*** This work has been accepted** as **Extended Abstract** in the ACM UMAP 2016 **conference.**

Analyzing MOOC Entries of Professionals on LinkedIn for User Modeling and Personalized MOOC Recommendations

Guangyuan Piao
Insight Centre for Data Analytics, NUI Galway
IDA Business Park, Galway, Ireland
guangyuan.piao@insight-centre.org

John G. Breslin
Insight Centre for Data Analytics, NUI Galway
IDA Business Park, Galway, Ireland
john.breslin@nuigalway.ie

ABSTRACT

The main contribution of this work is the comparison of three user modeling strategies based on *job titles*, *educational fields* and *skills* in LinkedIn profiles, for personalized MOOC recommendations in a cold start situation. Results show that the `skill-based` user modeling strategy performs best, followed by the `job-` and `edu-based` strategies.

1. INTRODUCTION

Massive Open Online Courses (MOOCs) are an online phenomenon which has been gathering momentum over the past few years. According to a recent study [3], over half of MOOC learners (62.4%) reported themselves as being employed full-time or self-employed. This indicates that MOOCs play a significant role in educating professionals. In this work, we investigate whether information in different fields of professionals' profiles from LinkedIn[1] (e.g., job titles) allows to produce useful user profiles which can be used for personalized MOOC recommendations.

2. RELATED WORK

To provide personalized MOOC recommendations, MOOC data traces (e.g., learning history, access logs, etc.) as well as learning content information have been used in the literature [1, 2]. Aher et al. [1] used data mining techniques to learn students' behaviors from data collected in a course management system for course recommendations. Apaza et al. [2] proposed recommending MOOCs based on historical grades of students in their college and inferred topics from content in course syllabus using probabilistic topic models.

Our work is different since we focus on user modeling based on user profiles and focus only on cold start situations. For instance, *does it make sense to provide MOOC recommendations based on a learner's job title(s) or skill(s)?*

Table 1: Parameter estimates

Parameter	B	Exp(B)	p-value
the degree below bachelor	-0.373	0.689	0.003
bachelor's degree	-0.206	0.814	0.003
master's degree	-0.194	0.824	0.004
PhD degree	0	1	.

Dependent variable: the # of MOOCs taken by learners

3. DATA COLLECTION

We created a Google Custom Search Engine (GCSE)[2] to retrieve LinkedIn profiles with the keyword "coursera". Overall, the dataset consists of 15,744 Coursera[3] MOOC entries for 5,668 professionals from LinkedIn. 5,134 out of 5,668 profiles contain degree information. 37% of learners in our dataset have bachelor's degrees while 47% and 12% of them have master's and PhD degrees. The average number of courses taken by these learners with different degrees (from the degree below bachelor to PhD) are 2.4, 2.8, 2.9 and 3.5, respectively. Based on *Negative binomial regression* and *Poisson regression*, which are often used for modeling count variables, we can observe the "*rich get richer*" phenomenon: a "*richer*" (with a higher degree) learner tends to take more MOOCs (see Table 1). More details about the dataset is available at the supporting website of this work[4].

4. USER MODELING STRATEGIES

Users are represented by a vector of weighted keywords from a specific field in their profiles. Thus, the profile of a user $u \in U$: $P_u = \big\{\big(k, w\big(u,k\big)\big) \mid k \in K, u \in U\big\}$ consists of a set of weighted keywords where with respect to the given user u for a keyword $k \in K$ its weight $w(u,k)$ is computed by a certain function w. Here, K denotes the set of keywords from a specific field of user profiles, and U denotes users. For instance, the fields in a LinkedIn profile about a user u can be summarized as: (1) job titles: `Software Engineer`, `Java Engineer` (2) education fields: `Information Engineering`, and (3) skills: `Java, C++, Microsoft Excel`. We use the well the known TF (Term Frequency)-IDF (Inverse Document Frequency) as the weighting scheme, i.e., $w(u,k) = \log\left(f_{k,u}\right) \times \log\frac{N}{1+|\{u \in U : k \in u\}|}$. $f_{k,c}$ denotes the number of occurrences of a keyword k in a specific field of a user u, N and $|\{u \in U : k \in u\}|$ denote the total number of

[1] https://www.linkedin.com

[2] https://www.google.ie/cse

[3] https://www.coursera.org

[4] http://parklize.blogspot.ie/2016/04/umap2016ea.html

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

users and the number of users where the keyword k appears in their user profiles. In the same way, we construct a course profile, which is represented by a vector of weighted keywords from users who have taken the course.

5. EXPERIMENTAL EVALUATION

Our main goal is to analyze and compare the applicability of different user modeling strategies in the context of MOOC recommendations. We do not aim to optimize recommendation quality, but are interested in comparing the quality achieved by the same recommendation algorithm when inputing different types of user profiles. In this regard, we apply a lightweight content-based algorithm as below.

Recommendation algorithm. Given a user profile and a set of candidate courses, the recommendation algorithm ranks the candidate courses according to their similarity to the user profile. The similarity is calculated by the *dot product* of the user and course profiles:

$$sim\,(u,c) = \frac{\vec{P_u}}{||\vec{P_u}||} \cdot \vec{P_i} \qquad (1)$$

which denotes "*how much of the course vector* $\vec{P_i}$ *is pointing in the same direction as the user vector* $\vec{P_u}$". Although it is reasonable to use the cosine similarity, the dot product outperforms the cosine similarity when representing user and course profiles using the TF-IDF weight for each keyword.

Given the dataset of learners with course entries in the previous section, we filtered learners with all information about current/previous job titles, educational fields and skills in their profiles. 4,401 profiles were left after filtering. Next, we randomly divided the dataset into training (4,080) and test sets (321) for the experiment. The TF-IDF weights for each keyword were obtained based on the distribution of each keyword in the training set. The ground truth of MOOCs for 321 users, was given by MOOCs in their LinkedIn profiles. All MOOCs in the dataset were used for constructing the candidate set (442) for recommendations. The recommender system then recommends MOOCs with highest similarities to a learner profile from 442 candidate MOOCs.

We compare the quality of different user modeling strategies to that of the *top-popular* recommendation strategy as a baseline, which is a common practice for cold start situations. *Top-popular* recommendation (`pop`) is a non-personalized model recommends the top-N items with the highest popularity amongst learners. The performance of the recommender system was evaluated by standard evaluation methods Mean Reciprocal Rank (MRR) and Success at rank N (S@N). MRR indicates at which rank the first item relevant to the user occurs on average. S@N stands for the mean probability that a relevant item occurs within the top-N recommendations.

Results. Figure 1 shows the recommendation performance in terms of MRR and S@05. We tested the statistical significance of our results with the bootstrapped paired t-test where the significance level was set to $\alpha = 0.01$ unless otherwise noted. The results show that `skill-based` profiles performs best, followed by `job-` and `edu-based` profiles. All of these user modeling strategies outperform the baseline method while `skill-` and `job-based` profiles perform significantly better than the baseline method in terms of MOOC recommendations. In detail, the `job-based` user modeling strategy improves MRR and S@05 26% and 43% respectively, while the `skill-based` approach improves MRR and S@05 37% and 52% respectively compared to the non-personalized approach. `edu-based` user modeling strategy improves MRR and S@05 11% and 26% respectively (although the difference is not statistically significant). The results show that job titles from work experience and skills of learners are useful for user modeling in the context of MOOC recommendations. Also, it indicates that any MOOC provider that has the functionality of signing up with LinkedIn (via OAuth[5]) as well as LinkedIn itself can exploit different fields of user profiles to provide personalized MOOC recommendations, especially in a cold start situation.

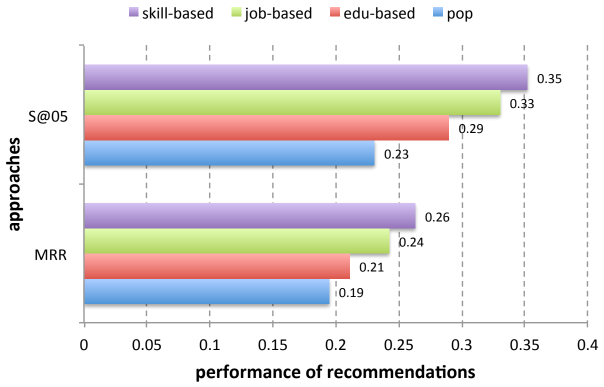

Figure 1: Results of MOOC recommendations with different user modeling strategies.

6. CONCLUSION

In this work, we investigated three different user modeling strategies based on the collected LinkedIn dataset. The dataset showed that a "*richer*" learner tend to take a greater number of MOOCs. In terms of user modeling strategies, our experiment showed that the `skill-based` user modeling strategy performs better than the `job-` and `edu-based` ones in the context of MOOC recommendations.

7. ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight Centre for Data Analytics).

8. REFERENCES

[1] S. B. Aher and L. Lobo. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems*, 51:1–14, oct 2013.

[2] R. G. Apaza, E. V. Cervantes, L. C. Quispe, and J. O. Luna. Online Courses Recommendation based on LDA. In *SIMBig*, pages 42–48. Citeseer, 2014.

[3] G. Christensen, A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E. J. Emanuel. The MOOC phenomenon: who takes massive open online courses and why? *Available at SSRN 2350964*, 2013.

[5] http://oauth.net/2/