will not help people when a tornado is about to hit the city. Of course people can continue using a reliable alert system from other companies if functionality is a big concern. But with its entire popularity, user base and zero cost, Twitter can push proprietary alert systems out of business. Once Twitter replaces those alert systems, it will be a single point of failure which the public cannot afford to lose. Already, TechRadium, a Texas based company that produces alert systems sued Twitter for patent infringement in “mass notification” concept.

1.5.7 Data Mining of Tweets

Twitter provides an excellent medium for spatiotemporal text mining and information retrieval. Here we summarize three research problems in the context of mining Twitter data: text classification, expert finding, and trend mining.

Text Classification. A useful research problem mining of tweets is to classify streaming tweets into topic-based groups. Mining short segments of text has been studied in the literature in various other contexts, e.g., query-query similarity [45], paragraph and sentence similarity [30]. A successful Twitter text classification needs to handle a diverse set of streaming short text messages with abbreviations, slangs, and no sound grammar use. Fortunately, the quality of mining results can be improved by incorporating the rich contextual information, such as the author bio, profile, hash tags, urls, previous tweets and status of the author in the underlying social network.

Expert Finding. Expert finding have been traditionally studied in the context of enterprise intranets [4]. One of the most promising fields of information finding on Twitter takes advantage of the sheer size of its huge user base. Identifying experts in topics of user interests is a challenging task, given the large number of users and wide variety of potential interests. Some applications use bios to group similar people, and user posts can be scanned to find people with same hobbies, background and profession. Besides user bios and previous tweets as the text-base, the spatial and temporal meta-data provide a constraint on the potential user-base, since we typically look for ideas constrained to location and time. Twitter has the potential of involving more than locating experts, it provides an environment for people to assert their expertise by actively joining the information flow and giving useful insights.

Trend Analysis. While expert finding focuses on authoritative sources, observing the patterns in a crowd would provide information with the power of collaboration potentially by millions of users. Applications of trend mining include identifying and monitoring emerging topics and events dynamically [44, 57], and sentiment analysis on user posts for products publicized on Twitter [46]. Canonizing some ideas through Twitter user posts has an inherent liability to manipulation, but it also offers a quick and effective way of getting to know how people react to, discuss and adopt new ideas. By aggregating users' ideas, we can effectively eliminate fringe cases, and find accurate information on a fact. The system strongly resembles the idea of democracy. Crowd mining is a luminous manifestation of the power of Web 2.0 applications. To make it more interesting, Twitter as an open platform enables briefer exchanges of information that would be lost in a lengthy blog or text.

1.6 Third Party Applications of Twitter

"The difference was at least as old as the digital computer. Forgers created a new technology and then moved on to the next project, having explored only the outlines of its potential. Honers got less respect because they appeared to sit still technologically, playing around with systems that were no longer start. Hacking them for all they were worth, getting them to do things the forgers had never envisioned". p76, The Diamond Age, Neal Stephenson.

While a very broad classification can be made on the PC/mobile phone application basis, Twitter applications are too diverse to fall into two categories. Some applications have both PC and mobile phone support, but in general, PC applications offer more variety. Hundreds of applications are available on Twitter and everyday some others pop up. Application rankings change overnight, but to get a glimpse of what would be a classification of them, some of most popular applications [62] and some others are classified into nine categories here:

1. Location : Those applications use maps to show status posts. They can be configured to show posts only from certain regions.

   Twittervision[17], Twittearth, Twitter Atlas, Twibs[20].

2. Similarity :Applications that give user information, find similar people. Some applications use bios to group similar people, and user posts can be scanned to find people with same hobbies, background and profession.

   TwitterCounter[4], Twitterholic[6], Twubble, Twittie me, Twellow, Twitrank.

3. Search&Monitor :Applications that search and monitor patterns on media. These applications are more oriented toward scanning posts instantaneously and more interested in posts than people.

   Twitturly[8], Twitscoop, Tweetscan[14], Tweetburner[15], Monitter[19], Twistori[21], Hashtags.

4. Processing :Applications that enable threading a conversation/media sharing process. These applications solve the problem of losing track of a conversation. Users can see post exchanges between other users.

   Tweet2tweet, Tweepler[11].

5. Sharing :Applications that enable sharing media, other than micro blogging posts. Examples include feeding your blog post excerpts to Twitter, picture sharing and even preparing polls on Twitter.

   Twitpic[1], Twitterfeed[5],Twtpoll,Twitdom[13].

6. Enhancers :Applications that enhance personal pages, or your posts. These range from giving you special characters to use in your post to filtering out all Twitter traffic for some time.

   Twitter Keys, Twalala.

7. Platform :Applications that connect multiple platforms. With these, users can change their status fields on many sites at once. With Hello.txt, user can change Facebook and Twitter status posts at once.

   Hellotxt[12], Digsby[3].

8. Access :Applications that provide access to Twitter. Those can be from mobile phones or browser extensions. Although they can also share data, their primary goal is to connect you to Twitter.

   Tweetdeck[2], Twitbin[22].

9. Hybrid applications :Applications that utilize more than one feature. Many applications use more than one feature, but our definition for hybrid states that, an application should integrate two services to perform a task. As Twitter grows, we will be able to see more mash-ups with other technologies.

   Twitterfall(Location, Similarity).

As a precaution, these categories should always be considered dynamic. Some applications can merge more features in future, and even some of them can migrate to another category. Given Twitter's dynamism, a classification will be tenuous and time-dependent at best. We welcome suggestions about our classification.

CHAPTER 2

USING TWITTER FOR CROWDSOURCING

2.1 Motivation

The ubiquitous systems vision [74] of embedding and weaving abundantly available tiny-computers to the fabric of our daily lives is close to fruition. With the advances in MEMS technology in the previous decade, it has become feasible to produce various types of sensors (such as magnetometers, accelerometers, passive-infrared based proximity, acoustics, light, heat) inexpensively, in very small-form factor, and in low-power usage. Furthermore, cellphone technology has seen an adoption rate faster than any other technology in human history [23]: as of 2009, the number of cellphone subscribers has exceeded 3.3 billion users. The rate of innovation in this field has been head-spinning. Nokia, Google, Microsoft, and Apple have all introduced cellphone operating systems (Symbian, Android, Windows Mobile, iPhoneOS) and provided APIs for enabling open application development on the cellphones. These modern cellphones, which are dubbed as *smartphones*, enable location-aware services as well as empowering the users to generate and access multimedia content.

Despite the availability of the devices to fulfill the ubiquitous computing vision, the-state-of-the-art falls short of this vision. We argue that the reason for this gap is the lack of an infrastructure to task/utilize these devices for collaboration and coordination. In the absence of such an infrastructure, the state-of-the-art today is for each device to connect to Internet to download/upload data and accomplish an individual task that does not require collaboration and coordination. Providing an infrastructure for publish/subscribe and tasking of these devices enables any node to search the data published by several nodes in one region to aggregate and decide on a question, as well as task several nodes in one region to acquire the needed data (if the data is not already being published to the infrastructure).

We propose that Twitter [71] can provide an "open" publish-subscribe infrastructure for sensors and smartphones and pave the way for ubiquitous crowd-sourced sensing and collaboration applications. The open publish-subscribe system of Twitter implies that different actors may integrate user data differently. Moreover, third parties can use the gathered data in unanticipated ways to offer new services with them. In addition to this open publish-subscribe infrastructure, the social networks angle of Twitter also provides a useful feature for the crowd-sourced sensing and collaboration applications. Finally, the wide popularity of Twitter and the big community behind it (more than 30 million users in US), is an important reason to target our crowd-sourcing system for Twitter: It is easier to give the community a tool than to give the tool a community.

More specifically, we provide the following contributions.

- We provide a detailed survey of Twitter with existing application domains on news and alert systems. In Section 2.2, we present emerging application domains for Twitter: including participatory sensing and social collaboration.

- We discuss sensor integration to Twitter in Section 2.2.1 and smartphone integration in Section 2.2.2. We point to a potential new architectural trend in sensor integration, that of inexpensive sensors using cellular data network to reach Internet in one hop.

- In Section 2.3, we present our design and implementation of a crowd-sourced sensing and collaboration system over Twitter. Central to our system is a Twitter-bot (with an integrated database system) that accepts questions, crowd-sources them, and aggregates the answers to reply back to the querier. The system also includes a smartphone client for automatically pushing sensor reading information to Twitter.

- In Sections 2.4 and 2.5.1, we showcase and evaluate the performance of our crowd-sourced sensing and collaboration system on two case-studies. The first one is a crowd-sourced weather radar [1], which help monitor fine-granularity weather conditions and act as a ground-truth. Our second application is noise mapping of a region by aggregating the automatic noise-sensing updates from smartphones.

- We present an analysis of our real-world Twitter experiments to give insights for the feasibility of our approach. We find that although we do not offer the user any incentives to reply, our queries receive at least 15% reply ratios. Surprisingly, 50% of the total replies arrive within the first 10 minutes of our query, and 80% of the replies arrives within the first 2 hours, enabling low-latency operations for crowd-sourcing applications. Our experiments also found that consistently the majority of replies come from users that access Twitter from their mobile phones.

2.2 Research Directions using Twitter

In this section we discuss sensor and smartphone integration to Twitter and identify research directions and emerging applications for these domains.

2.2.1 Integrating Sensors

Sensor technology has matured. Sensor networking is getting there With the advances in MEMS technology in the previous decade, it has become feasible to

[1] You can visit our weather radar `@rainradar` on Twitter. We display the answers to our weather radar on a map at `http://ubicomp.cse.buffalo.edu/rainradar`. The map is configurable to show results from previous days, and also is zoomable to show fine-grain locations of the replies.

produce various types of sensors (such as magnetometers, accelerometers, passive-infrared based proximity, acoustics, light, heat) inexpensively, in very small-form factor, and in low-power usage. Moreover there has been nearly a decade of research in wireless sensor networks (WSNs) and some real-world deployments of WSNs have been successfully demonstrated [2, 3, 60, 75]. But sensors are not getting the attention/visibility/impact they deserve As such, WSNs offer an untapped source of information about our physical world. However, WSNs have not achieved the broad impact and visibility it deserves. Not only are we far away from "a central nervous system for earth", there is no significant market penetration for WSNs yet.

We can compare the current state of affairs in WSNs to the era when mainframe computers were used in many places, but for the public computers remained a mystery.

Sensor internet integration seems to be the key, twitter solves this Arguably the greatest barrier against wider adoption of WSNs is the difficulty in locating sensors and subscribing to them. We propose that Twitter can provide an "open" publish-subscribe infrastructure for sensors, as well as the search/discovery of sensors with certain attributes.

Moreover, having access to a lot of sensors is also valuable in that it would be possible to reduce false-positives from sensors by cross-checks.

Below we list some ideas we are pursuing for sensor integration to Twitter.

Sensor tweet standards. In order to search and process sensor values on Twitter, we need to agree upon a standard for publishing these sensor readings. We offer a biography format on Twitter that describes a sensor in detail in Section 2.3.2.

The bio-code makes sensors easy to find. By just searching for the desired sensor functionalities using the Twitter API over the bios, one can reach all sensors within a locality that provides the desired functionalities.

We are currently developing a standard, *TweetML*, for tweeting sensor values. We will make use of the built-in hashtags feature in Twitter for easier accessibility and searchability of sensor value fields. As part of our current work, we are publishing data to Twitter from some existing WSNs deployments. One of these is the wine-cellar monitoring WSN deployment, and another is personnel tracking WSN deployment.

New WSN architectures. The popularity of Twitter have already resulted in the production of inexpensive specialized devices for microblogging. TwitterPeek [51] is a good example of this trend. TwitterPeek enables the user to tweet and follow Twitter from anywhere (no WiFi necessary) using the cellular data network to connect to Twitter. One can buy TwitterPeek for a low, one-time fee and get connectivity service for the lifetime of the device –without any bills ever. The reason TwitterPeek is able to offer a powerful device at a low price is because of the benefits of mass production.

TwitterPeek may signal a new direction for WSN devices. Instead of using low communication range devices that incur the challenges/complexity of maintaining a multihop network and still require a basestation to access Internet, TwitterPeek-like sensors can directly reach Internet at one hop. These devices may not only tweet their sensor readings, but can also be easily controlled over Twitter to reconfigure their sensing schedules and tune their parameters.

2.2.2 Integrating Smartphones

Smartphones provide significant advantages over traditional wireless sensor nodes. Firstly, smartphones are mobile. Wherever a smartphone user goes, smartphone can take sensor readings (with built-in sensors for acoustic, image, video, accelerometer, tilt, magnetometer, and potentially with other integrated custom sensors). The dynamic geolocation feature of smartphones enables these readings to be location and time-stamped. Thus, in contrast to WSN nodes that are tied to static locations, and do not scale for coverage of large areas, smartphones cover large areas due to their mobility.

Secondly, smartphones are personal and administrated by their users. In contrast to sensor networks where energy-efficiency of utmost importance, smartphones are recharged by their users and it is not necessary to try to squeeze every bit of energy. Moreover, since smartphones are personal, they provide the potential of interacting with the phone user for tasks requiring human intelligence and intervention, such as taking a picture of a requested location, answering a question for which the user is well-equipped.

Below, we identify 3 new application domains for smartphone integration to Twitter, with increasing level of complexity.

2.2.2.1 Participatory Sensing

Participatory sensing is the use of volunteering smartphones to collect data from a large region. Although there has been significant work on participatory sensing [11], using Twitter opens up novel improvements on this application domain. Twitter's open publish-subscribe system enables others to use the gathered data in unanticipated ways and offer new services over them. Moreover Twitter's social network aspect enables new features to be added to participatory sensing. For example, when one of the users have performed significant amount of participatory sensing but her friend and competitor (Twitter enables using lists for followers/friends) have not done anything for that week, our system can send a reminder message for that friend.

There is already a good support for enabling participatory sensing applications over Twitter. Some Twitter third party applications (including Twittervision17, Twittearth, Twitter Atlas, Twibs20) use maps to show status posts, and can be configured to show posts only from certain regions.

2.2.2.2 Crowd-Sourcing

Crowd-sourcing means distributing a query to several Twitter users in order to gather and aggregate the results and exploit the wisdom-of-crowds effect. Examples of crowd-sourcing may be a weather/rainradar (with better precision and ground-truth than meteorological weather radars), and polling for the best restaurant entree in town. Web offers a rich variety of successful crowdsourcing applications. Rent-a-coder [2] facilitates assigning programming tasks to freelance programmers, and with Open Mind [3] non-expert internet users collaborate to create intelligent software.

Crowd-sourcing depends on user participation. With Twitter's popularity, finding a user to ask a question is not a problem, and we find that users are willing to participate and answer questions. In our experiments up to 1/6th of our queries got answered, although we did not provide any incentive for answering. We think this is due to the sharing and participatory nature of Twitter culture.

It is possible and easy to provide incentives for encouraging participation. Using Twitter's list functionality a group of users might be classified as experts of a topic. Each topic may have multiple user groups with different expert levels. Upon answering questions, the users can get promoted to a higher level. Visibility of these lists to the public would will be a great incentive for users to collaborate. Another way to incentivize users is to give the users that answer more questions the right to send more questions to our crowd-sourcing engine.

For developing countries of the world, crowdsourcing can utilize interesting incentives. Eagle [24] developed the txtagle system to crowdsource translation, transcription and survey tasks to mobile phone users in Nigeria. With txteagle, users earn mobile phone airtime or mobile money upon completing tasks that are sent to them via text messages. Citizens' interest in shaping their own city is also a strong incentive. Brabham [9] proposes harnessing creative ideas for city planning from web

[2] http://www.rentacoder.com/

[3] http://www.openmind.org/

users. Another platform, SeeClickFix[4], creates a vital link for city inhabitants to report the problems to the government.

The social network nature of Twitter can be exploited to provide an extra incentive for crowd-sourcing. It is also possible to provide useful feedback to crowd-source participant based on others answers. For example, the participant may get to see how her answer fares with other answers. In the “best restaurant” query, participants may get to learn which other participants also favorite their restaurant of choice.

2.2.2.3 Social Collaboration

Social collaboration applications are more sophisticated than crowd-sourcing applications in that they require back-and-forth interaction in contrast to the asymmetric one-shot interaction involved in crowd-sourcing. Examples of social collaboration applications include pick-up soccer games, arranged ride-sharing, community-organization events, support groups for addicts, and support groups for exercising and weight-watching.

Recently, cultural institutions are developing platforms where users collaborate on creating rich media for an art exhibit. In the m-Dvara project by Coppola et al [14] visitors can record media and comment on the art pieces, and the new visitors can surf internet to read comments and see which art pieces are most recommended.

Governments can greatly benefit from the synergistic effect of large scale collaboration. In a test case in India, 5000 students from more than 100 Indian institutions worked on e-Government projects to win a prize, and as the students competed for the best applications, the government benefited from receiving applications for free [59].

[4] http://seeclickfix.com/

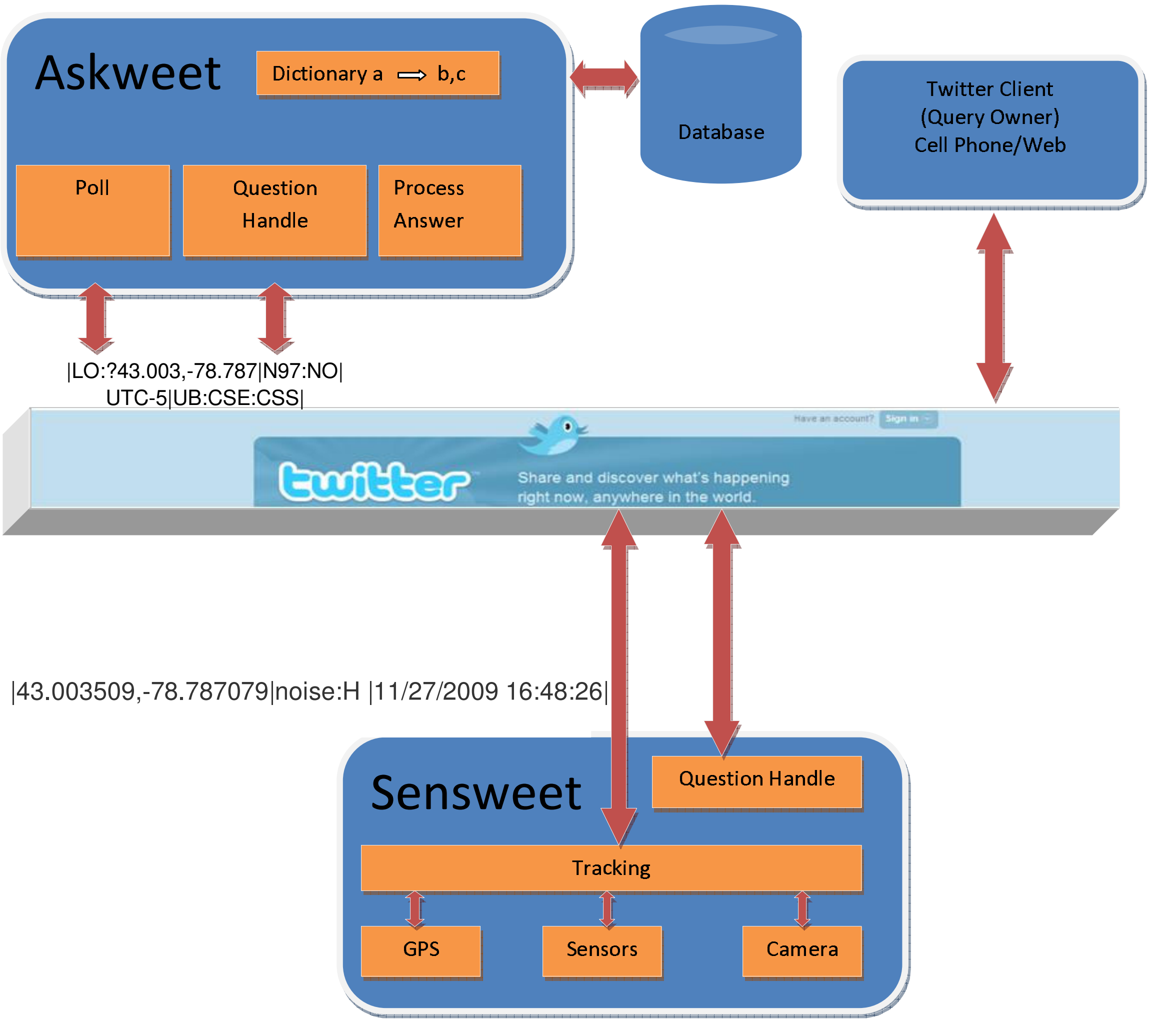

Askweet
Dictionary a ⟹ b,c
Poll
Question Handle
Process Answer
Database
Twitter Client (Query Owner) Cell Phone/Web
|LO:?43.003,-78.787|N97:NO| UTC-5|UB:CSE:CSS|
Have an account?
Sign in
twitter
Share and discover what's happening right now, anywhere in the world.
|43.003509,-78.787079|noise:H |11/27/2009 16:48:26|
Sensweet
Question Handle
Tracking
GPS
Sensors
Camera

Figure 2.1. Crowd-sourcing System Architecture

2.3 Our Crowd-sourcing System Architecture

In this section, we present the design of our crowd-sourced sensing and collaboration system over Twitter. Figure 2.1 illustrates the high level architecture of our crowd-sourcing system. Twitter acts a middleware for publish/subsribe as well as search & discovery. Our system is composed of three components namely *Askweet*, *Sensweet* and *Twitter clients*. *Sensweet* is a smartphone application that publishes real-time readings from the integrated-sensors to Twitter. *Askweet* is a program that listens to its Twitter account for questions and processes the questions and aggregates the replies it receives to these questions from *Sensweet* and the *Twitter clients.* We discuss the design of the Askweet and Sensweet components in more detail below.

2.3.1 Askweet

Askweet accepts a question, and tries to answer the question using the data on Twitter, potentially data published by Sensweets. If it is not possible to answer the question with existing data and/or if the question requires interaction, Askweet finds experts on Twitter (potentially using information retrieval techniques) and forwards the question to these experts. After obtaining answers from the experts, it replies the answers back to the asker. Askweet accepts a certain syntax from queries and replies, but it can also be extended and generalized to adopt modern natural language processing techniques.

The Askweet components of two case studies in this work run on a dedicated server, and keep all relevant data in a database to process questions and replies in a coordinated manner. Due to the parallelizable nature of processing queries and replies (a thread is assigned to each reply), it is easy to deploy Askweet on a cloud computing platform for elastic scalability. Since Askweet accounts have been recently whitelisted by Twitter and hourly request limits removed, it is possible to implement Askweet over Hadoop Map/Reduce framework to handle millions of queries and replies daily.

2.3.2 Sensweet

A Sensweet application uses the smartphones' ability to work in the background without distracting the mobile user. Sensweet applications sense the surrounding environment and send these data to the Twitter. While sending the data to Twitter, the Sensweet client formats the data according to the *bio-code* it advertises in the Biography section of its Twitter account. The main idea of using a bio-code is to allow worldwide users to search for the sensors they are looking for on-the-fly and enjoy a plug-and-play sensor network without registering through dedicated sites.

Here we provide a standard for a bio-code for Twitter to encode the values published by the sensor. To illustrate with an example, the Bio section of our noise-sensing application reads as: $|LO:?43.003,-78.787|N97:NO|UTC-5|UB:CSE:CSS|$. This bio-code consists of tuples separated with a vertical bar (|). In each tuple, descriptive fields are separated with a colon (:). The values that are separated with commas describe the phenomena the sensor(s) captures. The first tuple is always the location parameter: longitude and latitude (obtained from the built-in GPSs or entered manually). If the sensor is mobile (e.g., smartphone), a question mark will precede the longitude value. Even for mobile sensors a default location is added to give the queriers an idea of the region the sensor operates. The question mark hints that a more accurate location is included in the tweets. The second tuple explains the manufacturer of the sensor, product ID (if possible) and the sensor type(s) the sensor provides. The third tuple is optional, and describes the time zone that the sensor uses and can also include a timestamp. Although Twitter provides timestamping of tweets, this extra timestamp becomes important in case when a sensor need to store readings and send them later when it can connect to the Internet. The fourth tuple involves identification of the company/project that deploys the sensor, and defines a group id to locate other sensors that are part of that project.

Thus, the above bio-code is decoded as: Location is dynamic, but default location is UB North Campus Bell hall, Nokia N97 is used to capture GPS and accelerometer values in NY time zone for UB CSE Crowd-Sourced Sensing (CSS) Project.

2.4 Case Study: Crowd-Sourced Weather Radar

In this section we explain our crowd-sourced weather radar application. For the sake of simplicity, we choose a topic where everybody in Twitter can be an expert: the current weather condition. Our application contains two sub-applications, one

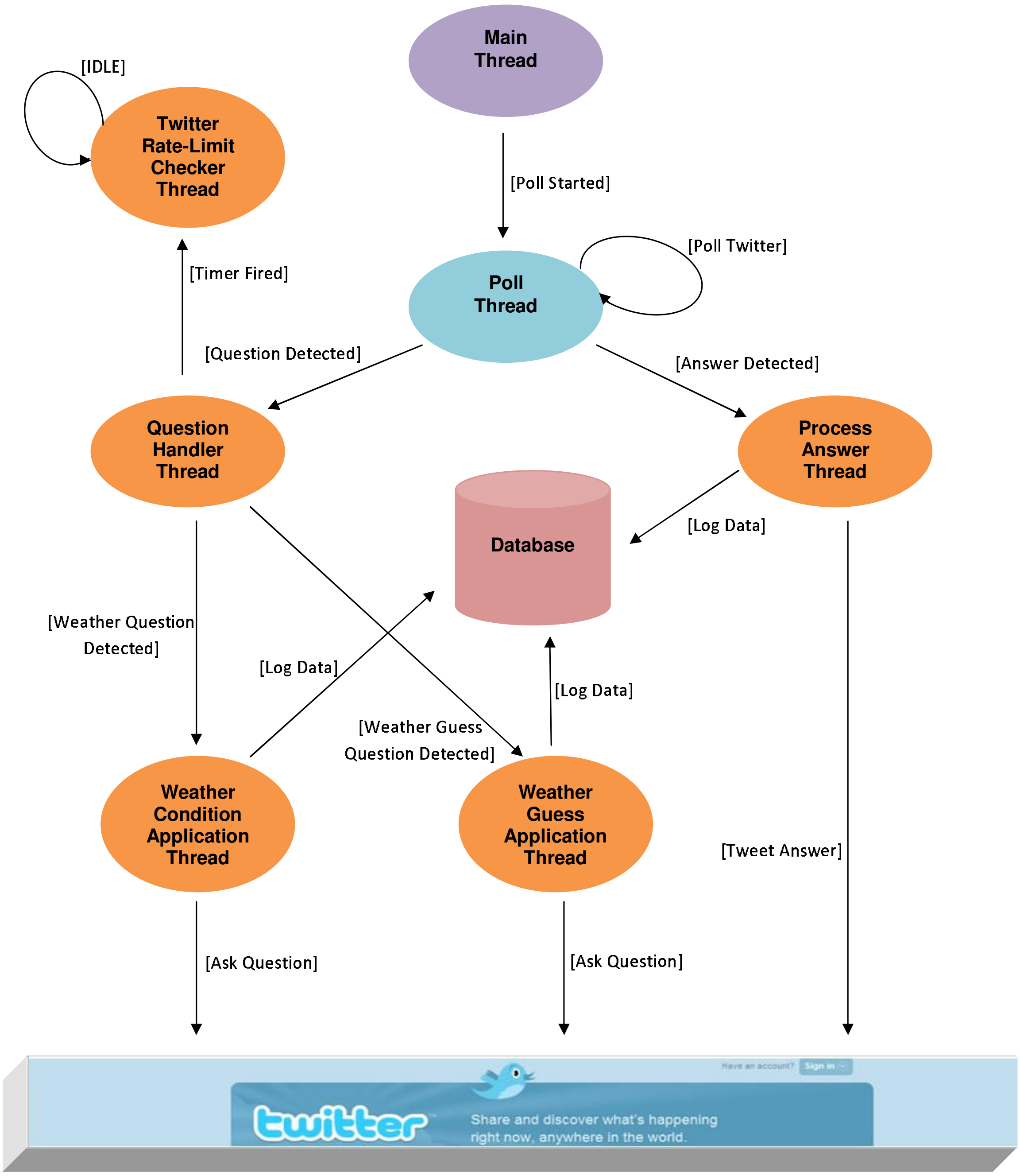
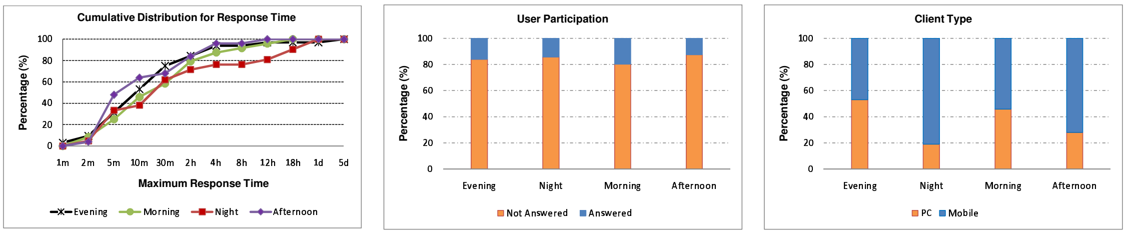


Figure 2.2. State transition diagram for Askweet component

of them obtains the current weather condition from users, and the other one obtains guesses from the users about the next day's weather condition.

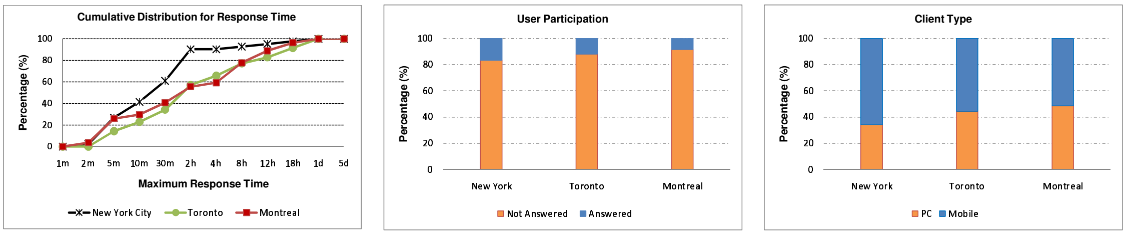
Weather radar application has its own question and answer format. The question messages sent by query owners are in the form of "?[Application Name] Loc:Location" where application name is either Weather or WeatherGuess. For instance "?Weather Loc:Buffalo,NY" might be a valid question for asking weather condition in Buffalo,NY. The forwarded query to the Twitter users is of the form: "How is the weather there now? reply 0 for sunny, 1 for cloudy, 2 for rainy, and 3 for snowy" Our weather radar application account can be visited at rainradar on Twitter. We display the answers to our weather radar on a map at http://ubicomp.cse.buffalo.edu/rainradar. The map is configurable to show results from previous days, and is also zoomable to show fine-grain locations of the replies.

We have implemented only the Askweet component of the crowd-sourced system since the Sensweet component can be any smartphone Twitter application. The Askweet component of our weather radar application is written in Java Programming language by using Twitter4J open source API library and total size of the source is about 2KLOC. Askweet listens to the incoming messages to its Twitter account and processes them with respect to their message types. The main function of Askweet component is to get a question, process it and/or forward this query to the multiple users who can answer it. After obtaining answers from Twitter users, Askweet sends the reply to the original querier.



(a) CDF for response time (b) User participation (c) Client types

Figure 2.3. Experimental results for NYC in different time slices



(a) CDF for response time (b) User participation (c) Client types

Figure 2.4. Experimental results for 3 cities

Our Askweet implementation is multithreaded for scalability, with each thread implementing a specific functionality. When the Askweet application is launched (Figure 2.2), it starts the poll thread that polls the Twitter account and gets the messages. Then the thread detects whether the message is a question or answer. Depending on the message type, it starts either a question handle thread or a process answer thread. Poll thread keeps on checking the account every minute continuously to get the new messages addressed to itself.

Question handle thread receives the question from the poll thread and detects if it is weather guess question or weather condition question. Depending on the question type it starts either a weather condition application thread or a weather guess application thread. Question handle thread also starts Twitter rate-limit checker thread in order to ensure that Askweet stays within Twitter's request limits. After this step, the question handle thread is terminated.

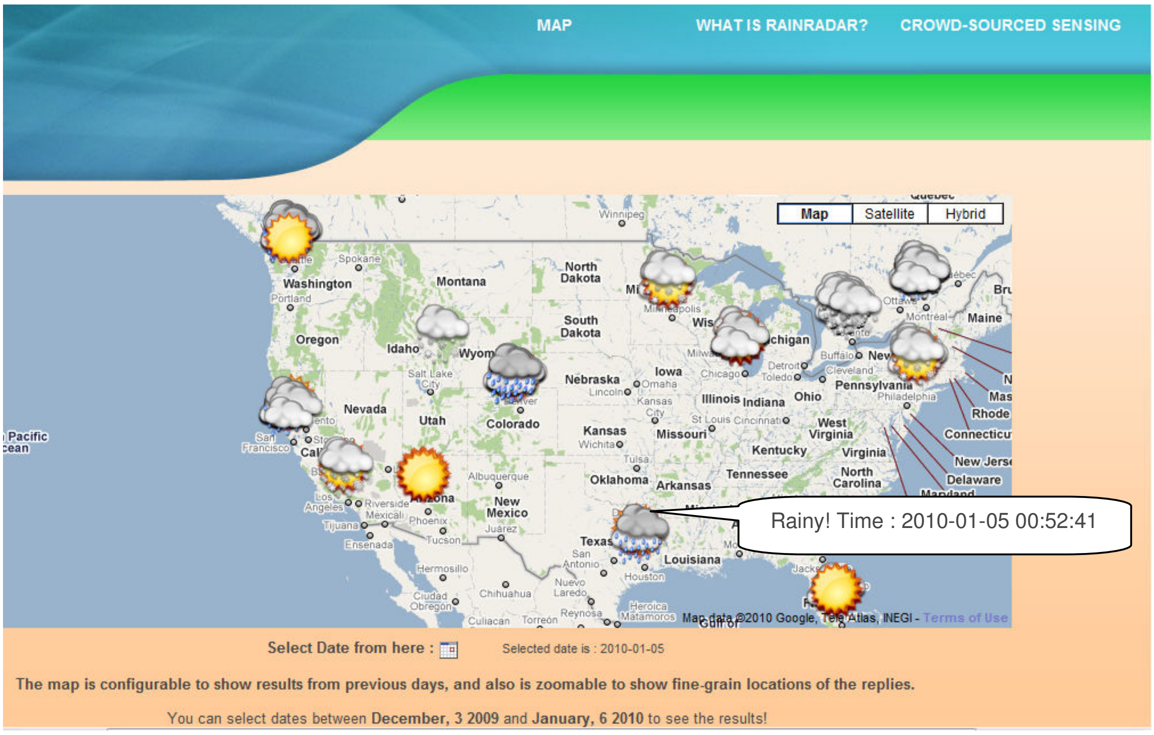
Weather guess application and weather condition application threads have almost the same functionality. Both of them get the question and parse the location from question text and search through Twitter to find users for the specified location. Then they send the question to the selected qualifying Twitter users. After that these application threads are terminated. Both of the applications keep all the relevant data in a database in order to observe the social collaboration and attendance. This database also helps the program not to spam any Twitter user with multiple requests within a week.

As we have a query count restriction on Twitter, we need a thread that checks Twitter to see if the system can proceed to post questions and inform the query owner about the received answers. If the system exceeds the rate limit, the thread locks question asking permit and releases the lock if otherwise. Process answer thread gets the answers from the poll thread and tweets the answer to Twitter. It also selects five of the answers to forward to the original querier.

2.4.1 Experiment Results

In this section, we present our experimental results for the weather radar application. We performed three types of experiments using weather radar. In the first one, we compare the user responses in different time slices of day for New York City (NYC). In the second, we compare user responses from three different cities: NYC, Toronto and Montreal. In the last one, we analyze the correlation of answers from our users with data from weather.com for one day (December 6, 2009).

In the first experiment, we compare the user response behaviors in NYC at different time slices. We observed that the response times in the afternoon and in the evening are better than those in the morning and at night (Figure 2.3(a)). An interesting phenomenon is that on the average 50% of the answers are received within the first ten minutes (Figure 2.3(a)). Figure 2.3(b) shows the user contribution to our experiments. We observe that Twitter user contribution to the experiment is highest in the morning which is nearly 20% (Figure 2.3(b)); we get a response from 20% of the queried users. For the other time slices, the contribution is around 15% (Figure 2.3(b)). Figure 2.3(c) shows the user distribution with respect to Twitter client types. At night time, an overwhelming majority of people use mobile Twitter clients to send their responses (Figure 2.3(c)). Overall, mobile client users consistently dominate over desktop/laptop users (Figure 2.3(c)).



Rainy! Time : 2010-01-05 00:52:41

Select Date from here :

Selected date is : 2010-01-05

The map is configurable to show results from previous days, and also is zoomable to show fine-grain locations of the replies.

You can select dates between **December, 3 2009** and **January, 6 2010** to see the results!

Figure 2.5. Screenshot of the Weather Radar Web Application

In the second experiment, we compare the user responses from different cities. We observe that users in NYC respond quicker than those in Toronto and Montreal, which have almost the same response patterns (Figure 2.4(a)). In Figure 4b, we compare the participation ratio of the users in these three cities. We see that users in NYC participate more than those in Toronto and Montreal (Figure 2.4(b)). In all these three cities, mobile Twitter client users dominate over desktop/laptop users and this ratio is highest in NYC (Figure 2.4(c)).

A screen shot of the weather radar map application for all cities is given in Figure 2.5.

Table 2.1. Comparison of user responses with weather.com

City	**Match for Current Day**	**Match for Next Day**
New York City	89%	56%
Toronto	79%	29%
Montreal	88%	54%

In the final experiment, we analyze the correlation of answers from our users with data from Weather.com. Since it is not practical to validate Twitter user responses with various fine-grain spatial (latitude, longitude) and temporal dimensions, the correlation is based on course-grain city wide level weather data for the entire day.

In the first column of Table 2.1, we list the correlation of user responses with the data from weather.com for the current day (the weather.com data and user responses are collected in the same day). If the weather.com reports "snowy" for the day, all responses except "snowy" are counted as "unmatched". If the weather.com reports a fuzzy condition such as "partly cloudy", all responses including "sunny" and "cloudy" are counted as "matched". In this experiment, we observe that for each city at least 79% of the answers match with the data from weather.com.

In the second column of Table 2.1, we list the correlation of user predictions for the next day with the data from weather.com. Here we collect the predictions of users in previous day (December 6) and find the correlations of those predictions with weather.com data collected on the next day (December 7). We observe that at least 50% of the user predictions match with weather.com for New York City and Montreal whereas it is 29% in Toronto.

2.5 Case Study: Smartphone Enabled Noise Map

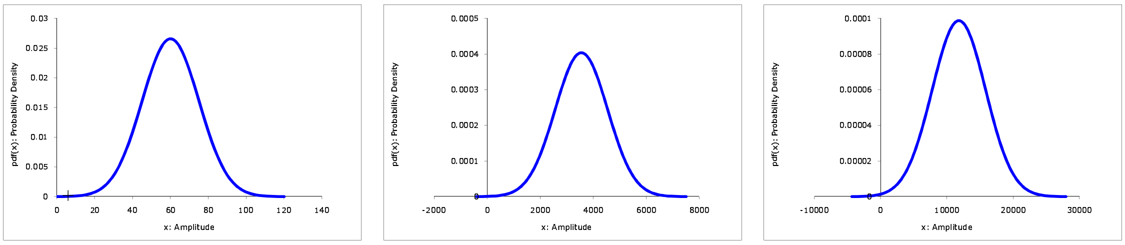
In this application, we measure the noise level of the surrounding environment via GPS enabled smartphones and provide a noise level querying service over Twitter. Here, the noise corresponds to all sound frequencies in the environment. We describe our implementations of the Askweet and Sensweet components for this application below.

Askweet component. We implemented the Askweet component similar to that of the weather radar application. The noise map application has its own query format of "?Noise Loc:Location". Any Twitter user can send a question to the Twitter account of Askweet (`twitter.com/askweet`) in order to query the noise level of a specific location. For example "?Noise Loc:Student Union, UB, Buffalo, NY" queries for the noise level of the Student Union at the University at Buffalo.

When Askweet gets a new query, it automatically tries to resolve the location by using Google's Geocoding Service (`http://code.google.com/apis/maps/documentation/`). After getting the latitude and longitude information from Google's Geocoding Service, Askweet searches previously known Sensweet clients in the database in proximity of the specified location. If Askweet finds a local client, it returns the latest noise level obtained from that client. If multiple Sensweet clients are found, the noise value with the latest timestamp is returned to the querier.

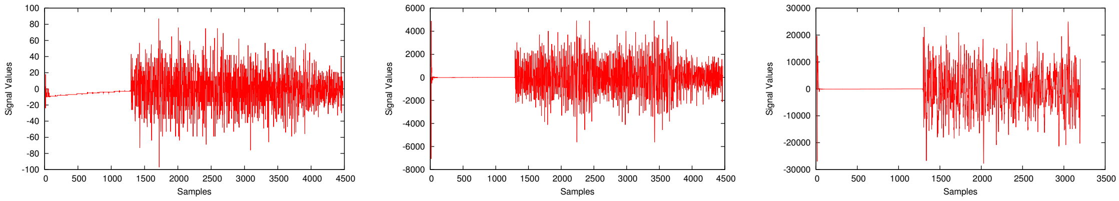
Sensweet component. We implemented a Sensweet client for the Nokia N97 Smartphone series. For implementing the Sensweet client we used Carbide C++ version 2.0.2, Nokia N97 Symbian S60 SDK V1.0 and Qt Tower 4.5.2. The total size of the source code for this Sensweet component is more than 1500 lines of code.

The Sensweet client detects the noise level of the surrounding environment and forwards this data to Twitter using our TweetML format mentioned in Section 2.3.2. The specific TweetML format ($|Loc|Noise : Val|Timestamp|$) for Noise Map application includes ordered values for location, sensor reading and timestamp. An example



(a) Low level noise (b) Medium level noise (c) High level noise

Figure 2.6. Normal distributions for different noise levels



(a) Low level noise (b) Medium level noise (c) High level noise

Figure 2.7. Representative samples for different noise levels

sensor reading can be “Noise:H” denoting that the current noise reading is “High”. Since Nokia N97 smartphones do not provide the noise level in decibel format, we implemented our own noise sensor driver to map noise samples into three categories: L as Low, M as Medium and H as High.

Our Sensweet client implements a timer for reading the GPS coordinates and using the microphone to record a one second noise sample in “Windows WAV” file format. Then, Sensweet client parses this WAV file to obtain the mean value for the amplitude of signals in the sample. In order to map the current sample into one of the noise categories {Low, Medium, High}, we used three normal distributions. For a given mean value x of amplitudes obtained from a one second sample, we calculate the following probability density function (pdf(x)) for each of the predefined three normal distributions:

$$pdf(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \qquad (2.1)$$

The μ in the formula represents the mean of the corresponding distribution and σ^2 represents the variance. The assignment is based on the highest value. Since there is no gain setting for the microphone of Nokia N97, our mapping is valid for any Nokia N97 smartphone device. For the smartphones having adjustable microphone gain, our mapping can be easily adapted by dividing signal values by the gain factor.

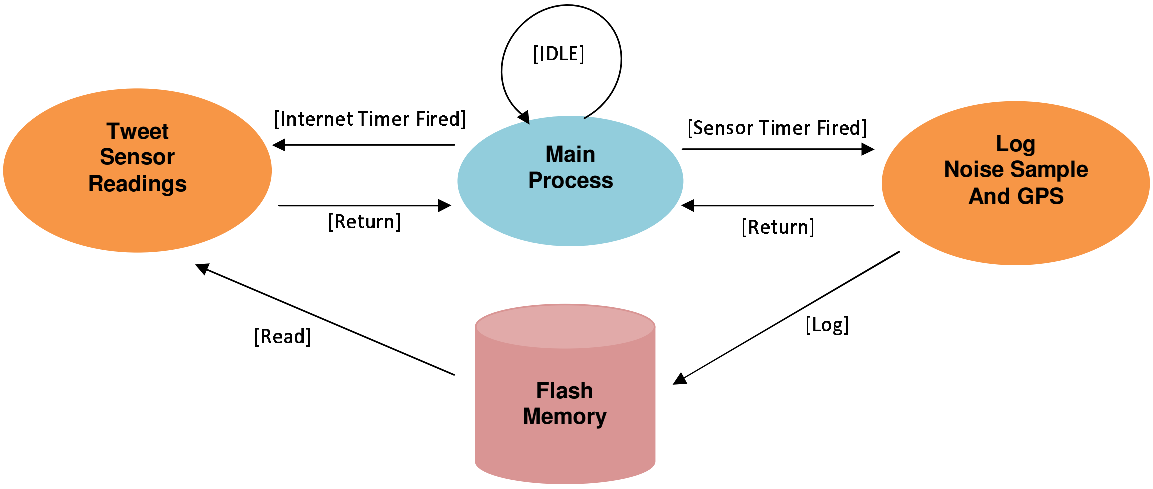


Figure 2.8. State transition diagram for Sensweet client

The state diagram of the Sensweet client for noise map application is given in Figure2.8. When the phone is started the Sensweet application is also launched as a background process and waits in the "idle" state. The GPS based location, noise level, and current timestamp is logged to the flash memory when the sensor timer is fired. We also keep another timer for forwarding sensor readings to Twitter. When the Internet timer is fired, main application reads the latest sensor readings from the flash disk and tweets it (`http://twitter.com/Sensweet`).

2.5.1 Experiment Results

Here we provide our experimental results for the noise map application.

In order to determine the normal distributions representing the “Low”, “Medium”, and “High” categories for noise levels, we performed experiments in six different locations with varying noise levels. In each location, we recorded more than 200 noise samples with a duration of one second.

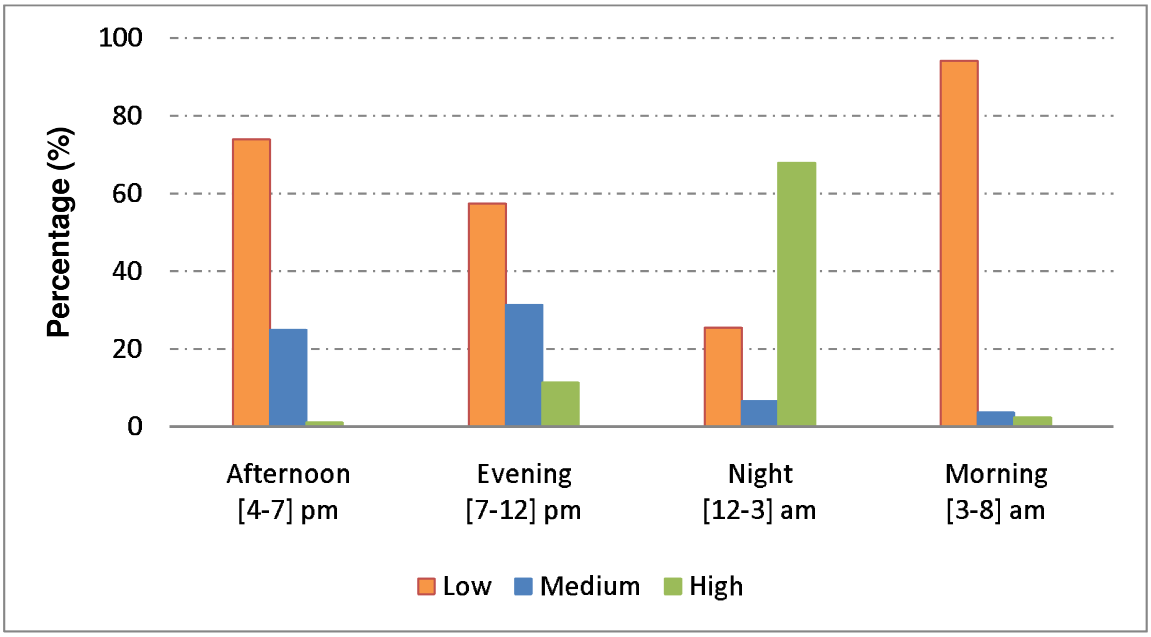


Figure 2.9. Daily noise fluctuation graph

We assign the “Low” category to the samples that we obtained during the silence in home and computer lab locations. The amplitude distribution for “Low” level noise is given in Figure 2.6(a). Here the amplitude (absolute value of signal values) of low level noise mostly fluctuates between [0,100], which also implies that signal values mostly fluctuate between [-100,100] (Figure 2.7(a)). For the “Medium” category we collect samples from the Student Union at UB and various meeting rooms at the CSE department where people talk to each other (noise mostly includes human voice). The “High” category is collected in bars and clubs in Buffalo with loud background music. The normal distribution of amplitudes for “Medium” and “High” categories are given in Figure 2.6(b) and Figure 2.6(c). Representative samples for these two categories are also given in Figure 2.7(b) and Figure 2.7(c) respectively.

In another experiment, we measure the noise fluctuation of our case study user for one weekend day over different time slices starting from Saturday 4.00 pm until Sunday 8.00 am (Figure 2.9). By analyzing the temporal noise fluctuation, it can be possible to predict some of the activities of the user during the day time. In the afternoon period the noise level fluctuates between “Low” and “Medium” level. During this time the user was at home and meeting with his friends. In the evening period the ratio of “Low” level decreases and ratio of other two levels increase. In this period, the user was having dinner with his/her friends in some place and going to a bar/club after that. In the night period the noise level is mostly “High” and the user was visiting a club. The noise level in the morning period is “Low” mostly since the case user was sleeping at home.

CHAPTER 3

MINING PUBLIC OPINION ON TWITTER

3.1 Benefits of Opinion Mining on Twitter

Since the 1824 poll that was conducted in the contest for the United States presidency, polls have been used to take a snapshot of public opinion, but they cannot reach many people nor capture opinions about the topics that are not asked in the questionnaire. Moreover, while events unfold rapidly and public opinion changes with those events, polls cannot account for the temporal changes in public opinion. With the advance of micro-blogging sites like Twitter [37, 41], we are now able to observe individual opinions and keep up with the changes in the public opinion. When carefully aggregated and classified, individual opinions can give us a better understanding of how some events are received by the public.

In this work, we propose efficient methods to identify and classify opinions in a large stream of information, and pinpoint related events that stimulate users to express their opinions.

In particular, the contributions of this work are as follows:

- We develop and utilize an emotion corpus to detect emotions in tweets. This method enables expanding opinion representation from binary options ("positive or negative") to multiple dimensions by providing more granularity in classification.

- We propose combining set and vector space models to observe the public opinion and detect changes over time. From the experimental results, we found that using these two methods together eliminates false positives and improves the accuracy of our findings.

- We develop a dynamic scoring function to give a synopsis of news (in terms of prominent words) that led to breakpoints in public opinion.

- We create a customized event tracking application that can notify users without flooding them with every new entry about the event. We show that our application is more user friendly than the Google Alert[1] service.

3.2 Related Work

Opinion Mining has received great attention recently and researchers started to investigate people's opinion about certain topics or news [21].

Existing opinion mining methods are usually grouped under two categories [38, 49] called document based and attribute based approaches. These approaches are focused on characterizing user opinions as positive or negative over domain specific web sites [17, 52] for different applications.

As a document level approach, Turney et al. [67] proposed determining polarity of documents by using semantic orientation of extracted phrases. As an example of attribute based approaches, Zhuang et al. [78] proposed a method for grouping movie reviews based on frequent opinion terms. Differing from these supervised approaches, we propose using a finer granularity classification (8 emotion classes) for opinions.

To account for the temporal changes in public opinion, a related work to our approach is proposed by Ku et al. [40] where the authors used the language characteristics of Chinese. In temporal dimension, their method captures opinions and shows changes in overall sentiment about candidates in a presidential election.

[1] http://www.google.com/alerts

3.3 Methodology

We begin our discussion for methodology by first explaining what indicates a change in public opinion in streaming tweets. For this purpose, we note two observations on Twitter data.

Observation 1: If an event results in a change of public opinion, more tweets contain emotion words. Furthermore, **emotion pattern** of tweets in that time period is different from the emotion pattern of the preceding period, but more similar to the emotion pattern of tweets in the following period, i.e, the news has an enduring impression on public.

- **Example Tweet:** (Transgression claims admitted by Woods.) *Tiger Woods - What a disappointment.*

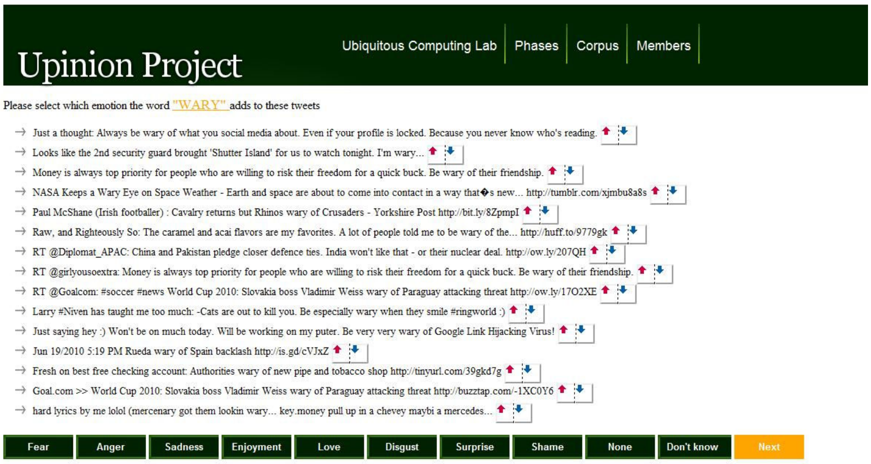
Observation 2: If an important story about the event appears, the **word pattern** of tweets is different from last period. On the other hand, the same word pattern repeats in the next period, i.e, tweets contain similar words in the next period as still the same topic is discussed.

- **Example Tweet:** (Companies start ending sponsorship agreements.) *Accenture Dumps Tiger Woods From Corporate Homepage.*

Following these observations, we conclude that, to claim a change in the public opinion, the **emotion pattern** and the **word pattern** must change according to these observations. We are looking for news that are both major events and opinion changers. In Section 3.3.1 we discuss how we find emotion and word patterns and use mentioned observations to detect opinion changes. We continue with finding topics related to the events in section 3.3.2

3.3.1 Opinion Detection

For the emotion pattern, we use an emotion corpus based method, while using set space model for the word pattern.



Please select which emotion the word "WARY" adds to these tweets

- Just a thought: Always be wary of what you social media about. Even if your profile is locked. Because you never know who's reading.
- Looks like the 2nd security guard brought 'Shutter Island' for us to watch tonight. I'm wary...
- Money is always top priority for people who are willing to risk their freedom for a quick buck. Be wary of their friendship.
- NASA Keeps a Wary Eye on Space Weather - Earth and space are about to come into contact in a way that�s new... http://tumblr.com/xjmbu8a8s
- Paul McShane (Irish footballer) : Cavalry returns but Rhinos wary of Crusaders - Yorkshire Post http://bit.ly/8ZpmpI
- Raw, and Righteously So: The caramel and acai flavors are my favorites. A lot of people told me to be wary of the... http://huff.to/9779gk
- RT @Diplomat_APAC: China and Pakistan pledge closer defence ties. India won't like that - or their nuclear deal. http://ow.ly/207QH
- RT @girlyousoextra: Money is always top priority for people who are willing to risk their freedom for a quick buck. Be wary of their friendship.
- RT @Goalcom: #soccer #news World Cup 2010: Slovakia boss Vladimir Weiss wary of Paraguay attacking threat http://ow.ly/17O2XE
- Larry #Niven has taught me too much: -Cats are out to kill you. Be especially wary when they smile #ringworld :)
- Just saying hey :) Won't be on much today. Will be working on my puter. Be very very wary of Google Link Hijacking Virus!
- Jun 19/2010 5:19 PM Rueda wary of Spain backlash http://is.gd/cVJxZ
- Fresh on best free checking account: Authorities wary of new pipe and tobacco shop http://tinyurl.com/39gkd7g
- Goal.com >> World Cup 2010: Slovakia boss Vladimir Weiss wary of Paraguay attacking threat http://buzztap.com/-1XC0Y6
- hard lyrics by me lolol (mercenary got them lookin wary... key.money pull up in a chevey maybi a mercedes...

Figure 3.1. Corpus

Emotion Corpus Based Method is based on vector space model for calculating document similarity. For the emotion detection in tweets, we use an emotion corpus that is based on 8 basic classes, E={**Anger, Sadness, Love, Fear, Disgust, Shame, Joy, Surprise**}, from [50]. We built a 309 word emotion corpus to populate those 8 classes. Our ongoing corpus building efforts include a web site (Figure 3.1) where we are crowdsourcing the task of assigning words to emotions. Each class represents a dimension in the Boolean emotion vector of a tweet. We look for emotion words in a tweet, and if found, set the corresponding class dimension in the emotion vector to 1, otherwise it remains 0.

- **Tweet:** *I was on main street in Norfolk when I heard about tiger woods updates and it made me feel angry,* on 2009-12-11. **Emotion vector:** $(1, 0, 0, 0, 0, 0, 0, 0)$.

For all the tweets in a chosen time interval, a centroid of all corresponding emotion vector dimensions is calculated, and this centroid is considered a document for each interval.

For a given time interval T that contains N tweets, let V={v_1, v_2, ..., v_N} be a set of vectors (with $l = 8$ dimensions each) generated from these tweets. We define centroid $\bar{v}$ for period T as:

$$\bar{v} = (\frac{\sum_{k=1}^{k=N} v_k^1}{N}, \frac{\sum_{k=1}^{k=N} v_k^2}{N} ..., \frac{\sum_{k=1}^{k=N} v_k^l}{N}) \tag{3.1}$$

After finding centroid vector for each interval, we define the opinion similarity between two intervals T_1 and T_2 by calculating cosine similarity between their centroid vectors:

$$Sim(T_1, T_2) = \frac{\bar{v}_1.\bar{v}_2}{|\bar{v}_1||\bar{v}_2|} \tag{3.2}$$

Set Space Model prescribes representing each interval by a single document which is the union of the tweets posted in that particular time interval. After removing the stop words and stemming the terms using Porter stemmer [2], we collect all terms in a hash set for each interval. We define the similarity between two intervals T_1 and T_2 by calculating Jaccard Similarity [5]:

$$Sim(T_1, T_2) = \frac{|(Set)T_1 \cap (Set)T_2|}{|(Set)T_1 \cup (Set)T_2|} \tag{3.3}$$

To find the changes, neither corpus based method nor the set space model alone is suitable. For the corpus based method, a change in the centroid can be misleading when the interval has very few emotion words compared to its neighbors. For the set space model, a change in similarity does not by itself imply an opinion change, because not all of the words are emotion words. In our method, we first analyze vector space similarity. If we detect a possible change, we validate it by analyzing

[2] http://tartarus.org/martin/PorterStemmer/

the Jaccard Similarity. Following the observations 1 and 2, if both methods detect the change, we report that point as a breakpoint.

T_n is a time break, if the followings are satisfied in both corpus based method and set space model:

$$Sim(T_{n-1}, T_n) < Sim(T_{n-2}, T_{n-1}) \tag{3.4}$$

$$Sim(T_{n-1}, T_n) < Sim(T_n, T_{n+1}) \tag{3.5}$$

3.3.2 Breakpoint Representation

After detecting the changes, we set out to identify the events that caused these changes. To this end, we look for the prominent words of an interval to represent the breakpoint. For the prominent word selection, we propose a TfIdf based dynamic scoring function. The algorithm should effectively find recently emerging keywords to guide users into catching breaking news and pay special attention to the words which emerge in a period and start appearing in more periods as time progresses.

The Streaming TfIdf Algorithm. To identify the events that caused breakpoints, we need to find keywords that represent the topics of these events. We propose the Streaming TfIdf algorithm for extracting event related keywords from an information stream of tweets.

Document Phase. For breakpoint representation, the same time interval length in the opinion detection is used, and for every time interval T_n, a document D_n contains the union of stemmed words from all tweets in that period. For word x in document D_n, Term Frequency Tf_{x,D_n} is calculated as:

$$Tf_{x,D_n} = \frac{Count_{x,D_n}}{\sum\limits_{m} Count_{m,D_n}} \tag{3.6}$$

For the total count of documents up to document D_n, Inverse Document Frequency of a word x in document D_n, Idf_{x,D_n} is calculated as:

$$Idf_{x,D_n} = \log(\frac{n}{|\{\forall k, k \leq n : x \in D_k\}|}) \tag{3.7}$$

Note that, n is not a fixed value. As we move from the oldest document to the newest document, the total number of documents, n, increases. By this parameter, the first appearance of a keyword will always have a bigger Idf value, and the following appearances of the word will have smaller values.

Based on the calculated Idf_{x,D_n} and Tf_{x,D_n}, we calculate the $TfIdf$ value as:

$$TfIdf_{x,D_n} = Tf_{x,D_n} \times Idf_{x,D_n} \tag{3.8}$$

Prominence Update Phase. For a keyword x that recently appeared in D_n, we define the Tf_{x,D_o} for the word x in document D_o where $o < n$ as:

$$tf_{x,D_o} = tf_{x,D_o} + F(D_o, D_n) \times tf_{x,D_n} \tag{3.9}$$

Here, we apply a decay function $F(o, n)$ to prevent the word x in the document D_n to increase the Tf value of x in a too old document D_o. This follows from the fact that, tweets are highly temporal, i.e, new events tend to affect user tweets only for a short period of time. As we move forward in the time domain, a keyword in a new period should not increase the prominence of a keyword in a way back period, because it is highly unlikely that appearance of a keyword is because of a very old event.

For the period numbers o and n, we define the decay function for two periods T_o and T_n as:

$$F(D_o, D_n) = 1/(n - o) \tag{3.10}$$

For the updated Tf values of the keyword x in document D_o, we re-calculate the $TfIdf_{x,D_o}$ as:

$$TfIdf_{x,D_o} = Tf_{x,D_o} \times Idf_{x,D_o} \tag{3.11}$$

We choose p words with highest $TfIdf$ values from each document, and call them prominent words of that document.

3.4 Experimental Results

In this section, we present experimental results of our methods on Twitter. We analyzed data about two topics, (1)Fort Hood shootings in Texas, USA, November 05, 2009 and (2)Tiger Woods, November 27, 2009 car accident. We used a Twitter search engine, Twopular [3] to collect data. We processed 258548 tweets, and found 23280 emotion words in those tweets. Figure 3.2 shows the tweet count of each day.

3.4.1 Opinion Detection

The length of time intervals is an important factor in our analysis. We evaluated unit lengths varying from 2 hours to 24 hours. Intervals shorter than 12 hours lead to biased results, because they contain too few tweets to form a meaningful sample. On the other hand, intervals longer than 24 hours are not suitable for the problem

[3] www.twopular.com

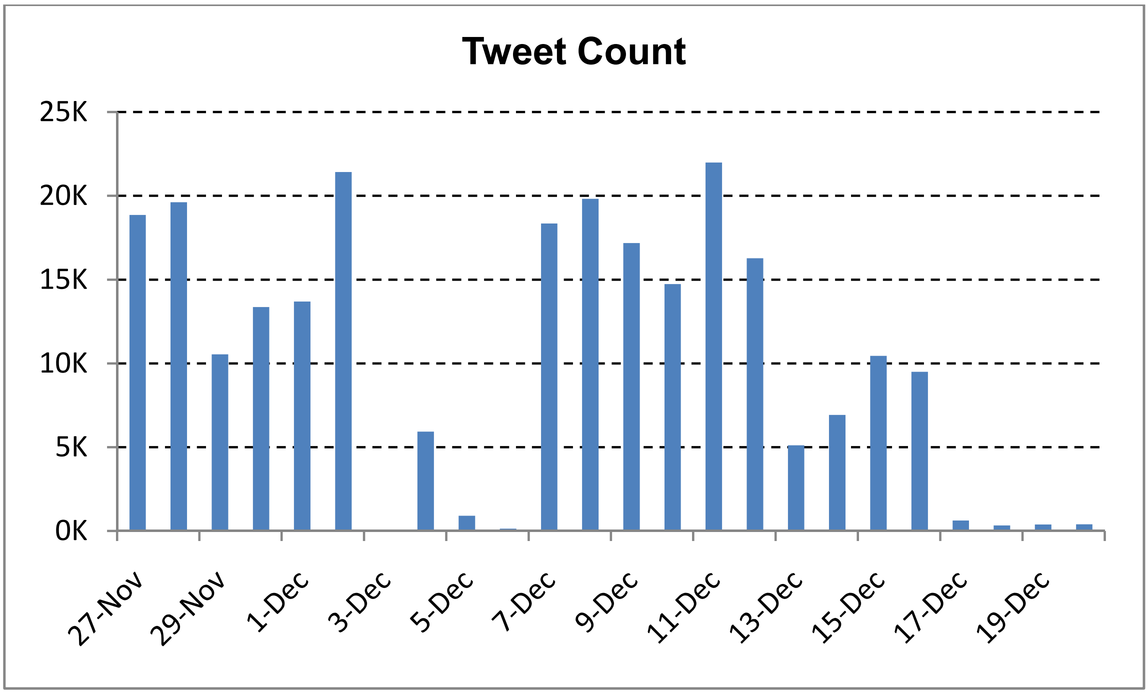


Figure 3.2. Tweet Count of Days

domain (media news cycle). We chose 12 hours, because it is the shortest interval to provide meaningful data besides enabling us to capture events in fine granularity.

In our data for 20 days, we found 8 possible breaks by Emotion Corpus Method (Figure 3.3) {**5, 10, 17, 23, 25, 27, 32, 36**}, and 5 of them {**5, 10, 23, 25, 27**} were also captured by Jaccard similarity (Figure 3.4). Figure 3.3 contains black bars that represent outlier intervals with very few tweets.

We tested our findings with a time line of Tiger Woods related events from CNN, ABCNews and ESPN [4]. Our 3 validated breaking points are related to the following events in successive order: (5)Transgression claims accepted by Tiger Woods, (10)more women alleged to have affairs with Woods, (23) Gatorade ends a sponsorship agreement with Woods, and Twitter users start writing thousands of jokes

[4] http://sports.espn.go.com/golf/news/story?id=4922436

about Woods with Santa Claus #hashtags nearing Christmas. Among the validated breakpoints, **25** and **27** are false positives.

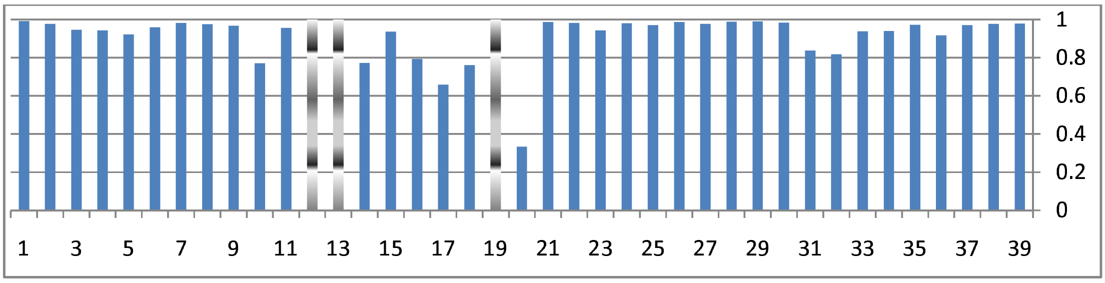


Figure 3.3. Emotion Vector Similarity of two successive intervals

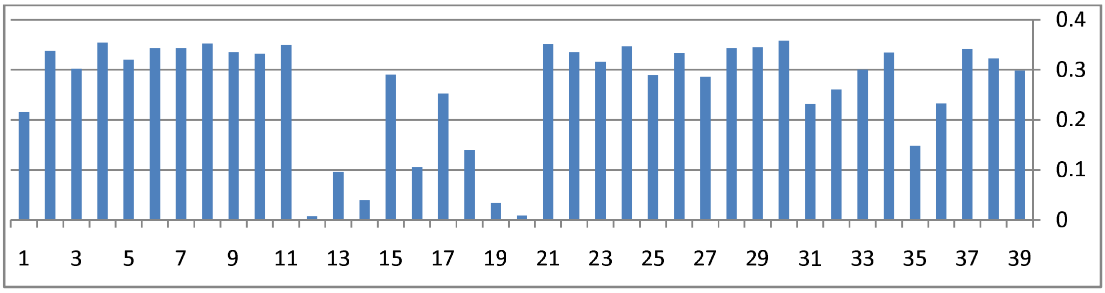


Figure 3.4. Jaccard Similarity of two successive intervals

3.4.2 Breakpoint Representation

Upon detecting opinion changes in the Tiger Woods case, we found frequent keywords of all periods, and by using the Streaming TfIdf algorithm, we extracted the prominent words from these keywords.

While creating documents for each 12 hour period, we put top F most frequent words into their respective documents. During this process, we used the Porter Stemmer to remove the commoner morphological and inflexional endings of words and analyzed the frequency distribution graph of the words. We found 50 to be the best choice because for values larger than 50, big clusters of words with low frequencies appear.

For the number of **prominent words** p, we used $p = 5$. The first document has the prominent words: **crash, report, florida, injur, golfer**. The prominent words can many times be self explanatory: **accenture, drop, stop, golfer, sponsorship**. This refers to the Accenture's decision to drop a sponsorship with Tiger Woods. The algorithm can successfully detect appearance dates of emerging topics. While prominent words of the 11th document with the traditional $TfIdf$ does not include the word “voicemail”, the Streaming TfIdf algorithm correctly identifies it as breaking news and adds it to the prominent words.

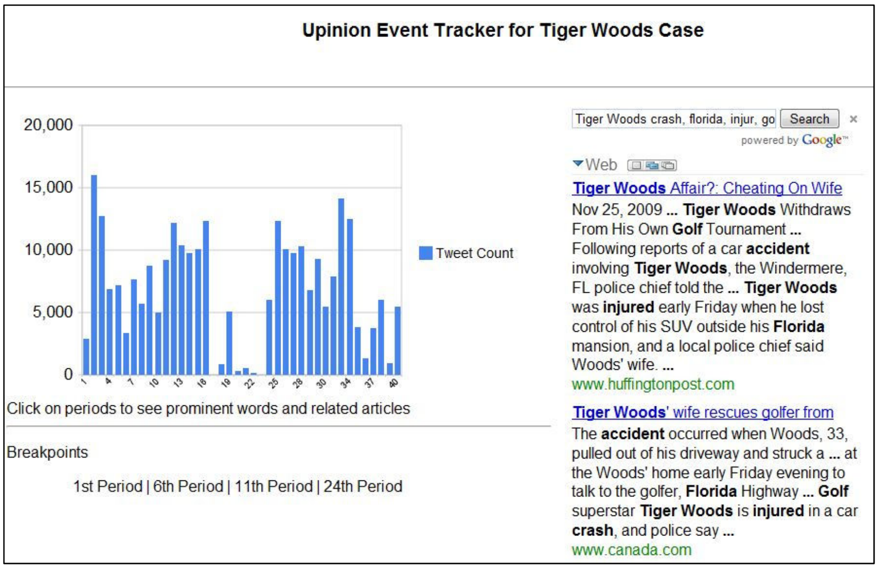
Apart from identifying the prominent words, the algorithm correctly discriminates against words that are not related to the events. In the $11th$ interval, the word ”Afghanistan” is in the set of prominent words. It is because of the tweets that protest Tiger Wood headlines while ”Afghanistan war” gets more violent. In the following days, the prominent word set of the document is updated and ”Afghanistan” disappears from the prominent word set, as it is not actually related to the event.

The breakpoint representation method identifies the significant periods as 6, 11 and 24. Note that, a break on the $(n)th$ bar in the similarity graphs (Figures 3.3-3.4) indicates an opinion change between $(n)th$ and $(n+1)th$ time periods. For these breakpoints, Table 3.1 gives us the prominent words for $(n+1)th$ intervals.

Run Time Analysis of our methods show a linear characteristic as the tweet count increases. In order to test scalability, we experimented with $5000, 10000$ and 20000 tweets and found the run time of our methods to be $24224, 45985$ and 92867 miliseconds on AMD Turion Dual-Core 2.00GHz processor.

Period	Prominent Words
1	***crash, florida, injur, golf, accident***
6	***crash, wife, accident, mistress, golf***
11	***voicemail, wife, f***, golf, cheat***
24	***drop, stop, santa, claus, gatorade***

Table 3.1. Prominent Values for Significant Periods


Upinion Event Tracker for Tiger Woods Case
20,000
15,000
10,000
5,000
0
1
4
7
10
13
16
19
22
25
28
30
34
37
40
Tweet Count
Click on periods to see prominent words and related articles
Breakpoints
1st Period | 6th Period | 11th Period | 24th Period
Tiger Woods crash, florida, injur, go
Search
×
powered by Google
Web

Tiger Woods Affair?: Cheating On Wife

Nov 25, 2009 ... **Tiger Woods** Withdraws From His Own **Golf** Tournament ... Following reports of a car **accident** involving **Tiger Woods**, the Windermere, FL police chief told the ... **Tiger Woods** was **injured** early Friday when he lost control of his SUV outside his **Florida** mansion, and a local police chief said Woods' wife. ...

www.huffingtonpost.com

Tiger Woods' wife rescues golfer from

The **accident** occurred when Woods, 33, pulled out of his driveway and struck a ... at the Woods' home early Friday evening to talk to the golfer, **Florida** Highway ... **Golf** superstar **Tiger Woods** is **injured** in a car **crash**, and police say ...

www.canada.com

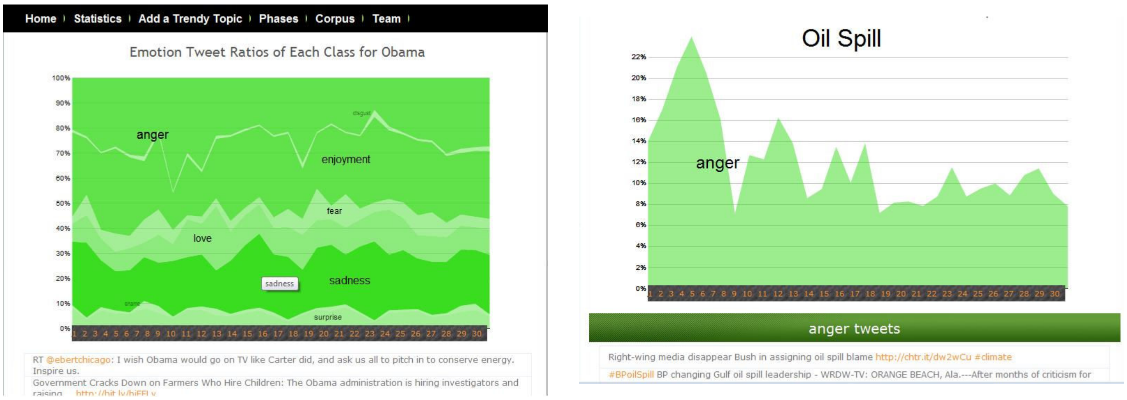
Figure 3.5. Upinion Application

3.5 Customized News Tracking

We developed a news tracking application using Twitter. The resulting application can be seen at the project web site[5], and its screen shot is given in Figure 3.5. In the web site, we are batch processing user entered topics, and providing users with an interface that shows all emotion changes of a topic(Figure 3.6(a)).

When clicked on a class, users can see the changes in this emotion class(Figure 3.6(b)). We also serve users with tweets that contain emotions from this class(Figure 3.7).

[5] http://upinion.cse.buffalo.edu



RT @ebertchicago: I wish Obama would go on TV like Carter did, and ask us all to pitch in to conserve energy. Inspire us.

Government Cracks Down on Farmers Who Hire Children: The Obama administration is hiring investigators and raising ... http://bit.ly/hiFFLv

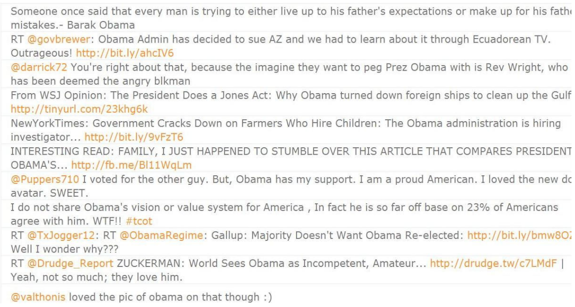
(a) All emotions for a topic

Right-wing media disappear Bush in assigning oil spill blame http://chtr.it/dw2wCu #climate

#BPoilSpill BP changing Gulf oil spill leadership - WRDW-TV: ORANGE BEACH, Ala.---After months of criticism for

(b) Anger emotion in Tweets

Figure 3.6. Screen shots of Upinion



Someone once said that every man is trying to either live up to his father's expectations or make up for his fath mistakes.- Barak Obama

RT @govbrewer: Obama Admin has decided to sue AZ and we had to learn about it through Ecuadorean TV. Outrageous! http://bit.ly/ahcIV6

@darrick72 You're right about that, because the imagine they want to peg Prez Obama with is Rev Wright, who has been deemed the angry blkman

From WSJ Opinion: The President Does a Jones Act: Why Obama turned down foreign ships to clean up the Gulf http://tinyurl.com/23khg6k

NewYorkTimes: Government Cracks Down on Farmers Who Hire Children: The Obama administration is hiring investigator... http://bit.ly/9vFzT6

INTERESTING READ: FAMILY, I JUST HAPPENED TO STUMBLE OVER THIS ARTICLE THAT COMPARES PRESIDENT OBAMA'S... http://fb.me/Bl11WqLm

@Puppers710 I voted for the other guy. But, Obama has my support. I am a proud American. I loved the new dc avatar. SWEET.

I do not share Obama's vision or value system for America , In fact he is so far off base on 23% of Americans agree with him. WTF!! #tcot

RT @TxJogger12: RT @ObamaRegime: Gallup: Majority Doesn't Want Obama Re-elected: http://bit.ly/bmw8O Well I wonder why???

RT @Drudge_Report ZUCKERMAN: World Sees Obama as Incompetent, Amateur... http://drudge.tw/c7LMdF | Yeah, not so much; they love him.

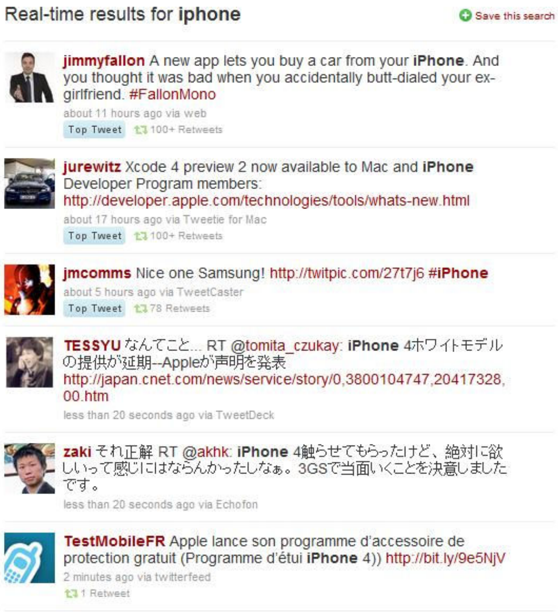
@valthonis loved the pic of obama on that though :)

Figure 3.7. Tweets for an emotion class

This way, we improve the search experience of users who would otherwise see many tweets that are spam, ads or simply in another language (Figure 3.8).

The application uses an interactive Javascript interface that lists the tweet counts of each period. The user can click on the period columns to see the events of a time period depending on the prominent words. For each period, we search for the articles that are published in the date range of the period. We are not storing those web links in a database, because the links can be removed or re-located over time.

Google Alert offers such a customized web service, and it provides a system which notifies users by email when a chosen keyword has a new entry on web. Whereas Google sends updates about every entry on a tracked keyword, our application ob-



Real-time results for **iphone**

jimmyfallon A new app lets you buy a car from your **iPhone**. And you thought it was bad when you accidentally butt-dialed your ex-girlfriend. #FallonMono
about 11 hours ago via web
Top Tweet 100+ Retweets

jurewitz Xcode 4 preview 2 now available to Mac and **iPhone** Developer Program members: http://developer.apple.com/technologies/tools/whats-new.html
about 17 hours ago via Tweetie for Mac
Top Tweet 100+ Retweets

jmcomms Nice one Samsung! http://twitpic.com/27t7j6 **#iPhone**
about 5 hours ago via TweetCaster
Top Tweet 78 Retweets

TESSYU なんてこと... RT @tomita_czukay: **iPhone** 4ホワイトモデルの提供が延期--Appleが声明を発表 http://japan.cnet.com/news/service/story/0,3800104747,20417328,00.htm
less than 20 seconds ago via TweetDeck

zaki それ正解 RT @akhk: **iPhone** 4触らせてもらったけど、絶対に欲しいって感じにはならんかったしなぁ。3GSで当面いくことを決意しましたです。
less than 20 seconds ago via Echofon

TestMobileFR Apple lance son programme d'accessoire de protection gratuit (Programme d'étui **iPhone** 4)) http://bit.ly/9e5NjV
2 minutes ago via twitterfeed
1 Retweet

Figure 3.8. Twitter search results for IPhone

serves the public opinion to identify breaking points and finds keywords of important events to notify users about them.

CHAPTER 4
CONCLUDING REMARKS

We presented a crowd-sourcing system architecture over Twitter, and demonstrated this system with two case studies: weather radar and noise mapping. Our experiments with crowd-sourcing on Twitter are promising. Even without an incentive structure, Twitter users volunteer to participate in our crowd-sourcing experiments (with around 15% reply rates) and the latency of the replies are low (50% replies arrive in 30 minutes and 80% replies arrive in 2 hours). Another promising finding is that a majority of replies were tweeted from smartphones.

Our experiments suggest that Twitter provides a suitable open publish-subscribe infrastructure for tasking/utilizing sensors and smartphones and can pave the way for ubiquitous crowd-sourced sensing and social collaboration applications. There are several open research questions remaining for fulfilling this vision. Security and trust issues remain as significant challenges.

In the Upinion project, we presented an efficient way to observe public opinion on temporal dimension. Our methods can identify break points, and find related events that caused these opinion changes. We tested our results with the timeline of Tiger Woods case and showed the accuracy of our results. We developed an application that can serve users with news pages depending on the time period. We are currently working on expanding the emotion corpus for eliminating outlier intervals in our analysis.

BIBLIOGRAPHY

[1] Analytics, Pear. http://bit.ly/uizmb.

[2] Arora, A., Dutta, P., Bapat, S., Kulathumani, V., Zhang, H., Naik, V., Mittal, V., Cao, H., Demirbas, M., Gouda, M., Choi, Y-R., Herman, T., Kulkarni, S. S., Arumugam, U., Nesterenko, M., Vora, A., and Miyashita, M. A line in the sand: A wireless sensor network for target detection, classification, and tracking. *Computer Networks (Elsevier) 46*, 5 (2004), 605–634.

[3] Arora, Anish, Ed. *ExScal: Elements of an Extreme Scale Wireless Sensor Network* (2005).

[4] Balog, K., Azzopardi, L., and de Rijke, M. A language modeling framework for expert finding. *Inf. Process. Manage. 45*, 1 (2009), 1–19.

[5] Berry, Michael W., Ed. *Survey of text mining: clustering, classification, and retrieval.* Springer, 2004.

[6] Bill, David. http://www.davidbill.org/ 2009/04/17/scholarly-crowdsourcing - twitter-does-history/.

[7] Blogging, Beat. http://beatblogging.org/2009/02/10/leaderboard-for-2-9-2009-crowdsourcing-edition/.

[8] Bohringer, M., and Richter, A. Adopting Social Software to the Intranet: A Case Study on Enterprise Microblogging. In *Proceedings of the 9th Mensch & Computer Conference* (2009), pp. 293–302.

[9] Brabham, Daren C. Crowdsourcing the public participation process for planning projects. *Planning Theory 8* (2009), 242–262.

[10] Broder, Andrei. A taxonomy of web search. *SIGIR Forum 36*, 2 (2002), 3–10.

[11] Burke, Jeff, and et al. Participatory sensing. In *ACM Sensys World Sensor Web Workshop* (2006).

[12] CNN. http://bit.ly/qcmkq.

[13] CNN. http://www.cnn.com/2009/tech/04/17/ashton.cnn.twitter.battle/index.html.

[14] Coppola, Paolo, Lomuscio, Raffaella, Mizzaro, Stefano, and Nazzi, Elena. m-dvara 2.0: Mobile & web 2.0 services integration for cultural heritage. In *SWKM* (2008).

[15] Crumb, Robert. http://www.insideria.com/riaimages/tuar_0103.png. crumb.

[16] Daggle.com. http://bit.ly/40hvom.

[17] Dave, Kushal, Lawrence, Steve, and Pennock, David M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW* (2003), pp. 519–528.

[18] Davidson, Taylor. http://www.taylordavidson.com/ writing/2009/05/19/ ambient-intimacy/.

[19] Dell. http://bit.ly/md6s.

[20] Demirbas, Murat. http://twitter.com/muratdemirbas.

[21] Diakopoulos, Nicholas, and Shamma, D. A. Characterizing debate performance via aggregated twitter sentiment. In *Conference on Human Factors in Computing Systems (CHI)* (April 2010).

[22] e2Campus. http://www.e2campus.com/ pr081203-pacific _facebook_twitter.htm.

[23] Eagle, N., and Pentland, A. Social serendipity: Mobilizing social software. *IEEE Pervasive Computing 04-2* (2005), 28–34.

[24] Eagle, Nathan. txteagle: Mobile crowdsourcing. In *IDGD '09: Proceedings of the 3rd International Conference on Internationalization, Design and Global Development* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 447–456.

[25] Ebner, M., and Schiefner, M. Microblogging-more than fun. In *Proceedings of IADIS mobile learning conference* (2008), vol. 155, p. 159.

[26] Economist. http://www.economist.com/world/ middleeast-africa/ displaystory.cfm?story_id=13856224, June 2009.

[27] Gawker.com. http://gawker.com/380288/ twitter-saves-american -arrested-in-egypt.

[28] Guardian, The. http://bit.ly/12kql4.

[29] Gutwin, C., and Greenberg, S. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW) 11*, 3 (2002), 411–446.

[30] Hatzivassiloglou, V., Klavans, J. L., and Eskin, E. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *In Proceedings Of The 1999 Joint SIGDAT Conference On Empirical Methods In Natural Language Processing And Very Large CORPORA* (1999), pp. 203–212.

[31] Holotescu, C., and Grosseck, G. Using microblogging in education. Case Study: Cirip. ro. In *6th International Conference on e-Learning* (2009).

[32] HowStuffWorks. http://computer.howstuffworks.com/internet/social-networking/networks/twitter2.htm.

[33] HubSpot. State of twittersphere.
http://cdnqa.hubteam.com/State_of_the_Twittersphere_by_HubSpot_Q4-2008.pdf, December 2008.

[34] HubSpot. State of twittersphere.
http://blog.hubspot.com/Portals/249/sotwitter09.pdf, June 2009.

[35] Insider, Business. http://bit.ly/cj0dy.

[36] Jansen, B.J., Zhang, M., Sobel, K., and Chowdury, A. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology 60*, 11 (2009), 2169–2188.

[37] Java, Akshay, Song, Xiaodan, Finin, Tim, and Tseng, Belle. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (New York, NY, USA, 2007), ACM, pp. 56–65.

[38] Jin, Wei, Ho, Hung Hay, and Srihari, Rohini K. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *KDD* (2009), pp. 1195–1204.

[39] Kolari, P., Finin, T., Lyons, K., Yesha, Y., Yesha, Y., Perelgut, S., and Hawkins, J. On the structure, properties and utility of internal corporate blogs. *Growth 45000* (2007), 50000.

[40] Ku, L.W., Liang, Y.T., and Chen, H.H. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs* (2006), pp. 100–107.

[41] Kwak, Haewoon, Lee, Changhyun, Park, Hosung, and Moon, Sue B. What is twitter, a social network or a news media? In *WWW* (2010), pp. 591–600.

[42] Maine. http://www.maine.gov/portal/cas/index.shtml.

[43] marketer, Chief. http://chiefmarketer.com/online _marketing/ 0616-marketing-twitter/.

[44] Mei, Q., and Zhai, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (New York, NY, USA, 2005), ACM Press, pp. 198–207.

[45] Metzler, D., Dumais, S., and Meek, C. Similarity measures for short segments of text. In *In Proceedings Of ECIR-07* (2007).

[46] Nasukawa, T., and Yi, J. Sentiment analysis: capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture* (New York, NY, USA, 2003), ACM, pp. 70–77.

[47] Nielsenwire. http://bit.ly/pd064.

[48] Nielsenwire. http://bit.ly/qk7hu.

[49] Pang, Bo, and Lee, Lillian. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1-2 (2007), 1–135.

[50] Parrott, W. Gerrod, Ed. *Emotions in social psychology: essential readings.* Psychology Press, 2001.

[51] Peek. www.twitterpeek.com.

[52] Popescu, AM, and Etzioni, O. Extracting product features and opinions from reviews. In *EMNLP-05* (2005).

[53] Publishing, Harvard Business. http://bit.ly/3kcljd.

[54] PuppetGov. http://blog.puppetgov.com/2009/07/07/ twitter-your-death-cities -rethink-high-tech-alert-systems/.

[55] Reichelt, Leisa. http://www.disambiguity.com/ambient-intimacy/.

[56] Robson, M. How teenagers consume media. http://bit.ly/j0oce.

[57] Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M., and Sperling, J. Twitterstand: news in tweets. In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2009), ACM, pp. 42–51.

[58] Sarno, David. http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html. Latimes.com.

[59] Shah, Neeta, Dhanesha, Ashutosh, and Seetharam, Dravida. Crowdsourcing for e-governance: case study. In *ICEGOV '09: Proceedings of the 3rd International Conference on Theory and Practice of Electronic Governance* (New York, NY, USA, 2009), ACM, pp. 253–258.

[60] Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Mainwaring, A. M., and Estrin, D. Habitat monitoring with sensor networks. *Commun. ACM 47*, 6 (2004), 34–40.

[61] TechCrunch. http://tcrn.ch/bjb5w7.

[62] TechCrunch. http://www.techcrunch.com/2009/02/19/the-top-20-twitter-applications/.

[63] Telegraph, The. http://bit.ly/g77bz.

[64] Telegraph, The. http://bit.ly/stex.

[65] Times, The New York. http://bit.ly/qxqzn.

[66] Trusov, M., Bucklin, R., and Pauwels, K. Monetary value of word-of-mouth marketing in online communities, 2009.

[67] Turney, Peter D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL* (2002), pp. 417–424.

[68] twitpic. http://twitpic.com/7mi3f.

[69] Twitpic. http://twitpic.com/135xa, January 2009.

[70] Twitter. http://twitter.com/pistachio.

[71] Twitter. www.twitter.com.

[72] Vark. http://vark.com/.

[73] Week, Business. http://www.businessweek.com/ blogs/whatsyourstoryidea/.

[74] Weiser, Mark. The future of ubiquitous computing on campus. *Commun. ACM 41*, 1 (1998), 41–42.

[75] Werner-Allen, Geoff, Lorincz, Konrad, Johnson, Jeff, Lees, Jonathan, and Welsh, Matt. Fidelity and yield in a volcano monitoring sensor network. In *in 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2006)* (2006).

[76] Wikipedia. http://en.wikipedia.org/wiki/crowdsourcing.

[77] Wikipedia. http://en.wikipedia.org/wiki/pareto_principle.

[78] Zhuang, L., Jing, F., Zhu, X.-Yan, and Zhang, L. Movie review mining and summarization. In *CIKM-06* (2006).