360° User Profiling: Past, Future, and Applications

Aleksandr Farseev[1], Mohammad Akbari[1], Ivan Samborskii[2] and Tat-Seng Chua[1]
[1]School of Computing, National University of Singapore
[2]Department of Computer Technologies, ITMO University

Users in social networks are often encouraged to complete their profile by providing personal attributes such as age, gender, interest, income, etc. Additionally, users are likely to join interest-based groups that are devoted to various topics: "2016 University Graduates", "Accordions Singapore", etc. These profiles and groups are often used as a basis for rendering better online services, marketing, and advertisement. However, in practice, the majority of users are reluctant to provide actual personal attributes, while the group participation is often relevant to their friendship connections, rather than interests. As a solution, user profiling at the individual and group level is explored to mine the demography and mobility information of a user. In this paper, we discuss different user profiling approaches on social networks, highlight the challenges, techniques, and future trends. We explain the weakness and strength of these methods and introduce an analytic platform to bridge the gap between social media users, business intelligence and the Big Data.

DOI: 10.1145/2956573.2956577 http://doi.acm.org/10.1145/2956573.2956577

1. Introduction

The past decade has recorded a rapid development and change in the Web and Internet. We are currently witnessing an explosive growth in social networking services, where users are publishing and consuming online contents. In such a context, millions of users publish their posts regularly on different online social platforms, such as Twitter, Facebook, Foursquare, and Instagram. In particular, more than 35 percent of American adults older than 65 years use social media, which records an increase of more than three times as compared to 2010 statistics. Further, it was reported that more than half of adult users are active in multiple online social forums, daily[1]. The increasing amounts of published user-generated content (UGC) provides opportunities as well as challenges for its consumers — users and data analysts. At the users level, the biggest problem is the information filtering, due to the growing amount of irrelevant content, such as advertisement and spam. At the analytic level, data-preprocessing, fusion, and modeling probably pose the greatest challenges that need to be tackled. In the following, we highlight the existing challenges:

Data Gathering: Due to the sensitivity of privacy, only a limited amount of data can be collected for each active internet user. Even worse, after the collection of necessary data,

[1]According to Pew Research Centers' Social Networking Fact Sheet, 2014, `www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`

it is a big challenge to align various social network accounts to the same user. Another problem is the lack of ground truth, which hinders the development of supervised learning approaches. *Collection of multi-source data and user accounts mapping are the crucial challenges*.

Multi-source data is often of different modality: Except for text, other data modalities often appear in users' timeline. For example, users may shoot and share pictures in photo-sharing services like Instagram, upload videos to Youtube, perform sports activities in MyFitnessPal, or geolocate themselves in Foursquare. Additionally, the nature of users' posts on conventional social media platforms, such as Facebook or Twitter, have become more multimedia. Concerning the above, user profile learning on these complex information requires *seamless integration of such heterogeneous data in a complementary way, which is the great challenge.*

Data comes from multiple sources: Users frequently enroll in different on-line social networks that capture them from various perspectives. For example, Linkedin conveys users' formal career path; while Foursquare uncovers users' offline activities. *Effective integration of multi-source data for individual and group profiling is a tough challenge*.

2. Data Gathering and Representation

As mentioned before, data gathering and representation are of crucial importance for social media research. In this section, we outline the techniques on data collection and representation. In order to cover the most popular modalities (visual, sensor, textual, and location data), we pick up the following social media sources as examples: Foursquare as a location data source; Twitter as a textual data source; Instagram as a visual data source; and MyFitnessPal[2] as a sensor data source.

2.1. Data Gathering

First, we address the problem of cross-source user account mapping (also known as cross-source user-identification [Vosecky et al., 2009]), which is an opened research problem in social media computing. Recently, some solutions were proposed, which are mainly based on machine learning and information retrieval techniques. The detailed discussion of these approaches are beyond the scope of this paper, and can be referred from the past works [Vosecky et al., 2009; Zhang et al., 2016].

2.1.1. Cross-source user accounts mapping. In this work, we propose two approaches for avoiding the user identification problem such that the research can be purely focused on data processing and modeling: (I) Use of existing mapping directories and (II) Posts co-occurrence monitoring.

(I) Use of existing mapping directories. The idea is simple: nowadays, there exist a multitude of so-called social media aggregation services (i.e. About.me[3]) that are commonly used to present users to public. Such services list the publicly available social network

[2] www.myfitnesspal.com
[3] www.about.me

accounts in a form of a name card so that other users can easily contact or follow them on one of the listed social networks. From the research perspective, these services are the great opportunity to obtain the user account mapping among the same set of users on a different social forums [Song et al., 2015b].

(II) Posts co-occurrence monitoring. It is known that many social media users often connect their accounts in different social networks via the so-called, cross-site linking functionality [Chen et al., 2014; Zhang et al., 2016]. These users often forward their messages from one social network to another, such that these posts become more publicly accessible and, thus, can be easily harvested for further processing, analysis, and learning. One example is the use of Twitter as a common "sink" for all the other data sources [Farseev et al., 2015b]. In such a setting, the crawling process consists of two steps: first, a set of active users who have recently posted tweets through the cross-linking functionality of multimedia social networks is collected; second, Twitter Stream API[4] (or Twitter REST API[5]) is used to perform the tweets streaming with respect to specified geographical regions, so that cross-social network account mapping can be performed when users publish tweets that contain posts from other social networks (i.e. post Instagram images on Twitter). By following the described user identification approach, in our previous work [Farseev et al., 2015b] *we collected and published our Big, Multi-Source, Social Dataset NUS-MSS*[6], which includes user-generated data collected from the same users of Twitter, Instagram, Foursquare, and Facebook with respect to three geographical regions: Singapore, London, and New York.

2.1.2. Multi-source data gathering. Based upon the collected cross-network user mapping list, the user-generated contents can be obtained from multiple sources. For example, Twitter streams can be monitored with the keywords specified as "`swarmapp.com`" and "`instagram.com`" to receive tweets posted from Foursquare (Swarm) and Instagram mobile applications, respectively. Specifically, in Twitter, each sampled check-in message contains a short link to the original check-in page, where the check-in details are available. Noticeably, the proposed data gathering approach can be enriched by using API of other social networks, since the user IDs in that networks are available after the cross-network mapping being performed.

The above-mentioned data gathering approaches can be categorized as "user-based", which means that the data is collected based on a pre-defined user list. However, an alternative approach to data gathering is based on an individual concept or word ("topic-based" approach), so that the collected multi-modal UGC would be related to the similar pre-defined topic. Such a collection can be performed by using advanced Twitter search with the keywords (or hashtags) specified as a query. For example, the tweets that are related to Diabetes can be harvested by querying Diabetes-related hashtags such as "#diabetes" or "#BGnow" [Akbari et al., 2016], while the sport-activity related tweets can be monitored by querying hashtags "#myfitnesspal", "#Runkeeper", or "#FitBit". The above three hashtags indicate the auto-generated reports of the sport-tracking mobile applications and provide the URL of the sports activity inside the Twitter post. The actual data that represents

[4]`dev.twitter.com/streaming/overview`
[5]`dev.twitter.com/rest/public`
[6]`ms.comp.nus.edu.sg/research/NUS-MULTISOURCE.htm`

the sports activity can then be crawled from the sports activity's web page.

The aforementioned data gathering solutions have its own advantages and drawbacks. For example, the user account mapping based upon existing mapping directories is simple, but limited by providing only those user accounts, where the cross-network mapping is explicitly revealed. Such a collection approach is biased towards a certain type of user personalities and online behaviors. At the same time, the approach based upon posts co-occurrence mining is less biased towards users' personality and online goals, but limited by the necessity of cross-site functionality to be enabled and used by users. Finally, the "topic-based" approach is useful for harvesting data that is related to certain topic concepts but suffers from small data amount per user. Further, the power law distribution of data makes it difficult to find topics which are discussed less frequently. The decision regarding which approach to select is of crucial importance and often depends on the final application [Farseev et al., 2015b; Song et al., 2015a; Akbari et al., 2016].

2.2. Data Representation

To model the multi-modal data efficiently, it must be represented in mutually-consistent form. Specifically, textual, visual, location, and other data sources must be integrated into one model in such a way that they compliment, but not contradict each other. Below, we list several approaches to do so:

Textual data: In the case of textual data, it is common to represent user messages in the form of latent topics, where the topics are extracted by applying training topic models such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003] or Probabilistic Latent Semantic Analysis (PLSA) [Vorontsov et al., 2015]. Latent topic modeling has been widely used as an effective approach for representing the semantic meaning of the data, which, in turn, is useful for user profile learning [Farseev et al., 2015b]. At the same time, the LIWC categories [Pennebaker et al., 2001] are known to be good in reflecting writer's personality and, thus, can be used for personality profiling. Finally, the manually-defined "writing style", "emotions-related", and "online activity level" representations such as that discussed in [Farseev et al., 2015b] are, in general, important for user profile learning, while its combination with LIWC and LDA data representations performs the best [Farseev et al., 2015b].

Location data: The location data in its original form (time-stamped geographical coordinates) is often processed in the form of mobility representations. For example, the so-called Areas of Interest (AoI) [Qu and Zhang, 2013] can be extracted from users' geolocation tracks, and, thus, uncover common areas of their activities. Further, different AoI metrics (such as the number of AoI, its size, the number of independent locations) can be extracted. Finally, time-dependent and position-dependent features such as "Pace popularity", "Physical and Rank Distance", "Place and Activity Transition", "Time Intervals Distribution" [Noulas et al., 2012] etc, can be utilized. At the same time, the location semantics (i.e. venue type such as Restaurant, Airport, etc.) brings another important information for user profiling [Farseev et al., 2015b] and recommendation [Wang et al., 2015; Farseev et al., 2015a] tasks. The idea here is to represent each user as a distribution of all his/her semantically marked geo-locations among pre-defined venue category list,

such as Foursquare venue categories[7]. Additionally, the latent topics [Blei et al., 2003; Vorontsov et al., 2015] can be extracted to infer higher-level semantic representations.

Visual data: It is common to extract different low-level and high-level data representations from images and key-frames of videos. One commonly used low-level representations is the HOC [Dalal and Triggs, 2005] or [Lowe, 1999] descriptors. At the same time, it is important to represent the semantic meaning of the data [Farseev et al., 2015b], using visual concepts learning from public datasets (such as ImageNet [Deng et al., 2009] or NUS-WIDE [Chua et al., 2009]). It has been shown, that visual data representations are of crucial importance for user profiling [Farseev et al., 2015b] and recommendation [Zhao et al., 2013].

3. Personal User Profiling

User profile learning plays an increasingly important role in many application domains, such as Recommendation [Zhao et al., 2015], Attributes Inference [Farseev et al., 2015b], and Timeline Analysis [Akbari et al., 2016]. Broadly speaking, user profile learning techniques can be divided into two categories: *static profiling* and *dynamic profiling*. We give some examples of research done in these categories next.

3.1. Static Profiling

Traditionally, the task of personal user profiling is treated as a supervised learning task, where different user attributes are inferred users' online behavior and their generated content from multiple social networks. In static profiling, data representation does not depend on the temporal aspect or related to it implicitly via the construction of aggregated representations over the whole dataset (these gives the name "static"). Till now, there were some research efforts dedicated to multi-source static user profile learning. For example in earlier work, Yuan et al. [Yuan et al., 2012] studied the approaches to integrate multi-modal data from clinical measurements to enhance the results of Alzheimer's Disease prediction. However, the model was formulated as a binary classification task, and, thus, not directly applicable to real-world scenarios. Later on, Song et al. [Song et al., 2015a] proposed a learning framework for Volunteerism tendency prediction for users of different social networks. The missing data was inferred by the proposed constrained Non-Negative Matrix Factorization (NMF) approach. The problem was also formulated for the binary classification. At the same time, [Farseev et al., 2015b] introduced an efficient ensemble learning solution, aiming to combine multi-source multi-modal data for demographic user profiling. The above model performed the best among various state-of-the-art baselines and demonstrated the necessity of learning from multiple sources to improve the user profiling performance.

3.2. Dynamic Profiling

The user profiling approaches from the previous section assume that users' attributes and available content are static, which means that they do not change over time. However, this

[7] `developer.foursquare.com/categorytree`

assumption is often not accurate for cases when data is delivered with a high velocity. Some recent efforts have been made on analyzing temporal behaviors of users aiming at learning a dynamic user profile. For example, the Netflix award winning algorithm timeSVD++ [Koren, 2010] models the temporal evolution of users and product characteristics aiming at intelligently distinguish transient factors from lasting ones. By learning a dedicated latent basis for each user at each particular time point, they significantly improved the effectiveness of the recommendation process. Similarly, probabilistic tensor factorization has been utilized to learn low-dimensional latent space models the time-evolving changes of users and items [Xiong et al., 2010]. Finally, [Akbari et al., 2016] proposed an optimization approach which learns the wellness profiles of users concerning a taxonomy of wellness events. The framework utilizes content information of Twitter tweets as well as the relation between event categories to extract personal wellness events (i.e. doing exercise, having a meal, etc.) from users' timelines. In particular, the authors modeled the inter-relatedness among distinct events as a graph Laplacian, which was employed as a regularizer in the learning process.

4. Group Profiling

Group profiling aims at learning the collective behavior of users groups that are also known as user communities in social media computing. An intuitive approach to perform user profiling at a group level is to discover communities of users and then apply profile learning approach to capture information and behaviors of the community members [Zhao et al., 2013]. In other words, community profiling often modeled as a two-stage framework including community discovery problem and aggregation of members' personal profiles. Group user profiling is necessary for various applications including facility planning, personalized recommendation [Zhao et al., 2013; Farseev et al., 2014; Farseev et al., 2015a; Farseev et al., 2016], and marketing [Qu and Zhang, 2013].

4.1. User Community Detection

As noted above, it is important to find representative user communities to leverage on a group knowledge for various applications. Such task is commonly approached by modeling users' relationship as a graph so that dense subgraphs of such graph can be treated as user communities. The graph can be constructed in a several ways: (a) based on *social connections between users* (i.e. follower/followee relationship) [Fortunato, 2010] that are often hidden behind the privacy settings; (b) based on *user generated content*, when the edges of the graph are weighted as a distance between data representations of users (i.e. cosine or Euclidian distance) [Farseev et al., 2014]; or (c) as a *combination of the above two methods*.

After the graph G is constructed, it is important to define — what exactly "user community" means. Traditionally, it has been modeled as a MinCut problem [Von Luxburg, 2007], where for a given number k of subsets, the MinCut, essentially chooses a partition $C_1, ..., C_k$ such that it minimizes the expression: $cut(C_1, ..., C_k) = \sum_{i=1}^{k} W(C_i, \overline{C}_i)$, where $\overline{C}_i$ stands for a compliment of C_i, and $W(A, B) = \sum_{u \in A, v \in B} \rho(u, v)$, and ρ is a distance function. Such a formulation allows us to find k user communities $C_1, ..., C_k$, that are grouped according to some criteria (defined by distance function ρ). The two most

commonly used definitions of the MinCut problem are RatioCut and the so-called NCut [Von Luxburg, 2007]. The idea behind RatioCut is based upon an assumption that the resulting communities could have similar size, while NCut formulation conditions the sum of edges' weights in each community to be minimized among all communities. Though both RatioCut and NCut are NP-Hard [Von Luxburg, 2007], there are many approximate solutions exist for the problem above. One of the existing MinCut approximations is the spectral clustering approach, which is defined as a standard trace minimization problem: $\min_{U \in R^{n \times k}} tr(U^\intercal L_{sym} U), s.t. U^\intercal U = I$. According to the Rayleigh-Ritz theorem, the spectral clustering optimization problem can be solved as the first k eigenvectors of the normalized graph Laplacian $L_{sym} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where W is the adjacency matrix, and D is the degree matrix of the graph G. It should be noted that in cases when data comes from multiple sources, the multi-layer generalization of spectral clustering and its variations can be used: $\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} tr(U^\intercal L_i U), s.t. U^\intercal U = I$.

In addition to the techniques above, other solutions of cross-source (cross-domain) clustering problem can be utilized. For example, in [Farseev et al., 2014], the authors compared different conventional clustering techniques to solve the cross-domain venue category recommendation task. Specifically, they detected user communities based on data from Twitter and used the obtained communities to perform cross-domain recommendation of Foursquare venue categories. The technique allows them to perform recommendation even to those users, who did not perform a Foursquare activity, which naturally solves the Cold Start problem. At the same time, Zhao et al. [Zhao et al., 2013] proposed an approach to perform multi-modal venue recommendation based on regularized Modularity Maximization clustering. The same authors subsequently proposed another solution of the same problem, which regularizes the Matrix Factorization approach [Zhao et al., 2015]. Finally, the cross-source community detection approach was proposed in [Farseev et al., 2016], where the multi-source user relations were modeled as a multi-layer graph, and the community detection was implemented as a regularized spectral clustering. Specifically, authors developed the multi-source user community detection algorithm, which utilizes both inter-source relationship and sources' ability to complement each other via efficient clustering guidance, based on the automatically-inferred social networks relationship graph structure.

5. Future Trends

While user profiling at the individual and group level has already attracted large research interests, many challenges remain to be addressed for efficient and effective user profile learning. First, there are *various rich data sources emerging*, such as data from *Wearable Sensors*, *External Sensors, and Smart Devices*. The high availability of such data supports the "Internet of Things" concept, and thus *shrink the gap between offline and online worlds*. User Profile Learning from such data together with data from Social Multimedia opens the opportunity to learn a user profile in 360° perspective and thus has the potential to improve the quality of users' daily life. However, such *data sources are often completely different in nature* as compared to the conventional ones. Moreover, the *temporal aspect assumes the crucial role due to the high velocity of incoming data and data samples interdependence*. All these brings new challenges into Social Multimedia Computing. Second,

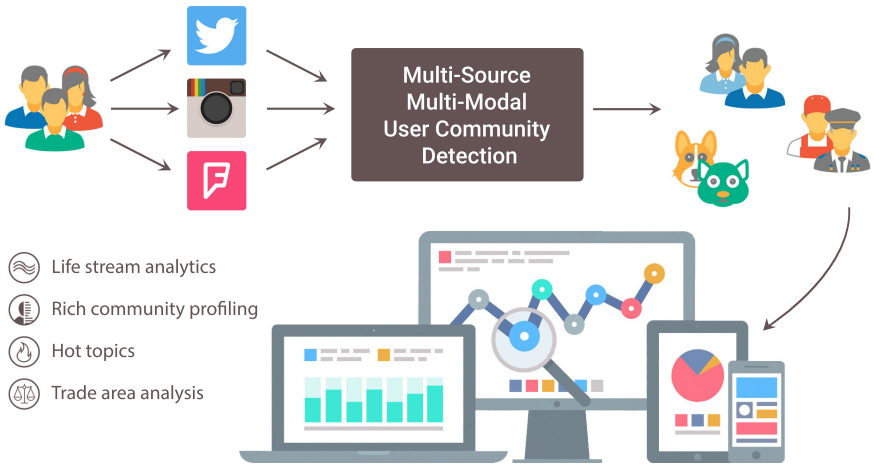

Fig. 1: Architecture of the bBridge Big Data Analytics Platform

most existing approaches to a group and individual profiling utilize either network topology [Fortunato, 2010] or user generated content [Farseev et al., 2015a]. However, both of these data sources are often insufficient for finding proper personal user profiles and communities due to extreme sparsity and high noise level in social media data. As users can perform several activities in social media (i.e., posting, liking, replying, etc.) *aggregating all the available information is essential to model user interests better at both individual and group level*. Third, in many real-world scenarios, some prior knowledge about the community affiliations of users might be available. For example, two papers which have been published in the same venue are often united by the same research field. Although, they were not linked or cite each other. The similar situation may happen in social networks, when very similar (interest-vice) users, indeed are not connected to each other via friendship relationship or explicit community structure. However, some prior knowledge may guide the community discovery process: in the first case, it can be publication venues graph, while in the second case, the users' interests distribution. However, *incorporating domain knowledge into learning and, especially, community detection process pose a great challenge*. Last but not least, most existing community detection techniques fail to provide any rationale and insight about the formation of obtained communities as well as the collective behavior of members. *Deriving a community profile from its members' personal profiles and community members' generated data is a challenging task.*

6. Application

To demonstrate the applicability of group and individual user profiling, many of the above approaches and our recently developed techniques are integrated into a big social multimedia data analytics platform named "bBridge"[8]. bBridge offers a Real-Time Group Analytics to Business and Public Sector Users. The provided features are Trade Area Analysis [Qu and Zhang, 2013], User Communities Detection and Profiling [Farseev et al., 2014; Farseev et al., 2016], Live Social Media Stream Analytics, Hot Topics Extrac-

[8]bBridge: The Big Social Multimedia Analytics Platform `bbridge.net`

tion [Vorontsov et al., 2015], and Brand Monitoring. bBridge is a useful tool for both businesses and social sectors. The architecture of bBridge is presented on a Figure 1. To support the declared functionality, the platform, first, detects user communities based on novel regularized spectral clustering approach [Farseev et al., 2016] that can perform an efficient partitioning of multi-layer user relations graph, which is constructed from heterogeneous social media streams: Facebook, Foursquare, Instagram, Twitter, and Mobility Tracks. The clustering regularization considers both inter-layer relatedness and its ability to complement each other, which smoothen the latent representation and allow for the detection of meaningful multi-source user communities. Second, bBridge profiles the detected user communities using the predicted attributes such as users' demographics [Farseev et al., 2015b], interests [Song et al., 2015b], and social status. The prediction is based on a novel Multi-Source learning framework [Farseev et al., 2015b]. Third, the platform builds the semantic profile of each community based upon multi-modal Topic Modeling framework BigARTM [Vorontsov et al., 2015]. Finally, bBridge presents the live stream analytics to business users at the group level, while providing personal-level analytics and interactive services at an individual level to mobile users. To the best of our knowledge, bBridge is one of the first full-stack solutions that offers multi-source multi-modal Big Data Analytics at both personal and group levels.

7. Conclusion

In this paper, we reviewed existing user profile learning approaches at both individual and group levels. Particularly, we went through the pipeline of user profile learning, the associated challenges along with its possible solutions on each step: data gathering, data representation, personal and group user profiling. We then discussed potential future trends in the research field, followed by introducing a big multimedia data analytics platform named "bBridge". The platform serves both business and personal objectives. We encourage further research on user profile learning by contributing our multi-source multi-modal dataset NUS-MSS[9], and believe that our common efforts will help to make further steps towards 360° user profiling.

ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

The Microsoft Windows Azure Cloud and Microsoft MSDN subscription were provided by "MB-Guide"[10] and "bBridge"[11] projects, as part of Microsoft BizSpark program.

REFERENCES

Akbari, M., Huc, X., Liqianga, N., and Chua, T.-S. (2016). From tweets to wellness: Wellness event detection from twitter streams.

[9] lms.comp.nus.edu.sg/research/NUS-MULTISOURCE.htm
[10] mb-guide.com
[11] bbridge.net

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

Chen, Y., Zhuang, C., Cao, Q., and Hui, P. (2014). Understanding cross-site linking in online social networks. In *Proceedings of the Workshop on Social Network Mining and Analysis*.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y.-T. (July 8-10, 2009). Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE.

Farseev, A., Kotkov, D., Semenov, A., Veijalainen, J., and Chua, T.-S. (2015a). Cross-social network collaborative recommendation. In *Proceedings of the ACM International Conference on Web Science*. ACM.

Farseev, A., Nie, L., Akbari, M., and Chua, T.-S. (2015b). Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM.

Farseev, A., Samborskii, I., Filchenkov, A., and Chua, T.-S. (2016). Cross-source community-based collaborative recommendation. In *Proceedings of the 25th International Conference on Information and Knowledge Management (To Appear)*.

Farseev, A., Zhukov, N., Gossoudarev, I., and Zarichnyak, U. (2014). Cross-platform venue recommendation based upon user community detection from social media.

Fortunato, S. (2010). *Community detection in graphs*. Physics Reports.

Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on Computer vision*. Ieee.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on*. IEEE.

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

Qu, Y. and Zhang, J. (2013). Trade area analysis using user generated mobile location data. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Song, X., Nie, L., Zhang, L., Akbari, M., and Chua, T.-S. (2015a). Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Song, X., Nie, L., Zhang, L., Liu, M., and Chua, T.-S. (2015b). Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. ACM.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*.

Vorontsov, K., Frei, O., Apishev, Murat, R., Suvorova, M., and Yanina, A. (2015). Non-bayesian additive regularization for multimodal topic modeling of large collections. In *In Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*.

Vosecky, J., Hong, D., and Shen, V. Y. (2009). User identification across multiple social networks. In *Proceedings of the First International Conference on Networked Digital Technologies (NDT)*. IEEE.

Wang, X., Zhao, Y.-L., Nie, L., Gao, Y., Nie, W., Zha, Z.-J., and Chua, T.-S. (2015). Semantic-based location recommendation with multimodal venue semantics. *Multimedia, IEEE Transactions on*.

Xiong, L., Chen, X., Huang, T.-K., Schneider, J. G., and Carbonell, J. G. (2010). Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *SDM*, volume 10, pages 211–222. SIAM.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., and Ye, J. (2012). Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Proceedings of the 18th International conference on Knowledge discovery and data mining (SIGKDD)*.

Zhang, Y., Wang, L., Li, X., and Xiao, C. (2016). Social identity link across incomplete social information sources using anchor link expansion. In *Advances in Knowledge Discovery and Data Mining*. Springer.

Zhao, Y.-L., Chen, Q., Yan, S., Chua, T.-S., and Zhang, D. (2013). Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Zhao, Z., Cheng, Z., Hong, L., and Chi, E. H. (2015). Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1406–1416. International World Wide Web Conferences Steering Committee.

Aleksandr Farseev is the PhD Candidate at the School of Computing, National University of Singapore. His main research interest is in area of Social Media Analysis. In particular, his research focuses on the multi-source user profile learning arising from the Web and social networks at both group and individual levels.

Mohammad Akbari is the PhD Candidate at the School of Computing, National University of Singapore. Ivan Samborskii is the Master Student at the Department of Computer Technologies, ITMO University, Saint-Petersburg, Russia. Dr. Chua Tat-Seng is the KITHCT Chair Professor at the School of Computing, National University of Singapore. His main research interest is in Multimedia Information Retrieval and Social Media Analysis.