Recommending Information Sources to Information Seekers in Twitter

Marcelo G. Armentano, Daniela Godoy, and Analía Amandi
ISISTAN Research Institute, Fac. Cs. Exactas, UNCPBA
Campus Universitario, Paraje Arroyo Seco, Tandil, 7000, Argentina
CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
{marmenta, dgodoy, amandi}@exa.unicen.edu.ar

Abstract

Finding high-quality sources in the expanding micro-blogging community using Twitter becomes essential for information seekers in order to cope with information overload. In this paper, we present a recommendation algorithm aiming to identify potentially interesting users to follow in the Twitter network. This algorithm first explores the graph of connections starting at the target user (the user to whom we wish to recommend previously unknown followees) in order to select a set of candidate users to recommend, according to an heuristic procedure. The set of candidate users is then ranked according to the similarity between the content of tweets that they publish and the target user interests. Experimental evaluation was conducted to determine the impact of different profiling strategies.

1 Introduction

Micro-blogging activity taking place in sites such as Twitter is becoming every day more important as real-time information source and news spreading medium. In the followers/followees social structure defined in Twitter a follower will receive all the micro-blogs from the users he follows, known as followees, even though they do not necessarily follow him back. In turn, re-tweeting allows users to spread information beyond the followers of the user that post the tweet in the first place.

Recent research efforts on understanding micro-blogging as a novel form of communication [Java *et al.*, 2007; Krishnamurthy *et al.*, 2008] revealed that few users in Twitter maintain reciprocal relationships with other users. This fact differentiates Twitter from other online social networks, such as Facebook, Hi5, or Orkut, in which people mainly make connections to keep in touch with people they consider as friends or acquaintances.

Although posts in Twitter or *tweets* are allowed to have any textual content within the limit of 140 characters, many users only publish information about a particular subject, such as sports, movies, music or about a particular rock band. These users can be considered as information sources or broadcasters. In contrast, many people use twitter to get information on particular subject, as a form of RSS reader, registering themselves as followers of their favorite artists, celebrities, bloggers, or TV programs. Users acting as information sources are characterized by having a larger number of followers than followees, as they are actually posting useful information or news. Information seekers, on the other hand, subscribe to this kind of users but rarely post tweets and, finally, friends are users exhibiting reciprocal relationships. With information seekers being an important portion of registered users in the system, finding relevant and reliable sources in the constantly increasing Twitter community[1] becomes a challenging issue.

To address this problem, we propose a followee recommender system that, according to an heuristic procedure, explores the topology of followers/followees network of Twitter to find candidate users to recommend and then these candidate users are ranked according to their similarity with the target user's interests. Three profiling strategies are analyzed and evaluated for modeling users' interests in Twitter based on two general approaches. The first approach models a user by analyzing the content of his/her own tweets whereas the second approach represents users by the tweets of their followees. For the second approach, two different types of profiles were considered: modeling a target user by the set of profiles of his/her followees, and by a set categories that can be discovered by clustering his/her followees according to the content of their tweets.

Unlike other works that focus on ranking users according to their influence in the entire network [Weng *et al.*, 2010; Yamaguchi *et al.*, 2010], the algorithm we propose explores the follower/followee relationships of the user up to a certain level, so that only the neighborhood of the target user is explored in the search of candidate recommendations. The influence rankings presented by studies on the complete Twittersphere have no direct utility for followee recommendation since people who are popular in Twitter would not necessarily match a particular user's interests. For example, if a user follows accounts talking about technology, he/she would not be interest in Ashton Kutcher, one of the most influential Twitter accounts according to [Kwak *et al.*, 2010].

[1] In 2010 Twitter grew by more than 100 million registered accounts (`http://yearinreview.twitter.com/whosnew/`. Accessed on March 2011)

In this article we study Twitter from a user modeling perspective. Our goal is to provide recommendations to information seekers about users who publish tweets that might be of their interest. Unlike traditional recommendation systems, we do not have any explicit information available about the user's interests in the form of ratings on items he/she likes or dislikes. The only information available for profiling a Twitter user is the structure of the followers/followees network and the tweets published in this network. Both of these elements are considered in this paper as a mean to recommend people who share the same content-related interests with the user who will receive the recommendations.

The rest of this work is organized as follows. Section 2 discusses how related work is related to our research. Section 3 describes the content-based approach to the problem of followee recommendation for helping information-seeking users in Twitter. In Section 4 experiments carried out to validate the approach using a Twitter dataset are reported. Finally, Section 5 discusses the results obtained and presents our conclusions and future work avenues.

2 Related Work

The problem of helping users to find and to connect with people on-line to take advantage of their friend relationships has been studied in the context of traditional social networks. For example, SONAR [Guy *et al.*, 2009] recommends related people in the context of enterprises by aggregating information about relationships as reflected in different sources within a organization, such as organizational chart relationships, co-authorship of papers, patents, projects and others. Liben-Nowell et al. [Liben-Nowell and Kleinberg, 2003] presented different methods for link prediction based on node neighborhoods and on the ensemble of all paths. These methods were evaluated using co-authorship networks obtained from the author lists of papers at five sections of the physics e-Print arXiv[2]. Authors found that there is indeed useful information contained in the network topology alone. Chen et al. [Chen *et al.*, 2009] compared relationship-based and content-based algorithms in making people recommendations, finding that the first ones are better at finding known contacts whereas the second ones are stronger at discovering new friends. Weighted minimum-message ratio (WMR) [Lo and Lin, 2006] is a graph-based algorithm which generates a personalized list of friends in a social network built according to the observed interaction among members. Unlike these algorithms that gathered social networks in enclosed domains from structured data (such as interactions, co-authorship relations, etc.), we face the problem of taking advantage of the massive, unstructured, dynamic and inherently noisy user-generated content from Twitter for recommendation.

Several studies dedicated to understand micro-blogging as a novel form of communication and news spreading medium have been recently published. Some of these research efforts have been dedicated to study the structure of Twitter network and its community structure. Java et al. [Java *et al.*, 2007] presented a characterization of Twitter users identifying three kinds of users:

- "Information Sources" are users who are characterized by having a much larger number of followers than they themselves are following.
- "Friends" are users who trend to use Twitter as a typical online social network and are characterized by reciprocity in their relationships.
- "Information Seekers" are users who rarely post a tweet authored by himself, but that regularly follows other users

In a posterior study presented by Krishnamurthy et al. [Krishnamurthy *et al.*, 2008] also three categories of users were identified. The first category is called "Broadcasters of tweets" and corresponds to the "Information Sources" category above. The second category, "Acquaintances", is equivalent to "Friends" category identified by Java et al. However, a different interpretation is given to the third category of users. Krishnamurthy et al. call users who follow lots of other users but that are followed by few users "Miscreant / Evangelists". They consider that users in this category are usually spammers or stalkers that contact lots of users expecting to be followed by them.

Kwak et al. [Kwak *et al.*, 2010] quantified these findings indicating that 77.9% of Twitter connections are unidirectional and only 22.1% of the relations are reciprocate. Moreover, 67.6% of users are not followed by any of their followees, indicating that these users probably use Twitter as a source of information rather than as a social networking site.

Other line of research has been devoted to measure the influence of users in Twitter. In [Kwak *et al.*, 2010] it was shown that ranking users by the number of followers and by their PageRank give similar results. However, ranking users by the number of re-tweets indicates a gap between influence inferred from the number of followers and that from the popularity of user tweets. Coincidentally, a comparison between in-degree, re-tweets and mentions as influence indicators [Cha *et al.*, 2010] concluded that the first is more related to user popularity. Analyzing spawning re-tweets and mentions, it was found that most influential users hold significant influence over a variety of topics but this influence is gained only through a concentrated effort (such as limiting tweets to a single topic). TwitterRank [Weng *et al.*, 2010], an extension of PageRank algorithm, tries to find influential twitterers by taking into account the topical similarity between users as well as the link structure. Garcia et al. [Garcia and Amatriain, 2010] propose a method to weigh popularity and activity of links for ranking users. User recommendation, however, can not be based exclusively on general influence rankings since people get connected for multiple reasons.

While the studies mentioned above focused on the analysis micro-blogging usage, other works try to capitalize the massive amount of user-generated content as a novel source of preference and profiling information for recommendation. Chen et al. [Chen *et al.*, 2010] proposed an approach to recommend interesting URLs coming from information streams such as tweets based on two topic interest models of the target user and a social voting mechanism. For each user two models are used: a Self-profile built with the words of the user tweets and a Followee-profile built by combining the

[2] http://www.arxiv.org

self-profiles of the user followees. Thus, a set of candidate pages posted by a user followees and followees of followees is filtered according to these models. In the social scheme filtering is based on a voting system within a user followee-of-followees neighborhood so that the most popular URLs within the group are recommended. *Buzzer* [Phelan *et al.*, 2009] indexes tweets and recent news appearing in user specified feeds, which are considered as examples of user preferences, to be matched against tweets from the public timeline or from the user Twitter friends for story ranking and recommendation. Esparza et al. [Esparza *et al.*, 2010] address the problem of using real-time opinions of movie fans expressed through the Twitter-like short textual reviews for recommendation. The work by Esparza et al. assumes that tweets contain preference-like information that can be used in content-based and collaborative filtering recommendation. Opinion mining and sentiment analysis applied to tweets are starting to be considered to replace explicit ratings required by traditional recommendation technologies [Pak and Paroubek, 2010; Davidov *et al.*, 2010].

Continuing in this direction, Naaman et al. [Naaman *et al.*, 2010] classify users into "informers" and "meformers". According to this work, "informers" users publish tweets containing mainly non-personal information while"meformers" users mainly post status updates about themselves and their daily routines. Ramage et al. [Ramage *et al.*, 2010] go a step forward using a partially supervised learning model that maps the content of tweets into different dimensions that correspond to substance, style, status and social characteristics of posts. "Substance" tweets contain information about events, ideas, things or people; "social" tweets relate to some socially communicative end; "status" tweets refer to personal updates; finally "style" tweets are those indicative of broader trends of language use. Perez-Tellez et al. [Perez-Tellez *et al.*, 2010] categorize tweets which contain a company name into two clusters corresponding to those which refer to the company and those which do not. They use a text enrichment technique, called Self-Term Expansion Methodology (S-TEM), aiming at improving the quality of the corpora. Several variations of this technique are presented and compared, such as enhancing S-TEM by considering additional information extracted from Wikipedia.

In contrast to the previous works that address the problem of suggesting potentially relevant content from micro-blogging services, we concentrate in recommending interesting people to follow. In this direction, Sun et al. [Sun *et al.*, 2009] proposes a diffusion-based micro-blogging recommendation framework which identifies a small number of users playing the role of news reporters and recommends them to information seekers during emergency events. Closest to our work are the algorithms for recommending followees in Twitter evaluated and compared in [Hannon *et al.*, 2010] using a subset of users. Multiple profiling strategies were considered according to how users are represented in a content-based approach (by their own tweets, by the tweets of their followees, by the tweets of their followers, by the combination of the three), a collaborative filtering approach (by the IDs of their followees, by the IDs of their followers or a combination of the two) and two hybrid approaches. User profiles are indexed and recommendations generated using a search engine, receiving a ranked-list of relevant Twitter users based on a target user profile or a specific set of query terms. Our work differs from this approach in that we do not require indexing profiles from Twitter users. Instead, a topology-based algorithm is used to explore the follower/followee network in order to find candidate users to recommend and a content-based analysis is then applied to generate the ranked list of recommendations.

3 Followees Recommendations in Twitter

The problem of followee recommendation in Twitter consists in identifying users posting relevant tweets for a target user, so that he/she can subscribe to these users and start receiving real-time information from them. The approach presented in this work can be decomposed into three main parts. First, we create the target user's profile which describes his/her interests or information needs. Second, we search for a suitable group of candidate users to be considered for recommendation and determine whether the information that they publish may be of interest to the target user. Finally, we rank these users and present the top-ranked users as followee recommendations. These parts of our approach are described in the following sections. Section 3.1 describes different strategies for building user profiles in order to describe a user's interests. Next, Section 3.2 describes the search for candidates based on the topology of the Twitter network. Finally, Section 3.3 explains how profiles are compared in order to determine which set of users recommend to the target user.

3.1 Content-based User Profiles

In our approach, the user profile for a target user u_T will model the information he/she likes to read, whereas for a candidate user u_C the user profile will model the information he/she publishes. For any user u, let $tweets\,(u)$ be the set of all his/her posts:

$$tweets\,(u) = \{t_1, \ldots, t_k\} \tag{1}$$

The interests of a target user can then be described using different content sources, such as the text of his/her own tweets or the content of the tweets published by his/her followees. The different strategies we used to create the target user's profiles are described in the following sections.

Profile Strategy T0

The simplest alternative to build a profile for a user in Twitter is to aggregate his/her own tweets under the assumption that users are likely to tweet about things that are of interest to them:

$$Profile^{T0}\,(u_T) = \sum_{i=1}^{k} t_i \tag{2}$$

The profile of a user is then a vector in which terms are weighted according to their frequency of occurrence in the text of the user tweets. Tweets are processed to obtain the vector of a given user posts, $Profile^{T0}\,(u)$, by applying a number of filters in a pipeline. First, tokens only composed of punctuation symbols are assumed to be emoticons and are

then removed. Second, common slang vocabulary and abbreviations are substituted. This kind of words are widely used in Twitter messages to overcome the limitation in the number of characters. The NoSlang on-line dictionary[3], containing 5,227 entries, was used to this end. In this step abbreviations are replaced with the corresponding complete words or phrases, for example "idn" is replaced by "i don't know" or "ntta" by "nothing to talk about". Finally, stop-words are removed and Porter stemming algorithm [Porter, 1980] is applied to the remaining words.

Profile Strategy T1

Information seekers are characterized by posting few tweets themselves, but they follow people who generate content more actively. Hence, as followee recommendation is oriented toward information seekers, an alternative method to model the interests of a user is based on who is he/she following, this is, which information the user wants to read about.

It is assumed that users select their followees expecting that their tweets will be of interest to them. Thus, a second type of profile is built based on the observation that a user has a number of followees:

$$followees\left(u_T\right) = \{f_1, \ldots, f_l\} \tag{3}$$

and information a user is interested in can be obtained from the profiles of his/her followees.

However, users might follow people twitting about different subjects. For example, a user may follow celebrities, politicians, sportsmen and other type of users. As a result, considering all followees as responding to a unique topic of interest is not enough to effectively model multiple user interests in diverse areas. Consequently, rather than creating a single vector representing all of the user's interests, this strategy creates multiple vectors, each of them representing a different followee. This profile strategy allows us to attain fine-grained profiles. The profile of a user is then defined as the set of the profiles of the user followees, each modeling a followee own tweets:

$$Profile^{T1}\left(u_T\right) = \left\{Profile^{T0}\left(f_1\right), \ldots, Profile^{T0}\left(f_l\right)\right\} \tag{4}$$

Profile Strategy T2

In a more realistic view of a user information preferences, it can be assumed that users are likely to follow people in different interest categories. For example, a user can be following some Twitter users because they talk about his/her favorite sport and others according to her/his political opinions. Hence, to assess a more precise description of the user interests a last type of profile tries to group the user followers into meaningful categories.

Coarser-grained profiles are created using a simple clustering algorithm detailed in Algorithm 1. The identification of categories to which a user's followees belong need to be incrementally discovered starting from scratch as the user starts following a new user in Twitter. In this clustering approach, as soon as the user subscribes to a followee it is assigned to the first cluster or category in the user profile. Each subsequent followee is incorporated into either some of the existent categories or to a novel category depending on its similarity with the current categories. Hence, a user's interest categories are extensionally defined in the user profile by highly similar followees that conform clusters. This partition reduces the total number of vectors representing all followees to a relatively smaller number of clusters, which can be further analyzed to discover topicality.

[3] http://www.noslang.com/dictionary

Algorithm 1 Incremental clustering algorithm

Input: The vector profiles, $Profile^{T0}(f)$, of all $f \in followees(u)$ of user u and a similarity threshold δ

Ouput: The profile of u grouping the followees in a set of followee categories $FC_u = \{fc_1, \ldots, fc_m\}$

INCREMENTALCLUSTERING

1: $FC_u \leftarrow \emptyset$ */*Create an empty profile for u*/*
2: $Q \leftarrow \emptyset$ */*Initialize a set to contain the clusters the new followee is similar to*/*
3: **for all** f_i such that $f_i \in followees(u)$ **do**
4: **for all** fc_j such that $fc_j \in FC_u$ **do**
5: Let c_j be the centroid of fc_j
6: $sim_j \leftarrow sim(c_j, fc_j)$
7: **if** $sim_j \geq \delta$ **then**
8: $Q \leftarrow add(\langle fc_j, sim_j \rangle)$
9: **end if**
10: **end for**
11: **if** $(Q \neq \emptyset)$ **then**
12: Sort instances in Q by decreasing order of sim_j
13: Let fc_k be the first cluster in Q
14: Include the followee f_i into fc_k */*The centroid vector of the cluster is updated*/*
15: **else**
16: Create an empty cluster fc_{new}
17: Include the followee f_i into fc_{new}
18: **end if**
19: **end for**
20: Return FC_u

The clustering algorithm returns a set of categories $FC_u = \{fc_1, \ldots, fc_m\}$ the current followees of the user u can be grouped into. Given the cluster or followee category fc_i , which is composed of the set of followees and their corresponding vector representations, the centroid vector c_{fc_i} is

$$c_{fc_i} = \frac{1}{|fc_i|} \sum_{f \in fc_i} Profile^{T0}(f) \tag{5}$$

Each time the user starts following another user, the new followee vector is incorporated to the current user profile within the most similar existing cluster. In order to predict which this cluster is, the closest centroid is determined by comparing the vector $Profile^{T0}(f_{new})$ of the new followee with all centroids in the existing clusters. This similarity measure determines the degree of resemblance between the vector representations and is calculated by the cosine similarity. As the result of vector comparison, the new followee f_{new} is assigned to the cluster with the closest centroid, i.e.

$$\arg \max_{j=1...k} sim\left(f_{new}, c_{fc_j}\right)$$

provided that the similarity is higher than a minimum similarity threshold δ. Vectors not similar enough to any existent centroid according to this threshold cause the creation of new singleton clusters.

In summary, two general approaches are evaluated in this paper for modeling a user's interests in Twitter according to if the user own tweets or the tweets of their followees are used to glean a profile. For the last approach, two different mechanisms to combine the vectors of the user followees were analyzed. The first consists in modeling a target user using a set of vectors, each of them representing the content of a user followee tweets. The second profile models a target user by a set of vectors corresponding to the centroids obtained after applying a clustering algorithm to the vectors representing the target user followee tweets.

3.2 Topology-Based Candidate Search

In order to recommend Twitter users, a set of viable candidates need to be first identified within the follower/followee network. The method employed to explore the Twitter network with the goal of gathering candidate users for recommending to a target user u_T is based on the following hypothesis: the users followed by the followers of u_T followees are possible candidates to recommend to u_T. In other words, if a user u_F follows a user that is also followed by u_T, then other people followed by u_F can be of interest to u_T.

The rationale behind this hypothesis is that the target user is an information seeker that has already identified some interesting users acting as information sources, which are his/her followees. Other people who also follow some users in this group (i.e. are subscribed to some of the same information sources) have interests in common with the target user and might have discovered other relevant information sources in the same topics, which are in turn their followees. Figure 1 illustrates this approach for candidate selection schematically.

More formally, the search of candidate users for recommendations is performed according to the following steps:

1. Starting with the target user u_T, obtain the list of users he/she follows, let's call this list $S = \bigcup_{\forall x \in followees(u_T)} x$.
2. For each element in S get its followers, let's call the union of all these lists L, i.e. $L = \bigcup_{\forall s \in S} followers\,(s)$.
3. For each element in L obtain its followees, let's call the union of all these lists T, i.e. $T = \bigcup_{\forall l \in L} followees\,(l)$.
4. Exclude from T those users who the target user is already following. Let's call the resulting list of candidates $R = T - S$.

Each element in R is a possible user to recommend to the target user as future followee. To relate to the previous hypothesis the group S will be mostly composed of information sources, L will be other users looking for information in the same way that u_T does and T will be further information sources. Users can appear more than once in R, depending on the number of times that they appear in the lists of followees or followers obtained at steps 2 and 3 above, this is a factor that can be later consider to boost its chances of being recommended.

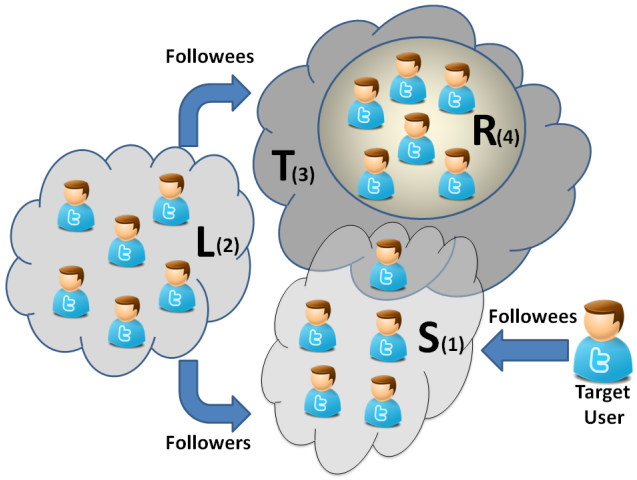

Figure 1: Strategy for exploring the followee/follower network to find candidate users

It is worth noticing that other strategies can be elaborated or combined with the search based on the topology of the network described above to include in the evaluation users who are not in the proximity of the target user. For example, candidate users can be taken from Twitter's *public timeline*. The public timeline is an information stream that contains the collection of the most recently published tweets and it is fed by all accounts that are not configured to be private. The public timeline can be considered as the current flow of information in Twitter and it is a good source to obtain active users in the social network.

3.3 Comparing User Profiles

Once a list of viable candidates R is available, the matching between the information each user $r \in R$ publishes in Twitter and the user interests need to be evaluated in order to obtain a ranked list of followee recommendations.

We determine the similarity between the profiles of a candidate user that need to be evaluated for recommendation u_C and the target user u_T, denoted $sim^{T0}\,(u_C, u_T)$, as the cosine similarity between the two vectors. The cosine of the angle conformed by two vectors in the space is calculated as the normalized dot product [Salton and McGill, 1983].

For strategy T1 and T2, in order to evaluate whether to recommend a candidate user u_C to the target user u_T, the information published by the candidate, $Profile^{T0}\,(u_C)$, needs to be compared with the profile of the target user, $Profile^{T1}\,(u_T)$, which is the information u_T is subscribed to receive in Twitter. The matching is then calculated as shown in equation 6.

Finally, the similarity $sim^{T2}\,(u_C, u_T)$ is evaluated in the same way, as specified in Equation 6.

$$sim^{T1}(u_C, u_T) = \max_{\forall i: f_i \in followees(u_T)} sim^{T0}\left(Profile^{T0}(f_i), Profile^{T0}(u_C)\right) \quad (6)$$

	Average	Maximum	Minimum
#followees	94.77±15.54	119	41
#followers	2.0±1.74	10	1
#tweets	102.44±57.56	199	11

Table 1: Summary of statistics of the users selected for testing the approach

For all strategies, all candidate users are ranked according to their similarity to the profile of the target user and the user is presented with a reduced number of followee recommendations.

4 Experimental Evaluation

4.1 Dataset Description

The Twitter dataset[4] used in this paper is a social graph of 835.541 follower/followee relations between 456.107 users and their corresponding tweets belonging to a time span of 2006 to 2009, reaching a total of 10.467.110 tweets. This dataset was created using a focused crawler based on a snowballing technique over a set of quality users, who post about a diverse range of topics and reasonably frequently. In the assemblage of this dataset, reported in [Choudhury *et al.*, 2010], the crawler was seeded with 500 users comprising politicians, musicians, environmentalists and so on; and next the social graph was expanded from the seeds based on the friend links between users.

From the entire dataset a test set $|U_{test}| = 100$ was created to empirically evaluate the content-based followee recommendation approach. Since the recommendation approach is intended to help information seekers in Twitter rather than users serving as information sources, the 100 users were selected on the basis of having their followees outnumbering their followers. Likewise, users who posted less than 10 tweets were excluded from the social graph so that valuable content-based profiles could be extracted for all users involved in the evaluation. The profiles of these target users were built analyzing the text of their tweets according to the strategies proposed in Section 3.3. Table 1 summarizes the characteristics of the U_{test} in terms of number of followees, followers and published tweets.

4.2 Methodology and Metrics

Experiments were carried out using a holdout strategy in which some the target user followees are hidden from the recommendation algorithm and then it is verified if they were discovered and suggested as future followees. In all experiments, the set of followees of each user were partitioned into a 70% for training, starting from which candidates are located and evaluated, and a 30% for testing, whose existence is verified in the list of top-*N* suggested followees for each user in U_{test}. If followees in the 30% group are suggested to the target user in spite of being concealed, it means that the algorithm was able to locate these users through the 70% non-concealed followees and their relationships. In order to make the results less sensitive to the particular training/testing partitioning of the followees, in all experiments the average and standard deviation of 5 runs for each individual user are reported, each time using a different random partitioning into training and test sets.

The quality of lists of top-*N* followee recommendations generated for the group of users used for testing was evaluated considering the standard precision:

$$precision(RE) = \frac{1}{|U_{test}|} \sum_{u \in U_{test}} \frac{|followees_{test}(u) \cap RE_u|}{|RE_u|} \quad (7)$$

where RE_u is the set of recommendations for a user $u \in U_{test}$, U_{test} is the set of users considered for testing (in this work $U_{test} = 100$ as described in the previous section), $followees_{test}(u)$ is the set of followees that were reserved for testing the top-*N* list of a single user u (not used as seeds for starting candidate search).

In other words, precision measures the average percentage of overlap between a given recommendation list and the user actual list of followees and it can be evaluated at different points in a ranked list of suggested followees. Thus, precision at rank *k* (P@*k*) is defined as the proportion of recommended followees that were relevant, i.e. were in the target user test set. In the reported experiments we evaluate precision for values of *k* equal to 1, 5, 10, 15 and 20, although *k* values of 1 and 5 are the most common sizes for recommendation lists reported in the literature as people tend to pay more attention to the first few results that are presented.

Another measure similar to precision is the number of hits in a recommendation list, this is the number of followees in the test set that were also present in the top-*N* recommended followees for a given test user. If $|U_{test}|$ is the total number of testing users, the hit-rate (HR) of the recommendation algorithm is computed as [Deshpande and Karypis, 2004]:

$$HR = \frac{number\ of\ hits}{|U_{test}|} \quad (8)$$

HR grants high values to an algorithm if it is able to predict the followees in the test sets of the corresponding users, while assign low values of the algorithm was not able to recommend the hidden followees.

One limitation of this measure is that it treats all hits equally regardless of where they appear in the list of the top-*N* recommended items. Average reciprocal hit-rank (ARHR) rewards each hit based on where it occurred in the top-*N* followees that were recommended by a particular strategy. If h is the number of hits that occurred at positions $p_1, p_2 \ldots, p_h$ within the top-*N* lists (i.e., $1 \leq p_i \leq N$), then the average reciprocal hit-rank is equal to:

[4]Originally posted at `http://www.public.asu.edu/~mdechoud/datasets.html`

$$ARHR(RE) = \frac{1}{|U_{test}|} \sum_{i=1}^{h} \frac{1}{p_i} \quad (9)$$

That is, hits that occur earlier in the top-*N* lists are weighted higher than hits that occur later in the list. The highest value of ARHR is equal to the hit-rate and occurs when all the hits occur in the first position, whereas the lowest value of the ARHR is equal to hit-rate/*N* when all the hits occur in the last position in the list of the top-*N* recommendations.

4.3 Experimental Results

Table 2 show the precision and hit-rate results for followee recommendations using the different profiling strategies and the mentioned pre-processing techniques for analyzing tweets. The number of candidates explored was on average 6,692.74±511.25 users reached through the user own followees.

It can be observed in the results presented that the users own tweets are not effective for identifying potentially interesting followees. This is probably due to the fact that information seeking users tend to be more passive in posting messages while behave more actively following other people to keep up with interesting information or news.

In contrast, the strategies profiling users based on the information published by their followees either separately or grouped into categories, were more effective in recognizing people to start follow among the candidates found. In fact, the strategy using a vector for each followee outperforms all others for the various sizes of the recommendation lists. When followee vector representations were aggregated into clusters, precision diminished significantly but also the number of similarity calculations is reduced since profiles are of smaller size.

Interestingly, the ARHR values shown in Figure 2 for the four profiling strategies allow to infer that hits are better positioned in the list generated using clustering of followees than in those produced with separate followee vectors. Therefore, the issue of improving the ranking of relevant recommendations will be then matter of future research, particularly exploiting the number of occurrences of the candidates in the set R as a voting mechanism.

It is worth noticing that in the previous results the effectiveness of the algorithm to identify followees is being underestimated given the testing methodology employed. Users suggested to the target user that are not in the test set are not necessarily irrelevant, although they are considered incorrect recommendations in the calculation of the precision and hit-rate metrics. In fact, the target users might not be in their list of followees either because they are not interested on receiving their tweets or because they have not yet discovered the recommended users in the Twitter network. In the last case, these recommendations are also appropriate and will be valuable for the users.

Figure 3 depicts the mean average similarities between the vectors of the users in the top-*N* lists with the corresponding target user profile. The low similarity of information publish by the recommended users and the target user tweets account for the poor results of the first profiling strategy. On the other hand, the high average similarities of users in the top-*N* lists generated by the two last strategies suggests that even the recommended users deemed as irrelevant publish information highly similar to the user profile and to the remaining recommended users in each list, most of which the user is already following. Hence, they are likely good recommendations in spite of being considered otherwise.

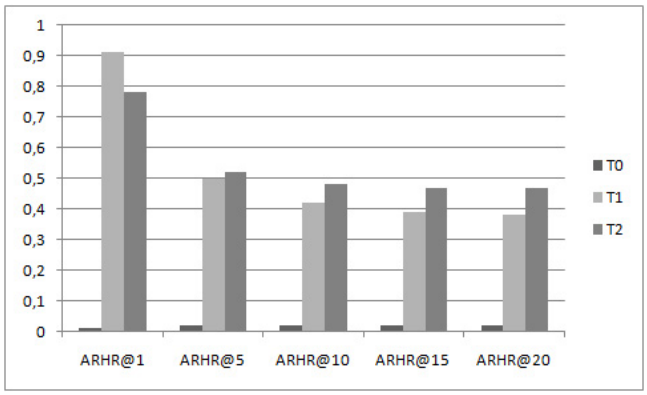

Figure 2: ARHR values of followee recommendations for different profiling strategies

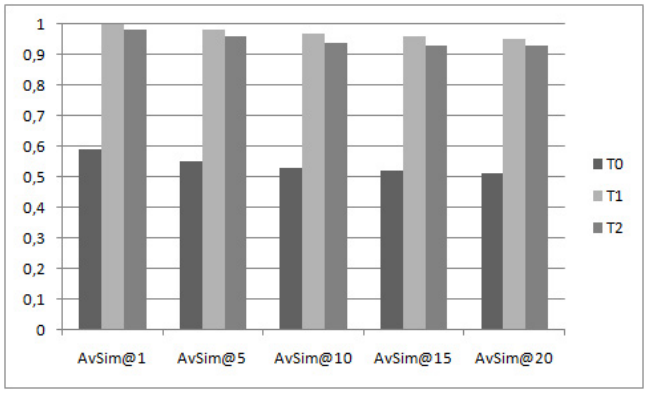

Figure 3: Average similarity of the recommended followees with the target users

5 Conclusions

In this paper an effective algorithm for recommending followees in the Twitter social network dedicated to information-seeking users was presented. This algorithm first explores the social graph in search of candidate recommendations and then ranks these candidates according to the inferred interest of the user that will receive the recommendations on the information the candidates tweet about. The search of suitable candidates was guided by the assumption that the users followed by the followers of a target user followees are potentially interesting and should be further evaluated from a content point of view.

Three different strategies were defined to create content-based profiles of users describing the information they like to received from the people they follow. Using the user's

	T0		T1		T2	
	Average	Std.dev.	Average	Std.dev.	Average	Std.dev.
P@1	0.01	0.01	0.91	0.07	0.78	0.13
P@5	0.01	0.01	0.75	0.08	0.49	0.12
P@10	0.01	0.01	0.61	0.07	0.31	0.09
P@15	0.01	0.01	0.51	0.06	0.22	0.07
P@20	0.00	0.00	0.42	0.06	0.17	0.05
Hits@1	0.01	0.01	0.91	0.07	0.00	0.13
Hits@5	0.03	0.03	3.75	0.38	2.45	0.62
Hits@10	0.06	0.06	6.11	0.70	3.12	0.94
Hits@15	0.08	0.08	7.6	0.96	3.32	1.04
Hits@20	0.09	0.08	8.39	1.2	3.38	1.06

Table 2: Precision and Hit-rate of followee recommendations for different profiling strategies

own tweets, maintaining a term vector for each followee, and grouping followees into categories by means of a clustering algorithm. Thus, candidates are ranked according to the similarity of their tweets with these models of the target user interests in order to recommend a list of top-*N* followees.

Experimental evaluation using a dataset containing a sample of Twitter social graph and the tweets of each user in this graph was carried out in order to validate the approach and compare the performance of the proposed profiling strategies. The achieved results show that the user own tweets are not a good source of profiling knowledge. In contrast, strategies using the posts of the followees of users, either individually or grouped into categories, for modeling their interests reached high levels of precision in recommendation.

Future work will be oriented to obtain further improvements in the performance of the approach by varying the text analysis techniques applied to tweets and the ranking scheme. In the first point, we are currently working on exploiting terms appearing in the URLs linked in tweets as well as words related to hashtags to expand the tweet textual representation. In the second point, the work envisioned consists in measuring the impact that factors such as the number of occurrences in the candidates set or the relation followers/followees that characterize good information sources have on ranking effectiveness.

References

[Cha *et al.*, 2010] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10)*, Washington DC, USA, 2010.

[Chen *et al.*, 2009] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, pages 201–210, Boston, MA, USA, 2009.

[Chen *et al.*, 2010] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10)*, pages 1185–1194, Atlanta, Georgia, USA, 2010.

[Choudhury *et al.*, 2010] M. De Choudhury, Y-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, 2010.

[Davidov *et al.*, 2010] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceeding of the 23rd International Conference on Computational Linguistics (COLING'2010)*, pages 241–249, Beijing, China, 2010.

[Deshpande and Karypis, 2004] M. Deshpande and G. Karypis. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, 22(1):143–177, 2004.

[Esparza *et al.*, 2010] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth. On the real-time web as a source of recommendation knowledge. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*, pages 305–308, Barcelona, Spain, 2010.

[Garcia and Amatriain, 2010] R. Garcia and X. Amatriain. Weighted content based methods for recommending connections in online social networks. In *Workshop on Recommender Systems and the Social Web*, pages 68–71, Barcelona, Spain, 2010.

[Guy *et al.*, 2009] I. Guy, I. Ronen, and E. Wilcox. Do you know?: recommending people to invite into your social network. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09)*, pages 77–86, 2009.

[Hannon *et al.*, 2010] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*, pages 199–206, 2010.

[Java *et al.*, 2007] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st*

SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pages 56–65, 2007.

[Krishnamurthy *et al.*, 2008] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *Proceedings of the 1st Workshop on Online Social Networks (WOSP'08)*, pages 19–24, Seattle, WA, USA, 2008.

[Kwak *et al.*, 2010] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 591–600, Raleigh, North Carolina, USA, 2010.

[Liben-Nowell and Kleinberg, 2003] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the12th International Conference on Information and Knowledge Management (CIKM'03)*, pages 556–559, New Orleans, LA, USA, 2003.

[Lo and Lin, 2006] S. Lo and C. Lin. WMR–A graph-based algorithm for friend recommendation. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, pages 121–128, Washington, DC, USA, 2006.

[Naaman *et al.*, 2010] M. Naaman, J. Boase, and C-H. Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'10)*, pages 189–192, Savannah, Georgia, USA, 2010.

[Pak and Paroubek, 2010] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

[Perez-Tellez *et al.*, 2010] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso. On the difficulty of clustering company tweets. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC'10)*, pages 95–102, Toronto, ON, Canada, 2010.

[Phelan *et al.*, 2009] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to recommend real-time topical news. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*, pages 385–388, New York, NY, USA, 2009.

[Porter, 1980] M. Porter. An algorithm for suffix stripping program. *Program*, 14(3):130–137, 1980.

[Ramage *et al.*, 2010] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, Washington, DC, USA, 2010.

[Salton and McGill, 1983] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Sun *et al.*, 2009] A. R. Sun, J. Cheng, and D. D. Zeng. A novel recommendation framework for micro-blogging based on information diffusion. In *Proceedings of the 19th Workshop on Information Technologies and Systems*, 2009.

[Weng *et al.*, 2010] J. Weng, E-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 261–270, New York, NY, USA, 2010.

[Yamaguchi *et al.*, 2010] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. TURank: Twitter user ranking based on user-tweet graph analysis. In *Web Information Systems Engineering*, volume 6488 of *LNCS*, pages 240–253, Hong Kong, China, 2010.