bBridge: A Big Data Platform for Social Multimedia Analytics

Aleksandr Farseev*, Ivan Samborskii**, and Tat-Seng Chua*
farseev@u.nus.edu; samborskiy.ivan@gmail.com; chuats@nus.edu.sg
*School of Computing, National University of Singapore; **Computer Technologies Department, ITMO University

ABSTRACT

In this technical demonstration, we propose a cloud-based Big Data Platform for Social Multimedia Analytics called bBridge [9] that automatically detects and profiles meaningful user communities in a specified geographical region, followed by rich analytics on communities' multimedia streams. The system executes a community detection approach that considers the ability of social networks to complement each other during the process of latent representation learning, while the community profiling is implemented based on the state-of-the-art multi-modal latent topic modeling and personal user profiling techniques. The stream analytics is performed via cloud-based stream analytics engine, while the multi-source data crawler deployed as a distributed cloud jobs. Overall, the bBridge platform integrates all the above techniques to serve both business and personal objectives.

Keywords

Multiple Sources Integration; Community Detection; User Profiling; Stream Analytics; Cloud computing

1. INTRODUCTION

The past decade has recorded a rapid development and change in the 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. 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 *information filtering*, due to the growing amount of irrelevant content, such as advertisement and spam. At the data analytics level, *data-preprocessing, fusion, and modeling* pose the greatest challenges to address. A comprehensive solution of these problems provide users with meaningful social-multimedia insights at both the individual and group levels [6]. For example, the group-level analytics can be utilized in Facilities Planning, Brand Monitoring, Marketing, and Geo-Enabled Advertisement domains, while the personal analytics, can be used for Personal Content Filtering and Interest-Based Recommendation.

In our previous works, we proposed the frameworks for venue category recommendation via user community detection [11], and multi-source user profile learning [6, 8]. However, the adopted user community detection techniques [11] cannot be directly applied to our current problem of multi-source community detection due to the data representation and multi-source integration issues. Thus, for this demonstration, we enriched the previously utilized techniques by implementing the so-called multi-source regularized spectral clustering [5]. Furthermore, to perform user community profiling, we exploited the multi-modal topic modeling [16] and our earlier developed multi-source user profiling [8] techniques. Eventually, all the above approaches were integrated into the Big Social Multimedia Analytics Platform bBridge, which delivers the live stream analytics to its users.

The overall data processing pipeline in bBridge is as follows: (a) the distributed cloud crawlers harvest recently posted UGC from multiple social networks [6]; (b) the community detection framework extracts user communities in a latent cross-social network space [5, 11]; (c) multi-source user profile learning [8] and multi-modal topic modeling [16] techniques construct comprehensive profiles of each detected user community and prepare it for further visualization; and (d) the stream analytics engine aggregates communities' multimedia UGC streams [1], which enables various T-SQL queries to be performed into live multimedia data [3, 13]. To the best of our knowledge, the bBridge is one of the first online solutions that offers Big Social Multimedia Analytics at both personal and group levels.

2. BBRIDGE

2.1 Platform Description

We now briefly describe the bBridge platform. The details of the adopted community detection approach were elaborated in previous works [5, 11]; the utilized user profile learning techniques were also described in our previous studies [8, 6]; while the multi-modal topic learning framework was outlined by Vorontsov et al. [16]. The block diagram of the bBridge platform is shown in Figure 1.

In the bBridge cloud platform, the user communities and its' corresponding profiles are re-computed every 12 hours, while the stream analytics and multi-source crawling are executed continuously. At any given time frame, the following

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

ACM ISBN 978-1-4503-3603-1/16/10..

DOI: http://dx.doi.org/10.1145/2964284.2973836

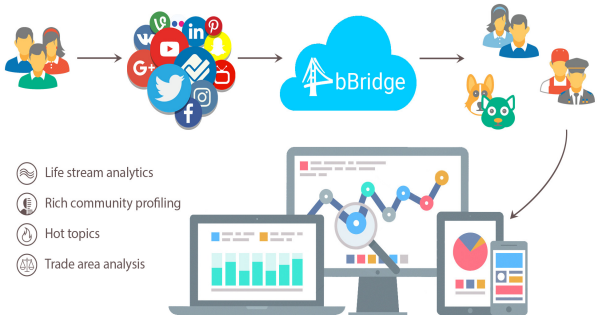

Figure 1: Block diagram of the bBridge platform

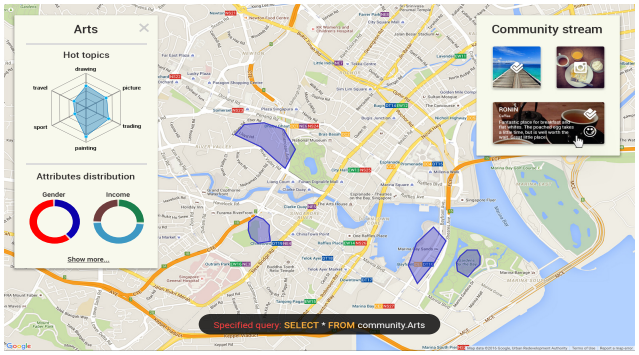

Figure 2: The bBridge online GUI snapshot

steps are taken (the continuously executing operations are marked with the "*"):

I. Multi-Source Distributed Crawling*: The crawling process consists of two steps: (a) a set of active users who have recently posted tweets through the cross-linking functionality [8] of multimedia social networks is collected; and (b) Twitter Stream APIs [15] are used to perform 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. publish Instagram images on Twitter). By following the described user identification approach, we continuously harvest the data from Twitter, Instagram, Foursquare, and Facebook. In the current technical demonstration setup, we continuously monitor about 170 thousand Twitter users from Singapore, which were active (post cross-linking tweets) near Singapore region during the recent one month time frame.

II. Multi-Modal Feature Extraction: We represented data as follows: visual data (Instagram and Twitter images) is represented in a form of image concept distribution [4]; location data (Foursquare check-ins) is represented in a form of distribution among venue categories and mobility patterns [6, 8]; textual data (Twitter tweets, Instagram image captions, and Foursquare check-in comments) is represented in a form of latent topics [16], LIWC distribution [14] and writing style features [8]. The detailed description of the adopted feature extraction processes is given in Farseev et al. [8].

III. Multi-Source Community Detection: The user communities are detected via regularized spectral clustering [5] on a multi-layer graph. Each layer in the graph is constructed based on UGC similarity between users, which is estimated as a cosine distance between data representations in each social network [6].

IV. User Communities Profiling: User Community Profiling is performed in three steps: (a) individual user profile learning [6] is applied to each community member, which results in a set of personal users' attributes, such as Age, Gender, Occupation Industry, Education Level, Income Level, Relationship Status, etc.; (b) the extracted attributes are summarized among all community members for further visualization in a form of "Pie" diagrams; and (c) top K (where $K = 6$) hot multi-modal latent topics [16] are detected for each user community for further visualization in a form of "Radar" diagrams.

V. Live Stream Analytics*: Except for sentiment analysis for every post [8], the live stream analytics is performed in a form of T-SQL queries [3] into a real-time social multimedia stream. bBridge adopts the Windows Azure Stream Analytics cloud service [1], which allows for using various query constructions such as "Like/Not Like patterns", "CASE statements", "Embedded procedures", etc. A typical query into bBridge Platform could be expressed as below:

```
SELECT Topics, Content FROM MultimediaStream
TIMESTAMP BY CreatedAt GROUP BY SLIDINGWINDOW(s, 30)
WHERE Sentiment='Positive' AND Community='Arts'
AND Region='Downtown'
```

The above query will output distribution among extracted latent topics, and content of all the multimedia posts published from "Downtown" district during the last 30 seconds by the "Arts" community with the "Positive" sentiment score. From the example, it can be seen that the aforementioned querying approach can be used to query multimedia data at both global and local levels.

2.2 Architecture

To take advantage of parallel processing, we have implemented bBridge platform as a multi-role cloud service [2]. Each crawler runs as a separate Windows Azure worker role, which enables us to harvest multiple social multimedia venues simultaneously. The community detection and profiling modules are implemented as multi-instance Windows Azure worker roles, so that it can be performed for different time frames at the same time. The stream analytics module utilizes Azure Stream Analytics engine [1] that can perform complex real-time queries [3] into the live social multimedia stream. All worker roles and stream engine are synchronized via the Windows Azure Service Bus.

3. THE DEMONSTRATION

Using our web GUI [10], bBridge users can select one of the recently detected communities in a specified geographical region. Based on the selected user community, the following features are provided: (a) observe geo-activity of the community members; (b) discover hot multi-modal topics that were discussed by the community members; (c) view the distribution of the community members among automatically inferred user attributes (such as age, gender, occupation industry, education level, relationship status, and income level); and (d) perform queries of different complexity into the live social multimedia stream, which is generated by the community members.

A snapshot of the GUI is shown in Figure 2. The UI shows the map of a Singapore region and the opened visualization console for the "Arts" community. The map shows

Table 1: User Profiling Results in Terms of Macro Accuracy

Attribute	Age Gr.	Gender	Rel.	Edu.	Occ.
Macro Acc.	0.509	0.878	0.659	0.689	0.107

the polygons that describe recent community members' geo-activity; the left pane displays the hot topics, discussed by the community members and the distribution of the community members among the detected personal user attributes; while the right pane lists the recent cross-network posts of the community members, which are filtered according to the specified query. bBridge users can select different user communities, view visualization, specify real-time queries into the community stream via the query UI control.

4. EVALUATION

To quantify the efficiency of the core technologies behind the bBridge platform, we utilize a subset of Multi-Source Cross-Region Social Dataset NUS-MSS [8] (Singapore region) for testing user community detection and personal user profiling performance.

As in Farseev et al. [7], the user community detection is implicitly evaluated via performing venue category recommendation, where the user-based collaborative filtering is only applied to users from the same pre-detected user community [11]. Multi-source user communities were detected on a multi-layer users' relatedness graph (constructed based on user-generated data similarities for each data source) via the Grassmann subspace analysis clustering [5]. The venue category recommendation was completed via our previously proposed community constrained collaborative filtering approach [11]. The user community detection results are presented in Figure 2. It can be seen that the user-based collaborative filtering, which is performed over pre-detected user communities, can achieve higher recommendation performance as compared to other multi-source recommendation approaches. This suggests that the bBridge user community detection approach can detect relevant and meaningful user communities from multiple social media sources.

The personal user profiling performance is explicitly evaluated by predicting the following individual user attributes: Age Group (<20, $20-30$, $30-40$, >40), Gender (Male/Female), Education Level (School/University), Occupation Industry (14 classes from NUS-MSS [8] test set), and Relationship Status (In a Relationship/Single). To do so, we utilized our previously proposed multi-source ensemble learning approach [8], which can handle imbalanced multi-source data for the inference purposes. The personal user profiling performance results are presented in Table 1. It can be seen, that the Age Group, Gender, Relationship Status, and Education Level attributes can be inferred with the sufficiently high Macro Recall (Macro Accuracy) [8], while the Occupation Industry inference [12] requires further elaboration. The above suggests that the automatically inferred personal user attributes can be used for visualization of the detected user communities in an aggregated form.

5. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Center @ Singapore Funding Initiative and administered by the IDM Programme Office. This research is partially supported by the Government of Russian Federation, Grant 074-U01.

We thank Prof. Konstantin Vorontsov (MIPT, Russia), Ms. Nadezhda Chirkova, Mr. Arseniy Pendryak, Mr. Maksym Sihida, and Mr. Alexei Vasilyev for assistance with demo evaluation and implementation.

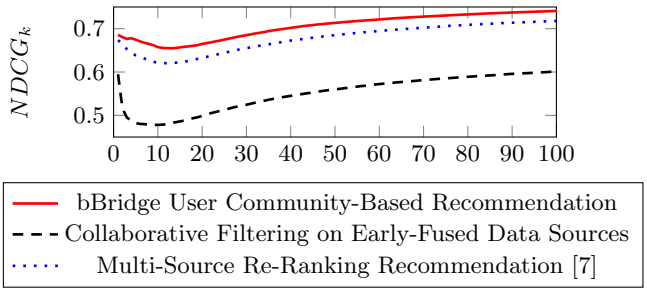

Figure 3: User Community Detection Results (via Venue Category Recommendation) in Terms of $NDCG_k$

6. REFERENCES

[1] M. Azure. Microsoft azure stream analytics service. https://azure.microsoft.com/en-us/services/stream-analytics/. Online; accessed 3 Aug 2016.

[2] M. Azure. Microsoft cloud services overview. https://azure.microsoft.com/en-us/documentation/articles/cloud-services-choose-me/. Online; accessed 3 Aug 2016.

[3] M. Azure. Stream analytics query language reference. https://msdn.microsoft.com/en-us/library/azure/dn834998.aspx. Online; accessed 3 Aug 2016.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009*. IEEE, 2009.

[5] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov. Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. *IEEE Transactions on Signal Processing*, 2014.

[6] A. Farseev, M. Akbari, I. Samborskii, and T.-S. Chua. 360° user profiling: past, future, and applications. *ACM SIGWEB Newsletter*, (Summer), 2016.

[7] A. Farseev, D. Kotkov, A. Semenov, J. Veijalainen, and T.-S. Chua. Cross-social network collaborative recommendation. *Proceedings of the 7th ACM International Conference on Web Science*, 2015.

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

[9] A. Farseev, I. Samborskii, and T.-S. Chua. bbridge: A big social multimedia analytics platform. https://bbridge.net. Online; accessed 3 Aug 2016.

[10] A. Farseev, I. Samborskii, and T.-S. Chua. bbridge: A big social multimedia analytics platform demo. https://demo.bbridge.net. Online; accessed 3 Aug 2016.

[11] A. Farseev, N. Zhukov, I. Gossoudarev, and U. Zarichnyak. Cross-platform online venue and user community recommendation based upon social networks data mining. *Computer Tools in Education*, 6, 2014.

[12] L. Nie, L. Zhang, M. Wang, R. Hong, A. Farseev, and T.-S. Chua. Learning user attributes via mobile social multimedia analytics. *ACM Transactions on Intelligent Systems and Technology*, 8, 2016.

[13] I. Nikolaidis, K. Iniewski, B. Malhotra, H. Jeung, T. Kister, S. Bressan, and K.-L. Tan. Maritime data management and analytics. In *Building Sensor Networks: From Design to Applications*. CRC Press, 2013.

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

[15] Twitter. The streaming apis. https://dev.twitter.com/streaming/overview. Online; accessed 3 Aug 2016.

[16] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the Workshop on Topic Models: Post-Processing and Applications*. ACM, 2015.