The Pursuit of Happiness: 
Searching for Worthy Followees on Twitter

John Hannon, Kevin McCarthy and Barry Smyth

CLARITY: Centre for Sensor Web Technologies,
School of Computer Science and Informatics,
University College Dublin, Ireland
firstname.lastname@ucd.ie

Abstract. We are living in an age of information overload, where it can be difficult to define which information is relevant and important to the end user at a point in time. In this paper, we introduce a solution to apportioning this constant flow of information by going to the source of the content, namely the producers. This paper examines an application for searching for pertinent friends on the popular microblogging service, Twitter\textsuperscript{1} and our approach to curtail the cold start problem that new users of the service face. We introduce our search technology which is capable of finding the producers of wanted content and suggest connecting to them as followees on Twitter. We also prove the usefulness of this technology through the results of a live user experiment carried out on these cold start users.

1 Introduction

Since its first emergence on the scene in early 2006, Twitter has grown from an early microblogging engine into a real time web behemoth. In the early days some 300,000 tweets were produced by early adopters of the system per month, contrast this with some 3-4 billion tweets estimated to be produced per month in 2011 \textsuperscript{2}. With this obvious information increase, creating ways to use this information appropriately and intelligently has become a hot topic for many researchers [3, 14]. The volume of content alone on Twitter is intriguing from a research standpoint, but can only 140 character per tweet contain enough information to provide us with a signal to use in the formation of recommendations?

In Twitter the main producers of this content are the users themselves who have subscribed to the system. Obviously, not all the information produced by each user is to everyone’s preference, so finding the producers, the so-called “diamonds in the rough”, is an interesting research challenge. In this paper, we examine the challenge of finding these users who would be of interest to follow on Twitter and we provide a way of searching for potential new connections based on the content produced by users. Typically these systems would use the content

---

\textsuperscript{1} http://www.twitter.com
\textsuperscript{2} http://blog.twitter.com/2011/03/numbers.html
created by a user to generate the recommendations of new potential followees [6], however in this work we provide a technique for new users of the Twitter service to find interesting people to follow. We do this by using the Twittomender recommendation engine and specifically its search engine capabilities, to search for users based on key terms contained within the content they produce (their tweets). We have evaluated this search engine in a live user trial and we provide an analysis based on the relevance of the results produced by our system.

2 Related Work

The research in this paper can be split into two categories, that of mining Twitter for information signals to form recommendations and secondly, friend finding on social networking sites. Recently, researchers of recommender systems have quickly realised the potential for using freely available public snippets (tweets) from users to aid traditional research practices. In the area of news recommendation, Phelan et al. describe their “Buzzer” system which uses various arrangements of user generated data, to suggest current news articles to a user and also news personalised to that user [14]. Garcia et al. harnessed Blippr\(^3\) (a Twitter-like micro-blogging service that focuses on products), to beat traditional collaborative filtering approaches for product recommendation. Also, in video segment recommendation; an aggregation of tweets being produced during football games at the FIFA World Cup is used to create personalised highlight reels based on frequency and content being produced at segmented time periods [7].

The main contribution of this paper is to the field of friend finding or friend recommendation in Social Networking Sites (SNS). Much research has been carried out into friend finding online and also in the enterprise environment [5, 9]. In previous work on Twitter followee recommendation [6], we collected a corpus of Twitter users along with their tweets, profile information and social connections and use both the content and the connections to form recommendations. We mined and aggregated these information signals of user tweets and connections to evaluate their interests and hopefully find similar people to recommend and build a friend recommendation system. Guy et al. describes their “Do you know?” system which uses overlapping similarities between users online profile pages e.g. two people co-authored paper, contributed to the same wiki, etc. It helps to promotes connections between colleagues or people who have worked together and increase one’s social graph. They also add recommendation explanation to each of their suggestions showing how this recommendation was formed to aid users in making a decision to form a connection [5]. Similarly, Hsu et al. examine how people share commonalities, in this case links in weblogs, these links represent a diverse set of things such as subscription or friendship between the bloggers. Their link recommendation engine models users as vertexes and recommends links between them based on each users interests [9]. Other related work focuses on the “Why?” of friend recommendation. Why did you decided to follow someone or form a connection? Jensen et al. [10] examine the

\(^3\) http://www.blippr.com/
online reputation of users and evaluated how much a user’s online presence or reputation is factored into the decision making process. Finding reputable users, users who are good contributors to the system and who actively engage with a community, are the types of users we aim to recommend.

We see our research centring on forming connections that aren’t fleeting, forming connections with users who share similar interests and produce similar content. Some research has been carried out into the potential disillusionment that can occur after someone decides to befriend or form a connection on a social networking site. Kwak et al. [11] examined a subset of Twitter users and explored the factors and reasoning behind a user unfollowing or breaking the connection between them and another user. They concluded that many connections were dissolved due to the content that was being produced by the users that were being following, content that could have been made transparent when the user was deciding to form their connection initially.

3 Tackling Cold Start In Twitter

Traditionally the cold start problem relates to a lack of initial knowledge to base a prediction or recommendation for a new user starting out using a system. Many researchers have examined the problem of cold start particularly in traditional collaborative filtering type recommender systems [8, 12, 17, 18]. This problem affects site such as; Amazon and Netflix who base their recommendations on purchase and/or viewing history, but when this information isn’t available they need alternative heuristic information such as using item popularity to form recommendations. Whilst a valid stepping stone approach, until a user’s profile becomes indicative of their preferences, it lacks the user centric personalisation aspect. This one size fits all approach is not optimal for each individual user. Twitter has a similar problem, whilst not focused solely on recommending a product or item but instead focused on initial user engagement with the system. To retain users, Twitter must have new users actively follow/befriend others and contribute through tweets. Initially, though who do you follow: celebrities? colleagues? classmates? Twitter has taken steps to jump start the user engagement as they need to get users into the habit of following other Twitter users and forming new social connections. Twitter’s approach augments the sign up process by allowing users to select categorised popular twitter users from News, Sport, Entertainment, etc. to initially follow. These popular users don’t always promote the conversational nature that is Twitter, as Cha et al. explain, these celebrities’ content isn’t always significant but attracts attention because they have a public persona outside of the Twittersphere [2].

We believe there is a better way to recommend new users to follow based on the content that users are producing and not solely on their follower count. Our recommendation system allows users to find others to follow based on the aggregated historical tweets from the Twitter users we have collected. We also aid the user in their followee choosing decision, by providing recommendation explanations, such as showing the users description of themselves and the top
terms they mention in their tweets or who they mention in their tweets. In this way users can decide if Joe Bloggs, who is a computer engineer, but who tweets mainly about pottery, not computers is a person they would like to follow. In the following section our followee recommender Twittomender is discussed and specifically it’s search functionality to aid these cold start users to initially find other users who share a similar interest with them.

4 Twittomender

Twittomender is a user recommendation web application that provides two main functions to suggest new users to follow on the microblogging site, Twitter [6]. The function we are going to examine in this paper is the user defined search capabilities of Twittomender. This service allows users to define their own search queries to locate users, in a content-based recommendation approach [13]. There are many types of search engine; semantic, those powered by communities, ranked lists, etc. [1, 4, 19] but all search engines have the same goal – to retrieve a desired or relevant piece of information for a given query. In the case of Twittomender, we hope to utilise the content produced by users on the popular social micro blogging site to mine rich information and provide a relevance score to produce followee recommendations.

![Fig. 1. Control Flow of User Search.](image)

The Twittomender system’s main function involves syncing a user’s account and producing followee recommendations through a range of collaborative and
content-based strategies. However for this to work efficiently, users must be active on Twitter, i.e. they must follow a number of other users, must have some followers themselves and must have produced some content (through tweets). Although this functionality is great for Twitter users who wish to increase the number of appropriate user streams they follow as seen in our initial Twittomender evaluation [6] (where on average 7 new user connections were formed), it does not perform satisfactorily for new/recently joined Twitter users. These users have not produced much content through tweets, nor are they following or being followed by enough users for collaborative or content-based followee recommendation techniques to perform as expected. For this reason we introduced the search capability to Twittomender. Figure 1 shows the control flow when using Twittomenders search functionality.

Twittomender mines a large number of user profile details and content by following the links between users on the social graph. When a new user is looking for followees, Twittomender can recommend profiles from our database of over 1 million Twitter users. Twittomender allows users to input a search query to find interesting potential followees. The terms of this query relate to the types of content the user would like to consume. These terms are used to find users who have used these key terms frequently within their own tweets. The search functionality is built upon Lucene4, the document search engine. We collected a new dataset of users for the Twittomender system than our previous trial [6], each user in the Twittomender database was modelled as a document but, with these documents content containing up to 200 tweets of that user. Modelling each user in this way lends itself to an information retrieval style approach to search. Using the TF-IDF [15, 16] scoring metric and query word stemming, search queries entered are scored for similarity amongst the document corpus to produce a result set. Then up to 20 top matching results are returned to the user as recommendations for followees. In Figure 2 we see the search results recommended to a user for two sample search queries. Figure 2 only shows unary search, but “AND” and “OR” query flags can be used to build more complex search queries. The term cloud shows the weight of a particular term a suggested user uses in their tweets about that topic. Users can quickly identify if a user talks about the topics that would interest them.

5 Evaluation: User Trial

To test the performance of our system we carried out a live user evaluation of the search functionality on Twittomender. We asked 80 participants to carry out up to 5 or more searches for users they would be interested in following on Twitter. These participants where all first year students from a computer applications course in Dublin City University. The break down of the 80 participants was 16 females and 64 males with the majority of participants aged 18-22. As we were evaluating the ability to aid new users form connections the test group of users was formed of all relatively new adopters of Twitter, with many recently joining

4 http://lucene.apache.org/
Fig. 2. Screenshot of User Search Results for (a) “Manchester United” (b) “recsys”

Twitter. Table 1 shows a summary of user statistics. We can see that on average participants of the trial produced 12 tweets, followed 20 Twitter users and were followed by about 8 people. When compared to the overall average for an active Twitter user, based on a dataset of active users taken from Twittomender, the average number of tweets produced was 78, they were followed by 465 users, and they were following 520 users [6]. So it is clear to bridge the gap between being a new user and an active user, some help is needed. During the course of the experiment each user completed just over 5 searches each and these queries were made up of almost three terms each on average.

Table 1. Average User & Evaluation Statistics

<table>
<thead>
<tr>
<th>No. Of Tweets</th>
<th>No. Of Followers</th>
<th>No. Of Followees</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.75</td>
<td>7.875</td>
<td>19.99</td>
</tr>
</tbody>
</table>
5.1 Search Relevance: MAP

In order to test the relevance of our returned results, we used the mean average precision (MAP) measurement. We looked at the results across all participants searches and averaged them across all of our participants. When users where shown a results list they voted on whether or not a result was relevant to their search query and whether they would follow this recommended user. In Figure 3 we can see on average across all of the searches, participants would choose to follow 20% of the recommendations on Twitter. On average this is 5 new users that each participant would follow per search. Also in this graph users indicated that 45% of the results were deemed relevant to the query that was entered. So, using the search functionality of Twittomender, users can find new, previously undiscovered and relevant Twitter profiles to follow. Precision is only one factor when evaluating a good search system. Next we look at the location of relevant results in the results list.

5.2 Relevance Positions

One key aspect that must not be overlooked when evaluating the performance of a search engine, is that of the relative position that results are returned to the user in the result set. Much study refers to the lions share of user intent being directed towards the upper tier of the results set. So in turn not only must results be relevant they must appear high in the result list to be classed as a high
standard of suggestion. In Figure 4 we have modelled the selections rated by the participants in the trial and their position within a search result set. In Figure 4 (a) we can see, of all the relevant suggestions that users would also follow on Twitter, that 50+% of results appear in the top 10 of the result set. And, we can see in Figure 4 (b) that likewise, when we examine just the relevant results (those result that a participant deems relevant to the query including those users they would follow and wouldn’t follow but are relevant), we again see that over half are contained within the top 10 of the result sets. These trends show that Twittomender is able to produce the most relevant results towards the top of the list returned to users.

![Fig. 4. (a) Position Of Relevant & Would Follow and (b) Position Of Relevant across all searches.](image)

### 6 Conclusions

In this paper we have described the problem associated with cold start users joining Twitter. We have described the Twittomender system, a followee recommender for Twitter. Specifically in this paper we focused in on the search functionality of Twittomender, which can help users who haven’t formed many connections or produced much content, find new users to follow on Twitter. We carried out a live user evaluation on 80 new or recently joined participants on Twitter. Each participant carried out their own user defined searches and on average across all their searches they found 45% were relevant and would follow 20% of the recommendations. Also, the search results show that the higher quality recommendations tend to appear towards the top of result lists where the lion’s share of users attentions is focused. Overall, we have shown that Twittomender’s search functionality can be used as a efficient tool in aiding these cold start users form early connections quickly and engage with Twitter.

As part of future work we plan to extend how we form our recommendations to users. Using the content of users’ tweets, we would hope to categorise users
based on their tweet activity, what they talk about frequently and suggest people groups of users. Recommendation would be formed based on their content if they have produced enough content to use Twittomender’s syncing functionality or the search functionality if the have not. Also, if Twitter allows access to their recommendation system we would hope to carry out a comparison between the recommendations of Twittomender and that of Twitter. Finally, I believe we have shown that using Twittomender to help increase your social graph connections with good quality producers has been clearly shown through our evaluations.

Acknowledgment

This work is supported by Science Foundation Ireland under grant 07/CE/I1147 and by Amdocs Inc.

References