Predicting User Participation in Social Media

Fredrik Erlandsson[1](✉), Anton Borg[1], Henric Johnson[1], and Piotr Bródka[2]

[1] Blekinge Institute of Technology, 37179 Karlskrona, Sweden
{fredrik.erlandsson,anton.borg,henric.johnson}@bth.se
[2] Wrocław University of Technology, 50-370 Wrocław, Poland
piotr.brodka@pwr.edu.pl

Abstract. Online social networking services like Facebook provides a popular way for users to participate in different communication groups and discuss relevant topics with each other. While users tend to have an impact on each other, it is important to better understand and analyze users behavior in specific online groups. For social networking sites it is of interest to know if a topic will be interesting for users or not. Therefore, this study examines the prediction of user participation in online social networks discussions, in which we argue that it is possible to predict user participation in a public group using common machine learning techniques. We are predicting user participation based on association rules built with respect to user activeness of current posts. In total, we have crawled and extracted 2,443 active users interacting on 610 posts with over 14,117 comments on Facebook. The results show that the proposed approach has a high level of accuracy and the systematic study clearly depicts the possibility to predict user participation in social networking sites.

1 Introduction

Online social networks are a large part of our society. Just Facebook alone attracts 1.3 billion users[1] with 640 million minutes spent each month. Facebook had a total revenue of $12,466 M in 2014[1]. Consequently, discovering trending topics or influential users is of interest for many researchers, e.g. for marketing [6]. Several studies have tried to identify user influence, however most have used page rank [21] or centrality [5,17] based approaches to identify influential users.

In this article we argue that users, on Facebook groups, are following each other and that it is possible to detect influential users. E.g. if user A, B, C and D share common interests, the chance is that if A, B, and C already have commented on a topic, D also will comment on it. Therefore, this paper relates to how users perform actions (e.g. comments or likes) on posts in Facebook pages. In addition, we use association learning to discover relationships between variables, or in our case users, in the dataset [9]. Given a list of posts from a specific domain we extract users actions such as comments and likes.

[1] http://www.statisticbrain.com/facebook-statistics/.

A. Wierzbicki et al. (Eds.): NetSci-X 2016, LNCS 9564, pp. 126–135, 2016.
DOI: 10.1007/978-3-319-28361-6_10

Using association rule learning on the data, we argue that it is possible to predict if a particular user will or will not participate on a post discussion based on the other users activity.

For evaluation, a systematic study is conducted, which include building association rules that can be used to predict if a specific user will be active in a particular post. The prediction is done based on the activeness of users within current posts. Moreover, the scope of the paper is limited to user interactions on a subset of Facebook users on posts with a similar topic.

The paper is organized as follows: In Sect. 2 related work is discussed. Sections 3 and 5 presents the data and the methodology. Association rule learning and the evaluation metrics are discussed in Sect. 4. Finally, the results are presented in Sect. 6 and discussed in Sect. 7.

2 Related Work

Online social networks and social media analysis are one of the hottest areas of research in modern network science. Like in many different areas, scientists struggle to predict the future of online social network. The main focus in social network area is on link prediction [16] but different teams around the world work also on: (i) popularity prediction in social media based on comment mining [12], (ii) personality prediction for micro blog Users [29], (iii) churn prediction and its influence on the network [4,22], (iv) community evolution prediction [7,23], (v) using social media to predict real-world outcomes [3], (vi) predicting information cascade on social media [11], (vii) users features prediction using relational learning [14,15], (viii) predicting patterns of diffusion processes in social network [13], (ix) predicting friendship intensity [2,20], (x) affiliation recommendations [25,26], and many others.

Association rule mining has been previously used in social network and social media analysis. In [18], the authors explores the association rule between a course and gender in the Face book 100 university dataset. This was performed to discover the influence of gender in studying a specific course. [27] introduces the scheme for association rule mining of personal hobbies in social networks, while [24] tackle the problem of mining association rules in folksonomies and try to find out how association rule mining can be applied to analyze and structure folksonomies.

However, while online social network analysis is popular, there is according to our review a lack of research on using association rules for predicting user participation in online social media discussions.

3 Data Model

The data used in this study has been obtained from the crawler described by [8]. This crawler gathers complete posts from Facebook. In this context, the term complete stands for posts that contains all likes and comments. In addition, if a post is crawled, the dataset contains all likes, comments and interacting users

up to the crawling time. Our current dataset, captured from public pages and groups on Facebook, consists of over 56 million posts, 560 million comments and 7.3 billion likes made by 820 million Facebook users. The crawled data is parsed and available from a SQL database, structured as described in [19], making all fields needed for our task available. In this study, we assume that the investigated posts will not get any new comments. We simplify the dynamics of social media by saying that the posts we are investigated are "dead" when the data was collected, in which the term of dead posts refers to posts that no longer attracts attention or new comments or likes.

We are limiting this study to only investigate a subset of groups available by the crawler. From these groups we exclude posts with less than 20 comments as these posts are considered to be of too low value and do not hold enough information.

3.1 Data Selection

To perform prediction of user interactions, we have selected the page OccypyTogether. This page was selected based on the following properties: it is active, it has a high number of users (~300 k), it has a reasonable high number of active users (~30,000 users with more than one comment) and it is political with a bias user group (most of the users are positive to the Occupy movement). From this page, only users that have made more than five comments are investigated. This ensures that the selected users are or have been fairly active in the community. The resulting dataset consists of 2,443 users interacting on 610 posts totaling in 14,117 comments.

4 Association Rules

As stated in Sect. 1, we are predicting user participation based on previous interactions with other users on common posts. We argue that if user A participates in all posts where B is participating, there is a high chance of A participating in a new post where B is already active. The method of matching items in different transactions is called association rule mining. We apply association rule mining to the domain of social media where we model the data as follows. Items correspond to users on Facebook and transactions correspond to posts. An user is considered to be active and part of the transaction, as an item, if the user comments on a post.

To build association rules from our dataset, we evaluated several implementations. [1] presented the Aprori algorithm, which was proven to be an efficient method for association rule learning. This algorithm is however proven to have efficiency issues in large datasets [10] and the identified implementation for Python is very slow (considering our dataset it was not possible to get a result within reasonable time). Hence, other algorithms were tested, and in particular the Eclat algorithm [28]. The Eclat algorithm quickly discards items with low frequency by considering a minimum of associations as input parameters.

From the selected dataset, described in Sect. 3.1, we firstly count the frequency of all posts where A and B are active respectively. Secondly, we count all posts where $A \cup B$, both participates. This gives us two measures, length (the number of participating users) and frequency (the sum of all posts where they are participating). These two steps can be summarized as, building frequent item-sets ($\mathcal{I}$). Finally, all possible rules from the computed $\mathcal{I}s$ are generated. In this step we also compute the evaluation metrics described below.

4.1 Evaluation Metrics

To understand the learned association rules, there exist a few metrics. First, we have *Support*, where we compute the frequency of a given item-set, $\mathcal{I}$, and divide it with the total number of transactions (posts) in $\mathcal{D}$. Or, the number occurrence of $\{A, B\}$ in our dataset, $\mathcal{D}$ divided by length of $\mathcal{D}$. As shown in (1).

$$support(\{A, B\}) = \frac{\{A, B\}}{|\mathcal{D}|} \tag{1}$$

Secondly, we have *Confidence*, which is an indicator saying that $\{A, B\} \Rightarrow C$ in the set of transactions in $\mathcal{D}$ is the proportions of transactions that contain $\{A, B\}$ also will contain C as illustrated in (2). Say that $\{A, B, C\}$ participates in 4 common posts and $\{A, B\}$ participates in 8 posts in total. This leads to $4/8 = 0.5$ i.e., the *confidence* that C will participate on a post where A and B already are active is 50 %.

$$confidence(\{A, B\} \Rightarrow C) = \frac{support(\{A, B, C\})}{support(C)} \tag{2}$$

Thirdly, we have *lift*, a ratio of the interdependence of the observed values. As we see from (3), if lift is 1, it implies that the rule and the items are independent of each other. However, if the lift is > 1, the lift indicates the degree of dependency of our item-sets.

$$lift(\{A, B\} \Rightarrow C) = \frac{support(\{A, B, C\})}{support(\{A, B\}) \times support(\{C\})} \tag{3}$$

Finally, we have *conviction*, as the ratio of the expected *support* that $\{A, B\}$ occurs without C as shown in (4). Notable, *conviction* is infinite (due to division with zero) when the *confidence* is 1.

$$conviction(\{A, B\} \Rightarrow C) = \frac{1 - support(\{A, B\})}{1 - confidence(\{A, B\} \Rightarrow C)} \tag{4}$$

The described measures enable understanding of the learned rules in $\mathcal{D}$, where higher number of all four measures indicate that the learned rule has relevance for prediction.

5 Methodology

The final dataset used in the experiment consists of 2,443 users interacting on 610 posts and writing 14,117 comments. The selected users are or have been fairly active in the community, which reflect how we build the association rules.

The algorithm used for the association rule mining is the Eclat algorithm. The Eclat algorithm learns about all the frequent item-sets in our data. By using Eclat, it is possible to define a lower bound threshold and in our dataset a good trade-off between resolution and speed is 4, where lower frequency is ignored. The used implementation of Eclat is modified to sort the item-sets by participants so only $\{A, B, C\}$ is considered. Other combinations e.g., $\{B, C, A\}$ and $\{C, A, B\}$ are consolidated in the item-set $\{A, B, C\}$. Association rules supporting the hypothesis of user participation based on other users activities were computed from the calculated frequency item-sets. The results are measured using the evaluation metrics presented in Sect. 4.1.

6 Results

The resulting frequent item-sets are depicted in Fig. 1. This figure illustrates the length of elements (number of collaborating users) with respect to frequency (the number of occurrence for each item-set). The main scatter-plot illustrates how the *frequency* decreases when the number of users (*length*) increases, a natural feature of frequent item-sets.

Figure 1 also depicts the distribution as histograms. The top histogram shows the distribution of frequency and the histogram on the right hand side shows the distribution of the length of the learned item-sets. The top histogram illustrates a significant density of user collaboration to occur at low frequency, between 4–6. This is natural as the frequency of user participation decreases for most of the users. Noticeable on the length distribution is the fact that the density is higher for two and three participating users than for just one. This is because there exist more combinations of users than the number of single users.

Association rules supporting the hypothesis of user participation based on other users activities were computed from the calculated frequency item-sets. Resulting in 55,166 rules. Table 1 shows descriptive statistics for all the computed rules. It can be noted that although the confidence median and mean is low, the high level of lift indicates high dependency of the learned rules, i.e., the computed rules show that out hypothesis is valid and users tend to follow each other. As our dataset is big, with many users and many posts, the low support mean and median is expected. Moreover, it is noticeable that users are not active in all posts but more on a subset of them.

Figure 2 depicts the distribution, Confidence, Lift, Conviction and Frequency respectively in our learned model. The figures are illustrated as violin-plots which represents the kernel density (shown as height and depth) in addition to normal box-plots with outer quartiles as thin lines, the inner quartiles as bold lines and the mean as a white dot.

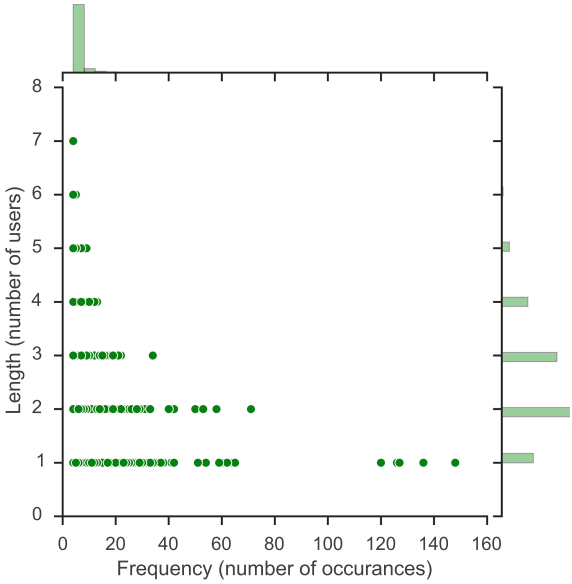

Fig. 1. Combined plot of number of users (Length) with respect to number of occurrence (Frequency). The upper and right axis illustrates histograms of the respective distributions

Table 1. Descriptive statistics of 55,166 computed rules.

	Mean	Median	Std
Support	0.05	0.02	0.07
Confidence	0.43	0.33	0.33
Lift	18.97	9.38	24.64
Conviction	1.83	1.32	1.18

Figure 2a shows a dense distribution of support at 0.025 and interestingly a higher density at 0.20. The confidence distribution is illustrated in Fig. 2b, interestingly there is a dense distribution around 1.0, i.e., there is a significant number of learned rules with high confidence that the rule is accurate. Figure 2c shows that the lift measure have a heavy tail distribution. Figure 2d illustrates a distribution of conviction to be concentrated between zero and five.

Table 2 presents learned rules in tree sections. Each section is sorted firstly by Confidence, Lift and Conviction respectively and secondly by the number of supporting users. The rule $\{u_{429},\ u_{578}\} \Rightarrow \{u_{19}\}$ should be interpreted as user–429 together with user–578 influence participation of user–19. Notable, when sorting by confidence and lift, the conviction is infinite (this is due to the confidence of 1.0) as shown how conviction is calculated in (4). All of the rules in Table 2 have high confidence and show high dependency (via the lift metric), i.e., the top five rules sorted by either Confidence, Lift or Conviction are relevant for predicting user participation.

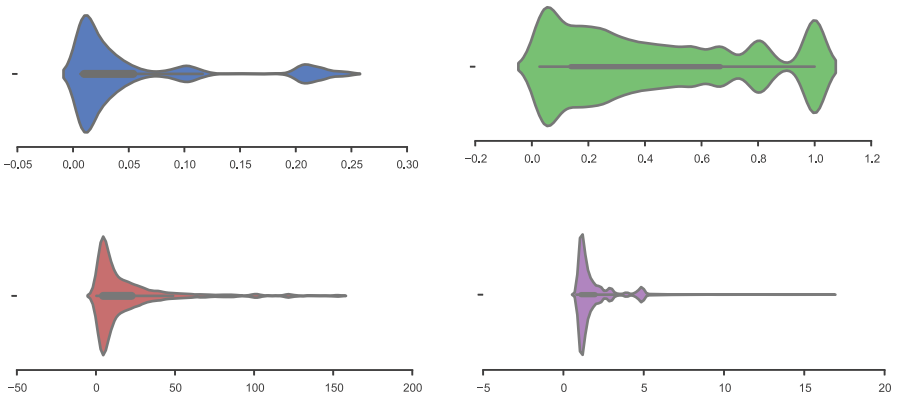

Fig. 2. Distribution of values in learned association rules.

The rule, $\{u_{580},\ u_{861},\ u_{1352},\ u_{1466}\} \Rightarrow \{u_{896},\ u_{1291}\}$ presented in Table 2 with confidence of 1.0 and lift of 152.5 strongly indicates that the left-hand-side user set influences the right-hand-side user set, i.e., when the left-hand-side user set is active on a post the right-hand-side user set also will be active. A confidence of 1.0 means that 100 % of the posts where the left-hand-side user set is active, the right user set also will be active. A lift value of 152.5, in this specific rule, shows that the right-hand-side user set is dependent on the left.

Considering rules where at least two separate users affect another user with a confidence of $\geq 95\,\%$. We can reduce the 55,166 rules to 4,959 rules, which have a median lift of 4.80 and a median support of 21 %. In other words, we have close to 5,000 rules that strongly indicates that users are affected by each other when it comes to participating in online social networks. From learned rules, we can also identify influential users, i.e., the users that exists on the left side of multiple rules.

7 Conclusion and Discussion

Users within online social networks creates a large amount of generated data in form of interactions (comments and likes). Not enough attention has been put on the prediction of how users influence each other and how to predict the behavior of users within Facebook groups. Therefore, we have, in this paper, crawled a significant amount of user data and then by using machine learning, implemented and examined how users influence each other. Based on the results and analysis, we are able to determine that users influence other users to participate and interact in new groups.

From the group OccypyTogether, 2,443 active users have been extracted. They interact on 610 posts with a total of 14,117 comments. From this dataset, the association rules were computed. Resulting in almost 5,000 rules with high confidence of correctness, $\geq 95\,\%$. These rules were proven to be dependent of the

Table 2. Top 5 rules sorted by different metrics

Rule	Confidence	Lift	Conviction
Confidence			
$\{u_{179}, u_{538}, u_{580}, u_{938}, u_{992}, u_{1090}\} \Rightarrow \{u_{11}\}$	1.00	10.17	inf
$\{u_{11}, u_{31}, u_{80}, u_{179}, u_{992}, u_{1093}\} \Rightarrow \{u_{580}\}$	1.00	4.80	inf
$\{u_{11}, u_{31}, u_{179}, u_{580}, u_{992}, u_{1093}\} \Rightarrow \{u_{80}\}$	1.00	9.53	inf
$\{u_{11}, u_{179}, u_{538}, u_{580}, u_{938}, u_{953}\} \Rightarrow \{u_{429}\}$	1.00	4.84	inf
$\{u_{179}, u_{1094}, u_{1096}, u_{1113}, u_{1171}, u_{1352}\} \Rightarrow \{u_{1378}\}$	1.00	101.67	inf
Lift			
$\{u_{580}, u_{861}, u_{1352}, u_{1466}\} \Rightarrow \{u_{896}, u_{1291}\}$	1.00	152.50	inf
$\{u_{580}, u_{861}, u_{1291}, u_{1352}\} \Rightarrow \{u_{896}, u_{1466}\}$	1.00	152.50	inf
$\{u_{31}, u_{80}, u_{179}, u_{580}\} \Rightarrow \{u_{11}, u_{992}, u_{1093}\}$	1.00	152.50	inf
$\{u_{19}, u_{64}, u_{673}, u_{685}\} \Rightarrow \{u_{54}, u_{581}\}$	1.00	152.50	inf
$\{u_{580}, u_{861}, u_{1291}, u_{1466}\} \Rightarrow \{u_{896}, u_{1352}\}$	1.00	152.50	inf
Conviction			
$\{u_{429}, u_{578}\} \Rightarrow \{u_{19}\}$	0.95	3.93	16.66
$\{u_{920}\} \Rightarrow \{u_{179}\}$	0.95	4.27	16.32
$\{u_{929}\} \Rightarrow \{u_{179}\}$	0.95	4.26	15.54
$\{u_{580}, u_{1093}\} \Rightarrow \{u_{179}\}$	0.94	4.22	13.21
$\{u_{580}, u_{938}\} \Rightarrow \{u_{179}\}$	0.94	4.22	13.21

active users, via the lift metric. Therefore, the hypothesis of user participation influences can be accepted. The results also proved that using association rule learning, influential users can be identified. Moreover, users on the left-hand-side, in a rule with high confidence and high lift, are influencing users on the right-hand-side to participate in the conversation.

At present, information on Facebook are filtered by a secret algorithm. This poses a potential validity threat to our results. Even external recommender systems might pose a threat as data might be bias since users can only see a subset of all posts.

For future work, it would be interesting to compare the results across different Facebook groups, e.g. politics-related Facebook group is different from news-related Facebook groups. Additionally, methods for association rule learning that supports number of occurrence and order of items in each transaction also needs to be investigated further. Finally, investigating the temporal aspects of users participation, e.g. whether users influence each other over time, or if a user participates throughout a discussion or only in the beginning, is something that needs to be considered and which could hopefully improve the prediction results.

Acknowledgement. This work was partially supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE] and by The National Science Centre, the decision no. DEC-2013/09/B/ST6/02317.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994)
2. Ahmad, W., Riaz, A., Johnson, H., Lavesson, N.: Predicting friendship intensity in online social networks. In: Proceedings of the 21st Tyrrhenian Workshop on Digital Communications: Trustworthy Internet. Springer, September 2010
3. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2010, vol. 01, pp. 492–499. IEEE Computer Society, Washington, DC (2010)
4. Au, W.H., Chan, K.C., Yao, X.: A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. Evol. Comput. **7**(6), 532–545 (2003)
5. Bródka, P.: Key user extraction based on telecommunication data (aka. key users in social network. how to find them?) (2013). arXiv preprint arXiv:1302.1369
6. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in Twitter: the million follower fallacy. ICWSM **10**(10–17), 30 (2010)
7. De Meo, P., Ferrara, E., Rosaci, D., Sarne, G.M.L.: Trust and compactness in social network groups. IEEE Trans. Cybern. **45**(2), 205–216 (2015)
8. Erlandsson, F., Nia, R., Boldt, M., Johnson, H., Wu, S.F.: Crawling online social networks. In: 2015 European Network Intelligence Conference (ENIC), September 2015
9. Flach, P.: Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, New York (2012)
10. Goethals, B.: Survey on frequent pattern mining. Technical report, University of Helsinki (2003)
11. Hakim, M., Khodra, M.: Predicting information cascade on Twitter using support vector regression. In: 2014 International Conference on Data and Software Engineering (ICODSE), pp. 1–6, November 2014
12. Jamali, S., Rangwala, H.: Digging digg: comment mining, popularity prediction, and social network analysis. In: International Conference on Web Information Systems and Mining, 2009, WISM 2009, pp. 32–38, November 2009
13. Jankowski, J., Michalski, R., Kazienko, P.: The multidimensional study of viral campaigns as branching processes. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 462–474. Springer, Heidelberg (2012)
14. Kajdanowicz, T., Kazienko, P., Indyk, W.: Parallel processing of large graphs. Future Gener. Comput. Syst. **32**, 324–337 (2014)
15. Kazienko, P., Kajdanowicz, T.: Label-dependent node classification in the network. Neurocomputing 75(1), 199–209 (2012), Brazilian Symposium on Neural Networks (SBRN 2010) International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010)
16. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
17. Musiał, K., Kazienko, P., Bródka, P.: User position measures in social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNA-KDD 2009, pp. 6:1–6:9. ACM, New York (2009)

18. Nancy, P., Geetha Ramani, R., Jacob, S.: Mining of association patterns in social network data through data mining techniques and methods. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds.) Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing, vol. 178, pp. 107–117. Springer, Berlin, Heidelberg (2013)
19. Nia, R., Erlandsson, F., Bhattacharyya, P., Rahman, M.R., Johnson, H., Wu, S.F.: Sin: a platform to make interactions in social networks accessible. In: 2012 International Conference on Social Informatics (SocialInformatics), pp. 205–214. IEEE, December 2012
20. Nia, R., Erlandsson, F., Johnson, H., Wu, S.F.: Leveraging social interactions to suggest friends. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 386–391. IEEE (2013)
21. Riquelme, F.: Measuring user influence on Twitter: a survey (2015). CoRR abs/1508.07951
22. Ruta, D., Kazienko, P., Bródka, P.: Network-aware customer value in telecommunication social networks. In: IC-AI, pp. 261–267 (2009)
23. Saganowski, S., Gliwa, B., Bródka, P., Zygmunt, A., Kazienko, P., Koźlak, J.: Predicting community evolution in social networks (2015). arXiv preprint arXiv:1505.01709
24. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining association rules in folksonomies. In: Batagelj, V., Bock, H.-H., Ferligoj, A. (eds.) Data Science and Classification, pp. 261–270. Springer, Heidelberg (2006)
25. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. ACM, New York, August 2005
26. Vasuki, V., Natarajan, N., Lu, Z., Savas, B., Dhillon, I.: Scalable affiliation recommendation using auxiliary networks. ACM Trans. Intell. Syst. Technol. (TIST) **3**(1), 1157–1162 (2011). Article No. 3
27. Yu, X., Liu, H., Shi, J., Hwang, J.N., Wan, W., Lu, J.: Association rule mining of personal hobbies in social networks. In: 2014 IEEE International Congress on Big Data (BigData Congress), pp. 310–314, June 2014
28. Zaki, M.J.: Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. **12**(3), 372–390 (2000)
29. Seng, B.: ICT for sustainable development of the tourism industry in Cambodia. In: Zu, Q., Hu, B., Gu, N., Seng, S. (eds.) HCC 2014. LNCS, vol. 8944, pp. 1–14. Springer, Heidelberg (2015)