Amandi). Information Systems. Elsevier Science. ISSN 0306-4379. Vol. 38(8), pp. 1116-1127. 2013.

The final publication is available at http://dx.doi.org/10.1016/j.is.2013.05.009

Followee Recommendation based on Text Analysis of Micro-blogging Activity

M. G. Armentano[a,b,*], D. Godoy[a,b], A. A. Amandi[a,b]

[a] *ISISTAN Research Institute, Fac. Cs. Exactas, UNCPBA, Campus Universitario, Paraje Arroyo Seco, Tandil, 7000, Argentina*

[b] *CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina*

Abstract

Nowadays, more and more users keep up with news through information streams coming from real-time micro-blogging activity offered by services such as Twitter. In these sites, information is shared via a followers/followees social network structure in which a follower will receive all the micro-blogs from his/her followees. Recent research efforts on understanding micro-blogging as a novel form of communication and news spreading medium, have identified three different categories of users in these systems: information sources, information seekers and friends. As the social network grows in the number of registered users, finding relevant and reliable users to receive interesting information becomes essential. In this paper we propose a followee recommender system based on the analysis of the content of micro-blogs to detect users' interests and in an exploration of the topology of follower/followee network to find candidate users for recommendation. Experimental evaluation was conducted in order to determine the impact of different profiling strategies based on the text analysis of micro-blogs as well as several factors that allows the identification of users acting as good information sources. We found that user-generated content available in the network is a rich source of information for profiling users and finding like-minded people.

Keywords: Micro-blogging, Text Mining, Recommender Systems

1. Introduction

Micro-blogging is a new form of communication that is gaining adherents every day. This service allows users to send and post short messages usually containing only text. These updates are shown in the user profile page and are also sent immediately to other users who have chosen the option of receiving them. Twitter is the recognized leader of the microblogging systems. This

*Corresponding author

Email address: marmenta@exa.unicen.edu.ar (M. G. Armentano)

online social networking site has attracted the attention of users as a means of disseminating news and information. Unlike many other on-line social networks such as Facebook, Hi5, Orkut, or MySpace in which users build his/her social connections based mostly on "friendship" relations, it has been demonstrated that only 22.1% of the relations in Twitter are reciprocate [1]. The fact that 77.9% of Twitter connections are unidirectional, in addition to the fact that 67.6% of users are not followed by any of their followees are clear indicators that these users probably use Twitter as a source of information rather than as a social networking site.

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 uses twitter to get information on a particular subject, as a form of RSS reader, registering themselves as followers of their favorite artists, celebrities, bloggers, or TV programs. For this last type of users finding high quality and reliable information sources in the constantly increasing Twitter community becomes a challenging issue.

The facts described above, in addition to the great explosion in the number of registered users in Twitter[1], make us believe that information-seeking users would benefit from a recommender system able to suggest information sources that they might be interest in following. In this article we study Twitter from a user modeling perspective. Our goal is to provide recommendations to information seekers about users that publish tweets that might be of their interest. Twitter itself has included a "who to follow" system[2], but the only information available about the information used to make recommendations is that *"the suggestions are based on several factors, including people you follow and the people they follow"*[3]. Unlike other works that focus on ranking users according to their influence in the entire network [2, 3], the proposed algorithm explores the network starting from the user own follower/following relationships up to a certain level, so that more personalized factors are considered in the selection of candidates for recommendation, for instance other users that are followed by someone that follows some of the same people than me.

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. For profiling a Twitter user we need to make use of context that can be derived from user interactions, content streams and the topology of the network. There are also other sources of information about a user, such as his/her own bio, but it has been demonstrated that many users either do not provide a bio, or have a bio which does not provide any topical

[1] In 2010 Twitter grew by more than 100 million registered accounts (`http://yearinreview.twitter.com/whosnew/`. AccessedonMarch2011

[2] http://twitter.com/who_to_follow

[3] http://blog.twitter.com/2010/07/discovering-who-to-follow.html

information [4]. Preliminary studies have also considered Twitter lists meta-data to derive the topics of interest of a Twitter user [5, 4]. In this work, we use both the structure of the followers/followees network and the tweets published in this network as a mean to recommend people that shares the same content-related interests with a the user who will receive the recommendations.

Several profiling strategies are analyzed and evaluated for modeling a user interests in Twitter based on two general approaches. The first approach models a user by analyzing the content his/her own tweets whereas the second approach represents users by the tweets of their followees. For the second approach, three different types of profiles were considered: modeling a target user by the set of profiles of his/her followees, by the aggregation of the profiles corresponding to his/her followees or by a set of categories that can be discovered by clustering his/her followees according to the content of their tweets.

The rest of this work is organized as follows. Section 2 discusses other research efforts related to our work. 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 social networks. For example, SONAR [6] recommends related people in the context of enterprises by aggregating information about relationships as reflected in different sources within an organization, such as organizational chart relationships, co-authorship of papers, patents, projects and others. Liben-Nowell et al. [7] 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[4]. Authors found that there is indeed useful information contained in the network topology alone. Chen et al. [8] 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) [9] 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 with the problem of taking advantage of the massive, unstructured, dynamic and inherently noisy user-generated content from Twitter for recommendation.

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

Other line of research has been devoted to measure the influence of users in Twitter. In [1] 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. In a posterior study [10], the temporal order of information adoption was also considered to detect effective readers of a tweet. In this study authors concluded that Twitter accounts corresponding to news media has significant influence in spreading information to effective readers. Coincidently, a comparison between in-degree, re-tweets and mentions as influence indicators [11] 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 [2], 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. Romero et al. [12] also considered a user's passivity when computing the propagation of tweets through the network. Garcia et al. [13] propose a method to weight 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. Pal and Counts [14] proposed a set of features for characterizing social media authors and applied a clustering approach over this feature space to detect authoritative sources. These features include both topological (for example, number of followers and number of friends tweeting on a given topic) and content metrics (number of tweets authored by the user, number of retweets of other's tweets, etc). Then, a ranking algorithm based in a Gaussian Mixture model is applied to select the most representative authors for three different topics (*iphone*, *oil spill*, and *worldcup*). Saez-Trumper et al. [15] distinguish trendsetters from influential users in the sense that a trendsetter is not necessarily popular or famous, but the one whose ideas spread over the network successfully before these ideas become popular. Authors presented a new ranking algorithm, *TS* (for TrendSetters), that combines temporal attributes of nodes and edges of the network with a Pagerank based algorithm to find trendsetters for a given topic. Ghosh et al. [16] identify the set of experts related to a topic by extracting the nouns and adjectives from Twitter lists' meta-data (names and description) and associating these terms with the listed users. This approach is based on the intuition that a user listed by many other users under a certain topic is very likely to be an expert on that topic.

While the studies mentioned above focus on analyzing 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. [17] 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 build with the words of the user tweets and a Followee-profile build by combining the 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 systems within a user followee-of-followees neighborhood so that the most popular URLs within the group are recommended. *Buzzer* [18] 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. [19] address the problem of using real-time opinions of movie fans expressed through the Twitter-like short textual reviews for recommendation. This work assumes that tweets carry on 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 [20, 21].

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. [22] 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 [23] 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 ranking of recommendations.

There is not much related work about systematic analysis of the content of the tweets. Perez-Tellez et al. [24] presented a text enrichment technique, called Self-Term Expansion Methodology (S-TEM), aiming at improving the quality of the corpora with respect to task of clustering blogs. They consider blogs to be "short-text" since they trend to exhibit low frequency of terms, a short vocabulary size and vocabulary overlapping of some domains. A main advantage of this approach is that it does not rely on external linguistic resources, but it uses the corpus to be clustered itself to perform the term expansion. The S-TEM methodology comprises a twofold process: the self-term expansion technique, which is a process of replacing terms with a set of co-related terms, and a Term Selection Technique with the role of identifying the relevant features. A similar approach is applied in [25] to categorize tweets which contain a company name, into two clusters corresponding to those which refer to the company and those which do not. Besides S-TEM, three other techniques are applied: Term Expan-

sion Methodology - Wiki (TEM-Wiki), which enhance S-TEM by considering additional information extracted from Wikipedia, Term Expansion Methodology with Positive examples - Wiki (TEM-Positive-Wiki), where the TEM-Wiki methodology is used for enriching only tweets that really refer to companies, and Full Term Expansion Methodology (TEM-Full), where ambiguous words are expanded with all those words that co-occur with it in the same class of the corpus.

In the direction of tweets classification, Naaman et al. [26] distinguish "informers" users, whose tweets contain mainly non-personal information, from "meformers" users, who mainly post statuses updates about themselves. Ramage et al. [27] 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 event, 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.

3. Followee Recommendation in Twitter

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. [28] and Krishnamurthy et al. [29] presented a characterization of Twitter users grouping them into three categories:

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

Kwak et al. [1] 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. This fact, in addition to the great explosion in the number of registered users in Twitter (in 2010 Twitter grew by more than 100 million registered accounts[5]), make us believe that these people seeking for information would benefit from a recommender system able to suggest information sources that they might be interest in following.

[5] http://yearinreview.twitter.com/whosnew/.AccessedonJanuary2011.

The problem of followee recommendation in Twitter is aimed to information seekers and consists in identifying users posting relevant tweets for a target user, so that he/she can subscribe to these users and starts receiving real-time information from them. The approach presented in this work can be decomposed into two main parts, first, finding a suitable group of candidate users to be evaluated for recommendation and, then, determining whether the information that they publish may be of interest for the target users. Section 3.1 describes the search for candidates based on the topology of the Twitter network and Section 3.2 explained the construction of profiles to evaluate candidates based on their posts.

3.1. Topology-Based Candidate Search

In order to recommend Twitter users, a set of viable candidates needs 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 that also follows some of the users in this group (i.e. is subscribe to some of the same information sources) have interests in common with the target user and might have discover other relevant information sources in the same topics, which are in turn their followees. Figure 1 illustrates these approach for candidate selection schematically. This scheme for searching candidates was successfully applied in our previous work [30] where we conducted an experiment with real users and test only topology-based metrics for ranking candidates.

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 P, i.e. $P = \bigcup_{\forall l \in L} followees\,(l)$.
4. Exclude from P those users that the target user is already following. Let's call the resulting list of candidates R.

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

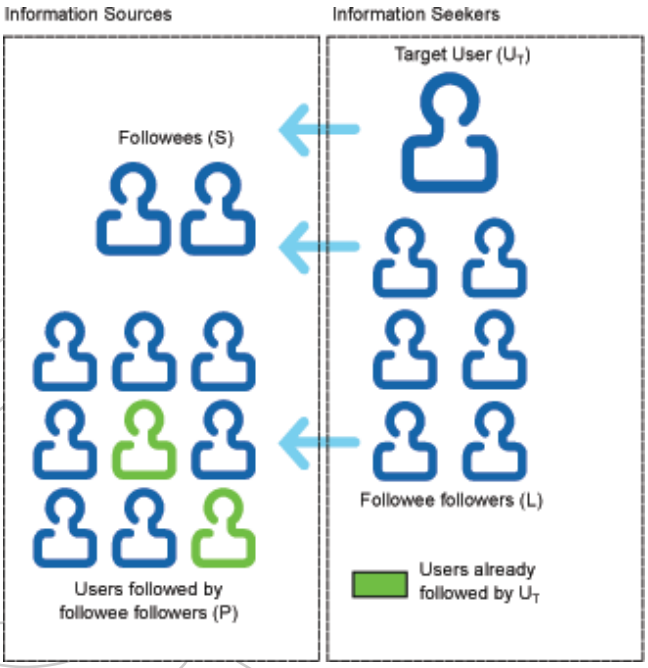


Figure 1: Strategy for exploring the followee/follower network to find candidate users

composed of information sources, L will be other users looking for information like u_T and P 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 considered to boost its chances of being recommended.

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 that are not in the proximity of the target user. For example, looking for users mentioned in an information stream provided by Twitter which is known as *public timeline* can be an alternative. This stream 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.

3.2. Content-Based Profiling

Once a list of viable candidates R is available, the matching between the information each user $r \in R$ publish in Twitter and the user interests need to be evaluated in order to obtained a ranked list of followee recommendations. The interests of a target user can be described using different content sources as well as relations in Twitter, such as the text of his/her own tweets or the tweets published by the users that he/she follows.

For a target user u_T, $tweets\,(u_T)$ is the set of all user posts, this is:

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

The simplest alternative to build a profile for a user in Twitter is aggregating his/her own tweets under the assumption that users are likely to tweet about things that interest 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. 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 $sim^{T0}\,(u_C, u_T)$, is simply 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 [31].

Information seekers characterize for posting few tweets themselves, but follow people that 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 build based on the observation that a user has a number of followees:

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

and the information a user is interested in can be seen as the aggregations of the profiles of his/her followees:

$$Profile^{T1}\,(u_T) = \sum_{i=1}^{l} Profile^{T0}\,(f_l) \tag{4}$$

where $sim^{T1}\,(u_C, u_T)$ measures the level of resemblance between $Profile^{T1}\,(u_T)$ and $Profile^{T0}\,(u_C)$. For the active user the profile models the information he/she likes to read, whereas for the candidate user u_C the profile models the information he/she published. Both vectors are compared in sim^{T1} using the cosine similarity.

This profiling strategy aggregates in a single vector the information published by the user followees instead of the user own tweets. However, it considers all followees as responding to a unique topic of interest, which is not enough to effectively model multiple user interests in diverse areas. For example, a user may follow celebrities, politicians, sportsmen and other type of users which information will coexist in the vector representing the target user profile.

Rather than a single vector representing all of the user interests, multiple vectors each representing a different followee allows to attained 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:

$$Profile^{T2}(u_T) = \left\{Profile^{T0}(f_1), \ldots, Profile^{T0}(f_l)\right\} \tag{5}$$

To evaluate whether to recommend a candidate user u_C to the target user u_T the similarity between the information that publish the candidate, $Profile^{T0}(u_C)$, needs to be compared with the profile of the target user $Profile^{T2}(u_T)$, this is the information u_T is subscribed to receive in Twitter. The matching is then calculated as:

$$sim^{T2}(u_C, u_T) = \max_{\forall i: f_i \in followees(u_T)} sim^{T0}\left(Profile^{T0}(f_i), Profile^{T0}(u_C)\right) \tag{6}$$

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 followees belong to 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 to either some of the existent categories or to a novel category depending on its similarity with the current categories. Hence, user 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.

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

$$c_{fc_i} = \frac{1}{|fc_i|} \sum_{f \in fc_i} Profile^{T0}(f) \tag{7}$$

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.

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: $sim_j \leftarrow sim(c_j, f_i)$
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: Include fc_{new} into FC_u
19: **end if**
20: **end for**
21: Return FC_u

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

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. The profile of a user using this strategy is then defined as the set of the centroids of the clusters identified:

$$Profile^{T3}(u_T) = \{c_{fc_1} \ldots, c_{fc_m}\} \tag{8}$$

Finally, the similarity $sim^{T3}(u_C, u_T)$ is evaluated as specified in Equation 6.

In summary, two general approaches are evaluated in this paper for modeling a user 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, three different mechanisms to combine the vectors of the user followees were analyzed.

The first consist in modeling a target user interests by joining the content of all their followees. The second models a target user using a set of vectors, each of them representing the content of a user followee tweets. The third 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. Finally, 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.

The text of tweets was processed for obtaining the vector of a given user posts $Profile^{T0}(u)$ applying a number of filters in a pipeline. First, tokens composed of punctuation symbols only are assumed to be emoticons and removed. Second, common slang vocabulary and abbreviations are substituted. The words are widely used in Twitter messages to overcome the limitation in the number of characters. The NoSlang on-line dictionary[6], 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". Ultimately, stop-words are removed and Porter stemming algorithm [32] is applied to the remaining words.

4. Experimental Evaluation

4.1. Dataset Description

The Twitter dataset[7] 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 [33], 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. Users who posted less than 10 tweets were excluded from the social graph so that valuable content-based profiles can be extracted for all users involved in the evaluation. The filtered dataset contained 100,727 unique users and 54,495 relationships.

From the entire dataset a test set 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 users in the test-set were selected on the basis of having their followees outnumbering their followers. Thus, a information source index (IS) was computed according to the following equation:

$$IS(u) = \frac{\frac{followers(u) - followees(u)}{followers(u) + followees(u)} + 1}{2}$$

[6] `http://www.noslang.com/dictionary`

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

	Average	Maximum	Minimum
#followees	87.20±47.34	212	1
#followers	0.9±1.89	27	0
#tweets	97.18±57.49	200	11

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

This index is a number in the interval $[0, 1]$ so that a user with a IS value close to 1 is a good information source whereas an information seeker will have a IS value near to 0. Based on this index we selected all user such that $IS(u) < 0.5$, indicating that user fits the behavior of a information seeker better than the behavior of a information source. The resulting test set had a size of $|U_{test}|$ =530. The profiles of these target users were built by analyzing the text of their tweets according to the strategies proposed in Section 3.2. Table 1 summarizes the characteristics of the U_{test} in terms of number of followees, number of followers and published tweets.

4.2. Methodology and Metrics

Experiments were carried out using a holdout strategy in which some of the target user followees are hide 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:

$$precision\,(RE) = \frac{1}{|U_{test}|} \sum_{u \in U_{test}} \frac{|followees_{test}\,(u) \cap RE_u|}{|RE_u|} \tag{9}$$

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, 15, 10 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.

Other 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 [34]:

$$HR = \frac{number\,of\,hits}{|U_{test}|} \tag{10}$$

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:

$$ARHR(RE) = \frac{1}{|U_{test}|} \sum_{i=1}^{h} \frac{1}{p_i} \tag{11}$$

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

Figures 2 and 3 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 in average 9856.5±701.92 users reached through the user own followees.

It can be observed in the figures 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. Likewise, summing all followees into a single vector performs poorly because the vector fails at representing the multiple interests of users.

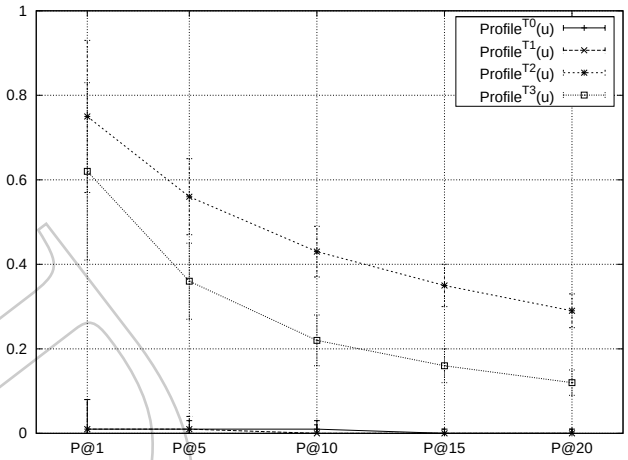


Figure 2: Precision of followee recommendations for different profiling strategies

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. Further experiments will be conducted to evaluate other settings of the similarity threshold δ, that was assigned to 0.5, and other well established clustering algorithms.

Pre-processing techniques apply to the text of tweets included the substitution of slang-vocabulary, removal of stop-words, and then stemming. Figure 4 shows the impact of the slang and stemming filters over a basic pre-processing of stop-words and emoticons removal. It can be deduced from the image that the impact of pre-processing strategies is not significant in terms of precision in the recommendations. These results might be due tothe fact that the dimensionality reduction achieved with stemming has not effect in the already small space product of short-texts, although this issue should be further studied.

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 uninteresting, 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 because either they are not interested on receiving their tweets or 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.

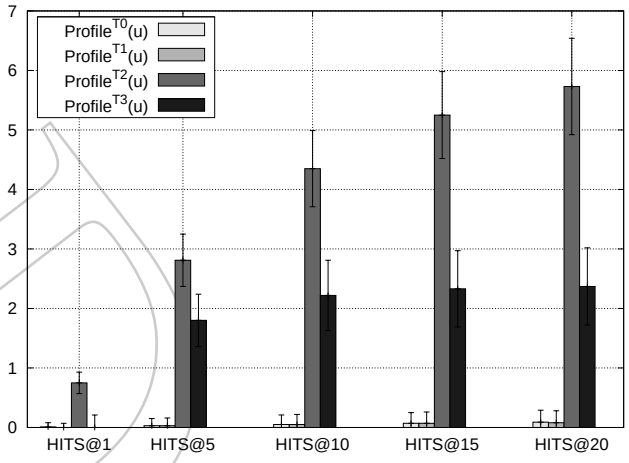


Figure 3: Hit-rate of followee recommendations for different profiling strategies

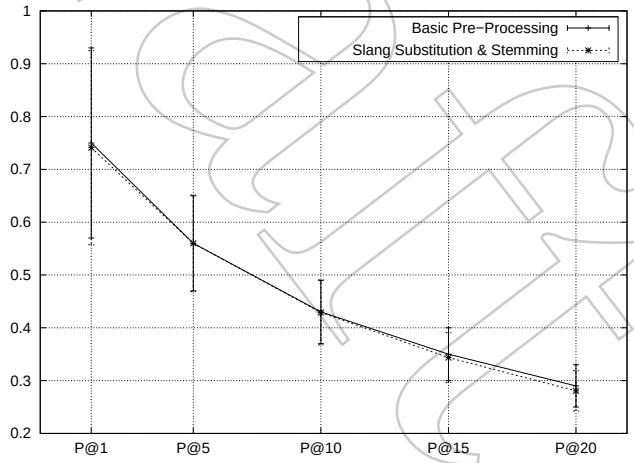


Figure 4: Impact of pre-processing techniques on the precision of recommendations

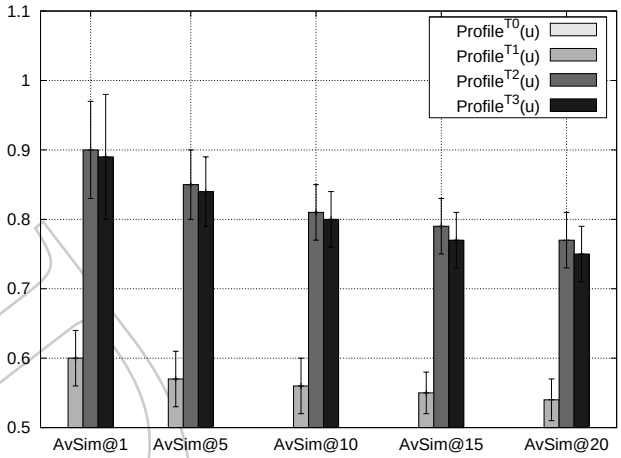


Figure 5: Average similarity of the recommended followees with the target users

Figure 5 depicts the mean average similarities between the vectors of the users in the top-N lists with the corresponding target user profile, while error bar indicates its standard deviations. The low similarity of information published by the recommended users and the target user tweets account for the poor results of the first profiling strategies. The small deviations in the 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.

Interestingly, the ARHR values shown in Figure 6 for the four profiling strategies allows 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 mechanisms.

4.4. Comparison with similar approaches

From the related work, the approach that we find closest to ours is *Twittomender*, proposed by Hannon et al. [23]. Similarly to our approach, *Twittomender* only uses the network structure and the content of tweets to generate recommendations. Since, results reported in [23] are not fully comparable to the results presented in this article as different datasets were used, we implemented Hannon et al. approach in order to use the same dataset we used in our experiments. In this section we present a comparison of the performance of both approaches.

Twittomender creates different indexes for all users in the dataset generated

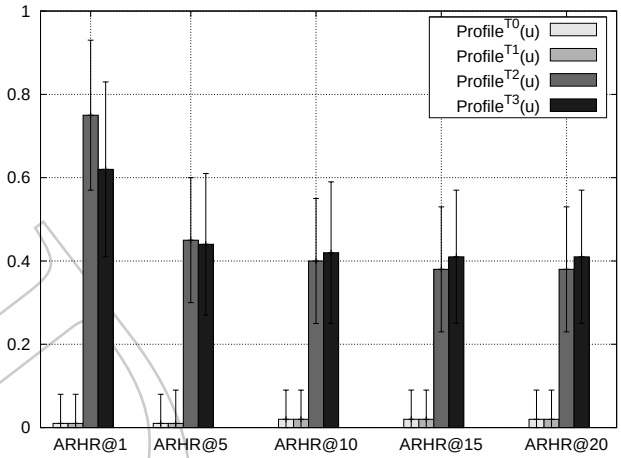


Figure 6: ARHR values of followee recommendations for different profiling strategies

from different sources of profile information. Four of these indexes are content-based, modelling users by:

- their own tweets (S1),
- the tweets of their followers (S2),
- the tweets of their followees (S3),
- a combination of the S1, S2 and S3 (S4).

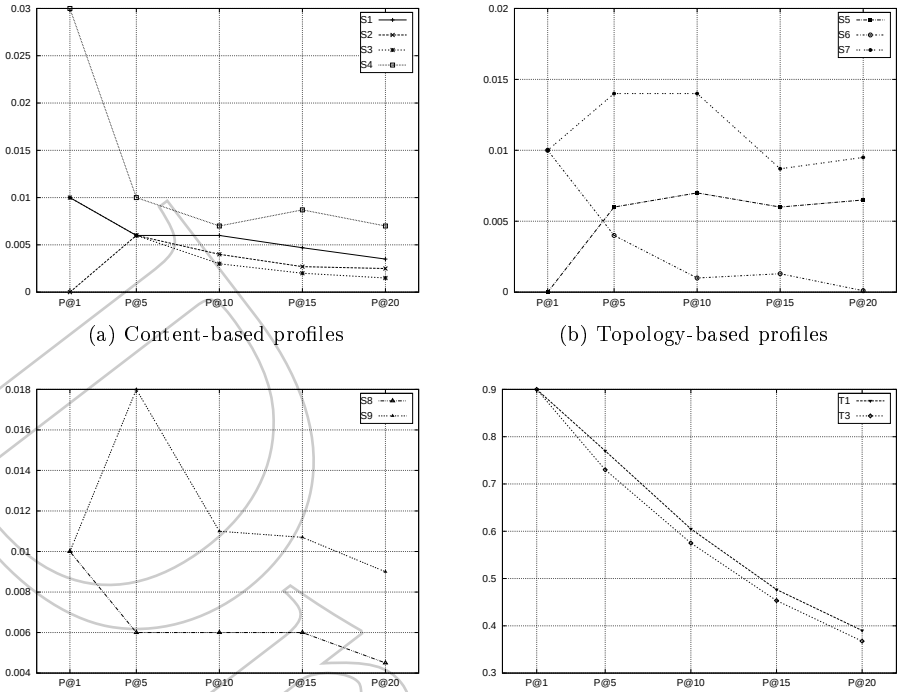
The three remaining strategies are topology-based and model users by:

- the IDs of their followees (S5),
- the IDs of their followees (S6)
- a combination of S5 and S6 (S7).

Additionally, two hybrid ensemble strategies compose a selection of previous basic component recommenders, S1 to S7, and the union of the recommendations from these independent strategies is scored and ranked:

- a combination of strategies S1 and S6 (S8)
- basing the scoring function on the position of the user in each of the recommendation lists so that users that are frequently present in high positions are preferred over users that are recommended less frequent or in lower positions (S9).

We run a set of experiments using the same methodology to that presented in Section 4.3, with a 70/30 hold-out technique per user profile. Figure 7 shows the



(a) Content-based profiles

(b) Topology-based profiles

(c) Hybrid profiles

(d) Our approach

Figure 7: Precision for different profiles used in *Twittomender* approach and our approach

precision of *Twittomender* profiles using different strategies, grouped by those which model a user using the content of tweets obtained from different sources (7a), those which model a user with the IDs of his/her followees and followers (7b) and the two hybrid approaches (7c). Figure 7(7d) show the precision obtained using our approach.

It can be seen that strategy S4 (combination of the tweets of the user's followees and followers) outperformed the other content-based strategies, while strategy S7 (combination of the IDs of the user's followees and the followers) performed better for topology-based strategies. On the other hand, the hybrid approach S9 showed a better general precision compared to all other strategies.

The results reported in [23] vary between $\sim$0.15 to $\sim$0.3 of precision, while the best precision obtained with our testing methodology was 0.018 for strategy S9 and a recommendation list of five users. There are several differences between the methodology used in this paper and that used in *Twittomender*. First, *Twittomender's* dataset consisted in 20,000 users with at most 100 tweets per user while the dataset used in this paper contained 456,107 users and 10,467,110 tweets. Although our approach might have a reduced serendipity, there are more chances of finding potentially interesting users in the neighborhood it explores.

Second, *Twittomender's* approach needs all profiles to be indexed, and recommendation lists are obtained by evaluating the full set of users in the dataset. In the case of the experiments reported in [23], candidate lists had always the same size: 19,000 users used for testing purposes. Our approach, on the other hand, selects a set of candidate users from the neighborhood of the target user being the number of candidates explored, in average, 9856.5 ± 701.92 users. A final and important reason justifying the differences between the precision results reported in this paper for *Twittomender* profiles and those reported in [23] is that we try to rediscover users whose connection to the target user was hidden during training. As stated before, Hannon et al. build the profiles of the testing users using all the information of their followers/followees, without hiding those connections that were expected to be rediscovered. Finally, an important shortcoming of Hannon et al. approach is the need of indexing tweets, which can be highly computationally expensive in a real-time environment such as Twitter.

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 recommendation 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-based point of view.

Four different strategies were defined to create content-based profiles of users describing the information they like to receive from the people they subscribed. Using the user own tweets, aggregating their followees vector representations, maintaining a 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 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 approach performance 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. Furthermore, following the ideas presented in [16], our user profiles could be combined with the information that can be extracted from users Lists metadata and the users listed in them. We

believe that this approach is complementary to ours, and that the knowledge we can obtaind about each user can be enhanced by a combination of both approaches. 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.

Acknowledgments

This research was supported by the National Scientific and Technical Research Council (CONICET) under grant PIP No. 114-200901-00381 and PIP No. 114-201101-00181.

[1] H. Kwak, C. Lee, H. Park, 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), Raleigh, North Carolina, USA, 2010, pp. 591–600.

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

[3] Y. Yamaguchi, T. Takahashi, T. Amagasa, H. Kitagawa, TURank: Twitter user ranking based on user-tweet graph analysis, in: Web Information Systems Engineering, Vol. 6488 of LNCS, Hong Kong, China, 2010, pp. 240–253.

[4] N. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, K. Gummadi, Inferring who-is-who in the twitter social network, in: Workshop on Online Social Networks (WOSN 2012), 2012.

[5] R. Pochampally, V. Varma, User context as a source of topic retrieval in twitter, in: Workshop on Enriching Information Retrieval (ENIR 2011), 2011.

[6] I. Guy, I. Ronen, 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), 2009, pp. 77–86.

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

[8] J. Chen, W. Geyer, C. Dugan, M. Muller, 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, Boston, MA, USA, 2009, pp. 201–210.

[9] S. Lo, 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), Washington, DC, USA, 2006, pp. 121–128.

[10] C. Lee, H. Kwak, H. Park, S. Moon, Finding influentials based on the temporal order of information adoption in twitter, in: Proceedings of the 19th international conference on World wide web, WWW '10, ACM, New York, NY, USA, 2010, pp. 1137–1138.

[11] M. Cha, H. Haddadi, F. Benevenuto, 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.

[12] D. M. Romero, W. Galuba, S. Asur, B. A. Huberman, Influence and passivity in social media, in: Proceedings of the 20th international conference companion on World wide web, WWW 2011, ACM, New York, NY, USA, 2011, pp. 113–114.

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

[14] A. Pal, S. Counts, Identifying topical authorities in microblogs, in: Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, ACM, New York, NY, USA, 2011, pp. 45–54.

[15] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, F. Benevenuto, Finding trendsetters in information networks, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2012.

[16] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, K. Gummadi, Cognos: crowdsourcing search for topic experts in microblogs, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12, ACM, New York, NY, USA, 2012, pp. 575–590.

[17] J. Chen, R. Nairn, L. Nelson, M. Bernstein, 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), Atlanta, Georgia, USA, 2010, pp. 1185–1194.

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

[19] S. G. Esparza, M. P. O'Mahony, 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), Barcelona, Spain, 2010, pp. 305–308.

[20] A. Pak, 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.

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

[22] A. R. Sun, J. Cheng, 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.

[23] J. Hannon, M. Bennett, 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), 2010, pp. 199–206.

[24] F. Perez-Tellez, D. Pinto, J. Cardiff, P. Rosso, Improving the clustering of blogosphere with a self-term enriching technique, in: Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD'09), Pilsen, Czech Republic, 2009, pp. 40–47.

[25] F. Perez-Tellez, D. Pinto, J. Cardiff, 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), Toronto, ON, Canada, 2010, pp. 95–102.

[26] M. Naaman, J. Boase, 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), Savannah, Georgia, USA, 2010, pp. 189–192.

[27] D. Ramage, S. Dumais, 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.

[28] A. Java, X. Song, T. Finin, 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, 2007, pp. 56–65.

[29] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about Twitter, in: Proceedings of the 1st Workshop on Online Social Networks (WOSN'08), Seattle, WA, USA, 2008, pp. 19–24.

[30] M. Armentano, D. Godoy, A. Amandi, Topology-based recommendation of users in micro-blogging communities, Journal of Computer Science and Technology 27 (3) (2012) 624–634.

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

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

[33] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, 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.

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