This is the final draft of the following article:
Aleksandr Farseev, Denis Kotkov, Alexander Semenov, Jari Veijalainen, and Tat-Seng Chua. Cross-social network collaborative recommendation. In *Proceedings of the ACM Web Science Conference*, pages 38:1-38:2. ACM, 2015. doi: 10.1145/2786451.2786504

Cross-Social Network Collaborative Recommendation

Aleksandr Farseev*, Denis Kotkov**, Alexander Semenov**,
Jari Veijalainen** and Tat-Seng Chua*
farseev@u.nus.edu; deigkotk@student.jyu.fi; {alexander.v.semenov, jari.veijalainen}@jyu.fi; chuats@nus.edu.sg
* National University of Singapore, School of Computing, COM 1, 13 Computing Drive, Singapore;
** University of Jyvaskyla, Dept. of Comput. Sci. & Inf. Syst., , P.O.Box 35, FI-40014, Finland

ABSTRACT

Online social networks have become an essential part of our daily life, and an increasing number of users are using multiple online social networks simultaneously. We hypothesize that the integration of data from multiple social networks could boost the performance of recommender systems. In our study, we perform cross-social network collaborative recommendation and show that fusing multi-source data enables us to achieve higher recommendation performance as compared to various single-source baselines.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

1. INTRODUCTION

The fast expanding multi-modal social media data serves as an important resource to perform relevant recommendation and comprehensive user profiling in many application domains. In particular, venue category recommendation (e.g. restaurant, museum, or park) is an important task in tourism and advertisement for suggesting interesting venues near users' current location. At the same time, most internet-active adults prefer to use multiple social services simultaneously to satisfy their different information needs[1], and thus, interests of such users can be better inferred from different perspectives using multiple online social networks (OSNs).

Up to now, only a few studies investigated multi-source data processing, while the usefulness of the multi-source data integration for venue recommendation remains unclear. For example, according to [1, 4], multi-source data integration may help to achieve higher recommendation performance. However, there has not been much research done on multi-source multi-modal recommender systems, while most existing works focus on either multi-source or multi-modal data.

In this paper, we investigate multi-source and multi-modal recommender systems and report the results of an initial experimental study. We seek to address the following research question: *is it possible to improve the recommendation performance by incorporating multi-source multi-modal data?*

To address this question, we employed a subset of NUS-MSS [2] dataset that includes user-generated content posted by the same users in three popular OSNs: Twitter, Foursquare and Instagram. We made the assumption that the social media data from the same user on different social network presents users activities in different perspectives, which are correlated to overall user profile. We therefore built a recommender system that exploits the multi-source multi-modal data and suggests categories of venues visited and unvisited by a particular user. Since location recommendation is especially useful when it is based on the user's current location, we recommend venue categories. It allows an individual to immediately discover the newly suggested places. The task can also be considered as user profiling for content and location-based recommendation. Furthermore, we also address the source integration problem and evaluate our proposed approach against single-source baselines.

2. MULTI-SOURCE DATA

To address the research question, we employed a subset of NUS-MSS dataset [2] by considering only those users who performed activities on each of the following three OSNs: Twitter, Foursquare and Instagram, from 10 July till 29 September 2014 in Singapore region. The resulting dataset used for the experimentation contains 3,058,833 tweets, 81,755 check-ins and 87,672 Instagram posts, generated by 4,172 users.

2.1 Feature extraction

We model the users as vectors in m-dimensional feature space: $u_{i,F} = (f_{i,1}, f_{i,2}, ..., f_{i,m})$, where f_k is the kth feature for user $u_{i,f}$ using features space F. Overall, we leverage 4 types of features: Linguistic Inquiry and Word Count (LIWC) from Twitter, Latent Dirichlet Allocation (LDA) from Twitter, Instagram image concepts and Foursquare venue categories[2]. Each user is, thus, represented by 4 feature vectors:

- $u_{i,LIWC} = (liwc_{i,1}, liwc_{i,2}, ..., liwc_{i,70})$, where $liwc_{i,k}$ is a normalized LIWC feature;
- $u_{i,LDA} = (lda_{i,1}, lda_{i,2}, ..., lda_{i,50})$, where $lda_{i,k}$ is the probability that the tweets of user u_i are about topic k;

[1]According Pew Research Internet Project's (www.pewinternet.org) Social Media Update 2013

[2]Detailed features description is given in [2]

WebSci'15 June 28 - July 01, 2015, Oxford, United Kingdom
©2015 ACM. ISBN 978-1-4503-3672-7/15/06...$15.00
DOI: http://dx.doi.org/10.1145/2786451.2786504.

- $u_{i,IN} = (in_{i,1}, in_{i,2}, ..., in_{i,1000})$, where $in_{i,k}$ is a normalized number of pictures posted by user i with a concept in_k;
- $u_{i,SQ} = (sq_{i,1}, sq_{i,2}, ..., sq_{i,546})$, where $sq_{i,k}$ is a normalized number of the times user i visited venue category sq_k.

3. RECOMMENDATION APPROACHES

We recommended venue categories based on each data source independently and based on the fusion of data from multiple sources.

3.1 Single-source recommendation

To recommend venue categories, we implemented user-based collaborative filtering. In the case of single-source recommendation, the list of suggested categories is sorted according to the ratings of items. Rating of each item for each user is calculated as follows:

$$\hat{r}(u_{i,F}, j) = \frac{\sum_{k \in U, k \neq i} sq_{k,j} \cdot cos(u_{i,F}, u_{k,F})}{\sum_{k \in U, k \neq i} cos(u_{i,F}, u_{k,F})}, \quad (1)$$

where $\hat{r}(u_{i,F}, j)$ is the rating calculated using feature vector $u_{i,F}$ such as $u_{i,LDA}$, $u_{i,LIWC}$, $u_{i,SQ}$ or $u_{i,IN}$; $u_{i,F}$ is a target user that receives a recommendation list with item j; $sq_{k,j}$ is the weight of item j for user k; and $cos(u_{i,F}, u_{k,F})$ is the *cosine similarity measure* between users i and k.

3.2 Multi-source recommendation

We performed different fusion approaches at the different stages of collaborative filtering.

As a simple baseline, we first employed an **early fusion** approach [3] to fuse multi-source data, where features derived from each source were concatenated into a single feature vector:

$$u'_i = (sq_{i,1}, ..., sq_{i,546}, lda_{i,1}, ..., lda_{i,50}, \\ liwc_{i,1}, ..., liwc_{i,70}, in_{i,1}, ..., in_{i,1000}). \quad (2)$$

Seeking to boost the recommendation performance, we developed a new **late fusion re-ranking** approach. We linearly combined the outputs from different sources, where the weight of each source is learned based on a stochastic hill climbing with random restart (SHCR) optimization algorithm. Each source is assigned a real-valued weight of between 0 to 1 and the rank of *ith* item in the final recommendation output is computed as follows:

$$Rank_f(item_i) = \frac{1}{n} \sum_{s=1}^{n} \frac{w_s}{Rank_s(item_i)}, \quad (3)$$

where $Rank_s(item_i)$ is the rank of *ith* item in recommendation list for source s; w_s corresponds to the weight of the source s; n is a total number of sources (in our case, $n = 4$). The venue categories in final recommendation list are sorted in increasing order according to their rank. We optimize the multi-source recommendation performance (measured by $F - measure@10$) in 1000 SHCR iterations that gives a good chance to find reasonable source weights.

4. EXPERIMENTAL RESULTS

In our experiments, the recommender system suggests a sorted number of categories to each user. We trained the model using the whole dataset and evaluated based on 736 users who have checked-in in at least 20 categories in the training set and 8 categories in the test set.

To measure the recommendation performance we use $F - measure@K = \frac{2 \cdot P@K \cdot R@K}{P@K + R@K}$, where $P@K$ and $R@K$ are precision and recall at K, respectively, and K indicates the number of selected items from the top of the recommendation list.

Figure 1 demonstrates that multi-source multi-modal data fusion helps to improve the recommendation performance. Specifically, the proposed late fusion re-ranking approach outperforms all the baselines starting from $K > 6$.

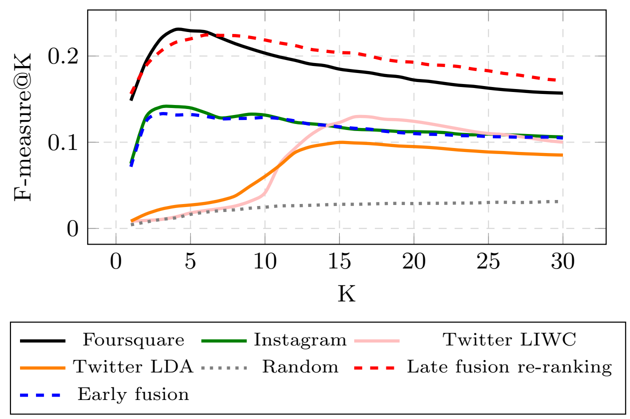

Figure 1: The evaluation of recommendation performance in terms of F-measure@K

The lower recommendation performance of multi-source approach for $K < 6$ could be explained by the significant noise level in Twitter data, while for $K > 9$ Twitter features are able to archive better generalization ability of the proposed model.

Another observation is the failure of the early fused model to improve the recommendation performance as compared to single source baselines, where the shape of the performance curve is similar to that of the Instagram approach. The possible reason is the unbalanced sparsity of different feature vectors, since features were derived from different multi-modal sources.

5. CONCLUSION

In this study, we investigated the impact of multi-modal data from different social media sources on the recommendation performance. Based on the NUS-MSS dataset [2], we incorporated multi-source multi-modal data and compared its performance with single-source baselines. Our results indicated that the fusion of multi-source multi-modal data is able to boost the recommendation performance significantly.

Our future work includes the development of new efficient source fusion solutions. Also we plan to work on new feature types, data completion techniques and multi-source recommendation models.

6. ACKNOWLEDGMENTS

The research in MineSocMed project was partially funded by the Academy of Finland, Grant #268078.

The research is also 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"[3] project and as part of Microsoft BizSpark program.

7. REFERENCES

[1] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *User Model. User-Adap. Inter.*, 23(2-3):169–209, 2013.

[2] A. Farseev, N. Liqiang, M. Akbari, and T.-S. Chua. Harvesting multiple sources for user profile learning: a big data study. In *ICMR*, 2015.

[3] P. Winoto and T. Tang. If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations. *New Generation Computing*, 26(3):209–225, 2008.

[4] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo, and V. S. Tseng. Urban Point-of-interest Recommendation by Mining User Check-in Behaviors. In *SIGKDD Workshop on Urban Computing*, 2012.

[3] mb-guide.com