Assessing the Contribution of Twitters Textual Information to Graph-based Recommendation

Evgenia Wasserman Pritsker
Information Systems Dep.
University of Haifa
evgeniaw@is.haifa.ac.il

Tsvi Kuflik
Information Systems Dep.
University of Haifa
tsvikak@is.haifa.ac.il

Einat Minkov
Information Systems Dep.
University of Haifa
einatm@is.haifa.ac.il

ABSTRACT
Graph-based recommendation approaches can model associations between users and items alongside additional contextual information. Recent studies demonstrated that representing features extracted from social media (SM) auxiliary data, like friendships, jointly with traditional users/items ratings in the graph, contribute to recommendation accuracy. In this work, we take a step further and propose an extended graph representation that includes socio-demographic and personal traits extracted from the content posted by the user on SM. Empirical results demonstrate that processing unstructured textual information collected from Twitter and representing it in structured form in the graph improves recommendation performance, especially in cold start conditions.

Author Keywords
Graph-based Recommendation; Social Media; Twitter; PPR; Information Extraction

ACM Classification Keywords
H.4 INFORMATION SYSTEMS APPLICATIONS: Miscellaneous; I.2.7 Natural Language Processing: Text analysis

INTRODUCTION
Social media (SM) forms a rich source for personal information about individuals. Some of this information is provided explicitly by the users, and is available in structured form, for example job and education details in Google Plus. However, in many cases, these structured fields are sparsely populated, since users rarely fully complete their online profiles. Furthermore, some SM services, e.g. Twitter, do not support extended structured user profiles [19]. Still, additional personal attributes can be extracted implicitly from the user generated content. Indeed, multiple recent works demonstrated that socio-demographic information [28, 30, 36], personality traits [10], emotions [6, 23] and mental state [7] can be successfully extracted from the text posted on social media. Ideally, this rich information extracted from SM should be incorporated into personalized recommendation systems to improve their performance.

IUI 2017, March 13-16, 2017, Limassol, Cyprus
©2017 ACM. ISBN 978-1-4503-4348-0/17/03/$15.00
DOI: http://dx.doi.org/10.1145/3025171.3025218

Previous research efforts mostly considered personalization within the SM application, for example the recommendation of new friends, posts, URLs and pictures [12, 5] to users. Only few works exist in which personal information extracted from SM is used to recommend items outside the context of SM. We claim that recommender systems can substantially benefit from this freely available user information. Indeed, several works incorporated personality information as part of user modeling in books and movies recommendation, albeit without directly using SM [14, 9]. Here, we assess the contribution of modeling diverse information extracted from SM to personalized recommendation in the movies domain.

Since rich personal information can be extracted from SM, a challenge that has to be addressed is how to effectively represent this information within a recommender system. It has been previously argued that graph-based approaches are advantageous for contextual recommendation; it is straightforward to model in the graph background information as entities of different types alongside historical user and item interactions. Graph-based similarity measures, such as the random walk Personalized Pagerank (PPR) [13], can then be applied to model transitive associations between the users and the various entities that are represented in the graph [17, 33, 34, 27, 24].

Recently, it has been shown that the modeling graph features extracted from SM structured auxiliary data, like tags and friendship, results in improvement of a graph-based recommender system [34, 33]. Encouraged by these results, we wish to model in the graph information extracted implicitly from textual content on SM, specifically Twitter. We conduct comparative recommendation experiments using a dataset of movie ratings, in which Twitter user IDs are provided. Our goal is to address the main following questions:

- To what extent does personal data inferred from user-generated content published on Twitter improve graph-based recommendation? In particular, we focus on evaluating a cold-start scenario, in which no or little rating history is available for a new user.
- How effective is recommendation using the extended graph compared with popular recommendation techniques?

Our results show that leveraging the text-based user information can significantly improve recommendation performance. The advantage of modeling SM-derived personal information is especially relevant for new users with little rating history.

RELATED WORK

Recommendation using Social Media Information

we are interested in evaluating the contribution of SM information to the recommendation process. Previously, Abel *et al.* [1] introduced a news recommender system in Twitter using hash-tag based, entity-based and topic-based user profiles. Similarly to their work, in our study we extract various types of user-related information from their Twitter posts, but our application domain is outside of Twitter. Some recent works have already exploited SM information for recommendation of items outside SM. [37] used various Facebook features including "liked" pages, n-grams derived from Facebook page names, and demographic data, to overcome the cold start problem in eBay purchases recommendation. They demonstrated significant correlation between SM profiles and online purchases. [20] extracted information from Twitter for mobile applications recommendation. Starting from the accounts of applications followers, they iteratively expanded the networks to the followers of these accounts, and created groups of users with similar interests using topic modelling. Their work also showed significant improvements in performance due to resolution of the cold-start problem. [31] leveraged SM information for movies recommendation. They investigated if information about users' favorite items on Facebook can replace explicit item ratings, and achieved positive results. Compared with these works, our innovation is two-fold: extraction of users' attributes from unstructured SM content and the modeling such diverse information in a graph-based framework.

Graph-based Recommender Systems

Multiple studies have previously used related graph-based approaches for recommendation, starting from early works using homogenous graphs (e.g.,[11]), continuing with bipartite user-item graphs (e.g., [2]) and up to heterogeneous graph representations that consist of multiple node and edge types, which is applied also in our study. For instance, Bogers [4] proposed a graph consisting of multiple nodes types: users, movies, tags, genres and actors, extracted from movie database websites. In this study we also construct a heterogeneous graph in the movies domain, but our focus is on representing user properties extracted from SM content. Konstas *et al.* [16] incorporated friendship and social tagging information from a music social network in a graph that included users, music tracks and tags as nodes, and relationships between users, users and tracks as graph links. They applied Personalized PageRank (PPR) for recommending music tracks to users, outperforming collaborative filtering methods. There is a number of additional works comparing graph-based PPR with classical recommendation methods, reporting superior results using PPR [22, 24]. Finally, while we do not follow this practice here, it is possible to automatically generate graph-based features as a way to enrich the available ones and incorporate them in a common learning process, so as to improve recommendation accuracy [32, 33].

Addressing Cold-Start Problem with Complementary Data

One of the key issues limiting the recommendation process is the cold start problem, when new users have not rated enough items. In order to overcome cold start conditions, researchers have incorporated user personality characteristics into a collaborative filtering (CF) movie recommendation framework, reaching superior performance over traditional CF [14]. The same trend was demonstrated in similar setting for image recommendation [35]. [9] experimented with publicly available database of the myPersonality project[1] to evaluate the usefulness of personality information for recommendation of books, movies and music. They concluded that the recommendation algorithms effectively exploited this information in extreme new user situation. Those works assumed that structured auxiliary data was available. In this work, we obtain user traits automatically and implicitly from SM texts.

Users	Items	Ratings	Sparsity	Avg. ratings/user $\pm\sigma$	Avg. ratings/item $\pm\sigma$
$2.7K$	$8.5K$	$73K$	99.86%	27.1 ± 9.9	8.7 ± 31.8

Table 1. Description of the dataset.

TOOLS AND METHODS

Dataset

An essential element of our study is a dataset which integrates users ranking information with pointers to their SM accounts. The MovieTweets dataset [8] ties a user's Twitter ID with a list of movies rated by that user. We use a snapshot of this database from October 2015. In order to support learning and evaluation, we only consider users who rated between $15 - 50$ movies. The upper bound was set in order to reduce the number of users who possibly rated a large number of movies without watching them; the lower bound was set so as to support solid experimental evaluation. In order to enrich the dataset with personal information, Twitter was crawled using Twitter API. For every user, we retrieved the textual content of her tweets (up to 3,200 tweets per user, according to the restrictions of Twitter's API). The users, whose language determined by Twitter API was different from English, were excluded from the dataset. Retweets of a user were discarded, since they were not authored by that user. The dataset obtained in this fashion contains 2,695 users, their collection of tweets, structured profile information and the list of the movies rated by the users on a scale from 1 to 10. Table 1 details some statistics about the dataset. The dataset is freely available to the research community upon request.

Data Extraction from Unstructured Text

Due to their limited length, tweets are not to be processed independently. Instead, we concatenate the tweets authored by a user, and apply text processing techniques to map the resulting aggregated document into several sets of personal dimensions. First, we use the Linguistic Inquiry and Word Count (LIWC) software [26], designed by psycho-linguistics to reveal personal traits based on text usage. LIWC includes roughly 90 personal dimensions in total that represent beliefs, fears, thinking patterns, social relationships, and personalities. The association of a given text with each LIWC dimension is based on pre-specified word dictionaries. The

[1] http://mypersonality.org

program practically reads a given text and reports the proportion of words that pertain to the various dimensions. Another approach which we adopt in our work is the association of the user-generated text with a set of automatically detected topics. Concretely, we consider the topics generated from millions of Facebook messages using the Latent Dirichlet Allocation model [3], which have been shown to also reflect the language of Twitter. They provided the research community with lists of words that constitute as many as 2,000 word clusters, namely topics. The text authored by an individual is mapped to these topics. Finally, a regression model for predicting a user's age and gender has been developed using word counts as features [30]. We use this model to extract this demographic information about the user from his tweets texts. The described tools have been successfully applied to analyze Twitter data in a previous work [29].

Personalized PageRank Algorithm (PPR)

We experiment with the graph-based PPR [13] for generating recommendations. In general, PPR is used to assess graph nodes relatedness with respect to pre-specified nodes of interest. It applies a random walk procedure, in which some probability mass is propagated to neighboring nodes over outgoing edges at each step of the walk, while redirecting a fixed ratio of the probability mass at each node to the initial node set. As a result of this process, graph nodes that are connected to the initial query distribution over multiple and shorter paths are assigned higher weights, indicating on higher relevancy. In our setting, we apply PPR to compute node relevancy with respect to a node which represents a target user.

GRAPH-BASED DATA REPRESENTATION

The proposed graph representation consists of a set of typed nodes and labeled and directed edges. For every edge in the graph, there exists an edge going in the other direction, denoting an inverse relation. This implies that the graph is cyclic and highly connected. We distinguish between the following entity classes, representing them as distinct node types:

- *Users (U)*. These nodes represent individual users that are included in our dataset.
- *Movies (M)*. These nodes represent the movies for which there exist at least one rating in the dataset.
- *LIWC variables (L)*. A group of nodes that represent each of the LIWC variables as discussed above, including 4 summary language variables (analytical thinking, clout, authenticity, and emotional tone), 3 general descriptor categories, 21 standard linguistic dimensions, 41 word categories tapping psychological constructs, 6 personal concern categories, 5 informal language markers and 12 punctuation categories.
- *Topics (T)*. 2000 distinct topics are modeled as defined in the work of [30].
- *Gender (G)*. A couple of nodes which represent the user's inferred gender: male or female.
- *Age (A)*. We discretized the predicted user's age into the following age ranges: Under 25, 25-34, 35+, in accordance with the main age groups identified among the dataset's users, and following *Movielens* [18] commonly used age ranges.

Edge type	Source node type	Target node type
watch (1:*)	*user*	*movie*
markLIWC (1:*)		*LIWCvar*
useTopic (1:*)		*topic*
fromGender (1:1)		*gender*
fromAge (1:1)		*age*

Table 2. Relation types in the movies recommendation graph. (1:*) and (1:1) denote one-to-many and functional relationships, respectively.

The graph edge types are detailed in Table 2, and illustrated schematically in Figure 1. The *watch* labeled edges connect a *user* to each *movie* ranked by her. The rank score is ignored in the current study (87% of movies are ranked positively in the given dataset). A user is linked to a certain *topic* node if the association, calculated using method provided in [30], between the user and topic is significant, i.e., if it exceeds $Avg \pm \sigma$, where Avg is the average user score for a given topic *T*, and σ is the respective standard deviation. Users are selectively linked to LIWC dimensions in the same fashion. Overall, the resulting experimental graph consists of as 10K nodes and as 100K edges.

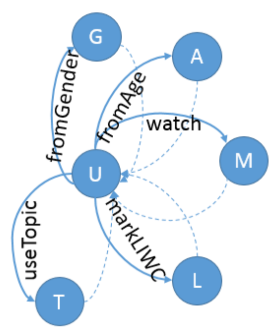

Figure 1. A schematic view of the relation types in the movies recommendation graph.

EXPERIMENTAL SETUP

We perform a set of link prediction experiments. For every user *u*, who rated a set of movies I_u, we select a subset of these movies $I_q \subset I_u$, assuming that only these ratings are known to us. The links between the graph nodes that represent *u* and each of the remaining movies, $m \in I_u \setminus I_q$, are then removed from the graph. The relevancy of each of the graph nodes is then assessed using the PPR measure with respect to the source node *u*. The result of this process is a ranked list, which includes *movies* ranked by their PPR score. Since u is directly connected with the movies I_q as well as to nodes that represent personal information about u, these nodes may be viewed as the user profile in this setting. Prediction performance is evaluated using mean average precision (MAP), considering all of the movies $m \in I_u \setminus I_q$ as correct answers. While this evaluation setting is imperfect, mainly, as other highly ranked items may be of interest to the user, it allows a fair comparison between alternative ranking methods.

We apply the PPR using a dumping factor 0.85. In addition, several popular recommendation techniques were implemented, including User-to-User collaborative filtering (u2u-knn), Item-to-Item collaborative filtering (i2i-knn) and Weighted Matrix Factorization method (WRMF) [15, 25]. We applied these methods as implemented in the MyMediaLite Recommender System Library[2], using default parameters. Using these algorithms, the user profile simply consists of the query movie set I_u, as it is non-trivial to incorporate the diverse personal information extracted from SM into the underlying matrix representation. We also report recommendation results using a non-personalized random baseline, in which the candidate movies are ranked randomly, and the non-personalized yet highly informative baseline of popularity-based ranking, having the candidate movies ranked according to the number of users who watched each movie.

Num. of Known Ratings	0	1	2	3	4	5
Graph-based Recommendation						
G:UM	-	0.101	0.158	0.217	0.281	0.350
G:UMGA	0.101	0.116	0.170	0.226	0.288	0.356
G:UML	0.101	0.141	0.189	0.241	0.300	0.366
G:UMT	**0.110**	**0.156**	**0.204**	0.254	**0.311**	**0.375**
G:All	**0.110**	**0.156**	**0.204**	**0.255**	**0.311**	**0.375**
Baseline Algorithms						
u2u-knn	-	**0.156**	0.168	0.171	0.172	0.171
i2i-knn	-	0.016	0.028	0.039	0.052	0.062
WRMF	-	0.134	0.147	0.157	0.164	0.162
Random	0.003					
Popularity	0.146					

Table 3. Movies recommendation (MAP) in cold-start scenario.

EXPERIMENTAL RESULTS AND DISCUSSION

Our assumption is that personal user information extracted from SM forms useful background knowledge which can come into play when the rating information is limited. We therefore focus on simulating movie recommendation in 'cold start' conditions. Specifically, the number of movies for which rankings are known for each user are restricted to the range $|I_q| = \{1, 5\}$. (We keep the set of 'test' movies fixed, and gradually increase I_q.) We report our results using multiple variants of graph-based recommendation, considering different subgraphs that include various combinations of node types; e.g., *G:UM* uses a graph that contains only *user* and *movie* nodes, and no personal information, the graph variant *G:UMT* incorporates also *topic* nodes, and so forth.

We now turn to evaluate our main research question: does personal information automatically drawn from content posted on SM improve prediction performance? As can be seen from our results in Table 3, the answer to this question is positive. Each of the graph variants improves over the variant *G:UM*, which models pure user-item ranking information. Indeed, the modeling of personal information is more pronounced in cold start conditions. When there are no ratings, graph-based recommendation that uses only background information can provide a recommendation, where all other approaches fail. In the case that a single rating is available for the evaluated user, the improvement due to the modeling of SM personal information reaches 35%. Impressively, even in this challenging scenario, the SM-augmented graph outperforms the informative popularity baseline. Furthermore, graph-based recommendation using the extended graphs consistently outperforms the collaborative filtering algorithms. We believe that these algorithms (mainly *i2i-knn*) are affected by the sparsity of the ratings matrix. It has been previously shown that graph-based recommendation is advantageous in high sparsity conditions [22]. Finally, we note that the association of a user with topics leads to larger performance improvement, overtaking the modeled demographic dimensions and LIWC variables. Somewhat surprisingly, extending *G:UMT* graph with either LIWC or demographic information did not lead to additional performance gains. This suggests that the information captured by the demographic and LIWC variables correlates with the large set of fine-grained topics that we used in our experiments. Another possible explanation is that the number of topics nodes significantly exceeds the number of other information nodes types and this may affect the probability flow in the graph. The issue is currently under investigation as part of continuing research.

[2] http://www.mymedialite.net/

CONCLUSION AND FUTURE WORK

This study integrates two promising research directions, namely graph-based recommendation using background knowledge, and integrating personal information extracted from unstructured content on SM into the recommendation process. Empirical evaluation using diverse personal dimensions extracted from tweets demonstrated consistent improvements in recommendation performance when SM information was incorporated, beating popular recommendation methods. The improvement gains due to the modeling of SM data were more pronounced when little was known about the user. Encouraged by these promising results, we intend to pursue the following research directions:

- The graph edges may be weighted according to their importance. One may assign each edge a parametric weight according to its type, having the weight tuned either manually or using learning [21]. Likewise, the rating values may be reflected by the edges weights.

- In addition to gender and age, it is possible to extract from SM information about one's relationship status, IQ, number of children, religion, race, income, education, etc. As relevant extraction tools continue to develop rapidly, they can be readily incorporated in the graph-based model, ideally resulting in further improvements of personalized recommendation performance.

ACKNOWLEDGMENTS

We thank anonymous reviewers for helpful comments on previous versions of this document. The work was partially supported by the Israeli Science Foundation grand ISF-1108/2014 and by Magnet InfoMedia grant.

REFERENCES

1. Abel, F., Gao, Q., Houben, G.-J., and Tao, K. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*. Springer, 2011, 1–12.

2. Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, ACM (2008), 895–904.

3. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

4. Bogers, T. Movie recommendation using random walks over the contextual graph. In *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems* (2010).

5. Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 1185–1194.

6. De Choudhury, M., Gamon, M., and Counts, S. Happy, nervous or surprised? classification of human affective states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media* (2012).

7. De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. (2013).

8. Dooms, S., De Pessemier, T., and Martens, L. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, vol. 2013 (2013).

9. Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., and Cantador, I. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction* (2015), 1–35.

10. Golbeck, J., Robles, C., Edmondson, M., and Turner, K. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, IEEE (2011), 149–156.

11. Gori, M., Pucci, A., Roma, V., and Siena, I. Itemrank: A random-walk based scoring algorithm for recommender engines. In *IJCAI*, vol. 7 (2007), 2766–2771.

12. Hannon, J., Bennett, M., and Smyth, B. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, ACM (2010), 199–206.

13. Haveliwala, T. H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering 15*, 4 (2003), 784–796.

14. Hu, R., and Pu, P. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*, ACM (2011), 197–204.

15. Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, Ieee (2008), 263–272.

16. Konstas, I., Stathopoulos, V., and Jose, J. M. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM (2009), 195–202.

17. Kuflik, T., Minkov, E., and Kahanov, K. Graph-based recommendation in the museum. In *DMRS* (2014), 46–48.

18. Lam, S., and Herlocker, J. Movielens 1m dataset, 2012.

19. Li, J., Ritter, A., and Hovy, E. Weakly supervised user profile extraction from twitter. *ACL, Baltimore* (2014).

20. Lin, J., Sugiyama, K., Kan, M.-Y., and Chua, T.-S. Addressing cold-start in app recommendation: Latent user models constructed from twitter followers. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM (2013), 283–292.

21. Minkov, E., and Cohen, W. W. Improving graph-walk-based similarity with reranking: Case studies for personal information management. *ACM Transactions on Information Systems (TOIS) 29*, 1 (2010), 4.

22. Minkov, E., Kahanov, K., and Kuflik, T. Graph-based recommendation integrating rating history and domain knowledge: Application to on-site guidance of museum visitors. *Journal of the Association for Information Science and Technology (JASIST) 68*, 2 (2017).

23. Mohammad, S. M., and Kiritchenko, S. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence 31*, 2 (2015), 301–326.

24. Musto, C., Lops, P., Basile, P., de Gemmis, M., and Semeraro, G. Semantics-aware graph-based recommender systems exploiting linked open data. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, ACM (2016), 229–237.

25. Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., and Yang, Q. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*, IEEE (2008), 502–511.

26. Pennebaker, J. W., Francis, M. E., and Booth, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71* (2001), 2001.

27. Pham, T.-A. N., Li, X., Cong, G., and Zhang, Z. A general graph-based model for recommendation in event-based social networks. In *2015 IEEE 31st International Conference on Data Engineering*, IEEE (2015), 567–578.

28. Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ACM (2010), 37–44.

29. Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., Ungar, L. H., and Schwartz, H. A. Developing age and gender predictive lexica over social media.

30. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one 8*, 9 (2013), e73791.

31. Shapira, B., Rokach, L., and Freilikhman, S. Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction 23*, 2-3 (2013), 211–247.

32. Tiroshi, A., Berkovsky, S., Kaafar, M. A., Chen, T., and Kuflik, T. Cross social networks interests predictions based ongraph features. In *Proceedings of the 7th ACM conference on Recommender systems*, ACM (2013), 319–322.

33. Tiroshi, A., Berkovsky, S., Kaafar, M. A., Vallet, D., Chen, T., and Kuflik, T. Improving business rating predictions using graph based features. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, ACM (2014), 17–26.

34. Tiroshi, A., Berkovsky, S., Kaafar, M. A., Vallet, D., and Kuflik, T. Graph-based recommendations: Make the most out of social data. In *User Modeling, Adaptation, and Personalization*. Springer, 2014, 447–458.

35. Tkalcic, M., Kunaver, M., Košir, A., and Tasic, J. Addressing the new user problem with a personality based user similarity measure. Citeseer (2011).

36. Volkova, S., Bachrach, Y., Armstrong, M., and Sharma, V. Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).

37. Zhang, Y., and Pennacchiotti, M. Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web*, International World Wide Web Conferences Steering Committee (2013), 1521–1532.