How university entrants are choosing their department? Mining of university admission process with FCA taxonomies.

NIKITA ROMASHKIN, DMITRY IGNATOV and ELENA KOLOTOVA, National Research University – Higher School of Economics, Moscow, Russia

The aim of this paper is to present a case study in the analysis of university applications to the Higher School of Economics (U-HSE), Moscow. Our approach uses lattice-based taxonomies of entrants' decisions about undergraduate programmes. These taxonomies were built by means of Formal Concept Analysis (FCA). FCA is a well-known algebraic technique for object-attribute data analysis. Admission data as well as formalised survey data were used to reveal possibly significant factors of entrants' decisions. In this paper we argued that institutional characteristics of the admission process are highly correlated with entrants' choice. The obtained results are helpful to the university to correct the structure and positioning of its undergraduate programmes.

Key Words and Phrases: University admissions, formal concept analysis

1. INTRODUCTION

The aim of this paper is to present a case study in the analysis of university applications to the Higher School of Economics (U-HSE), Moscow.

Decision-making process of potential students is a popular topic among researchers of higher education, because of it value for understanding university positioning. Nevertheless, not many works cover speciality choice issue - much more of papers study factors influencing university choice. Many papers on speciality choice try to investigate factors influencing prospective students decision on particular faculty selection (e.g., [Chang et al. 2006]). Some investigations are focused on speciality choice during last courses of study (e.g., [O'Herrin et al. 2003]). We found paper [Akbulut and Looney 2007] most relevant to our study - it tests a model aimed at identifying and explaining the mechanisms that shape student choice of computer science.

To analyse the data we mainly use a well-known algebraic technique called Formal Concept Analysis (FCA) [Ganter and Wille 1999]. There are applications of FCA in such fields as linguistics [Priss 2005], social networks analysis [Freeman and White 1993; Roth et al. 2006; Kuznetsov and Ignatov 2009], machine learning [Kuznetsov 2004], data mining [Poelmans et al. 2011] and software engineering [Tilley et al. 2005].

A common usage of FCA implies building of so-called concept lattices based on corresponding formal contexts. From a graph-theoretic point of view, a formal context is a bipartite graph where one part is a set of objects and the other part is a set of attributes. An edge between an object and an attribute means that an object has an attribute. Thus in most cases we represent a context as a biadjacency matrix or in terms of FCA a cross table. In terms of bipartite graphs a formal concept is a maximal biclique of a context. In a cross table a formal concept is a maximal rectangle filled with crosses with respect to any permutation of rows and columns. A formal concept consists of an extent and an intent. An extent is the set of objects

This work was partially supported by the Scientific Foundation of the State University Higher School of Economics, grant #10-04-0017 and by the Russian Foundation for Basic Researches, grant # 08-07-92497-NTSNIL_a

Author's address: N. Romashkin, email: romashkin.nikita@gmail.com; D. Ignatov, email: dmitrii.ignatov@gmail.com; E. Kolotova, email: kolotova.e@gmail.com

which have all attributes from an intent. Similarly, an intent is the set of attributes common for all objects from an extent. A set of all concepts are ordered by a generalization relation $\leq$ defined as follows. If $A \subseteq C$ (equivalently, $D \subseteq B$) stands for a concept extent A and another concept extent C then $(A, B) \leq (C, D)$, that is the concept (C, D) is more general than (A, B). A set of all concepts and a relation $\leq$ forms a concept lattice. A concept lattice can be viewed as an overlapping taxonomy for underlying categories (concepts).

The rest of the paper is organized as follows. In the next section, we present our case study, describe the admission process to U-HSE and our data, explain the data preprocessing step, and discuss the results of our analysis in the end. Section 3 concludes the paper.

2. CASE STUDY: ADMISSION PROCESS TO U-HSE

2.1 Background

Assuming probable confusion of the Russian educational system, we must say a few words about the Higher School of Economics[1] (U-HSE) and its admission process.

Nowadays U-HSE is acknowledged as a leading university in the field of economics, management, sociology, business informatics, public policy and political sciences among Russian universities. Recently a number of bachelor programmes offered by U-HSE has been increased. Currently U-HSE offers 20 bachelor programmes. We consider only bachelor programmes in our investigation. In order to graduate from school and enter a university or a college every Russian student must pass a Unified State Exam (Russian transcription: EGE), similar to US SAT–ACT or UK A-Level tests. During 2010 admission to U-HSE, entrants were able to send their applications to up to three programmes simultaneously. Some school leavers (major entrants of U-HSE bachelor programmes) chose only one programme, some – two, and some – three. Then entrants had to choose only one programme to study among successful applications.

2.2 Data

2.2.1 *General Information.* We used data representing admission to U-HSE in 2010. It consists of information about 7516 entrants. We used mainly information about programmes (up to three) to which entrants apply[2]. Exactly 3308 entrants successfully applied at least to one programme, but just 1504 become students. Along with this data we also used the data of entrants' survey (76% of entire assembly).

Further in the paper we mostly used data for the Applied Mathematics and Informatics programme to demonstrate some results. The total number of applications to the Applied Mathematics and Informatics programme was 843, of which 398 were successful but only 72 of them were actually accepted into the program. It might seem confusing only 72 out of 398 eligible prospective students decided to enroll, but since the admission process was set up in two stages, and at each stage only 72 entrants were eligible to attend the program, some of them decided to go for a different programme or university. As a result, the number of entrants whose applications were successful in any way came down to 398. Such situation is typical for all the bachelor programmes at U-HSE.

2.2.2 *Preprocessing.* FCA requires object-attribute data. In our case objects are entrants and programmes they apply to are attributes. Together they are treated as a context. A series of contexts were constructed. Namely, we built a context for every programme where objects were entrants applying to that programme and attributes were other programmes they applied to. We built a separate context for every programme because it is meaningless to consider all programmes at once as programmes are very different in size and the resulting lattice would represent only the largest of them.

[1] http://www.hse.ru/en/

[2] U-HSE is a state university, thus most of student places are financed by government. In this paper we consider only such places.

Likewise, we built a context for every programme where objects were entrants and attributes were programmes to which entrants successfully applied as well as the programmes that the entrants decided to enroll into, including those at other universities.

These contexts were then used to build concept lattices. Since the resulting lattices had too complicated a structure to interpret, we filtered concepts by their extent size (extent size is the number of objects, in our case it is the number of entrants), thus remaining concepts express only some of the more common patterns in entrants decisions.

2.3 Results

2.3.1 *Entrants' choice of programmes to apply.* To which programmes entrants often apply simultaneously? Trying to answer this question for every programme, we built diagrams[3] similar to figure 1. Such diagrams help us to reveal common patterns in entrants choices. Let us explain what the diagram in figure 1 represents. This diagram (also known as a Hasse diagram or a line diagram) is a diagram of the partial order on formal concepts. Typical applications of FCA imply building formal concept lattices discussed earlier, but here we filter concepts by extent size to avoid complexity caused by noise in the data; this reduction technique is well-known to the data mining community as so called "iceberg lattices" [Stumme et al. 2002]. Thus the order on remaining concepts is no longer a lattice, it is a partial order. Meaning of the labels on the diagram is obvious. A label above a node is a programme, a label below a node is a percent of entrants to Applied Mathematics and Informatics programme who also applied to programmes connected to a node from above. For example, the most left and bottom node on the diagram means that five percent of applied math's entrants also apply to Mathematics and Software Engineering. Then if we look at the nodes above the current node we may notice that ten percent Applied Mathematics and Informatics applicants also apply to Mathematics programme, and 70 percent also applied to Software Engineering.

Now let us try to interpret some knowledge unfolded by the diagram in figure 1. 70 percent of entrants who applied to Applied Mathematics and Informatics also apply to Software Engineering. The same diagram for Software Engineering states that 80 percent of Software Engineering applicants also apply to Applied Mathematics and Informatics. How this fact can be explained? Firstly it can easily be explained by the fact that these two programmes require to pass the same exams. Therefore there were not any additional obstacles to apply to both programmes simultaneously. Another possible explanation is that it is uneasy for entrants to distinguish these two programmes and successful application to any of them would be satisfactory result.

Analysing diagrams of other programmes' applications we found that equivalence of required exams is probably the most significant reason to apply to more than one programme.

2.3.2 *Entrants' "Efficient" choice.* If an entrant successfully applied to more than one bachelor programme he or she must select a programme to study. Unlike the previous case, entrants have to select exactly one programme which gives us more precise information about entrants preferences. For that reason we define this situation as an efficient choice, efficient in the sense of more expressive about true entrants preferences. Figure 2 presents the efficient choice of entrants to Applied Mathematics and Informatics programme. The meaning of diagram labels is almost the same as in figure 1. Programmes without plus sign (+) are successful applications, programmes with preceding plus sign are programmes chosen to study by entrants. Label "- Other -" means that the entrant canceled his application preferring another university or not to study this year altogether.

Together with diagram in figure 1 this diagram provides us with more precise knowledge about preferences of entrants to the Applied Mathematics and Informatics programme. More than two thirds of entrants

[3]As any other data mining technique FCA implies an intensive use of software. All diagrams mentioned in this paper have been produced with meud (https://github.com/jupp/meud-wx). Meud is mainly developed by Nikita Romashkin and currently is in far pre-released state.

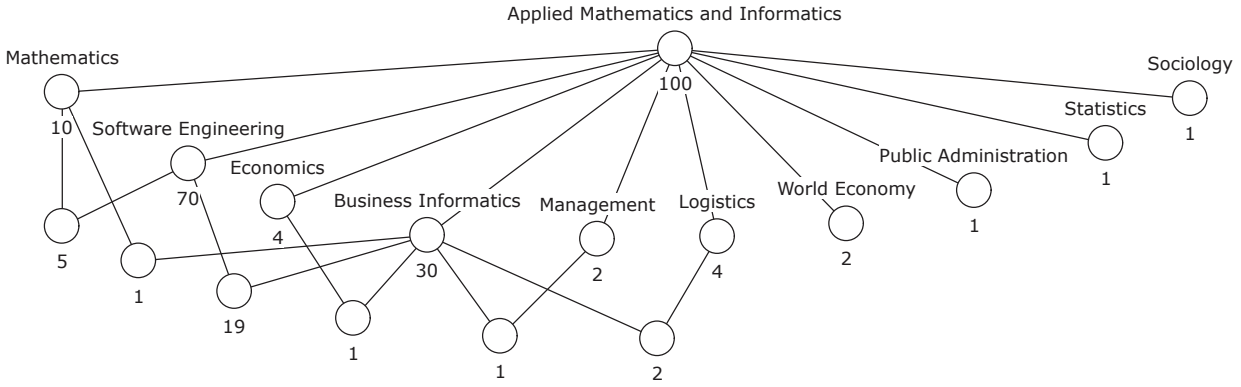

Fig. 1. Other programmes which entrants of Applied Mathematics and Informatics programme also apply.

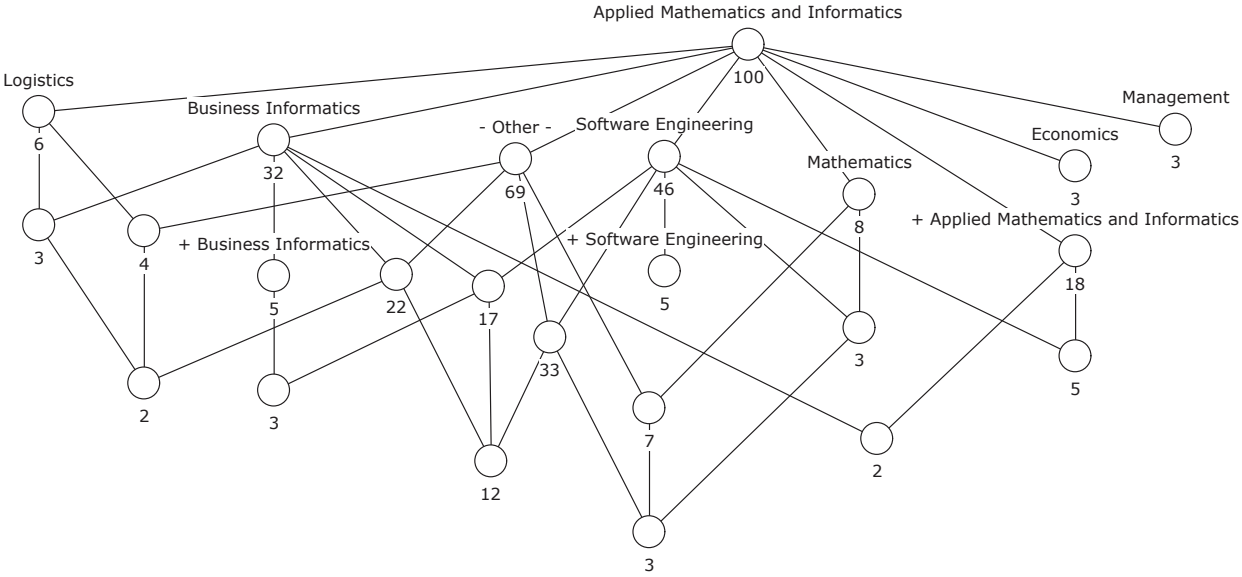

Fig. 2. "Efficient" choice of entrants to Applied Mathematics and Informatics programme.

who successfully apply to the Applied Math programme nevertheless prefer to study at another university. Whereas just 18 percent of successful applicants then become students on the Applied Mathematics and Informatics programme. Exactly 5 percent prefer to study Software Engineering and 5 percent of entrants who choose Applied Mathematics and Informatics also successfully applied to Software Engineering. It can be interpreted as equality of entrants preferences concerning these two programmes. Additionally, 5 percent prefer Business Informatics and only two percent of entrants who prefer Applied Mathematics and Informatics also successfully apply to Business Informatics, therefore in the pair Business Informatics and Applied Mathematics and Informatics the latter one is less preferable by entrants.

Here we should note that the sum of nodes percents with labels containing plus sign and node "- Other -" must equal to 100%, however here it does not because we excluded some nodes during filtering.

We built diagrams of "efficient" choice for every programme. Analysis of these diagrams helps us to recognise some relations between programmes in terms of entrants preferences. For example, some programmes in most cases is rather backup than actual entrants preference. Some programmes are close to each other by subject of study, these relations are also expressed by diagrams. With help of formalised survey data we found some possible factors of entrants' choice among some particular programmes. These knowledge can help our university to understand entrants' attitude to its undergraduate programmes and thus correct the structure and positioning of them.

3. CONCLUSION

We demonstrate a possible usage of FCA taxonomies in the field of educational data mining. After presenting some basic notions of FCA, we provide a case study of university admission process mining by means of FCA.

We present examples of diagrams obtained with help of FCA and provide a possible interpretation in two slightly different cases. Our main statement is that FCA taxonomies are a useful tool for representing object-attribute data which helps to reveal some frequent patterns and to present dependencies in data entirely at a certain level of details. Some interesting patterns then can be analysed separately and more carefully with help of other supportive data, for example, formalised survey data.

ACKNOWLEDGMENTS

We would like to thank Jonas Poelmans, Katholieke Universiteit Leuven and Galina Makarova for their help and support.

REFERENCES

Akbulut, A. Y. and Looney, C. A. 2007. Inspiring students to pursue computing degrees. *Commun. ACM 50*, 67–71.

Chang, P.-Y., Hung, C.-Y., lng Wang, K., Huang, Y.-H., and Chang, K.-J. 2006. Factors influencing medical students' choice of specialty. *Journal of the Formosan Medical Association 105,* 6, 489 – 496.

Freeman, L. and White, D. 1993. Using galois lattices to represent network data. *Sociological Methodology Volule 23,* 1, 127–146.

Ganter, B. and Wille, R. 1999. *Formal concept analysis: Mathematical foundations*. Springer, Berlin-Heidelberg.

Kuznetsov, S. O. 2004. Machine learning and formal concept analysis. In *ICFCA* (2004-02-19), P. W. Eklund, Ed. Lecture Notes in Computer Science Series, vol. 2961. Springer, 287–312.

Kuznetsov, S. O. and Ignatov, D. I. 2009. Concept stability for constructing taxonomies of web-site users. In *Satellite Workshop "Social Network Analysis and Conceptual Structures: Exploring Opportunities" at the 5th International Conference Formal Concept Analysis (ICFCA'07), Clermont-Ferrand, France*. 19–24.

O'Herrin, J. K., Becker, Y. T., Lewis, B., and Chen, H. 2003. Why do students choose careers in surgery? *The Journal of surgical research 114,* 2, 260–.

Poelmans, J., Elzinga, P., Viaene, S., and Dedene, G. 2011. Formally analysing the concepts of domestic violence. *Expert Syst. Appl. 38*, 3116–3130.

Priss, U. 2005. Linguistic applications of formal concept analysis. In *Formal Concept Analysis* (2005-07-20). 149–160.

Roth, C., Obiedkov, S. A., and Kourie, D. G. 2006. Towards concise representation for taxonomies of epistemic communities. In *CLA* (2008-04-15), S. B. Yahia, E. M. Nguifo, and R. Belohlavek, Eds. Lecture Notes in Computer Science Series, vol. 4923. Springer, 240–255.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. 2002. Computing iceberg concept lattices with titanic. *Data & Knowledge Engineering 42,* 2, 189–222.

Tilley, T. A., Cole, R. J., Becker, P., and Eklund, P. W. 2005. *A Survey of Formal Concept Analysis Support for Software Engineering Activities*. LNAI Series, vol. 3626. Springer-Verlag, 250–271.

Valtchev, P., Missaoui, R., and Godin, R. 2004. Formal concept analysis for knowledge discovery and data mining: The new challenges. In *Concept Lattices*, P. W. Eklund, Ed. Lecture Notes in Computer Science Series, vol. 2961. Springer, 352–371.