Similarity Functions for Collaborative Master Recommendations

Alexandru Surpatean
Department of Knowledge Engineering
Maastricht University
6211 LH, Maastricht, the Netherlands
a.surpatean@student.maastrichtuniversity.nl

Evgueni Smirnov
Department of Knowledge Engineering
Maastricht University
6211 LH, Maastricht, the Netherlands
smirnov@maastrichtuniversity.nl

Nicolai Manie
University College Maastricht
Maastricht University
6211 KH, Maastricht, the Netherlands
nicolai.manie@maastrichtuniversity.nl

ABSTRACT

A memory-based collaborative system for recommending Master programs has been recently developed for University College Maastricht (UCM). Given the academic profile of a Bachelor student, the system recommends Master programs for that student based on the similarity of her profile to the profiles of the alumni students. The system is operational since September 2011 and is already popular among the UCM students.

This paper considers the question of how to improve the quality of Master recommendations. For that purpose we study several academic profile representations and similarity functions. We identify the best representation strategy and show how to combine recommender systems based on different similarity functions to achieve superior Master recommendations.

Keywords

Student Similarity, Collaborative Recommender System, Master Program Recommendation

1. INTRODUCTION

University College Maastricht (UCM) is a Bachelor program offering a liberal-arts and sciences education. In this study, students can build their own curriculum consisting of approximately 40 out of 157 offered educational modules: courses, skill trainings, and projects. Thus, the academic profiles of the UCM Bachelor students are diverse. To manage the diversity, UCM employs academic advisors, whose task is to help students choose courses in the light of the final goals: desired type of Master programs, jobs, etc.

To facilitate the students and advisors at UCM, we have developed a memory-based collaborative system for recommending Master programs. Given the academic profile (list of past, current, and future academic modules) of a Bachelor student, the system suggests Master programs for that student based on the similarity of her profile to the academic profiles of the alumni students. The tool allows the Bachelor student to modify her own profile, and thus to explore different alternatives in her study and how they influence her Master program possibilities.

This paper considers the question of how to improve the quality of Master recommendations. For that purpose we study two representations of the academic profiles of the students (binary and ECTS[1]-based) and two classes of similarity functions (cosine and Tversky index). We show that: (1) the best representation is ECTS-based and (2) there is no best similarity function. Nevertheless, we introduce an approach to combine recommender systems based on the similarity functions under study so that the resulting combination achieves superior Master recommendations.

[1] European Credit Transfer and Accumulation System.

2. MASTER RECOMMENDATIONS

The Master Recommendation problem is given as follows. Let S be the set of all the students, C be the set of all the Bachelor modules, and M be the set of all the Master programs. The *academic profile* of a student $s \in S$ is a vector p_s of values $p_{sc} \in p_s$ corresponding to modules $c \in C$. The values p_{sc} are binary or ECTS-based: if the student s followed or plans to follow module c, then p_{sc} equals 1 or the number of ECTS for c; otherwise, p_{sc} equals 0. The set of all the academic profiles p_s is denoted by P.

In the context of our formalization, the Master Recommendation problem is to find a subset $M' \subseteq M$ of Master programs that fit the academic profile p_s of a student $s \in S$, given data $D \subseteq P \times M$ of academic profiles of alumni Bachelor students, labeled by the Master programs they have chosen. The problem is essentially a classification problem, as each alumni profile is labeled by one Master program, not by a set of programs or preference on them. In this respect, our problem differs from standard recommendation problems where such sets/preferences are available [1].

To solve this Recommendation problem we need a Recommender System h: $P \rightarrow 2^M$. We have designed our recommender system h as a memory-based collaborative recommender system [1]. The system memory consists of the training data $D \subseteq P \times M$ of the academic profiles of UCM alumni students labeled by the Master programs they have chosen. The system operates in a collaborative way [1,2]: given the academic profile p_s of a student $s \in S$, the recommender system h returns the set M' of Master programs of the alumni students whose academic profiles are among k-closest in the training data D to the profile p_s.

To specify completely the recommender system h we need similarity functions over the set P of academic profiles. In this context we note that, for UCM, the set C of modules is much larger than the set of modules a student takes. This implies that the module variables p_{*c} are asymmetric. Thus we need similarity functions for asymmetric binary variables and we choose two such functions: cosine similarity and Tversky index. Given two academic profiles $p_s, p_a \in P$ the functions are defined as follows:

$$\cos(p_s, p_a) = \frac{p_s \bullet p_a}{\|p_s\| \|p_a\|}$$

$$Tversky(p_s, p_a) = \frac{p_s \bullet p_a}{p_s \bullet p_a + \alpha(p_s \bullet \bar{p}_a) + \beta(\bar{p}_s \bullet p_a)}$$

The cosine similarity is a *symmetric* function for asymmetric variables. The Tversky index is an *asymmetric* function for asymmetric variables. If $\alpha=\beta=1$, then it equals the Jaccard distance; if $\alpha>\beta$, we have emphasis on the student to be advised; if $\alpha<\beta$, we have emphasis on the alumni students.

3. EXPERIMENTS

We evaluated our recommender system using the leave-one-out method. The UCM data for the system consists of academic profiles of 223 alumni. The total number of Bachelor modules, past and present, to define the academic profiles is 329. The number of unique Master programs to recommend is 147. Among the alumni, 106 followed the same Master as at least one other alumnus. Thus, the academic profiles of these 106 alumni were used in test folds.

Table 1 shows the accuracy of our recommender system. We note that a recommended set of Master programs is correct if the Master program of the student from the test fold is in the set. Since the parameter *k* increases the size of the recommended set, the accuracy grows with *k*. In addition we note that *k* is an upper bound on the size of the recommended set of Master programs.

We observe that, in the case of recommendations for Master programs, the similarity functions perform better on average when applied on the ECTS representation, as opposed to the binary representation. Moreover, the prediction accuracy when considering a set of between 2 and 80 recommending neighbors is significantly better when applied on the ECTS representation.

k	*cos*		*Tversky* (α=1, β=0)		*Tversky* (α=1, β=1) Jaccard		*Tversky* (α=0, β=1)	
	binary	ECTS	binary	ECTS	binary	ECTS	binary	ECTS
1	13.2	12.3	13.2	12.3	14.2	10.4	4.7	4.7
11	53.8	57.6	55.7	56.6	50.9	53.8	53.8	54.7
21	63.2	73.6	63.2	67.9	62.3	69.8	64.2	71.7
31	72.6	81.1	69.8	75.5	71.7	79.3	75.5	77.4
41	80.2	87.7	78.3	81.1	78.3	84.9	83.0	82.1
51	85.9	91.5	83.0	90.6	85.9	91.5	84.9	89.6
61	88.7	93.4	85.9	91.5	88.7	93.4	89.6	92.5
71	91.5	93.4	89.6	91.5	91.5	93.4	91.5	92.5
81	92.5	93.4	90.6	92.5	92.5	94.3	92.5	94.3
91	94.3	95.3	91.5	94.3	93.4	94.3	93.4	95.3
101	94.3	95.3	93.4	94.3	94.3	95.3	94.3	96.2
111	95.3	95.3	96.2	94.3	95.3	95.3	95.3	98.1
121	96.2	95.3	97.2	94.3	96.2	95.3	96.2	98.1
131	97.2	97.2	97.2	95.3	97.2	96.2	97.2	99.1
141	97.2	98.1	97.2	97.2	97.2	96.2	98.1	99.1
151	97.2	98.1	97.2	98.1	97.2	97.2	100	99.1
161	98.1	99.1	98.1	98.1	98.1	98.1	100	99.1
171	98.1	99.1	98.1	98.1	98.1	99.1	100	99.1
181	98.1	99.1	98.1	98.1	98.1	100	100	99.1
191	100	100	98.1	99.1	100	100	100	100

Table 1. Recommender accuracy versus *k*, as *k* increases.

Furthermore, we observe that the performance of the classifiers is relatively different in distinct ranges of the *k* number of neighbors. If we compare the three ECTS-based Tversky indexes in Table 1 we notice that: for *k* between 1 and 11, Tversky with α=1 and β=0 performs better than the other two; for *k* between 31 and 81 recommendations, the Jaccard index outperforms the other variants; while for *k* from 91 to 171, Tversky with α=0 and β=1outperforms the other Tversky variants. We conclude therefore that there is no clear best function when taking a big range for *k* into consideration.

4. COMBINED RECOMMENDING SYSTEM

Figure 1 shows the accuracy curves of two versions V_1 and V_2 of our recommendation system built on the ECTS-based Tversky similarity functions with $\alpha=\beta=1$ and $\alpha=0$, $\beta=1$. The convex hull of these curves is a set of points that contain the curves. We can build a combined recommender system whose accuracy curve is that of the convex hull. To illustrate the idea assume we need a recommender system whose accuracy curve contains the line segment (p_1, p_2) in Figure 1. This means that we need a recommender system V_3 whose accuracy for some *k* defines a point p_3 on (p_1, p_2). We design such a system by a very simple approach similar to that from [3]. If we need to determine Master programs for the academic profile p_s, we flip a loaded coin with heads probability equal to 1 - *distance*(p_1, p_3) / *distance*(p_1, p_2). If the face-up side is heads, the recommending set of V_1 for point p_1 is given; otherwise, the recommending set of V_2 for point p_2 is given. In the long run, it is straightforward to prove that the accuracy of the recommender system V_3 for the *k*-value of point p_3 will be equal to the accuracy of point p_3.

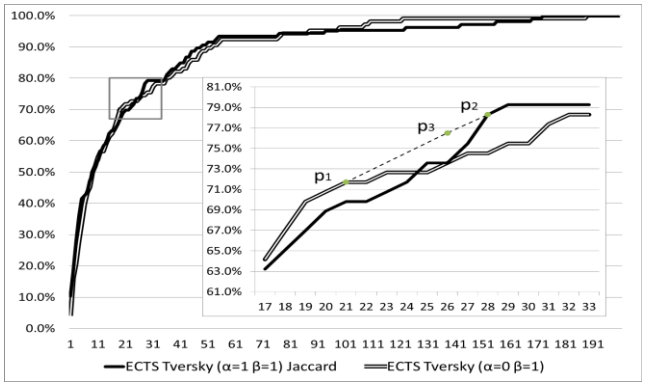

Figure 1. Accuracy Curves of the Recommender System.

5. CONCLUSION AND FUTURE WORK

This paper showed how to improve the quality of Master recommendations. It determined the best representation of academic profiles and introduced a new approach on how to combine recommender systems based on different similarity functions to achieve superior performance. Future work will focus on implementing and testing this approach.

6. ACKNOWLEDGMENTS

We would like to thank the members of the Maastricht University Leading in Learning program for their financial support, as well as Prof. Harm Hospers and the staff of UCM for their help and support.

7. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17* (6). 734-749.

[2] Hastie, T., Tibshirani, R. and Friedman, J. *The elements of statistical learning*. Springer, 2009.

[3] Provost, F. and Fawcett, T. Robust classification systems for imprecise environments. *Proceedings of AAAI/IAAI*, 1998, 706-713.