THE REPUBLIC OF TURKEY
BAHCESEHİR UNIVERSITY

A GRAPH-BASED FOLLOWEE RECOMMENDATION APPROACH FOR SOCIAL NETWORKS

Master of Science Thesis

SERDAR ÖZAY

İSTANBUL, 2013

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCE COMPUTER ENGINEERING

A GRAPH-BASED FOLLOWEE RECOMMENDATION APPROACH FOR SOCIAL NETWORKS

Master of Science Thesis

SERDAR ÖZAY

Supervisor: ASSIST. PROF. DR. TEVFİK AYTEKİN

İSTANBUL, 2013

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES SCHOOL COMPUTER ENGINEERING

Name of the thesis: A graph-based followee recommendation approach for social networks
Name/Last Name of the Student: Serdar Özay
Date of the Defense of Thesis: 02-08-2013

The thesis has been approved by the Graduate School of Natural And Applied Sciences.

Assoc. Prof. Tunç BOZBURA
Acting Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Assist. Prof. Dr. Tarkan AYDIN
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Comittee Members	Signature
Assist. Prof. Dr. Tevfik AYTEKIN	----------------------------------
Assoc. Prof Dr. Alper TUNGA	----------------------------------
Assist. Prof. Dr. Egemen ÖZDEN	----------------------------------

ABSTRACT

A GRAPH-BASED FOLLOWEE RECOMMENDATION APPROACH FOR SOCIAL NETWORK

Serdar Özay

Computer Engineering

Thesis Supervisor: Asst. Prof. Dr. Tevfik Aytekin

September 2013, 42 of Main Text

Today, billions of people are using social network web sites to communicate. These sites consist of huge information about today’s people relationships. The study aims to understand categorical behaviors of a social network user and recommend new celebrities to follow by analyzing the user’s social networks acts.

We work on the two biggest micro-blogging web sites for this purpose: Twitter and Tensei Weibo because of accessibility and data attributes.The first goal of the study is to represent the social network data on a graph. We believe that the graph representation is the most natural way to represent the social network data. To do that, we calculate numerical distance values between two people on a social network using different approaches. We use different attributes to calculate the distance, which comes from user acts on social network web sites.

After building the graph, all shortest paths are calculated between each user who are not connected already. Dijsktra shortest path algorithm is used to find the distance. On the recommendation stage, simply the closest celebrities, that are not followed already, are recommended to target user.

We perform experiments to evaluate the performance of graph based followee recommendation approach and its variations and discuss the results.

Keywords: recommendation systems, social network, Twitter

ÖZET

SOSYAL AĞLARDA, GRAF TABANLI TAKİPÇİ ÖNERİM YAKLAŞIMI

Serdar Özay

Bilgisayar Mühendisliği

Tez Danışmanı: Yard. Doç. Dr. Tevfik Aytekin

Eylül 2013, 42 Sayfa

Bugün milyarlarca insan internet üzerinden sosyal ağ uygulamaları üzerinden haberleşmekteler. Araştırmacılar için bu insan ilişkilerini anlamak için büyük ve yeni bir veri tabanının oluştuğu anlamına geliyor. Bu çalışma da öncelikle sosyal ağ siteleri üzerinde ki kullanıcıların davranışsal verileri kullanılarak, kullanıcıların birbiriri arasında ki ilişkileri bir ağ üzerinde sayısallaştırmak hedeflenmiştir. Bu yapı oluştururulduktan kullanıcıya, son çıktı olarak, takip edebilecekleri yeni bir ünlü önermek hedeflenmiştir.

Sağladığı erişim imkanları ve içerdikleri data özelliklerinden kaynaklı olarak, bu çalışma Twitter ve Tensei Weibo mikroblog siteleri üzerinde yapılmıştır..

Çalışma sırasında, ünlü olarak adlandırılan kullanıcılar kişilerin kategorik verilerini çıkarmada ayırt edici özellikle olabileceği düşünülerek merkeze alınmıştır.

Çalışmanın ilk amacı sosyal ağ uygulama verilerinin bir graf üzerinde sunumunun sağlanmasıdır. Graf sunumunun sosyal ağların doğal sunumu olduğu düşünülmüştür. Bu maksatla, sosyal ağda ki, iki kişi arasında ki uzaklığın sayısal bir değeri dönüştürülmesi için farklı yaklaşımlar kullanılmıştır. Her yaklaşım da, kullanıcının sosyal ağlar içerisinde davranışlarından yada var oluşundan kaynaklı farklı özellikleri baz almıştır.

Grafın ayağa kaldırılmasından sonra, uzaklık hesabı yapılmamış tüm yollar yani birbirine direkt bağlı olmayan kullanıcılar arasında ki uzaklık, Dijkstra kısa yol algoritması uygulanarak hesaplanmıştır. Kullanıcıya yeni bir ünlü önermek için son olarak, bu kısa yol verilerine bakılarak, kullanıcının takip etmediği en kısa mesafede ki ünlüler kullanıcıya önerilmiştir.

Anahtar Kelimeler: öneri sistemleri, sosyal ağlar, Twitter

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

It was seen earlier stage of Internet evolution that it is used mostly to access knowledge. Many people like W3C Director Tim Berners-Lee are called that stage is web 1.0. (Getting, 2007) The web was nearly “read-only”. It allowed us to search for information and read. Web 1.0 was about connecting computers and making information available on the world.

After that stage, some application appeared on Internet, which tries to recognize the users. The most known samples were shopping chart sites. It is started to say simply ‘hello’ with user special name differently. These were the first meet of computer systems and the people. All of the read only communication turns to dynamic system quickly at that starting point. Today, it is reached complicated, which uses machine-learning algorithms, production recommendation application.

Web 2.0 term first was coined in by Darcy DiNucci (1999) and was popularized by Tim O'Reilly (2005) at the O'Reilly Media Web 2.0 conference. Web 2.0 suggests a new version of the World Wide Web. They called that for all changes comes after from ‘read only’ age of Internet. It is not only about communication of between people and computer but also it is between people to people and machine to machine. Some of the people split and called that stage of Internet as web 3.0.

Today, an internet application does not used only that own database to server to people. It is also connected to many different applications to serve better service to end-users. The machine-to-machine communication is realized on web service frameworks.

Of course, the people to people communication was on the startup of Internet. One of sample is mail service. But with the social network applications, the most of the web application’s abilities is given to people use. On these abilities, all the users can be a service provider to their followers. Today anyone, with a small computer or a smart phone, can be a great newspaperman or a writer or a journalist, showman and etc. without almost any technical background.

Today, billions of people communicate with each other on social networks. Only Facebook, maybe most known service, have 1 billion users (Smith 2013). The information about people relationships on the networks is huge. It gives many opportunity today's researchers.

We prefer to work on microblogging services because of the relationship with information and accessibility. So the study is based on Twitter and Tencent Weibo social networks. But the methodology can be implemented to other social networks.

The main purpose of the study is to represent people as mathematical data. To do that we aimed to represent people on a network graph first. We believe, graph representation is the nature of the social networks and it can be built a meaningful graph, which represent the users with using the distance to other users.

Of course, after that many consequences can be produced from the graph data but we have chosen the aim of work is to recommend new users to follow on the social network. If the methodology is successful on recommendation, it means that graph representation is successful also.

2. LITERATURE SEARCH

Since the early days of Twitter, it offers special facilities; so many researchers focused their attention on this area. Twitter, which is a simple service, allows people to publish a 140-characters text message to other people whom called followers. Twitter is not only one of the biggest social network service on the Internet but also it is the fastest way to spread the information out among huge crowds easily and quickly.

In Twitter, people can follow whoever they want. Information is generally public. So many news channels, politicians, celebrities and the other sources, which wants keep in touch with people uses Twitter actively.

For these reason Twitter also has provided many facilities for researchers to understand the relationships and communication between people, unlike the other social networks. There are many research topics on Twitter but this study is interested in only one of them, 'followee recommendation' to people.

This topic is not only related to recommend a new person to Twitter user but also it provides new opportunities to understand which person close to whom or how can we find communities.

One of the early studies in this regard is the paper of the Lo and Lin (2006). In this work, the researchers offer a recommendation algorithm named as “weighted minimum-message ratio” (WMR) which generates a limited, ordered and personalized friend list by looking real message interaction number among web members. The algorithm simply looks the message counts, which sent by users to each other, and it assumes that a strong relationship when there is more message count**.** After that it builds a graph with these count values and prepare a ranked list with looking distances and message counts. An example can see on table 2.1 and figure 2.1.

Figure 2.1: Message Count Graph

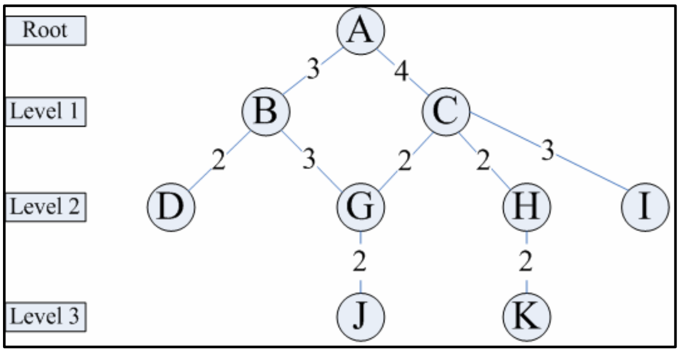


Table 2.1: Ordered recommendation list to node

Recommend Member	Member	Recommend Score
1	G	6.00
2	I	4.00
3	H	3.43
4	D	2.14
5	J	1.62
6	K	0.76

In this paper we see 'Graph-based' word on title. It is imported because the works about followee recommendation are generally divided into three basic models: Content-base, (Collaborative) graph-based or hybrid. When it is said the method is content-base, the research model works with information on the content like message content or the users' personal information like age, country or the other meta-data. These models also include neural language process or text retrieval methods. If the research works on graph base model, like mentioned paper above or in our thesis, the model does not look at the content, it looks at who sends message, number of messages, what is the retweet number of a message, number of followers or who follows whom. Then generally it builds a graph to understand this data. Hybrid method uses two methods at the same time on that model.

Another paper about on this subject is written (Chen and Geyer et al, 2009). This research is about IBM Sonar software solution and it works on only closed company based social networks. In this study, three different methods were compared with SONAR about recommending a new person to an user to be friend. One of the methods in this paper a content base method, the second method named as Content-plus-Link and the last one is friends of Friends (FOF). Especially Friends of friends' algorithm is important for this thesis study. We used it to compare our methods. It simply takes the users' friends of friends and rank these users by looking how many times repeated on their friend lists**.** It is the one of the simple methods of Collaborative based recommendation algorithm. This study shows that FOF gives much better results than the other two content-based methods.

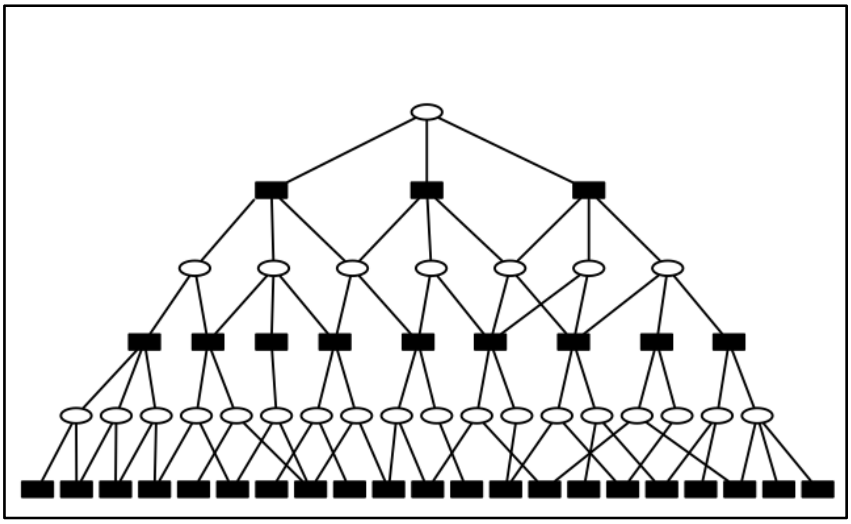
A remarkable research on Twitter is written (Kwak et all, 2010). The main objective of this research is to study on topological characteristics of Twitter. They use a fairly large sampling. (41 million users and 106 million tweets). An interesting fact which found on that research, the pagerank and follower count properties effects were similar in order to identify influential on Twitter. But retweet counts show different properties. In this research it is clearly defined when ranking the popularity of a user, or understand an user domain, we can use follower count. This paper is one of the most cited research in this area.

One of the researches about the problem of finding authoritative users in a micro-blogging service is *TURank* (Yamaguchi et al, 2010). For this purpose they build up a graph network with using their own algorithm, which named as TURank (Twitter use rank). In TURank, users and tweets are represented in a user-tweet graph which models information flow, and ObjectRank is applied to evaluate users' authority scores. ObjectRank algorithm is an improvement over pagerank.

The Turank algorithm seems more successful than FollowNum, RTNum, PageRank, HITS methods as a result of the study. However, the authorities such as weather tracking services, which have huge follower count, show poor results because of algorithm gives great importance to retweet count.

Also another article about how data influence on twitter (Wu et all, 2011). Especially data collection method of this study is rally useful for working with huge graph data like twitter. They use snowball-sampling (figure 2.2) method for this purpose, which comes from the science of sociology. The method is also used on this thesis as will be described later.

Figure 2.2: Snowball Sampling



Another article is written about finding user influence (Cha et al, 2010). In this work a comparison of three measures of influence is made: indegree, retweets, and mentions. As the result of this research, they couldn't find any mutual relationship between retweet count and follower count. Also there is no rule found like every people have huge follower have a big influence. As the most retweeted uses are new channels while the celebrities are not retweeted as much as.

Other study (Hannon et all, 2010) is used text retrieval methods to examine different data sets, which collected from Twitter. The work is focused on two main categories; Content base, Collaborative filtering. At content base method, the researchers have been worked on data of tweet's contents, tweets of followees, tweets of followers and mixture of all. Also they worked with followee list and follower list under title of Collaborative filtering. All the users, which are sampled on the paper, represented as vectors with each of the dataset are applied text retrieval methods. After that with looking closeness of these vectors, the recommendation process is completed.

Another interesting study on followee recommendation is written (Armentano et al, 2012). Researchers of this paper are not interested in content of Twitter data as our study. Simply the method of the study takes an user’s followees of followers, and after that again take the followees of the last taken list of users. After that to order this list, study looks some of values. One of them is found with counting, how many times per user exist on this list, the other is common friend count between the recommended user and source user and the other is a rate of follower count and followee count of recommended user. This algorithm is also applied in our study to compare our main algorithm model because of there is some similarity on model. Especially, for this reason to understand our results success this study is used on this thesis.

After that study, the same researchers has been published another article as future work. Its name is Towards a Followee Recommender System for Information Seeking Users in Twitter (Armentano, 2012). On this work, same technique is compered with content base algorithms. As the result of the study the content base and topology base techniques produced as similar results.

One the works on this topic gives another approach about followee recommendation (Lu et all, 2012). In this study the tf-idf ranking method is used which comes from text retrieval techniques and it is a content base approach. This study briefly re-ranks tweets in user’s time-line, by constructing a user profile based on user’s previous tweets and by measuring the relevance between a tweet and user interest.

3. A GRAPH BASED RECOMMENDATION METHOD ON TWITTER

3.1 TWITTER AS A SOCIAL NETWORK PLATFORM

3.1.1 What is Social Network?

When we analyze today's online social network applications, we can easily observe some common features and properties. One of the most common property, all of them was started with a web-based application, although today they are using all known internet based information channels like new generation phones or tablets and its applications. Another most known function, is publishing a public or semi-public user profile page. But the most distinctive property of a social network from a blog or community page, they consist who connected with who information and they used it in their applications.

The first recognizable social network site SixDegree.com was published in 1997 and they created most used feature of a social network. They simply service to people to create a friend list and surf the other's friends list. Of course a similar services have been giving by some other application like some dating site or some chatting applications like AIM or ICQ but they did not allow to surf on your friends of friends list. Everybody was closed on their unique universe. SixDegree was changed that. In a time, they also were upgraded and they published a new service, send message to friends. But they cannot survived after 2000 (Boyd et all, 2007).

In 1997 to 2001 some of similar services was supported some of feature of SixDegree, but the next wave in social network sites was become in 2001 with Ryze.com. It was started to its life at 2001 and its main purpose is leverage people on business networks. Ryze couldn't be popular, but it was a pioneer in the others like Ryze, Tribe.net, LinkedIn, and Friendster from the same area. And today as you know LinkedIn became a powerful business service in this field.

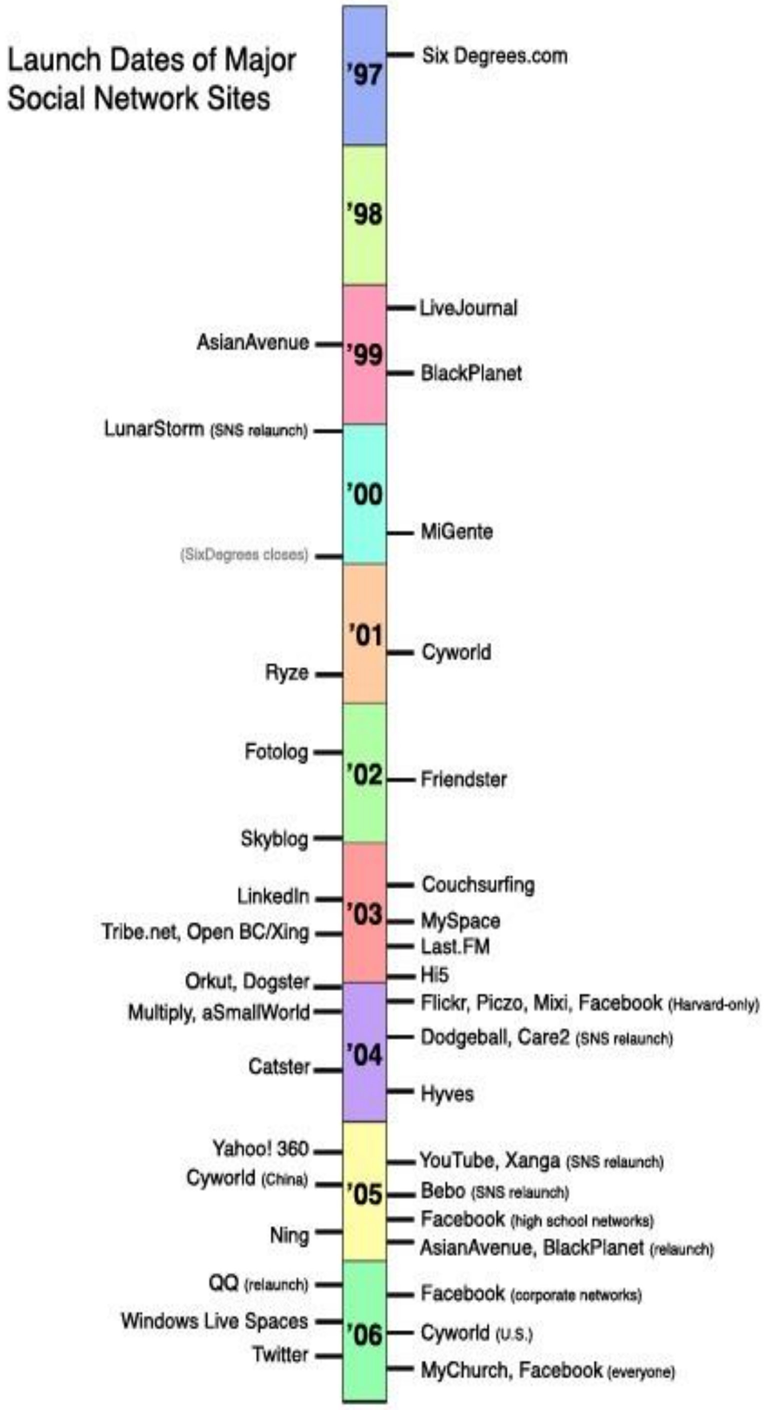
Another social network site Friendster which is also remarkable, was launched on 2002. It was a dating site but unlike the others it doesn't to try to introduce people to strangers. It has been helping to find new partner from friend's to friends lists. On initial design it allowed to view people who were closer than user's four degree away. But for this reason, to see additional profiles, users started to add interesting or famous fake account ("etc. Brown university,"). These fake account was become centers of interest in time. Many people started to follow these account. The Friendster's popularity grown rapidly in USA. It also increased out of USA in the Philippines, Singapore, Malaysia, and Indonesia. 3 million users was become web site users. But they didn't handle this rapid grown. Also they don't understand user's need and they deleted to fake accounts. In a time, the users lost their interest and trust to company.

Another most know social site, Myspace was started its life a competitor of Friend-ster. It was borned on 2003. Main competitors were Friendster, Xanga and AsianAvenue. They wanted to attract the Friendster users. After rumors about the Friendster will become fee-based system, MySpace grown rapidly. Another characteristic feature of MySpace it gives to user to personalize their pages. These features were liked by especially music bands.

Firstly Indie-rock bands from Los Angeles region, was began creating profile, and tried to access to followers with MySpace pages. After that MySpace contacted to local musicians to see how to help them. Bands were not main existence reason of MySpace, but bands and their fans helped Myspace to grown. Then, News Corporation purchased MySpace for $580 million significant in July 2005. From 2005, until early 2008, Myspace was the most visited social networking site in the world. In June 2012, Myspace had 25 million unique U.S. visitors.

While MySpace grown in US and abroad, Friendster spread in the Pacific Islands, Orkur became the most used social network site in Brazil and India. In japan Mixi is loved, in Sweden LunarStorm became popular. Dutch users embraced Hyves and Hi5 was loved in small countries in Latin America, South America and Europe. And the others Bebo, Grono, QQ come. Social networks web sites spread quickly the entire quickly world on this time interval.

Figure 3.1: History of Social Sites



Reference: Social Network Sites: Definition, History, and Scholarship, Journal of Computer-Mediated Communication, 2007

Alongside these open services, another social network sites, Facebook was designed to support college students. Facebook started to it life 2004 in Harvard university. To use this site, users have to hardvard.edu email address. In the future, they accept the other university students in Facebook. After 2005, Facebook expanded to everyone. Unlike the others, Facebook is given the change to users to publish public page for all users.

Also they had given opportunities to developers to build their own Facebook application with using Facebook API. Today Facebook is the world's largest social network, with more than 1.06 billion monthly active users. They have 618 million daily active users for December 2012. The company has generated $1.58 billion in revenue for the last quarter in 2012 (Tam D., 2013).

3.1.2 An Overview of Twitter

Twitter is a free microblogging service. It is founded in 2006 by Jack Dorsey and Biz Stone. Its gives ability to users, publish their 140 characters message to people who are subscribed user's message. The messages are published on user profile page. And the user can decide the messages can see on private network, which created by user or public.

The service rapidly has been grown since its first day of life. On 2012, it has been reached to 500 million registered users. The twitter users send 456 tweets per second. (Hold R., 2013)

Twitter is an online social networking service and microblogging service that enables its users to send and read text-based posts of up to 140 characters, known as "tweets". It was created in March 2006 by Jack Dorsey and launched that July. The service rapidly gained worldwide popularity, with over 140 million active users as of 2012,generating over 340 million tweets daily and handling over 1.6 billion search queries per day.[1] Also in Turkey 7.2 million people use Twitter service.

The users can be easily send and publish your message in Twitter like a web blog. But the biggest success of the twitter is to give an opportunity the application users for build own social in-formation channel. The users can follow anyone who can be a real friend or a famous person is using twitter application. At the same time the users can be followed by anyone. For this reason, twitter has become an important media on internet. Today, many famous person and company are using Twitter to direct communication

[1] http://en.wikipedia.org/wiki/Twitter

with public. For example our president Abdullah Gul is using actively your twitter account and 3,600,534 people is following him over twitter. [2]

For this reasons, Twitter network is given many opportunity to understand people behaviors and social network to researcher. Today an important count of researchers is interesting in the data and people behaviors in Twitter. Some of them are trying to cluster people, some of them research the user influence. To find most powerful users whose messages read and forwarded by user, have also economical meaning in today’s market.

3.1.2.1 Basic definitions

Before to analysis Twitter, we have to speak some basic feature of twitter network application to understand the information twitter contains,

Follower or Following: A user can follow someone or someone can follow the user. This gives us important information about relationship between two people.

Trendy (Celebrities) People in Twitter: Some famous people in twitter have too many followers. We can easily say that following the famous people give some certain information about interesting groups (politician, sport, music, travel, TV etc.) of a user.

Direct message (DM): Sending a direct message to your followers.

At (@): It is used when a person refering a user in him updates. It can prefix him username with @ to display him Twitter account in update.

Retweet: In Twitter, people can re-send a incoming broadcast tweet messages that send by followed by users. So that re-tweeted message contains information about relationship of users and common likes.

Hash (#) : When a user want to tell something specific about a subject, it can be prefix with #. If the other users also interest the subject and use same hashtag, it can be build

[2] https://twitter.com/cbabdullahgul

online forum. It can see the most popular hash tags on Twitter page. Everyday thousands of people communicate on these hash tags.

Direct message: You can send direct message to someone with adding @ char to an user.

Blocking/Spam : It can be block any Twitter account, when user wants. Also a user can complaint an account as spam. The account can be closed or suspended if the complaint is deemed justified by Twitter authorized.

Message: The information, which is written and sent by people using Twitter network.

Public / Private Account: A twitter user can open your account everybody to see, so that way if you are not in user friend list you can see the user's message, followers or following. If account is set private only user's friend can access user's information.

3.2 FOLLOWEE RECOMMENDATION

3.2.1 Some Recommendation Methods

To understand how our algorithm success, we used some methodologies as competitor. These are listed at below.

3.2.1.1 Common followees

The method is taken from the study, which interested on the recommendation problem (Armentano et al, 2011). As we explain in literature search part of the thesis, the method takes an user's followees of followers. After that it is taken followees of the last taken list of users and it is built a ranked list with a rank value which calculated with formula. The formula is which is build by researchers of paper given formula 3.1

$$D_{cf} = \frac{C_{rp}}{C_a} \frac{C_{fw}}{C_{fe}} \frac{C_m}{100} \tag{3.1}$$

C_{rp} = Count of how many times per user exists on last joined list

C_a = All user count on the list

C_{fw} = Count of follower

C_{fe} = Count of followee

C_m = Count of mention

Mention is used in Twitter terminology. It means similar reference. In Twitter message, users can mention the others user with using @ tag like @username. We can not take mention count from Twitter API, so we used 1 for all. Also C_{fw} and C_{fe} values used if only our database have. Because of each of list can be reach 800K size. And try to take all person followers and followee count is huge work. But In the study we especially try to recommend 50 celebrities, this simplification does not have to change anything. Also when we use this methodology, we ignore people who are not in our celebrities list. Because our approaches try to recommend only celebrity, not friend. That can be tricky. Armentano‘s (2012) method does also interests ordinary peoples and recommends them. But it is also rewarded celebrities person as you can see in formula.

3.2.1.2 Friends of friends

It is a basic model to recommend new person in social network system. The hypothesis of model simply says that people like friend’s preferences. To realize on a social network, we get source person’s followee’s and we rank in a list their followee preferences. Rank value is increased with repeating count in list of followee of followees.

3.2.2 Graph Based Recommender

Our plan is simply to take a set of people data from social network web sites and build a graph, which consist meaningfully distance data about their relationship of users between each other. After that we try to recommend new followees with looking these distance. It is targeted to work on Twitter social network web site initially. After we have gotten some results, we also applied our approaches to Tencent Weibo social network.

Initially, we try to calculate a distance value from Twitter user's data, which represents closeness of two users; one of them follows to other. It is targeted, if we can find a distance value like that than we can calculate distances for all people in the network. So we build a graph, which shows the users with their distance to each other.

3.2.2.1 Shortest path algorithm

It is an important problem in graph theory. As the same reason, it is also important in its applications such as transportation, communication and electronic. In generally, a graph represent as G(V,E). V represents vertex or nodes, E represents the edges on the graph. Main problem can be defined as to find the shortest path between two nodes. The graph properties can be change for different problems. The edges can be weighted or unweighted or the weights can be negative or positive. For all state, the shortest path problem solutions can change. Another important point on working shortest path problem is performance is property of the solutions. Sometimes finding shortest path can takes too long time or can goes to infinite.

3.2.2.1.1 Dijkstra shortest path algorithm

The algorithm is known as the fastest shortest path algorithm on positive weighted graph. When it runs, it calculates all shortest paths from source vertex to others, not only source to a one point. The algorithm is appropriate solution because of our graph data positive weighted and also all shortest path calculation is useful to find our 'bundled' distance metric.

When it is used a naive implementation of the priority queue in algorithm, it gives a run time complexity $O(V^2)$, where V is the number of vertices. Implementing the priority queue with a Fibonacci heap makes the time complexity $O(E + V \log V)$, where E is the number of edges. (Cormen et all, 2009)

The algorithm in simple terms works like;

1. Create a hash distance list and input all vertex and distance pair which distances set infinity at first except starting (source) point which is set to zero
2. Create a visited vertex list, and put source vertex

3. Get the all one edge away neighbour vertexs from visited vertexs
4. Get the shortest distance with adding distances which lies from source vertex to the neighbour vertexs. It is used here, hash distance list to do not calculate again and again distance value from source to target neighbour.
5. Add the shortest distance value into hash distance list with new shortest away vertex and total distance value
6. Add the new vertex to visited vertexs list
7. Turn to step 3 until all the vertex puts in visited vertex list

3.2.2.2 Calculation of distance

The first problem was to find the distance between two connected people in Twitter on building graph. Initially, we analyzed data types which are obtained from Twitter to use in that calculation. After that, we have used some different approaches to calculate the distance.

To calculation of distance between two people who does not connected directly, we preferred to use Dijkstra algorithm. It is the fastest algorithm for our problem. And with working big data, like our problem; it is a necessity, not a choice.

3.2.2.2.1 Mutual distance

Initially, we try to find a distance with building a formula which inference from some basic acceptance. One of them is if a user follows a person, we call him as followee, who have more followers (like celebrities) than the user's other followees, this relationship is more important to understand user interest than other. Because of we accept that the celebrities give more information than the others. So we have to reward it. To join this property in our calculation, we define R_f parameter

$$R_f = \begin{cases} \log F_c & follower\ count\ know \\ 2 & else \end{cases} \tag{3.2}$$

F_c = target user followers count

R_f = user closeness metric for target

Another value is calculated from retweet property. If a user retweets a target user's message, it gives an idea of closeness of interests on these two people. We can follow weather channels but we don't retweet their message. Or we follow many students from our university but if we don't have a similarity, we don't retweet messages generally. So we join that property.

$$R_r = \begin{cases} \log F_c & follower\ count\ know \\ 2 & else \end{cases} \tag{3.2}$$

F_c = target user followers count

R_r = user closeness metric for target

If a user retweet a target user message, R_r values calculated as above. As you see follower size of target user is rewarded like calculating $R_f$for the same reason. We preferred $Log10$ multiplier because of some celebrities have millions of followers.

In our thinking, two people of common followers or followees also could have to give an idea to understand how they close. If two people follow same peoples, when one of them finds a new followee, the other can also interest to the new followee. But how can we calculate a value with looking common followers or followings count. We use adjusted round index calculation metric to find it. (Vinh, 2009)

Adjusted round index generally use in to compare clustering result. When we cluster the data we need a measure of agreement. A measure gives the similarity between of two cluster. In our problem we used it to understand the similarity of two people on Twitter.
Adjusted round index value was calculated two times. One is worked with using common followers and the other for common followings of the people.

$$ARI = \frac{x_1 x_4 - x_2 x_3}{(x_1 + x_2)(x_3 + x_4) + (x_1 + x_3)(x_2 + x_4)}$$

(3.3)

x_1 = Common followings/follower count

x_2= A followings/followers count - Common followings/followers count (A-B)

x_3= B followings/followers count - Common followings/followers count (B-A)

x_4= All working set (1,230,496 users) - (A U B)

At last, we calculated the distance value with all these values,

$$Distance = \frac{1}{(R_f + R_r + (10(R_{cfw} + R_{cfr}))} \quad \textbf{(3.4)}$$

R_{cfw} = ARI for followings

R_{cfr} = ARI for followers

$R_f and\ R_r$Is calculated at formula 3.1 and 3.2 on top.

3.2.2.2.2 Unit distance

It was set all distance 1 in that methodology when one person follows to other. We have to remember that our graph built as directed. As you can see in the result section, this method disclosed to interesting hit ranges. The range besides was huge because of the setting all distance values to one. In an example, if you have 30 followees, you have 30 edges which weights are one. If each of the followees follows 30 followees too, then you have 900 edges which weight is two. As you can understand, trying to find the shortest path and order with these each other with looking the distance values will not be possible. So we built ‘bundled’ methods.

3.2.2.2.3 Bundled unit distance

As explained above, if it is set all weight one there will be a problem about to understand relationship on people which are same distance away from source user. To solve that problem, we planned, not only look the target user distance, but also all the distances to target user’s direct followers.

In using unit distance, called D_m, we find direct Dijkstra distances. After that we calculated all shortest paths on network again. But, at this time we calculated all shortest paths to target user’s followers (one to many). And we calculated a sum values from these values. We called as D_b. We use these values in our recommender systems.

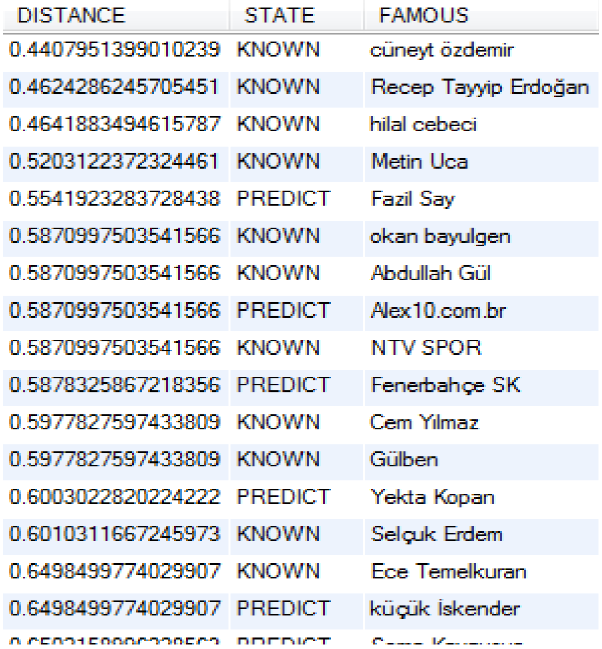
3.2.2.2.4 Bundled mutual distance

In here, it has been used same bundled methodology as explained above but it has been used mutual distance to calculate Dijsktra direct distance. (D_m) So we didn’t set all graph to one as above instead we use mutual distance values at which have been calculated as first. Than we have found all target user’s followers distance as bundled method and we take sum of these (D_b).

3.2.2.3 Recommend a new followee

At final stage, it has been built 4 different graphs for different distances calculation as above. To recommend a new followee to a user, we ordered celebrities with taking their distance to our source user. It is recommended to closest celebrities, which does not already exist in the source user’s followee list.

Figure 3.2: Recommendation list results



DISTANCE	STATE	FAMOUS
0.4407951399010239	KNOWN	cüneyt özdemir
0.4624286245705451	KNOWN	Recep Tayyip Erdoğan
0.4641883494615787	KNOWN	hilal cebeci
0.5203122372324461	KNOWN	Metin Uca
0.5541923283728438	PREDICT	Fazil Say
0.5870997503541566	KNOWN	okan bayulgen
0.5870997503541566	KNOWN	Abdullah Gül
0.5870997503541566	PREDICT	Alex10.com.br
0.5870997503541566	KNOWN	NTV SPOR
0.5878325867218356	PREDICT	Fenerbahçe SK
0.5977827597433809	KNOWN	Cem Yılmaz
0.5977827597433809	KNOWN	Gülben
0.6003022820224222	PREDICT	Yekta Kopan
0.6010311667245973	KNOWN	Selçuk Erdem
0.6498499774029907	KNOWN	Ece Temelkuran
0.6498499774029907	PREDICT	küçük İskender
[illegible]	PREDICT	[illegible]

4. EXPERIMENTAL SETUP

4.1 DATA COLLECTION

Big data will be the first problem, which has to solve by the researchers who interest to analyze social network data. Today, the user volumes of social media sites and transaction counts have reached incredible size. If we try to explain with real numbers that situation; Twitter had been 140 million active users in 2012. Daily sent message count was reached 340 million and 1.6 billion search transaction has been done on Twitter in those days. The other system, which this thesis also studied about, Tencent Weibo has been 100 million active user at the end of the 2012 and it has been also 277 million registered users.

4.1.1 Twitter Data

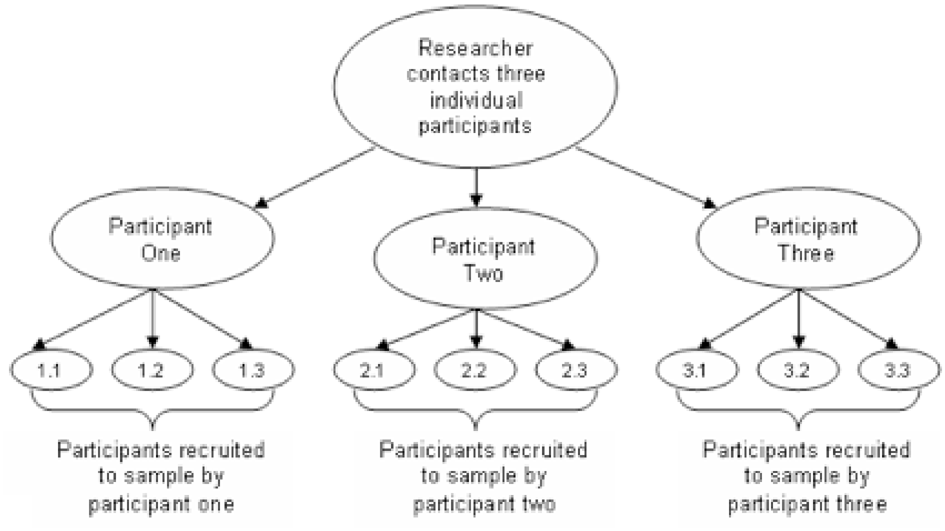
The simple idea of the this study is to build a network graph from social network data which consisting the people followers and followings relations and the some other data about their relationships. Our main aim is to analysis and catches people interests with working this graph. But Twitter has huge network information, we had to do some sampling on Twitter and Weibo networks to build small but meaningful network graph.

4.1.1.1 Snowball sampling

Snowball sampling method is generally used in sociology and statistics research. It can be called as chain sampling, chain-referral sampling or referral sampling.[3] It is non-probability sampling technique. The process of snowball sampling is much like asking your subjects which you selected, to nominate another person with the same trait as your next subject.

[3] http://en.wikipedia.org/wiki/Snowball_sampling

Figure 4.1: Snowball sampling tree


Researcher contacts three individual participants
Participant One
Participant Two
Participant Three
1.1
1.2
1.3
2.1
2.2
2.3
3.1
3.2
3.3
Participants recruited to sample by participant one
Participants recruited to sample by participant two
Participants recruited to sample by participant three

In our study, we need to take a relational sample part of the network to build meaningful network, so snowball sampling has been seen usefully.

It has been selected 50 famous trendy people from different interest areas (politics, sport, movie, literature, popcorn culture) on the top of the snowball sampling from Turkey country. The main goal is to build maximum discrete network, which users who have different interests. These people are called as ‘first generation’ on this study.

After that, we have been collected followers of the famous person account. We firstly pulled 20 people per famous account (second generation) randomly. And also we pulled again 3 follower people (third generation) from these 20 people randomly. So this way we have been built project working people set.

In this process we used Twitter API. The API only can give access public marked users, because of this public / private restriction, the target user of this study size has been reached 2855 people. We also collected of all network data of the sample set to use our calculation. So in sample data, we reached 7,256,115 edges and 1,367,270 nodes.

In sampling process one of the problems is to take only public account, other one is to eliminate fake or robot accounts. So for each person, we have pulled user’s followers than we have taken followers person randomly. After that, it has been checked followers and following count of the account, to understand of which account is real which is not.

In the planning of the project, it was planned to pull 10000 people to analysis. But we had restricted by Twitter API firstly. Twitter only gives to take 350 requests per hour. We have to do too many requests for one person, like getting followers or messages or get followings etc. Also it was understood that it becomes harder when the data amount grows. We will do many processes, which takes CPU when processing the data. So for that reasons I pulled down working set size to down.

4.1.1.2 Structure of Twitter data

It has been taken different types of data for each user. The data can be taken by anyone who uses Twitter API for public marked users. Some of them are given as follows;

User’s followers : the first 5000 of people who take the user messages
User’s followings : the first 5000 people who tracked by a twitter user
Follower count : all followers count
Following count : all followings count
Retweet count : We can pull last 20 retweets of a user with Twitter API. We have pulled that information and if a user more than one retweets for one user, we also evaluate value to one because of only getting last 20-retweet limit.
User name : name and surname of user
Screen name : the name which seen on twitter as nick
Location : where user lives

Also we have found with calculation

Common followers counts : count of people that follows both of the two users
Common followers counts : count of people that follows both of the two users
Friend State: The users who follow each other

As a note; it is quite difficult to calculate the common followers and followings count of the two users. Because of we have over 8000K edges on relation table on database. When I try to calculate with using MySQL engine for two user whose have 5000 followers for each, it takes 160 second to find common follower count. Total process time for all user who in working set was more than 20 days. After that I turn back to IR basic algorithms, and I use that,

```
while (aindex < a.size() && bindex < b.size()) {
        if (a.get(aindex).intValue() == b.get(bindex).intValue()) {
            count++;
            aindex++;
            bindex++;
        } else if (a.get(aindex) < b.get(bindex)) {
            aindex++;
        } else {
            bindex++;
        }
    }
```

The method was worked. It takes only 5-6 hours to find commons for all users.

4.1.2 Tecent Weibo Data

When we have been working about Twitter data, we noticed there was a international challenge about recommendation user on Tencent Weibo social network, which can be called Chine's twitter. Competition was called Kdd Cup.[4] They have been published two challenge, the one of tasks, which we interested, is given as "The prediction task involves predicting whether or not a user will follow an item that has been recommended to the user. Items can be persons, organizations, or groups and will be defined more thoroughly below".

They have published a huge amount of real social network data which taken from Tencent Weibo. We only used edges, training and celebrities from the data. But still the it was too huge. Dataset consists over 50 million edges and the 1300 K training data. We used 50 celebrities in Twitter sampling as start point, but here we have 6095 celebrities (they called as item because they try to recommend, not only celebrities but

[4] https://www.kddcup2012.org/

also news, games, advertisements, products e.g.). We couldn’t cut the item data. It was not logical. Also 50 million edges are too big for calculation in our hardware environment. So we executed reverse of snowball sampling every sample. In our twitter work, we have been observed that the most meaningful data is 4 people away from source for our algorithm. For every person who is in KDD cup training set, we build new relational network looking. We take only 4 people away from target person and we executed our algorithm in that small graph. Also we didn’t use all training data which have 1.3 million sample, we only tested some of sets each one contains 1000 users.

4.2 EVALUATION METHODS AND METRICS

To understand how our algorithm works or what is our algorithm success. We have to answer two simple questions: what will be our methodology and which metric does we use in it?

4.2.1 Evaluation Methodology

When we look recommendation studies about Twitter, we can see two titles on experimental evaluation methodology. One of them is lived-user experiments and the other, offline experiments that works with true data.

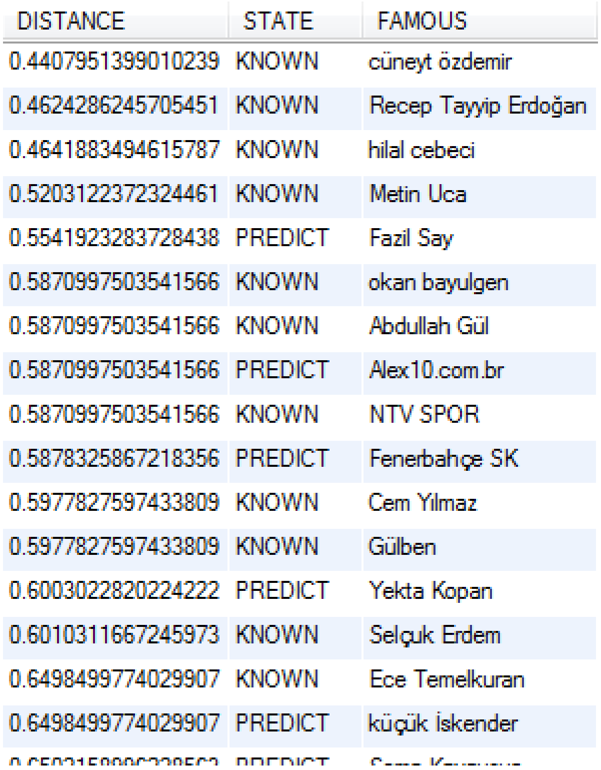
In live user experiment methodology, generally is given a ranked recommendation list to users who invited this experiment. Than the users rerank the list or give another success rate. For an example use, In Armentano (2012) work, they work with students who studied their last lesson. The students first created a twitter account and they have chosen 20 Twitter users who published information or news about a set of particular subjects of their interest. In the second part, researchers recommended to students new Twitter users to follow and they calculate a success rate from asking to students which recommendations are success.

In offline experiment, a group of the users are taken from real Twitter network, and hide some known following relation from users. After that algorithm are evaluated if it was able to rediscover those hidden connections using the remaining connections.

4.2.2 Evaluation Metrics

There are some useful techniques to understand the algorithm success in machine learning researches. But in recommendation problem on Twitter, we didn’t hire these because of the order of recommendation list is important as much as accuracy of result. (Figure 4.2) But as you know the classical measurement in machine learning, like precision/recall, don’t interest of order, they count only success and fails. The closest problem to our problem can be rating of success on a search engine result page.

Figure 4.2: Ordered recommendation



DISTANCE	STATE	FAMOUS
0.4407951399010239	KNOWN	cüneyt özdemir
0.4624286245705451	KNOWN	Recep Tayyip Erdoğan
0.4641883494615787	KNOWN	hilal cebeci
0.5203122372324461	KNOWN	Metin Uca
0.5541923283728438	PREDICT	Fazil Say
0.5870997503541566	KNOWN	okan bayulgen
0.5870997503541566	KNOWN	Abdullah Gül
0.5870997503541566	PREDICT	Alex10.com.br
0.5870997503541566	KNOWN	NTV SPOR
0.5878325867218356	PREDICT	Fenerbahçe SK
0.5977827597433809	KNOWN	Cem Yilmaz
0.5977827597433809	KNOWN	Gülben
0.6003022820224222	PREDICT	Yekta Kopan
0.6010311667245973	KNOWN	Selçuk Erdem
0.6498499774029907	KNOWN	Ece Temelkuran
0.6498499774029907	PREDICT	küçük İskender
[illegible]	PREDICT	[illegible]

4.2.2.1 Discounted Cumulative gain

Discounted cumulative gain (DCG) is a known measure to understand of a ranked list. It uses to analyze the web search engine algorithm results or other text retrieval related applications results. Simply the algorithm is based to give an award which falling from top to the bottom of a recommendation list with looking order. (Armentano, 2011) The DCG calculate at a particular rank position p as:

$$DCG_p = rel_1 + \sum_{2}^{p} \frac{rel_i}{\log_2 i}$$

(4.1)

i = order
rel_i = the graded relevance of the result at position i

if we look the formula, we see that need to per sample in the list must take a success degree which set by expert user.(rel_i) Generally, the degrees will be like 3 relevant, 2 and 1 between relevant and irrelevant,0 non relevant . But in our case it can be use as binary {0,1}.

4.2.2.2 Average Precision

One of the known method is precision and recall algorithms which used in machine learning and informational retrieval problems. Precision (Formula 4.2)
interests, how relevant the retrieved results are.

$$precision = \frac{|\{relevant\ docs.\} \cap \{retrieved\ docs.\}|}{|\{retrieved\ documents\}|} \qquad \textbf{(4.2)}$$

Recall (Formula 4.2) tries to answer the question that the system retrieve many of the truly relevant documents.

(4.3)

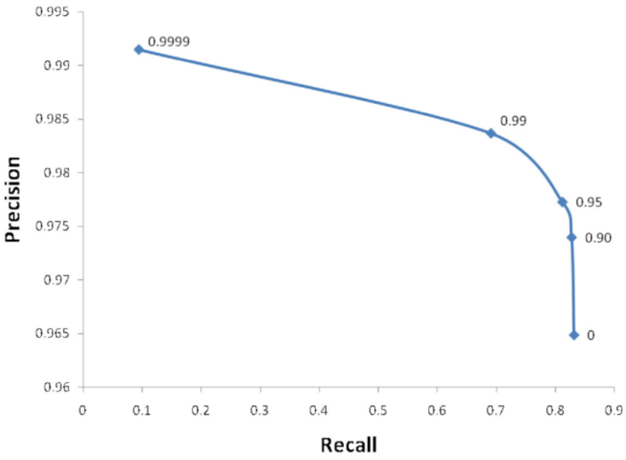
$$recall = \frac{|\{relevant\ docs.\} \cap \{retrieved\ docs.\}|}{|\{relevant\ documents\}|}$$

But as mentioned earlier, the algorithms don’t interest the order of recommendation list, only it evaluate the success of hits in retrieval results. Average precision is an algorithm solve that problem which also uses both recall and precision in. (Formula 4.3)

$$AVP = \int_0^1 p(r)dr \qquad \textbf{(4.4)}$$

If we look at to average precision formula, we see average precision is calculating precision and recall values in continuously. It simply gives the area under the recall-precision curve.

Figure 4.3: Precision Recall Curve



If we assume $r(t)$ is differentiable almost everywhere, then

$$\int_0^1 p(r)dr = \int_0^1 p(t)dr(t) \qquad \textbf{(4.5)}$$

In practice this integral is replaced with a finite sum;

$$\int_0^1 p(t)dr(t) = \sum_1^n p(i)\Delta r(i) \qquad \textbf{(4.6)}$$

n=Total prediction count

where Δr(i) is the change in the recall from i − 1 to i. And we can write Δr(i) as follows.

$$\Delta r(i) = \frac{y_i}{m} \tag{4.7}$$

m = total relevant documents count
y= hit success (0 or 1)

So that average precision formula is formed as below

$$\text{AVP} = \sum_{1}^{n} p(i)\frac{y_i}{m} \tag{4.8}$$

(Zhu, M, 2004)

As you can see, this measure interests in not only precision but also recall value. So when we recommend a ranked list to a user, average precision will work fine. If we look an example:

Table 4.1: Average precision calculations

Recommendation	**Success**	**p(i)**	**Δr(i)**
1	1	1/1	1/3
2	1	2/2	1/3
3	0	2/3	0
4	1	3/4	1/3

AVP calculated like as below

$$AVP = {}^{1}/_{3}\left({}^{1}/_{1} + {}^{2}/_{2} + {}^{3}/_{4}\right) = 0{,}92$$

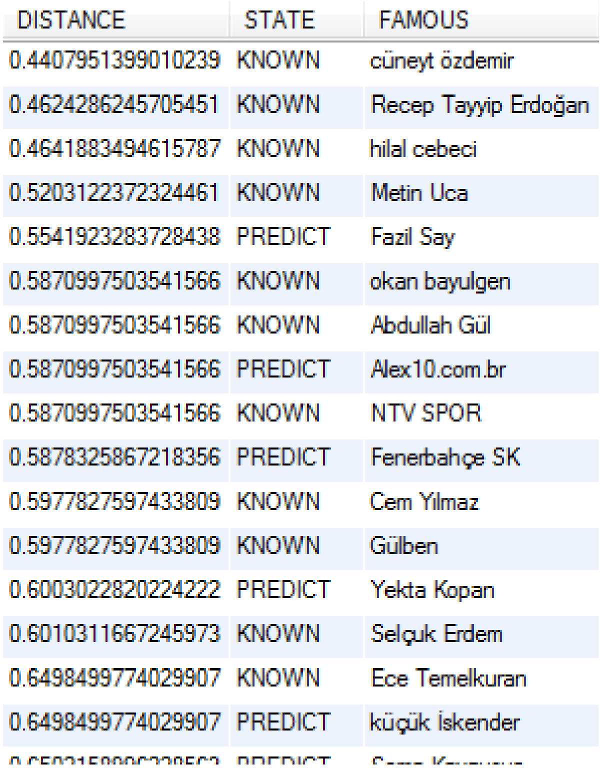
4.3 RESULTS

Our approaches are applied to two different social network data sets, Twitter and Tencent Weibo.

4.3.1 Twitter Social Network

As you can remember in experimental setup section, we have collected three generation people sets from Twitter with using Twitter API. Total set has been 2855 people, which collected using snowball sampling from fifty celebrities people from Turkey as start. When we analyzed the third generation people on this dataset, we have seen, 1000 of them already followed some of our start point celebrities' people. We used these people as test set in our study. We aimed to recommend these celebrities to the third generation people set. To do that when we were building the graph, we cut the existing connection between third generation people to these famous people (first generation). After that we calculated the distance from these users to all celebrities people on the new graph, and try to recommend these lost celebrities to user.

Figure 4.4: Distances



DISTANCE	STATE	FAMOUS
0.4407951399010239	KNOWN	cüneyt özdemir
0.4624286245705451	KNOWN	Recep Tayyip Erdoğan
0.4641883494615787	KNOWN	hilal cebeci
0.5203122372324461	KNOWN	Metin Uca
0.5541923283728438	PREDICT	Fazil Say
0.5870997503541566	KNOWN	okan bayulgen
0.5870997503541566	KNOWN	Abdullah Gül
0.5870997503541566	PREDICT	Alex10.com.br
0.5870997503541566	KNOWN	NTV SPOR
0.5878325867218356	PREDICT	Fenerbahçe SK
0.5977827597433809	KNOWN	Cem Yılmaz
0.5977827597433809	KNOWN	Gülben
0.6003022820224222	PREDICT	Yekta Kopan
0.6010311667245973	KNOWN	Selçuk Erdem
0.6498499774029907	KNOWN	Ece Temelkuran
0.6498499774029907	PREDICT	küçük İskender
[illegible]	PREDICT	[illegible]

You can see an example of recommendation list on above. All the celebrities are ordered with closest distance to source user. State column shows that the celebrity is already followed by test user (it showed as ‘know’) or a new predict of the system is. On result calculation of our approaches, only ‘known’ celebrities are taken as success hits. The new predicts was joined result calculation as failed.

On ordering ‘bundled’ approach list, we ordered with using two distance values. It has been ordered first with normal distance values (D_m) when it is ascending. Then we reordered with using bundled distance values (D_b) on whose of normal values are same. But at this time we ordered these with descending because of the result are gone well.

We used average precision methodologies to understand success of the work. The average precisions calculation was worked four times for each distance calculation approaches. It is also calculated for common followee which is used earlier study about recommendation person on Twitter.

The recommendation process is repeated for different recommendation list size like 1,3,5,12,15. Results are given on below table.

Table 4.2 : Average precision results

List Size	Mutual	Unit	Bund. Unit	Bund. Mutual	Common	Set Size
1	%37.1	%63	43.5%	%27	51.2%	1098
3	%54.1	%68	64.15%	%46	49.8%	827
5	%61.2	%71	69.92%	%53	55.4%	629
12	%68.3	%77	73.79%	%62	45.1%	254
15	%69.8	%82	77.79%	%65	56.3%	170

List Size: Recommendation list size
Mutual: Mutual distance
Bund. Unit: Bundled unit distance
Bund. Mutual: Bundled mutual distance
Common: Common Followers Method
FOF: Friend of friends

As you see, when recommendation list size grows, out test set size is decreasing. Because of our test people set don’t have same amount of celebrities on their followee list. For example we have only 170 people in our test set who follows 15 celebrities or higher.

An interesting point on the result is unit distance values. It has ben seen too high. But it is tricky. As you remember, we found many users from same distance when we use unit distance (all distance set 1 between two person). The orders of distance which are same come from database. So the result comes from database incoming order. So, it can be called as random. It would goes to very low average precisions values when the order of people changes on database.

Table 4.3: Sample recommendation list which found using unit distance

	distance
1st Celebrity	2
2st	2
3.th	3
4.th	3
5.th	3
6.th	4
7.th	4

4.3.2 Tencent Weibo Network

As you remember from data collection section, Tencent Weibo data was taken from KDD CUP competition. It has been given 1.3 million training sample. The each sample consists how act people when an item (they called item, because they not recommend only celebrities, also used channels, companies etc) recommend to follow. Does it follow or not. We used data like Twitter. We have only taken samples who accept to follow. And we ordered by user id.

The data is too huge. It needs too big calculation source for process. So we have taken 3 set, each consist 1000 people, from training set. And we tried to recommend the items to these. Also the network data is too huge for use our method. So we take parts of network for each user. Also for the same reason, it couldn’t use common followee methodology as competitor. We preferred to use Friends of Friend.

When we tried to recommend a celebrity on KDD Cup data, we use only bundled unit distance because of it success and simplicity.

In our Twitter working set, because of the celebrities count small, the methods could recommend the celebrities who are 2 unit distance away. But in here, because of too many items are exist and they are away from source between 1 to 2 unit distance, the method can not recommend items which are bigger than 2 unit distance away. So we recommend 3 three list which are 1,2 or 3 unit distance away. The results are given on Appendices A section.

5. CONCLUSION

On working with Twitter dataset, we tried different calculation metrics to find the distance with two Twitter data. As you can see, our approaches have successful to common followee method. In the next dataset we use Friend of friend methodology as a competitor but we didn’t prefer to use here. Because of common followee method has given more successful rates than FOF, which is proofed (Armentano et all, 2011)

When we look our distance metrics result, we see unit distance is most successful. But as we said in earlier, it is tricky. The results come from randomly which cause is from order of user’s on database table. If it is not regarded, bundled unit is most successful distance metric. Before to start to analyze that, it maybe better to look mutual results of our works

In mutual distance, we tried to combine some logical truths. The results have shown that it works better against to common followee methods especially on large recommendation list. So it can be said it is successful. We think the logical truth, which are based, is useful to new developments on followee recommendation. It can be improve by working on combine of the parameters. It can be one of continues works which born from this study. But it is important to aware, it is seen that will needs more calculation cost if it exceeds the bundled unit distance results.

We said that unit distance success results rolls on huge interval. But it needs to bundle to fix that. But what the means of bundling is? It can be important inference, which found in the study.

In logically, when a person follows a user in Twitter, it can be said the user has closeness to the followee user. And it can be easily says after that, the person have to closer or similar to other people who follows the followee. But as you remember, it has been ordered in our methodology with descending of total distance from our target user to the people who follow the target followee. So what is that mean? In here we have to remember our method first step, it is ordered with using unit o mutual distances values (D_m) first, after that it is used bundled distance values to reorder which have same D_m values. The method results say that looking descending bundled is working for the same distance away people recommendation. But why it is descending? In our opinion, the reason of total distance big is better that to recommend a person who has less closer followers to our target is better. So in other words if we recommend a person who does not follow by user's real friends it is better. It has surprise effect.

In working KDD Cup competition data, when we look our test results, we can say our method gives more successfully than friend of friend methodology. It is important because our method have big similarity with FOF. If the graph is cut from smaller distance away than two when it uses unit distance, our method turns almost to FOF method. Because FOF also interests with two distance away people. But we have to remember that while FOF ranks people with repetition count on total list, we rank them with bundling method.

As you can see in our results, our method shows that 2-3 distance away people are also important set to recommendation studies. FOF couldn't interest these. Each of time, we see that is ineffective with working 2-3 away data more than with working 1-2 away sets. But the results said that the 2-3 away sets data still consist powerful recommendation people set. Continue of this study, it can be research to combine these two recommendation lists.

REFERENCES

Books

Cormen, T. H., Leiserson, C. E., Rivest R. L., Stein C., 2009, *Introduction to Algorithms,* Third Edition, The MIT Press

Periodicals

Armentano , M. G., Godoy D., Amandi A., 2012, Topology-Based recommendation of Users in Micro-Blogging Communities, *Journal of Computer Science and Technology 01/2012*

Armentano , M. G., Godoy D., Amandi A., 2011, Towards a Followee Recommender System for Information Seeking Users in Twitter, *Proceedings of the International Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings; 01/2011*

Boyd, D. M., Ellison N. B., 2007, Social Network Sites: Definition, History, and Scholarship, *Journal of Computer-Mediated Communication, 13(1), article 11.*

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, 2010. K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. *4th International AAAI Conference on Weblogs and Social Media, ICWSM*

Hannon, J., Bennett, M. Smtyh, B., 2010, Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches, *4th ACM Conference on Recommender Systems (Rec-Sys'10),* pp. 199–206

Geyer W., Dugan, C., Muller, M., Guy I., 2009, "Make New Friends, but Keep the Old" – Recommending People on Social Networking Sites, *CHI '09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 201-210

Kwak, H., Lee, C., Park, H, Moon, S., 2010, What is Twitter, a Social Network or a News Media?, *WWW '10 Proceedings of the 19th international conference on World wide web,* pp. 591-600

Lo, S. and Lin, C., 2006. WMR—A Graph-based Algorithm for Friend Recommendation, *WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence,* pp. 121-128

Lu, C., Lam,W., Zhang, Y., 2012, Twitter User Modeling and Tweets Recommendation Based on Wikipedia Concept Graph, *Intelligent Techniques for Web Personalization and Recommender Systems AAAI Technical Report WS-12-09*

Vinh, N. X., Epps, J., Bailey, J., 2009, Information Theoretic Measures for Clustering Comparison: Is a Correction for Chance Necessary?, *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM*. pp. 1073–1080

Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J., 2011, Who says what to whom on Twitter, Proc. WWW ACM

Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa. H., 2010, TURank: Twitter user,ranking based on user-tweet graph analysis, *Web Information Systems Engineering, volume 6488 of LNCS*, pp. 240–253

Other References

Getting, B., 2007, Basic Definitions: Web 1.0, Web. 2.0, Web 3.0, Practical Ecommerce, http://www.practicalecommerce.com/articles/464-Basic-Definitions-Web-1-0-Web-2-0-Web-3-0, [accessed 17-08-2013]

Holt R., 2013, Twitter in Numbers, http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html, [accessed 17-08-2013]

Kddcup, Kddcup from Tencent, http://www.kddcup2012.org/, [accessed 18-07-2013]

O'Reilly, T., 2005, What is Web 2.0, Oreilly, http://oreilly.com/web2/archive/what-is-web-20.html, [accessed 17-08-2013]

Smith, C., 2013, How many people use the top social media, app, services?, DMR, http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/, [accessed 17-08-2013]

Tam, D., 2013, Facebook by the numbers: 1.06 billion monthly active users, http://news.cnet.com/8301-1023_3-57566550-93/facebook-by-the-numbers-1.06-billion-monthly-active-users/, [accessed 17-08-2013]

Twitter, 2013, Abdullah Gul's Twitter page, https://twitter.com/cbabdullahgul, [accesed 18-07-2013]

Wikipedia , Twitter , http://en.wikipedia.org/wiki/Twitter, [accessed 18-07-2013]

Wikipedia, Snowball sampling, http://en.wikipedia.org/wiki/Snowball_sampling, [accessed 18-07-2013]

Zhu, M, 2004, Recall, Precision and Average Precision, Working paper, http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf

APPENDICES

APPENDICES A: Experimental Result Details

Table A.1: One-two distance away average precision values for first people set

List Size	Av. Pre.	Set Size
1	2,30	827
3	8,72	217
5	16,45	118
12	29,06	53
30	30,89	40

Table A.2: Two-three distance away average precision values for first people set

List Size	Av. Pre.	Set Size
1	1,82	827
3	6,60	217
5	12,19	118
12	19,07	53
30	20,55	40

Table A.3: Three-four distance away average precision values for first people set

List Size	Av. Pre.	Set Size
1	0,24	827
3	1,38	217
5	2,33	118
12	5,97	53
30	6,38	40

Table A.4: Average precision values on FOF for first people set

List Size	Av. Pre.	Set Size
1	2,56	827
3	9,84	217
5	15,03	118
12	20,84	53
30	24,24	40

Table A.5 : One-two distance away average precision values for second people set

List Size	Av. Pre.	Set Size
1	0,99	827
3	4,76	217
5	6,59	118
12	16,51	53
30	18,31	40

Table A.6: Two-three distance away average precision values for second people set

List Size	Av. Pre.	Set Size
1	1,23	827
3	3,87	217
5	4,75	118
12	10,22	53
30	11,39	40

Table A.7: Three-four distance away average precision values for second people set

List Size	Av. Pre.	Set Size
1	0,62	827
3	1,54	217
5	1,13	118
12	4,25	53
30	5,41	40

Table A.8: Average Precision values with FOF for second people set

List Size	Av. Pre.	Set Size
1	1,24	827
3	8,47	217
5	14,01	118
12	14,41	53
30	11,14	40

Table A.9 : One-two distance away average precision values for third people set

List Size	Av. Pre.	Set Size
1	1,66	827
3	6,96	217
5	10,77	118
12	19,91	53
30	24,95	40

Table A.10: Two-three distance away average precision values for third people set

List Size	Av. Pre.	Set Size
1	1,02	827
3	4,85	217
5	7,73	118
12	14,85	53
30	19,01	40

Table A.11: Three-four distance away average precision values for third people set

List Size	Av. Pre.	Set Size
1	0,25	827
3	0,94	217
5	2,05	118
12	4,83	53
30	7,55	40

Table A.12: Average Precision values with FOF on third people set

List Size	Av. Pre.	Set Size
1	1,81	827
3	6,13	217
5	8,63	118
12	11,92	53
30	14,40	40