Querying For Relevant People

In Online Social Networks

by

Ke Xu

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2010 by the
Graduate Supervisory Committee:

Hari Sundaram, Chair
Aisling Kelliher
Jieping Ye

ARIZONA STATE UNIVERSITY

December 2010

ABSTRACT

Online social networks, including Twitter, have expanded in both scale and diversity of content, which has created significant challenges to the average user. These challenges include finding relevant information on a topic and building social ties with like-minded individuals.

The fundamental question addressed by this thesis is if an individual can leverage social network to search for information that is relevant to him or her. We propose to answer this question by developing computational algorithms that analyze a user's social network. The features of the social network we analyze include the network topology and member communications of a specific user's social network. Determining the "social value" of one's contacts is a valuable outcome of this research. The algorithms we developed were tested on Twitter, which is an extremely popular social network. Twitter was chosen due to its popularity and a majority of the communications artifacts on Twitter is publically available. In this work, the social network of a user refers to the "following relationship" social network. Our algorithm is not specific to Twitter, and is applicable to other social networks, where the network topology and communications are accessible.

My approaches are as follows. For a user interested in using the system, I first determine the immediate social network of the user as well as the social contacts for each person in this network. Afterwards, I establish and extend the social network for each user. For each member of the social network, their tweet data are analyzed and represented by using a word distribution. To accomplish this, I use WordNet, a popular lexical database, to determine semantic similarity between two words. My mechanism of search combines both communication distance between two users and social relationships to determine the search results.

Additionally, I developed a search interface, where a user can interactively query the system. I conducted preliminary user study to evaluate the quality and utility of my method and system against several baseline methods, including the default Twitter search. The experimental results from the user study indicate that my method is able to find relevant people and identify valuable contacts in one’s social circle based on the query. The proposed system outperforms baseline methods in terms of standard information retrieval metrics.

ACKNOWLEDGMENTS

I would like to say my thanks to all the various people who, during the several months in which this work lasted, have directly or indirectly contributed to a very gratifying educational and research experience at ASU.

First, I would like to thank Dr. Hari Sundaram for his continuous and priceless guidance and support all through my two and a half years of Master's research program. Those research discussions with him had been extremely inspiring and insightful for me. I am always surprised and fascinated with the broad ideas he has and gain very much from his meticulous analyzes. He is a keep-busy hard working person. I am always curious how he can manage so much work and complete them as well. He is always my example and my support through the difficulties of my Master research life. More things I want to highlight about him are his intuitive thinking, his ability to abstract the problems and many guides he had given to make me feel proud and confidence.

The next person I would like to thank is Dr. Aisling Kelliher. Dr. Kelliher is more like a co-advisor for me. She, Dr. Sundaram, and I are in the same research group, Reflective Living, Arts, Media, and Engineering. She is an energetic and funny person. It is very pleasing and very comfortable to talk with her. She also gives me many advice and suggestions, especially those related to data visualizations and user interfaces. I still remember the first discussion with her about the Info-Viz, which inspired me and introduced me to the world of Media plus Arts plus Engineering.

I also would love to say my thanks to Dr. Jieping Ye. I took Dr. Ye's Machine Learning and Data Mining classes. His classes are the best I have had in ASU, which are approachable and well arranged. I like to hear his lectures even though I am not so good at mathematics. His lectures always give me more power and willingness to ask more and learn more. Therefore, I am so glad Dr. Ye agreed to be my committee member and for accommodating such a tight and demanding schedule.

I thank all my lab mates: Munmun De Choudhury, Ryan Spicer, Yu-Ru Lin, Shawn Nikkila, Silvan Linn, Lu Zhang, Heng Chen, Brandon Mechtley, and Zhen Li. Our group is more like a family. I am living here instead of working here and we are more like brothers and sisters. It is always great and unforgettable for me to ask or discuss questions with them. They had given me many helps not only for research but also for my life. Here I will also thank my friends Dr. Yinpeng Chen, Meng Chen, Jeff Zhang, Jessie Wang, Wenhui Yu, Jiqing Zhang, Liqing Zhou, Jin Zhang, Xi Fang and Hao Liu for their support and kindness helps as well.

At last, I will thank my parents in China who support me to be here mentally and economically. I will keep my hard working and continue to be their proud son.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

1.1 The online social network, Twitter 1

1.2 Why is this problem interesting 3

1.3 Problems Addressed 6

1.3.1 Semantic analysis of the tweets 6

1.3.2 Finding the right user 7

1.3.3 Determine the social value of one's neighbors 8

1.4 Summary of Contribution 8

1.5 Organization of thesis 9

2 RELATED WORK 11

2.1 Document finding 11

2.2 People finding 12

2.3 Semantic analysis 15

3 DATA COLLECTION 16

3.1 Introduction 16

3.2 Data collection plan 16

3.2.1 Identifying data types and sources 16

3.2.2 Identifying what will be involved 17

3.2.3 Tools and methods will be utilized 17

3.3 The Data crawler 20

3.4 Collected data 22

3.4.1 Data bases 22

3.4.2 Data elements 22

4 METHODS AND INTERFACE 24

4.1 Semantic analysis 24

4.1.1 Why Semantics 24

4.1.2 WordNet 25

4.2 Semantic distance using WordNet 26

4.2.1 The algorithms 27

4.2.2 Experimental Setup 29

4.3 Different query methods 32

4.3.1 The Baseline method 32

4.3.2 WordNet based method 33

4.3.3 WordNet + Social Graph 35

4.3.4 WordNet + Profile 36

4.3.5 WordNet + Social + Profile 37

4.4 Social value 37

4.5 Interface 39

5 USER STUDY AND EXPERIMENTS 42

5.1 Introduction 42

5.2 Design of user study 43

5.3 Results 45

5.3.1 Result of questionnaire 45

5.3.2 System testing and evaluation 50

5.3.3 Social Value evaluation results 54

5.3.4 Limitations of the user study 56

5.3.5 Useful suggestions and facts from user study 57

6 CONCLUSIONS AND FUTURE WORK 59

6.1 Research summary 59

6.2 Improvements 60

6.3 Future directions 61

7 REFERENCES 63

LIST OF TABLES

Table Page

3.1: Data elements table for each user 23

3.2: Data elements table for each tweet 23

5.1: Table for overlaps between top 5 most valuable friends given by participants and top five socially valuable friends provided by system … 54

5.2: Table for rank correlation between systems provided ranking and participant re-ranking........ 55

LIST OF FIGURES

Figure Page

1.1: Statistic of tweets per day from 2007 to 2010 2

4.1: Use word "play" to show the example that one synset can have multiple words and one word can exist in multiple synsets 26

4.2: F-measure comparison for different algorithms: Shevade for Shevade's algorithm, L-C for Leacock-Chodorow's algorithm, Resnik for Resnik's algorithm.. 32

4.3: Process of calculating one user's similarity score by using WordNet based method.. 34

4.4: Example for composition of a vector and the scope of the people-people similarity 36

4.5: Interface of the People Finding system 40

5.1: Pie chart of Twitter usage frequency results 45

5.2: Histogram of the question for what the participants tweet about........ 46

5.3: Histogram of the question regarding the motivation for using Twitter 47

5.4: Histogram of the question for when user will follow others 48

5.5: Pie chart of the question for how many of the participant's followers are known in real life 48

5.6: Pie chart of the question for main composition of participants' Twitter friends 49

5.7: Histogram for average relevant count for the five methods; y-axis is the average amount of relevant results the user selected from choices of ten. 51

Figure Page

5.8: Overall normalized average rank values, average from all participants 53

5.9: Overall normalized average rank values using standard queries and user specific queries 53

1 INTRODUCTION

1.1 The online social network, Twitter

Social networking and micro-blogging have become a popular phenomenon in recent times. For example, Twitter is used globally by wide demographic of users. Twitter allows individuals to broadcast information quickly and briefly via messages known as tweets, which has been accepted as a way of quickly sharing information.

We chose Twitter for this research due to the following reasons. First, Twitter can have a wide range influence and serve as catalysis for social change. Protests during the 2009 Iranian presidential election have been nicknamed the "Twitter Revolution" because of the protesters' reliance on Twitter [1]. Second, tweets require little time investment due to their 140-character limit, which motivates individuals to use the service. An additional motivation for individuals to use Twitter is the fact that they can tweet anytime and anywhere. The prevalence of desktop software, phones applications, and Internet browser add-ons enable them to do so. One example is that when Bill Gates released a mosquito during a TED 2009 conference, Dave Morin, the manager of Facebook, wrote a tweet about this anecdote at the very moment that it occurred [2]. Third, Twitter has an enormous user base consisting 75 million users at the beginning of 2010. According to the Twitter BLOG (Feb 22 2010), only 5,000 tweets were posted per day in 2007. This number grew by 1,400% in 2009 to 35 million per day (Figure 1.1). It further grew to 50 million tweets per day in Feb 2010, which translated to an average of 600 tweets per second. In addition, most of this communication data is public available.

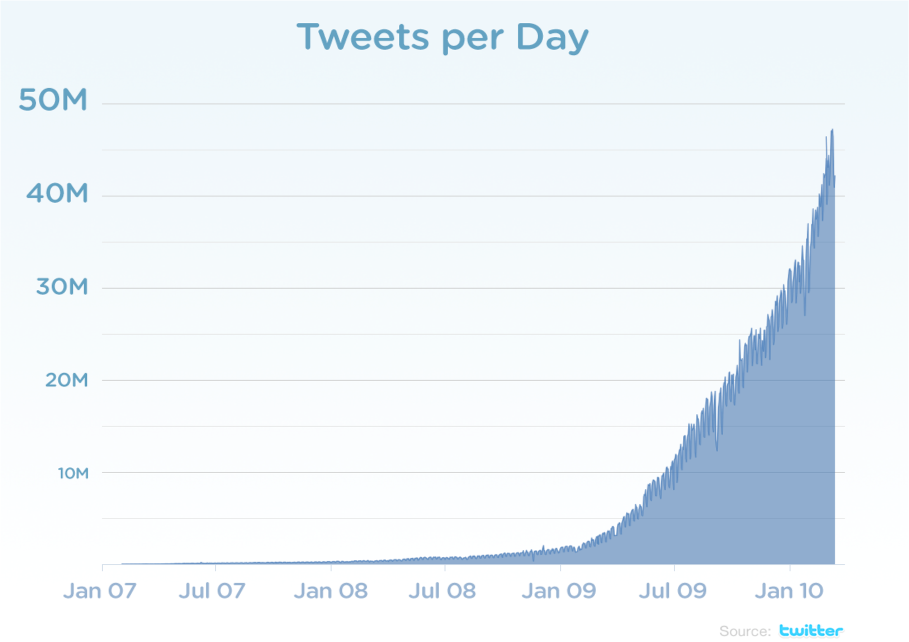


Figure 1.1 Statistic of tweets per day from 2007 to 2010[3]

Twitter is not only a platform for people-people sharing and interactions, but it also serves as a public information center. Deal websites use Twitter as a mailing list to post deals and coupons; celebrities use it as a fan club to share gossip and interact with their fans; and colleagues use it as a bulletin board to share professional information or ask technical questions.

Although Twitter is used in many ways, the only way to be fully involved is to follow other users. Following a person on Twitter makes that individual's tweet history visible. People can read a random Twitter user's tweets and view his or her basic personal information from their Twitter personal homepage if the user's account is not private. Knowing who to follow is a problem since following other users is a very important part of Twitter.

In the next section, we shall discuss the motivation of this research. In section 1.3, we shall present the problems in doing this research. We shall summarize the contributions of the thesis in section 1.4. Section 1.5 will presents the organization of this thesis.

1.2 Why is this problem interesting

The problem how to find relevant individuals to follow on Twitter is very meaningful and interesting to deal with. Solving this problem can guarantee Twitter users to have valuable interactions and sources of information.

There are some key problems faced by users with current state-of-the-art technologies available. First, for a certain keyword, the search function on Twitter only performs word matching. The returned results are based on only the most recent tweets posted on the Twitter so that they might not be the true reflection of a user to be a topical authority on a topic. Additionally, the returned accounts hardly have social relationships to the user who is involved in the query-based search.

Second, the “browse interests” function of Twitter can only provide a list of accounts according to some general topics. The listed user accounts have no further details to enable users to judge if these accounts are relevant. The given accounts through “browse interests” always contain many celebrity-like accounts, such as companies, music bands, movie stars, singers, and sportsmen. Some users may not care about such accounts to serve as a reliable source of information on a topic.

Third, Twitter provides another way for searching for people: searching for accounts according to account names. Consequently, users can only search for the Twitter accounts whose screen names they already knew. It is more like exchanging e-mail addresses between people who are friends in the physical world. That is to say, searching people according to account names cannot help find unknown people on Twitter if they are not friends in real life or do not know each other's account names from before.

It is intuitive that people like to accept unknown persons who have shorter distance with respect to their ego-network, rather than totally unknown persons on the Internet. This intuition springs from observations of trust in terms of reliable information consumption sources, or in terms of user-user similarity of interests. Following these ideas, many online social networks, like Facebook [4] and Flickr [5], suggest potential accounts that have mutual friends with the user (i.e. highly embedded with respect to their egocentric network). As a result, the trust between users can transmit through the social connections. It may be more reasonable to follow a person who is a "friend's friend" than to follow a random people. Our approach to this problem takes these ideas in to account. We focus on the group of users in the user's egocentric network and extended social network instead of randomly sampled users from the Twitter universe.

Interpersonal ties are very important in social network. In mathematical sociology, interpersonal ties are defined as information-carrying connections between people. Interpersonal ties, generally, come in three varieties: strong, weak, or absent [6]. Weak social ties [7] are responsible for the majority of the structure of social networks as well as the transmission of information through these networks. Specifically, more information that is novel flows to individuals through weak rather than strong ties. Because close friends tend to move in the same circles, the information they receive overlaps considerably with what people already know. Acquaintances, by contrast, know the people not in close relationships thus they receive information that is more novel. Levin et al. [8] presented empirical work examining how weak ties provide useful knowledge.

Hence, our work utilizes the idea of helping Twitter users to understand which of the friends in their social networks hold more important interpersonal ties, and possibly to find the weak ties. Our approach not only provides the relevant accounts according to the user's query, but also provides the direct friends who contribute most for the returned relevant account lists. We define “social value” as the importance of the friends in user’s social network who are responsible for the interpersonal ties leading to valuable information on a topic.

Finding the right people is important; finding which of a user's friends are more valuable is significant as well. Both of the two findings are likely to make the search for new information more precise and efficient. Finding relevant persons who are topical authorities is a shortcut to getting more useful and related information. Knowing which friends are more socially valuable is seeking the bridge to newer and bigger social networks. All these are benefits from the aspect of egocentric social network that tend to be targeted to the querying user and consequently more meaningful to the users themselves.

The goal of this thesis is to solve the problems described above, that is, finding the relevant topical authorities who are more social related to the particular user in question, and finding the friends of the same user who are socially valuable. To reach this goal, we build a query-based system to find the accounts that may be relevant to the users. In the system, we also find out which of the user’s direct friends are more “valuable” to the user for providing relevant accounts. People enter queries into the system using a set of keywords. They may want to find experts who know the topical area well or just search for someone who is vocal about the topic. Thus, they may discover and consume more information from those specific accounts that are recommended by our system after following them. Our system analyzes the contents of tweets semantically and thereafter recommends relevant individuals to a user based on his or her query. Not only the tweet content serves as the evidence for finding relevant people, but also social network information is likely to be a valuable feature in the recommendation process. In addition, the query user's tweeting history will also play a significant role to affect the returned results. The "top friends", who are ranked on how valuable to the user they are, are also provided by our system, which provides the possibility for the querying user to find out which friend of the user may serve as the bridge to the important interpersonal ties within the user's egocentric social network.

1.3 Problems Addressed

In this section, we shall discuss the intuition behind our solutions to the problems, including analyzing tweets, strategies of finding relevant people, and methods for finding socially valuable friends.

1.3.1 Semantic analysis of the tweets

We attempt to analyze the pure tweet texts, by using existing ontology such as WordNet [9] [10], the popular lexical database, to exploit semantic relationships between words. For example, assume there is a tweet that has the word "rose". The user posted this tweet when the he was intending to talk about flowers and gardening. Therefore, the tweets that contain "flower", "plant" or "garden" might be the similar tweets and their owners could have more possibilities of being the potentially relevant users. Since the ontology like WordNet encapsulates people's understanding and knowledge of the world in the format of different semantic relationships like “is-a”, “location-of” and “part-of” etc., and we know “rose is a flower”. Therefore, we can use those ideas to categorize and analyze the tweets. For instance, the tweets contains “rose” and the ones contains “flower” are more similar.

1.3.2 Finding the right user

We approach the problem of comparing and calculating similarity distances from the aspect of semantic network. First, for each user, we gather users’ historical tweets and count the distribution of the words. Then we translate the words into synsets[1] by using WordNet. Next, we compare and compute the similarity for each pair of these synsets. After the calculation, results will combine the synsets similarities with other supporting information like social network topologies and personal profiles. For example, if user A had mentioned the words “football”, “player”, and “quarterback” for several times in his historical tweets, those tweets will be collected and considered relevant to football. It seems user A may be a football fan. If other user who like football very much types “touchdown”, which is also relevant to football, as a keyword to query. The keyword will be processed into the querying user’s personal social network to calculate the semantic distances with every other account’s historical tweets to see if any of these accounts are football fans. As expected, the system flags user A if he has a higher rank than most of the other users and he is in the social network of the querying user.

The social network topologies and personal information will also play roles in our query methods. People who hold important positions in the social network may be more powerful and useful; the users who are more similar with

[1] synset: A set of one or more synonyms (explanation from WordNet)

others based on their tweet history and personal interests should be paid more attention. Those are all the aspects our approach will consider about and will take into accounts.

1.3.3 Determine the social value of one's neighbors

We seek for the person who contributes more than others do to the returned result of relevant users. The system returns the relevant users based on the query and finds those users from the querying user's personal social network. The querying user may not have a "follower" relationship with some of the returned users, which means that the he or she may not know who these users are. However, these unknown accounts have a "follower" relationship with at least one of the user's neighbors. In other words, the user's neighbors help provide the individuals who may be relevant to the user. The neighbor who has higher social value is the one who bring effects that are more influential to the returned result of querying relevant accounts. Our approach finds out social-value friends according to this idea.

1.4 Summary of Contribution

We now summarize the original contributions of this thesis in attempting to solve the problem of finding the people who may be relevant on the Twitter network and providing the social values of user's neighbors. The goal of this work is to provide better functional methods to enrich the social network experiences such as using Twitter. We now briefly summarize our original contributions in attempting to solve the problem of relevant people finding:

- We compared several algorithms for tweet semantic similarities that used for finding potentially relevant users for individual Twitter user
- Proposed a keyword based and user personalized queries for search

- Provided different search methods which were combined with social network and personal information
- Determined the social value of one's direct friends (neighbors)
- Developed an application for searching relevant users on Twitter

1.5 Organization of thesis

The rest of the thesis organizes as follows. In the next chapter, we shall discuss in detail about the related works that how the other researchers worked on helping people to find related things such as documents and persons. In that chapter, we shall focus on how those methods work and if those ideas can be incorporated into our work. We also discuss about the limitations of those work, and the differences with our work.

In chapter 3, we shall present the technique used for collecting data for our research. The technique focuses on implementing the data crawler utilized Twitter API and the ways of gathering data. We will discuss the source of the data collection and introduce the data types. Additionally, we introduce the databases we build for storing collected data.

Chapter 4 will discuss the methodologies we have used for our work, including comparing several different WordNet similarities algorithms with experiments. We will also introduce different query methods used in our system for finding and ranking relevant people. The application interface for querying relevant people will be introduced as well. We shall present details of the algorithms, algorithm comparison experiments, and interface functions in this chapter.

Chapter 5 will be the detail introduction for the design of our user study and the user study results with their analysis. We shall describe why the questions are important and how these questions will lead and inspire the improvements and directions of our work. The results of the user study will be provided, as well as the evaluated comparisons for the query methods. The participants' discussions during the user study will be included since they bring many fresh thoughts for our work.

Chapter 6 contains the conclusion of our work and the future directions of this research as well as possible improvements and potentials.

2 RELATED WORK

There has been a lot of prior work for helping people to find things, such as finding information on the Internet, finding similar documents, and finding relevant people. Semantic analysis plays an important role for finding relevant documents and relevant people. In this chapter, we will briefly discuss the related work in three areas: document finding, people finding, and semantic analysis.

2.1 Document finding

Trappey et al. [11] developed a document classification and searching methodology based on neural network technology that helped companies manage patent documents more effectively. The classification process began by extracting key phrases from the document sets in the way of automatic text processing and significance key phrases determining according to their frequencies in the text. In order to maintain a manageable number of independent key phrases, they applied correlation analysis to compute the similarities between key phrases. The back-propagation network model was adopted as a classifier. The target output identified a patent document's category based on a hierarchical classification scheme, the international patent classification (IPC) standard. The idea of key phrases extracting and text processing according to their frequencies is a way we can use for the micro-blog texts as well since they both have key phrases and key words as the representatives of the documents/tweets.

Berry et al. introduced Latent Semantic Indexing (LSI) [12] for retrieving textual materials. Because of the tremendous diversity in the words, people use to describe the same documents, lexical matching methods maybe incomplete and imprecise. LSI tries to overcome the problem of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval.

LSI is based on a mathematical technique called Singular Value Decomposition (SVD). LSI is used to match queries to documents in information retrieval applications. LSI has been shown to improve retrieval performance for some collections, when compared to traditional vector space retrieval.

However, document searching is different from relevant people searching. For example, the same person may create documents in different categories; different persons may create documents in the same category. Documents are searched only from their contents. That is to say, algorithms calculate the similarity measures from the pure contents of every single document. Hence, the result of document finding may not reflect the authors' personal interests well. However, our work wants to find the relevant people although they may create tweets in various areas. As a result, personal information and personal social network topology play important roles in our approach.

2.2 People finding

Artiles et al. [13] had provided a method for people searching strategies in the web documents. They retrieved the web pages by using person names, classified the pages according to the people, and manually annotated the relevance. Results of applying clustering algorithms are also provided as a baseline for the ambiguity resolution problem. Their idea and motivation were from the statistics of people name searching on the Internet (30% of search engine queries include person names [14]) and people names were very ambiguous. The limitation of their work is that they searched and retrieved data from web pages. A person's web pages may be incorrect or misleading if they were not created by the person himself or herself. Moreover, people do not prefer to use their full real name publicly on the Internet. Web pages with misleading or wrong information can make Artiles' work less trustworthy. On the other hand, the information from the social network are more reliable since only the person himself or herself can create and edit his or her personal information and use those information to have online social interactions. These social interactions contain more real information and true data as well.

Dunlop et al. [15] developed and evaluated clustering techniques for finding people. Their motivation was quick match-ups for the persons who were in the same company or university. The persons had close research areas or similar working fields but did not know each other. The paper reported an investigation into the use of information retrieval (IR) techniques to automatically matched people according to their web pages. It provided three clustering algorithms to evaluate, including balanced clustering, single link clustering, and group average clustering. The paper had the similar motivation for matching the persons according to their personal interests or related fields. Still the information source was the people's web pages and there were no other personal options. This research work can help ours with the methods of matching people considering the clustering algorithms and other IR techniques.

Chen et al. compared similarities and network cues for recommending people in the enterprise social network at IBM [16]. Their work matched up the users according to their job positions, working projects and other information related to the enterprise. It provided people recommendation within a restrained network, which was similar to our work. Therefore, we are not taking the whole Twitter population as the user group to select users. We choose the data from the seed users' direct friends, the direct friends of those direct friends, and so on. This process can make our collected data trustworthy and useful because people share more social connections in their own circles than randomly picked user since Chen et al. found that similarity was a stronger cue in recommending new contacts

Guy et al. expanded Chen's idea by proposing that user similarity might be derived from the "people, things, and places" they shared [17], which is a meaningful fact that can help our research. We follow the factors "people" and "things". We take the advantages of personal social graph and personal tweets as the "people" and "things" factors when calculating the similarity between Twitter users. It is more meaningful that we care the "things" users talk in their tweets as the most important features. The "place" factor is not an appropriate element for our approach right now because the location information and time zone information are optional on Twitter. After checking users' location and time zone data in our database, we found many of them are blank and some are incorrect. As a result, location and time zone information are not qualified to be considered as features to take effects right now.

Golder and Yardi evaluated structural patterns on Twitter and found that structural paths involving reciprocated links were generally a strong signal in recommending users [18]. Our method takes the social network as a directed graph, considering the following relationships as the edges of the graph. In our approach, there are no priorities to the bidirectional links in the social graph. The bidirectional edges will be considered as two separate edges.

There are also some work related to "finding people" but they actually found people images in pictures [19] [20]. Those work used sampling and probabilistic methods for finding people figures in a static image. It is not exactly the same area as ours. Our “finding people” is finding the potentially relevant accounts in Twitter network according to the user specific queries.

2.3 Semantic analysis

One of the big differences between our relevant people searching and other related work is that we try to understand the meanings behind people’s tweets. That is to say, extract useful information from their words in tweets. There are also plenty of previous works in the area of understanding what people mean when naming categories or giving tags to images [21] [22]. The authors provided a framework to annotate images using personal and social network contexts. The system intelligently annotated tags for images according to the user contexts, event contexts and social network based recommendations. The authors used natural language processing (NLP) tools, such as WordNet, to analysis concepts and features in linguistic relationships to help annotating images. The idea of utilizing WordNet is exactly the same method our approach will take.

Budanitsky and Hirst [23] discussed several different proposed measures of similarity or semantic distance in WordNet. They compared those algorithms by examining their performance in a real-world spelling correction system. Their purpose was comparing the performance of several measures of semantic relatedness that had been proposed for use in NLP applications. Their work inspired our research that we could test and compare semantic algorithms to evaluate which fits our approach better; this idea will be discussed more in chapter four.

3 DATA COLLECTION

3.1 Introduction

There are two reasons to collect users' personal data, social network data, and tweet data before performing the people searching. First, our algorithms need too much data to collect if requiring them dynamically. In our approach, the relevant people are found from the querying user's personal social network. For example, if a user has 100 contacts in his social network, all the tweet data of these 100 accounts and the overall social graph topology are required. Second, calling Twitter API for a user's tweet history only returns the latest 200 tweets. If one user posts tweets very often, we may not get his/her older tweets by calling the API since his recent tweets may be more than 200. As a result, building a data crawler and collecting data persistently are necessary. That is to say, we need store the tweet data in advance.

In this chapter we will introduced the data collection process for our work. Section 3.2 will introduce the data collection plan, including data type, data source, collecting methods and the tool we utilize. Section 3.3 will briefly discussed the data crawler and its functions. In section 3.4, we will simply introduce what data we had collected.

3.2 Data collection plan

We will discuss the sources and types of the data we collected, the subjects involved in the data collection, and the tools utilized for the data collection

3.2.1 Identifying data types and sources

Twitter is the source we use to collect data. An important initial step in data collection is to make an inventory of the types of data and clarify where or from whom they will be collected. There may be two types of data: existing data, or called pre-data, and program-generated data. In our case, the data collected from Twitter was pre-data. Twitter provides public API for developers to collect public data. If a Twitter user does not set his account private, his tweets and personal information will be considered as public data.

3.2.2 Identifying what will be involved

Strictly speaking, all the people who have accounts in Twitter without setting privacy protections are involved in the data collection. In our work, Twitter is the data source and we do not directly collect the data from Twitter users. Therefore, we use Twitter API to collect data. Twitter API permits 150 anonymous requests per hour for each host IP but 150 requests are far not enough for our approach. As a result, we had applied and been approved to join the REST API white list. Then we can make up to 20,000 rate-limited API requests per hour.

In our approach, the seed users are the members of Reflective Living research group in Arizona State University. The members in Reflective Living group are graduate students and faculty. Their research areas include analysis of large-scale social networks so that they are moderately to highly active on online social networks, such as Twitter. Their direct contacts on Twitter, or say their neighbors, and the direct friends of the neighbors are the targets to be collected. Next section will discuss more details about data collection.

3.2.3 Tools and methods will be utilized

We utilized Twitter public API [24] and open-source Twitter library to collect data. Twitter provides its APIs to collect public data. The Twitter API has three parts: two REST APIs and a streaming API. The REST API allows developers to access the core Twitter data, including updated timelines, status data, user personal information, and so on. We use the REST API in our approach. We need Twitter users' raw data, which includes their tweet data and personal information. We implemented our data crawler by using python with the Python Twitter Library [25]. This library provides a pure python access to the Twitter API, that is to say, it is a python wrapper around the Twitter API. Twitter exposes a web services API [26] and the Python Twitter library is intended to make it even easier for python programmers to use. We implemented our crawlers beyond the bases of these APIs and libraries with our modifications.

Generally, there are two data collecting methods for collecting the data of a social network. One is random sampling, which will collect all the data once in one timescale and then collect in another random timescale and so on. That is to say, this method will collect random tweets whose time stamps may be the same or very close to each other. However, there are no other connections between these tweets. The authors of the tweets hardly have close relationships. The advantage of this method is that, as its name, it can collect random tweets posted by random users. However, these tweets have no further relevance. Moreover, the authors of these tweets can hardly have social relationships.

The other method for the data collection in social network is snowball sampling. This method starts from a seed as the root node, follows the social network structure, collect the data from the root node then continue to collect the data from the nodes in the next level of the social graph. This method can guarantee the social network structure remained in the collected data. That is exactly what we want: full user information and complete social network structures. By using snowball-sampling method, the user database was growing exponentially with the number of levels expanded in social network.

Twitter had changed its authentication method for accessing data. Since September 1, 2010, Twitter stopped to support basic authentication method. All the work should change to the OAuth protocol [27]. A user can grant to a Web application tickets to access protected resources hosted in another site by using OAuth protocol, without trusting any set of credentials. Our data crawler had to change for continuing collecting the data. We implemented the new Twitter crawler by using the new Python Twitter API wrapper, tweepy [28], which supported OAuth protocol and basic authentication both. The first data crawler started to work from May 2010. Then Twitter changed their authentication method from Sep 1 and the new Twitter data crawler using OAuth protocol started working around Sep 10.

Here are some statistics of the collected data. Up to the day for user evaluation, the data crawler had crawled 22,809,205 tweets. Remind that Twitter API only provides a user's latest 200 tweets as the result of one request. Therefore, because of the different tweeting frequencies among users, the time range of tweets for each user might vary. The oldest tweet time stamp from the database was 2006-04-14 03:25:58 and the newest one was 2010-10-25 03:00:04. The timestamps were directly crawled from Twitter and there might be time zone differences. However, it would not influence the major time ranges. Additionally, we had collected 89,022 users in total (accounts had set as private cannot be collected and not included) which are in seed users' two-hop-friend social network. There are 7,209,589 friend relationships in our crawled database (bidirectional friendships are counted as two friendships).

3.3 The Data crawler

The crawler included several sub crawlers; each of them had their own special functions. The first data crawler was implemented by using python with Twitter Python Library and the second data crawler changed the library to Tweepy. However, we kept the same name and the same function for the two crawlers. The data crawling process followed the snowball sampling method. We used one account as the seed of the seed users since every seed user is this account's direct friend. The user crawling process started from the seed users and went deeper level by level. Following are the main functions of the data crawler.

- Function **crawlUserInfo:**

This function will crawl the information belongs to a seed user's friends from the seed user's direct social network and extended social networks. There is an argument that constrains the looping times for the crawler, performed as the number of extended levels. For example, if we set the number as 2, it will only crawl the seed user's direct friends (neighbors) and the friends of those neighbors, which are the contacts in the extended social network (one level extended). The friendship was defined as the "following" relationship in Twitter. This function only collects the personal information from the Twitter API, such one user's screen name, real name, location, and number of followers and so on. We had modified and improved the original python library for requesting the friend list. We had implemented our own function for collecting user information from Twitter which is more efficient (100 friends are returned by using just one request count) than the previous ones (either only return the first 100 friends or each friend will cost one request count).

- Function **crawlSocialGraph:**

This function fetches the full social network topology for all collected Twitter users. It goes over the database w contains all the Twitter users collected by the crawlUserInfo function. Then it sends API requests to collect each user's friend list. The friend relationship is defined as the “following” relationship in Twitter. Each of the friend relationship is considered as a directed edge in the social graph and these edges will be stored in the SocialGraph database. There are two attributes for each edge: user_from and user_to, denoting the direction of the friendship. That is to say, there are no bidirectional edges in our database. If user A and user B follow each other, there will be two separate edges in the social graph. In our approach, we did not collect the “followed” relationship. Because people can be followed by any account, especially spam accounts. People may not trust these accounts or may not be interested in them.

- Function **getUserTweets:**

This function collects tweet data posted by the users. It goes over the user database, sends API requests to Twitter, and gets each user's latest tweet data. Tweet data mainly includes ids of the tweets, account names of the authors, timestamps when the tweets were posted, and the contents of the tweets. There is a limitation that Twitter API only returns each account's latest 200 tweets. That is to say, the crawler need keep running to collect more historical tweet data. This function has a parameter to set the repeating time to collect all users’ tweet data and the start index of the collection. The default parameters make the crawler keep crawling all the time and always start from the first user in the database. For those users who have set their account private, our crawler cannot collect their tweet data.

3.4 Collected data

We will briefly introduce the data we had collected in this section, including the databases and the data elements.

3.4.1 Data bases

For storing the data we had collected from the Twitter, we had built several databases for easy querying and utilization. There are four databases so far.

- **SocialGraph** database: it contains all the friend relationships in our dataset.
- **UserInfo** database: it stores the basic user information, crawled from the crawlUserInfo function.
- **Tweets** database: it has all the tweet contents of the users.
- **Users** database: it is a helper database, which stores some flag values to identify crawling histories, such as if a user had been crawled for its social graph information.

3.4.2 Data elements

For each user, we had the following data elements displayed in Table 3.1. For each tweet, in the mean time, we had the following data elements displayed in Table 3.2.

Table 3.1 Data elements table for each user

Element	**Data type**	**memos**
name	Pre-data, String	User's real name
User_id	Pre-data, number	User's specific id, distributed by Twitter
Screen_name	Pre-data, String	User's account name
location	Pre-data, String	User's input when they register the account
Time_zone	Pre-data, String	User's input when they register the account
S_count	Pre-data, number	Number of user's existed statues
F_count	Pre-data, number	Number of accounts who follow this user
Frd_count	Pre-data, number	Number of accounts the user follows

Table 3.2 Data elements table for each tweet

Element	**Data type**	**memos**
Tweet_id	Pre-data, number	The id of the tweet's author, relates the user_id
Screen_name	Pre-data, String	User's account name, main display when using the Twitter
Time_Created	Pre-data, timestamp	The timestamp this tweet was posted
Content	Pre-data, Text	The content of the tweet
Reply_to	Pre-data, String	The screen_name the tweet replied to (if available)

4 METHODS AND INTERFACE

In this chapter, we shall present our approach to deal with the problem of querying for relevant people based on queries. In the first section, we shall briefly explain why we use semantic analysis and introduce WordNet, the lexical database we utilize for our semantic analysis. In section 4.2, we compare five WordNet similarity algorithms with related experiments to determine which algorithm is optimal for our approach. Section 4.3 will discuss a set of different methods for finding the relevant people according to the query word and other social network related information. Section 4.4 will discuss the social value, which is one of the outcomes of our research. We also design and implement an interactive interface, which is introduced in section 4.5.

4.1 Semantic analysis

We shall discuss why we need semantics analysis in our methods and how we incorporate semantics in our system. We shall also introduce WordNet, a lexical ontology database.

4.1.1 Why Semantics analysis

Semantics analysis is very important for understanding what users want to express in their tweets. There are many Twitter applications, such as TweetCloud [29], which only considers words comprising a tweet as symbols or tags. In TweetCloud, users can view a tag cloud of their tweets with the font size signifying how frequently a word has appeared. Unfortunately, TweetCloud and other similar visualizations only count the frequencies for each word and do not take the stems of the words into consideration. For example, “wolves” and “wolf” are treated as two different entities. Those applications also do not recognize semantic links; for example, football is a sport. We need to understand the meanings behind people’s tweets and the relationships between tweets. Therefore, simply performing frequency counting and word matching are not enough. We need to understand the topic behind the tweets to find relevant people for the user. For example, if a user queries about “touchdown”, our system should know that the use may talk about football and returns the relevant accounts who also posts football related tweets.

4.1.2 WordNet

We incorporated and utilized semantics in our system by utilizing the WordNet ontology. WordNet is a large online lexical database. It also has its desktop database, which stores the concepts of words and their relationships. Our approach accesses the local version WordNet data (Windows version 3.0) to process our analysis.

WordNet is composed by synsets. WordNet organizes English words into several categories including nouns, verbs, adjectives, etc. Those words are stored as synonym sets, also known as synsets, which represent lexical concepts. Synsets are interlinked by means of conceptual semantic and lexical relations. One synset may contain several words, and one word may be in several synsets. For example, Figure 4.1 shows that synset 07007945-N (the id indicates that the synset belongs in the database of nouns and has a unique id) has the words “play”, “drama”, and “dramatic play”. “Play” has several meanings which includes “an act of playing for stakes in the hope of winning” (ID 00430140-N), “verbal wit or mockery” (ID 06780882-N) and “a deliberate coordinated movement requiring dexterity and skill” (ID 00556313-N), etc. Therefore, the word “play” can belong to multiple synsets. The format of a synset and the similarity distance between concepts are very important to our research since each synset represents a concept. In our approach, we use the most frequent synset to represent a word.

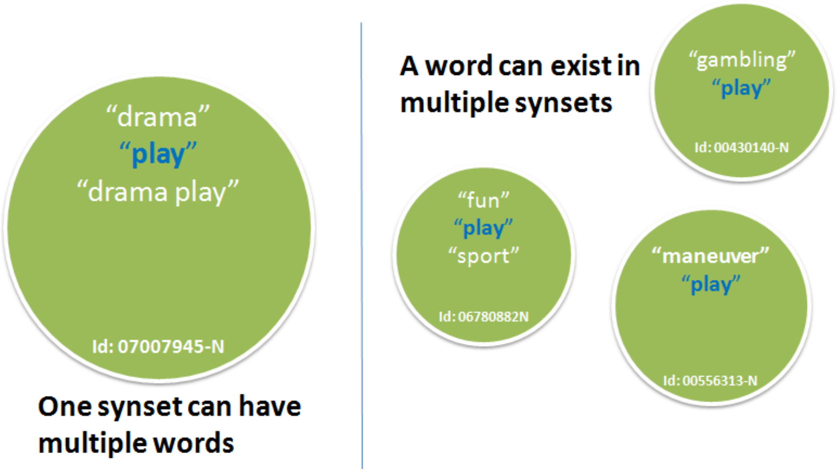


Figure 4.1 Use word "play" to show the example that one synset can have multiple words and one word can exist in multiple synsets

Synsets are further organized into generalization (hypernyms) and specialization (hyponyms) hierarchies. These hierarchical organizations bring out the concepts of likelihood and other relationships among the words. For example, the word "rose" implies the concept of "flower" with some likelihood, and vice versa. That is to say if there is a user use the word "rose" in his tweets, then he or she probably talking about topics related to flowers. We will do this by computing the semantic similarity/distance measures between concepts to understand the relationships between them.

4.2 Semantic distance using WordNet

In this section, we shall describe a method we have developed to measure semantic distances/similarities between two synsets in the WordNet hierarchy. This measure is not a metric due to it not being symmetric. In addition, WordNet has specific generalizations and relationships that are associated with the synsets. In this section, we compare five semantic algorithms for computing the similarities between synsets in order to choose the one that is best for our approach.

4.2.1 The algorithms

Leacock-Chodorow [30]**:**

Leacock and Chodorow's idea utilizes the length of the shortest path between two synsets for measuring similarity. However, it only focuses on the IS-A links and scales the path length by the overall depth, D, of the taxonomy.

$$\mathrm{sim}_{\mathrm{LC}}(\mathrm{s}_1, \mathrm{s}_2) = -\log \frac{\mathrm{len}(\mathrm{s}_1, \mathrm{s}_2)}{2\mathrm{D}} \quad < 1 >$$

Where $\mathrm{len}(\mathrm{s}_1, \mathrm{s}_2)$ is the shortest path between two synsets $\mathrm{s}_1$and s_2. D is the overall depth of the nouns dataset.

Resnik [31]**:**

Resnik's approach was the first to bring ontology and corpus together. It is important that the similarity between two concepts might be influenced by "the extent to which they share information". Similarity between two concepts, as defined in Resnik's algorithm, is closely related to their lowest super-ordinate, defined as lso (s1, s2).

$$\mathrm{sim}_{\mathrm{R}}(\mathrm{s}_1, \mathrm{s}_2) = -\log \mathrm{p}\big(\mathrm{lso}(\mathrm{s}_1, \mathrm{s}_2)\big) \quad < 2 >$$

Where p(s) is the probability of encountering an instance of a synset s in some specific corpus. We will discuss the corpus that we use for our work in a later section.

Jiang_Conrath [32]**:**

Jiang and Conrath's algorithm also uses a similar notion of information content, but the form of the conditional probability is different. It reflects encountering a child-synset instance by a given parent-synset instance. Hence, not only does the lowest super-ordinate play a role in this algorithm, but the two synsets themselves both play roles as well. Note that the algorithm gives out the semantic distance, which is the inverse of the similarity.

$$\mathrm{dist_{JC}}(\mathrm{s_1}, \mathrm{s_2}) = 2\log \mathrm{p}(\mathrm{lso}(\mathrm{s_1}, \mathrm{s_2})) - (\log \mathrm{p}(\mathrm{s_1}) + \log \mathrm{p}(\mathrm{s_2})) \quad < 3 >$$

In addition, p(c) is the probability of encountering an instance of a synset c in some specific corpus.

Lin [33]**:**

Lin’s similarity measure follows from his theory of similarity between arbitrary objects. It looks similar as Jiang-Conrath's algorithm but in different fashion, as follows:

$$\mathrm{sim_{Lin}}(\mathrm{s_1}, \mathrm{s_2}) = \frac{2\log \mathrm{p}(\mathrm{lso}(\mathrm{s_1}, \mathrm{s_2}))}{\log \mathrm{p}(\mathrm{s_1}) + \log \mathrm{p}(\mathrm{s_2})} \quad < 4 >$$

Moreover, Lin's algorithm gives the concept similarity as the outcome, not the semantic distance.

Note that Resnik's, Jiang-Conrath's, and Lin’s algorithms require a probability function p(c) which is the probability of encountering an instance of a synset in some specific corpus. In our approach, the “specific corpus” refers to a database of tweets, which were retrieved using the Twitter API. This data consists of tweets that belong to a set of seed users, the friends of these seed users (immediate social network), and these neighbors’ friends (extended social network)

Shevade [34]**:**

In Shevade’s algorithm, a synset’s parent and children play important roles in the similarity calculation. This algorithm focuses on the WordNet hierarchy by calculating it recursively. For example, there is a synset alpha in WordNet. If synset alpha is not a root or a leaf synset in the WordNet hierarchy, it always has two kinds of directly related synsets: the parent (hypernyms) synset and the children (hyponyms) synsets. Each kind of synset implies a concept with a different weight – w_1 (the parent) and w_2 (all of the children). Hence, the implication that synset alpha is true by given another synset beta is true is computed as <5>. It can be understood as follows: in the WordNet hierarchy, a synset alpha is given, algorithm computes the possibility to find a synset beta start from synset a by recursively checking alpha's parent and children synsets.

$$\mathrm{I(aplha \rightarrow beta)} = w_1 \, \mathrm{I(parent(alpha) \rightarrow beta)} + \frac{w_2}{k} \sum_{i=1}^{k} \mathrm{I}\,(c_i \rightarrow \mathrm{beta}) \quad < 5 >$$

$$\mathrm{children(alpha)} = \{c_1, c_2, \dots\}$$

$$\mathrm{I(beta \rightarrow beta)} = 1$$

$$w_1 + w_2 = 1$$

I is the implication strength, k is the number of children that synset alpha has and c_i is the *i-th* child of synset alpha. Moreover, w_1 and w_2 are the weights for the parent and children respectively. If synset alpha is the root node or leaf node in the WordNet hierarchy, then:

$$\mathrm{I(alpha \rightarrow beta)} = 1 \quad \text{if} \;\; \mathrm{alpha = beta}$$

$$\mathrm{I(alpha \rightarrow beta)} = 0 \quad \text{if} \;\; \mathrm{alpha \neq beta}$$

The distance between the two synsets is $1 - \mathrm{I(alpha \rightarrow beta)}$.

4.2.2 Experimental Setup

We needed to select the best semantic similarity calculation algorithm in order to find relevant people. Therefore, we performed the following experiment to evaluate the algorithms, which described in the previous sections, in order to select the one that best fits our work.

Data:

We chose 355 users, which included the seed users and their direct friends, as the experimental user group. The experiment used those users' synset distribution, which contains the synset records extracted from their historical tweets, to test the algorithms. The ground truth was manually constructed by carefully analyzing and picking the relevant accounts. The ground truth for each query word in the experiment is different.

There were 50 words used as queries for testing the algorithms in this experiment. First, we collected the 1000 most frequent words among the tweets of the users. We defined the distance between two words as the path length of their related synsets in WordNet. This resulted in a 1000×1000 distance matrix. Sixty-five of these words needed to be removed since their related synsets are single nodes in WordNet (for example, "David"). As a result, the size of the distance matrix used for clustering was 935×935. We then used a hierarchical clustering method [35] (average linkage clustering) to aggregate the 935 words into 50 clusters. The distance between two clusters is defined as the average of distances between all pairs of words, where each pair comprised of one word from each cluster. After clustering, we picked the most frequent word in each cluster as the queries for our experiment.

For each algorithm, we used different scope value to evaluate the performances. The scope value was the counted number of the result an algorithm returns. For example, scope 20 means we only count the top 20 users to see if they are relevant.

Results:

For each algorithm, we ran the experiments with scope values set as 10, 20,40,60,80, and 100. We calculated the precision and recall for each algorithm with each query word in each scope value. Therefore, there are 50 queries × 6 scope values × 5 algorithms = 1500 pairs of precision and recall results. Precision and recall measures are defined as:

$$\text{Precision} = \frac{\text{User}_{\text{ret}} \cap \text{User}_{\text{rel}}}{\text{User}_{\text{ret}}} \quad < 6 >$$

$$\text{Recall} = \frac{\text{User}_{\text{ret}} \cap \text{User}_{\text{rel}}}{\text{User}_{\text{rel}}} \quad < 7 >$$

Where User_{ret} is the number of retrieved users and User_{rel} is the number of relevant users. We use the F-measure to evaluate the performance of the algorithms:

$$\text{F}_{\text{measure}} = \frac{2\text{PR}}{\text{P} + \text{R}} \quad (\text{P} = \text{Precision}, \text{R} = \text{Recall}) \quad < 8 >$$

Figure 4.1 shows the experiment results by plotting all the five algorithms' F-measure values with different scopes. The five algorithms' overall F-measures are averaged by the number of queries. Clearly, Shevade's algorithm performs much better than the other algorithms. Therefore, we decide to pick Shevade's algorithm into our relevant people finding methods as the WordNet based method.

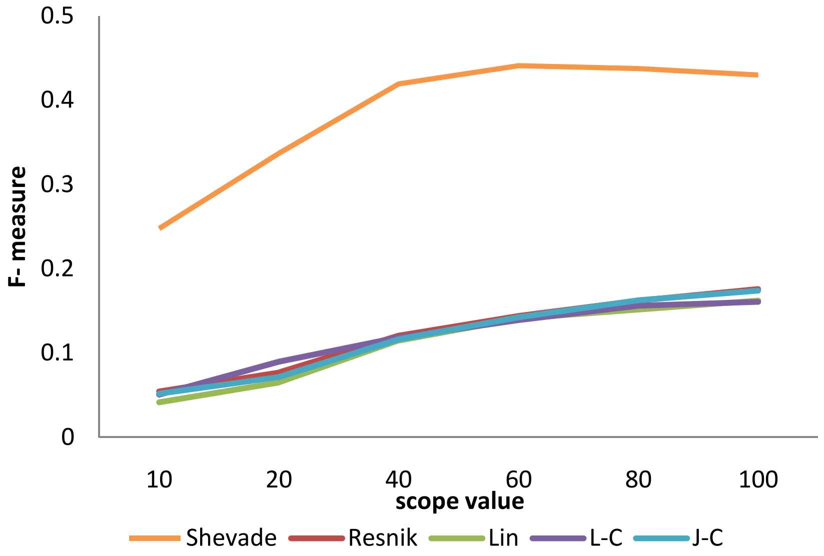


Figure 4.2 F-measure comparison for different algorithms: Shevade for Shevade's algorithm, L-C for Leacock-Chodorow's algorithm, Resnik for Resnik's algorithm, J-C for Jiang-Conrath's algorithm, Lin for Lin's algorithm

4.3 Different query methods

We introduce several query methods for querying relevant users in this section. Each method has its own emphasis when performing the people searching process, such as social graph information and interpersonal similarities.

4.3.1 The Baseline method

In our system, the baseline algorithm is the default Twitter search function, the same one used on Twitter's Homepage. Our system uses a built-in Twitter Java API to call the Twitter search function. Twitter's search function performs simple word matching by searching for the latest tweets of random users. The returned results are not ranked since they are sorted by their post time. Additionally, the authors of the tweet results can be anyone in the Twitter universe and is not limited to the querying user's personal social network.

We use Twitter to be the baseline method to compare against our methods. The results returned from the baseline method will be the users who have just posted tweets and their tweets contain the queries. Therefore, it is possible to see that the returned results have the same content if many people are tweeting the same sentence at the same time, which may be the case with some slogans. Then the returned result might be all the Twitter users who tweets the slogan. This result may be end up being useless. Although the baseline method has these disadvantages, it is useful to compare our methods against it. It is also interesting to see if users may think the results of baseline method are relevant.

4.3.2 WordNet based method

We utilized WordNet to analyze the semantic meaning of users' tweets. From the experiment introduced in section 4.2.2, we compared and evaluated five semantic similarity algorithms for semantic analysis in WordNet. According to the results, Shevade's algorithm performed best. We utilize this algorithm to the WordNet based query method. This method takes the query word as its only input. First, our method translates the query word into its related synset. Each account in the user's direct social network and extended social network has a synset set, which contains all the synset extracted from their historical tweets. Then each account in the user’s social network receives similarity score against the user who performed the query. The score is calculated by measuring the overall similarity between the query synset and the specific user's synset set. Finally, the algorithm normalizes the score based on the size of the synset set. The pseudo code for this process is shown below along with the process of calculating a user's similarity score (Figure 4.3).

```
for (user in social_network){
    for (synset in synsetlist){
    score += similarity(query, synset);
    }
  score = normalize(score, synset.size);
  result.add(user,score)
}
sort(result)
```

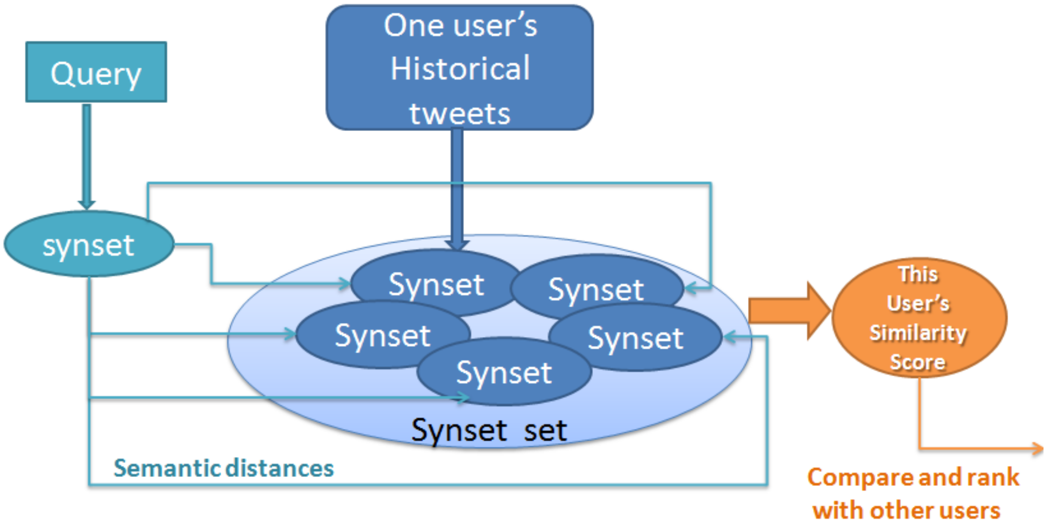

Query
One user's Historical tweets
synset
Synset
Synset
Synset
Synset
Synset
Synset set
Semantic distances
This User's Similarity Score
Compare and rank with other users

Figure 4.3 Process of calculating one user's similarity score by using WordNet based method

If there is more than one word in the query, such as a phrase, the system will separate the query into several words. Each word is used as input in the same method and the final result is the average of the scores for each individual word.

There is a reasonable limitation for this method; it cannot search for any word. WordNet does not contain any words that are considered slang or are not part of the English dictionary. For example, many abbreviations and special words like "LOL" or ": D" cannot be found in WordNet and our system will give a "word not found" error message. On the other hand, we feel it meaningless to search for these types of words because individuals are not interested in others who are experts in "ROFL."

4.3.3 WordNet + Social Graph

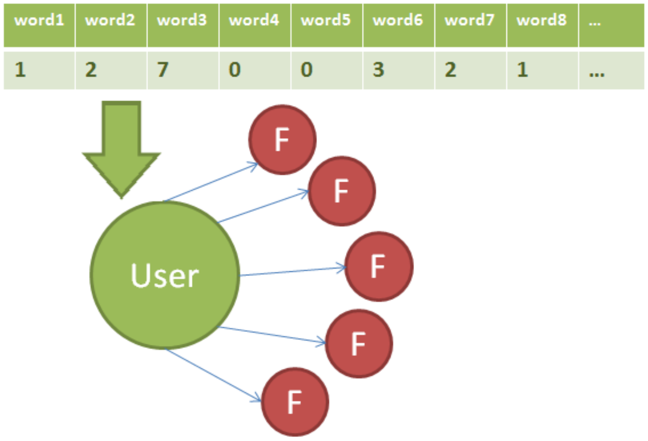
Only considering the query word by itself may be limited. Therefore, we incorporate other factors into our methods for finding relevant people. The social network topology is one of the factors that can be used. The PageRank algorithm [36], used by Google's Internet search engine, assigns a numerical weighting to each element of a hyper-linked set of documents with the purpose of measuring its relative importance within the set.

We consider the Twitter social network as a directed graph that can be used in the PageRank algorithm. We consider every user in our database a node, and relationships between friends as directional edges. There are no weights on any of the edges. PageRank is a probability distribution, which represents the likelihood that a person will randomly click on links and arrive at any particular page. A 0.5 PageRank value implies that there is a 50% chance that a person clicking on a random link will be directed to the document with a 0.5 PageRank. In our approach, the 0.5 social PageRank for one user implies that this user has a 50% chances to be the friend of a randomly chosen user in our database.

In our approach, the friendship between two users is considered as one edge where the direction is from the user's friend to the user. This means that the seed users have higher PageRank values because they are more important in the social graph than those leaf users. Providing the users who have more friends or have a higher possibility to find more friends is meaningful. In this method, we care about which user has more in-links in the graph since they have more contacts to search through. We have a full map of all the users and their relationships and we use this social network to calculate each node's social PageRank value. Therefore, there are two ranked lists of the users after every query with, one derived from the WordNet based method and one derived from the PageRank values. We set equal weights onto these two ranked lists in order to combine the results. The final ranks are the average of the two lists.

4.3.4 WordNet + Profile

The social network is useful in terms of PageRank, but it may be biased by its topology as well. This is because users may dislike or are ambivalent towards the one who can provide more contacts. Therefore, we explored another possible option, which is person-person similarity. We refer to it as a "profile" here because we want to compile a user's information as a set of features, such as user interests, time zone, and location information. However, the location and time zone information may not be accurate and many users do not provide location information on Twitter. Therefore, the only useful feature left is the people-people similarity. The query method based on WordNet only considers the query word itself. The result of the WordNet based method remains the same if the same query word is used regardless of the user. To address this issue, we take the users' historical tweet data to compute the similarity between users.



word1	word2	word3	word4	word5	word6	word7	word8	...
1	2	7	0	0	3	2	1	...

.

Figure 4.4 Example for composition of a vector and the scope of the people-people similarity

We utilized the vector space based method used in the documents similarity calculation. The contacts in the user's social networks and the user himself are considered as vectors. The dimensions of the vectors correspond to all of the words that were used. We use the term count model for the tf-idf [37] weights because the global parameter is not fixed, and changes between users. We use the cosine similarity to compute the people-people similarity.

$$\mathrm{similarity} = \cos(\mathrm{a},\mathrm{b}) = \frac{\mathrm{a} \cdot \mathrm{b}}{||\mathrm{a}||\,||\mathrm{b}||} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{a_i b_i}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{a_i}^2}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{b_i}^2}} \quad < 8 >$$

Each user has a list of similarity values with the contacts in his direct and extended social network. The returned results will be the combination of the result obtained from the WordNet based method and the result obtained from the profile similarity method. We once again use the same weights for these two lists.

4.3.5 WordNet + Social + Profile

The last method combines all of the features that were discussed above. What we do is combine the social PageRank, profile similarity, and WordNet based methods. We thought it would be interesting to see if the combination of these three features performs better than any individual feature. Perhaps it would perform worse because due to the good part being diluted.

4.4 Social value

A "social value" is the outcome of our research work. The social value is a measure for understanding which direct friends (neighbors in the social network) are more valuable to the user when searching for relevant people based on a certain query. The user's neighbors who contribute more and have higher-ranking accounts in the results are more useful to the user, because they provide the most Twitter users in the returned result that are relevant. These neighbors may hold the more important interpersonal ties, for instance, weak ties.

Our system provides the friends with the five highest social values from the user's social networks. The system will give relevant users based on the query word and query method. The user may not have a “follower” relationship with some of the returned users, which means that the user may not know who these users are. However, these unknown accounts have a “follower” relationship with at least one of the user's neighbors. In other words, the user’s neighbors help provide individuals that may be relevant to the user. There are two preconditions for the friends with the highest social values: 1) they are direct friends of the user and 2) they have higher, normalized average rank values calculated from the returned relevant people result provided by the system. Although the system only provides the top ten users as the relevant accounts, the friends with the highest social values are evaluated from the top 100 returned relevant accounts. The normalized average rank values for each friend is computed as:

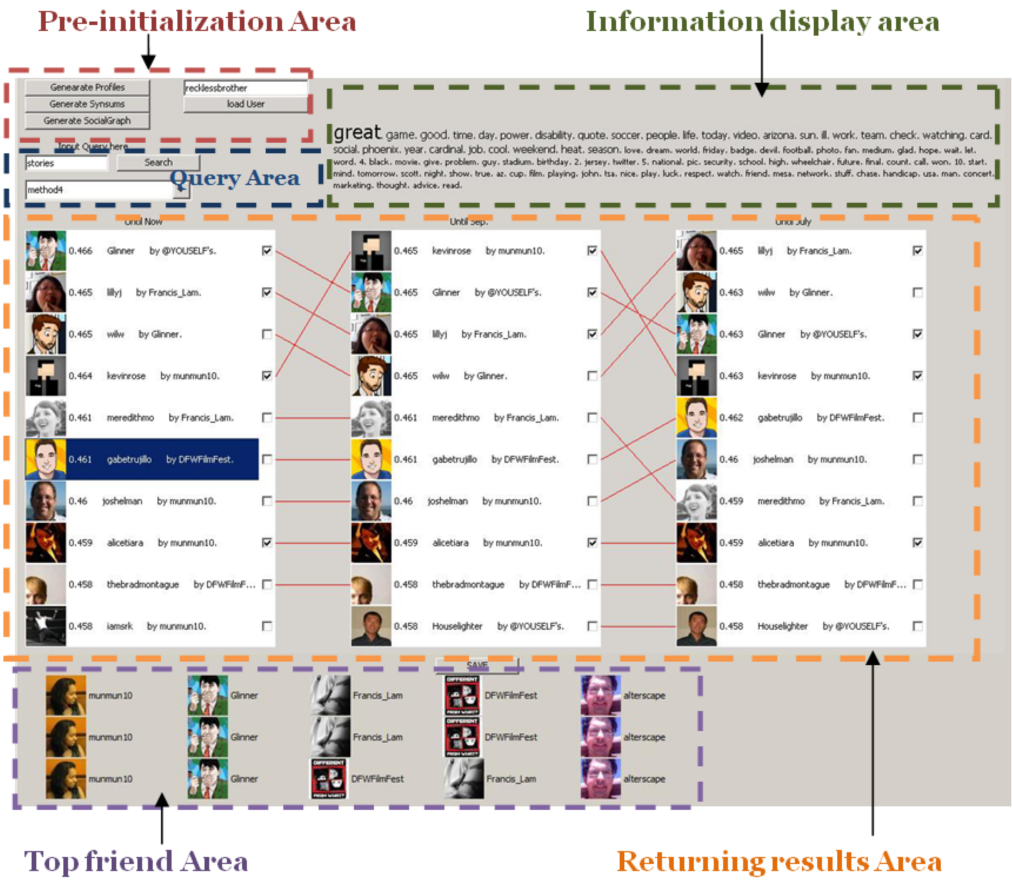
$$\mathrm{NAR}_{\mathrm{friend}_{\mathrm{f}}} = \frac{1}{\mathrm{NN}_{\mathrm{R}}}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}_{\mathrm{R}}} \mathrm{R}_{\mathrm{i}} - \frac{\mathrm{N}_{\mathrm{R}}(\mathrm{N}_{\mathrm{R}}+1)}{2}\right) \quad < 9 >$$

Where R_{i} is the rank in which the *i-th* user is from the friend, f. N is the total number of users that are returned which is 100 for our work because we count the top hundred returning accounts. N_{R} is the number of users that the friend, f, has provided. Performance perfect result would be zero for this measure and will approach one as performance worsens.

One potential issue with our approach is that ranking users by social value may not be the best way to inform an individual that which of the friends are valuable to him or her. It is possible that Twitter users do not consider their friends' values by ranking them by their importance. In other words, people may not evaluate which friend is the most valuable by comparing their ranking. In our user study, we attempt to find out if people would like to use these rankings in order to evaluate their valued friends. This will be discussed in chapter 5.

4.5 Interface

We built a functional user interface to help users understand the results from our method. The user interface also has a user-relevance-feedback mechanism in which the user can select which results were truly relevant. The application was developed in Java. We also used Twitter4J [38], which is an API that allows us to retrieve user thumbnails from Twitter. The interface has five main components: the pre-initialization area, the query area, the information display area, the results area, and the top friends area. Figure 4.2 shows the screenshot of the interface, which displays the results of one query.



recklessbrother

stories

method4

great game. good. time. day. power. disability. quote. soccer. people. life. today. video. arizona. sun. ill. work. team. check. watching. card. social. phoenix. year. cardinal. job. cool. weekend. heat. season. love. dream. world. friday. badge. devil. football. photo. fan. medium. glad. hope. wait. let. word. 4. black. movie. give. problem. guy. stadium. birthday. 2. jersey. twitter. 5. national. pic. security. school. high. wheelchair. future. final. count. call. won. 10. start. mind. tomorrow. scott. night. show. true. az. cup. film. playing. john. tsa. nice. play. luck. respect. watch. friend. mesa. network. stuff. chase. handicap. usa. man. concert. marketing. thought. advice. read.

Until Now				Until Sep.				Until July			
0.466	Glinner	by @YOUSELF's.	☑	0.465	kevinrose	by munmun10.	☑	0.465	lillyj	by Francis_Lam.	☑
0.465	lillyj	by Francis_Lam.	☑	0.465	Glinner	by @YOUSELF's.	☑	0.463	wilw	by Glinner.	□
0.465	wilw	by Glinner.	□	0.465	lillyj	by Francis_Lam.	☑	0.463	Glinner	by @YOUSELF's.	☑
0.464	kevinrose	by munmun10.	☑	0.465	wilw	by Glinner.	□	0.463	kevinrose	by munmun10.	☑
0.461	meredithmo	by Francis_Lam.	□	0.461	meredithmo	by Francis_Lam.	□	0.462	gabetrujillo	by DFWFilmFest.	□
0.461	gabetrujillo	by DFWFilmFest.	□	0.461	gabetrujillo	by DFWFilmFest.	□	0.46	joshelman	by munmun10.	□
0.46	joshelman	by munmun10.	□	0.46	joshelman	by munmun10.	□	0.459	meredithmo	by Francis_Lam.	□
0.459	alicetiara	by munmun10.	☑	0.459	alicetiara	by munmun10.	☑	0.459	alicetiara	by munmun10.	☑
0.458	thebradmontague	by DFWFilmF…	□	0.458	thebradmontague	by DFWFilmF…	□	0.458	thebradmontague	by DFWFilmF…	□
0.458	iamsrk	by munmun10.	□	0.458	Houselighter	by @YOUSELF's.	□	0.458	Houselighter	by @YOUSELF's.	□

munmun10	Glinner	Francis_Lam	DFWFilmFest	alterscape
munmun10	Glinner	Francis_Lam	DFWFilmFest	alterscape
munmun10	Glinner	DFWFilmFest	Francis_Lam	alterscape

Figure 4.5 Interface of the People Finding system

The pre-initialization area is in the upper left side of the interface. The buttons in this area are used for generating and loading necessary data files for first-time users (i.e. these files contain the user’s social graph information) and loading the user's information.

The results area consists of three columned lists, which are the three results at three different time granularities. The first column represents the results got from the data history until now (when using the system), the second one indicates the results got from the data history until one months before the previous column’s timescale, and the third column uses the data until two months before the previous column’s timescale. These red lines provide a visual indicator to the user, which enables him or her to quickly show how the ranking evolved over each time period. The numbers in front of each Twitter screen name are the scores computed from the current method. The users in each column are ordered by their rank. The only exception is the result of the baseline method. In this case, the tweets are ordered by time and only the screen name and tweets are displayed. User selects the check boxes behind each account if he or she considers the Twitter user as relevant.

The result items in the columns are interactive. Each Twitter user in the columns can be selected by single clicking, and the information display area will show a tag cloud of that user for that time period. The fonts of the different tags change based on the frequencies of the words. Each Twitter user in the list can be double-clicked as well which will result in their personal Twitter homepage to be displayed in an external web browser. These functions help make the interface more convenient and helpful. These functions are designed for those who may want to check a Twitter user's latest tweets or personal information to better inform their decision when determining if that user is relevant or not.

The top friends area will show the top five friends who contribute most of the returned results based on the current query's words. There are three rows of top friends based on the different timescales.

5 USER STUDY AND EXPERIMENTS

5.1 Introduction

We designed a user study for our work for evaluating our algorithm and method. The goal of this research is to provide methods for users to find relevant people based on their queries. It is a new query method for specific users where the performance is judged subjectively. As a result, we designed the user study to help us understanding more about users' Twitter using experiences and their feedback about our work. Their responses can help us to improve our algorithms, interfaces and other aspects of the work.

The demographics of the participants are the following. There were 10 participants in total. They were graduate students and faculties at Arizona State University. Their majors varied, including Computer Science, Electrical Engineering, Arts Media, and Industrial Design. There were 5 male and 5 female participants, demonstrating substantial gender diversity. The median age of the participants was larger than 18 and less than 30. Meanwhile, 5 participants were native English speakers and 5 were non-native English speakers, making our user study culturally diverse. These participants are moderately to highly active on Twitter. However, the scope of the participants was limited. It only concluded students and faculties in a university and ten participants was still a small sample. We will discuss the limitations more in section 5.3.4.

In this chapter, we shall first present a brief introduction of our designed user study, which includes a questionnaire and a system evaluation. We shall also introduce the motivation and goal of the user study. In section 5.3, we will present the results of the user study and the data analysis of the participants' system evaluation, including tables, figures, and discussions.

5.2 Design of user study

The Office of Research Integrity and Assurance of Arizona State University approved the user study. All participants had signed and given us their permissions to use their data and responses. The user study was anonymous and voluntary. No participant personal information was released.

The user study contained three main parts: one questionnaire about prior Twitter experience, attitudes about social-value users, and testing/evaluation for the system.

The questionnaire for Twitter using experiences was composed of both multiple choice and interview-style questions. The questionnaire studied aspects of users' Twitter use habits or personal routines on Twitter. We shall pick some sample questions and introduce what information we expect to gather from them. For example, one question read "how frequently do you tweet on the Twitter on average?". The choices varied from "more than twice a day" to "less than once a month". This question helped us to analyze the tweeting frequencies of the users, which can provide a guideline for adjusting the timescale setting of the system. Other questions were designed to gather information about users' motivation for following an account on Twitter; for instance, "when do you decide to follow someone on Twitter?". The options included "recommended by TweetDeck", "using search engine", "friends' friends" and so on; the user specified answer is also available. Some questions concerned the social graph composition of the participants. Questions asked what kind of people are the most common in the participant's friend list on Twitter. This question allowed us to discover the distribution of the participants' friends on Twitter.

Questionnaires also asked participants' attitudes regarding the social value of their Twitter friends. During the user study, we ask the participants to list the five most valuable Twitter friends in their own social network. Further questions seek their options of the top five social-value friends provided by our system: based on the top social-value user provided by the system, the users are requested to answer some "social capital" generation style questions. The advantage of such implicit feedback is that the user is not mentally biased when thinking about explicit ranking of their ego-network; rather, they evaluate the "social value" in more realistic and qualitative terms. There also exists the possibility that people do not usually rank their friends. We want to find out how the participants think about their social-value friends and how they evaluate friends' values.

The second part of the user study asked participants to use the system. There were five fixed queries and five personally specific queries for each participant. Participants entered a query and saw the returned lists of relevant Twitter accounts, then were asked to determine which of them were relevant. Supportive data provides to the participants were similarity scores, word tags of the account's tweeting history and each account's personal Twitter homepage. For each query, participants used only one of the possible methods and gave their feedback, as it would take too much time for the participants to finish 10 queries multiplied by 5 methods. As a result, only one method was used during the user study; the other four methods, on the other hand, ran in the background. The experimental result analysis would use the data gathered from all methods. The relevant accounts selected by the participants may also be found in the results generated by other methods. The participant did not know which method they were using when we did the queries, as this knowledge might bias their judgments of the personal relevance of the returned accounts. The users were randomly numbered and all of them had an ordered sequence of methods to use, making every query and every method equally likely to be used. For example, if user's first query used method one, he would run the rest of the queries in sequence of method 2, method 3, method 4, and method 5. The second user would do the first query by using method 2, and ran the rest by using method 3, 4, 5, and 1 in order.

5.3 Results

5.3.1 Result of questionnaire

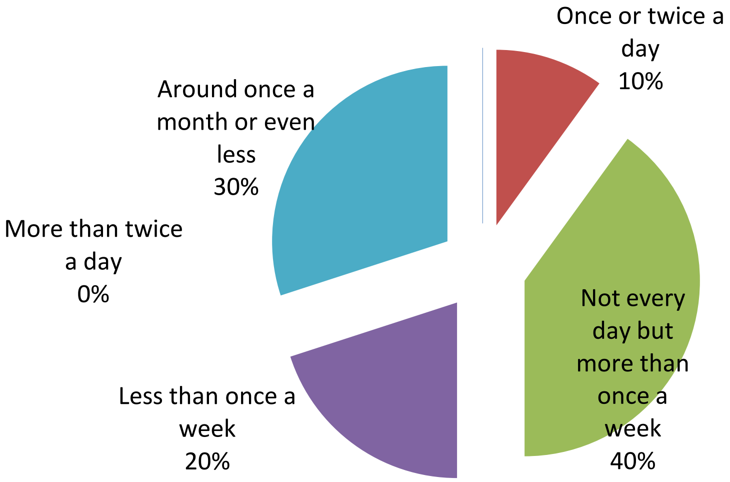


Figure 5.1 Pie chart of Twitter usage frequency results

Figure 5.1 shows the result of participants' Twitter usage frequencies. According to the questionnaire answers, not everyone tweeted very often; about half of the participants tweeted less than once every week. Some participants mentioned this during the user study, pointing out that Twitter is useful, but not a daily chatting tool for them. They generally used Gtalk, MSN or other similar

IM software for their instant online interactions. E-mails were also very popular, as they are more private and safe. Nevertheless, we are curious to see what the participants are talking about on Twitter; Figure 5.2 shows the histogram representing the subject of participant's tweeting.

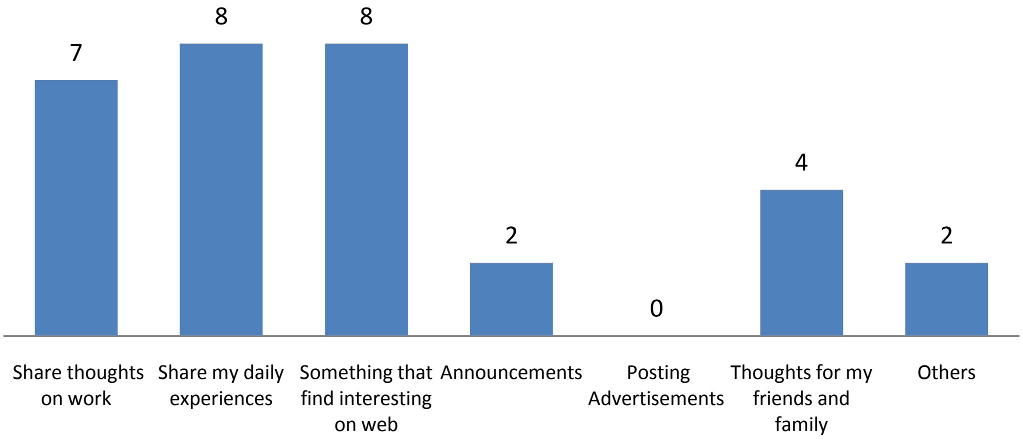


Figure 5.2 Histogram of the question for what the participants tweet about

We find that there are three main usages for the participants: sharing interesting stuff, sharing thoughts on work, and sharing their life experiences. They were not especially interested in public announcements, and no one wanted to post advertisements. Two participants elected to give their own answers: "Share my awesome insights" and "networking". We can categorize the first one as part of "sharing thoughts", and the second one as "connect with friends or family". Hence, from this statistic we can state that participants are using Twitter as a tool for sharing interesting things and interacting with their friends. To confirm the point above, the next question asks the reason that the participants use Twitter. Figure 5.3 shows the distribution of answers to this question.

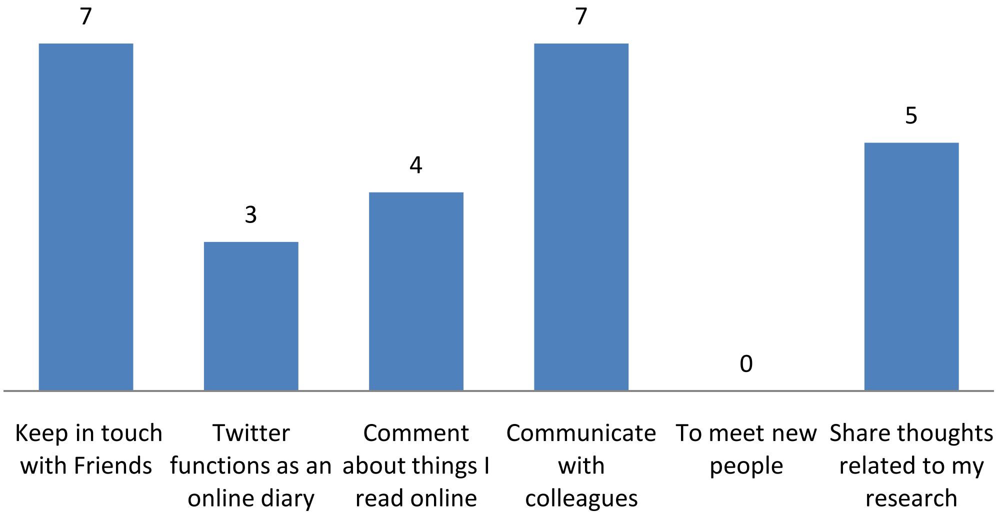


Figure 5.3 Histogram of the question regarding the motivation for using Twitter

The results show that keeping in contact with friends and communicating with colleagues were the highest two. Around half of the participants also chose to use Twitter as a diary or comment board; both are similar ways as sharing information. However, no one selected “meet new people”, which came as a surprise. This option showed that participants were not quite comfortable with using Twitter as a tool for meeting new people. According to participants' comments, it is hard to follow strangers on Twitter because there is not enough information provided about them. This is a good sign for our work, since our purpose is to fill this blank and locate people who are trustworthy relevant to the Twitter users. Therefore, we want to know when participants decide to follow a new account. Figure 5.4 shows their answers.

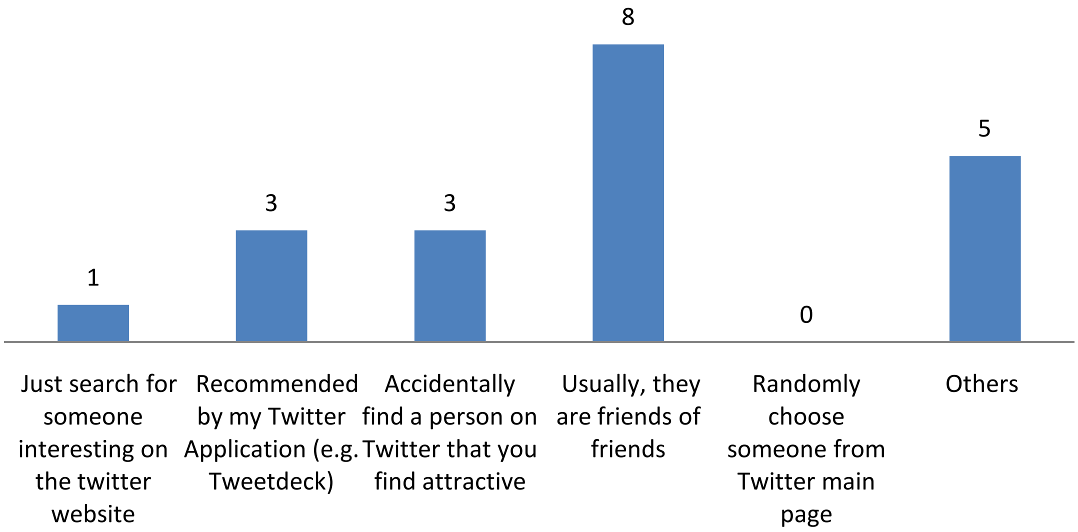


Figure 5.4 Histogram of the question for when user will follow others

We can clearly see that no one chose to follow a new account by randomly picking someone from the Twitter home page, and only one person would search on Twitter for his or her interests. Three people accepted the recommendations from Twitter applications. However, the most frequently chosen reason for accepting a friendship is still through a friend's friend. It proves that trust between people can be transmitted through social connections. Our system applies this idea to narrow down the searching population to egocentric social networks rather than the entire Twitter universe.

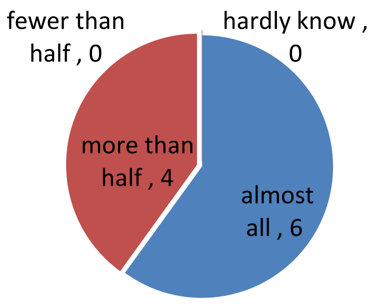


Figure 5.5 Pie chart of the question for how many of the participant's followers are known in real life

Figure 5.5 shows how many following accounts in the participant's Twitter network are also real life friends. It seems that the participants still rely primarily on their real life friendship. The result indicates us that all of the participants know more than half of their Twitter friends in real life. Figure 5.6 shows that 90% of the participants indicates that their friends in real life are their main source of the Twitter friends.

Results displayed in figure 5.5 and figure 5.6 may be limited. Our participants are all college students and faculty, who may prefer personal friends than celebrities since they spend most of their time on research work and interact with the colleagues. We believe that celebrities and public organizations' accounts will have greater representation with general Twitter users.

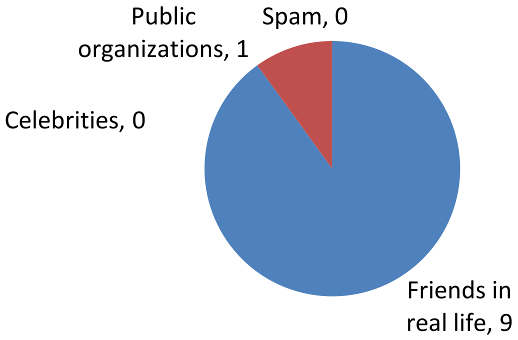


Figure 5.6 Pie chart of the question for main composition of participants' Twitter friends

The last question regarding Twitter use is what people may search for on Twitter. People, products, and events are all roughly equally represented. The participants were eager to have better searching functions on Twitter, with 100% of them indicating that they would use a service to search for their needs on Twitter, if there is one.

All the results from the questionnaires indicate that participants care more about Twitter accounts associate with real-life friends. They care about their own social networks more than other factors. It is possible that the participants intentionally constrain their Twitter social network to their daily friend group, where all the people share similar or related backgrounds and interests. Finding potentially useful accounts that users may not be directly aware of was identified as very useful. Finding which of their direct friends can lead them to these potentially useful accounts were also determined to be useful. We will discuss the social value related user study results later in section 5.3.3.

5.3.2 System testing and evaluation

Each participant used ten queries in the people search; five of the queries are standard queries and the other five are user-specific queries. We picked the five standard queries from among the top 100 frequently used words, which are gathered from the tweets of seed users and their direct friends by counting the number of occurrence of the words. The five standard words were "fun", "movie", "research", "job", and "travel". The author, the advisor, and another co-worker discussed and selected these words; none of these three people was included in the user study. The participants only used one method to do each query, with the methods' names hidden from the participants to avoid biasing their judgments. The other four methods ran in background. With ten users and ten queries per user, five different methods for searching people and three different time periods, there are 10 × 10 × 5 × 3 = 1500 results for the evaluation.

The baseline method, which is method 1 in the user study, had no rank relationships between the returned results since its results were listed by time. Therefore, we cannot use normalized average rank value to compare it with the other four methods. Instead, we use the average number of relevant accounts selected by participants, to make the comparison. During the user study, the baseline method also returned Korean, Japanese, and Spanish tweets as results because it only returns the latest tweets that contain the query keywords. The results may vary at different times, even one second after the previous query. Some participants did not select any of the returned accounts from baseline method as relevant as they could not find any clue for the accounts to be relevant. Figure 5.7 presents the results, the values are the average number of relevant accounts picked out of 10 for each method.

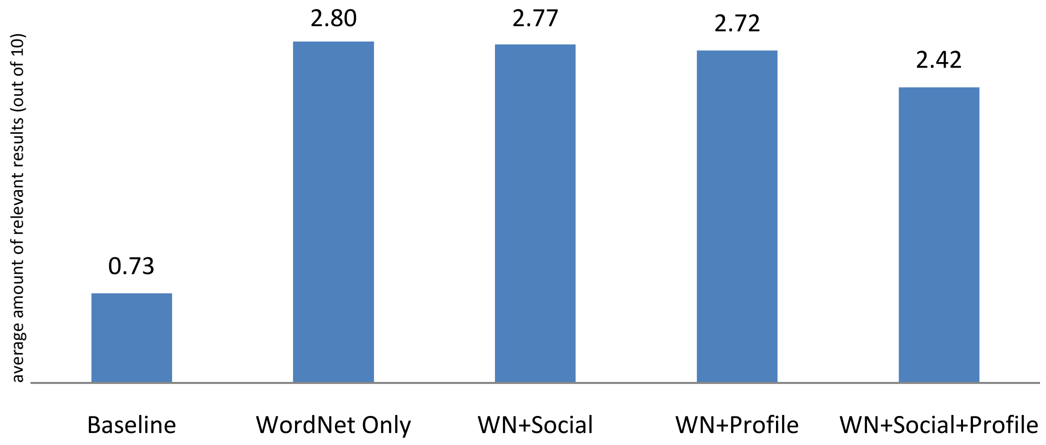


Figure 5.7 Histogram for average relevant count for the five methods; y-axis is the average amount of relevant results the user selected from choices of ten.

We can clearly see it from Figure 5.7 that baseline method is worse than the other four methods, from the low average relevant count. Next, we will compare the four methods to see which one is the participants' favorite method. A participant p has one query q by using method m; system will return three lists of accounts according to timescale t1, t2, and t3. The participant will select which accounts are relevant from the three lists. The three lists contain the relevant accounts with their rankings in the list.

We use the normalized average rank value to compare the performances of the methods. The normalized average rank value for a user u, query q and method m in timescale $\mathrm{t_i}$ will be:

$$\mathrm{NAR}_{\mathrm{user_u query_q method_m t_i}} = \frac{1}{\mathrm{NN_R}}\left(\sum_{i=1}^{\mathrm{N_R}} \mathrm{R_i} - \frac{\mathrm{N_R}(\mathrm{N_R}+1)}{2}\right) \quad < 10 >$$

Where $\mathrm{N_R}$ is the number of relevant accounts picked by user u from the results in the list of time $\mathrm{t_i}$ for query q by using method m. N is 10 for our approach. Then the normalized average rank value for the query q and method m for user u will be the mean of the three normalized average rank values for the three timescales.

$$\mathrm{NAR}_{\mathrm{user_u query_q method_m}} = \frac{\sum_{i=1}^{3} \mathrm{NAR}_{\mathrm{user_u query_q method_m t_i}}}{3} \quad < 11 >$$

The normalized average rank value for a user u using method m will be:

$$\mathrm{NAR}_{\mathrm{user_u method_m}} = \frac{\sum_{q=1}^{N} \mathrm{NAR}_{\mathrm{user_u query_q method_m}}}{N} \quad < 12 >$$

Where N is the number of the total queries user had used for the methods.

Finally, the normalized averaged rank value will be averaged over the number of the participants to determine the overall normalized average rank value for each method. Figure 5.8 shows the overall values. We find WordNet + Social method has the only value below 0.3, which is the best of the four methods. This result matches the results analysis of our previous Twitter use questionnaires. We summarize that the participants in our user study pay close attention to their social networks. They prefer to accept new friendships between their friend's friends and they like to interact with the people they know in real life. The WordNet + Social Graph method will rank higher for the accounts that have more friends and the accounts that are closer to the participant. This feature is exactly what the participants like. Figure 5.9 show the overall average normalized rank value by using standard queries only and using user specific queries only. The WordNet + Social Graph method still performs the best.

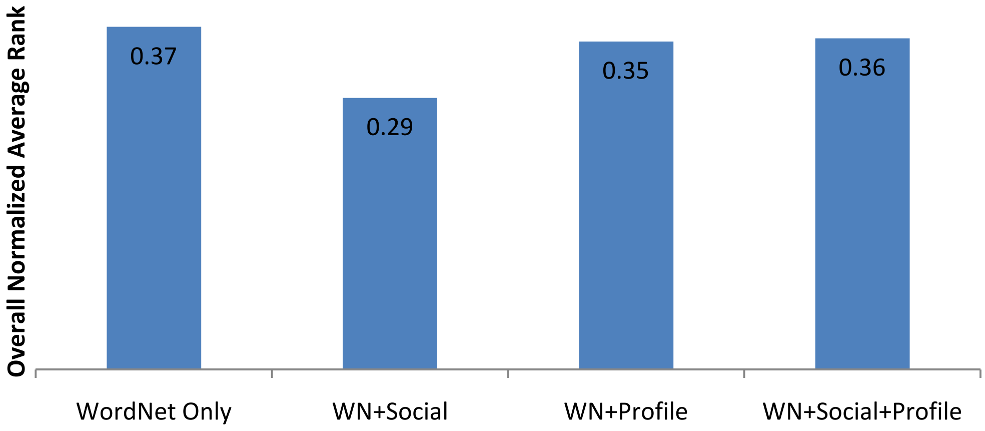


Figure 5.8 Overall normalized average rank values, average from all participants

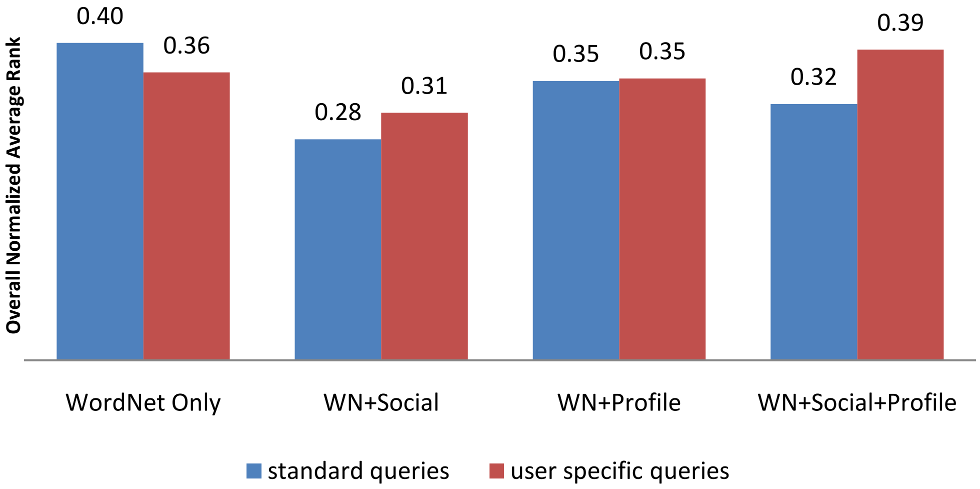


Figure 5.9 Overall normalized average rank values using standard queries and user specific queries

5.3.3 Social Value evaluation results

We compared the socially valuable friends given by the users and the socially valuable friends provided by the system to see if participants considered the result valuable. During the user study, participants were required to list the top five socially valuable friends from their Twitter network. We also collected, counted, and presented the overall top five socially valuable friends from all the different queries. We compared between these two "top 5" lists to see if there are overlaps. Table 5.1 shows the overlap results for the ten participants.

Table 5.1 Table for overlaps between top 5 most valuable friends given by participants and top five socially valuable friends provided by system

Participant 1	1
Participant 2	3
Participant 3	1
Participant 4	2
Participant 5	4
Participant 6	1
Participant 7	0
Participant 8	1
Participant 9	0
Participant 10	2

However, there are not as many overlaps as we expected, although participant 5 has four friends overlapped and participant 2 has three friends as well. We believe the reason is either that the top 5 friends provided by system are not valuable in participants' view, or that these friends are valuable but not considered to be in the top 5. That is to say, there may be no ranking relationships between the valuable friends.

We also compared the ranking provided by system and ranking re-ranked by the user for the top social-value friends. Table 5.2 shows the correlation between the ranking of the top 5 social-value friends provided by the system, and the ranking after participant's re-ranking. The rank correlation results are computed by using Spearman’s rank correlation coefficient [39]. Values closer to 1 indicate that the two rankings matches more closely while a value of -1 means the rankings are totally contradictory. The average rank correlation is 0.12, which means there is almost no correlation, which also recall the question that if the participants are not use rankings to evaluate their valuable friends.

Table 5.2 Table for correlation between system-provided ranking and participant re-ranking

	Rank Correlation
Participant 1	-0.3
Participant 2	0.9
Participant 3	-0.9
Participant 4	0.6
Participant 5	0.1
Participant 6	-0.3
Participant 7	0.1
Participant 8	0.5
Participant 9	0
Participant 10	0.5

To confirm whether participants evaluate their valuable friends by ranking or not, we provide the following questions from the user study. These questions are based on the top 5 social-value friends provided by the system. The first two questions ask the participants questions about “social capital” generation to determine if the friends are valuable. Question 1: Will you seek help from any of these users in case of emergency? Nine of the ten participants chose YES. Question 2: Do you think these users can help you in job-hunting? All the participants chose YES for this question.

These results provide meaningful evidence to suggest that the top friends provided by the system are also valuable to the participants. It could explain our statement that users may not use ranking to judge their friends' social value. Therefore, the most valuable friends are not strictly limited to the five friends provided by the participants, because the top friends provided by the system are still recognized as valuable, though there is not too much overlap with the participants' lists.

5.3.4 Limitations of the user study

Two main limitations for our user study need to be clarified: the scope and the number of the participants.

First, the participants in our user study are graduate students and faculty at Arizona State University. This factor could limit the range of questionnaire responses. For example, 90% of the participants felt that their friends in real life are the major component of their Twitter contact lists. If the participants came from every field of society, the number for celebrities or other figures might be higher. Furthermore, the trend that participants in our user study care more about their social network does not necessarily apply to the whole Twitter user group. It might only suggest that Twitter users who are also students and faculty in universities care more about their social network. If the participants are recruited from other areas, their priorities and standards for choosing relevant Twitter accounts might be very different, and the best-performed query method might change as well. As a result, the user study should widen its recruiting scope in the future work.

The second limitation is the number of participants. We only had 10 participants take part in our user study. This may constrain the generalizability of the user study's results. It is possible that the ten participants coincidentally have the same Twitter use habits and similar standards for choosing friends. For example, the result of the Twitter use frequency question shows that more than half of the participants tweet less than once a week. This cannot reasonably represent the overall tweet frequency of all the Twitter users. Since there are 75 million Twitter users and 35 million tweets per day in Feb 2010, the average tweet frequency is roughly calculated as 75/35 = once every 2.14 days, Hence, more participants are needed in future user studies.

5.3.5 Useful suggestions and facts from user study

During the user study, there were many other interesting facts and observations, which may prove very useful for our research and in future improvements.

It is interesting to observe how the participants select the relevant users. In our approach, we provide several options in the interface to help the users' decisions. They can see the tweets' word tags by single clicking the list item, and can see an account's personal homepage by double-clicking the list item. Among the ten participants, three of them read each word tag carefully, thought repeatedly and made careful selections. Four of the participant just took a glance at the word tags, but paid more attention to the account sources. They might judge it relevant or not by checking whether the source friend was relevant to the query. The rest of the participants combined those two habits together, taking full considerations to all the evidence, and made the choices. It surprised us to see that one participant checked the relevant boxes without reading the word tags.

We provide three times periods of the returned accounts but most of the users will keep their choices for all three lists. If user A is in one list, it will be checked for all lists in which A exists. It shows that most of the participants think of the user as a permanently relative item.

The selections of the user-specific queries are also intriguing. The chosen words generally reflected personal interests. One participant summarized the five user-specific queries, as “they are all of my life”. Some liked searching for more detailed concepts like “Korean restaurant” or “film camera” while some used generic words, like “food” and “games”. One participant pointed out that there should be priorities in multiple-word queries. It would be better if there were user preferences for words. One query for “Arizona photos” returned results more relevant to “Arizona”, not “photos”.

We received many suggestions during the user study for selecting the relevant users. One participant liked accounts representing small groups rather than individual people, but other users preferred the opposite. It would be better if they could choose the type of accounts they prefer to see.

Another interesting observation is that the participants were following some of the listed accounts during the user study. They said these experiences encouraged them to spend more time using Twitter and to follow more people. We were glad to see that our original goals for the system were met, increasing our confidence that our work is useful for Twitter users.

6 CONCLUSIONS AND FUTURE WORK

In this chapter, we shall first present a summary of the work described in this thesis. Then in section 6.2, we shall discuss some potential improvements inspired from the user study. In section 6.3, we shall conclude this chapter by discussing some future research directions.

6.1 Research summary

In this thesis, we explore a novel way for querying relevant people in online social networks, such as Twitter. The goal of the work is to help the Twitter user to find more relevant and valuable Twitter accounts based on their queries and their personal social network. It also helps them to understand their social networks by providing their most socially valuable friends. We implemented application interface and design user study to help Twitter users to evaluate our system and methods. The following are the key ideas in our approach:

- Developed a novel keyword-based system to find potentially relevant accounts on Twitter
- Developed a set of data crawlers to collect Twitter data
- Compared and evaluated several different WordNet similarity algorithms
- Compared several query methods to explore user awareness for choosing relevant accounts
- Discovered social values of direct friends in Twitter user's social network
- Implemented application interface and design user study for system evaluation
- Explore that participant users are more aware of their social network and friends in real life.

6.2 Improvements

In this section, we will discuss about the potential improvements. Almost all of these ideas are inspired from or directly suggested by the user study.

- Weighted multiple query

For multiple words queries, such as phrases, users may have their own preferences to set which word is more important. In our system, it is only set as equal weights. It will be better to provide related interface functions for changing the weights of the words by the users. For example, if a person would like to query “Arizona photo” and he or she want to get the results related to “Arizona” more. He or she may choose 80% of the importance to “Arizona” and leave 20% of the importance to “photo” by using the interface function to effect the search.

- Ambiguous words

An English word may have more than one meaning. For example, when a query word is “apple”, people may hardly know it means the fruit apple or the company apple if no further clue is provided. It happened in our user study as well, which influenced the accuracy of the algorithms. Fixing this problem can provide better semantic understanding. One possible solution is that users may provide more words to make it clear which meaning they want to express. Moreover, the system will remember and analyze the co-occurrences of the words. For example, if a user queries “apple” and “mac” together, it probably means apple the company. This requires more semantic analysis and trainings for the system to understand the possible combinations of the words.

- Account filtering

During the user study, some participants had very clear standards for choosing relevant users. For example, some did not like to follow public organization accounts. However, other participants announced that they like the community accounts better than individual accounts. One possible solution to filter accounts is checking the account's follower/following relationships ratios. Normally the celebrities are those who have fewer followers than the followings, same for the public accounts used by companies and groups. The spam accounts are the opposite. General users may have somehow equal numbers for followers and followings. However, this idea need more works to figure out some important questions, such as what is the threshold of the ratio to judge a Twitter account is an ordinary user but not a celebrity account and what the differences between celebrity and company accounts and so on.

- Better participant recruiting for user study

According to the limitations of our user study, the scope of the participants is limited to the graduate students and faculty in Arizona State University. The number of the participants is only ten. Participants in other areas of society and more participants are needed for better user study results and more accuracy evaluations for the algorithms and methods.

6.3 Future directions

From the literature reviews and the user studies we have, we can find the research work in the area of finding relevant people on social networks is not enough yet. We will highlight some possible potential research directions in this section.

1. Combine other social networks

Twitter itself may have constrains for its limited information. In the other hand, Digg.com [40] has clearer topic based categories for interests; Facebook has better complete personal information and more concentrated social networks.

It will be great to combine the different social network together to have more powerful database for query the relevant people. At the same time, privacy issue will remains as one potential problem for combining the different social network information.

2. Efficient algorithms for large scaled semantic analysis

The processing time will be one potential problem for the systems. If our approach increases the scope of the user's egocentric social network, for example, one more level for extending the social network, the number of Twitter accounts and their tweets will increase exponentially. As a result, the analysis process will cost much more time. It is very useful to improve the semantic analysis speed by introducing algorithms that are more efficient.

3. New words reorganization

Another limitation for our system is that it cannot query "any" word. The participants will get empty result if they tried to type "BOA" or "Jennifer Aniston" as the queries. Our system need this kind of words be recognized, such as turning them into "bank" and "celebrity". The words that are not in WordNet include celebrity names, abbreviations for texting, street names and so on. Understanding these words would perfect our system.

7 REFERENCES

[1] "Iran Elections: A Twitter Revolution? - washingtonpost.com.", Accessed June 17, 2009, http://www.washingtonpost.com/wp-dyn/content/discussion/2009/06/17/DI2009061702232.html

[2] "Bill Gates releases mosquitoes into audience - Health - Infectious diseases - msnbc.com." , Accessed February 4, 2009, http://www.msnbc.msn.com/id/29022220/

[3] "Twitter Blog: Measuring Tweets." Accessed February 22, 2010, http://blog.twitter.com/2010/02/measuring-tweets.html

[4] "Facebook ,"*http://www.facebook.com/.*

[5] "Flickr," *http://www.flickr.com/.*

[6] "Interpersonal ties - Wikipedia" http://en.wikipedia.org/wiki/Interpersonal_ties

[7] M. Granovetter, "The Strength of Weak Ties: A Network Theory Revisited," *SOCIOLOGICAL THEORY*, vol. 1, 1982, pp. 105--130.

[8] D.Z. Levin, R. Cross, and L.C. Abrams, "The strength of weak ties you can trust: the mediating role of trust in effective knowledge transfer," *MANAGEMENT SCIENCE*, vol. 50, 2004, pp. 1477--1490.

[9] G.A. Miller, "WordNet: A Lexical Database for English," *COMMUNICATIONS OF THE ACM*, vol. 38, 1995, pp. 39--41.

[10] C. Fellbaum, *WordNet: An electronic lexical database*, The MIT press, 1998.

[11] A. Trappey, F. Hsu, C. Trappey, and C. Lin, "Development of a patent document classification and search platform using a back-propagation network," *Expert Systems with Applications*, vol. 31, 2006, pp. 755-765.

[12] M.W. Berry, S.T. Dumais, G.W. O'Brien, M.W. Berry, S.T. Dumais, and Gavin, "Using Linear Algebra for Intelligent Information Retrieval," *SIAM REVIEW*, vol. 37, 1995, pp. 573--595.

[13] J. Artiles, J. Gonzalo, and F. Verdejo, "A testbed for people searching strategies in the WWW," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, Salvador, Brazil: 2005, p. 569.

[14] R. Guha and A. Garg, "Disambiguating people in search," *13th World Wide Web Conference (WWW 2004), ACM Press*, 2004.

[15] M.D. Dunlop, "Development and Evaluation of Clustering Techniques for Finding People," *Proc. Of the third Int. Conf. on PAKM*, 2000.

[16] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old," *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09*, Boston, MA, USA: 2009, p. 201.

[17] I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel, "Same places, same things, same people?," *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, Savannah, Georgia, USA: 2010, p. 41.

[18] S.A. Golder and S. Yardi, "Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality," *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on*, Los Alamitos, CA, USA: IEEE Computer Society, 2010, pp. 88-95.

[19] S. Ioffe and D.A. Forsyth, "Probabilistic Methods for Finding People," *INTERNATIONAL JOURNAL OF COMPUTER VISION*, vol. 43, 2001, pp. 45--68.

[20] S. Ioffe and D. Forsyth, "Finding People by Sampling," *the Seventh IEEE International Conference on Computer Vision*, Kerkyra , Greece: 1999, pp. 1092-1097.

[21] B. Shevade, H. Sundaram, and L. Xie, "Modeling personal and social network context for event annotation in images," *Proceedings of the 2007 conference on Digital libraries - JCDL '07*, Vancouver, BC, Canada: 2007, p. 127.

[22] B. Shevade, H. Sundaram, and M. Yen-kan, "A Collaborative Annotation Framework," *PROC. INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO 2005*, 2005.

[23] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures," *IN WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES, SECOND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 2001.

[24] "Twitter API wiki," *http://apiwiki.twitter.com/.*

[25] Author: dewitt@google.com, "Python-Twitter," *http://code.google.com/p/python-twitter/.*

[26] "Twitter API," *http://twitter.com/help/api.*

[27] Eran Hammer-Lahav <eran@hueniverse. com>, "The OAuth 1.0 Protocol," Apr. 2010.

[28] J. Roesslein, "Tweepy," *http://joshthecoder.github.com/tweepy/.*

[29] "TweetCloud," *http://tweetcloud.com/.*

[30] C. Leacock and M. Chodorow, “Combining local context with WordNet similarity for word sense identification,” *WordNet: A Lexical Reference System and its Application*, 1998.

[31] P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448--453.

[32] J. Jiang and D. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” *International Conference Research on Computational Linguistics (ROCLING X)*, 1997, p. 9008.

[33] D. Lin, “An Information-Theoretic Definition of Similarity,” *IN PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 1998, pp. 296--304.

[34] B. Shevade, “FRAMEWORKS FOR COLLABORATIVE MEDIA ANNOTATION,” Master Thesis, Arizona State University, 2008.

[35] T. Hastie, R. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning*, Springer, 2003.

[36] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” *STANFORD INFOLAB*, 1999, p. 17.

[37] G. Salton and J. Michael, *McGill. 1983. Introduction to modern information retrieval*, McGraw-Hill, New York, 1983.

[38] Y. Yamamoto, “Twitter4J,” *http://twitter4j.org/en/index.html,*

[39] J.L. Myers and A. Well, *Research design and statistical analysis*, Psychology Press, 2003.

[40] “Digg,” *http://digg.com/.*