Content and User Recommendation In Social Networks

Nathan Hale

Exeter College

University of Oxford

A dissertation submitted in partial fulfilment of the degree of

Master of Science

Summer 2012

Abstract

Social networking services have exploded in popularity over the past several years and the amount of content available on those services has grown correspondingly. With this much available content it is easy for users to be overwhelmed. Thus, it is in the interests of both the service operators and the users of such services to have access to effective methods for recommending novel content and users to connect with.

This dissertation proposes one such method, adapting and extending a method used in web search to the problem of recommending both users and content within social networks. It also demonstrates a novel (??) procedure for incorporating links that would otherwise render a social graph non-bipartite into a proper bipartite graph.

It is then demonstrated that this method is effective at recommending both content and users that a particular distinguished user may be interested in. This efficacy is demonstrated through both user evaluation and calculated metrics. A number of different variations to the methods are explored in order to refine the results further.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Over the past decade, a number of internet services dedicated to 'social networking' have been created. These services allow people to share what they are doing with their friends and to see what their friends have done. They also frequently allow users to see the publicly shared content of other users who they find interesting, perhaps because they know them, perhaps because that person is a celebrity, or perhaps because that person consistently shares content that is interesting in some way such as breaking news, funny jokes, or interesting articles.

The popularity of these social networking services has grown tremendously over the past several years, and that growth has been accompanied by a corresponding growth in both the amount of content available on these networks and the number of services available. Facebook, the largest social network, had over 901 million monthly active users at the end of March 2012 and had over 125 billion friend connections between those users[1]. Twitter is another popular social network which has over 140 million active users as of May 2012, with more than 383 million users[2] having created accounts at some point since the service was founded in 2006. Dozens of other smaller social networks have been created as well, aimed both at general audiences and niche audiences.

With this explosion of content and users it has become much more difficult for users to find novel content of interest to them and to find new users to connect with who might consistently produce such content. Because this may negatively impact the user experience it is in the best interests of both users and the operators of social

[1]http://newsroom.fb.com/content/default.aspx?NewsAreaId=22
[2]http://www.guardian.co.uk/technology/2012/may/15/twitter-uk-users-10m

networking services to provide methods to allow users to easily find interesting content and users.

Existing approaches to content and user recommendation (as described in Section 2.2.4) have focused on either a network-based approach, which examines the topology of the network and the connections between users, or a content-based approach, which seeks to take semantic and syntactic information from the content itself and use that to recommend other content. In many cases these approaches are still hamstrung by the same issues that vex individual users, namely the vast amount of data that exists in these social networks which can make analysing them a very difficult task.

Most social networks already provide some form of content recommendation to their users, usually in the form of a list of other users that they may know or be interested in which is created based on the connections that the user already has, i.e. a network-based approach. This is a step in the right direction, but it recommends only users, failing to recommend interesting individual pieces of content. On existing social networks where some content is shown to users, content is usually only recommended in terms of what is popular amongst all users, without considering the individual user's own interests.

The purpose of this dissertation is to provide a novel and more effective method of recommending both users and content to users of social networking services despite the large amount of data involved. This dissertation will use a combination of a network-based approach and a content-based approach since both have been shown to have value. Taking both methods into account is also very helpful in making the process faster because it is possible to recommend both users and content while running only one algorithm.

In developing this method, it is hoped that such a technique could be put to use either directly by a social networking service or by a third party in order to improve the experience for their users by leading them to more interesting and relevant content. Furthermore, existing techniques for evaluating the effectiveness of such recommendations are very lacking, so it is hoped that new techniques can be used to improve the ability to evaluate both this scheme and future research in the field.

1.2 Structure

This dissertation is structured into six chapters, including this introduction describing the motivation behind the project and the problems it is intended to address.

Chapter 2 gives important background on what social networks are, with particular focus on features that most networks have in common. Two networks in particular, Facebook and Twitter, are discussed in depth to show how these generic features are implemented. This also provides insight on which of these features may be useful in recommending content and users. After giving this background information, existing research on social networks and content recommendation is examined as a means of seeing which network features might be useful for developing recommendations and what techniques have already been used.

Chapter 3 describes the general methodology used by this dissertation. The method is described without reference to any particular social network to emphasize that the technique being used can be applied to nearly any social network having the general features described in Chapter 2. The suitability of this method, known as the Co-HITS algorithm, to social network graphs is described, as are the methods used to transform the graph of the social network into a graph suitable for use by this algorithm.

Chapter 4 describes the specific implementation used by this dissertation. It describes how the data about the social network was selected and obtained and it describes how the challenges of the huge size of this data were dealt with. It then describes the implementation of creating the social graph and implementing the Co-HITS algorithm to run efficiently on this large graph.

Chapter 5 describes the results obtained using this method. Evaluating the results is not straightforward since there is not yet a canonical method of evaluating results in this research area. A number of techniques are proposed and the results of the algorithm for each evaluation technique are presented. Additionally, this chapter describes a number of modifications to the basic algorithm for recommendation and shows how each of these variations affects the results.

Chapter 6 concludes the dissertation. This chapter provides a discussion on how the performance of the specific method developed here could be improved and what future work could be done in this area to improve the overall method by modifying it more fundamentally. The dissertation ends with some concluding remarks on the method developed, its efficacy, and its suitability for real-world implementation.

Chapter 2

Background and Related Work

This chapter begins by describing what social networks are and what features one can expect to find in them. In particular, the features of two of the largest social networks, Facebook and Twitter, are discussed in detail, including how those features might be useful in recommending novel content to users. Using this description for background, the discussion then moves to research that has already been done in this field and which techniques for content recommendation have shown promise so far.

2.1 Social Networks

Facebook and Twitter are certainly the most popular social networking services, with more than 1 billion user accounts between them. There are dozens of other social networks, however, including many with millions of users such as Google+, Pinterest, and LinkedIn. It would be impractical and of limited value to discuss the nuances of all of these services and to what degree they implement particular features, but some background is necessary in order to understand the work undertaken in this project. Because Twitter and Facebook are the most popular social networking services and demonstrate all of the common features of social networks this discussion will be limited to only these two services.

Before describing these two services and their features in detail, it is valuable to begin with a brief discussion of the general features that are common to all social networks. Examining these general features provides a good basis for comparison of how the general features are implemented by the two test cases. It is also helpful as a reference to show that while the study undertaken for this project was specific to one particular social networking service, the features used were actually quite generic and could easily be extended to other services.

2.1.1 What are Social Networks?

At its core, a social network is a set of users, a set of edges connecting those users, and content produced by those users. Depending on the network and the relationships between users, the edges may be either directed or non-directed, and sometimes a given network may include edges of both types. A non-directed edge indicates that both users are subscribed to each other's content while a directed edge from user A to user B indicates that while user A is subscribed to content produced by user B and will see all public content from user B, user B will not see any content from user A.

The types of content produced also vary depending on the network. Twitter famously limits content to 140 characters of text, a limit designed to ensure that it can fit within the standard text message size limit of 160 characters. Still, those 140 characters can contain links to any other sort of content, and as the service has developed various conventions have been created in an ad hoc manner by its users in order to mark up the text to provide additional meaning. Facebook users, meanwhile, can share much longer text updates, photos, links, and videos.

Another common feature of social networks is the ability for users to create profiles to tell other users about themselves and to express their interests and desires. The importance of that profile depends on the network, and in some cases it is the primary form of content produced by users. LinkedIn, for example, is a social network designed for professional networking and job-seekers, and its user profile pages are essentially just extended CVs.

Social networks also frequently provide the ability for users to re-share or forward content from other users so that their connections can see that content. This allows links and pictures to make their way across the network and be seen by users who are not themselves connected to the original user who produced a piece of content. And, importantly for the task of recommending content, it provides a good insight into the types of content that the person re-sharing the content finds interesting.

The degree to which networks support these features varies widely, but most networks—and certainly the majority of the most popular networks—implement all of these features.

2.1.2 Facebook

Facebook was originally founded as a way for students at Harvard University to keep in contact with each other. Early versions of the website were rudimentary compared to the current behemoth—the only real features were the ability to become friends with other users and to fill out a profile indicating the user's birth date, home town, current location, interests, and similar biographical details. From those early beginnings the network expanded first to other universities and then to secondary schools before membership was eventually open to everyone.

As one of the oldest social networks which still enjoys a large user-base, Facebook has evolved quite a bit over the years since it was created. As other social networks have challenged it by introducing new features, Facebook has responded by adding versions of those same features with slight modifications to adapt them to the Facebook environment. Over the years of growth the modern version of Facebook has slowly taken shape as features such as photo sharing, the news feed, and the 'like button' have been introduced.

2.1.2.1 Features

The Facebook network is built around the concept of friends. The edges in the Facebook network are connections between users who (mostly) know each other in real life and are bi-directional. One user sends a friend request to a second user and if the second user accepts then the connection is created and each user can see content produced by the other.

Suppose that there is a user Alice who is friends with a user named Bob but not with a user named Carol. Bob, however is friends with Carol. Alice might share a piece of content C_A with her friends, in which case Bob will see the update but Carol will not. Similarly, if Carol shares a piece of content C_C, Bob will see it but Alice will not. Finally, if Bob shares a piece of content C_B, both Alice and Carol will see it since both users are friends with Bob. This visibility is shown in Figure 2.1.

The primary forms of content on Facebook are status updates, photographs, and links. Status updates are just text updates, sometimes telling other users what that person is doing, sometimes asking for advice, and sometimes just expressing emotions. Photographs and links are both self-explanatory, and are usually accompanied by captions or explanatory text from the user describing his or her thoughts on the content being shared. Each of these types of content is accompanied by a set of comments made by friends. Additionally, friends may 'like' the content by pressing the 'like' button to indicate some sort of agreement, support, or genuine like of the associated status, link, or photograph without needing to leave a comment.

More recently, perhaps in response to competition from Twitter, Facebook has introduced the ability to share (forward) content from another user and to subscribe

$$\begin{tikzcd}
Alice \arrow[r, shift left] \arrow[d, dashed, leftrightarrow] & C_A \arrow[l, dotted, shift left] \arrow[dl, dotted] \\
Bob \arrow[r, shift left] \arrow[d, dashed, leftrightarrow] & C_B \arrow[l, dotted, shift left] \arrow[ul, dotted] \arrow[dl, dotted] \\
Carol \arrow[r, shift left] & C_C \arrow[l, dotted, shift left] \arrow[ul, dotted]
\end{tikzcd}$$

Figure 2.1: An example of how visibility of content works on Facebook. The dashed edges indicate the friendship connections between users. Directed edges from users to content indicate authorship, and dotted directed edges from content to users indicate which users can see that content.

$$\begin{tikzcd}
Alice \arrow[dr] \arrow[d, dashed, leftrightarrow] & \\
Bob \arrow[d, dashed, leftrightarrow] & C_A \arrow[ul, dotted] \arrow[l, dotted] \\
Carol &
\end{tikzcd}
\qquad
\begin{tikzcd}
Alice \arrow[dr] \arrow[d, dashed, leftrightarrow] & \\
Bob \arrow[r, shift left] \arrow[d, dashed, leftrightarrow] & C_A \arrow[ul, dotted] \arrow[l, dotted, shift left] \arrow[dl, dotted] \\
Carol &
\end{tikzcd}$$

(a) Before Bob reshares C_A **(b)** After Bob reshares C_A

Figure 2.2: Visibility of content on Facebook before and after content is reshared. The dashed edges indicate the friendship connections between users. Solid edges between users and content indicate content authorship and resharing, and dotted edges between content and users indicate visibility. This figure does not include the situation where Alice subscribes to the public updates of Carol.

to public posts of another user or group, essentially creating a directed edge between those users. Returning to the example of Alice, Bob, and Carol, if Alice shares content C_A, Bob will see it but Carol will not since Alice and Bob are friends while Alice and Carol are not. Bob can re-share this, however (provided that Alice's privacy settings allow this) at which point Carol will be able to see it. This visibility relationship is shown in Figure 2.2. Additionally suppose that Carol subscribes to the public updates of Alice. In this case, if Alice shares a photograph publicly, Carol will be able to see it without Bob having to re-share it since she subscribes to Alice's public updates.

As alluded to above, various privacy controls have been put in place over the years to give users control over who can see what content. Though the original idea of Facebook was that friends on Facebook would correspond to friends in real life, the actual mode of use didn't follow and people often found themselves in nominal 'friendships' with acquaintances or people from different social circles with whom they didn't necessarily want to share all of their content. Thus, users can choose to share specific content only with specific sets of people and to restrict the re-sharing of their content. Additionally, users must decide whether posts will be public, visible only to friends, or visible only to specific friends.

Other popular features of Facebook allow users to create events and send invitations to their Facebook friends and the ability to run applications and games developed by third-party developers which leverage the user's social data to provide additional experiences.

2.1.2.2 Content Recommendation

Many of the basic features of Facebook would be very applicable for use in a content recommendation scheme. The basic edges between users present an obvious starting point, but other features could be leveraged to put a more accurate weight on edges between users. Users who frequently comment on one another's posts are more closely connected than users who never interact, for example.

Similarly, if a user likes a post or comments on it then it indicates that the content itself was relevant to that user, information which could be used to recommend future content of interest to that user. Attending common events would be another method of determining just how close two nominal 'friends' actually are—users who go to the same events in the real world are obviously much more likely to be good friends than users who do not.

The large amount of information on a Facebook user's profile is another piece of information that could be leveraged to recommend content. If two users declare similar interests on their user profile then it would be more likely that those users would be interested in content generated by one another.

Facebook itself has developed a system based on many of these components to help it recommend other users that it believes you may know and want to become friends with. One such project is described in [?].

The major drawback to studying content recommendation on Facebook is that the data is difficult to access. Most of the content created by users and the edges between those users are private and thus inaccessible to researchers. Therefore most studies focusing on Facebook, including that in [?], are undertaken at least in part by employees of Facebook participating in research and development to help advance the interests of the company.

The private nature of this data made it unsuitable for study in this dissertation, though many of the other components of the network would have provided a rich set of features to use in content recommendation. Facebook is discussed here to show that it has analogues to most of the major features of Twitter and that all of the research undertaken in this dissertation could easily be applied to Facebook if the data were available.

2.1.3 Twitter

2.1.3.1 History

Created in 2006, Twitter is a social network noted for the brevity of its content. The basic concept is simple: users provide updates on what they are doing or thinking in up to 140 characters. The basic interface has changed very little since its inception and the feature set is still very similar to its beginnings, in marked contrast to the large number of features that are integrated into Facebook.

The original intent was that users could update their status and receive statuses from friends using SMS messages from their mobile phones, and thus the length of any given Twitter message was limited to be smaller than the maximum size of an SMS message, 160 characters.

The nature of the network began to change as smartphones—mobile phones capable of running applications and communicating with the internet—became ubiquitous following the release of Apple's first iPhone in 2007. Twitter provides an extensive Application Programming Interface (API), which has allowed application developers to create rich and visually pleasing interfaces to the service that have helped boost its popularity by making it even easier to use. As smartphones grew in popularity the need for and popularity of status updates via SMS faded, though the 140 character limit has remained.

Today, Twitter has over 140 million active users worldwide, and those users produce hundreds of millions of status updates each day. As a result, numerous attempts have been made to make sense of this vast source of public data about the feelings and activities of millions of people around the world. Research into this area is discussed in greater depth in Section 2.2.2.

$$\begin{tikzcd}
Alice \arrow[r, shift left] \arrow[d, dashed] & C_A \arrow[l, dotted, shift left] \\
Bob \arrow[r, shift left] & C_B \arrow[l, dotted, shift left] \arrow[ul, dotted]
\end{tikzcd}$$

Figure 2.3: An example of how visibility of content works on Twitter. The dashed edge indicates that Alice follows Bob, but Bob does not follow Alice. The dotted edges represent who will see each piece of content in their Twitter stream. Note that Alice or Bob could always see the other users's content by visiting their Twitter feed, regardless of any follower/followee relationship between them.

2.1.3.2 Features

Twitter updates are commonly known as *tweets*, and the act of producing one is known as *tweeting*. Within the research literature on Twitter, one person's *Twitter feed*, the collection of all of the tweets that they have produced, is commonly called a microblog in reference to the previously mentioned 140 character limit imposed on each tweet. Edges between users are directed. If a user Alice follows a user named Bob but Bob does not follow Alice, Alice will see all of Bob's tweets, but Bob won't see any of Alice's tweets. This is shown in Figure 2.3.

The terminology which will be used by this dissertation will be to refer to Alice as one of Bob's *followers* and to refer to Bob as one of Alice's *followees*. In contrast to Facebook, most tweets are visible to the public; the act of following Bob simply causes Bob's tweets to appear along with the tweets of Alice's other followees in Alice's *Twitter stream*, which is the set of tweets of everyone that Alice follows. If Bob cared to see all of Alice's tweets he could simply visit her Twitter feed, which collects all of Alice's tweets.

Initially, this was the complete feature set of the service; it was possible to post updates and to subscribe to the updates of others. As Twitter became more popular its users began to develop a shorthand for indicating things to other users. As that shorthand became more codified many of these features were given native support by Twitter. The most important of these features are hashtags, mentions, and retweets.

Since the service only permits the sharing of text directly, it does not internally support sharing pictures, videos, or links. However, web addresses contained within the text of a tweet are highlighted and automatically turned into links, allowing pictures, videos, or any other content to be shared by hosting it externally and providing a link. A number of image hosting websites were developed specifically to support user uploads of photos which can then be linked from Twitter. The biggest issue with linking is that many web addresses are by themselves more than 140 characters, and even those which are not often would not leave much room for comment if included in their entirety. As a result, URL shortening services (such as bit.ly, t.co, and ow.ly), which take a complete URL and hash it down to 5 or 6 characters, have become popular. An example of such a link is seen in Figure 2.5.

Hashtags Hashtags were created organically by users of Twitter[1] as a way of tagging tweets on a similar topic. Thus, someone might tag a tweet about the University of Oxford by appending '#oxford' to the message. Or, in keeping more with the brevity that is prized within tweets, they might integrate the hashtag into the message itself, something like 'Currently visiting the University of #Oxford'. Figure 2.5 contains a screenshot of a tweet which includes a hashtag.

Tagging tweets in this way clusters tweets which all discuss a particular topic and makes it easy to search. As this feature became widely adopted by users it was integrated to have native support by Twitter. Today, such hashtags appear as highlighted links within a tweet and clicking the link brings up other tweets that have used the same hashtag recently.

Twitter collects these hashtags and provides a list of topics that are currently popular amongst users. While these hashtags must not contain spaces or punctuation in order to be parsed properly, they are often stylized to contain entire phrases and questions to which users provide answers, such as #MyFavouriteSongIs for discussing music. While certainly commonly used for banalities and idle chatter, these hashtags can also take on great importance for providing information during breaking news events. For example, during the Iranian political unrest of 2009, hashtags such as #IranElections became popular for users in that country to disseminate information about protests [?].

Mentions and @replies Mentions are another important feature of Twitter, and like hashtags, they also started out as a convention amongst users before being adopted as a native feature by Twitter[?]. Mentions are simply a way of directing a tweet to the attention of a particular person. This is accomplished by pre-pending an @ symbol to their user name and including it in the tweet. Thus, to mention Bob, Alice would type '@bob' in her tweet. Mentions are collected by Twitter and can be seen

[1] https://support.twitter.com/articles/49309-what-are-hashtags-symbols

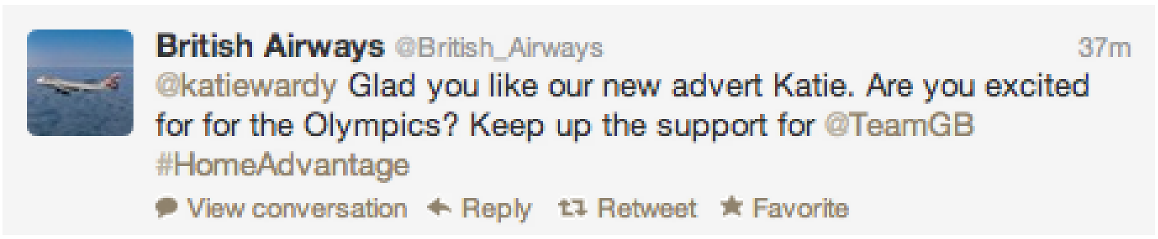
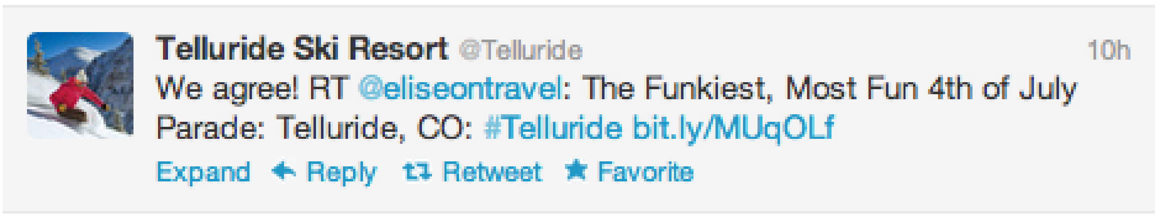

British Airways @British_Airways 37m
@katiewardy Glad you like our new advert Katie. Are you excited for for the Olympics? Keep up the support for @TeamGB #HomeAdvantage

Figure 2.4: This tweet is an @reply to a user named 'katiewardy' but also contains a mention of a user named 'TeamGB', the official Twitter page for the United Kingdom's Olympic Team

even if the user being mentioned, henceforth referred to as the *mentionee*, does not follow the mentioner. These mentions do not appear on a person's main Twitter feed, but are instead accessed from a special mentions page.

One distinction that is made is between an @reply and a mention. An @reply is the colloquial term given to mentions where the @username syntax appears at the beginning of the tweet as in the following sample tweet from Alice: "@bob Hope to meet you after the presentation today!". Because this is an @reply, this tweet would not show up in the main Twitter streams of Alice's followers (unless they also follow Bob). Since Bob doesn't follow Alice, this @reply also would not show up in his Twitter stream, only in his mentions. In contrast, a mention is a tweet that contains the @username syntax anywhere except at the beginning of the tweet, such as in this example tweet from Alice: "Hoping to meet @bob after the presentation today!". These tweets are visible on the Twitter streams of Alice's followers just like any other tweet, but are also collected for Bob along with all other tweets that mention him. Figure 2.4 is a screenshot of a tweet which includes both a mention and an @reply.

Retweets The final major feature of Twitter is yet another syntactic convention that was developed naturally by users and then given native support by Twitter: the retweet. A retweet is when a user repeats a tweet by another user to pass it along to their own followers, analogous to the re-sharing feature on Facebook. The common syntax is 'RT @username [original tweet]', with the RT standing for retweet. The retweet is often sent along with a comment from the person retweeting about how they feel about what was said. While no specific convention exists for where to place this comment, two conventions often seen are to place a comment before the RT and to place it after the original tweet with some indication of the divide between the original tweet and the comment such as '//' or '...'. Figure 2.5 contains a screenshot


Telluride Ski Resort @Telluride 10h

We agree! RT @eliseontravel: The Funkiest, Most Fun 4th of July Parade: Telluride, CO: #Telluride bit.ly/MUqOLf

Figure 2.5: An example of a Twitter retweet. Here, the resort has forwarded the positive tweet by user 'eliseontravel' along with a comment of 'We agree!' to their own followers. This tweet is also an example of a tweet containing a hashtag which was used in the original tweet: '#Telluride'. The 'bit.ly/MUqOLf' at the end is a shortened link to an external website.

of a tweet which includes a retweet which follows the convention of placing a comment before the RT.

2.1.3.3 Content Recommendation

All of these features of the Twitter network are useful in content recommendation, and many of them have been used in existing studies as described in section 2.2.4. The primary advantage of the Twitter network as an object of study, however, is that the data is freely available and accompanied by an extensive API to allow the data to be queried and searched. This API has been used by numerous researchers to download large sets of tweets and the social graph accompanying it indicating which users follow which other users.

In fact, while studying Twitter it is actually a case of having far too much data to deal with easily than of not having enough data. Even though each tweet is no more than 140 characters, if the assumption is made that 100 million tweets are produced on a particular day (a very low estimate with today's usage), then that single day will contain roughly 14GB of data. Fortunately, previous researchers such as [?] and [?] have already created some parsed versions of the data that can be used. More recently, a track of the Text Retrieval Conference (TREC)[2] has been opened with regards to microblog research along with an enormous database of tweets from 2011. Though these datasets are all extremely large and thus useful for research, they still represent only a small fraction of the total data present.

This vast amount of readily available data meant that Twitter was by far the best option to study in this project, even though most of the concepts would be simple to

[2]http://trec.nist.gov/tracks.html

apply to other social networks given the data since most of them have their major features in common with Twitter. As such, the remainder of this chapter will focus on Twitter since it is the object of study for this dissertation.

2.2 Existing Research on Social Networks

Twitter's simple and compact format and vast quantity of data has made it a popular target for research. The main topics of research are on the nature of the Twitter network itself, whether it is possible to gather useful information from Twitter, methods for determining what topics interest a particular Twitter user, and methods for ranking and recommending both tweets and users.

2.2.1 The Twitter Network

One of the earliest studies of the Twitter network was that of [?], from 2007, only a few months after the service launched. The study was undertaken when Twitter had fewer than 100,000 users and during the two month study period only 1.3 million tweets were made. Clearly things have changed since this study, but it does still provide a number of interesting results. For example, they found that Twitter showed a high degree correlation—users with a large number of followers also tended to have a large number of followees. In the same vein, the distribution of indegree and outdegree had a similar power exponent to that of the web and the blogosphere.

The most important findings of [?] were the different categories of users and of content. They found that the content produced mostly fell into four different categories: daily chatter, conversations, sharing information, and reporting news. Importantly, conversation comprised nearly one-eighth of the tweets in their dataset. Presumably if a similar study were repeated today it would find a fifth content category: spam. Users were found to fall into three categories: information seekers, users who may tweet infrequently but still commonly check the site for the tweets of others; information sources, users who post information that others find valuable; and friends. At the time of the study friendship links constituted most of the links between users, though with more than 100 million users today, this is unlikely to be the case.

A later study by [?] indirectly speaks to how much growth Twitter underwent between 2007 and 2010. At the time of that study in August 2009 Twitter had grown to nearly 54 million active users with 1.9 billion links between them and more than 1.7 billion tweets. The threshold for popularity investigated here, one million followers, is more than ten times greater than the total number of users in the 2007 study of [?].

[?] found that while Twitter users with one million followers might seem intuitively to be the most influential, the number of followers alone was in general a poor measure of the influence that a user has.

[?], a 2011 study, investigated which links in the twitter network (e.g. follower links, retweet links, etc.) were most likely to preserve topical relevance. They found that the most important link type for preserving this topical relevance is retweet links. Crucially, they also found that "traversing even a single follow link dramatically reduces the probability of topical relevance". In carrying out this study they repeated some of the experiments of [?] and found that most of the results from there still held true, despite the massive growth of the service.

A very important study for the purposes of this project is that of [?], an in-depth examination of characteristics of the Twitter network. In many ways it is similar to the early 2007 study of [?], but with a much more mature network. They studied things such as how likely a user was to reciprocate when someone followed them, the relationship between followee/follower numbers and number of tweets, the distribution of followee/follower numbers, the nature of trending topics, and the reasons behind and impacts of retweets, among other topics.

An important result from this study is that the number of followers/followees that users have behaves according to a strong power law with a long tail. There are hundreds of thousands of users with fewer than ten or twenty followers or followees, while only 40 users at the time of the study had more than one million followers. And of the users, 67% were not followed by any of their followees! The results also showed that how many tweets a user had was a strong predictor of both how many followers and how many followees that person had, at least up to about 100 follower or followees.

Those results are interesting, but the crucial outcome of [?] for this project is the large dataset that was produced in order to find those results, of which the complete social graph (more than 1.2 billion edges) which was included and used for this project is the most important component. The social graph from this study was the main one used in the project, as discussed in Section 4.1.

2.2.2 Gathering Information From Twitter

Given the vast size of the Twitter network and the vast amount of data produced by its users in talking about what they are doing and how they feel, it is not surprising that studies have been undertaken to try to mine this data to gain more insight into what people believe and what they're discussing.

The study undertaken by [?] attempted to predict the results of opinion polls based only on Twitter data. This process makes perfect sense when one considers that an opinion poll is really just an attempt to determine how a large population feels about something by sampling a small portion of that population and asking them how they feel. Analysing Twitter data, then, is akin to asking a much larger sample of that population.

Using a dataset of one billion tweets collected in 2008 and 2009, the research investigated three different possible opinion applications: election polls, job approval polls, and consumer confidence. For each application, tweets were analysed from a relevant period of time, such as in the run-up to the presidential election. For each tweet that contained certain trigger words that indicated relevance to a particular poll, the sentiment was analysed by simply seeing if a tweet had positive or negative words. All of the scores both positive and negative for a given period were summed and amalgamated and then compared to opinion polls from the same period of time.

The results were not perfect but did show a strong correlation for both job approval ratings and consumer confidence. Predicting election polling proved to be a more difficult task, and the correlation on those polls was not nearly as good. The strength of these results based on a very rudimentary method of sentiment analysis suggests that better techniques for analysing the tweets could improve the results.

A similar study was performed in [?], this time to try to predict the box office gross revenue of films. This study also reported some promising results from mining the sentiment of vast numbers of tweets in order to determine peoples' opinions.

2.2.3 Determining User Interest

A number of studies have focused on determining which topics are most often discussed by a user and which topics a user is most interested in reading about. The techniques from this area are very useful to the work undertaken in this project, particularly in determining the similarity of different pieces of content, e.g. as used in Section 3.3. Section 4.3.3 describes the use of named-entity recognition (NER) for purposes of determining which tweets are related, but any of the techniques used in this section could be used instead, probably with better results, albeit at the expense of greater run times.

One study, [?], used an ontology-based approach to determine a user's topics of interest. In this case, the ontology was the category structure of Wikipedia. For each tweet by a particular user, named entities were found using a named-entity recognizer. Those entities were then looked up in Wikipedia in order to disambiguate them and determine their categories. After repeating this process for all tweets by a user the process yields a pretty good set of the top ten topics of interest to that user.

A more mathematical approach was taken by [?]. They used a partially supervised machine learning algorithm and achieved good results when categorizing the tweets. Their particular machine learning approach was Labelled Latent Dirichlet Allocation, which finds distributions of words which tend to occur in similar documents. These sets of similar words are taken to be the topics, and the topics of interest to particular users can determined by analysing their tweets for these words. The authors were unsure at the outset if this technique, commonly used for long documents such as news articles, would work for documents as short as tweets, and their results indicate that it is quite successful.

A more recent method, [?], attempted to determine the topics of interest for a particular distinguished user by finding which of their followees were most important and setting the topics discussed by those users as the topics of interest for the distinguished user. The study used an approach based primarily on the Twitter network, including follower/followee relationships and retweet/mention relationships in order to determine the list of top users. As in other studies, retweets were found to be the most reliable and were thus weighted the most heavily. Upon identifying the top users, they took an approach similar to that of [?], using Wikipedia to look up terms and disambiguate them, though they used nouns returned from a part-of-speech tagger rather than named-entity recognition.

2.2.4 Content and User Recommendation

Two main approaches exist in current research on recommending content and users, the network-based approach and the content-based approach. In the network-based approach, the connections between users and their interactions with one another via retweets and mentions form the primary basis for recommendation. In the content-based approach, meanwhile, the analysis is primarily focused on the tweets themselves, such as by determining which tweets are most similar to one another. The two approaches are by no means mutually exclusive and are often used in complementary ways.

The research described here shows that when recommending users and content, each approach has had some degree of success.

2.2.4.1 Recommending Users

[?], the wide-ranging study that provided the dataset used for this project, included some analysis of potential network-based techniques for ranking users, though it was not the primary focus of the study. They used several techniques: ranking users in terms of number of followers, running the PageRank algorithm on the social graph, and ranking users in terms of how many times their content had been retweeted. As expected based on the results of [?], the number of retweets was the best of these ranking schemes. PageRank and the number of followers metric produced nearly identical results, suggesting based on the results of [?]—namely that a user's number of followers is a poor measure of their influence—that PageRank is not a good technique to use for recommending users.

Still, PageRank has been used by other studies, such as in [?]. They called their method TwitterRank because it differed slightly from original PageRank. Rather than visiting all users with uniform probability over links as in the random walk of PageRank, TwitterRank performs a topic-specific random walk using topics derived using Latent Dirichlet Allocation. Their results indicate that this modification allowed TwitterRank to outperform both the PageRank metric and a metric based on the indegree of nodes when ranking tweets. The performance improvement over page rank was small, however.

A hybrid approach utilizing both network and content information was taken by [?]. The goal of the research was to provide user recommendations in some sort of a ranked order. A network based approach was taken first in order to discover potentially interesting users. From that unordered list of potentially interesting users a ranked list is created by comparing the users based on the content that they produce to a reference document of what the user is interested in. Two strategies specifically used for building the user profile are the user's own tweets and the tweets of the user's followees.

Another piece of work focused on recommending users is presented in, [?], which recommends users based purely on content. As with a number of other approaches, the major choice explored is what to use as the reference content to which other users are compared. Several different strategies are compared, some of them overlapping with those examined by [?]. The tweets of the user, their followees, and their followers were all considered, as were the lists of users' followers and followees. Additionally, combinations of these methods were employed in a hybridized approach. For each means of building them, the user profile documents were compared as TF-IDF valued word vectors (see Section 3.4.2 for more details on TF-IDF weighting). The hybrid approaches performed the best, generally speaking.

2.2.4.2 Recommending Content

Attempts to rank individual tweets have also been made, such as in [?], which used a content-based approach to this recommendation. Here, various content features are gathered for each tweet and then a learning algorithm is employed to evaluate which of these features should be ranked most highly. The various features that were evaluated can be divided into three basic groups: those based on the content of the tweets themselves, those based on Twitter-specific features, and those based on the authority of the users who created the tweet.

The features for the content of the tweets were represented by the Okapi BM25 or cosine similarity score between a tweet and a query being evaluated. The twitter-specific group included features such as the length of the tweet, whether it contained hashtags, whether it contained a URL, and how many times it had been retweeted. Finally, the account authority group included metrics such as how many followers the author of the tweet has and their PageRank score within the twitter network. Surprisingly, the mere presence of a URL was found to be the most important feature, and the length of the tweet was also found to be very predictive.

2.2.4.3 Recommending Both Users And Content

The work of [?] used a machine learning algorithm to recommend both users and specific content in a system that they called TWITOBI. The approach is primarily based on the content of tweets, but some information about who follows whom in the network is utilized. A probabilistic model based on the expectation maximization algorithm is trained to properly rank both tweets and users via a similar technique.

The sample size in terms of number of users was fairly small, consisting of only 8,405 users, but the sample size in terms of number of tweets was very large, with more than 12 million tweets by those 8,405 users. No discussion of how those users were decided upon is included, which when combined with the small number of users makes some of the results a bit suspect.

2.2.4.4 Common Threads

Many of the methods for recommending users and content utilize similar ideas in making their recommendations. The literature indicates via multiple studies that retweets are a very important predictor of what a user is interested in and that retweets are also an excellent indicator of how much influence a user has over others. It is also clear from existing research that the PageRank algorithm, while useful to some degree, does not perform very well at predicting which users will be most influential on a given user.

Most recent research has not focused specifically on either the network-based approach or the content-based approach, preferring instead to take useful features from each. This dissertation takes the same tack, utilizing useful features from both the network and the content. In many cases, features used successfully in the research reviewed here were not used for various reasons, but would make excellent additions to the methodology and implementation described in chapters 3 and 4.

Chapter 3

Methodology

This chapter describes the methodology used in this dissertation to recommend both content and users to a particular user referred to here as the *distinguished user*. Items are recommended based on both the social network structure surrounding the distinguished user and on the actual content produced by the distinguished user and by others in the surrounding network.

The various aspects of the method are intentionally described as generally as possible and without reference to any particular social network. This generality is important because while the experiments described in Chapter 4 and Chapter 5 are specific to Twitter, the methodology described here could just as easily be applied to Facebook, Google+, or many other social networks.

3.1 Simple Approach

The nature of a social network—a group of users connected to one another by various relationships—suggests that creating a graph is the best way to analyse the network. There are many algorithms suitable for the task of analysing such a graph and much research has gone into applying these algorithms to Twitter and other social networks. Additionally, research has been done to analyse the nature of social graphs to determine things such as the average degree of nodes (i.e. users) in the network and the average distance between nodes. This research was discussed at length in the previous chapter.

The obvious approach to take is to build a graph where each user is a vertex and the relationships between those users are the edges, as in Figure 3.1. Thus, if user A follows user B then there will be a directed edge from vertex A to vertex B. When using the network-based approach for recommendation, this basic graph of the social network is often used, e.g. as in [**?**]. If it is desirable for the analysis to

$$\begin{tikzcd}
A \arrow[dr] \arrow[dd] & \\
& B \arrow[dl, leftrightarrow] \\
C &
\end{tikzcd}$$

Figure 3.1: A simple graph of a social network containing three users. User A follows user C, user A and user C both follow user B, and user B follows only user C.

take the meaning of the content into account then this basic graph can be extended by creating edges between users who frequently discuss similar topics. This can be extended further by weighting the edges based on the strength of the relationship as determined by factors such as common activity as described in Section 2.1.2.2. [?], for example, augmented the basic social graph with additional links based on influence calculations.

Once a weighted graph such as this has been constructed a number of different algorithms can be employed including random walks, as in [?], the PageRank algorithm, as in [?], or simply searching for the strongest links to users who are not already connected to the distinguished user.

The obvious limitation to this simple approach is that it has no way of recommending particular pieces of content—the best it can hope to do is to select content from the top users using some similarity metric to compare that content to content that the user has already indicated interest in. Since the goal of this dissertation is to recommend both users and content this simple approach is clearly not sufficient.

3.2 Co-HITS

3.2.1 Background and Suitability

Deng et. al. [?] developed an algorithm that they dubbed Co-HITS to use for ranking web queries and documents. The name of this algorithm is in reference to the famous HITS algorithm [?] for ranking websites. Co-HITS is similar to HITS, and in fact with particular values of the personalized parameters (see Section 3.2.3) the algorithm reduces to the standard HITS algorithm. As with normal HITS, Co-HITS is an online

$$\begin{tikzcd}
A \arrow[r, leftrightarrow] \arrow[dr, leftrightarrow] & C_{A_1} \\
 & C_{A_2} \\
B \arrow[r, leftrightarrow] & C_B \\
C \arrow[r, leftrightarrow] & C_C
\end{tikzcd}$$

Figure 3.2: This simple bipartite version of the graph of a social network shows how a social network can be made to be bipartite by incorporating generated content into the graph and creating edges between users and the content generated by those users. This also demonstrates that this process alone is not enough to provide a useful graph since this graph is very disconnected.

algorithm, meaning that scores must be obtained for each unique query at query time rather than calculating the score once and then storing the results.

Two major differences between Co-HITS and HITS are the introduction of an initial score and that Co-HITS is designed to operate on a bipartite graph. The initial score factor is particularly important because it makes it possible to include content information in the algorithm by initializing scores based on content similarity. Co-HITS also does away with the concept of differing scores for hubs and authorities, instead keeping only one score.

The obvious objection to using this algorithm for recommending users and content is that the graph of a social network as in Figure 3.1 isn't bipartite at all. However, the two classes described by [**?**], search queries and web documents, can be adapted to the task of content and user recommendation by reformulating the social network graph not with users as vertices and connections between users as edges, but with both users and content as vertices and with edges between user vertices and the vertices representing content produced by that user rather than between users.

This new formulation of the social graph is clearly bipartite, as demonstrated in Figure 3.2, making the Co-HITS algorithm applicable to it. Unfortunately it is also clearly a very disconnected graph since users are only connected to the content that they produce, leaving no connections to other users or other content. This is addressed in section 3.3; for now it is enough that this graph is bipartite and thus that the Co-HITS algorithm can be applied to it.

3.2.2 Algorithm

This section contains a summary of the Co-HITS algorithm as described in [?] but applied to the case of ranking users and content in social networks rather than ranking web queries. Some of the variable names have been changed from that paper's notation to make it more clear what the variables refer to in the context of social networks.

Let the set of all content be called T and each member of that set be called t. Similarly, let the set of all users be called U and each member of that set be called u. Then let the probability of transitioning from a particular user u_i to a particular piece of content t_j be denoted as w_{ij}^{ut} and the probability of transitioning from content t_j to user u_i be denoted as w_{ji}^{tu}. The initial score (discussed in section 3.4) for a given user node and content node will be indicated as u_i^0 and t_k^0, respectively. In both cases the initial scores are normalised such that $\sum\limits_{k \in T} t_k^0 = 1$ and $\sum\limits_{i \in U} u_i^0 = 1$. Given these definitions, the generalized Co-HITS equations are given in [?] as follows:

$$u_i = (1 - \lambda_u) u_i^0 + \lambda_u \sum_{k \in T} w_{ki}^{tu} u_k$$

$$t_k = (1 - \lambda_t) t_k^0 + \lambda_t \sum_{j \in U} w_{jk}^{ut} x_j$$

The paper goes on to describe further refinements under the assumption that only one set of vertices is desired to be scored. However, the goal for this project is to score both sets of vertices, so that refinement, consisting of substituting one equation into the other, will not be discussed. Note that the personalised parameters λ_t and λ_u will be discussed in Section 3.2.3.

From these general Co-HITS equations, [?] describes two frameworks to arrive at final scores for each node in the graph, the iterative framework and the regularization framework. The approaches are similar to the multiple approaches to determining PageRank in that one involves iteration and propagation of scores while the other involves matrix operations and more intense computation.

As the name implies, the iterative framework involves iteratively propagating the scores between the nodes using the above equations until achieving convergence. The paper states that their empirical results assert that convergence usually occurs within approximately ten iterations. Experiments done for this project and discussed in

Chapter 4 indicate that this is accurate. The initial scores are normalised so that the scores for each set T and U sum to one. A convenient consequence of the above equations is that after each iteration of the algorithm the sums each of each set, $\sum\limits_{k \in T} t_k$ and $\sum\limits_{i \in U} u_i$, will still sum to one without need for an additional normalisation step.

The regularization framework, meanwhile, involves transforming the transition probabilities into a transition matrix and using that matrix along with some other derived equations to solve for the final results directly. The final step of this calculation involves a matrix inversion. As described in the paper, this operation can be done efficiently provided that the matrix is sparse and small. That works well for the test case described in [?], because their graph is indeed very sparse as well as being quite small (only 50,000 entries). This method is not well suited for a social network after application of the projection procedure described in section 3.3 because this graph is quite dense and extremely large since millions of pieces of content are produced each day, making the calculation computationally much more difficult and this framework less tractable than the iterative framework.

3.2.3 Parameters

The Co-HITS equations described above each include a so-called personalisation parameter $\lambda \in [0, 1]$, which determines how much weight the initial score has on the final outcome. The closer a given λ parameter is to 0, the more weight the initial score is given in calculating the score after each iteration.

Setting both λ parameters to 0 simply means that the final score will be equal to the initial score. Setting both λ parameters to 1, meanwhile, makes the algorithm into something much more akin to the standard HITS algorithm—the scores of each vertex are determined entirely by the scores of the nodes transitioning into that vertex. If only one of the λ values is set to 0 then the algorithm will converge after only one iteration once the scores of the corresponding set of the bipartite graph have been propagated across to the other set. Finally, [?] indicates that if only one of the λ values is set to 1 then the algorithm becomes something akin to the Personal Page Rank algorithm.

Conceptually it makes sense to set the λ parameters according to the confidence in the initial scores. If they are believed to be quite accurate for one or both sets of vertices in the bipartite graph then more weight should be placed on the initial score by moving the λ value closer to 0. If, on the other hand, very little is known about the accuracy or values of the initial scores for a particular set of vertices then the λ value should be closer to 1.

For the particular data set used in [?], λ values of 0.7 and 0.4 were found to have particularly good results. Chapter 4 of this dissertation has more information on the values that were found to have good results for the social networking dataset used here.

3.3 Projecting to a Bipartite Graph

One of the key insights of this project was how to transform the social graph of a social network, which is very clearly not a bipartite graph, into a bipartite graph suitable for use with the Co-HITS algorithm or other algorithms specific to bipartite graphs such as [?]. Section 3.2.1 has already described how a social network can be made into a simple bipartite graph by taking users and content as vertices and connecting users to the content which they produced.

Unfortunately, this basic process leaves a very disconnected graph and discards all of the network information about which users are connected to which other users, which is a very important piece of information. It also ignores the connections between pieces of content, such as whether they are forwards of a particular piece of content and whether they discuss similar topics to other pieces of content.

This information can be recovered by projecting the network connections between users so that they are represented by connections between user vertices and content vertices. If a user Alice is connected to a user named Bob and a user named Carol in the social graph (as in Figure 3.3a), then this projection would connect content C_A produced by Alice to both Bob and Carol in the output bipartite graph, Figure 3.3b. This process makes sense intuitively because it is clear that if Bob and Carol see this content—which they probably will, given their connection to Alice—then it is exerting some influence on them.

Besides the obvious and explicit connections between users in a normal social network there are also implicit connections between the individual pieces of content that those users create. For example, in the case of Twitter, two tweets which include the same hashtag (e.g. #Oxford) have an implicit connection, as in Figure 3.4a. The projection procedure can easily be extended to apply to content by connecting users who have produced content to all of the other pieces of content which are linked to their own. Thus if Bob mentioned #Oxford in a piece of content, he would

$$\begin{tikzcd}
Alice \arrow[r, leftrightarrow] & C_A \\
Bob \arrow[r, leftrightarrow] \arrow[u, dashed] & C_B \\
Carol \arrow[r, leftrightarrow] \arrow[uu, dashed, bend left] & C_C
\end{tikzcd}$$

(a) Before projection

$$\begin{tikzcd}
Alice \arrow[r, leftrightarrow] & C_A \arrow[dl, leftrightarrow] \arrow[ddl, leftrightarrow] \\
Bob \arrow[r, leftrightarrow] & C_B \\
Carol \arrow[r, leftrightarrow] & C_C
\end{tikzcd}$$

(b) After projection

Figure 3.3: Demonstration of projecting network connections, indicated in (a) by the dashed lines, to make the graph bipartite and more connected.

$$\begin{tikzcd}
Alice \arrow[r, leftrightarrow] & C_A \\
Bob \arrow[r, leftrightarrow] & C_B \arrow[d, leftrightarrow, dashed, "\#Oxford"] \\
Carol \arrow[r, leftrightarrow] & C_C
\end{tikzcd}$$

(a) Before projection

$$\begin{tikzcd}
Alice \arrow[r, leftrightarrow] & C_A \\
Bob \arrow[r, leftrightarrow] \arrow[dr, leftrightarrow] & C_B \\
Carol \arrow[r, leftrightarrow] \arrow[ur, leftrightarrow] & C_C
\end{tikzcd}$$

(b) After projection

Figure 3.4: Demonstration of projecting connections between related pieces of content, indicated in (a) by the dashed lines, to increase the connectivity of the bipartite graph.

be connected to all other content in the graph which also mentioned #Oxford, as indicated in Figure 3.4b.

This projection also makes sense because if Alice creates a piece of content then she is likely to be interested in other content which is related to it. This process of projecting the content links isn't limited to just hashtags. Any two pieces of content determined to be similar could have their implicit links projected in the same way. That could mean that they both mention similar named entities, that it can be determined that they discuss the same topic, that they link to the same external web address, or any number of other means of determining the similarity of two pieces of content.

A number of other projections can be made to add other features of the social network to the bipartite graph. For example, [?] reveals that in Twitter retweets are one of the most important indicators of what is interesting to a user. Retweets—or more generally, re-shares of the content of others—can be included in the bipartite graph in a number of ways such as by strengthening the weight of edges connecting the sharer to the content of the original author, by connecting the content produced by the original content author's connections to the sharer, or by connecting the content produced by the original content author to the connections of the sharer.

Projecting out the links between users and the links between content makes the bipartite graph much more connected and also models the real-world influence that particular network characteristics exert. Depending on the test being run, some or all of these edges may be directed rather than undirected and they may be weighted or not. Some analysis of those design decisions is presented in Chapter 4.

3.4 Method for Initializing Scores

A key factor in the accuracy of the final scores when using the Co-HITS algorithm is the initial scores provided for each set of vertices. How much of an effect the initial scores have is dependent on the λ parameters, but unless both are set to 1 the initial scores will still have a major impact on the final results. It is therefore quite important to choose a good method for determining those initial scores. In the absence of any known metric for determining the scores it is necessary to set the initial scores for a particular set to a uniform value, which indicates that all of them are equally likely to be relevant at the outset.

Fortunately, in social networks, it is possible to make some very accurate initial determinations about content and users which may be relevant.

3.4.1 User Scores

Much work has been done on the so-called link prediction problem which seeks to predict which links will be created between users in a social network, i.e. which other users are most relevant to a given user. [?], [?], and [?] are all examples of studies that have been done on this subject.

Because this process of determining original scores is only the initial step in the recommendations, for this project it makes sense to choose a method which is as fast as possible while still maintaining a high degree of accuracy. [?] studies a number of different methods of link-prediction, ranging from very fast and easy to calculate to very complex calculations. Fortunately, one of the most accurate methods is also one of the easiest to calculate, which is the method of Adamic and Adar from [**?**].

If $\Gamma(x)$ is defined as the set of all connections (neighbours) of vertex x, then the similarity score for two users x and y is given as:

$$\textstyle\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|}$$

The initial score for all of the user vertices in the bipartite graph can be computed by iterating through all of the users and comparing them to the distinguished user with this metric. This function handily penalizes the effect that more popular users have on the scores by virtue of appearing more frequently while weighting less popular users more highly.

Depending on which particular social network is being examined, it may make sense to use slightly different versions of the Γ function. For example, in Facebook it might make sense to consider only the neighbourhood of close connections, while in Twitter it might make sense to find the intersection of the followees of the distinguished user and the followers of the user being scored rather than the intersection of the followers of both.

Both the simple common neighbours function, given by:

$$|\Gamma(x)\cap\Gamma(y)|$$

and the Jacard Coefficient, given by:

$$\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|}$$

were found by [**?**] to present good results as well and could be used in cases where a more lightweight function is desired.

3.4.2 Content Scores

Much research has also been done on comparing the similarity of content, usually in the form of document similarity. Such algorithms can be a straightforward comparison of the number of words which overlap in the two pieces of content or they can be much more complex and use ontologies or other methods of deriving semantic meaning to compare the two pieces of content.

As with the initial user scores, it is desirable that the initial content scores be fast and easy to compute. A natural choice is the cosine similarity metric, which is both very common and very easy to compute. This metric requires that the pieces of content (referred to as documents in the remainder of this section) be represented as vectors in the standard term frequency–inverse document frequency (tf-idf) form.

In this form, the documents are represented as vectors where each dimension corresponds to a different term that appears somewhere in one of the documents in the collection. The value of each dimension is the tf-idf value. In its most common form, the tf-idf value for a given term is computed as:

$$tf(t) \times idf(t) = |t| \times \log(\frac{|D|}{df(t)})$$

where $tf(t)$ is the frequency of term t within the document in question (represented above by $|t|$), and $idf(t)$ is the inverse document frequency for that term, which is found by dividing the number of documents in the collection (represented above as $|D|$) by the number of documents in the collection which containing term t (represented as $df(t)$). The denominator of the idf function can lead to a division by zero if a term does not appear in the collection, so either that needs to be detected and the corresponding tf-idf value set to zero or each df(t) values should be incremented by one.

Once tf-idf vectors have been created for both of the documents $\hat{a}$ and $\hat{b}$ being scored, the cosine similarity metric itself can be computed as:

$$\frac{\hat{a} \cdot \hat{b}}{|\hat{a}||\hat{b}|} = \frac{\sum_i a_i b_i}{\sqrt{\sum_i (a_i)^2}\sqrt{\sum_i (b_i)^2}}$$

This equation for the similarity between two vectors lies in the range $[-1, 1]$ in the general case. However for document vectors the individual terms can never be negative and so the range of the similarity lies in the range $[0, 1]$.

For actually computing the initial scores of the bipartite graph it is necessary only to iterate through all of the pieces of content and compare each one to a reference document using the cosine similarity metric. The only remaining question, then, is the choice of the reference document which best indicates which documents will be relevant. Based on [**?**], the best option would likely be the set of content which the distinguished user has re-shared, since that is the best indicator of the type of content that user would like to see. If that set is empty, then some secondary choices might be the set of all documents produced by connections of the distinguished user or the set of all documents produced by the distinguished user.

While the scores produced by this method are not as accurate as the methods for initialising user scores discussed in Section 3.4.1, the resulting scores are still quite reasonable and the scoring process is very fast. Plus, tf-idf scoring and cosine similarity are very widely used methods and are supported by a wide range of tools and libraries. The fact that the initial scores produced for the content are not quite as accurate as those for users can be dealt with by simply moving the value of the λ_t parameter closer to 1, dampening the impact of the initial score.

Chapter 4

Implementation

This chapter describes how the method of Chapter 3 was actually implemented for this project, a non-trivial step considering the major difficulties associated with the vast amount of data used. It begins by describing how the dataset was selected and the challenges that the large dataset presented in terms of storing and accessing the data. From there it moves on to describe how the bipartite graph was constructed, including selection of tweets and users, construction of edges, and initialization of scores. From there, the actual implementation of the algorithm is described both in words and in pseudocode before finally describing the method of retrieving the results from the database.

Java was chosen as the programming language for this project due to the wide adoption of libraries available to interface with it and the ease of development with those libraries. The Lucene information retrieval library[1], for example, is originally written in Java and was used to index the tweets in a way that made tf-idf cosine-similarity calculations easy. It also facilitated easy search of the tweet database which was useful during the development process to learn more about the nature of the dataset. In addition to Lucene, the Java Database Connectivity library (JDBC) was used to interact with the PostgreSQL database and the Stanford named-entity recognizer was used for the named entity recognition.

4.1 Selecting a Dataset

When beginning the project, one of the first questions was which dataset to use. There were four possibilities: downloading data via the Twitter API specifically for this project, the Choudhury data ([?]), the Stanford data ([?]) using the social graph

[1]http://lucene.apache.org/core/

of Kwak et. al. ([?]), and the database created for the Text Retrieval Conference (TREC). Each had advantages and disadvantages.

Given the availability of existing datasets that would meet the needs of the project, the possibility of downloading a social graph and set of tweets specifically for this project was dismissed early on, leaving the task as one of selecting which of the three datasets was best.

The smallest dataset was the Choudhury data. It consisted of tweets collected between late 2008 and late 2009 and contained more than 10 million tweets and more than 800,000 edges. This database is much smaller than the network was at the time it was created; it contains a tiny fraction of the tweets which were produced during that time and a much smaller fraction of the number of users. It was created by starting a seed set of users and then building the database out from there, but does not include the complete set of followers or followees for any user, though it is a connected graph. For example, a number of users have a list of some followers and then are shown to follow no one. This lack of completeness made this dataset unacceptable for this project.

The second dataset was the TREC dataset[2], consisting of 16 million tweets collected over a two week period in late January and early February of 2011. It is not intended to be a complete list of the tweets from that time period, merely a representative sample which includes spam tweets amongst the important tweets. The major problem with this dataset, however, is that it did not include any social graph information, just the tweets. Since a major portion of this project was based on the impact of the social graph on the recommendations, this clearly was not possible to use.

The Stanford Twitter dataset contains 476 million tweets from more than 17 million users. These tweets were collected over a period of 7 months between 1 June, 2009 and 31 December, 2009. The authors of [?] estimate that these tweets covered approximately 20-30% of the tweets published in that time period[3]. By itself, this dataset would be unusable for the same reason as the TREC database since it does not include any social graph information. Fortunately the authors link to the published social graph of [?], which was crawled via the Twitter API contemporaneously with the collection of the Stanford tweet data. Though the vastness of this dataset presented many challenges during the project, it also was the best choice for putting together a recommendation system to be as accurate as possible.

[2]http://trec.nist.gov/data/tweets/

[3]http://snap.stanford.edu/data/twitter7.html

Of these three datasets, only the TREC tweet data and the social graph of [?] are currently available to the public. Twitter changed their terms of service, requiring care to be taken with the data and researchers to remove tweets from their databases if they were deleted by their original authors. The TREC data is accompanied by a tool which queries the Twitter service to accomplish this, which is why it remains, though even it requires that the United States National Institute of Standards and Technology issue a password to allow the data to be downloaded.

The tweets from the Stanford dataset were able to be used here because one of the researchers within the Oxford University Department of Computer Science had downloaded it before this change came into effect.

4.2 Data Storage

Considering the vast amount of data available in the dataset, an efficient data storage mechanism was vital. The major consideration was the ease and efficiency of retrieving the data. PostgreSQL was selected to store the information because it has excellent support and is a mature database management system capable of being used with large amounts of data. It has a good interface with the Java programming language and is also open source, making it possible to tweak the implementations of certain features if necessary, though that was not done here.

The biggest source of data to be put into the PostgreSQL database was the social graph. The tweets took up far more disk space, but not all of them needed to be stored in the database because most of them were not used, as described in Section 4.3. The complete social graph, meanwhile, was approximately 24 GB, with each line containing a pair of user ids separated by a tab delimiter. Unfortunately the tweets data from Stanford had each tweet indicated only by the user's name, meaning that the data from the social graph which had only user id pairs had to be translated before being stored.

The final `namenetwork` table, whose schema is shown in Table 4.1, used the actual usernames as looked up from another database that accompanied the social graph data. Indices were created around both the username and followername columns, with the username column being clustered since it is used the most frequently and has the most data returned from it since the most popular users have far more followers than followees.

Still, it was helpful for certain queries to have fast access to the followees of a given user, so a second copy of the namenetwork table was created to make these queries

Field Name	Field Type	Description
username	VARCHAR	The name of a user
followername	VARCHAR	The name of a user who follows the user in the username column for this row

Table 4.1: Schema of the 'namenetwork' table

Field Name	Field Type	Description
username	VARCHAR	The name of a user
count	INT	The number of followers of the user in the username column for this row

Table 4.2: Schema of the 'followercount' table

faster—this version was called `namenetwork_followers` and had the same schema as the namenetwork table but was clustered around the followername field rather than the username field.

The usernames were all converted to lower case, and indeed all usernames which could potentially be parsed in mixed case found while indexing the tweets were converted to lower case. Additionally, the social graph data contained a number of duplicate user names. This happened because the program which aggregated the data apparently ran over an extended period of time over which some user accounts were deleted and re-created. Additionally, a large number of user ids corresponded to the name 'n/a'. All of these duplicate names were removed from the database.

For the initialization of the user scores (cf. Section 4.4) using the method of Adamic and Adair (cf. Section 3.4.1) it was necessary to know how many followers each user had. This could be calculated on the fly by running a simple SQL `COUNT()` query on the namenetwork table, but this would sacrifice performance, so it made more sense to do this query once and create a table for it. The `followercount` table, whose schema is shown in Table 4.2, was created using the following SQL command:

```
CREATE TABLE followercount AS
    (SELECT username, COUNT(followername) AS count
        FROM namenetwork GROUP BY username);
```

A hash index was created on the username column of the followercount table to allow for quickly looking up the number of followers that a user has.

The tweets themselves were not all used in any given run of the algorithm, and in fact only a tiny portion of them were used. One guideline for the project was that the recommendations themselves be only for a particular time period on the order of

Field Name	Field Type	Description
tweetid	INT	The unique id of the tweet that this vertex represents
name	VARCHAR	The author of the tweet that this vertex represents
date	TIMESTAMP	The timestamp of when the tweet that this vertex represents was published
tweet	VARCHAR	The text of the tweet that this vertex represents
score	FLOAT	The Co-HITS score for this tweet
original_score	FLOAT	The original Co-HITS score for this tweet before the algorithm's first iteration.

Table 4.3: Schema of the 'tweet_vertices' table

one to two weeks in order to reflect fleeting interests and recommend things that are of interest to the user at that time. For example, during the Olympics someone may be very interested in tweets about the Olympics, but they would probably not be so interested a few weeks after the Olympics have ended. For this project one million tweets were indexed when the algorithm was initialized, corresponding to about 1 week's worth of the data. Considering that the full dataset contains nearly half a billion tweets, this is a very small portion.

It was useful to store this small portion of tweets in two different ways for two different purposes. For purposes of initializing the tweet scores (again, cf. Section 4.4) it was very useful to index the tweets in a Lucene index because such indices automatically count the frequency of each term in each document and in the document collection as a whole, making tf-idf cosine similarity scoring easy.

At the same time, it was also valuable to be able to write SQL statements to quickly select particular tweets and to update their Co-HITS scores quickly, a task much better accomplished via a relational database such as PostgreSQL. This storage of the tweets represented the tweet vertices in the graph, and thus information was included about the Co-HITS score and the original Co-HITS score for each of these nodes. The schema for the `tweet_vertices` table is shown in Table 4.3. This table was indexed on the tweetid field using a btree index. After the initial index was built, an operation which only needed to happen one time for the series of tweets under investigation, the index was clustered.

The tweets were stored in the database, so it obviously makes sense that the user vertices would be stored in the database as well. The schema for the `user_vertices` table is shown in Table 4.4. Similar information needed to be tracked with these

Field Name	**Field Type**	**Description**
name	VARCHAR	The name of the user that this vertex represents
score	FLOAT	The Co-HITS score for this user
original_score	FLOAT	The original Co-HITS score for this user before the algorithm's first iteration.

Table 4.4: Schema of the 'user_vertices' table

vertices as with the tweet vertices, most notably the current Co-HITS score and the original Co-HITS score. The index on this table was a btree index on the name field, and as with the tweet vertices this index was clustered after it was populated since this only needed to happen one time.

Storage in a PostgreSQL database facilitated some easy implementations of certain operations. For example, by storing the original score in the vertex it was possible to quickly determine whether the original scores had been initialized and normalized by simply checking whether the sum of the original_score column was equal to one:

```
SELECT SUM(original_score) FROM user_vertices;
```

Similarly, the scores could be normalized with only one SQL command:

```
UPDATE tweet_vertices
    SET original_score=(original_score / S.sum)
    FROM (SELECT SUM(original_score) FROM tweet_vertices) S;
```

And if the algorithm needed to be reset so that it could be run again, it was a simple matter of copying over the original_score field to the score field. In the actual implementation, three additional columns were present in the two vertex tables, one for each of the distinguished users. These columns stored the original scores for each of the users and made it possible to calculate this value only once for each user, facilitating faster testing.

The final component that was stored in the PostgreSQL database was the list of edges between the two different classes of vertices. The schema of the `edges` table is shown in Table 4.5 and is quite basic. The name of the user and the id of the tweet that are connected is shown, along with the type of the edge. The directionality (if any) of that edge is indicated purely by its type, but by default all edges were treated as non-directed. Experimentation with making edges of particular types directed is addressed in Section 5.2.2.3.

Field Name	Field Type	Description
name	VARCHAR	The name of the user that this edge connects to.
tweetid	INTEGER	The unique identifier of the tweet that this edge connects to.
type	INTEGER	The type of edge that this represents (see Table 4.6).

Table 4.5: Schema of the 'edges' table

4.3 Building the Graph

Building the graph is a very important part of the process. This was taken care of in the GraphManager class (cf. Appendix **??**), and in particular in the `createGraph()` method and the other methods called from there. First the users and tweets are selected and indexed and then all the appropriate edges between the two sides of the bipartite graph are created.

4.3.1 Selecting Users and Tweets

How to build the graph itself is not as simple a question as it would seem to be at first. Which users should be present in the graph? Which tweets should be present? How the graph is created and the decisions about which nodes are present dictates how connected the graph will be and thus how the scores from each set of vertices will interact with one another. Having users who are not connected to the rest of graph serves no purpose because no information about their suitability can be gained.

Tweets were added by opening one of the files containing tweet data and parsing until the appropriate number of tweets had been indexed. Certain tweets in the data were invalid and were thrown out. The tweet ids were created by beginning at 1 and then incrementing for each tweet that was added to the database (i.e. each valid tweet). Proceeding in this manner ensured that the tweets were from the same time period and had a unique identifier.

In order to ensure that the graph was maximally connected, only users who would have connections were added. Practically speaking, this meant that users were added as the tweets were placed into the tweet_vertices database. All users who had authored a tweet were added, but so were all users who had been retweeted or mentioned. Because of the types of edges created, as described in Section 4.3.2, these users are guaranteed to at least be connected to one tweet and will likely be connected to several others.

The average degree of a user node and a tweet node can be taken by dividing the number of edges by the number of users and the number of tweets, respectively. For the set of tweet data used in this project, the first 500,000 tweets of the December data from the Stanford tweet dataset, there were ___ edges, ___ tweet vertices, and ___ user vertices, meaning that the average user was connected to __ tweets and the average tweet was connected to __ users.

4.3.2 Edge Creation

Once the users and tweets have been added to the database, the next step is to add the edges between the users and the tweets. There were twelve different types of edges used in this project, though the list is by no means exhaustive of the possibilities. The different edge types are listed in Table 4.6.

The authorship edges are simple and are actually created when the users and tweets are initially indexed. For each tweet added to the database its author is added if that user is not already present and then an edge is added between the tweet and the author. These edges are simple.

The remainder of the edge types are more complicated and fall broadly into two categories. The first category is the follower, retweet, mention, and @reply edges, which correspond to the network edges connecting users to other users and the edges connecting users to particular tweets. The second category of edge is the content edges which correspond to the edges between tweets.

4.3.2.1 Network Edges

The follower edges are the most straightforward to describe and are simply the result of following the procedure for projection of user connections described in Section 3.3.

The retweet and mention edges are more complicated but are very analogous to one another. The simplest of these edges is the simple retweet or mention edge which connects the tweet to the person being mentioned or retweeted; as with the authorship edges, these are created at the time that the tweets are initially indexed. These edges capture the intuition that a person is likely to be interested in a tweet that mentions them or comments on something they say. The followees and followers versions of the retweet and mention edges, meanwhile, are more complicated.

The followees edge creates a connection between the person who is retweeting or mentioning someone and the followees of the person being mentioned or retweeted. This is because that user has proven by their retweet or mention that they are very

Edge Type	**Description**
Authorship	Connects a user to a tweet authored by that user
Follower	Connects a user to a tweet authored by one of their followees, as described in Section 3.3
Retweet	Connects the tweet to the person being retweeted
Retweet Followees	Connects the tweets of the followees of the person being retweeted to the retweeter
Retweet Followers	Connects the tweets of the person being retweeted to the followers of the retweeter
Mention	Connects the tweet to the person being mentioned
Mention Followees	Connects the tweets of the followees of the person being mentioned to the mentioner
Mention Followers	Connects the tweets of the person being mentioned to the followers of the mentioner
At Reply	Connects the tweet to the person being @replied to
At Reply Content	Connects the tweets of the person being @replied to to the tweeter
Hashtag	Connects all tweets which use a given hashtag to the authors of those tweets
Content	Connects all tweets discussing similar content (cf. Section 4.3.3) to the authors of those tweets.

Table 4.6: The types of edges in the graph

interested in content from the user who authored the original tweet, so it stands to reason that they are correspondingly more likely to be interested in the same things as that person.

The followers edge connects the tweets of the person who was retweeted or mentioned to all of the followers of the person who retweeted or mentioned them. If a user is interested enough in someone to follow them and see their tweets then it stands to reason that they would also be interested in someone that their followee is very interested in.

The @reply edges are quite straightforward, though they probably provide less impact on the final rankings because they generally only reinforce the existing follower and followee connections. These edges connect the tweet to the person the @reply was directed at and connect the tweets of the person the @reply was directed towards to the person who sent it.

4.3.2.2 Content Edges

There are far fewer content edges, but they convey a great deal of information. These edges are based on the system of projecting edges between tweets into the bipartite graph by connecting the edges to users instead using the procedure detailed in Section 3.3.

Hashtag edges represent the connections between tweets using the same hashtag. These edges connect all the tweets using a particular hashtag to all of the authors of tweets containing those hashtags. Figure 3.4 demonstrates this process. The hashtags are converted to lowercase for normalization purposes and because Twitter itself treats hashtags as being case insensitive.

The content edges connect tweets in the same way, but using the implicit links between tweets with similar content as the basis for the projection. How the similarity in content is determined can vary between implementations and range from something complicated such as Latent Dirichlet Allocation topic models to something much simpler.

4.3.3 Determining Content Similarity

Determining which tweets contain similar content to one another is an important part of the algorithm used here because it provides many connections between users and tweets that are not connected via straightforward network connections, allowing novel tweets to be recommended by the algorithm.

The method for determining this similarity need not be very complicated, though it certainly can be. Using topic models and other techniques from statistics and machine learning would certainly generate good results for determining content similarity, but at the cost of being quite computationally expensive.

For this project a much simpler and faster method was used: named-entity recognition (NER). Using the Stanford Named-Entity Recognizer [?], each tweet was run through the tool in order to find any named entities such as people, places, and organizations. This was done as the tweets were being processed for other content information such as hashtags.

Each named entity found was converted to lower case and stored in memory along with the id of the tweet in which it appeared. If another tweet was found containing the same named entity then that tweet's id was stored along with the ids of the other tweets already found with that id. Once all of the tweets had been processed, each list of tweetids was connected to the authors of the respective tweets.

This process is far from perfect, but provides a good approximation of the people, places, and things that users are talking about. In experiments done prior to incorporating this tool into the algorithm, the NER of the Stanford library was found to be generally accurate, despite the lack of context provided by the space constraints of tweets.

The biggest problem with the use of the recognizer in this algorithm is that it returns compound entities as separate entities, meaning that entities such as 'United Kingdom' would be returned as two separate words. Similarly, names when given as first and last names (e.g. 'David Beckham') would be returned in two chunks. This presents a question as to whether these words should be treated individually or combined into one entity.

Leaving them as separate words gives some poor results such as 'David' or 'United' being counted as entities even though they are not very useful, while combining them separates cases where people are referred to by last name, meaning that 'Beckham' and 'David Beckham' would be two different entities even though they refer to the same thing. Similarly, phrases such as 'David Beckham, United Kingdom' would be returned as one big entity, which is clearly nonsensical.

For the purposes of this project, the former approach of leaving all entities as separate was taken. This leads to a small number of garbage edges but with the benefit of more correct connections and a simpler implementation. Put another way, this choice increases the recall of the connections between content while decreasing the precision.

4.4 Initializing Scores

Broadly speaking, the initialization of scores was implemented as described in Section 3.4. The user scores were initialized according to the method of Adamic and Adar ([?]) as described in [?] and the tweet scores were initialized via cosine similarity.

4.4.1 Tweet Scores

Calculating the initial tweet scores was mostly trivial. For purposes of finding a reference document which indicated the interests of the distinguished user against which all tweets could be compared, several possibilities were used. The first thing checked was whether a particular user had retweeted anything during the time period under study since retweets are the best predictor of what a user is interested in. If they had retweeted any content then those retweets were combined into one document with the'RT' removed and used as the reference document. Unfortunately, given that the data presented was a small subset of the overall data, most users had no retweets and those who did have any usually only had one, which decreases the accuracy of this technique by focusing on one particular interesting tweet.

For those users with no recorded retweets in the timeframe under study, the composite document can be formed using the tweets of all of their followees. These can conveniently be recovered based on the edges by selecting all follower edges which connect to the distinguished user.

The initialization of the tweet scores uses the previously discussed Lucene library to aid with content similarity. Lucene stores term frequency information within its index, simplifying the calculation of tf-idf information for use with cosine similarity. All of the tweet vertices are selected from the database and then iterated through with a tf-idf document vector created for each tweet and then compared to the reference document using cosine similarity. Once all of the scores have been added to the database, they can be normalized to add to one by using the SQL statement shown in Section 4.2.

4.4.2 User Scores

Calculating the initial user scores was one of the more time-intensive parts of the process and led to the creation of extra columns in the user_vertices table to allow all of the users being investigated to have their initial scores stored so that they did not need to be calculated each time a different user was used for experimentation.

Clustering the namenetwork table and creating the followercount table were both motivated largely by speed gains while calculating the initial user score.

Recall the scoring method of Adamic and Adar from Section 3.4. If $\Gamma(x)$ is defined as the set of all connections (neighbours) of vertex x, then the similarity score for two users x and y is given as:

$$\textstyle\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|}$$

For initializing the user scores, user x is always the distinguished user and user y is the user currently being scored. In the experiments of both [?] and [?] utilizing this formula the networks under study were non-directed, so the simple metric for the Γ function of neighbouring vertices was adequate. In the Twitter network, however, connections are directed.

Clearly it does not make sense to look for the intersection of the followers of two users to determine their similarity, since users have no control over this. A better method would be to look for the intersection of the followees of two users since it would at least indicate that they were interested in seeing the same things. Still, it would not indicate whether one user actually created content of interest to the other user.

For this project, the function used was the intersection of the followees of the distinguished user with the followers of the user being scored. If users that the distinguished user has interest in express interest in another user, then it stands to reason that that user will be more likely to be of interest to the distinguished user.

For each user z in this intersection, the score component from that user is calculated according to $\frac{1}{\log|\Gamma(z)|}$, and this is summed for all such users z. In this formula, $|\Gamma(z)|$ represents the size of the neighbourhood $\Gamma(z)$, i.e. the number of followers that the user z has, as found in the followercount table. This formula discounts the scoring effect of very popular users because they will naturally have more overlap in their follower lists with the followees of the distinguished user.

This process is accomplished in the project by running a SQL query to retrieve the follower counts of all of these users and then summing them in the code. That SQL query is:

```
SELECT C.count FROM
  ( (SELECT username AS name FROM namenetwork_followers
          WHERE followername=?)
```

```
      INTERSECT
    (SELECT followername AS name FROM namenetwork
           WHERE username=?)
  ) S, followercount C
  WHERE S.name=C.username;
```

where followername is parametrized with the username of the distinguished user and username is parametrized with the username of the user being scored.

The one user for whom this method does not work is the distinguished user—that user's score would not be properly returned using this method. Instead, the distinguished user has their initial score set to some factor of the largest score found before normalization. For this implementation this factor was 1.25, meaning that the distinguished user's initial score before normalization would be set to 1.25 times the highest score from any other user. This is designed to provide a strong score impact from the things that the distinguished user is connected to since that provides a great deal of information about content and users they are likely to be interested in.

4.5 Implementation of the Co-HITS Algorithm

The iterative version of the Co-HITS algorithm itself can be implemented to run in linear time and space once the graph and its associated indexes have been created. The process of indexing these various fields during the creation of the graph is $\mathcal{O}(n \cdot \log(n))$, but this only needs to be performed one time for the graph and once it is complete any number of users can have their recommendations built off of that graph with only linear time complexity. Furthermore, it is easy to update the graph incrementally as new users or tweets are added.

More specifically, the complexity of the algorithm is $\mathcal{O}(E \cdot I)$, where E is the number of edge vertices and I is the number of iterations. For each iteration the edges are sorted first by the user name, when calculating the tweet scores, and then by the tweet id, when calculating the user scores. This sort is performed when the tweets are indexed, however, not when running the algorithm. When updating the tweet scores, the following SQL statement is used to select the edges:

```
SELECT U.score as user_score, U.name, E.type, T.tweetid
  FROM user_vertices U, edges E, tweet_vertices T
  WHERE T.tweetid=E.tweetid AND E.name=U.name
  GROUP BY U.name, E.type, T.tweetid;
```

For updating the user scores, the following SQL statement is used to select the edges:

```
SELECT U.name, E.type, T.score as tweet_score, T.tweetid
  FROM user_vertices U, edges E, tweet_vertices T
  WHERE U.name=E.name AND E.tweetid=T.tweetid
  GROUP BY T.tweetid, E.type, U.name;
```

Note that in both cases the scores of the opposite vertex type from that being updated must be selected so that the update can be performed appropriately. For example, in the case of updating the tweet scores, the user score for each vertex must be known so that the score contribution from that vertex over the edges emanating from it can be calculated.

Each vertex is considered in order and its score contribution to the various nodes on the other side is considered. The sorting makes it easy to determine the total number and type of edges emanating from a particular vertex because all of the edges from that vertex are considered one after another. Knowing the total number of edges leaving a particular vertex is necessary when calculating the transition probability. Upon iterating through all edges sorted by one component (either users or tweets), the scores for the other half of the graph can be updated with the newly calculated values.

Pseudocode of the implementation is shown in Algorithm 1, while the actual implementation is shown in the code listing in Appendix **??** in the `updateTweetScores()` and `updateUserScores()` methods of the GraphManager class.

It is clear from examining Algorithm 1 that each edge is actually traversed twice, and since this algorithm is repeated twice for each iteration, the edges are actually traversed 4 times for each iteration. This is an unfortunate consequence of the need to know the number of edges leaving a particular vertex before calculating that vertex's effect on the vertices to which it connects.

When updating the scores, the following SQL query is used:

```
UPDATE tweet_vertices
  SET score=(? + (original_score * ?))
  WHERE tweetid=? AND ABS(score - (? + (original_score * ?))) > 0;
```

Only the version for updating the tweet scores is shown, but the user score update proceeds in a nearly identical manner. The two question marks in the SET clause and in the ABS() portion of the WHERE clause have identical meanings: the first is

Algorithm 1: The Co-HITS Implementation for updating the tweet scores. Updating user scores is identical, but with changes to the appropriate variable names.

```
tweetScores = $\emptyset$;
curUser $\leftarrow$ first_edge.username;
curUserEdges = $\emptyset$;
while database cursor for edges not at end do
    edge $\leftarrow$ edge at current database cursor position;
    if curUser $\neq$ edge.username then
        for edges e $\in$ curUserEdges do
            tweetProbability $\leftarrow$ 1 / (size of curUserEdges);
            scoreEffectFromThisEdge $\leftarrow$ tweetProbability * curUser.score * $\lambda_t$;
            if tweetScores contains key edge.tweetid then
                curScore $\leftarrow$ tweetScores.get value for key edge.tweetid;
                curScore $\leftarrow$ curScore + scoreEffectFromThisEdge;
                tweetScores.update(edge.tweetid , curScore);
            else
                tweetScores.put(edge.tweetid , scoreEffectFromThisEdge);
            end
        end
        clear curUserEdges;
    end
    add edge to curUserEdges
    curUser = edge.username
    move database cursor to next edge;
end
update the tweet scores in the database according to tweetScores
```

the score contribution from the other vertices as calculated by Algorithm 1, and the second is the appropriate λ value as discussed in Section 3.2.3—λ_t for updating the tweet vertices and λ_u for updating the user vertices.

The part of the WHERE clause with the check of the absolute value is intended to help with determining whether the algorithm has converged. The query will indicate that no rows were updated if the score does not change, making it easy to determine convergence. In their paper introducing the Co-HITS algorithm, [?], Deng, et. al. say that the iterative algorithm generally converges in under 10 iterations. This was found to be generally true in this project, depending on how it was defined.

Using the above SQL statement never resulted in absolute convergence in the sense that no additional vertices were updated, and relaxing the requirement of how close the value had to be in order to determine convergence simply resulted in the normalization of the scores being knocked out of balance. Still, after ten iterations, and usually much fewer, it was clear by examining the values as they changed that the changes that were still occurring were very small, not actually affecting the final outcome. For this project, the algorithm was allowed to run for up to fifteen iterations before terminating.

Updating the scores in this manner also adds an additional complexity factor. It would be possible to perform this in $\mathcal{O}(U)$ or $\mathcal{O}(T)$ time (for users and tweets, respectively) if the vertices were traversed in order and updated in that way, but in practice this is slower than performing a series of batch updates using the B-Tree index which runs in $\mathcal{O}(U \log U)$ time.

4.6 Retrieving Results

Retrieving the results of the algorithm is made easy by the structure of the tables within PostgreSQL; it is simply a matter of running a SQL command. For this project it was thought that users could find and become aware of users with a lot of followers very easily. By extension it was also assumed that users could easily find content by these users.

Thus, when retrieving the recommendations, only those whose follower count fell below a certain threshold were considered. Similarly it made no sense to recommend content that the user had already seen, users who the distinguished user already followed were ignored, as were their tweets. It may be reasonable to include tweets from existing followees in the recommendations for some applications (and Twitter's own new recommendation system does this) but for evaluation purposes within this project they were ignored.

The SQL statements for selecting the highest scoring tweets is:

```
SELECT T.name, T.score, T.tweet FROM
  tweet_vertices U,
  (
     (SELECT T.name FROM tweet_vertices T, followercount F
         WHERE T.name=F.username AND
                     F.count < 100000 ORDER BY T.score DESC)
   EXCEPT
     (SELECT username AS name FROM namenetwork_followers
         WHERE followername=?)
  ) S
  WHERE T.name = S.name ORDER BY T.score DESC;
```

The SQL statement for selecting the highest scoring users is:

```
SELECT U.name, U.score FROM
  user_vertices U,
  (
     (SELECT U.name FROM user_vertices U, followercount F
         WHERE U.name=F.username AND
                     F.count < 100000 ORDER BY U.score DESC)
   EXCEPT
     (SELECT username AS name FROM namenetwork_followers
         WHERE followername=?)
  ) S
  WHERE U.name = S.name ORDER BY U.score DESC;
```

These two SQL statements both take 100,000 followers as the threshold of followers below which the content is deemed useful for recommendation purposes, but this can easily be adjusted.

By default this does not filter out two classes of tweets that a user is unlikely to be interested in: his own and those containing @replies to other users. Similarly, for the sake of brevity, the SQL statement for retrieving users does not filter out the distinguished user, though it would be simple to do so. In both cases, these tweets are filtered out before the recommendations are delivered to the user.

Chapter 5

Results

Having detailed both the method used for recommendation and the actual implementation of that method, the discussion now moves to the method for evaluating the results and the presentation of the results themselves in order to demonstrate the efficacy of the method.

First the method of evaluating the results is described; for such a large dataset this is a challenging question, and it is not yet well-defined for this research area. The selection of the distinguished users is then described and their effect on the evaluation methods is considered before arriving at several metrics for evaluating the results.

Next the baseline results are shown using the established metrics. Finally, variations on the algorithm are experimented with and their results presented in order to establish which parameters and edge types are the most valuable for this task and whether the algorithm can be improved over its baseline result.

5.1 Evaluation Methodology

5.1.1 Challenges

A review of the existing research on content recommendation discussed in Section 2.2.4, reveals no consensus on how best to evaluate the results of a recommendation method. This is due primarily to two separate but related issues: the lack of a common dataset and the lack of ground truth judgements on the relevance of either content or users.

None of the research projects on Twitter content recommendation reviewed in Section 2.2.4 used the same set of Twitter data for their experiments. All of the researchers used the Twitter API to download their own datasets of varying sizes and using various procedures. Because of this the datasets turn out to be very different in terms of size, user composition, and tweet content, which makes results difficult to compare between papers. The creation of the microblog track of the Text Retrieval Conference (TREC) was accompanied by a very large collection of tweet data which may be commonly used in the future, but as discussed in Section 4.1, it was not sufficient for this project.

The bigger issue, however, is the lack of a set of ground truth judgements of the relevance of tweets and users in any dataset. One obvious cause of this is the lack of a canonical dataset upon which recommendations are made. Perhaps if ground truth judgements were available such a dataset would come into common usage, but the fact is that the job is too complex to realistically be performed reliably for any decent sample size. The content and users that are relevant for one user would be vastly different from the content and users relevant for another. And any database of a usable size would have millions and millions of tweets and users, far too many for all of them to be judged for even one user.

5.1.2 Distinguished User Selection

With these challenges in mind, it was necessary to develop methods of evaluating the performance of the algorithm that did not rely on a pre-judged set of relevant content and users. The first step was to choose distinguished users to whom the content should be recommended. For privacy reasons, those users are not named here. Overall, three users were chosen.

Two users were chosen because they are known personally by the author and are long-time users who are present in the dataset. These facts made them ideal candidates for the user studies discussed in Section 5.1.3. Both of these users are computer scientists, so technology is a major interest, but there are some distinguishing interests as well. These two users will be known as users K_i and K_j, with the choice of K meant to indicate that they are the users **K**nown to the author.

The other user was chosen because his interests are clear and easy to distinguish and because he was well connected to the small sample of tweets that were used during initial implementation. This user has more than 7,000 followers as of this writing, up from approximately 3,700 three years ago when the dataset was collected. As of this writing he has published nearly 25,000 tweets. His interests are also in technology, but with a business focus. A large number of his current followees are technology entrepreneurs, but he also posts tweets and has followees related to popular culture such as films, television, and music. This user will be known as user U, indicating that he is **U**nknown to the author.

5.1.3 User Study

The goal of the user study is primarily to determine whether the content recommended to the users was of interest to the distinguished user.

A user study is perhaps the closest method of evaluation to a set of ground truth relevance judgements, though on a smaller scale. For the tweets ranked by the users it is possible to say with certainty how relevant they are and thus to evaluate the precision and recall of the algorithm. The user study could obviously only be performed on the users known to the author, which was part of what motivated the choice of known users as distinguished users.

One major drawback of the user study is that it recommends content to these users based on what was happening three years ago. Thus, the recommendations are based on the interests of three years ago which may no longer be relevant. Before filling out the user survey these users were asked to put themselves in the frame of reference of what might have been interesting to them three years ago, but this is obviously inexact, so the efficacy of user study depends largely on the idea that interests change slowly.

The user studies consisted of two components: recommending users and recommending individual tweets.

For recommending individual tweets, the top one hundred tweets, retrieved as described in Section 4.6, were retrieved after the algorithm was run. These tweets were then presented to the user in a random order and each user was asked to rank each of the one hundred tweets from 1 to 5 using the values from Table 5.1. From these scores, the recall@rank and precision were evaluated by considering tweets scored as either 4 or 5 to be relevant tweets.

Recommending users proceeded in a largely similar manner. For each recommended user their Twitter profile and ten most recent tweets were retrieved from Twitter and presented to the user in random order. The users under study were then asked to rank the recommended users on the same scale as for individual tweets, with the results being used in the same way to determine recall@rank and precision scores by considering scored as either 4 or 5 to be relevant users.

5.1.4 Comparison to Current Tweets

The user study is effective for evaluating the results for the known users but is subject to biases and changes in interests and does nothing for evaluating the results

Score	Description of tweets with this score
1	Not relevant at all, e.g. non-english tweets
2	Useless tweets
3	Average tweet; not particularly useful, but not completely without value
4	Relevant or interesting tweet
5	Very relevant tweets

Table 5.1: User ranking scores

for unknown or unavailable users. As such, it was also necessary to develop more automated techniques.

One such technique is to assume that a user's interests remain static between the time period represented by the dataset and the present day. This makes it possible to compare the tweets recommended by the system to the tweets that the user is actually interested in while using a separate data source to avoid overfitting.

As described by [?] and mentioned elsewhere in this dissertation, retweets are the most effective means of determining which content a person is interested in. Thus, each tweet recommended by the system can be compared to a composite document consisting of the most recent retweets from the present day twitter stream of the user for whom the recommendations are being created. These numbers can be compared to a baseline created by comparing each tweet in the system to the reference document and taking an average of their similarity scores.

For this project, the method of determining similarity was the cosine similarity metric as described in Section 3.4.2. Because of the large number of tweets in the database the baseline against which these numbers were compared was determined by taking the average of the similarity scores for ten thousand tweets rather than for the entire collection.

This method can also be used to evaluate users, though the comparison is a bit less accurate. In this case, the most recent tweets from the recommended user are retrieved from Twitter and compared against the same reference document of retweets as used when evaluating tweets. The similarity of these can again be compared to a baseline by repeating this process for some number of users chosen at random. This requires far more requests to the Twitter API, however, making it subject to being rate limited, reducing the number of users who can be evaluated when generating the baseline.

As with all of the means of evaluation, this is an imperfect measurement. Given their technical interests and the fast-paced nature of that field, many of the technologies that these users are interested in today may not have existed at the time the data was collected. Still, it does allow a comparison to show that the recommendations provided have value above random recommendation and because it is automated it allows far more documents to be ranked than the user study.

5.1.5 Comparison to Current Network State

Much of the research on link prediction in social networks focuses on predicting whether links will be created between users in the future, and the evaluation involves comparing the predicted links to those actually formed later. For this project, the only data available on the state of the social graph is that of the dataset and that of the present day.

Thus, one evaluation method used was to compare the users recommended for each distinguished user to the set of actual followees of that person in the present social network. The number of recommended users likely to be in the set of present-day followees is extremely small given that most users follow a small set of people, but by comparing the number of users in the top twenty five recommended users who overlap with the present-day followee list to the probability of a random user appearing in that list it is possible to demonstrate that the recommendations have value.

5.2 Results

5.2.1 Baseline Results

5.2.2 Experimental Variations

5.2.2.1 Varying λ Parameters

5.2.2.2 Varying Edge Types

5.2.2.3 Varying Edge Directionality

Chapter 6

Conclusion

6.1 Enhancements and Future Work

One additional type of content edge that might be useful but which was not included in this project is a URL edge which connects tweets containing the same URL—in a shortened URL form using a service such as bit.ly, most likely—to the authors of those tweets. It wouldn't be feasible to expand all of the shortened URLs and to compare them that way, but since a given URL will always map to a particular shortened URL (at least for a period of several years) it is not necessary to expand them.

6.1.1 Parallelization and Performance Improvement