Learning and Adapting User Criteria for Recommending Followees in Social Networks

Antonela Tommasel
Facultad de Ciencias Exactas, ISISTAN, UNICEN-CONICET, Campus Universitario, Tandil (B7001BBO), Buenos Aires, Argentina. E-mail: antonela.tommasel@isistan.unicen.edu.ar

Daniela Godoy
Facultad de Ciencias Exactas, ISISTAN, UNICEN-CONICET, Campus Universitario, Tandil (B7001BBO), Buenos Aires, Argentina. E-mail: daniela.godoy@isistan.unicen.edu.ar

The accurate suggestion of interesting friends arises as a crucial issue in recommendation systems. The selection of friends or followees responds to several reasons whose importance might differ according to the characteristics and preferences of each user. Furthermore, those preferences might also change over time. Consequently, understanding how friends or followees are selected emerges as a key design factor of strategies for personalized recommendations. In this work, we argue that the criteria for recommending followees needs to be adapted and combined according to each user's behavior, preferences, and characteristics. A method is proposed for adapting such criteria to the characteristics of the previously selected followees. Moreover, the criteria can evolve over time to adapt to changes in user behavior, and broaden the diversity of the recommendation of potential followees based on novelty. Experimental evaluation showed that the proposed method improved precision results regarding static criteria weighting strategies and traditional rank aggregation techniques.

Introduction

Nowadays, online social networks play an important role in the life of millions of users, who actively use them not only for sharing content, but also for finding new friends. In face-to-face relationships, exposure to people of similar age, socioeconomic status or educational level in schools, universities, workplaces, or neighborhoods can promote the development of relationships with liked-minded persons (Golder & Yardi, 2010). However, finding new friends in an online context where most users do not know each other personally, and the only information known about others is through their short profiles, might be a challenging task.

The accurate suggestion of potentially interesting friends arises as a crucial issue, accentuated by the overload of available information and networks size. Hence, several approaches (Armentano, Godoy, & Amandi, 2011; Brzozowski & Romero, 2011; Hannon, Bennett, & Smyth, 2010) have been proposed to suggest users worth following in social networks, mostly based on the principle of homophily, that is, people tend to strengthen their connection with similar individuals (McPherson, Smith-Lovin, & Cook, 2001). In turn, similarity can be expressed in terms of users' interests, network topology, personality, popularity, geographic location, published content, or even emotions.

The decision to start following other users in social networks might involve, possibly simultaneously, several reasons. Interestingly, those reasons might differ according to the characteristics, experience, behavior, or life circumstances of each user. For example, users might choose to follow some users because they share mutual friends, others because they are celebrities, or others because they publish interesting information, among other possible explanations.

Understanding how users select their friends or followees emerges as a key design factor of strategies for personalized recommendations. Interestingly, most followee recommendation approaches have been only based on topological or content factors. Mostly, such approaches assume that those factors are equally important to each user, disregarding how users' interests and goals can affect followee selection, and thus, whether such factors need to be combined or adapted to each user. Moreover, traditional approaches ignore the fact that followee preferences might change over time (Liu & Turtle, 2013). To cope with dynamic interests, user profiling approaches need to not only track changes of users'

Received May 26, 2016; revised December 2, 2016; accepted January 21, 2017

interests to recognize new interests and forget old ones, but also to include the knowledge about past experiences in the decision-making process.

In dynamically changing environments, such as social networking sites, data distribution can change over time, yielding the phenomenon of "concept drift" (Gama, Žliobaitė, Bifet, Pechenizkiy, & Bouchachia, 2014). Concept drift refers to changes in the conditional distribution of the output (e.g., users' interests regarding their followees), while the input distribution stays unchanged (e.g., the pool of social network users). Learning algorithms, such as recommendation systems, operating in these settings need mechanisms to detect and adapt to the evolution of data over time, otherwise the quality of recommendations will degrade.

This work hypothesizes that the diverse criteria for recommending potential followees have a distinctive impact on the accurate prediction of followees (Tommasel & Godoy, 2015). In this regard, this study aims at verifying whether combining and adapting the importance of diverse recommendation factors to each user's characteristics helps to improve the quality of followee recommendations. Hence, a method for adapting the followee selection criteria to the decisions of each user regarding the characteristics of previously selected followees is proposed. Furthermore, the method has the ability of evolving such criteria over time according to changes in user behavior or interests.

Related Work

Several approaches have been proposed to suggest users worth following in social networks mostly based on a unique and independent factor. For example, considering network topology, Golder and Yardi (2010) found that transitivity and mutuality are significant predictors of the formation of new ties, whereas reciprocity had no significant effect. Brzozowski and Romero (2011) showed that structural closures significantly outperform recommendations based on traditional collaborative filtering, behavioral, and similarity features.

Regarding content-based recommendations, Hannon et al. (2010) suggested that, noisy as Twitter content can be, it could provide useful profiling information. Schaal, O'donovan, and Smyth (2012) tried to quantify whether topological neighbors share interests over similar topics. The authors concluded that the combination of topic similarity and full texts was more useful than hashtags, highlighting the potential of topic proximity for selecting followees.

Most approaches combining several factors assume that all of them are equally important to each user, that is, factors' weights are not personalized according to users' characteristics. For example, Armentano et al. (2011) combined topological and popularity factors by computing their average and product. However, the best results were obtained when the factors were individually used.

Chen, Cui, and Jin (2016) recommended followees based on a variation of the latent factor model that penalized mistakes in the first ranking positions. Two factors were considered: statistics of tweet's content and social information, including both the explicit followee–follower relationships, and the social interaction affinity of users (i.e., how frequently users interact with others). Experimental evaluation based on data from Sina Weibo showed that combining social and content information obtained the best results, although no significant difference was observed between the results of each individual factor.

Yuan, Murukannaiah, Zhang, and Singh (2014) proposed recommending followees based on topological or content-based factors, and sentiment homophily towards topics. Experimental evaluation considered political tweets, and compared the performance of each individual factor with every possible combination of them, assigning equal weights to factors. Regarding the individual factors, topological features obtained the best results, which were improved when adding sentiment-based features. Combining the three factors did not further enhance results, meaning that content was not an important predictor in the evaluated data set.

Additionally, Tommasel, Corbellini, Godoy, and Schiaffino (2016) combined topology and content with users' personality. An empirical analysis demonstrated that the accurate appreciation of traditional factors in combination with a quantitative analysis of personality is crucial for guiding the followee search. However, the importance of each factor was not personalized.

Closely related to this work are the studies (Agarwal & Bharadwaj, 2013; Garcia & Amatriain, 2010) that adapted the similarity between users by personalizing the weight of different factors. Agarwal and Bharadwaj (2013) proposed an evolutionary algorithm to learn the individual preferences of users towards gender, age, language, hometown, relationship status, religion, educational status, visited places, career interests, wall post behavior, and profile visits, among others. The genetic algorithm learned the weights of each characteristic under the guidance of ratings in the training set. The obtained weights represented the priorities that users assigned to each characteristic, and thus, their contribution to the recommendation process. Although recommendations were improved when considering personalized weights, the approach presented some limitations. First, weights were never updated, thus disregarding changes in user behavior over time. Second, due to the high computational cost of genetic algorithms, weights must be learned offline, hindering its applicability in frequently updated online systems.

Finally, Garcia and Amatriain (2010) combined user popularity (the ratio between followees and followers) and activity (the number of posted tweets). Each factor received a personalized weight according to the percentage of followees satisfying certain constraints. A factor was considered relevant to a user if its weight exceeded a threshold. For a potential followee to be recommended, his/her scores had to match the scores of the relevant factors for the target user. Experimental evaluation based on Twitter showed that assigning personalized weights to each feature lead to more accurate predictions than individually considering each

factor. Although the improvements over the individual features were small, the results reinforced the idea that user recommendation cannot be solely based on one criterion, as users can connect for several personal reasons. Like the previous approach, changes in user interests that were not considered as weights were never updated.

Methods

This work presents an adaptive method for personalizing followee recommendation aiming at suggesting potentially interesting followees by searching an optimal combination of followee recommendation factors. Such combination is unique to each user, as it is based on her/his behavior and preferences, as reflected in the previously selected followees. The method involves computing the personalized weights for each recommendation factor, updating them over time, and ranking potential followees according to their similarity with the target user.

Similarity Computation

Extensive research, (Gerani, Zhai, & Crestani, 2012; Vogt & Cottrell, 1999; Wu, 2012), among others, has shown that linear combination is one of the simplest and most effective methods for combining multiple scores. For example, for a recommendation system, Equation (1) depicts the overall similarity (denoted *Similarity*) between users u and v as a linear combination of their similarity regarding each followee recommendation factor $sim_i\ (u, v)$ and their corresponding weights α_i.

$$Similarity(u, v) = \sum_{i=1}^{n} \alpha_i * sim_i\ (u, v) \quad (1)$$

Linear combination has a low computational cost, allowing to efficiently perform online recommendations. Moreover, it is flexible, as different weights can be assigned to individual similarity scores to improve the final one. However, it might be difficult to assign optimal weights to all scores or, in this case, followee recommendation factors. Finally, it also allows including new recommendation factors without changing the combination strategy.

Computing and Updating Factor Weights

The method presented in this work tackles the problem of how to compute each factor's weight α_i, and then how to update them. As recommendation systems aim at finding the most similar potential followees, weights should be defined such that they accurately capture user preferences. For this purpose, the characteristics of previously selected followees are used for defining the similarity weights. For each user it is analyzed whether the past followee selections responded to any of the followee recommendation factors.

Followees are assumed to be chosen by their relevance regarding a determined factor if the similarity between them and the target user for such a factor is higher than a threshold. Then, the preference of users regarding the different factors can be defined as the proportion of selected followees having a similarity with the target user above a certain threshold, and that the similarity for the other factors did not surpass their associated thresholds. Finally, the computed percentages are used as the similarity weights that will be further updated as new followees are discovered and accepted. This weight definition guarantees that the overall similarity between two users ranges between $[0, 1]$.

Figure 1 depicts an example of how the weights corresponding to a certain user are updated. Consider the case in which the target user is presented with five potential followees, accepting all but one. For each accepted followee, it is analyzed whether he/she can be considered to be selected by any recommendation factor. The second step is to guarantee that the followee is not relevant for any other factor. Note that it is also possible for a followee to be accepted by either a mix of factors (which can be found by disregarding the second comparison) or none of them. Finally, weights are updated to reflect the interest of the target user regarding this particular set of followees. Once the factors' weights are obtained, they are used for computing the similarity between new potential followees and the target user, and continuing with the recommendation process. F1

Ranking Recommended Followees

For generating a recommendation list, candidate followees have to be ranked. The most important premise upon which recommender systems are defined is that similar users are likely to have similar interests. Consequently, algorithms rely on similarity metrics to generate recommendations by ranking followees according to their Equation (1) scores. As all candidates are similar to the target user, they are likely to be similar to each other, thus recommendation systems tend to overspecialize (Adomavicius & Tuzhilin, 2005), limiting the range of items that are presented to users. Hence, such algorithms will never uncover certain items, which, although less similar to target users, are nevertheless important to them (Hurley & Zhang, 2011). For instance, in such systems, users that mainly choose followees for the content they publish would be never recommended users that do not publish interesting content but are topologically closed.

To address the overspecialization of recommendations, it is desirable to also suggest novel and/or diverse items to avoid always recommending the same type of items, and enabling the discovery of new and interesting items. For example, when considering several factors for followee recommendation, if users tend to select all the followees by only one factor, it would be desirable to recommend some relevant items regarding the other factors.

Novelty refers to how different an item is regarding the already known ones. Conversely, diversity refers to how different items are with respect to each other (Vargas & Castells, 2011). The two concepts are related, as when items are diverse, they are also novel. Generally, the purpose of recommendations is related to the notion of discovery; hence,

DOI: 10.1002/asi

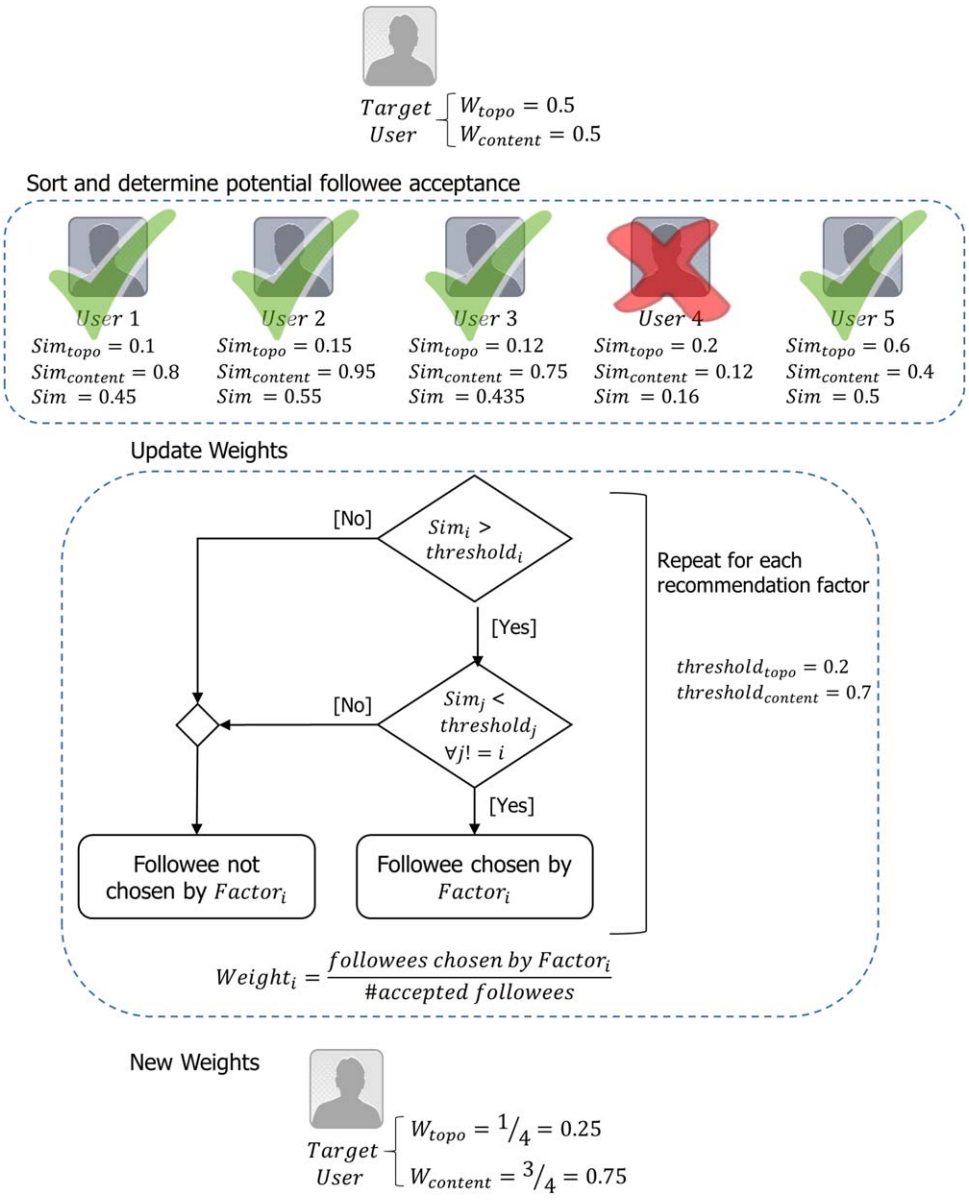

FIG. 1. Example of weight update. [Color figure can be viewed at wileyonlinelibrary.com]

recommendations should expose users to relevant items that they would not have found by themselves, that is, novel items. Moreover, predicting user interests inherently involves uncertainty, since it might be based on implicit or incomplete evidence of interests, which are also subject to changes over time. Consequently, avoiding a too-narrow choice of items enhances the possibilities of users to be pleased with the recommended items.

Similarity-based algorithms can be modified to balance both the relevance of a candidate followee (the similarity to target users as in Equation [1]) and the diversity or novelty of recommendations. The novelty of an item can be measured in terms of the degree to which it is unusual with respect to the target user interests (the previously selected followees) by means of a similarity-based model considering new and previously known interests. This similarity-based definition of novelty contrasts with Boolean-based novelty ones in which items are assigned a novelty of 1 if the item was already known, and 0 otherwise (Vargas & Castells, 2011). On the contrary, similarity-based definitions reflect the partial knowledge regarding user interests by which items might be familiar to users even if no direct interaction between them is observed.

The novelty of a potential followee can be computed as Equation (2) shows, where u represents the target user, pf represents the potential followee, $followees(u)$ represents the previously selected followees of u and *Similarity* is computed as in Equation (1).

$$novelty(pf) = \frac{\sum_{i \in followees(u)} abs(Similarity(u,i) - Similarity(u,pf))}{|followees(u)|} \quad (2)$$

DOI: 10.1002/asi

The rationale for considering the absolute difference between the similarities among previously selected followees, the potential followee and the target user is that if previously selected followees are similar to the target user, and the new potential followee is dissimilar to the target user, the new potential followee will also be dissimilar to previously selected followees. The higher the absolute differences, the higher the dissimilarity, and thus the higher the novelty introduced by the potential followee. Consequently, the novelty of a potential followee can be assessed without computing the actual dissimilarity between the potential followee and each previously selected followee, which would result in high computational complexity.

As the proposed novelty definition considers the average of the absolute differences, it might occur that, in particular cases, both similar and dissimilar users would yield the same novelty score. For example, potential followees who resemble already known followees would entail more novelty than potential followees whose similarities lie between the known followee similarities. In such cases, the definition in Equation (2) could be weighted by the minimum absolute similarity difference.

To generate recommendation lists combining similar items with novel and diverse ones, the relevance and novelty conveyed by the potential followees are linearly combined. The novelty's weight is computed as the percentage of previously selected followees for whom the novelty score was higher than a threshold. Similarly, relevance's weight is computed as the percentage of previously selected followees for whom novelty was lower than the threshold. Both weights are updated as previously described.

Experimental Evaluation

This section presents the experimental evaluation performed to assess the effectiveness of the proposed method, describing the selected recommendation factors, implementation details, the used dataset, and finally, the obtained results.

Factors for Followee Recommendation

Although the presented method could be applied to any arbitrary number of factors, this work focuses on the two main followee recommendation factors and their variants: topology and content.

Topology

Most link prediction algorithms are based on network topology. Typically, they compute the similarity between nodes based on their neighborhoods. Particularly, two topological metrics, which are usually applied to Twitter networks, were included in this study. First, *Common Followees* ($\frac{\Gamma_{out}(x) \cap \Gamma_{out}(y)}{\Gamma_{out}(x) \cup \Gamma_{out}(y)}$) that measures the overlap of the followee sets, that is, to what extent two users follow the same people. It assumes that if two users follow the same people, they are likely to have shared interests. Second, *Sørensen Index* ($\frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}$) that measures the number of shared neighbors, but penalizes it by the sum of the neighborhoods. In these metrics, x and y denote nodes, $\Gamma(x)$ denotes the set of neighbors of x, $\Gamma_{out}(x)$ denotes the set of followees of x, and k_x is the degree of node x.

Content

As shown in Hannon et al. (2010), the content of social networks is a valuable factor for link prediction, as users are likely to follow others sharing the same information preferences (Romero & Kleinberg, 2010). Users' interests can be characterized by means of profiles based not only on the content of the published tweets, but also in the retweeted or favorite tweets. Whereas the first alternative indicates users' interests in terms of the information they create and publish, the last two alternatives indicate users' interests in terms of the information they consume, that is, the information they read and consider interesting. These profiles will be referred as *publishing profile* and *reading profile*, respectively.

The set of tweets t for a user u_j can be denoted as:

$$tweets(u_j) = \{t_i, \ldots, t_n\} \tag{3}$$

The *publishing profile* of a user is built by considering all user tweets under the assumption that users tend to tweet about things that are relevant to them. Formally, the profile of user u_j can be defined as:

$$pub-profile(u_j) = tweets(u_j) \tag{4}$$

Although the publishing profile can adequately capture user's interests regarding the information they want to share with their followers, it cannot capture their interests regarding the information they consume, that is, the information that their followees publish and that they deemed interesting. In Twitter, if users tend to consume tweets regarding a certain topic, it is likely that they would follow users tweeting on those topics due to homophily (McPherson et al., 2001). However, as followees might tweet on several topics, which might not all be of interest to users, it is necessary to identify the specific tweets in which each user is interested. Twitter provides two mechanisms for expressing interest in tweets posted by others: tweets can be marked as favorites (analogous to bookmarking a website) or retweeted (reposted or forwarded messages). Retweets are considered the best mechanism to show interest in other users' tweets, as it makes tweets visible to the followers of the user who made the retweet. Hence, retweets convey the information users are interested in consuming. Once user profiles are built, the similarity between two profiles can be computed using cosine similarity. For example, content-based followee recommendations should match the *reading profile* of users with the *publishing profile* of their potential followees.

DOI: 10.1002/asi

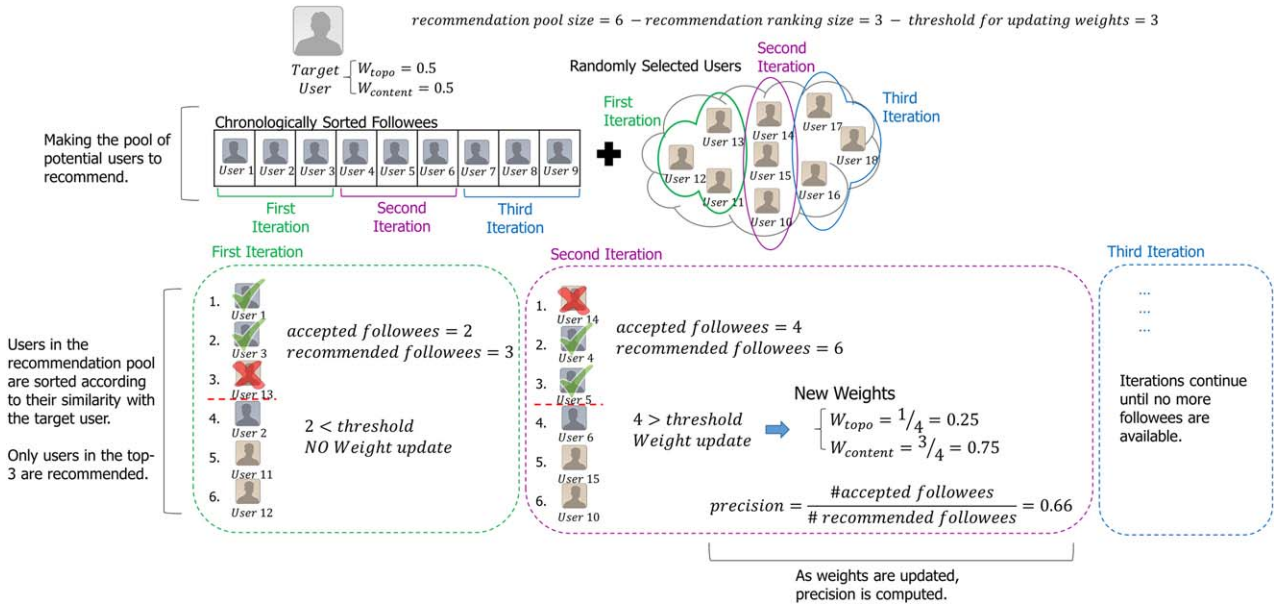

FIG. 2. Methodology evaluation. [Color figure can be viewed at wileyonlinelibrary.com]

The *reading profile* of a user u_j can be built by considering any of the following alternatives:

$$read-profile_{RT}(u_j) = tweets_{RT}(u_k)\ \forall\ k \in followees(u_j) \quad (5)$$

$$read-profile_{Fav}(u_j) = tweets_{Fav}(u_k)\ \forall\ k \in followees(u_j) \quad (6)$$

In turn, they can be combined to include all the favorite and retweeted tweets of user u_j that were posted by any of their followees:

$$read-profile_{RT-Fav}\ (u_j) = tweets_{RT}(u_k) \cup tweets_{Fav}(u_k)\ \forall\ k \in followees(u_j) \quad (7)$$

In all cases, user profiles comprise all terms appearing in each of the considered tweets following the traditional vector space model (Salton & McGill, 1983), in which each vector dimension corresponds to an individual term weighted by its frequency of appearance. As the method is intended to recommend followees in a real-time setting, more advanced weighting schemes requiring knowledge of the full tweet collection (e.g., TF-IDF) are not applicable. This is mainly due to two reasons. First, in a real-time setting, posts would be constantly arriving, and thus there is no fixed document corpus on which to base the statistics computation. Second, if the data collection changes every time a new post arrives, statistics would need to be periodically updated, resulting in a very inefficient approach. Consequently, even when some information regarding the overall relevance of terms could be lost, in highly dynamic environments it is preferable to use simpler but efficient weighting schemes.

Experimental Settings

The iterative methodology for evaluating the performance of the proposed method is exemplified in Figure 2. F2
For each target user, her/his actual followees and an equal number of randomly selected non-followed users were added to a pool of potential followees to recommend. To simulate the behavior and preferences of target users over time, actual followees were added to the pool in the same order in which the user started following them.

Initially, the evaluation assumes that at first users give equal importance to the factors by assigning equal weights to them (in the example topology and content received, 0.5). Next, the recommendation algorithm computes the similarities between the target user and each potential followee in the pool as previously described (as shown in the different iterations). The quality of recommendations was evaluated by selecting the top-N recommended users and computing precision. Precision can be defined as the percentage of relevant recommendations (i.e., the number of actual followees that was discovered by the algorithm) regarding the total number of recommendations. As there is no explicit feedback from target users available, the quality evaluation assumes that users that were not originally followed are uninteresting to the user. This assumption might not be completely accurate, as recommended users might not be in the followee list simply because the target user was unaware of their existence. Hence, the number of recommended users that were not on the original followee list, that is, false positives, might be overestimated, leading to an underestimated precision. The overall precision was computed as the aggregation of the scores of the multiple target users in each iteration.

After the number of discovered followees is known, it is determined whether it corresponds to update the factor

DOI: 10.1002/asi

weights. Although ideally weights should be updated with every newly accepted recommendation, in the performed evaluation not every accepted followee triggers a weight update. Instead, a minimum number of newly accepted followees is required for updating the weights (3 in the example). As in the first exemplified iteration, only two relevant users were discovered, weights were not updated. Conversely, by the second iteration four relevant followees were discovered. Hence, as the number of relevant followees found (four) was higher than the threshold (three), weights were updated. This restriction is imposed both to collect sufficient information regarding user preferences (a unique accepted followee might not be sufficient for evidencing changes in users' interests and preferences), and avoid deteriorating performance with frequent updates. In all cases, the reported precisions represent those obtained immediately after the weights were updated. Finally, as the method intends to analyze the evolution of user preferences, the process of creating the pool of users to recommend, computing similarity, and selecting the top-ranked users is iteratively repeated by selecting new pools until no more actual followees are available for the target user.

In the reported evaluation, the size of the pool of potential followees to build in each iteration was set to 20 (10 were actual followees, and the remaining 10 were random non-followed users). In each iteration, 10 followees were recommended. Factors' weights were updated every 10 accepted recommendations. Additionally, followee similarities towards the target user were standardized to make all similarity scores comparable. Although all similarities ranged between $[0, 1]$, the scores of each factor could be concentrated among different subranges, hindering their accurate comparison.

Typically, the selection of similarity thresholds to determine whether a potential item or followee is interesting depends on a specialist who fixes a value, or in a trial/error process, in which multiple values are tested until the result is satisfactory (da Silva, Stasiu, Moreira Orengo, & Heuser, 2007). If the chosen threshold is high, there is a risk of not finding interesting items. Conversely, a low threshold will find many irrelevant items. The difficulty of the problem increases when the diverse similarity metrics have a different score distribution. Hence, the selection of similarity thresholds should be guided by the characteristics of the social network under analysis. For example, in an information-centric network (a social network that is guided by the desire of consuming information as Twitter is) the content similarity between users will be higher than the topological similarity. Conversely, on a friendship-based social network (e.g., Facebook), relationships will mostly answer to topological factors. These characteristics will condition the distribution of user similarities, which, in turn, indicate the range of followee similarities for each user. Consequently, similarity thresholds could be defined based on the statistical distribution of similarities in the data set. The rationale is that as users tend to relate with people in a certain range of similarity, other users scoring in the same range should be preferred over users with dissimilar scores.

For the evaluated data set, the distribution of the content-based similarity showed that the median and mean values were similar and that the first and third quartiles were at the same distance from the median as the standard deviation was from the mean, implying that similarities were homogeneously distributed over the full range of possible values. Hence, the chosen value was set to 0.7, which represents the third quartile. This means that for a potential followee to be associated with a factor, his/her similarity should be higher than that of the 75% of the followee distribution. Following the same idea, the minimum similarity threshold for the topology factor was set to 0.2. In both cases, outliers were removed from the analysis.

As the novelty factor is introduced to balance between always recommending the same type of items and allowing the discovery of new and interests items, the novelty's threshold is defined considering the outliers of the similarity distribution. Outliers can be defined as observations that lie at an abnormal distance from other values in the sample distribution, that is, that are dissimilar to the majority of the other data points. In this work, outliers were detected using the Tukey's method (Tukey, 1977), setting $k = 1.5$ as suggested by the author. One of the advantages of Tukey's method is that it is applicable to both normal and skewed data, since it does not make any distributional assumptions, and it is independent from the mean and standard deviation. The novelty threshold was computed considering the novelty of those followees that could be considered outliers. As the novelty score is unique for each recommendation factor considered, the resulting threshold is the average of the proportions obtained for each similarity distribution. The resulting novelty threshold was 0.05.

The used data set was obtained by crawling a set of $3,453$ target users through the TwitterAPI[1]. Approximately a half of them were originally included in (De Choudhury et al. 2010), comprising politicians, musicians, environmentalists, and other users who frequently tweet about a diverse range of topics. The remaining target users were selected from their followee set to increase user diversity, as the selection was made regardless of the popularity or posting activity of users. To guarantee both meaningful content-based profiles and extensive topological networks, several restrictions were imposed on target users to be selected. First, users must have had more than 10 followees. Second, users must have had more than 10 published tweets. Third, the user account must have been listed as English, and the first set of retrieved tweets must also have been written in English, as detected by TextCat[2]. For all target users, all tweets, followees, followers, favorite tweets, and user account information were retrieved. The same data were retrieved for each of their followees. Table 1 summarizes T1 the data statistics. In the case of the average values, the

[1]https://api.twitter.com

[2]http://odur.let.rug.nl/vannoord/TextCat/

DOI: 10.1002/asi

TABLE 1. Data collection statistics.

Total number of users	3,449
Total number of tweets	3,227,782
Average number of tweets per user	935.86 (±1,200.21)
Total number of followee relations	1,650,208
Average number of followee relations per user	478.46 (±2,440.53)
Total number of follower relations	23,626,904
Average number of follower relations per user	6,850.36 (±187,662.64)

standard deviation is shown in parentheses. As shown, the number of tweets, followees, and followers are distributed over a great range of values. Interestingly, 25% of the seed users have a number of followees lower than 36, and 50% of the users lower than 125. Additionally, the mode of the followee distribution was 12. This implies that the dataset covers a wide spectrum of users, ranging from users only seeking information (i.e., users with a low number followees) to celebrities (i.e., users with a high number of followees).

Comparison With Other Approaches

Based on the two considered recommendation factors, the proposed method (named *adaptive*) for personalizing followee recommendations was compared to two types of approaches. First, state-of-the-art recommendation approaches that are not based on adapting to user interests. Second, a set of techniques that can be used for combining several ranking lists into a generated consensus ranking (Dwork, Kumar, Naor, & Sivakumar, 2001). Even though these techniques do not exactly personalize the importance of each factor to user interests, they represent an alternative to the linear combination of factors. Additionally, the proposed method was compared to a version without the novelty factor (named *adaptive-no-novelty*).

State-of-the-Art Approaches

The presented approach was compared to several state-of-the-art techniques. Particularly, the experimental evaluation considered the alternatives proposed in (Hannon et al. 2010), which continues to be widely used as baselines for friend recommendation in social networks (Kumara & Sundarrajb, 2016; Rodríguez, Torres, & Garza, 2016). In all cases, factor weights were neither personalized according to user interests nor evolved over time. Particularly, the following alternatives were considered for comparison:

- Content-based similarity analysis based on the publishing profile of both target users and potential followees (termed $pure\text{-}content_{PUBLISHING}$).
- Content-based similarity analysis based on the reading profile of target users and the publishing profile of potential followees (termed $pure\text{-}content_{RT}$, $pure\text{-}content_{Favs}$, and $pure\text{-}content_{RT\text{-}Favs}$).
- Topology-based similarity based on the coincidences between the set of followees of the target user and that of the potential followees (termed *pure-topology*).
- A hybrid strategy in which topology and content-based factors are combined in equal proportions (termed *half-topology-content*, where content can represent any of the previously described profiles).

Rank Aggregation Techniques

Rank aggregation refers to the problem of combining ranking results from different sources to obtain a unique ranking. In this study, each source corresponds to a ranking generated by a recommendation factor. The goal of rank aggregation is to find the aggregated ranking that minimizes the distance to each of the ranked input lists (Sculley, 2007). These techniques focus on the intrinsic characteristics of the similarity rankings obtained for each particular user. Thus, even though these techniques assign the same relative importance to all input rankings, the results are conditioned by the similarity distributions in each ranking. This implies that the absolute importance of each recommendation factor varies from user to user, even when their explicit characteristics are not considered. In this work, three categories of rank aggregation techniques (Schalekamp & van Zuylen, 2009) were considered: positional, comparison sort, and hybrid techniques.

Positional Techniques. Positional algorithms aim at finding a permutation in which each item's position is close or similar to the average position of the item in the input lists. Particularly, four techniques were evaluated (Sculley, 2007): *Borda Count*, *Footrule Spearman*, *Median Rank Aggregation*, and *Pick-a-Perm*. *Borda Count* relies on the absolute position of items in the ranked lists, rather than on their relative rankings. For each item i, its score is computed as $B(i)=\sum_{k=1}^{r} n-\pi_k(i)$, where n represents the number of elements in each ranking, r is the number of input rankings, and $\pi_k(i)$ is the position of item i in rank π_k. Then, the aggregated ranking is built by sorting the items in decreasing order according to their score, aiming at minimizing the sum of the distances from the position of elements regarding their mean position. *Footrule Spearman* aims at finding a ranking that minimizes the average Footrule distance ($\mathcal{F}(\pi,\pi')=|\pi(i)-\pi'(i)|$, where π and π' are rankings) between the aggregated rank and each of the input rankings. It can be cast as a bipartite matching problem, which can be solved by the Ford–Fulkerson algorithm. *Median Rank Aggregation* combines the set of input rankings by considering the median rank of each item. To compute the ranking on items drew from the set of ranked lists $\pi \in R$, first initialize the scores $M(i)=0$ for each item i. Then, starting at $n = 1$, update the scores $M(i)=M(i)+c(i,n)$, where $c(i,n)$ computes the number of lists in R for which the ranking of i is equal to n. The first item i with a score higher than a predefined threshold gets rank 1, the second item gets rank 2, and so forth until all items are ranked. Finally, *Pick-a-Perm* returns an input permutation at random. Note that these

DOI: 10.1002/asi

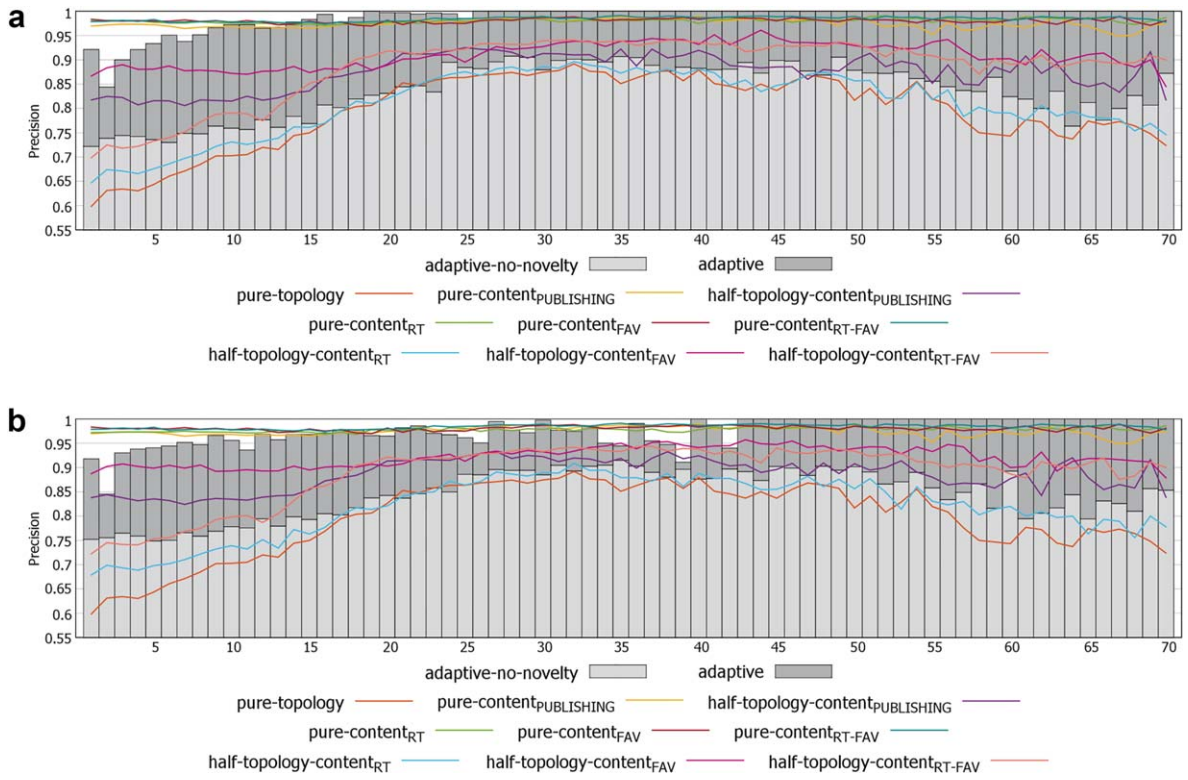

FIG. 3. Comparison of precision results regarding static weighting techniques. [Color figure can be viewed at wileyonlinelibrary.com]

techniques are similar to one of the alternatives proposed in Hannon et al. (2010).

Comparison Sort Techniques. These techniques use a comparison relation to sort the elements, which is not necessarily transitive, implying that different sort algorithms can generate different aggregated rankings (Schalekamp & van Zuylen, 2009). Items' relations can be defined as $i \preccurlyeq j$, where i is ranked above j for the majority of the input rankings. Also, $j \prec i$ if $i \preccurlyeq j$ and $j \preccurlyeq i$. Particularly, three sorting algorithms were considered: *Quick Sort*, *Merge Sort*, and *Insertion Sort*. *Quick Sort* recursively sorts the items by choosing an item i as pivot, and ranking item j higher than i if $j \prec i$, or lower than i if $i \prec j$. *Merge Sort* recursively sorts the items by dividing them into two equal parts, then recursively sorting each part, which are finally merged into the final ranking. Lastly, *Insertion Sort* starts with an empty list, to which items are added one by one. When adding item i to the list, it is placed in the highest position so that $i \prec j$ for every item j in a lower position than i.

Hybrid Techniques. These techniques combine both positional and comparison based algorithms. Two techniques were included: *Copeland's* method and *Markov Chains*. *Copeland's* method sorts items based on the number of items they would beat in a pairwise majority contest (Schalekamp & van Zuylen, 2009). The majority tournament can be defined as a directed graph comprising a node for each element and an edge between nodes i and j if $i \prec j$. Then, to obtain the aggregated ranking, items are sorted according to their in-degree. The *Markov Chain* method represents the items in the input lists as nodes in a graph, with transitions probabilities between nodes defined according to the relative rankings of items in the lists (Sculley, 2007). Then, the aggregated ranking is found by sorting in decreasing order the nodes according to their probability of being visited in a random walk on the graph, that is, the stationary distribution of the Markov chain. The transition matrix was defined as follows (Dwork et al., 2001): if the current state is node i, then the next state is chosen by first selecting a ranking uniformly from all rankings, and then choosing a node j uniformly from the set of nodes that it is ranked better than i.

Findings

This section presents the results obtained when assessing the effectiveness of the proposed technique for personalizing the importance of different followee recommendation factors according to each user's behavior and interests.

Comparison With State-of-the-Art Approaches

Figure 3 shows the evolution of the average recommendation precision for all the weight updates performed for each of the evaluated alternatives. As regards the state-of-the-art approaches, the best results were achieved when considering any *pure-content* alternative, which reached precisions higher than 0.95, with differences up to a 58% regarding the worst-performing approach, that is, *pure-*

DOI: 10.1002/asi

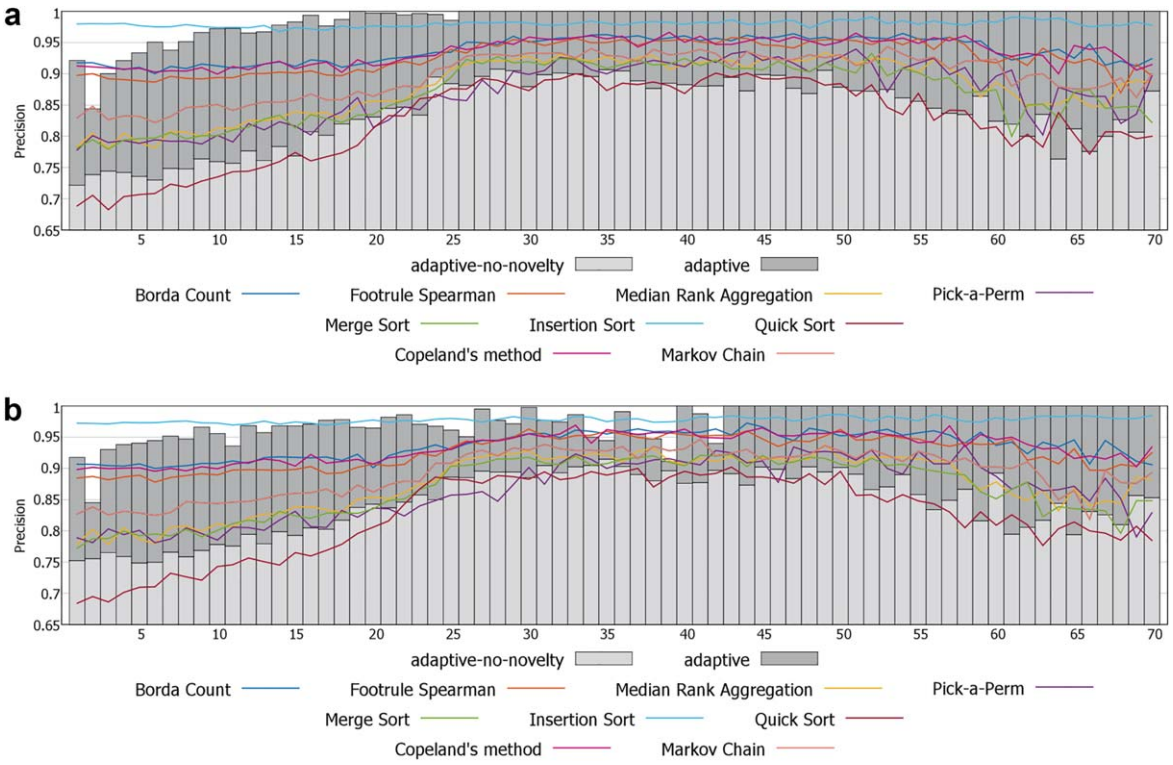

FIG. 4. Comparison of precision results with rank aggregation techniques. [Color figure can be viewed at wileyonlinelibrary.com]

topology. These results indicated that the majority of followee relations in this dataset could be discovered when only considering the content factor. However, there were also followees who could not be found with a pure content oriented strategy. Topology-based results further highlighted the fact that the majority of the followee relations are content driven.

Regarding the variations of the proposed method, the *adaptive-no-novelty* achieved the worst results. Although those results slightly outperformed the results achieved for the *half-topology-content* alternatives, they are lower than the best *pure-content* results. The best improvements regarding the *half-topology-content* techniques were obtained for the first weight updates, when the method starts to learn user preferences. Hence, it can be stated that although the combination of weights is adapted to each user, it is not sufficient for further improving results. Moreover, it can be inferred that although users have a particular preference for a certain type of followees, they also select some followees who do not exactly match such preferences. Consequently, the search and ranking of users should not be only guided by the similarity ranking, but also by the novelty component. Note that when adding the novelty component, that is, the *adaptive* alternative, the results are superior to the others. As the figures show, the adaptive alternative was able to achieve an optimal precision after 26 and 43 weight updates when considering *Common Followees* and *Sørensen*, respectively. These results evidenced the importance of not only recommending similar followees, but also recommending novel or diverse followees. Finally, the figures also show the stability of precision once the preferences of users were learned and adapted.

Interestingly, as regards the *Sørensen* metric, between the 25*th* and 35*th* weight updates, the precision results decreased for all the evaluated alternatives. This could be associated with an unexpected change in user behavior or preferences. Although the change affected the precision of state-of-the-art approaches, as the weights are never changed, the effect was only temporary. On the contrary, a sudden change in user interests could have a profound impact on the adaptive alternative, as it depends on the previously correct predictions. Remarkably, although the *adaptive* precision decreased for one weight update, it was able to learn the new target user preferences, and, in the next weight update, it outperformed all the other alternatives. These results further highlighted the importance of adapting the factor weights to user preferences over time.

Comparison With Rank Aggregation Techniques

Figure 4 shows the comparison between the *adaptive-no-novelty* and *adaptive* alternatives, and the chosen rank aggregation techniques. As the figure shows, rank aggregation techniques behaved similarly for both topology metrics. The best results were obtained for *Insertion Sort*, which achieved similar results to the *pure-content* alternatives. *Insertion sort* achieved a precision higher than 0.95, with differences up to 30% regarding the worst rank-aggregation technique, that is, *Quick Sort*. Note that *Pick-a-Perm*, which randomly chooses a ranking, did not achieve the worst results. F4

DOI: 10.1002/asi

All rank aggregation techniques achieved better results than the static weighted alternatives purely based on topology and those mixing topology and content in equal proportions. These results might suggest that rank aggregation techniques are more suitable for integrating multiple sources of information than static weighting schemes. The results showed that there was no clear superiority of the techniques belonging to a particular category, as most of them achieved similar results (e.g., *Borda Count*, *Footrule Spearman*, and *Copeland's method*, or *Median Rank Aggregation* and *Merge Sort)*.

In summary, although rank aggregation techniques achieved good results, they were unable to accurately find all interesting users, and thus, of achieving perfect precision. Conversely, by explicitly considering the past interests of users, our method was able to achieve optimal results.

Summary of Results

Regarding the ability of the presented method for predicting factor's weights, Figure 5 shows the differences between the predicted weights for each of the four combinations of factors, and the weights computed considering the complete set of followees for each user, which represent their real preferences. The best predictions were achieved when considering *Common Followees*. The average difference for the content-based factor was 0.062 with a standard deviation of 0.042. When considering the topological factor, the average difference was slightly higher, reaching 0.076 with a standard deviation of 0.05. Contrarily, the worst predictions were achieved when considering *Sørensen*. In this case, the average difference was 0.064 for the content factor with a standard deviation of 0.041. For the topological factor, the average difference was slightly higher, reaching 0.081 with a standard deviation of 0.041. However, the difference between both topology metrics is lower than 6.51%, which could indicate that both metrics can accurately represent the topological interests of users. Furthermore, the differences are below 0.1 for 76% of the target users, which emphasizes the usefulness of the proposed method not only for adequately capturing users' interests, but also for adapting to the changes in user preferences over time.

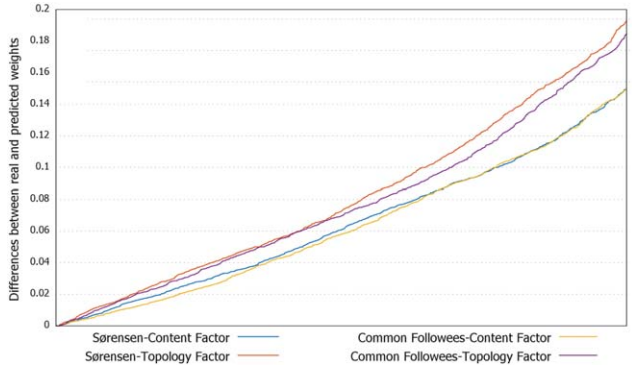

FIG. 5. Differences among the final and the real weights. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. Summary of precision improvements (%).

	Adaptive followee recommendation		
	Minimum	Maximum	Average
pure-topology	4.48	60	25.78
pure-content	−13.96	11.11	3.35
half-topology-content	9.10	53.99	26.38
best rank aggregation - Positional	4.48	30.43	14.83
best rank aggregation - Sorting	−13.90	11.11	2.88
best rank aggregation - Hybrid	−0.46	25.00	12.54
adaptive-no-novelty	9.38	49.42	24.82

Table 2 summarizes the precision improvements of the proposed method over the best-performing static weighted alternatives, the best of each type of rank aggregation techniques, and the proposed method without considering the novelty component for the *Common Followees* metric. Baseline precisions were outperformed in most cases, excepting when considering *pure-content* and the best hybrid rank aggregation alternatives. The maximum improvement was achieved regarding *pure-topology* (60%). Although the improvements regarding *pure-content* might seem low, it is necessary to consider that the *pure-content* alternative started with a very high precision, but the adaptive alternative was still able to improve it. In summary, precision of recommendations can be improved when considering an adaptive method for defining the weights of recommendation factors. The results emphasize the importance of adapting the factors' relevance or weights to changes in user preferences over time, and considering diversity in followee recommendations.

Finally, the statistical significance of the results was tested to determine whether the improvements reported by the *adaptive* technique were significant and not due to random or sampling error. As normality tests failed, statistical significance was evaluated using nonparametric tests. Two hypotheses were defined: the null (personalizing the weights of the recommendation factors according to user interests had no significant impact on its precision) and the alternative (personalization had a significant and nonincidental impact on precision) one. When performing the Mann–Whitney test and analyzing the results, the obtained significance levels allowed rejecting the null hypothesis for all cases, implying that personalizing the weights of recommendation factors had a significant and nonincidental effect on the results regarding the precision achieved when recommending followees considering either static weighting alternatives or rank aggregation techniques.

Conclusion

In social networks, the recommendation of potential followees to suggest arises as a crucial issue. Thus, the criteria used to guide the search and ranking of followees has to be

carefully evaluated. This work proposed a method for adapting the followee selection criteria to the decisions of each particular user regarding the characteristics of previously selected followees. The method has the ability of evolving such criteria over time according to changes in users' followee preference. Moreover, it considers not only the similarity but also the novelty or diversity of potential followees. Experimental evaluation showed that the proposed method helped to improve precision results regarding static weighting strategies and rank aggregation techniques. Hence, personalizing the importance of the followee selection criteria according to user behavior was shown to have a significant and positive effect on the quality of the performed recommendations. Furthermore, results highlighted the importance of adapting to changes in user preferences over time.

As regards future work, other recommendation factors such as personality, emotions, language, or geographical location could also be analyzed. Moreover, additional alternatives for computing the personalized weights of the different factors could be explored. Finally, strategies for more intelligently adapting to changes in user interests might be introduced. For example, instead of simply adapting to the most recent data by forgetting old data at a constant speed regardless of whether data are changing, strategies for detecting changes in data can be implemented to adapt more rapidly to them.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17, 734–749.

Agarwal, V., & Bharadwaj, K.K. (2013). A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Social Network Analysis and Mining, 3, 359–379.

Armentano, M., Godoy, D., & Amandi, A. (2011). A topology-based approach for followees recommendation in Twitter. In *Proceedings ITWP'11 at IJCAI* (pp. 22–29).

Brzozowski, M.J., & Romero, D.M. (2011). Who should I follow? Recommending people in directed social networks. In *Proceedings of the 5th ICWSM*, Barcelona, Spain, 2011.

Chen, H., Cui, X., & Jin, H. (2016). Top- followee recommendation over microblogging systems by exploiting diverse information sources. Future Generation Computer Systems, 55, 534–543.

da Silva, R., Stasiu, R., Moreira Orengo, V., & Heuser, C. (2007). Measuring quality of similarity functions in approximate data matching. Journal of Informetrics, 1, 35–46.

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K.S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th ICWSM*, 2010.

Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th* WWW (pp. 613–622). NY, USA.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46, 44.1–44:37.

Garcia, R., & Amatriain, X. (2010). Weighted content based methods for recommending connections in online social networks. In *Proceedings of the 2nd RSWeb* (pp. 68–71). Barcelona, Spain.

Gerani, S., Zhai, C., & Crestani, F. (2012). Score transformation in linear combination for multi-criteria relevance ranking (vol. 7224, pp. 256–267). In Advances in information retrieval. Berlin: Springer.

Golder, S., & Yardi, S. (2010). Structural predictors of tie formation in twitter: Transitivity and mutuality. In *SocialCom/PASSAT* (pp. 88–95). IEEE.

Hannon, J., Bennett, M., & Smyth, B. (2010). Recommending Twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the 4th ACM RecSys'10* (pp. 199–206).

Hurley, N., & Zhang, M. (2011). Novelty and diversity in top-n recommendation — analysis and evaluation. ACM Transactions on Internet Technology, 10, 14:1–14:30.

Kumara, S., & Sundarrajb, R. (2016). Social media user recommendation. In *Proceedings of the 8th STAIRS*, 284:197. IOS Press.

Liu, X., & Turtle, H. (2013). Real-time user interest modeling for real-time ranking. JASIST, 64, 1557–1576.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27, 415–444.

Rodríguez, F., Torres, L., & Garza, S. (2016). Followee recommendation in twitter using fuzzy link prediction. Expert Systems, 33, 349–361.

Romero, D.M., & Kleinberg, J.M. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proceedings of the 4th ICWSM*, DC, USA.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Schaal, M., O'donovan, J., & Smyth, B. (2012). An analysis of topical proximity in the Twitter social graph. In *Proceedings of the 4th SocInfo* (vol. 7710, pp. 232–245).

Schalekamp, F., & van Zuylen, A. (2009). Rank aggregation: Together we're strong. In ALENEX (pp. 38–51). SIAM.

Sculley, D. (2007). Rank aggregation for similar items. In *Proceedings of SDM*, Minnesota.

Tommasel, A., & Godoy, D. (2015). An adaptive technique for weighting multiple factors in followee recommendation algorithms. In *Proceedings of the CPCR+ITWP at IJCAI*, *CEUR Proceedings*.

Tommasel, A., Corbellini, A., Godoy, D., & Schiaffino, S. (2016). Personality-aware followee recommendation algorithms: An empirical analysis. Engineering Applications of Artificial Intelligence, 51, 24–36.

Tukey, J.W. (1977). Exploratory data analysis. Addison-Wesley series in behavioral science. Reading, MA: Addison-Wesley.

Vargas, S. & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM* RecSys (pp. 109–116). New York.

Vogt, C.C., & Cottrell, G.W. (1999). Fusion via a linear combination of scores. Information Retrieval, 1, 151–173.

Wu, S. (2012). Linear combination of component results in information retrieval. Data & Knowledge Engineering, 71, 114–126.

Yuan, G., Murukannaiah, P.K., Zhang, Z., & Singh, M.P. (2014). Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM RecSys*, CA, USA.