The Hong Kong Polytechnic University
7 December 2016

Domain-specific user preference prediction based on multiple user activities

Author: YUNFEI LONG, **Qin Lu**,Yue Xiao, MingLei Li, Chu-Ren Huang.

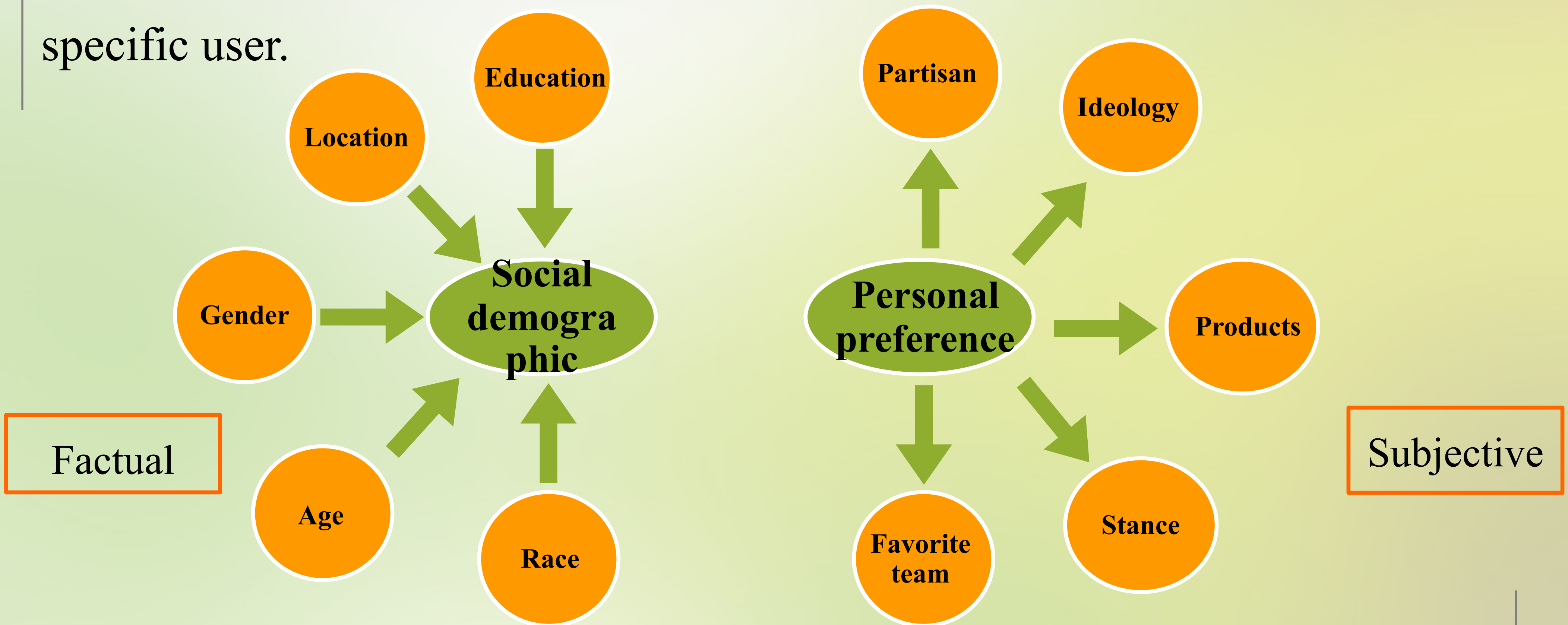
www.comp.polyu.edu.hk/

Dept. of Computing, The Hong Kong Polytechnic University

Outline

Introduction

- **What is User profile?** A visual display of personal data associated with a specific user.



- **Why acquiring user profiles?**
  - Personalized recommendation
  - Personalized opinion/emotion
  - Prediction of stances

Current Problems

- **How to acquire user profiles from the web?**
  - **Explicit: Structured profile components in web pages.**
  - **Implicit: User activities (Posted text, Social network and Interested topics).**
- **The Challenges:**
  - Structured information is sparsely available: use of unstructured data.
  - Hard to explore multiple components of user activities at the same time: Proposed an integrated framework.
  - Lack of user activity and user profile data: Build dataset for benchmarking/experiments.

Related Works

- **Previous Methods for User profile prediction:**
  - Linear classification learning based: Feature + classifier (Rao and Yarowsky. 2010)
    - Features: BoW, POS, excitement, social linguistic (agreement, abbreviation, and punctuation…)
    - Classifiers: SVM, Logistic regression, Naïve Bayes, etc.
    - Need labor intensive feature engineering.
      - *Features especially social linguistic related features need professional designing.*
    - Hard to incorporate non-text features.
    - Mostly in user social demographic detection: Gender, Age, Race…(Rao and Yarowsky. 2010)

Our Objectives

- Build a user profile data include three parts of user activities: **user posted comments**, **user social network** and **user interested topics**.
- Build an **integrated model** to learn **user preference** from three part of activities.
- **Solution based on two premises**
  - Homophily theory: similar individuals have similar preferences.
  - Embedding theory: similar users are represented by similar vectors if they are making similar comments, having similar followers, or sharing similar interested topics.

Corpus Construction

- **Data Source:**
  - **Hupu**(虎扑) basketball discussion forum . (http://cba.hupu.com/)
  - All discussing threads from March 2012 to April 2016.

Statistics	Number
Users	17011
Comments	423758
Connected users	76447
Topics/threads	38455

Statistic of collected Hupu corpus

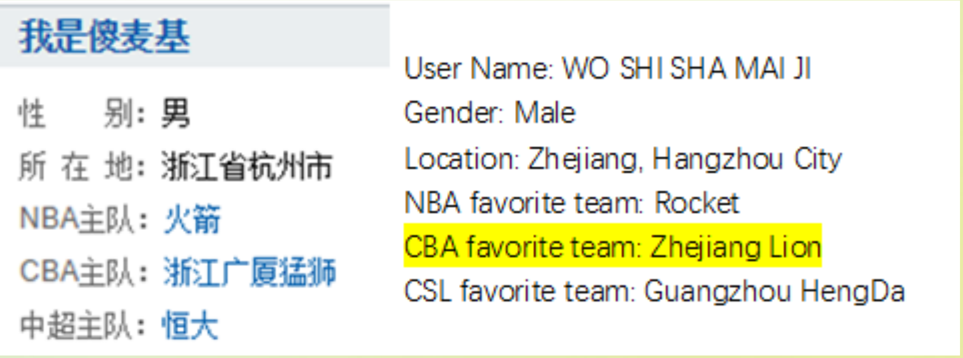
Model Name	Min	Max	Average
Comments	1	1747	25
Friends	1	1500	69
Topics	1	742	17

User activity information

- Reflect the unbalanced activities problem in social media data.

Corpus Construction

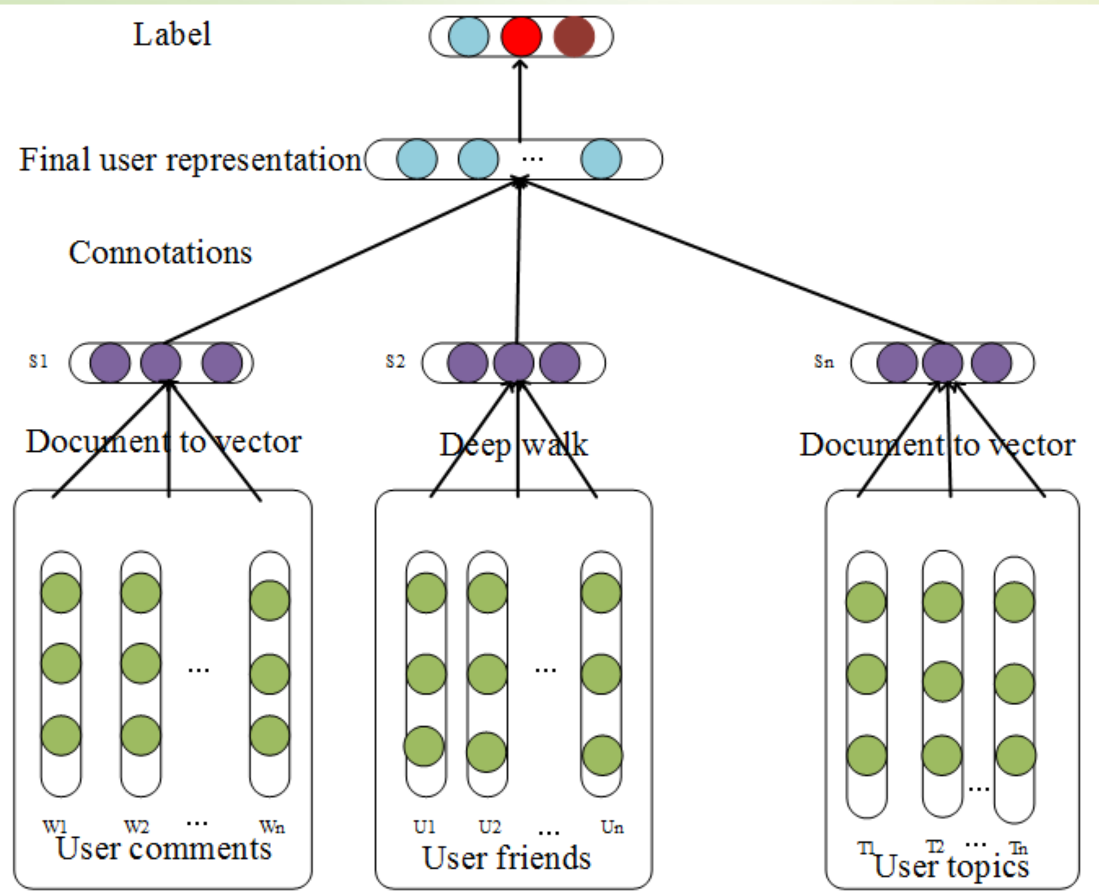
- **User preferred teams:**
  - Basketball, user can select one of CBA's 20 team as he/she's favorite team.



  - 17,011 users have favorite team, not uniformly distributed.
    - Popular teams like Guangdong Southern Tigers and Beijing Ducks have 4,221 and 3,159 loyalists.
    - The least popular teams like Beijing Fly Dragons and Jiangsu Kings only have 31 and 42 loyalists.
    - Reflects the unbalanced labeling problems in social media data.
    - More reliable as golden answer

Proposed model: Model framework

- **Task: User preferred team prediction.**
- **Model framework:**



- ➢ Soft max classifier
- ➢ Concatenate all three subparts
- ➢ Deep walk
- ➢ Doc2 vec
- ➢ User comments
- ➢ User friend
- ➢ User interested topics

User representation

- **User representation by three parts of user activities:**

$$u_i \propto \{G_{W_i}, G_{S_i}, G_{I_i}\}$$

  - User comments, user social network and user interested topic are represented as $G_{W_i}, G_{S_i}, G_{I_i}$ respectively.
- **Integrated User representation:**
  - Let e_{W}, e_S, e_I represent **Comments(words), Social Networks and interested topics**) $e_U = e_{\mathrm{W}} \bigoplus e_S \bigoplus e_I$
- **Comments representation:**
  - Input: user generated comments.
  - Document embedding problem
  - Embedding algorithm: Document to Vector. (Doc2vec, Mikolov 2014)

User representation

- **Social network representation:**
  - Input: User, Friends in user social network.
  - User and User's friend build a net work, we can use network embedding to obtain this representation.
  - Embedding algorithm: DeepWalk (Perozzi 2014)

- **Interested topic representation:**
  - Input: Topic threads in its list
  - User's interested topic can be treated as 'words', except no word order.
  - Embedding algorithm: Document to Vector. (Doc2vec, Mikolov 2014)

Performance Evaluation

- **Compare to other user presentation models:**
  - Task: prediction user's favorite team, a 20 class classification task.
  - Classifier: Soft-max classifier, no parameter tuning.
  - The dimension size of all experiments are set to 300 except BOW.

Model Name	Precision	Recall	F-score
Bag-of-words	0.2472	0.2695	0.2579
Latent Dirichlet Allocation	0.1745	0.2742	0.2133
Average word embedding	0.2241	0.2914	0.2538
Document-to-vector (Doc2vec)	0.3503	0.3689	0.3594
Singular Value Decomposition	0.3709	0.3176	0.3422
Probabilistic Matrix Factorization	0.3776	0.3564	0.3667
Integrated User Representation model (our proposed model):	**0.4182**	**0.3521**	**0.3822**

Experiment: Impact of user activity components

- **Examine the effect of each user activity: Comment (C), Network (N) and Interested topics (I).**

Feature Combination	Precision	Recall	F-score	P-value
Comment(C)	0.4014	0.3487	0.3707	0.0007
Network(N)	0.1996	0.2520	0.2227	<=10e-15
Interested Topics(I)	0.1875	0.2516	0.2149	<=10e-15
N+I	0.2880	0.2601	0.2738	<=10e-15
C+I	0.4011	0.3507	0.3741	0.002
C+N	**0.4066**	0.3502	0.3763	0.032
C+N+I	0.4052	**0.3632**	**0.3815**	**N/A**

Performance of different user activities

Performance Analysis

- Both user comments, user social network and user interested topic contain user preference information.
- Social networks and topics are not sufficient for the preference prediction task. ----May be due to the data sparseness issue in this dataset.
- Precision in the all integrated version is slightly lower than using only user comments and social network data. ---When recall is improved, more noise may also be introduced to have adverse effect on precision.

Conclusion

- A user profile corpus contain multiple activities. public available
  - To be made publically available after final check.
- A novel **integrated model** to learn user profiles from multiple user activities.
  - **Different embedding methods for different component**
  - **Best performance to all baseline methods**

- **Future Works:**
  - Deal with data sparseness and imbalance problem.

Thank you!

Selected Reference:

- B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.
- D. Rao and D. Yarowsky, “Detecting latent user properties in social media,” in *Proc. of the NIPS MLSN Workshop*. Citeseer, 2010.
- Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” *arXiv preprint arXiv:1405.4053*, 2014.
- S. Volkova, G. Coppersmith, and B. Van Durme, “Inferring user political preferences from streaming communications.” in *ACL (1)*, 2014, pp. 186–196.
- S. T. Dumais, “Latent semantic analysis,” *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.