TOWSON UNIVERSITY

OFFICE OF GRADUATE STUDIES

DATA COLLECTION IN A SOCIAL NETWORK WITH WEIGHTED SEED SELECTION AND DATA ANALYSIS BASED ON RULE-BASED METHODS

By

Changhyun Byun

A Dissertation

Presented to the faculty of

Towson University

in partial fulfillment

of the requirement for the degree

of Doctor of Science

Department of Computer & Information Sciences

Towson University

8000 York Road, Towson, MD 21252

May, 2013

TOWSON UNIVERSITY

OFFICE OF GRADUATE STUDIES

DISSERTATION APPROVAL PAGE

This is to certify that the dissertation prepared by Changhyun Byun, entitled "Data Collection in a Social Network with Weighted Seed Selection and Data Analysis Based on Rule-Based Method," has been approved by the thesis committee as satisfactorily completing the dissertation requirements for the degree Doctor of Science in Information Technology.

May 10, 2013

Chair, Dissertation Committee, Dr. Yanggon Kim

Date

MAY 10, 2013

Committee Member, Dr. Josh Dehlinger

Date

May 10, 2013

Committee Member, Dr. Sungchul Hong

Date

5/10/2013

Committee Member, Dr. Siddharth Kaza

Date

Janet V DeLany

5/29/2013

Dean of Graduate Studies

Date

ACKNOWLEDGEMENT

I am grateful to a number of people who have guided and supported me throughout the research process and provided assistance for my venture. I would first like to gratefully and sincerely thank Dr. Yanggon Kim for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at Towson University. His mentorship was paramount in providing a well-rounded experience consistent my long-term career goals. For everything you've done for me, Dr. Kim, I thank you.

I would also like to thank Dr. Ohoe Kim for his assistance and guidance in getting my graduate career started on the right foot and providing me with the foundation for becoming a creative researcher. He encouraged me to not only grow as a scientist of computer science but also as an independent researcher. I believe that his actions provided me with the unique opportunity to gain a wider breadth of experience while still a graduate student.

I would like to thank my committee members, Dr. Josh Dehilinger, Dr. Sungchul Hong, and Dr. Siddharth Kaza, whose work demonstrated to me that concern for global affairs supported by an "engagement" in comparative literature and modern technology, should always transcend academia and provide a quest for our times.

I am very much indebted to my family, my parents, mother-in-law, and father-in-law who supported me in every possible way to see the completion of this work. I also would like to thank my sister, brother-in-law, and the rest of my family for their love and encouragement. Most importantly, I would like to thank my wife, Misuk Ha, for her love, support, and patience during the past four or so years it has taken me to graduate.

Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

This dissertation is dedicated to

Misuk, Hannah, and to all my family.

I sincerely love you all.

ABSTRACT

DATA COLLECTION IN A SOCIAL NETWORK WITH WEIGHTED SEED SELECTION AND DATA ANALYSIS BASED ON RULE-BASED METHODS

By

Changhyun Byun

In recent years, with the increasing popularity of diverse online social network sites, such as Facebook, Twitter, Blogger, YouTube, LinkedIn, and MySpace, a massive amount of data has become available. Analyzing sets of data in social media can lead to some understanding of individual and human behavior, detection of hot topics, identification of influential people, and/or discovery of a group or community. However, it is difficult to discover useful information from social data without automated information processing because of three main characteristics of social media data sets: the data is large, noisy, and dynamic. In order to overcome these challenges of social media, data-mining techniques can be used by data seekers to discover a diversity of perspectives that would otherwise not be possible.

To apply data-mining techniques to social data, the target data set must be prepared from social networks before the analyzing process. For these reasons, Twitter enables researchers and data analyzers to access a variety of data in Twitter by providing Application Programming Interface (API). However, there is a restriction on data collection from Twitter: the method call of Twitter API is limited. Furthermore, it is impossible to collect enough data to apply data analysis techniques and filter out unnecessary data, such as spam messages without an automated data collector and filter. In order to overcome these data access problems, we aim to design and implement our own Twitter data-collection tool, which includes data filtering and analysis capabilities. This allows us, as well as other researchers and data seekers, to build their own Twitter dataset.

First, in this research we introduce the design specifications and explain the implementation details of the Twitter Data Collecting Tool we developed. To introduce and explain the implementation details and the design specifications of the Twitter Data Collecting Tool, the Unified Modeling Language (UML) diagram is used.

We next propose a new algorithm that selects the best seed nodes with limited resources and time to collect the data related to a specific topic and keyword efficiently. The algorithm also evaluates various user influence and activity factors, and updates the seed nodes dynamically during the gathering process. After the gathering process, we compared two results, one from this algorithm and one from a specialist.

In the final chapter, we provide an analysis of Twitter data gathered by the Twitter Data Collecting Tool in a case study about the Super Bowl 2012 and Super Bowl 2013. The case study aims to address the question of how people use Twitter and to assess the power of Twitter in creating consumer interest in brands and commercials. The main objective of this study is to find the relationship between Twitter and Super Bowl advertisements by analyzing data on Twitter.

This research shows that the Twitter Data Collecting Tool allows researchers to gather users' information, follow relationships and tweets from Twitter. Furthermore, the data collection result with the seed selection algorithm proved that the efficiency of the algorithm for collecting more keyword-related data is higher than the existing approach. In addition, data-mining techniques and rule-based data analysis are applied to the gathered data. With these results, we could prove that the Twitter Data Collecting Tool is able to gather a huge amount of data from Twitter and filter the data so it can be used in research areas. This paper will be valuable to those who may want to build their own Twitter dataset, apply customized filtering options to get rid of unnecessary, noisy data, and analyze social data to discover new knowledge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application Programming Interface
CLI	Command Line Interface
eWOM	Electronic Word-of-Mouth
GUI	Graphical User Interface
MVC	Model View Controller
UML	Unified Modeling Language
U&G	Uses and Gratification
WOM	Word-of-Mouth

Chapter 1 .

Introduction

In recent years, with the increasing popularity of diverse online social network sites, such as Facebook, Twitter, Blogger, LinkedIn, YouTube, and MySpace, a massive amount of data has become available. Particularly, Twitter has become one of the fastest growing social media sites since its launch. Developed in 2006, Twitter is a web-based social networking site and microblogging service that allows users to post "tweets" using up to 140 characters per post. Twitter has reached a user population above 500 million total users and more than 200 million active users in 2012. Additionally, the number of messages that Twitter users exchanged per day has increased from 300,000 in 2008 to 200 million in 2011, and to 340 million by February 2013 ("200 million Tweets," 2011; "What is Twitter?" 2012).

Technology and the Internet allow users of social media to access and share data frequently and rapidly. The growth and popularity of social media in our daily life have transformed the ways we communicate as a citizen, consumer, and audience member. People post their messages on social network sites for a variety of purposes, such as sharing information, conversations, updating real-time statuses, reporting the news, and expressing opinions.

Analyzing sets of data in social media can lead to some understanding of individual and human behavior, detection of hot topics, identification of influential people, or discovery of a group or community. Additionally, various business companies have made substantial efforts to accommodate such swift trends and have paid attention to the competitive advantages of using social media in marketing. It is not unusual to see icons for Facebook and Twitter on many businesses' websites and advertising messages. Social media became a "blessing" for marketers, allowing them to reach a wide range of target audiences efficiently. Simultaneously, the same social media sites create serious challenges in the marketing world, since consumers are further fragmented with various media platforms and they can jeopardize brand equity and brand images with active social media use (Kaplan and Haenlein 2010). Therefore, it became imperative for marketing and advertising professionals to better understand the complex behaviors and minds of consumers: Why do consumers choose to use Twitter, how do they use it, and what interactions do they have with other users? Facing the growing importance and popularity of this new medium, this study is intended to build a better understanding of the usages of social media among consumers, particularly Twitter use on Super Bowl advertising and to investigate how electronic word-of-mouth (eWOM) on this topic is spreading.

However, it is difficult to discover useful information from social data without automated information processing because of three main characteristics of social media data sets: the data is large, noisy, and dynamic. In order to overcome these challenges of social media, data-mining techniques can be used by data seekers to discover a diversity of perspectives that would otherwise not be possible. Data mining techniques are widely used to handle large sets of data and to discover new knowledge and useful information in a data set that is not readily obtainable and not always easily detectable. Hence, applying data mining techniques to online social media benefits many groups, such as market researchers, psychologists, sociologists, businesses, and politicians, fascinating insights into human behavior, marketing, business or political views (Cortizo et al., 2009; Sottara et al., 2011).

To apply data-mining techniques to social data, the target data set must be prepared from social networks before the analyzing process. For these reasons, Twitter enables researchers and data analyzers to access a variety of data in Twitter by providing API. Numerals of researchers collected Twitter data through Twitter APIs and applied data-mining techniques into their own data set to detect issues, such as earthquakes (Okazaki and Matsuo, 2010) and influenza by using Twitter (Aramaki et al., 2011) or recommending tags to users (Correa and Sureka, 2011). However, there is a restriction on data collection from Twitter: the method call of Twitter API is limited by 350 calls per hour for one authorized developer account (Twitter, 2013). Furthermore, it is impossible to collect enough data to apply data analysis techniques and filter out unnecessary data, such as spam messages without an automated data collector and filter. In order to overcome these data access problems, we aim to design and implement our own Twitter data-collecting tool, which includes a data filtering capability. This allows us, as well as other researchers and data seekers, to build their own Twitter dataset.

Over the years numerous studies have attempted to apply data mining techniques to social media. Seeds analysis, as a young field, emerged as one of representative studies related to data mining in social communities. Seeds analysis technique can be used to maximize the spread of influence through a social network, to observe and track popular events in social communities, to discovering popular events in a time-variant social community, or to investigate information diffusion processes for “word-of-mouth” and “viral marketing” effects (Domingos and Richardson, 2001; Richardson and Domingos, 2001; Lin *et al*.,2010). The key of selecting most influential nodes is finding a small set of seed nodes in given data set that maximize the spread of influence maximize influence.

Among the many considerations to gather the data related to a specific topic, there is no doubt that selecting seed nodes used for staring point of data gathering process is the most important step to gather more relative data for a specific topic. Thus, we propose an algorithm to find suitable seed nodes, which can maximize the efficiency of data gathering process to collect more topic-related data from Twitter. The algorithm considers user influences and activities to find the best initial seed nodes dynamically with limited resources and time.

The remainder of this paper is constructed as follows: In Chapter 2, the related work done so far is summarized. In Chapter 3, we will introduce the design specifications and explain the implementation details of the Twitter Data Collecting Tool we developed. In Chapter 4, we propose an algorithm to find suitable seed nodes, which can maximize the efficiency of data gathering process to collect more topic-related data from Twitter. In Chapter 5, we provide an analysis of Twitter data gathered by the Twitter Data Collecting Tool in a case study about the Super Bowl 2013 and Super Bowl 2013. The case study aims to address the question of how people use Twitter and to assess the power of Twitter in creating consumer interest in brands and commercials. The last part, Chapter 6, concludes the work by summarizing this paper and suggesting future research directions.

Chapter 2 .

Literature Review

In this chapter, we introduce the key concepts, terminology, and methodologies that are used in this research.

2.1. Data Crawling Tool in Twitter

2.1.1. TwitterEcho-Opensource Twitter Crawler

Bošnjak et al. presented an open source crawler, TwitterEcho, which is used to retrieve data from Twitter (Twitter, 2013). It allows data seekers to collect data from a focused community of interest. TwitterEcho adapts a centralized distributed architecture and includes three main components: clients, servers, and modules. Figure 1 shows the architecture and main components of the TwitterEcho Crawler.

A client consists of two modules. The first module collects tweets, user profiles, and simple statistics. The second module collects social network relations. The number of clients can be increased to retrieve more data from Twitter. The server manages the crawling process and allocation of user lists to each client. It also maintains the database in which the downloaded data is saved. Modules consist of user expansion, user selection, and user inspection. The user expansion module analyzes downloaded tweets, extracts the user's mentions, and adds the user's followers to the list of tentative users. The user selection module identifies users' accounts to be monitored by analyzing profiles and identifying languages. The user inspection module also monitors events, such as deletion, suspension, and the activity of users' accounts.

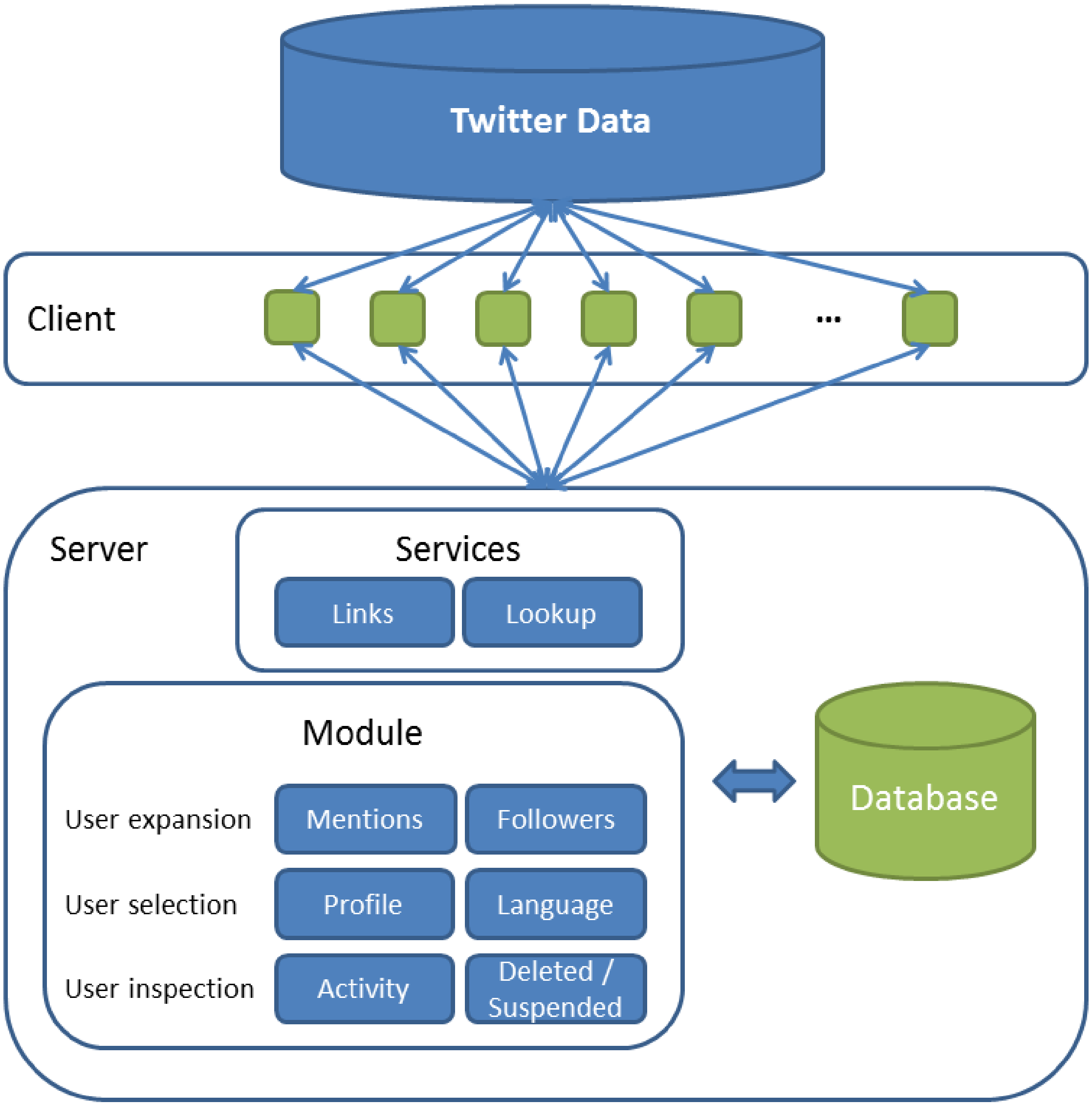

Figure 1. Architecture of Twitter Echo

Despite the fact that TwitterEcho is able to gather a huge amount of data from Twitter, there are still some problems with the program. First of all, it would be more efficient to retrieve data for a focused community of interest, starting with multiple seeds. However, the program does not support this function. Secondly, TwitterEcho starts with a seed node and keeps inspecting its followers and their followers' data. Since TwitterEcho does not restrict the level of followers, it is likely to have an enormous amount of noisy data. Finally, retrieving data is restricted to Portuguese only. Even though we can adapt additional modules to TwitterEcho for a specific area, it needs time finding or implementing these additional modules.

2.1.2. Twitter User Data Using Resource of Cloud

Another approach to collecting Twitter data is the use of the computation power of cloud computing (Noordhuis et al., 2010). Noordihuis et al. gathered Twitter data and applied the PageRank algorithm to rank Twitter users using the computation power of cloud computing.

There were five steps in this cloud system. First, a queue and table are set up to maintain all user IDs that need to be crawled. Then, users' and followers' IDs are saved in the SimpleDB, which is a service for storing structured data in the cloud. Furthermore, different users' information is gathered for different instances simultaneously by using their own web service. In the fourth step, the PageRank Algorithm is applied to rank users. Finally, a web interface enables public users to access their data. As a result, 50 million users and 1.8 billion followers' information were crawled, and Twitter users were analyzed using the PageRank algorithm.

Even though they showed cloud services are feasible to gather huge amounts of social data, there are also some problems with their approach. First, they did not save gathered data into local storage because the additional usage of storage would be costly. Moreover, whitelist accounts were used to gather Twitter data through the Twitter API. However, due to current Twitter policy, getting a new whitelist account is not feasible. Furthermore, it is not suitable for some specific research areas, such as data mining. Because they crawled all user data from Twitter, there is a lot of noisy data that many researchers are not interested in. For that reason, it would consume a lot of resources to analyze and filter this data. Additionally, they did not gather tweets data that can be used for analyzing users' tendencies and opinions.

2.1.3. Crawling Twitter Data Using Whitelist Accounts

Kwak et al. gathered Twitter data to study the topological characteristics of Twitter and its power for information sharing (Kwak et al., 2010). By using Twitter APIs, they gathered all users' profiles, trending topics, and tweets that mentioned the trending topics. As a result, they successfully crawled the entire Twitter site, including 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. However, they used whitelist accounts that are useful to gather huge amounts of data from Twitter but are no longer available. Also, Twitter is not suitable for some research areas as mentioned in Section 2.2.

2.2. Methodologies of Analyzing Sentiment of Messages in Social Networks

Analyzing sentiment of messages in social networks has been studied in various ways. Dey and Haque proposed a localized linguistic approach to extract opinion expressions from noisy text that are generated from online chat, emails, blogs, customer feedback, and reviews (Dey and Haque, 2008). Based on the pre-processed noisy texts, they analyzed opinion expressions using a classifier algorithm and candidate opinion words from Wordnet. Hu and Liu suggested a technique to identify opinion sentences in each review about product features and to decide whether each opinion sentence is positive and negative (Hu and Liu, 2004). They determine its semantic orientation using a set of adjective words, which are identified by a natural language processing. O'Connor et al. presented an approach to polarity classification by counting the number of positive and negative expressions in a tweet and selecting the category with more terms (O'Connor et al., 2010). Label propagation, using graph-based methods, can also be used to find opinion from social network sites (Zhu and Ghahramani, 2002; Baluja et al., 2008; Talukdar and Crammer, 2009). Choi and Cardie used domain specific lexicon and relations among words and opinion expressions to classify polarity of messages in a social network site (Choi and Cardie, 2009). However, the accuracy for polarity classification still needs to be improved as the accuracy in other research shows polarity classification of 80%.

2.3. Rule-Based Engine for Filtering and Analyzing Data

A rule is a syntactic construct that is used to check if certain consequences satisfy the premises (Sottara et al., 2010). A rule engine, also known as an inference engine, is a computer program which analyzes patterns by comparing facts and rules, and executes actions declared in the rules. The rule engine can be embedded in the Twitter Data Collecting Tool to filter data. There are many kinds of rule engines available, such as Jess, Jena, Drools, and etc. Among the free and Java-based rule engines, Drools (JBoss, 2013) and Jess (Friedman, 2013) have been considered as potential rule engines in the research.

Drools is a Java-based, object-oriented rule engine, which is open-source, so we can freely use and modify the code in Java. It uses an optimized version of the RETE Algorithm (Forgy, 1982), called the RETE-Object-Oriented algorithm to support high performance. It has its own writing rules, which is called DRL (Drools Resource Language) and is flexible enough to match the semantics of problem domain with DSLs

(Domain Specific Languages), graphical editing tools, web based tools and developer productivity tools. It has useful features which include rule debugging and rule authoring tools like IDE plug-in.

The result is described in the following table. Table 1 gives comparison for the selected inference engines measured on different performance metrics.

TABLE 1. **A Comparison of Rule Engines**

Category	Rule Engine	
	Drools	**Jess**
Algorithm	RETE Algorithm	RETE Algorithm
OWL-DL Entailment	No	Yes
Java Support	Yes	Yes
Rule Support	DRL (Own Rule Format)	SWRL
Version	5.0	7.1
Licensing	Free / Open source	Academic use only

We have chosen the Drools engine to filter and analyze tweets in the Twitter Data Collecting Tool. This is because it is a java-based open-source, which offers different semantics for encoding rules based on the programming language used for expressing conditions and actions, and has a syntax that favors readability.

2.4. Measuring and Maximizing User Influence in Twitter

Cha et al. (2010) measured the influence of users in Twitter using three interpersonal activities: in-degree, retweets, and mentions. In-degree influence is the number of followers of a user, which indicates the size of audience for that user. Re-tweet influence is the number of retweets contacting one's name, which indicates the ability of that user to generate content with pass-along value. Mention influence is the number of mentions containing one's name, which indicates the ability of that user engage others in a conversation. While they showed that user influence can be measured by three perspectives, they analyzed the influence based on the data that has been gathered, which needs a lot of pre-processing time. Also, they did not consider other factors that are related to user's influence in Twitter.

Influence maximization is to find individuals who have the most influence in a social network (Shang *et al*., 2012). Researches have paid a lot of attention to this problem. Domingos and Richardson firstly studied influence maximization problem and proposed a probabilistic solution (Dimingos and Richardson, 2001, Richardson and Domingos, 2002). Kempe et al. presented diffusion models to solve the problem and proved the influence maximization problem to be NP-hard (Kempe *et al*., 2003). They also proposed a greedy approximation algorithm that guarantees performance of the optimal influence, which outperforms the existing node-selection approaches. Chen et al. proposed a new greedy algorithm and heuristics to maximize influence in a social network (Chen *et al*., 2010, Chen *et al*., 2009). However, existing influence maximization approaches need to be improved to reflect the characteristics of a social network and find initial seed nodes for a specific topic with limited time and resources.

2.5. Data Analysis in Mass Media

2.5.1. The Audience's Media Use in the Media Convergence Era

Early studies on the audience's media use indicate that traditional media use, such as television viewing, declined because of the increase in the Internet use (Coffey & Stipp, 1997). Later, researchers observed that the Internet use may not necessarily reduce time people spent on other media (Cooper & Tang, 2009). When facing an expanded media menu, media users increasingly use two or more media simultaneously, or use one media to encourage or enhance a different medium use.

A common assumption held by social psychologists and technology adoption researchers is that few media can fulfill all the goals audiences seek. Accordingly, the audiences select certain media based on their perceived functionality (Ferguson & Perse, 2000; Lin, 2006; Papacharissi & Rubin, 2000). Literature on this indicates similar motives or functions on why people choose certain types of media. Papacharissi and Rubin (2000) summarized five motivations—interpersonal utility, pastime, information-seeking, convenience, and entertainment—that predict the Internet use. Lin (2006) found that motives such as infotainment, escape, and interpersonal communication were related to overall Webcast viewing interest. Further, Hong and Raney (2007) demonstrated that entertainment, information, and perceived interactivity explained why people visited sports sites.

Technology adoption literature further suggests that perceived usefulness, perceived ease of use, attitude, social norms, and perceived behavioral control have an impact on the adoption and actual use of new technologies (e.g., Lederer, Maupin, Sena, and Zhuang, 2000; Porter and Donthu, 2006). Porter and Donthu (2006) suggest that perceptions regarding usefulness and ease of use influenced online media use. Kaye (1998) found that the perceived informational learning and interaction utilities could predict use of the popular web sites. Lin (2006) demonstrated that perceived utilities acted as a significant predictor of the adoption of Webcasting.

2.5.2. The Concept of Active Audience, Uses, and Gratifications

As an influential theory in media research, the uses and gratification perspective (U&G) assumes that different people can use the same medium for different purposes. The theory holds that multiple forms of media compete for users' attention, and audiences select the medium that meets their needs, such as the desire for information, emotional connection, or status (Tan 1985; Baran and Davis 2011). At the core of this theory is the concept of an active audience, which assumes that the audience's communication behavior is goal-directed and purposeful in that people choose certain media based on their needs, wants, or expectations.

Considered to be one of the "old" theories in media research, U&G has recently been revitalized for studying technologies and media consumption behavior. They include research on the web (Ko, Cho, and Roberts 2005; Roy 2009), on blogging (Kaye 2007; Hollenbaugh 2010), and social-networking sites, such as Facebook and Twitter (Muntinga, Moorman, and Smit 2011; Chen 2011). Researchers found that interactivity, recreation, entertainment, diversion, information involvement, connectedness, and personal relevance are all major motivations for browsing or using the Internet and social media platforms. In other words, online users have various degrees of engagement with social media that range from simple and passive (e.g., simple consumption by reading) to active (e.g., producing and posting contents).

2.5.3. Sports, New Media, and Twitter

Transformed sports consumption patterns through new technologies have generated solid literature on the growing importance of the Internet and social media in sports media. Overall, literature on sports and new media has focused on small segments of audiences, such as sports bloggers, sports fans, online streamed sports events viewers, sport video gamers and the Internet sports community members (Conway, 2010; Crawford & Gosling, 2009; Plymire, 2009; Oates, 2009; Schultz and Sheffer, 2008; Jones, 2010; Tang and Cooper, 2012; Cheever, 2009).

One line of literature on this focuses on the gender differences in the social media use. Lam (2001) found that compared to their female counterparts, males were more likely to use streaming videos or Webcasting, while Lin (2004) observed that being female (instead of male) was a better predictor of the adoption of Webcasting. Surveying about 5,000 U.S. adults, Schultz and Sheffer (2011) reported that there was almost no gender difference for the overall social media use for sports. However, among those who identified themselves as “regular” or “occasional” users of the social media for sports, they found that women showed significantly higher levels of social media use for sports than men. Based on previous literature (see Pedersen and Macafee, 2007), it is speculated that women use social media in a sports context for social interaction, while men use it more for information seeking. Also, they suggested that older audiences (age 45 over) are increasingly turning to social media for sports and have as much interest in sports as younger audiences. In their analysis of the audience’s multiplatform experiences with the 2008 Beijing Olympic Games, Tang and Cooper (2012) found that men and women spent similar amounts of total time watching the Olympics on various platforms, including an

NBC broadcast outlet. At the same time, men spent more time watching sports events and newscasts and are more motivated to watch sports than women. Overall, previous literature shows inconsistent results in terms of the relationship between gender groups and their use of new media.

Another line of literature on sports and new media is about the social and information functions that new media play in the world of sports media. Sanderson and Kassing (2011) observed three specific ways in which blogs and Twitter significantly played a role in sports media: transformative, adversarial, and integrative. With a transformative function, blogs and Twitter allowed professional athletes to bypass “the voices” of journalists and to gain more control over what information gets presented to their fans and the public, as well as, how they are represented. In addition, blogs and Twitter made relational conflict between athletes and sport reporters or sports organizations more visible to the public and enabled athletes to be more assertive in rebutting or nullifying any inaccurate elements in sports reporting. While these transformative and adversarial functions of blogs and Twitter benefit athletes, the audiences also benefit by receiving sports information from multiple sources. Sanderson and Kassing (2011) noted that blogs and Twitter integrate “the perspectives of athletes, sports journalists, and fans into sports media stories” (p. 120).

Chapter 3 .

Twitter Data Collecting Tool

Faced with limited options, we designed and developed our own Twitter Data Collecting Tool. In this chapter, we present requirements, architecture, and implementation details of this system.

3.1. System Requirements

Problems that we discussed in the previous section encouraged us to define the following requirements for the Twitter Data Collecting Tool:

- Continuously and automatically collects data from Twitter: Once seed IDs and authorized accounts are configured, the tool must be able to collect tweets and relations related without any human interaction until the stop event has occurred or the target amount of data is collected.
- Runs with multiple seed nodes: In the gathering process, users may want to start the data collecting process from certain Twitter users. To fulfill this need, the system should be able to accept multiple seed nodes.
- Handles a multitude of authorized keys: The tool must be able to handle more than one authorized key. This increases the total number of Twitter API calls.
- Stores collected data in a database: The system has to save all collected data, such as users’ follower/following relations and tweets, into a database.

- Supports intuitive user interface: The tool must support user interface to interact with users. A straightforward and intuitive interface must be provided to start and manipulate the program.

3.2. System Architecture

Figure 2 shows you the constituent controllers of the Data Collecting Tool. The given data collector consists of the Account Handler, the Data Gathering Controller, the Database Handler, Data Gatherer Thread Pool, as well as, the Data Filtering Handler.

The Account Handler is in charge of managing the Twitter developer’s account. As long as Twitter limits the number of method calls for one developer’s account to 350 calls an hour, a program feature to handle multiple developers’ accounts is necessary to build a large database. The main role of the handler is to provide the Data Gathering Controller with current available resources consistently by managing developers’ account objects in a wait queue and a ready queue. When a developer’s account hits its limit of method calls, the module moves the account object to the wait queue and keeps it there until the limit is recharged.

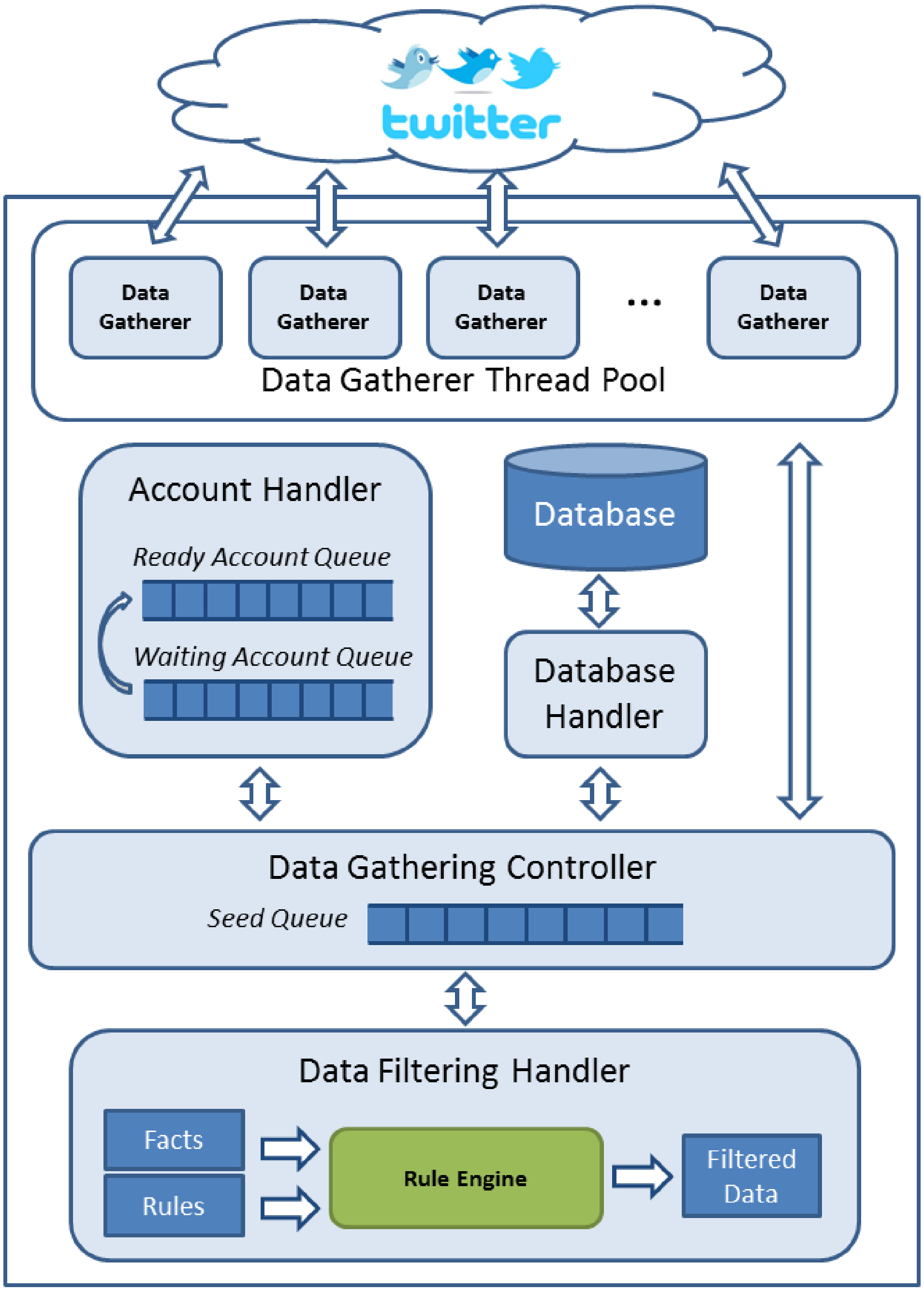

Figure 2. Architecture of the Twitter Data Collecting Tool

Account Handler

3.2.1. Data Gathering Controller

The Data Gathering Controller acts as a backbone of the tool. It receives available developer's account objects from the Developer Account Controller module, and then sends a request to the Data Gatherer threads with specific requirements, such as gathering seed-related users' information or gathering certain users' tweets. When the threads return with results to the Data Gathering Controller, it passes the data to the Database Handler to save the information in a database.

3.2.2. Database Handler

The Database Handler is required in the data-collecting tool to store data gathered in the database and to retrieve stored data back, as well. The Database Handler manipulates multiple database connection objects using a connection pool technique. The handler is designed to endure massive update transactions streamed from the Data Gathering Controller.

3.2.3. Data Gatherer Thread Pool

Each thread has its own data gathering rules to collect users' account information and users' tweets or follower relationships. The gathering rules also contain data filtering options to sift noisy data from the collected data. To run each thread, at least one developer account should be received from the Data Gathering Controller.

3.2.4. Data Filtering Handler

Related works described in the previous section show that there is a huge amount of noisy data that data analyzers are not interested in. Furthermore, analyzing sentiment in social networks has been issued in data mining of social networks. To solve these problems, we embedded a rule-based engine in the Twitter Data Collecting Tool to filter the data and analyze the sentiment of the messages.

The Data Filtering Handler filters the data gathered by the Data Gathering. For the filtering process, we imbedded an open source rule engine, the Drools rule engine, in the Twitter Data Collecting Tool. Once a user enters a keyword in the Tool through the user interface, the Data Filtering Handler generates rules to filter data. Then, the rule engine generates new knowledge through pattern matching process and then outputs the filtered data.

3.2.5. Database Architecture

We designed and implemented a database to store and handle plenty of data easily and effectively. Figure 3 shows an Entity-Relation Diagram for the database. To reflect following/follower relationships between seed nodes and follower nodes, two tables, tbl_user table and tbl_follow_relation table, are defined. A combination of the seed ID and user ID can identify each of the following relations. The user table, tbl_user table, is designed for storing the user's detailed information, such as screen name, first name, last name, location, etc. To save data about tweets, a table, called tbl_tweets table, is defined as well. The tbl_tweets table contains information about tweet ID, user ID, actual message, and date posted. A relationship between user and tweet is formed to track the message creator and its seed.

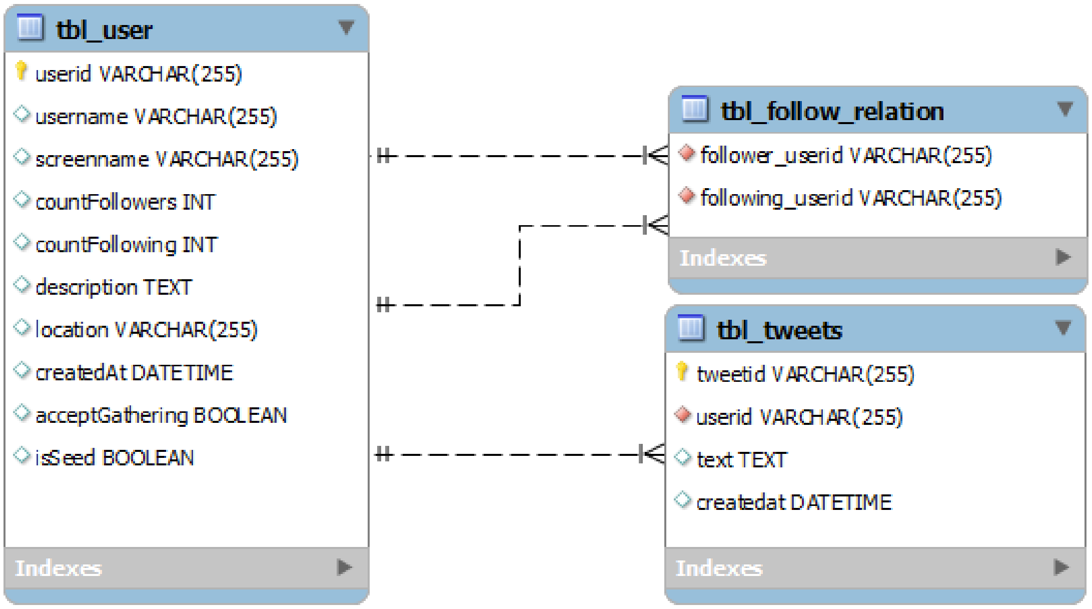

Figure 3. Database Design

3.3. Design of the Twitter Data Collecting Tool

To introduce and explain the implementation details and the design specifications of the Twitter Data Collecting Tool, UML (Unified Modeling Language) diagrams are used. Two of the major UML diagrams, class diagram and sequence diagram, are shown in Figure 3 and Figure 5. The system provides two different types of user interfaces, GUI (Graphical User Interface) and CLI (Command Line Interface).

3.3.1. Object Specification

The MVC (Model View Controller) framework is applied to this system to support various types of user interfaces, such as a standalone interface and a web-based interface. Followed by the MVC framework, objects of the system are divided into three packages: Model, View and Controller, as shown in Figure 4.

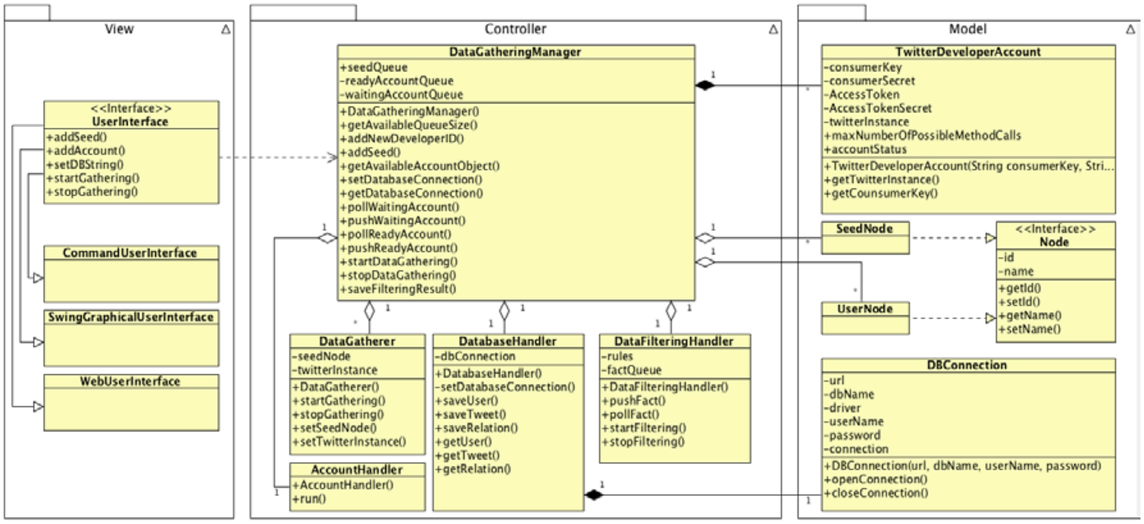

Figure 4. Class Diagram of the Twitter Data Collecting Tool

Only interface-related objects are defined in the View package. Three different types of user interfaces, the CommandUserInterface, SwingGraphicalUserInterface and WebUserInterface, realize the UserInterface interface so that the system is able to provide various interfaces to a user, as well as a new type of interface that can be defined and realized easily.

The Controller package consists of a DataGatheringManager class, a DataGatherer class, an AccountHandler class, a DatabaseHandler class, and a DataFilteringHandler class. The DataGatheringManager acts as the backbone of the system and has duties to transform messages from a user interface to a handler and to control all handlers in the Controller package. The dataGatheringManager class receives information about Twitter accounts, seed IDs, and filtering rules and keywords from one of user interfaces using public operations, such as addSeed() and addNewDeveloperID(). The DataGatherer class is supposed to connect to the Twitter database and collect data through Twitter API. The DatabaseHandler class takes responsibilities to store data into a local database and to retrieve stored data from the local database. The AccountHandler class is in charge of checking if a Twitter account exceeds its limits. If the account has hit its maximum allowed method calls, the AccountHandler class pushes the account into the waiting queue. If the account has not hit the limits the AccountHandler class pushes the account into the ready queue so that the DataGatherer class can always get an available account instance from ready queue. The DataFilteringHandler class is keeping an eye on the fact queue. If there is any case filtered by the rule engine, the DataFilteringHandler class sends the DataGatheringManager class the matched data.

The Model package contains four basic classes and one interface. The four basic classes, the TwitterDeveloperAccount class, the SeedNode class, the UserNode class, and the DBConnection class, are defined as message types. The SeedNode and UserNode are implementations of the Node interface.

3.3.2. Sequence of Data Collecting Process

In this section, we explain the process of data gathering and filtering in the Twitter Data Collecting Tool. A sequence diagram depicts the entire process in Figure 5. First, the process starts with a user's input from a user interface. A user enters seed node IDs, Twitter developer account information, and filtering keywords in the user interface. Once the user enters the input, it invokes the DataGatheringManager class. Then, the DataGatheringManager class executes the DatabaseHandler class, the DataFilteringHandler class, the AccountHandler class to set up a database connection and options for gathering and filtering. Based on the connection and gathering options, the DataGatherer class gathers users, relation and tweets data from Twitter and save them in a designated database. Finally, the DataFilteringHandler class filters the tweets using the keywords. The gathering process ends when the user calls stopGathering() method from user interface.

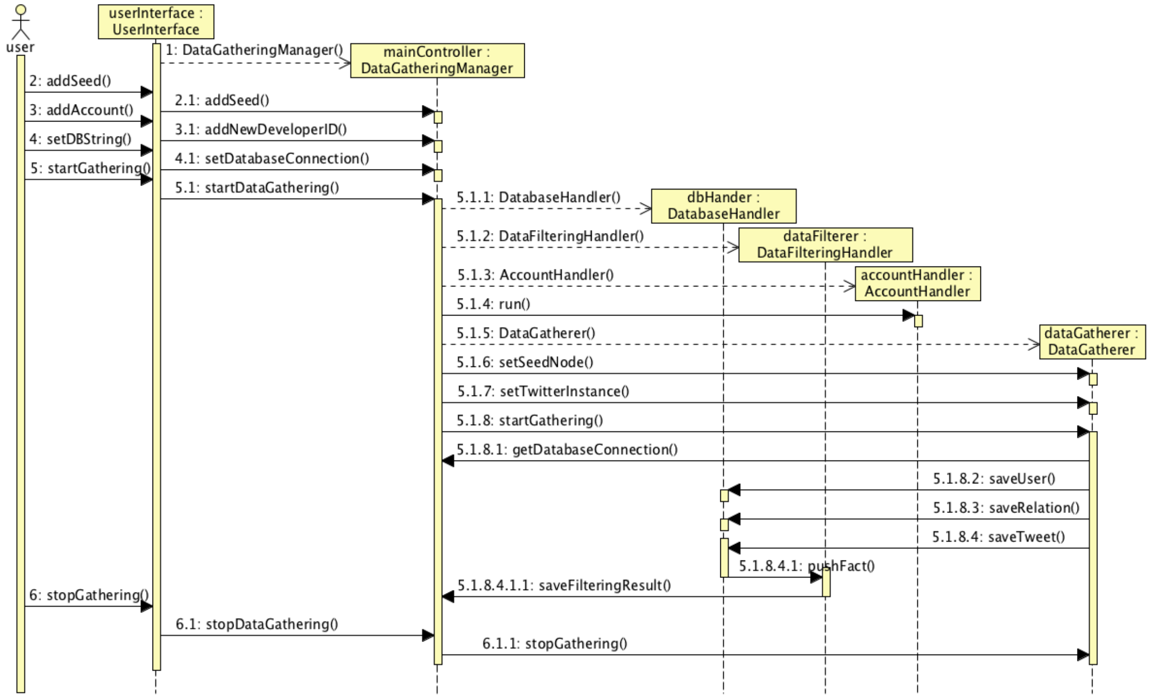
user
userInterface : UserInterface
1: DataGatheringManager()
mainController : DataGatheringManager
2: addSeed()
3: addAccount()
4: setDBString()
5: startGathering()
2.1: addSeed()
3.1: addNewDeveloperID()
4.1: setDatabaseConnection()
5.1: startDataGathering()
5.1.1: DatabaseHandler()
dbHander : DatabaseHandler
5.1.2: DataFilteringHandler()
dataFilterer : DataFilteringHandler
5.1.3: AccountHandler()
accountHandler : AccountHandler
5.1.4: run()
5.1.5: DataGatherer()
dataGatherer : DataGatherer
5.1.6: setSeedNode()
5.1.7: setTwitterInstance()
5.1.8: startGathering()
5.1.8.1: getDatabaseConnection()
5.1.8.2: saveUser()
5.1.8.3: saveRelation()
5.1.8.4: saveTweet()
5.1.8.4.1: pushFact()
5.1.8.4.1.1: saveFilteringResult()
6: stopGathering()
6.1: stopDataGathering()
6.1.1: stopGathering()

Figure 5. Sequence Diagram of the Twitter Data Collecting Tool

3.3.3. Graphical User Interface

We implemented GUI of the Twitter Data Collecting Tool as shown in Figure 6. It is designed for users to be able to enter filtering and gathering options easily. On the left side of the Basic Option tab, users can add or remove seed IDs by clicking the Add or Remove button. On the right side, users can add or remove Twitter developers' account information. In this menu, users can add multiple Twitter developers' account information to gather the data faster. Moreover, the Twitter Data Collecting tool also provides a keyword-filtering function. In the Filtering Option tab, users can add multiple filtering keywords. Based on the keywords that users entered, the Twitter Data Collecting tool generates rules to filter the gathered data using the keywords. At the bottom of the interface, three buttons are given to manipulate the tool, which are Start Analysis, Start Gathering and Stop Gathering buttons.

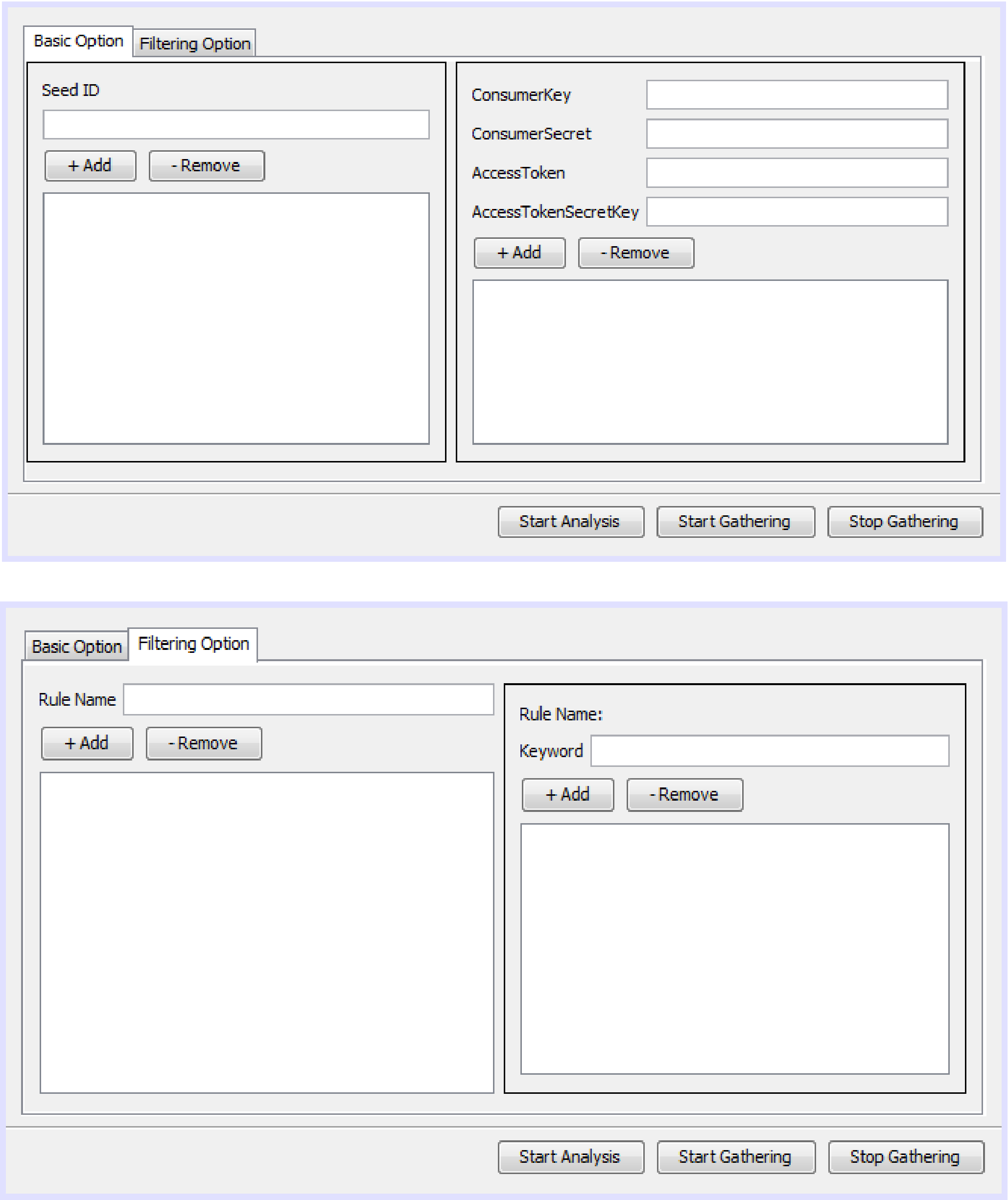

Figure 6. GUI of the Twitter Data Collecting Tool

Chapter 4 .

Seed Analysis in a Social Network

In this chapter, we present a method to maximize data gathering results using weighted seed selection. This chapter consists of three subsections. Section 4.1 shows architecture of the Twitter Data gathering tool with the Seed Analysis module, Section 4.2 describes how the algorithm quantifies activities of a node, and Section 4.3 presents an evaluation of the performance of building the Twitter dataset with the Twitter Data Collecting Tool with our influential node-selecting algorithm.

4.1. Architecture of Twitter Data Gathering Tool with Seed Analysis Module

We extended the existing architecture of the Twitter Data Collecting Tool described in Chapter 3. The given data collecting tool Figure 7 shows you the constituent modules of the Twitter Data Collecting Tool including the Seed Handler.

The novel module, the Seed Handler, is in charge of selecting an initial node and calculating each node's activities based on frequency of use of a certain keyword in its most recent postings. The Seed Handler module is running on two main algorithms that are the Initial Node Selecting Algorithm, and the Node's Activity Calculating Algorithm to satisfy its purpose of existence.

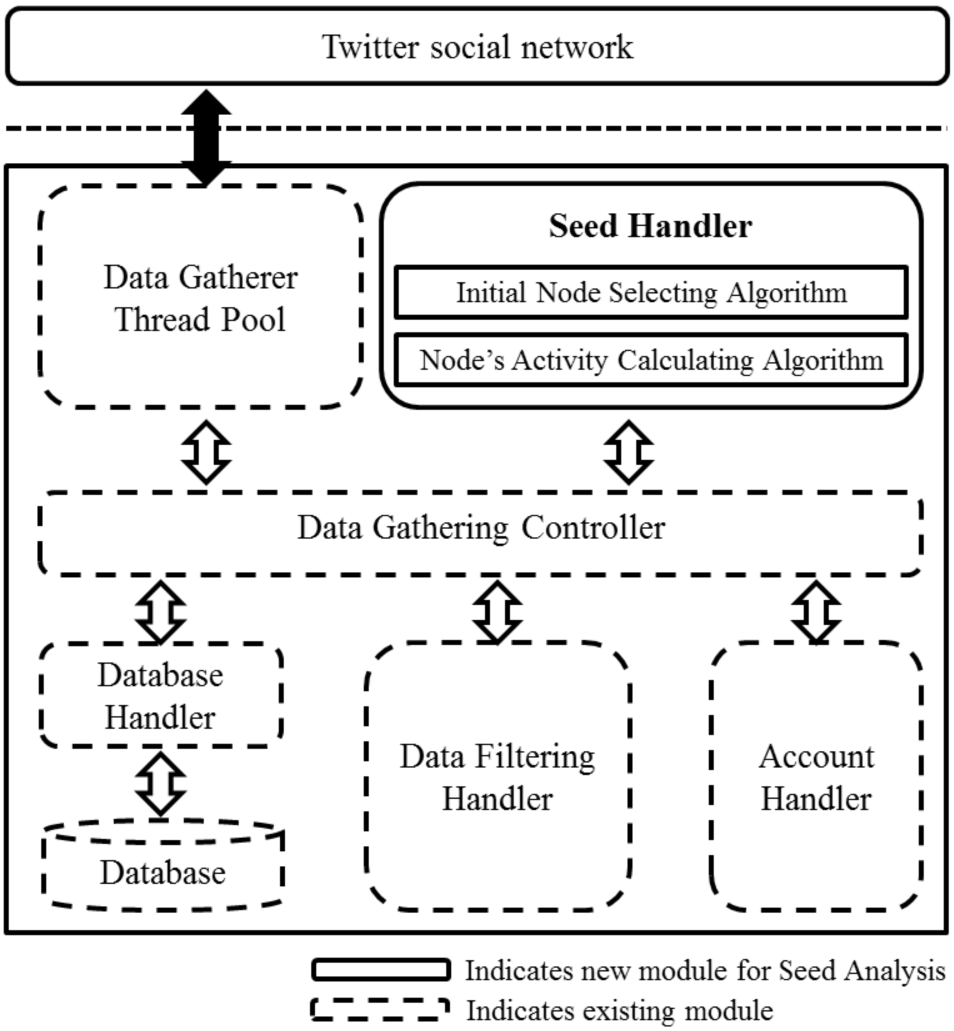
Twitter social network
Data Gatherer Thread Pool
Seed Handler
Initial Node Selecting Algorithm
Node's Activity Calculating Algorithm
Data Gathering Controller
Database Handler
Database
Data Filtering Handler
Account Handler
Indicates new module for Seed Analysis
Indicates existing module

Figure 7. Architecture of the Twitter Data Gathering Tool with Seed Analysis Module

4.2. Algorithm for Selecting Influence Nodes

For the most part, recent critical debates about social analysis have tended to center around the question of finding the most influential node in social network, also known as finding seed nodes. Social influence analysis aims either to verify the existence of social influence or to quantify the strength of the influence (Domingos and Richardson, 2001; Richardson and Domingos, 2002). Dynamic Social Network Analysis is about to model how friendships drift over time using a dynamic model or to investigate how different pre-processing decisions and different network forces such as selection and influence affect the modeling of dynamic network (Chen *et al*., 2009)**.**

These two related research areas in social network encouraged us to build an algorithm finding seed nodes by calculating node's activity in dynamic social network. Our approach of gathering keyword-related data more efficiently is collecting data from only qualified nodes. This goal can be achieved by giving activity weight to each node and checking if the node has enough activity weight before collecting tweets from the node.

Figure 8 shows the data gathering process from selecting an initial node through storing tweets into database. Once the tool is initiated, a list of candidates for an initial node is organized based on their most recent tweet, which contains a certain keyword at least once. When the list of candidates is built, an algorithm calculates activity weight of each candidate, and the list is organized by the number of followers of each candidate node. The first node in the list, which has highest the number of followers and is a qualified node, will be the initial node. If the first node in the list is not the qualified node, then check the next available node to see whether it is qualified. This process of finding a qualified node is iterated until an initial node is selected. Rebuilding a new list of candidates is required in case of the absence of a qualified node in the list. Once an initial node is selected, the tool generates a list of the initial node's follower information, such as each follower's unique id, language, number of followers, number of friends, etc. Then, each follower's activity weight is calculated, and only tweets from qualified users are collected, until there is no more follower information in the list.

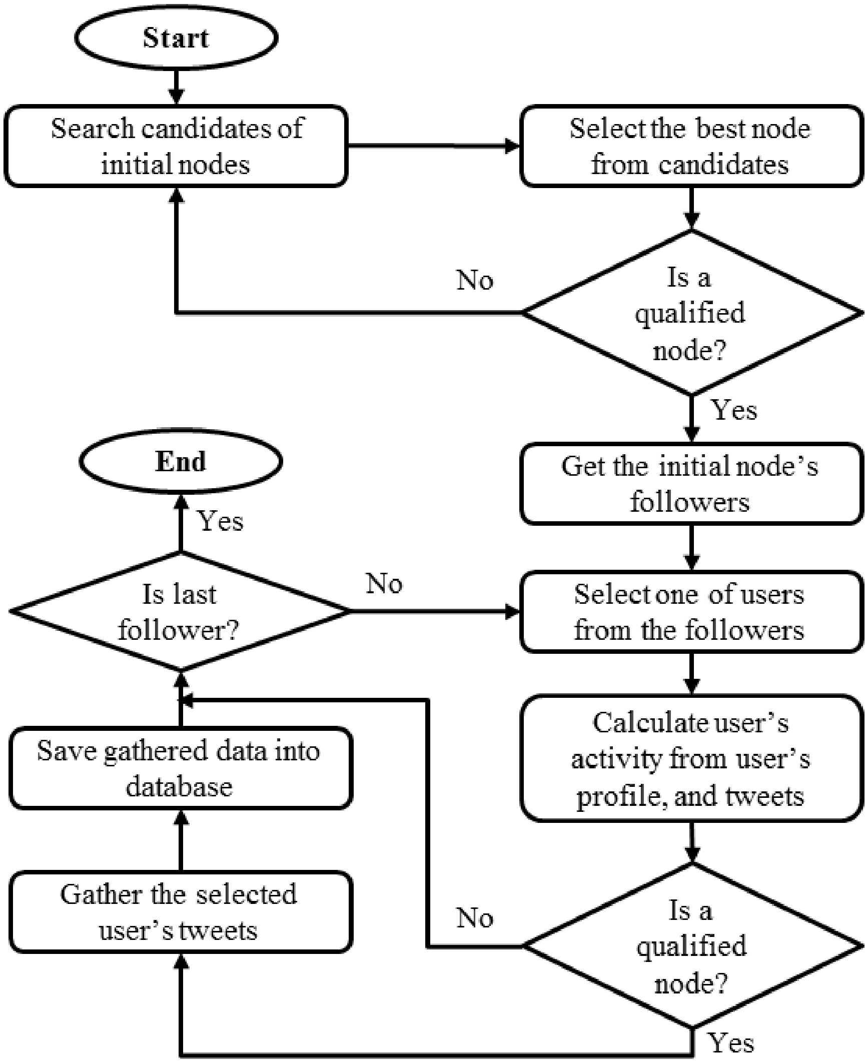
Start
Search candidates of initial nodes
Select the best node from candidates
Is a qualified node?
No
Yes
Get the initial node's followers
End
Yes
Is last follower?
No
Select one of users from the followers
Calculate user's activity from user's profile, and tweets
Save gathered data into database
Gather the selected user's tweets
No
Is a qualified node?
Yes

Figure 8. Flow Chart of Data Gathering Process by Qualified Nodes

Figure 9 shows the simple algorithm to find a node, which has more followers than others. If the selected node is not a qualified node, the tool removes the node from the list and runs the algorithm to find a suitable node again.

Notation: *A* is a set of nodes such that recently posted their tweets about a certain keyword. *B* is a pointer indicating the best initial nodes in the set *A*.

```
Set $B$=$A$[0].
Loop $I$ from 1 to $A$.size – 1.
    If $B$.no_of_followers < $A[I]$.no_of_followers, then
        Set $B$= $A[I]$.
    End If.
End Loop.
End of Algorithm.
```

Figure 9. Algorithm of Selecting Initial Node from a List of Candidates

In a user's profile, there are properties to be considered as factors of the user's activities in Twitter, such as number of followers, number of friends, number of keyword-related tweets, date tweeted, and favorite count. Among the user's properties, we use the number of followers, the number of keyword-related tweets, and date tweeted as main factors for calculating user's activity. Figure 10 shows the algorithm of calculating user's activity based on existence of a target topic in its tweet texts within a 30 day time period. If the target topic is not mentioned by a user within the time period, the user's tweets are not collected.

Notation: *T* is a set of tweets such that recently posted by a user within 30 days from search date and time. *K* is a string variable containing a keyword. *W* is a float variable containing user's activity value calculated by this algorithm. *M* indicates the number of tweets and *N* implies the number of tweets containing the keyword *W*.

```
Set $M$=$T$.size.
Set $N$=0.
Loop $I$ from 0 to $M$
    If the tweet ,T[$I$], contains the keyword K, then
        Set $N$=$N$+1.
    End If.
End Loop.
If $M$ is not 0, then.
    Set $W$=$N$/$M$.
End If.
End of Algorithm.
```

Figure 10. Algorithm of Calculating User's Activity Weight

4.3. Experiments of Weighted Seed Selection Algorithm

In this section, we present an evaluation of the performance of building the Twitter dataset with the Twitter Data Collecting Tool with our influential node-selecting algorithm on the real Twitter network. The following sub sections show an experimental test bed and the result of test.

4.3.1. An Experimental Test Bed

- **Keyword:** The presidential election of the United States of America was held at the end of 2012. There is no doubt that the presidential election is the most popular event in the U.S.A. The name of one of the candidates, Obama, has been chosen as a keyword to build a new dataset for a popular event to analyze user's behavior in Twitter about elections in the future.

- **Dataset:** Dataset will be built from real Twitter network in real-time. Two different types of data gathering approaches will be used. One approach is using the seed analysis algorithm we proposed in this paper, and another method is to start gathering process from an initial node manually selected by a data analysis specialist. Particularly, the specialist chooses possible seed candidates from Twitter accounts, such as BarackObama, MittRomney, VP, TheDemocrats, and etc. From the candidate list, one of Twitter account is selected for each attempt. For instance, Obama's Twitter account is picked as an initial node for first attempt because Twitter users following Obama tend to post tweets about Obama more than otherwise. Then, Target users are arbitrarily selected from all of Obama followers for each attempt to generalize result of data collection.

4.3.2. Experimental Results

Data gathering results by two different types of data gathering approaches are illustrated in TABLE 2 and TABLE 3. Each data gathering approach attempts three-times to see if the algorithm performs evenly with different sets of candidates of initial node. As shown in two table records, the average portion of keyword-related tweets in the dataset built by our approach is much larger than another approach (average of 9.83% keyword-related tweets in our approach compared to average of only 0.15% keyword-related tweets in the manual pick approach). In other words, this result means that gathering data from qualified seed node and follower nodes collects more keyword-related tweets than otherwise.

TABLE 2. Data gathering results from seed analysis algorithm

Attempts	**Total number of tweets**	**The number of keyword-related tweets (%)**	**The number of no keyword-related tweets (%)**
1	14,149	1,906 (13.47%)	12,243 (86.53%)
2	12,045	758 (6.29%)	11,287 (93.71%)
3	11,615	1109 (9.55%)	10,506 (90.45%)
Average	**12,603**	**1,258 (9.98%)**	**11,345 (90.01%)**

TABLE 3. Data gathering results from selecting an initial node by specialist

Attempts	**Total number of tweets**	**The number of keyword-related tweets (%)**	**The number of no keyword-related tweets (%)**
1	11,847	13 (0.11%)	11,834 (99.89%)
2	11,666	17 (0.15%)	11,649 (99.85%)
3	14,082	27 (0.19%)	14,055 (99.81%)
Average	**12,532**	**19 (0.15%)**	**12,513 (99.85%)**

To perceive the statistical significance of differences between data gathering results from two different algorithms, we applied the Chi-square test to the data gathered by two different data gathering approaches and derived the results, as shown in the TABLE 4 and Figure 11. The result of the Chi-square test shows that there is less than 0.001% chance that this deviation is due to chance alone. This implies that the algorithm we developed gathers significantly more data than gathering data from an initial seed selected by a specialist.

TABLE 4. Result of Chi-square test to data gathered by two different data gathering approaches

Approach	Keyword-related		Not-keyword-related	
	Freq.	*%*	*Freq.*	*%*
Seed Analysis	1,258	98.5%	11,345	50.1%
Selected by Specialist	19	1.5%	12,513	49.9%
Total	**1,277**	**100%**	**23,858**	**100%**

[a.] $\chi^2 = 1259.12$; Degree of freedom = 1; Probability < 0.001

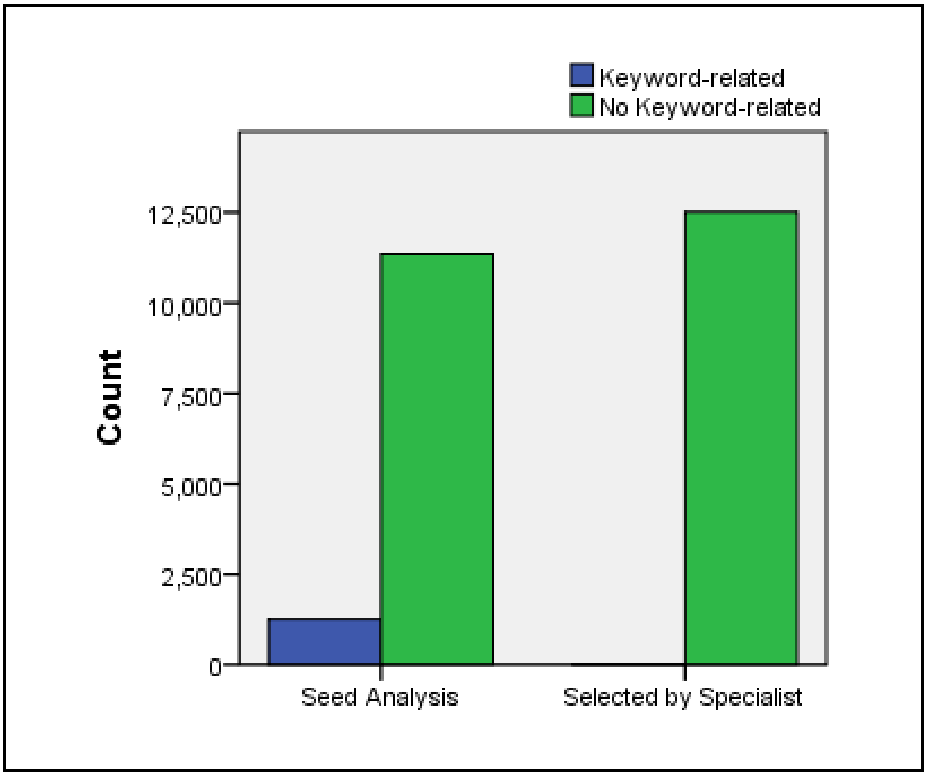

Figure 11. Graph visualization of the result of Chi-square test

Organizing user nodes by each node's activity weight allows us to discover the number of *k* seed nodes. TABLE 5 illustrates top 5 seed nodes from data gathering result using the seed analysis approach. To protect user's privacy, first four characters of each user's screen name are shown. During data collecting process, the ordered list of the nodes can be dynamically changed due to the dynamic seed selecting algorithm.

TABLE 5. Seed nodes discovered by our algorithm on dynamic twitter network

Number	Screen name	Follower Count	Friend Count	Activity Weight
1	Noma*******	13604	6345	0.99
2	Kath*******	2937	2972	0.36
3	Rock*******	7638	5236	0.36
4	Want********	13475	13364	0.33
5	Boud*******	1051	1175	0.32

Chapter 5 .

Case Study

5.1. Super Bowl 2012 Advertising in the Twitter

This exploratory study aims to address the question of how people use Twitter and to assess the power of Twitter in terms of creating consumer interest in brands and commercials. The following research questions will be addressed:

- To what extent do Twitter users get engaged in spreading Super Bowl commercials? How quickly or slowly are the conversations about Super Bowl advertising developed?
- To what extent do Twitter users have network relationships with advertised brands?
- How positive/negative are tweets about Super Bowl advertising? Do we see more positive tweets or negative ones? How do they differ by advertised brands?

This study aims to answer these questions by examining car commercials/brands advertised during the Super Bowl game in 2012. This section is interdisciplinary and unique in a sense that it uses a data mining approach to retrieve and analyze massive amounts of tweets exchanged about Super Bowl advertising from January 30 to February 11, 2012. The trends of Twitter traffic development over a study period, users' relationships with others, and the nature of texts exchanged for a certain brand or commercial are examined.

5.1.1. Data Gathering Plan

We planned to analyze the effects of Super Bowl 2012 commercials on the preference of car manufacturers that advertised their commercials during the event. To gather the data about car companies and Super Bowl commercials, we decided to gather data from January 30 to February 15, which includes one week before and two weeks after the Super Bowl game. This is necessary to track the trends of messages about Super Bowl advertising since marketers released their commercials to social media sites (e.g., YouTube) prior to the actual broadcast of the game, hoping to create more buzz and interest from consumers. Also, we found the eleven car-related companies.

Tweets related to Super Bowl commercials will be retrieved, filtered, analyzed, and visualized using data-mining tools. Since the amount of tweets on Super Bowl games and commercials is too huge, this study will only analyze the data related to car commercials. During the 2012 Super Bowl game, eleven car brands were promoted, making up the most promoted category. Therefore, this study will focus on tweets that are related to the Super Bowl and car categories.

5.1.2. Data Gathering Execution

Twitter data will be collected and filtered through 8 steps: selecting a seed, finding a Twitter account for each selected seed, gathering all followers' IDs for each seed node, picking random followers for each seed node, gathering tweets from picked followers, gathering profile information of each picked follower, saving all retrieved data into a local database, and filtering tweets with an imbedded rule-based engine. The detailed data handling process is as follows:

- Select seeds: As we mentioned in the introduction, 11 car-related companies were selected—Volkswagen, Toyota, Kia, Hyundai, Honda, Chrysler, Cadillac, Acura, Lexus, Chevrolet, and Audi. Table 6 is a list of car companies and the commercial titles that they broadcast during the 2012 Super Bowl Game.

TABLE 6. List of Car-related Commercials and Companies on Super-Bowl 2012

Commercial name	Brand/Company name
National-Transactions	Acura
Vampire Party	Audi
Green Hell	Cadillac
Silverado 2012	Chevrolet
Happy grad	Chevrolet
Projections/Brand	Chevrolet
Halftime in America	Chrysler
Broderick’s day off	Honda
Think Fast	Hyundai
Cheetah	Hyundai
Optima	Kia
GS Beast	Lexus
Connections	Toyota
It’s Reinvented	Toyota
The Dog Strikes Back	Volkswagen

- Find Twitter IDs for all selected seeds: To start gathering data from selected seed nodes, Twitter IDs for each seed node were identified. Twitter API provided a method to lookup a user’s ID using a user’s screen name or first or last name.

- Retrieve all followers' IDs for seed nodes: Using the Twitter data-collecting tool, find and retrieved all followers' IDs for seed nodes.

- Pick 1,500 followers randomly for each seed: We picked followers randomly to generalize the case.

- Gather tweets of randomly picked followers, posted between Jan 01 2012 and April 29 2012: The Super Bowl was held on Feb 05 2012. However, gathering data from before the game created a baseline to compare before and after findings. To study maintainability of Super Bowl commercial effects, data after the big game might be required.

- Gather personal information of picked followers: To determine a user's preference, the user's personal information was collected. If the user allowed the developer to gather their personal information, we can gather the user's name, location, number of followers, number of followings, and even date account created.

- Save all gathered data into a database: We saved all data gathered into database we designed in Section 3.2.5.

- Filter tweets using an embedded rule engine: We filtered tweets using Super Bowl Commercials related keywords and Drools, which are embedded in the Twitter Data Collecting Tool. Also, we filtered the tweets again using positive and negative keywords to rank the 11 car-related companies.

5.1.3. Results of Data Collection from Twitter

Table 7 shows the number of data for each category and the total number of data we gathered using the Twitter data-collecting tool we developed. As we selected 11 car-related companies, 11 rows of seed nodes are gathered and stored in the database. About 1.2 million following relationships between seeds and user nodes are gathered, and all users' 1 million user IDs are collected without data duplication. From 1,500 selected users for each seed nodes, 16,500 selected users in total, about 17 million tweets were retrieved from Twitter. Including 16,500 selected users, 88,121 of users had their information stored in the database.

TABLE 7. Data Collected from Twitter during the Study Period

Category	All data	Super Bowl related	Car-commercial related
The number of seed nodes (company)	11	-	-
The number of following relationships between seed nodes and followers	1,208,484	-	-
The number of followers that follow the seed nodes	942,024	-	-
The number of tweets that the selected users wrote.	16,931,580	25,492	4,970
The number of personal information of selected users	88,121	-	-
Total	19,170,220	25,492	4,970

5.1.4. Data Filtering and Analysis

Instead of using a traditional content analysis, this study used data-mining techniques and tools that allowed us to analyze massive amounts of messages about cars to derive useful knowledge.

TABLE 8. Company Name and Keywords to Filter Tweets about the Commercials

Company name	Keyword
Chevrolet	*Chevrolet*, *Chevy *
Chrysler	*Chrysler*, *Clientwood*
Hyundai	*Hyundai*, *Veloster*, *Genesis C*
Audi	Audi *, * Audi *
Acura	*Acura*, *NSX*
Toyota	*Toyota*, *Camry*
Lexus	*Lexus*
Cadillac	*Cadillac*
Kia	*Kia*, *Optima*
Honda	*Honda*, *CR-V*, *Matthew Broderick*
Volkswagen	*Volkswagen*, *new beetle*

In the first phase, we analyzed the relevance between tweets about the car-related companies and Super Bowl commercials. First, we filtered tweets that are related to Super Bowl car commercials using keywords and rules. The keywords are based on the name of companies that aired commercials during the Super Bowl and actors' and car names from the commercials. This is because users mostly mentioned the companies, cars, actors, or characters from the commercials in their tweets. For this reason, we made a rule that if a tweet includes at least one of the keywords, it is a tweet about the company. Table 8 shows the keywords for each company.

Based on the number of tweets filtered by keywords and the embedded rule engine, we visualized the tweets using WEKA (Hall el al., 2009), as shown in Figure 12. However, there are a lot of noisy data that are unrelated to the Super Bowl commercials in the data. For that reason, we filtered the data again using the rule engine and Super Bowl Commercial related keywords, such as Super Bowl, commercial, and ads. The filtered data is visualized again as shown in Figure 13. The number of tweets that are related to car companies and their commercials during the Super Bowl is higher than other times, which means the Super Bowl commercials created buzz on Twitter, and many Twitter users were interested in them. In addition, car-related tweets also significantly increased, accounting for 14.7% of total tweets that appeared over the study period (see TABLE 8).

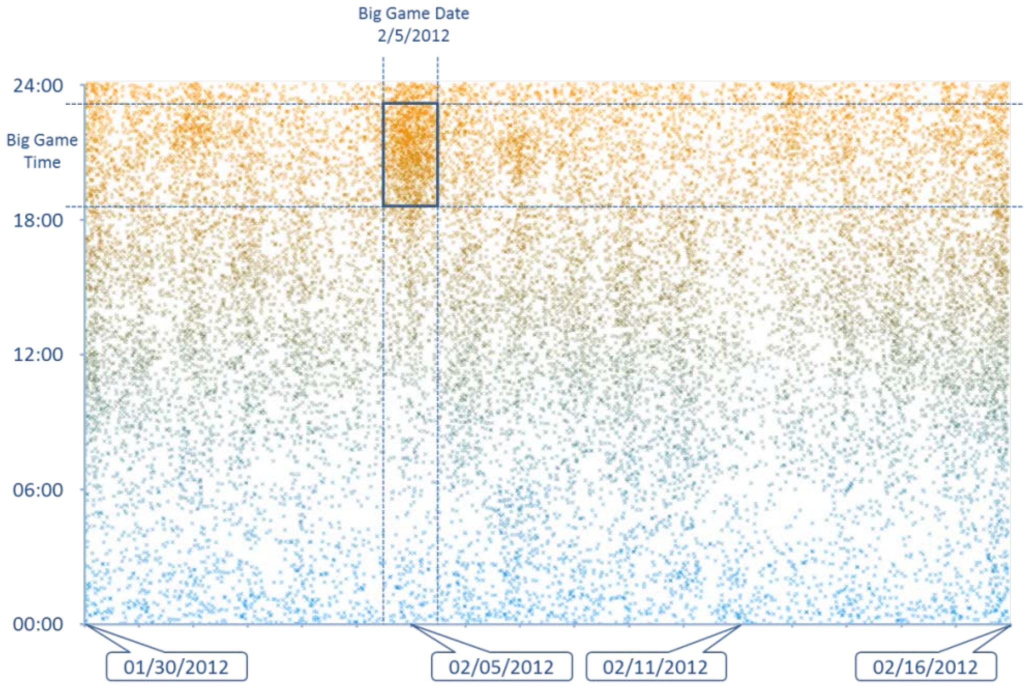

Figure 12. The Increased Number of Tweets during the Super Bowl before Filtering

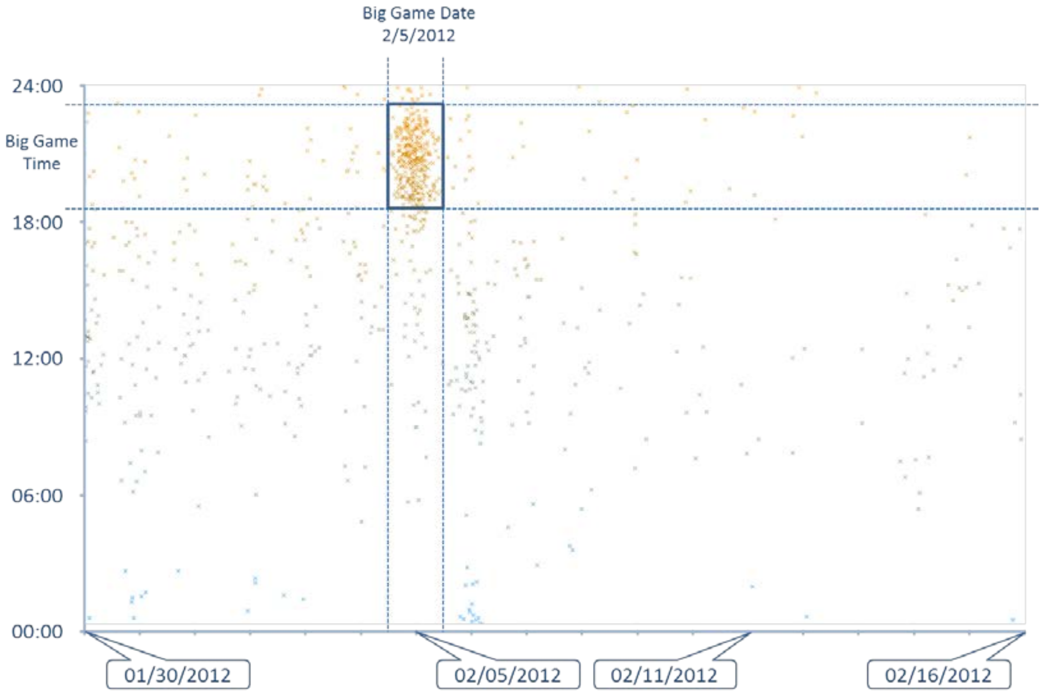

Figure 13. The Increased Number of Tweets during the Super Bowl after Filtering

TABLE 9. Frequencies of Tweets about Car-related and Non-Car-related Tweets by Date

Date	**Car-related (n = 28804)**		**Non-car-related (n = 938487)**	
	Freq.	%	Freq.	%
Jan/30/2012	1,900	6.6	59,184	6.3
Jan/31/2012	1,820	6.3	62,382	6.6
Feb/01/2012	2,215	7.7	63,841	6.8
Feb/02/2012	1,943	6.7	63,694	6.8
Feb/03/2012	1,855	6.7	60,660	6.5
Feb/04/2012	1,010	3.5	48,564	5.1
Feb/05/2012	4,231	14.7	71,145	7.8
Feb/06/2012	2,260	7.8	61,167	6.6
Feb/07/2012	1,870	6.5	64,139	6.8
Feb/08/2012	2,185	7.6	66,899	7.1
Feb/09/2012	2,141	7.4	64,590	6.9
Feb/10/2012	1,833	6.4	63,415	6.7
Feb/11/2012	846	2.9	55,120	5.8
Feb/12/2012	824	3.2	67,657	7.1
Feb/13/2012	1,711	6.1	66,030	7.0
Total	28,804	100	938,487	100

$X^2 = 3{,}190.10$; $df = 14$; $p < .0001$

Figure 14 depicts the differences between the number of tweets related to the Super Bowl (“Super Bowl tweets”) and the number of tweets not related to the Super Bowl (“non-Super Bowl tweets”). In this chart, each bar indicates the number of tweets posted on each day between Jan 30th to Feb 13th. Twitter users posted tweets about the Super

Bowl more on game day than any other day. When we compared the number of tweets on game day (Feb. 5) by day parts, the time from 5 p.m. to 11 p.m. marked the highest tweet posts. This implies that people tweeted as they watched the game.

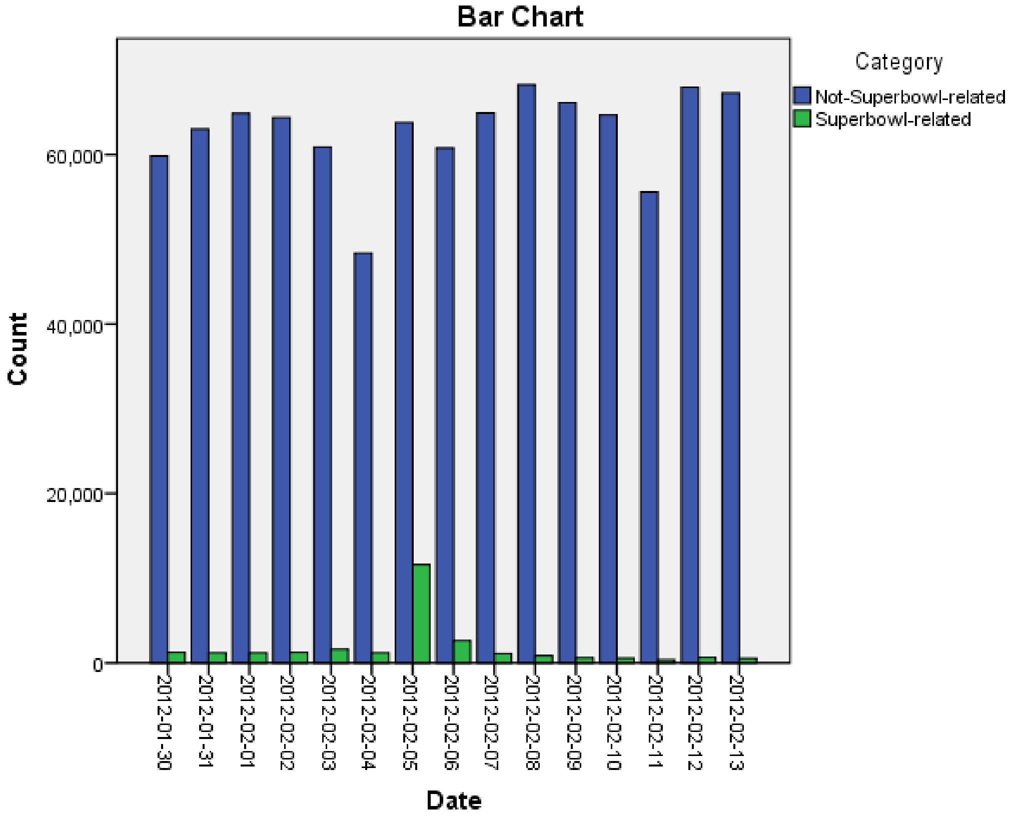

Figure 14. Comparison between "Super Bowl Tweets" and "Non-Super Bowl Tweets"

In the second phase, we inferred the meaning of the tweets about Super Bowl commercial using the rule engine. For this experiment, we retrieved tweets about Super Bowl 2012 commercials as we did in the first phase. Then, we derived users' sentiments by analyzing the tweets. User's sentiments are categorized into four different groups, which are positive, negative, positive-negative-mixed, and neutral sentiments. If a user's tweet contains positive words only or contains both positive and negative words, but has more positive words than negative words, then the tweet is categorized into the positive group. The concept of categorizing negative tweets is the same as categorizing positive tweets, but it focuses on the number of negative keywords in each tweet instead. Occasionally, users post tweets with both positive and negative words equally. In this case, if a tweet includes both positive and negative words and the amount of each sentiment is equal, the tweet is characterized as a positive-negative-mixed tweet. Otherwise, if there is no word reflecting a user's positive or negative opinion in a tweet, it is classified as a neutral tweet.

The categorizing user's sentiments process is implemented in the rule-based module as a rule. Figure 15 shows a structure of the rule and partial positive and negative words we used. Among 6,457 words, 2,304 words annotated as positive and 4,153 as negative, used in the OpinionFinder subjectivity lexicon, due to our specific dataset about Super Bowl commercials, we selected 72 positive words, such as "Amazing," "Awesome," etc., and 52 negative words are selected for the negative group, such as "Hate," "Horrible," etc. The rule, categorizing user's sentiments, consists of two sub-rules: One is counting positive words in a tweet and another one is counting negative words in the tweet. Once a tweet is forwarded to this counting words rule, both the number of positive words and the number of negative words are contended and stored in the tweet object. Particularly, "not" word is considered as a negative word to handle negation words. However, "another" word contains "not" text so an optional rule called "Remove Another" is defined to support this exceptional case. Only partial cases of negation words are filtered by this approach. The rule can be extended to improve accuracy of handling negation words.

```
rule "Positive"
   when
      t:Tweet (tweet matches ".*[a,A]mazing.*") or
                     …
      t:Tweet (tweet matches "[t,T]op.*") or
      t:Tweet (tweet matches "hot")
   then
      t.setPositiveValue(t.getPositiveValue()+1);
end

rule "Negative"
   when
      t:Tweet (tweet matches ".*[a,A]mbiguous.*") or
                     …
      t:Tweet (tweet matches ".*[n,N]ot.*") or
      t:Tweet (tweet matches ".*[p,P]oor.*") or
      t:Tweet (tweet matches ".*[t,T]errible.*") or
   then
      t.setNegativeValue(t.getNegativeValue()+1);

end

rule "Remove Another"
   when
      t:Tweet (tweet matches ".*[a,A]nother.*")
   then
      t.setNegativeValue(t.getNegativeValue()-1);
      t.setPositiveValue(t.getPositiveValue()-1);
end
                     …
```

Figure 15. A Rule for Categorizing User's Sentiments

To ensure the reliability of the algorithm, we investigated the accuracy of it by comparing the data set results categorized by a human and the other data set results categorized by our algorithm. Fifteen hundred tweets are randomly selected from the whole dataset and are coded by two coders who are not involved in the classification algorithm process. After these 1,500 tweets are coded into the four categories, the result of the classification by our algorithm is compared with that of human analysis to see the accuracy of our classification algorithm. Table 10 shows the accuracies of our rule-based algorithm compared to categorization result by human. The results indicate that the accuracy of the rule-based approach has an acceptable rate (more than 80%). This means that our sentiment categorization algorithm is reliable for analyzing user's sentiments in their tweet especially tweets about Super Bowl advertisement.

TABLE 10. Accuracies of the Rule-based Sentiment Categorization Algorithm Compared to Categorization Results of Human Analysis

Category	Positive	Negative	Neutral	Mixed	Total
Number of tweets categorized by rule-based algorithm	452	75	941	32	1,500
Number of tweets categorized by human	470	70	941	19	1,500
Number of tweets matched with results of rule	357	25	819	6	1,207 /1,500
Number of tweets unmatched with results of rule	113	45	122	13	293 /1,500
Accuracy (Matched)	**76%**	**36%**	**87%**	**32%**	**80.47%**

Figure 16 shows percentages of users' sentiments of each car-related company. In all cases, half of the tweets about each car-related company do not contain user's sentiments about their commercial. As well as, tweets containing positive sentiments about car commercials are posted on Twitter more than tweets containing negative sentiments.

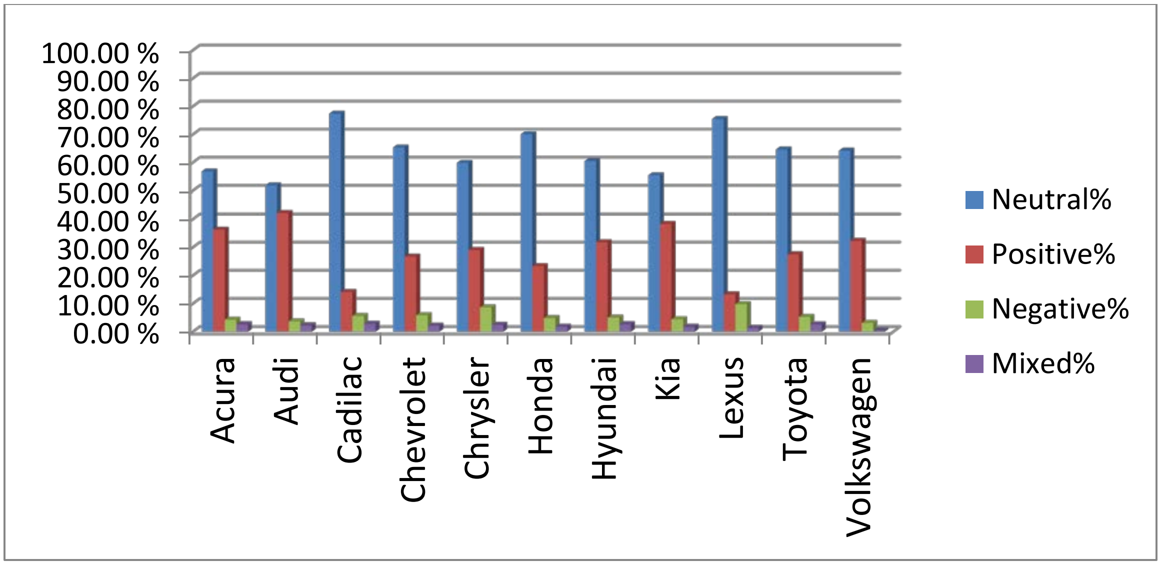

Figure 16. Percentages of Twitter Users' Sentiments by Car-related Company

Figure 17 shows followers-groups which are categorized into following-only group and following-with-others group for each company respectively. The former groups are marked as dark color and the latter groups are marked as light color. According to the results, we can assume that the seeds that have more following-only users than following-with-others users, such as Audi, Chevrolet and Lexus, have higher following "loyal" users. On the other hand, we also can assume that the seeds that have more following-with-others users than following-only users, such as Acura, Cadillac and Hyundai, have lower following "loyal" users.

To perceive statistical significance of differences, we applied the Chi-square test to the followers-groups' data and derive the results as shown in the Table 11 and Figure 18.

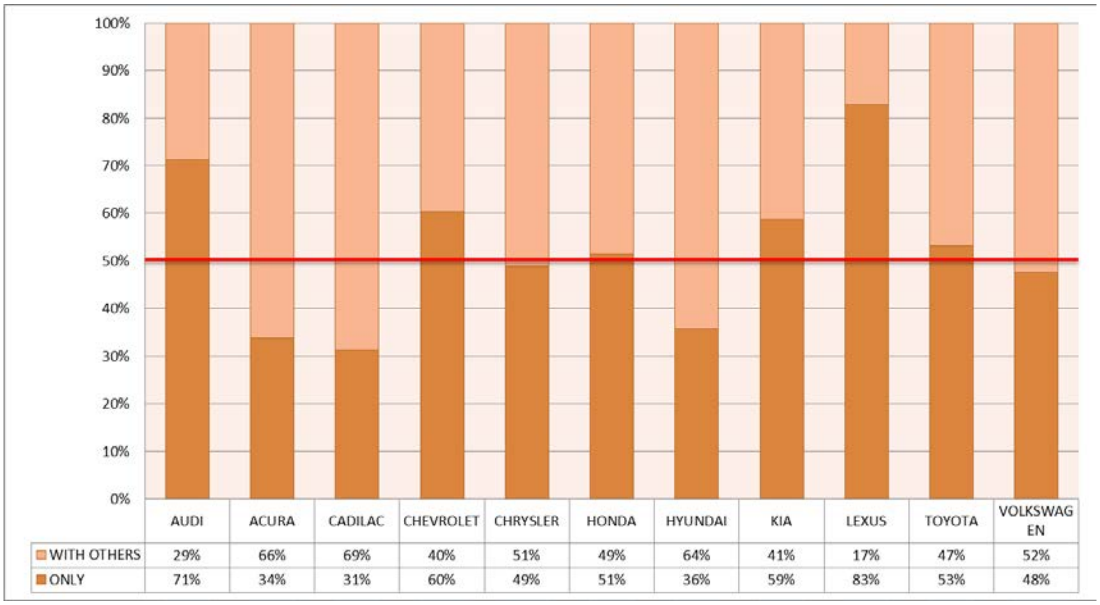

Figure 17. Following-Only Group and Following-with-Others Group

TABLE 11. Frequencies of Following-Only Group and Following-with-Others Group

Date	**Following-Only (n = 575)**		**Follow with Other (n = 535)**	
	Freq.	%	Freq.	%
Audi	71	12.3	29	12.3
Acura	34	5.9	66	5.4
Cadillac	31	5.4	69	12.9
Chevrolet	60	10.4	40	7.5
Chrysler	49	8.5	51	9.5
Honda	51	8.9	59	11.0
Hyundai	36	6.3	64	12.0
Kia	59	10.3	41	7.7
Lexus	83	14.4	17	3.2
Toyota	53	9.2	47	8.8
Volkswagen	48	8.3	52	9.7
Total	**575**	**100**	**535**	**100**

$X^2 = 100.79$; $df = 10$; $p < .0001$

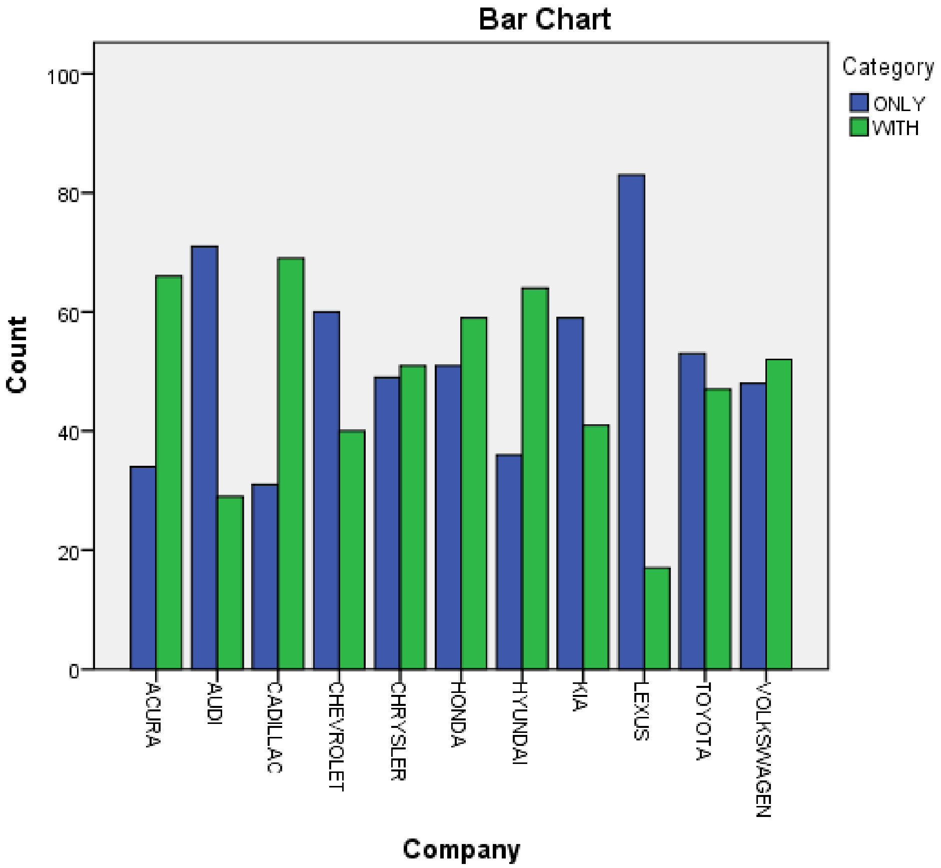

Figure 18. Comparison between “Following-Only Group” and “Following-with-Others Group”

In the last phase, we analyzed and visualized the relationship between the users and the car-related companies using the Fruchterman Reingold Algorithm (Fruchterman and Reingold, 1991) in Gephi (Bastian, 2009). Gephi is an open-source visualization tool for network analysis, dynamic and hierarchical graphs. In this case, Gephi was used to visualize the relationship between users and car companies, and to group similar users.

The Fruchterman Reingold Algorithm is a force-directed layout algorithm. This means that each node is a user and the length of an edge is decided by the number of tweets that the user wrote to seed nodes. Therefore, the more tweets a user wrote to seed a node, the closer the user node is to the seed node.

To visualize data as a graph, we converted our data into a Gephi data type and ran the algorithm. Figure 19 shows a visualized graph indicating relationships between seed nodes and user nodes. We grouped users by car companies. Each user group indicates a group that is interested in a car-related company.

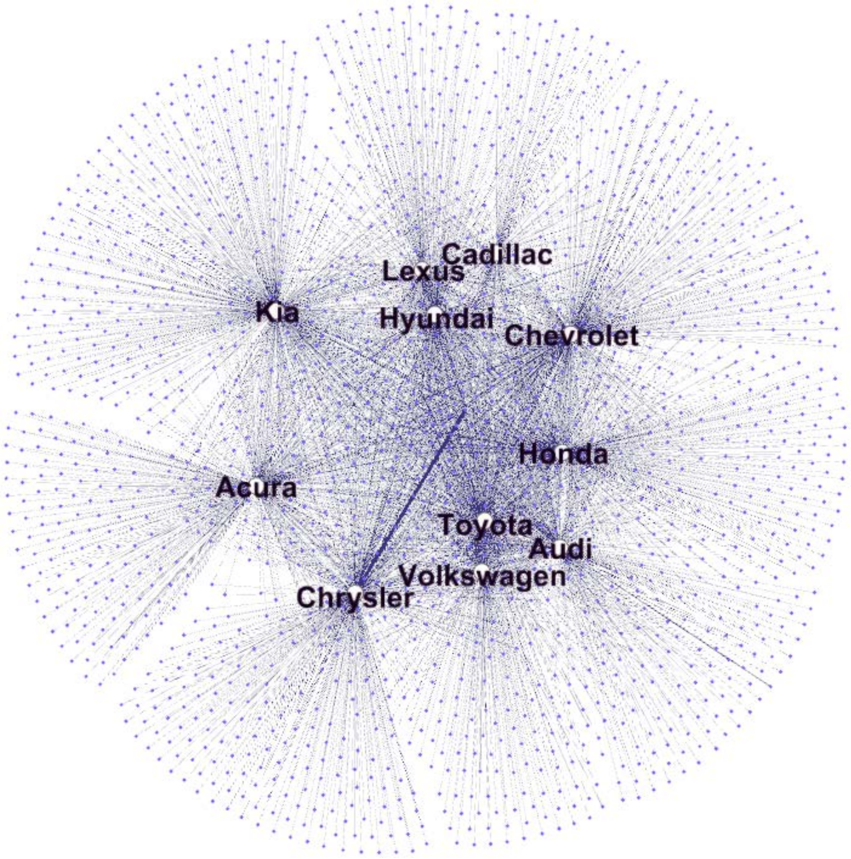

Figure 19. Grouping Results Graph

In addition, we derived relationships of the companies that compete with each other. The more inter-related users between 2 companies, the closer two seed nodes are. For example, the location of 4 seed nodes, Lexus, Audi, Cadillac, and Hyundai, are very close, as shown in Figure 19, because there are many common users who talked about the four brands. Based on these results, we can infer the relationships between the companies.

5.2. Super Bowl 2013

In 2013, 108.41 million viewers, a 3 million decrease from 2012’s viewership, tuned in to Super Bowl XLVII (“Super Bowl Ratings Decline,” 2013). According to Nielsen measures, the game earned an average overnight household rating of 46.3, meaning that 46.3% of households with TVs were watching this program, making it the second-highest-rated Super Bowl in 27 years, and still the biggest TV event of the year (“CBS Claims Record Super Bowl Ratings in Early Tally,” 2013). With such a large viewership, it is no wonder that the Super Bowl is also one of the most popular advertising venues, despite a hefty average cost of $3.5 million per 30-second spot. Highlighting the growing presence of the Super Bowl on social media, Advertising Age, a major advertising and marketing trade journal, partnered with Bluefin Labs, a Cambridge-based social TV analytics company, and reported in 2012 that more than 12.2 million social media comments were tracked during and after the game, primarily on Twitter and Facebook. Reportedly, this was a 578% increase from 2011, which had 1.8 million posts and exchanged comments (Dumenco, 2012).

The 2013 Super Bowl XLVII was considered one of the most exciting games, yielding a final score of 34-31 with 7 touchdowns, 6 field goals, and one safety. Particularly, this year’s Super Bowl game kept the audience’s attention to the last minute, making the time between 10:30 pm and 10:45 pm the most watched part of the game (“CBS Claims,” 2013). In addition, the game yielded several NFL records, such as the touchdown by a 109-yard kick-return (4th touchdown), a touchdown by a quarterback, and a 34-minute blackout due to a power outage, marked the dynamic nature of the game. Therefore, the

2013 Super Bowl game would provide good case material to examine multimedia experiences between television and Twitter.

5.2.1. Study Questions

Instead of relying on the audience's response, this study will use a data-mining approach to filter and analyze a massive amount of tweets and examine the following research questions:

- RQ1: How does the overall number of tweets differ between a game day and non-game day?
- RQ2: How does the overall number and usage of tweets differ in relation to the nature of the game, such as scoring moments and other dynamic moments? Do some scoring moments have more tweets? What is the relationship between non-scoring moments and the number of tweets at those times?
- RQ3: What major topics were exchanged through Twitter during the 2013 Super Bowl game?
- RQ4: How does the overall number of tweets differ by geographic location?

5.2.2. Main Method of Data Analysis

As survey, experiment, and content-analysis methods are conventional research methods in the mass communication field, the data-mining techniques are well-adopted in the computer science field and allow researchers to handle a huge amount of data and discover knowledge and information from those data sets. As the amount of tweets on Super Bowl games is enormous, it is not possible to retrieve, filter, analyze, and visualize them without automated-tools and well-defined approaches through these data mining techniques. For this reason, some data-mining techniques and computational data analyzing approaches, such as topic detection, were applied in this study.

5.2.3. Study Period

The Super Bowl XLVII was held on Sunday, Feb. 3, 2013. To compare the amount and patterns of tweets on this game day, we also chose two other Sundays (one week prior to the game and one week following the game), January 27 and February 10. A Strategies for Effective Tweeting report (2013) mentioned that Twitter engagement rates for brands are 17% higher on weekends compared to weekdays, meaning Saturday and Sunday typically generate much larger amount of tweets than weekdays. According to this statement, these two Sundays were chosen as references to compare to the Super Bowl game day.

5.2.4. Data Analysis

Over the three Sundays from 6:00 pm to 11:59 pm, a total of 305,369 (i.e., data population) tweets was identified. Out of them, 12.7% (a total of 38,714 tweets) were Super Bowl-related tweets (data sample).

TABLE 12. Total Number of Tweets collected from Twitter over Three Sundays from 6:00 pm to 11:59 pm

All Three Sundays		**01/27/2013**		**02/03/2013**		**02/10/2013**	
All	Super Bowl Related	All	Super Bowl Related	All	Super Bowl Related	All	Super Bowl Related
305,369	38,714 (12.7%)	84,149	4,200 (5.0%)	122,707	29,696 (24.2%)	98,513	4,818 (4.9%)

RQ 1 asked how the overall Twitter usage differed between a game day and non-game day. When comparing the number of tweets from each Sunday, Feb. 3 generated a larger number of tweets than both other Sundays in terms of total number of tweets, as well as, Super Bowl related tweets. The total number of tweets for Feb. 3 was 122,707, while only 84,149 on Jan. 27 and 98,513 on Feb. 10. As Table 12 shows, Super Bowl related tweets accounted for 24.2% of total tweets exchanged on Feb. 3, while they accounted for 5.0% and 4.9% on Jan. 27 and Feb. 10, respectively. This indicates that the overall tweet amount peaked on game day and the portion of Super Bowl related tweets increased as well. Figure 20 and Figure 21 show this finding visually in a minute interval by one minute span from 6:00 pm to 11:59 pm for each Sunday.

RQ 2 asked how the overall number and usage pattern of tweets differ in relation to the nature of the game. Figure 20 shows that there were some moments with higher Twitter usage than others over the course of the game. While there were little fluctuations in Twitter usage on Jan. 27 and Feb. 10, the overall Twitter usage on Feb. 3 was much dynamic.

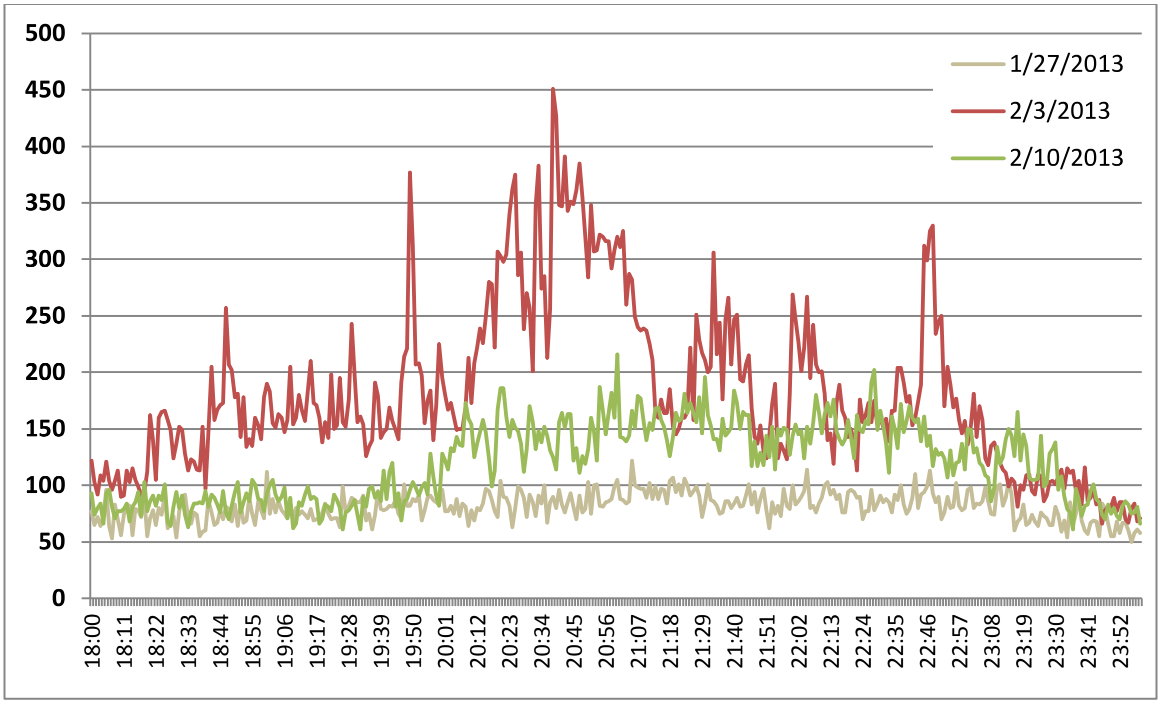

Figure 20. The Pattern of All Tweets exchanged on Three Sundays

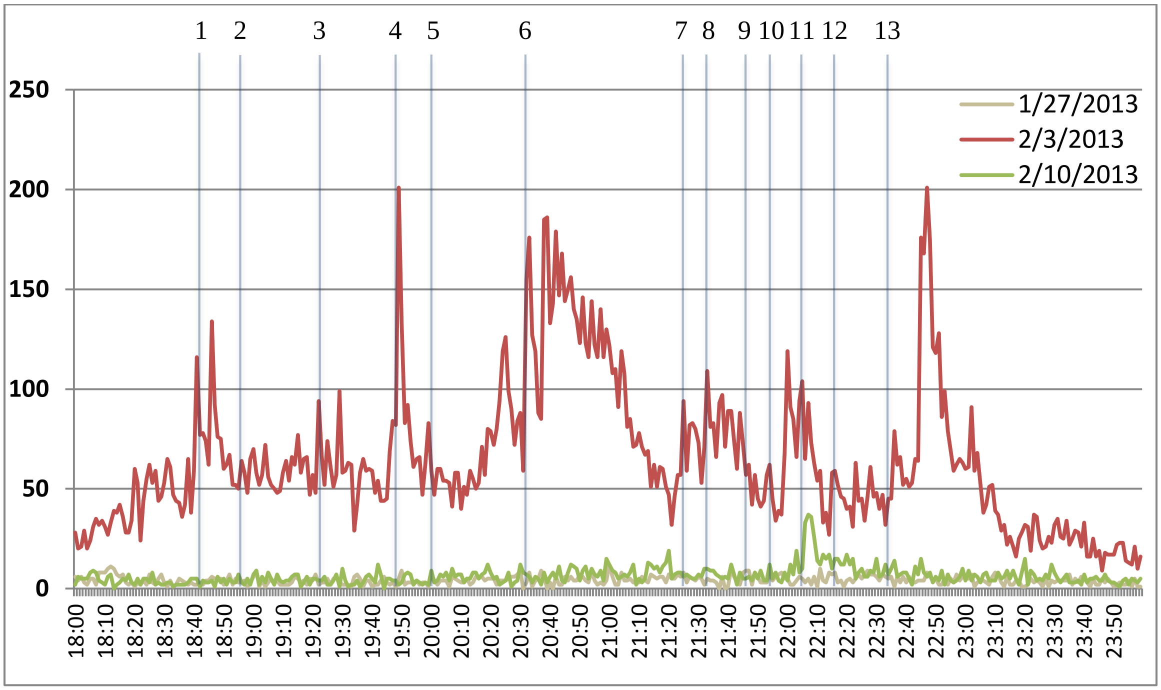

Figure 21. Number of Super Bowl-related Tweets exchanged on Three Sundays

TABLE 13. List of Scoring Moments of the 2013 Super-Bowl

#	Event	Score		Time line
		Ravens	49ers	
1	1st Touchdown	7	0	18:40
2	1st Field Goal	7	3	18:54
3	2nd Touchdown	14	3	19:21
4	3rd Touchdown	21	3	19:47
5	2nd Field Goal	21	6	19:58
6	4th Touchdown	28	6	20:31
7	5th Touchdown	28	13	21:24
8	6th Touchdown	28	20	21:32
9	3rd Field Goal	28	23	21:44
10	4th Field Goal	31	23	21:53
11	7th Touchdown	31	29	22:04
12	5th Field Goal	34	29	22:15
13	1st Safety	34	31	22:32

To understand these differences, we ran another analysis only by using Super Bowl related tweets. Figure 21 shows the pattern of Super Bowl related tweets during the Super Bowl broadcast, and the numbers from 1-13 on the top of Figure 21 marked each scoring moment, as listed in Table 13. As Figure 21 indicates, the number of tweets increased at every scoring moment. Specifically, it dramatically increased after the 3rd and 4th touchdowns. Additionally, this figure shows that the increase in tweets following touchdowns was higher than the increase in tweets following field goals.

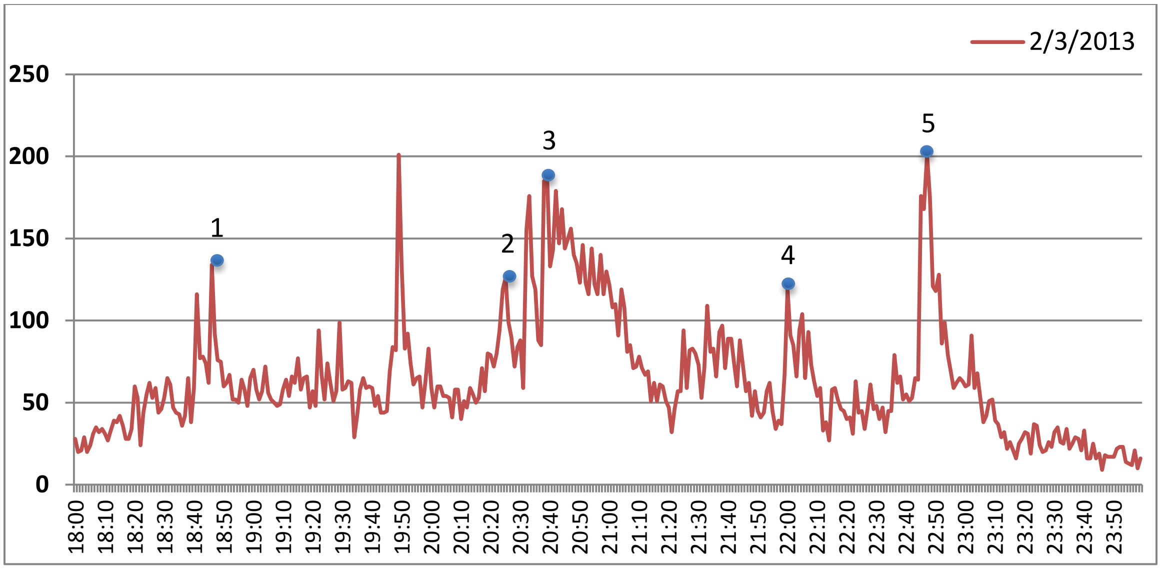

Figure 22. Number of Tweets that are not related to Scoring Moments on the 2013 Super Bowl Game

The analysis further revealed that the number of tweets increased with some events that were not related to scoring moments (see Figure 22). By looking at the times with increased Twitter activity, major non-scoring moments, such as a Super Dome power outage period, a half-time show, and commercial breaks, were found. To check a direct relationship between the increased time of Twitter usage and these non-scoring moments, we counted the number of all Super Bowl related tweets within 3 minutes after those events occurred. Table 14 provides a list of five major non-scoring events and the number of tweets exchanged within a 3-minute span. Our analysis further revealed that the amount of Twitter usage on specific moments has decreased after 3 minutes. On average, the event-related tweets accounted for 42% (A total 1,905 tweets were related to those five non-scoring events) of all tweets (4,515 tweets) exchanged over those five events.

TABLE 14. Five Moments that created Buzz on Twitter during the Game (by Total Number of Tweets)

#	Time line	Related Event	The number of all tweets within 3minutes after event*	The number of events related tweets within 3minutes after event*
1	18:46	Commercial break with Go Daddy, Doritos, and Audi spots	671	222 (33.1%)
2	20:25	Halftime Show	973	371 (38.1%)
3	20:38	Superdome Power Outage	1,234	490 (39.7%)
4	22:00	Commercial Break with Dodge “Farmers” spot	743	328 (44.1%)
5	22:47	The End of the Game	894	494 (55.2%)
Total Number of Tweets			4,515	1,905 (42.2%)

Among these five non-scoring moments, the end of the game and Super Dome power outage generated the most tweets, 494 and 490, respectively. When we examined the portion of each moment-related tweet, however, the end of the game (55.2%) and Dodge’s “Farmers″ commercial spot (44.1%) generated the most tweets, followed by the power outage (39.7%), half-time show (38.1%), and the commercial break featuring the “Go Daddy,” “Doritos,” and “Audi” spots (33.1%).

TABLE 15. List of Top 100 Frequently Used Words on Twitter on Game Day (Feb. 3)

#	word	count	#	word	count	#	word	count
1	superbowl	5885	35	need	642	68	ram	367
2	ravens	3571	36	play	624	69	xlvii	367
3	super	3274	37	way	615	70	looks	364
4	bowl	3167	38	down	612	71	won	361
5	Beyoncé	2796	39	half	612	72	godaddy	360
6	commercial	2380	40	baltimore	591	73	adbowl	338
7	game	2329	41	first	574	74	fans	334
8	ers	2163	42	day	555	75	san	333
9	power	1771	43	los	530	76	child	333
10	time	1517	44	year	527	77	jones	332
11	lights	1515	45	jeep	502	78	powertotweep	332
12	ad	1475	46	twitter	497	79	performance	331
13	love	1278	47	flacco	473	80	tweet	330
14	ray	1194	48	oreo	467	81	con	327
15	win	1070	49	farmer	454	82	del	326
16	que	1051	50	blackout	452	83	orleans	326
17	see	1008	51	team	412	84	joe	318
18	show	1008	52	spot	410	85	pepsi	318
19	best	1002	53	tonight	401	86	call	317
20	lewis	999	54	niners	399	87	run	315
21	know	934	55	please	397	88	today	312
22	think	919	56	ads	394	89	work	312
23	halftime	907	57	youtube	390	90	light	311
24	audi	778	58	video	386	91	any	311
25	football	771	59	touchdown	380	92	hope	311
26	brandbowl	765	60	dodge	380	93	baby	309
27	superdome	762	61	car	378	94	home	308
28	nfl	742	62	budweiser	377	95	thought	297
29	people	731	63	went	371	96	congrats	297
30	watch	708	64	wait	370	97	guy	297
31	outage	701	65	life	369	98	world	296
32	watching	699	66	guys	367	99	away	294
33	god	692	67	night	367	100	stadium	294
34	commercials	685						

RQ 3 asked to identify major topics on Twitter during the Super Bowl game. As explained in the methods section, this analysis was formed using a two-step process. First, the top 100 frequently used words were identified by using QB-Text Analyzer and were grouped into seven topic categories, which are Super Bowl, Commercials, Power Outage, Dodge Commercials, Miscellaneous, Mixed, and Other. Table 15 provides a list of the top 100 most frequently used words on all of Twitter. Super Bowl, Ravens, Super, Bowl and Beyoncé ranked top in the list, indicating 2013 Super Bowl created a buzz on Twitter, and frequently used words on Twitter in a specific time period are related to real time events or activities. Figure 23 shows the breakdown of tweets by topic. Among these seven topics, five were directly related to the Super Bowl game (e.g., Super Bowl, commercials, Dodge commercial, power outage, and mixed). The tweets belonging to these five topics accounted for 62% of all tweets. Among these five topics, the Super Bowl was the most discussed topic (35.7%), followed by the “mixed” category (15.4%, more than one topic discussed), the “commercials” (7.4%), the “power outage” (2.6%) and the “Dodge’s Farmers commercial” (1.1%). On the other hand, 23.6% was identified as “Miscellaneous” (i.e., non-Super Bowl game related tweets) and 14.3% was identified as “Other.”

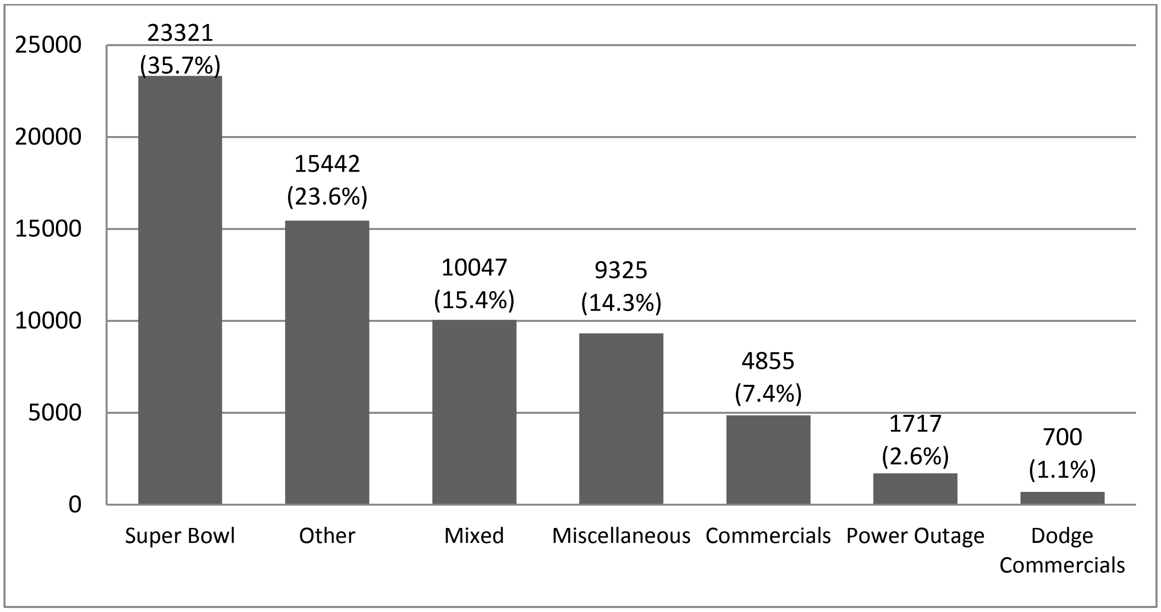

Figure 23. Number of Top 100 Frequently Used Words on Twitter by Topic

Table 16 shows sample tweets categorized by each of seven topics. For example, the following tweets are samples from the "Super Bowl" category.

- My #SuperBowl prediction: Coach Harbaugh will win.
- Let"s go 49ers! #sf #49ers #sanfrancisco49ers #questforsix #twitpics http://t.co/GIncT5qc
- In 7 hour we will know who will be Super Bowl champs! GO RAVENS

These are samples from the "commercials" category:

- RT @SaxonHolt33: Commercials better be better than last year
- @freep for the commercials. all the way. i don't even care about who is playing.
- Ready to see some car commercials! #BigGame
- What did you think of the @Audi "Bravery" spot? #SuperBowlAds

TABLE 16. Sample Tweets for Each Category (the 2013 Super Bowl on Feb. 3)

Category	Tweets Sample
Super Bowl	– 30 minutes to kickoff. We"re goin" in. http://t.co/CXpEBXo1 @SuperBowl XLVII – My #SuperBowl prediction: Coach Harbaugh will win. – Let"s go 49ers! #sf #49ers #sanfrancisco49ers #questforsix #twitpics http://t.co/GIncT5qc – In 7 hour we will know who will be Super Bowl champs! GO RAVENS – RT @MSN: Count down to the kickoff of Super Bowl XLVII & follow all of the action from New Orleans http://t.co/7IoEaB9i

Commercials	– RT @SaxonHolt33: Commercials better be better than last year – @freep for the commercials. all the way. i don"t even care about who is playing. – RT @FakeSportsCentr: Ray Lewis spotted in the locker room chopping onions and hiding them in his shoulder pads for the anthem – Ready to see some car commercials! #BigGame – What did you think of the @Audi "Bravery" spot? #SuperBowlAds
Dodge	– KathyatChrysler So proud to be part of the @chrysler @Jeep @Dodge @RamTrucks @driveSRT @OfficialMOPAR and @FIATUSA family. #LoveMySRT "s – And so god made a farmer... – God made a farmer.. #myfather – #sogodmadeafarmer commercial wins hands down. Go #Idaho #dodgeram – The @Ram commercial of Paul Harvey"s "Sp God Made a Farmer". #YES
Other	– @EmblemThree #E3AnnouncementTomorrow make my life and follow me please 40 – @UCFFootball represented on both sidelines tonight. Go Knights! – Twitter Chat: How SMBs Can Take Advantage of Consumer Holidays: Holidays are terrific opportunities to call atte... http://t.co/FVUQ9eMV – https://t.co/OQDT8EZy@brianlbos @Conlonke the video not loading on my iPad. Working for u? – The Home Depot Twitter Party, Thursday, Feb 7th, 8-9pm ET #DIHWorkshop http://t.co/f2psuqoV via @aboutamom
Power outage	– RT @AP: Lights out: Power outage stops game at #SuperBowl early in third quarter: http://t.co/jIou1OHK #NFL #SB47 -RD – RT @AFP: #BREAKING: Partial power outage dims lights at Super Bowl #NFL – I"m bewildered about the current power outage. #FutureTurnOnTheLights – RT @BestBuy: Looks like #SB47 could use a visit from some @GeekSquad Agents with @Insignia LED light bulbs for the power outage. – #Baltimore #Ravens Stadium Power Outage #SB47 Soaps http://t.co/0ibNLmE1 #TurnOnTheLights

Miscellaneous	– @hello_sarahx0 bring tissues – @kkinnggg meh, the good die young! – How to Stencil on Valentines Cookies http://t.co/9URYHrwV http://t.co/araju7ig – Staying warm in style thanks to the crew-neck sweatshirt by West Mid! http://t.co/b76LJGlv – Ah yeah here we go!
Mixed	– In case you missed it... Dodge: "Farmer" Spot http://t.co/znYE88Mi via @wsj #brandbowl – RT @PimpBillClinton: Let''s hope the lights go out again so we don''t have to watch Ray Lewis dance like a ferret''s trying to eat his dick ... – Enough with the godaddy ad already. For anyone who''s not into watching PDA - the audio just tops off the nastiness. – RT @TheLisaMcGrath: Always nice to see R guys represented in the Super Bowl. Congrats to @RayRice27 & @asilvestro45 on your first Su ... – @RachYoungggg I would hope that we''d be closed on Martin Luther King, Jr.''s b-day before Super Bowl Monday. Why not a Saturday game?

As the 2013 Super Bowl game marked an unprecedented power outage for about 30 minutes, the tweets exchanged during this time were analyzed. Table 17 lists Twitter’s top 100 frequently used words exchanged between 8:40 pm and 9:10 pm EST. and Figure 24 shows the portion of those tweets by each topic.

TABLE 17. Twitter's Top 100 Frequently Used Words during the power outage of 2013 Super bowl (8:40 pm to 9:10 pm)

#	word	count	#	word	count	#	word	count
1	superbowl	1175	35	bill	124	68	superbowlblackout	68
2	power	1037	36	lightsout	119	69	commercials	67
3	lights	922	37	half	118	70	happen	64
4	super	649	38	people	116	71	thought	64
5	beyonce	593	39	dark	115	72	powertotweep	64
6	bowl	583	40	see	114	73	make	62
7	superdome	531	41	pay	112	74	lost	62
8	game	341	42	problem	108	75	shut	60
9	outage	328	43	halftime	106	76	turned	60
10	right	312	44	commercial	106	77	guy	59
11	time	267	45	way	99	78	jeep	59
12	ers	247	46	wings	96	79	tweet	58
13	audi	234	47	want	95	80	fans	58
14	ray	226	48	football	92	81	night	58
15	ravens	214	49	she	90	82	big	58
16	lewis	194	50	buffalo	88	83	blew	57
17	orleans	190	51	luz	87	84	wait	57
18	blackout	186	52	los	87	85	god	56
19	stadium	182	53	watch	87	86	illuminati	56
20	light	173	54	ad	87	87	killed	56
21	nfl	172	55	twitter	87	88	watching	55
22	show	165	56	minutes	83	89	electricity	55
23	think	154	57	field	82	90	looks	54
24	play	154	58	beyonces	79	91	switch	54
25	LEDs	143	59	cbs	76	92	days	53
26	sending	143	60	win	76	93	keep	53
27	went	143	61	man	74	94	players	53
28	know	140	62	need	74	95	breaking	52
29	dome	140	63	poweroutage	74	96	electric	52
30	bane	130	64	performance	71	97	nola	52
31	mbusa	128	65	say	71	98	del	52
32	oreo	127	66	blame	70	99	used	51
33	turn	127	67	bring	69	100	cut	51
34	love	126						

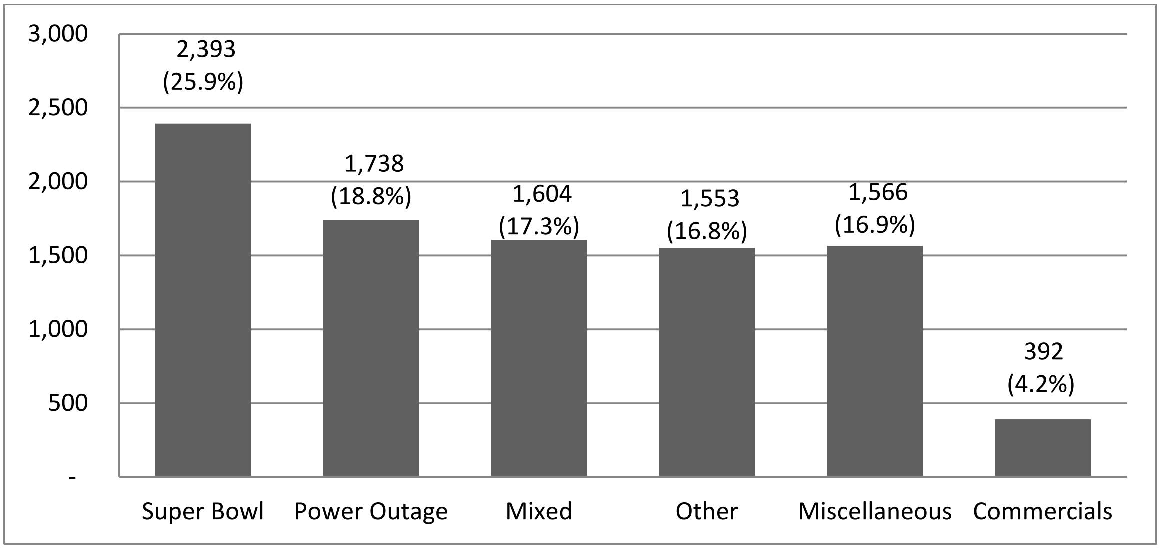

Figure 24. Number of Top 100 Frequently Used Words during Power Outage by Topic (8:40 pm to 9:10 pm)

The words "power" and "lights" were ranked second and third in the list, respectively, and "outage" was ranked ninth. This also indicates that there is a strong relationship between frequently used words on Twitter and real time events and activities. When the power outage happened, the Dodge commercial had not aired yet. Accordingly, all these tweets were categorized into six topics by excluding "Dodge commercial" topic. During this power outage period, the Super Bowl game-related category was the most discussed topic, accounting for 25.9% of all tweets (2,393 tweets out of 9,246 tweets). It was followed by the "power outage" (18.8%) and "mixed" category (17.4%). This suggests that Twitter users still discussed the teams' performances and the game itself more than the power outage. However, the percentages of power outage related tweets dramatically increased from 2.63% to 18.8% as Figure 23 and Figure 24 show. During the game, 2.63% of total tweets were about the power outage. Comparatively, they were 18.8% of tweets during the power outage period. Table 18 shows sample tweets categorized in each of the six topics. For example, the following tweets are samples from the “power outage” category.

- RT @SeinfeldToday: Kramer gets lost in the Super Dome on his way to his seats at the Super Bowl. Accidentally knocks out all the lights.
- That’s the night the lights went out in the Super Dome! #superbowlproblems
- RT @NYTSportsLive: Before that last play got off, most of the lights in the Superdome went dark and the power just died in the press box.
- RT @NYTSportsLive: And folks, we have a bit of a power outage on our hands at the Superdome.
- RT @jakeleehoward: Lights go out in the stadium and my brother yells "Here comes Bane!" #DarkKnightRises

TABLE 18. Sample Tweets for Each Category during the Power Outage (8:40 pm to 9:10 on Feb. 3, 2013)

Category	Tweets Sample
Super Bowl	– the second half of #superbowl47 will be played Hurricane Katrina style – Some angry #49er fans just tried to cause a diversion lol #superbowlxlvii – This games pretty much over #superbowl #ravenswin – @Beyoncé at the #superbowl2013 tho – RT @theneill84: #Beyoncé took all the energy with her when she left the field #superbowl

Commercial	– My 2 favorite commercials so far are #audi and #jeep. #superbowl – RT @ariannahuff: Ahead so far -- in the game: Baltimore. In the commercials: Hyundai. – @mrlevine perhaps it"s his revenge for the Oprah Jeep commercial. – @darrenrovell @audi @mbusa Hell yeah Audi > MB all day!!!! – RT @WPXIScott: Can"t we just start replaying the best commercials so far instead of listening to THIS????
Other	– RT @Midtown_Houston: See what happens when you don"t pay your electric bill. – This poor sideline guy is only used to speaking for 30 really rushed seconds at a time. He"s sooo off his game. #SuperBowl – Keep it going Guys RT @BWWings: We just really don"t want football season to end. Can you blame us? – ʺ@W_Gragg7: I think Buffalo Wild Wings turned off the lights. ????”lol – I don"t even need to watch the TV to know what"s going on right now, it"s all on my timeline
Power outage	– RT @SeinfeldToday: Kramer gets lost in the Super Dome on his way to his seats at the Super Bowl. Accidentally knocks out all the lights. – That"s the night the lights went out in the Super Dome! #superbowlproblems – RT @NYTSportsLive: Before that last play got off, most of the lights in the Superdome went dark and the power just died in the press box. – RT @NYTSportsLive: And folks, we have a bit of a power outage on our hands at the Superdome. – RT @jakeleehoward: Lights go out in the stadium and my brother yells "Here comes Bane!" #DarkKnightRises
Miscellaneous	– Ooops. RT @houstonfowler: $40 Million Iranian Offshore Gas Platform Sinks During Installation http://t.co/B21TfKiB via @gCaptain – RT @The_Nipple: @MikeA787 i laughed a little to hard at that one – Anonymous is taking responsibility for this. – Tell Facebook To Defend The World"s Children http://t.co/6DMlOmvX via @causes – RT @ZodiacFacts: As a #Sagittarius Most Compatible Zodiac Signs for you: Leo, Aries, Gemini, Aquarius, Libra
Mixed	– "That"s the night that the lights went out in..." New Orleans.

	#SuperBowl2013 #SuperBowl – RT @MalloryMcMorrow: GREAT brand placement in the dark #superbowl @MBUSA #lightsout #braverywins http://t.co/A84zb56D – All the football haters are loving this power outage #superball ppl that don"t take it as a serious sport, go #ravens – In case you missed the @Jeep commercial that aired a few minutes ago during halftime, here it is. Well worth your time. http://t.co/vQetX6jy – RT @BeauTAU1FuLYRE: They gonna run out of commercials if they don"t hurry and get these lights back on!

RQ 4 asked whether the overall number of tweets differs by geographic location. The total numbers of tweets were categorized by U.S. states using the method explained in the previous section. As Figure 25 shows, California, New York and Texas were three states that generated the most tweets during the study period, while Hawaii, Kentucky, Minnesota and Missouri had the lowest number of tweets. Also, the number of Super Bowl related tweets increased on game day, Feb. 3, 2013, in most states except for Hawaii, Kentucky, Minnesota, and Missouri. This analysis suggests that the overall amount of tweets and Super Bowl-related tweets is more associated with the size of a state than with the states that two NFL teams belong to for the 2013 Super Bowl. For example, California, New York and Texas are heavily populated states, while North Dakota, Vermont and Wyoming are relatively smaller than other states. However, this finding would require further analysis since it is beyond the scope of this study.

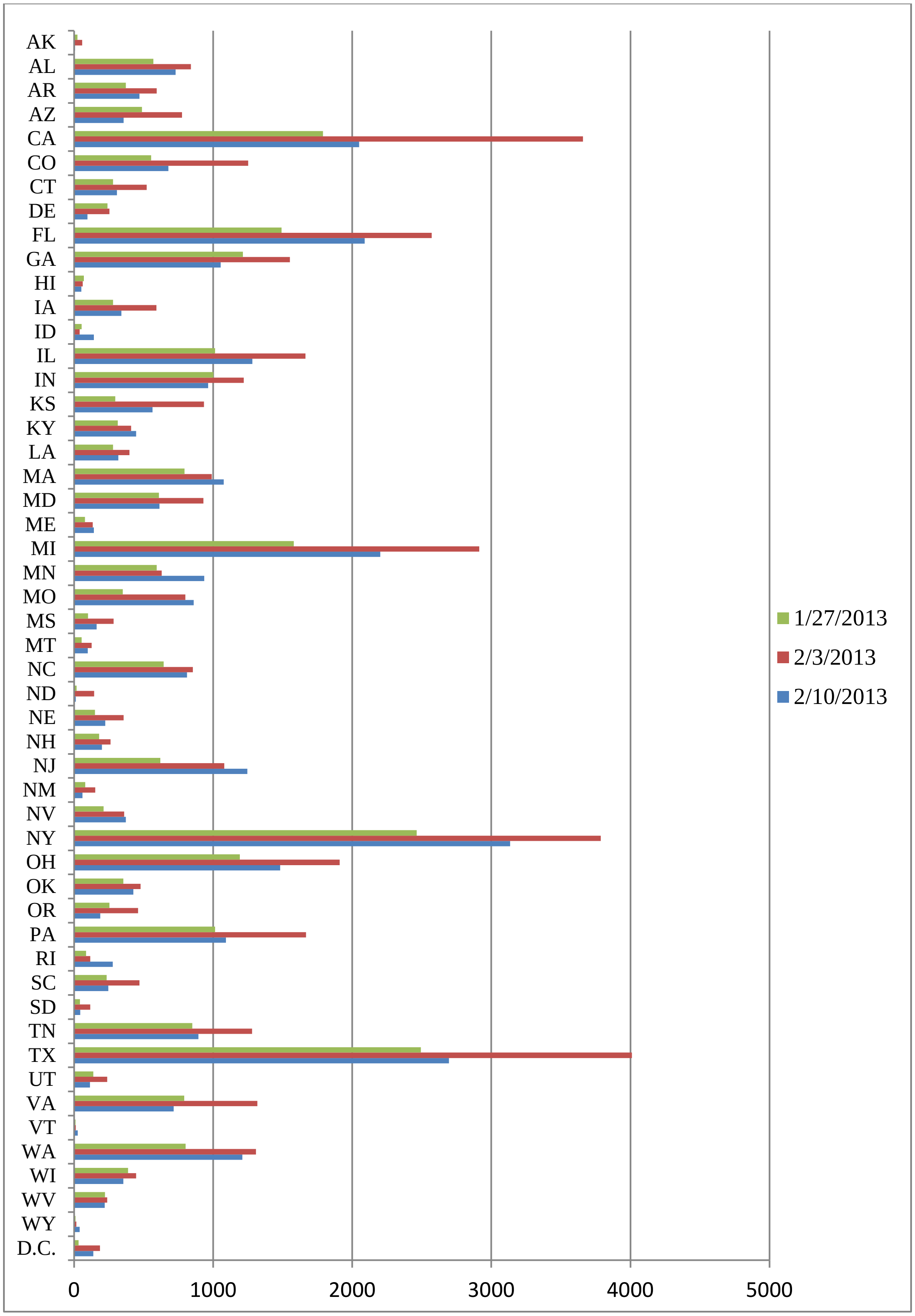

Figure 25. **Total Number of Tweets by State on January 27, Feb. 3, and Feb. 10, 2013**

5.2.5. Discussion

The purpose of this case study was to examine Twitter usage during a sports broadcasting. Instead of relying on the audience's response, this study used a data-mining approach from the computer science field and utilized existing data. Specifically, a data collection tool and a text analysis tool, QB-Text Analyzer, were useful and relevant since they enabled filtering and retrieving of a massive amount of tweets and the key topics discussed in those tweets.

Over the three Sundays (January 27, Feb. 3, and Feb. 10, 2013) from 6:00 pm to 11:59 pm, a total of 305,369 (i.e., data population) tweets was identified. Out of those, 12.7% were identified as Super Bowl-related tweets. When comparing the three dates, the Super Bowl game day had a larger portion of Super Bowl-related tweets (24.2%) than the other two Sundays with 5.0% and 4.9%, respectively. This indicates that the overall number of tweets peaked on game day and the portion of Super Bowl-related tweets increased as well. Another major finding of this study is the Twitter usage pattern on game day is much more dynamic than those from the other two Sundays. The nature of the game seems to affect the Twitter usage. The analysis revealed that more exciting scoring moments generated a larger amount of tweets. For example, seven-point touchdowns generated more tweets than three-point field goals. Even if this study didn't analyze specific motivations behinds such Twitter usage, we speculate that as the audience got excited about the game, they became more engaged and exchanged information and opinions on Twitter.

Besides major scoring moments, Twitter usage also increased around interesting events or activities. Out of five non-scoring moments, the end of the game and the Super

Dome power outage generated the most tweets. This suggests a similar connection between the nature of the game and Twitter usage. It was hard to predict the winner of the 2013 Super Bowl game up until the last minute with a two-point safety and a possible touchdown by San Francisco. The dynamic nature of the game was well reflected by the unusually high ratings, which also translated to increased Twitter usage. Also, an unprecedented power outage during the game also marked an increase in Twitter usage. Even though tweets about the power outage during this period increased, it was an unexpected finding that Super Bowl game-related tweets (25.9%) still exceeded the tweets about the power outage (18.8%).

The major purpose of this study is to provide a snapshot of how Twitter users talked about the Super Bowl game in real time and an understanding of the connection between a broadcast and new social media. As this study shows, tweets related to specific events or moments were exchanged almost instantaneously as those events occurred. The pattern analysis by minute indicated that people mostly tweeted within a 3-minute span following the event. In other words, this reinforces the unique characteristics of Twitter as an instant communication medium. Another contribution of this study is that this study is based on all possible tweets about the Super Bowl game from individual users who used Twitter to engage with the game, not just Twitter feeds of high profile organizations, such as NFL. This macro-level analysis sheds light on when people are more engaged during a broadcast.

This study can be further extended and applied to other major events and Twitter usage during their broadcast. For example, any historical, cultural, or social events can be used, such as the broadcast of the Academy Awards ceremony (The Oscars), the

President Inauguration Ceremony, or natural disasters. Such analysis would provide more opportunities for us to understand how and when people tweet.

As an exploratory study, it is not without limitations. This study did not provide an in-depth analysis of the Twitter contents at a micro-level. Accordingly, the motivations or reasons on why people tweeted at certain moments are unknown. What we can guess from this study are the key topics of tweets. In addition, the relationships among Twitter users discussing this topic were not analyzed in this study, and this study does not provide an insight to the understanding of how closely people are connected or who a key player is in the networked world.

Future studies could consider such limitations to enhance a better understanding of Twitter and Twitter usage. One suggestion is to compare tweets and retweets in terms of their contents, motivations, and user relationships. One of the unique characteristics of Twitter as a social media platform is the capability of retweeting. The analysis of differences in tweets and retweets might allow us to explore underlying motivations for retweets. This would further provide implications for various organizations and business communities to develop effective communication through Twitter.

Chapter 6 .

Conclusions

To apply data-mining techniques to social data, the target data set must be prepared from social networks before the analyzing process. However, it is impossible to collect enough data to apply data analysis techniques and filter out unnecessary data, such as spam messages without an automated data collector and filter. In order to overcome these data access challenges, we designed and developed our own Twitter Data Collecting Tool.

The Twitter Data Collecting Tool allows researchers to gather users' information and follow relationships and tweets from Twitter. It is characterized by the following features. First, it is able to collect the data continuously and automatically. Secondly, it is able to start the collection process with multiple selected nodes. Thirdly, it is able to handle a multitude of authorized developers' accounts. Fourthly, it is able to save the collected data into a database. Fifthly, it minimizes waste of hourly available methods calls. Finally, it is able to filter data using Drools that is embedded in the Twitter Data Collecting Tool.

Selecting seed nodes used for starting point of data gathering process is the most important step to gather more relative data for a specific topic. Thus, we proposed a new algorithm to find better initial seed nodes with limited time and resources to gather the data that is related to a specific topic or keyword that data seekers are interested in. The algorithm evaluates user's activities and updates the seed node list dynamically.

To evaluate the performance of the algorithm, we presented an evaluation of the performance of building the Twitter dataset with the Twitter data gathering tool with our influential node selecting algorithm on the real Twitter network. We compared two results, one from this algorithm and one from a specialist after the data gathering process from Twitter. The result proved that the efficiency of the algorithm for collecting more keyword-related data is higher than the existing approach.

We provided an analysis of Twitter data gathered by the Twitter Data Collecting Tool in a case study about the Super Bowl 2012 and Super Bowl 2013. By using the Twitter Data Collecting Tool, we gathered Twitter data about Super Bowl 2012 commercials, especially those related to cars. As a result, 11 rows of seed nodes, 1,134,798 rows of following relationships, 968,641 rows of follower IDs, 5,157,887 rows of tweets and 88,121 rows of user information have been gathered. After that, we filtered the tweets using Drools and retrieved the tweets about Super Bowl Commercials 2012. Furthermore, we inferred the meaning of tweets using the rule engine and ranked the 11 car companies that advertised their commercials at Super Bowl 2012. In addition, data mining techniques and an external tool called Gephi are applied to the gathered data. As a result, new meaningful knowledge was discovered based on the results that are made by the Twitter Data Collecting Tool. With these results, we could prove that the Twitter Data Collecting Tool is able to gather a huge amount of data from Twitter and filter the data so it can be used in research areas.

The major purpose of the case study about Super Bowl 2013 was to provide a snapshot of how Twitter users talked about the Super Bowl game in real time and an understanding of the connection between a broadcast and new social media. As this study shows, tweets related to specific events or moments were exchanged almost instantaneously as those events occurred. Another contribution of this study is that this study is based on all possible tweets about the Super Bowl game from individual users who used Twitter to engage with the game, not just Twitter feeds of high profile organizations, such as NFL.

Future work on the Twitter Data Collecting Tool can extend computing and storage capacity to gather more data from Twitter faster, because it consumes a lot of computing and storage resources. Using scalable clouding resources will be an easy way of extending these resources. Another improvement to the Twitter Data Collecting Tool is to adapt a built-in data-mining module. Applying data-mining techniques to the social network data has so much potential. For example, applying natural language processing techniques or text mining to the Twitter data can be used to analyze or detect social opinions (Yassine and Hajj, 2011, Dziczkowski et al., 2009). Therefore, a tool that supports collecting and mining the social data is needed for researchers to be able to use this untapped resource.

The weighted seed selection algorithm presented in this paper supports only one keyword. In future work, the algorithm need to be improved to find an initial node based on multiple keywords. Also, this algorithm can be enhanced to analyze user influence in a social network that extends dynamically.

References

[1] Aramaki, E., Maskawa, S. and Morita, M. (2011), "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing in Edinburgh, Scotland, UK, 2011, Association for Computational Linguistics, pp.1568–1576.

[2] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnic, Jay., Kumar, S., Ravichandran, D. and Aly, M. (2008), "Video suggestion and discovery for youtube: taking random walks through the view graph", Proceeding of the 17th international conference on World Wide Web, Bejing, China, 2008, ACM, New York, pp.895-904.

[3] Bastian, M., Sebastien, H., and Mathieu, J. (2009), "Gephi: An Open Source Software for Exploring and Manipulating Networks", Proceedings of the Third International ICWSM Conference in San jose, California, U.S.A, 2009, AAAI, pp.361-362.

[4] Bošnjak, M., Oliveira, E., Martins, J., Mendes, E. and Sarmento, L. (2012), "TwitterEcho - A Distributed Focused Crawler to Support Open Research with Twitter Data", Proceedings of the 21st international conference companion on World Wide Web in Lyon, France, 2012, ACM, New York, pp.1233-1240.

[5] Boyd, Danah. M. and Nicole B. Ellison (2007), "Social Network Sites: Definition, history, and scholarship," Journal of Computer-Mediated Communication, 13(1), pp.210-230.

[6] Byun, C, Kim, Y, Lee, H, & Kim, K. K. (2012). Automated Twitter data collecting tool and case study with rule-based analysis. Proceedings of the 14th International

Conference on Information Integration and Web-based Applications & Services, pp.196-204.

[7] Byun, C, Lee, H, & Kim, Y (2012). Automated Twitter data collecting tool for data mining in social network, Automated Twitter data collecting tool for data mining in social network, pp.76-79.

[8] Byun, C., Park K., Yun, J. and Kim, Y. (2011), “Design and Implementation of the Context-aware Collaboration Framework with the XCREAM (XLogic Collaborative RFID/USN-Enabled Adaptive Middleware)”, The Thrid International Conference on Smart IT Applications, 2011.

[9] Campbell, Colin, Leyland F. Pitt, Michael Parent, and Pierre R. Berthon (2011), “Understanding consumer conversations around ads in a web 2.0 world,” Journal of Advertising, 40(1), pp.87-102.

[10] Choi, Y. and Cardie, C. (2009), “Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009, Association for Computational Linguistics Stroudsburg, PA, USA, pp.590-598.

[11] Chu, Shu-chuan and Yoojung Kim (2011), “Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites,” International Journal of Advertising, 30(1), pp.47-75.

[12] Coffey, S. & Stipp, H. (1997). The interactions between computer and television usage. Journal of Advertising Research, 37(2), pp.61-67.

[13] Cooper, R. & Tang, T. (2009). Predicting audience exposure to television in today's media environment: An empirical integration of active-audience and structural theories. Journal of Broadcasting & Electronic Media, 53(3), pp.400-418.

[14] Correa, D. and Sureka, A. (2011), "Mining Tweets for Tag Recommendation on Social Media", Proceedings of the 3rd international workshop on Search and mining user-generated contents in Glasgow, Scotland, UK, 2011, ACM, New York, pp.69-76.

[15] Cortizo, J. C., Carrero, F. M., Gomez, J. M., Monsalve, B. and Puertas. P. (2011), Introduction to mining social media, International Journal of Electronic Commerce, Vol. 15 No. 3, pp.5-8.

[16] Domingos, P. and Richardson, M. (2001), "Mining the Network Value of Customers" Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining in San Francisco, California, USA, 2001, ACM, New York, pp.57-66.

[17] Dziczkowski, G., Bougueroua, L. and Wegrzyn-Wolska, K. (2009), "Social Network – An tutonoumous system designed for radio recommendation", International Conference on Computational Aspects of Social Networks in Fontainebleau, France, 2009, cason, pp.57-64.

[18] Forgy, C.L. (1982), "Rete: a fast algorithm for the many pattern/many object pattern match problem", Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pensilvenia, USA.

[19] Fruchterman, T. and Reingold, E. (1991), "Graph drawing by force-directed placement", Software-practice and experience, Vol.21 No.11, pp.1129-1164.

[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten. I. H. (2009), “The WEKA data mining software: an update”, SIGKDD, Vol .11 No.1, pp.10-18.

[21] Kwak, H., Lee, C., Park, H. and Moon, S. (2010), “What is Twitter, A Social Network or A News Media?”, Proceedings of the 19th International Conference on World Wide Web in Raleigh, North Carolina, USA. 2010, ACM, New York, pp.591-600.

[22] King, I., Li, J. and Chan, K. T. (2009), “A brief survey of computational approaches in social computing”. Proceedings of the 2009 international joint conference on Neural Networks in Piscataway, NJ, USA, 2009, IEEE Press, pp.2699–2706.

[23] Noordhuis, P., Heijkoop, M. and Lazovik, A. (2010), “Mining Twitter in the Cloud”, Proceeding of IEEE 3rd International Conference on Cloud Computing in Miami, Florida, USA, 2010, IEEE Press, pp.107-114.

[24] Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), “Earthquake shakes Twitter users: real-time event detection by social sensors”, in Proceedings of the 19th international conference on World wide web in Raleigh, North Carolina, USA. 2010, ACM, New York , pp.851-860.

[25] Sottara, D., Mello, P. and Proctor, M. (2010), “A Configurable Rete-OO Engine for Reasoning with Different Types of Imperfect Information”, IEEE Transactions on Knowledge and Data Engineering, Vol.22 No.11, pp.1535-1548.

[26] Yassine, M. and Hajj, H. (2010), “A Framework for Emotion Mining from Text in Online Social Networks”, IEEE 10th International Conference on Data Mining Workshops in Sydney, NSW, Austrailia, 2010, IEEE Press, pp.1136-1142.

[27] Friedman E, “Jess, the Rule Engine for the JavaTM Platform”, Available at http://www.jessrules.com/jess/index.shtml (accessed 2 Feburaly 2013).

[28] Super Bowl Commercials, “10 Best Super Bowl commercial”, Available at: http://www.superbowl-commercials.org/14261.html (accessed 10 Feburaly 2013),

[29] Twitter, “Twitter Rate Limiting in v.1.1”, Available at: https://dev.twitter.com/docs/rate-limiting/1.1 (accessed 25 Feburaly 2013).

[30] Dey, L. and Haque, S. M. (2008), “Opinion Mining from noisy text data”, Proceedings of the second workshop on Analytics for noisy unstructured text data in Singapore, 2008, ACM, New York, pp.80-90.

[31] Hu, M. and Liu, B. (2004), “Mining and Summarizing Customer Reivews”, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining in Seattle, Washington, USA, 2004, ACM, New York, pp.168-177.

[32] O’Connor, M., Balasubramanyan, B., Routledege, B. M. and Smith, N.M. (2010), “From tweets to polls: Linking text sentiment to public opinion times series”, Proceedings of the International AAAI Conference on Weblogs and Social Media in Washington, DC, USA, 2010.

[33] Zhu, M. and Ghahramani, Z. (2002), “Learning from labeled and unlabeled data with label propagation”, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[34] Talukdar, P.P. and Crammer, C. (2009), “New Regularized Algorithms for Transductive Learning”, Proceedings of the European Conference on Machine

Learning and Knowledge Discovery in Databases, Bled, Slovenia, 2009, Springer-Verlag, Berlin, Heidelberg, pp.442-457.

[35] Cha, M., Haddadi, H., Benevenuto,F., and Gummadi, K. P., “Measuring User Influence in Twitter: The Million Follower Fallacy”, 4th International AAAI Conference on Weblogs and Social, ICWSM, 2010, pp.10-17.

[36] Chen, W., Wang, Y., and Yang, S., “Scalable influence maximization for prevalent viral marketing in large-scale social networks” in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Washington,D.C., 2010, pp.1029-1038.

[37] Chen, W., Wang, Y., and Yang, S., “Efficient influence maximization in social networks” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Paris, France. 2009, pp.199-208.

[38] Kempe, D., Kleinberg, J., and Tardos, E., “Maximizing thespread of influence through a social network,” in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Washington D.C., 2003, pp.137–146.

[39] Shang, X., Chen, X., Jiang, Z., Gu, Q., and Chen, D., “Factor Analysis for Maximization Problem in Social Networks”, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing, Nanjing, China, SNPD, 2012, pp.95-101.

[40] Domingos, P. and Richardson, M., “Mining the network value of customers,” in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), San Francisco, CA, 2001, pp.57–66.

[41] Porter, C.E. & Donthu, N. (2006). Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. Journal of Business Research, 59(9), pp.999-1007.

[42] Richardson, M. and Domingos, P., “Mining knowledge-sharing sites for viral marketing,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Edmonton, Alberta, Canada, 2002, pp.61–70.

[43] Lin, C.X., Zhao, B., Mei, Q., and Han, J., PET: A Statistical Model for Popular Events Tracking in Social Communities. International Conference on Knowledge Discovery and Data Mining 2010, July 25-28, 2010.

[44] Kaplan, A.M. and Haenlein, M. (2010), “Users of the world, unite! The challenges and opportunities of social media,” Business Horizons, 53, pp.59-68.

[45] 200 million Tweets per day (2011), Twitter, Retrieved from http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.

[46] What is Twitter? (2012), Twitter, Retrieved from https://business.twitter.com/en/basics/what-is twitter3.

[47] Ferguson, D.A. & Perse, e.M. (2000). The world wide web as a functional alternative to television. Journal of Broadcasting & Electronic Media, 44(2), pp.155-174.

[48] King, I., Li, J. & Chan, K. Tong. (2009). A brief survey of computational approaches in social computing. Proceedings of the 2009 international joint conference on Neural Networks. pp.2699-2706.

[49] Lederer, A.L., Maupin, D.J., Sena, M.P. & Zhuang, Y. (2000). The technology acceptance model and the world wide web. Decision Support Systems, 29(3), pp.269-282.

[50] JBoss, “Drools. Java Rule Engine”, Available at: http://www.jboss.org/drools/ (accessed 1 Feburaly 2013).

CURRICULUM VITA

NAME: CHANGHYUN BYUN

PERMANENT ADDRESS: 10 CROTONA COURT, LUTHERVILLE TIMONIUM, MARYLAND 21093

PROGRAM OF STUDY: DOCTOROL PROGRAM

DEGREE AND DATE TO BE CONFERRED: DOCTOR OF SCIENCE, TOWSON UNIVERSITY, TOWSON, MARYLAND, U.S.A., MAY 2013, DEPARTMENT OF COUMPUTER AND INFORMATION SCIENCES

Secondary Education: Master Degree, Towson University, Towson, Maryland, U.S.A., May 2009, Department of Computer and Information Sciences

Education:

Towson University, Maryland, USA
Doctor of Science Degree, Computer Science Graduation May 2013
Jan 2013
G.P.A – 4.00/4.00

Towson University, Maryland, USA
Master of Science Degree, Computer Science Graduation May 2009
G.P.A – 3.60/4.00

Credit Bank System, National Institute for Lifelong Education, South Korea
Bachelor of Engineering Degree, Computer Science Graduation Feb 2007
G.P.A – 4.50/4.50

Shinheung College, Republic of Korea
Associate Degree, Computer Science Graduation Feb 2006
G.P.A – 4.42/4.50

Major: Computer Science

Professional publications:

- **Journal Publication**

C. Byun, H. Lee, J. You, Y. Kim, “Dynamic Seed Analysis in a Social Network for Maximizing Efficiency of Data Collection”, Invited, International Journal of Networked and Distributed Computing (IJNDC) 2013.

C. Byun, H. Lee, Y. Kim, K. Kim, “Twitter Data Collecting Tool with Filtering and Analysis Module”, International Journal of Web Information Systems (IJWIS) 9.3, 2013.

- **Conference Proceeding**

C. Byun, H. Lee, J. You, Y. Kim, “Dynamic Seed Analysis in a Social Network for Maximizing Efficiency of Data Collection”, 14th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2013), Honolulu, Hawaii, U.S.A. July 1-3, 2013.

H. Lee, C. Byun, K. Kim, Y. Kim, “Super Bowl Advertising in Social Media: The Analysis of Tweets through a Data Mining Approach”, American Academy of Advertising Global Conference (AAA 2013), Honolulu, Hawaii, U.S.A. May 31-June 2, 2013.

C. Byun, H. Lee, Y. Kim, K. Kim, “Automated Twitter Data Collecting Tool and Case Study with Rule-based Analysis”, International Conference on Information Integration and Web-based Applications & Services (iiWAS 2012), Bali, Indonesia. December 3-5, 2012.

C. Byun, H. Lee, Y. Kim, “Automated Twitter Data Collecting Tool for Data Mining in Social Network”, Research in Applied Computation Symposium (RACS 2012), San Antonio, TX, U.S.A. October 23-26, 2012.

K. Park, C. Byun, J. Yun, J. Chang, Y. Kim, “Context-Aware Inference (CAI) Model on Smart Computing Environment”, 2012 International Conference on Information Science and Applications, Kyonggi University, Suwon, Republic of Korea, 23-25 May 2012.

C. Byun, K. Park, J. Yun, Y. KimK, J. Chang, “Design and Implementation of the Context-aware Collaboration Framework with the XCREAM (XLogic Collaborative RFID/USN-Enabled Adaptive Middleware)”, The Third International Conference on Smart IT Applications, Seoul, Korea, August 2011.

J. Yun, K. Park, C. Byun, “Mobile Real-time Tracking System based on the XCREAM (XLogic Collaborative RFID/USN-Enabled Adaptive Middleware)”, Software Engineering Research,Management and Applications 2011, Baltimore, Maryland, U.S.A., August 10–12, 2011.

K. Park, J. Yun, C. Byun, Y. Kim, J. Chang, “The XCREAM Framework and Collaboration Validity Tests”, Computers, Networks, Systems, and Industrial Engineering Jeju Island, Korea, 23-25 May 2011.

K. Park, J. Yun, C. Byun, Y. Kim, J. Chang, “XCREAM: Collaborative Middleware Framework for RFID/USN-Enabled Applications”, Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS2010), Paris, France, November 8 - 10, 2010.

Computer Skills:

Programming	Java, C/C++, .Net C#, ASP .NET, ASP, PHP, JSP, HTML, Javascript, CSS, Python, Powershell, Perl, XML, JSON, AJAX, JQuery, SQL
Operating System:	Windows, Linux, Mac
Database:	MySQL, Oracle, MS-SQL
Server/Services:	Domain Name System(DNS), Windows Deployment Server (WDS), Internet Information Services (IIS), File Transfer Server(FTP), Apache Tomcat Webserver, Glassfish Webserver, IP-Cop Proxy Server, Squid Proxy, Windows Active Directory
Software:	Netbeans, Eclipse, MySQL Workbench, Excel, Visio, Powerpoint, Photoshop, Flash

Awards:

- **Graduate Student Scholarship Award** (2009-2013) –Towson University, Maryland

- **KUSCO-KSEA Scholarships for Graduate Students** (Mar 2011) - Korean-American Scientists and Engineers Association, Virginia

- **Graduate with top honors**-summa cum laude (May 2006) - Shinheung College, Gyeonggi-do, South Korea

- **Future Leader Scholarship (**Sep 2003) - The Korea Scholarship Foundation for the Future Leaders, Seoul, South Korea

- **Undergraduate Student Scholarship Award** (2003-2005)- Shinheung College, Gyeonggi-do, South Korea