What can closed sets of students and their marks say?

Dmitry Ignatov and Serafima Mamedova and Nikita Romashkin and Ivan Shamshurin, University – Higher School of Economics, Russia

This paper presents an application of formal concept analysis to the study of student assessment data. Formal concept analysis (FCA) is an algebraic framework for data analysis and knowledge representation that has been proven useful in a wide range of application areas such as life sciences, psychology, sociology, linguistics, information technology and computer science. We use the FCA approach to represent the structure of an educational domain under consideration as a concept lattice. In this paper, we aim at building lattice-based taxonomies to represent the structure of the assessment data to identify the most stable student groups w.r.t the students achievements (and dually for courses marks) at certain periods of time and to track the changes in their state over time.

1. INTRODUCTION

Formal Concept Analysis (FCA) [Wille 1982; Ganter and Wille 1999] is an algebraic data analysis technique for building categories (formal concepts) defined as object sets sharing some attributes, irrespectively of a particular domain of application (for example, [Ignatov and Kuznetsov 2009]). FCA provides an analyst with a convenient and algebraic definition of a formal concept as a unit of human thinking. This definition is closely related to the philosophical notion of the "concept" characterized extensionally by the set of entities it covers and intensionally by the set of properties they have in common. In our study the set of objects (entities) comprises students and the set of attributes (properties) contains students' marks on different courses. The concepts form a taxonomy called a concept lattice w.r.t. specialization (generalization) relation on formal concepts. This taxonomy allows an analyst to study relationships between different groups of objects (dually, attributes) and find some interesting and potentially useful implications between their attributes. In contrast to conventional clustering techniques, FCA provides us with a kind of cluster, called formal concept, that captures similarity of objects in it by the set of shared attributes, and dually for attributes. In conventional clustering techniques we commonly have only the value of objects' similarity which comprises the cluster. The aim of this paper is to reveal some homogeneous groups of students w.r.t. their marks in term assessment data and to trace evolution of such groups in different terms of study by means of FCA. We have to add that all FCA and data mining tools such as concept stability, iceberg lattices, intensionally and extensionally related concepts were successfully used for building taxonomies of epistemic communities [Roth et al. 2006; Kuznetsov et al. 2007] and web users [Kuznetsov and Ignatov 2009]. Therefore, the main point of the paper is to show how this powerful inventory can be useful for the Educational Data Mining domain.

2. BASIC NOTIONS

Before we describe our main analyses, we briefly introduce the FCA terminology. A triple $\mathbb{K} = (G, M, I)$ is called a *formal context*, where G is a set of *objects*, M is a set of *attributes*, and the binary relation $I \subseteq G \times M$ shows which objects have which attributes. The derivation operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

This work is partially supported by the Russian Foundation for Basic Research, grant # 08-07-92497-NTSNIL_a.
Corresponding author's address: D. Ignatov, Zelenaya ulitsa, 17-52, Kolomna, Russia, 140410; email: dignatov@hse.ru;

$$A' = \{m \in M | gIm \text{ for all } g \in A\};$$

$$B' = \{g \in G | gIm \text{ for all } m \in B\}.$$

In that way, A' is the set of attributes common to all objects of A and B' is the set of objects sharing all attributes of B.

The double prime operator $(\cdot)''$ forms a closure operator, i.e., $(\cdot)''$ satisfies the properties – extensity, monotony, and idempotency. So, sets A'' and B'' are said to be *closed*.

A *(formal) concept* of the context $\mathbb{K} = (G, M, I)$ is a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A = B'$, and $B' = A$. One can see that $A'' = A$ and $B = B''$ in this case. The set A is called the *extent* and B is called the *intent* of the concept (A, B).

A concept (A, B) is called a superconcept of (C, D) if $C \subseteq A$ (which is equivalent to $B \subseteq D$). In this case, (C, D) is a subconcept of (A, B), and we write $(C, D) \leq (A, B)$. The set of all concepts ordered by $\leq$ forms a lattice and called the concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$ of the context $\mathbb{K}$.

2.1 Iceberg Lattices and Stability Indices

The main drawback of concept lattice taxonomies is their huge size even for relatively small contexts. For example, for a context $\mathbb{K} = (G, M, I)$, such that $|G| = |M| = 10$ in the worst case we have as a result $2^{10} = 1024$ concepts. However, real data are often sparse and the number of concepts is rather moderate to visualize their taxonomy. The purpose of an analyst is to reveal some interesting groups of individuals defined as concept extents. There are some remedies to cope with the huge size of concept lattices and to help find relevant concepts.

One of such well-known approaches in Data Mining is to find all frequent concepts, so called "iceberg lattices" [Stumme et al. 2002]. The *iceberg lattice* of a concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$ is a set $\{(A, B) | |A|/|G| \geq \theta\}$, where $(A, B) \in \underline{\mathfrak{B}}(\mathbb{K})$ and θ is a given threshold such that $0 \leq \theta \leq 1$. An iceberg lattice is just an upper part of the concept lattice or its order filter. However, one should be careful not to overlook small but interesting groups, for example, groups not yet represented by a large number of objects, or, groups that contain "emergent" objects which are not among members of any other group. There is an additional problem: the presence of noise in data may result in many similar concepts in the concept lattice. Considering the upper part of the lattice does not solve the problem, since this part may contain a lot of such similar nodes.

To solve the problem of selecting "sound" concepts (or their intents), many FCA practitioners use the notion of concept stability[Kuznetsov 2007; Kuznetsov and Ignatov 2009]. Let $\mathbb{K} = (G, M, I)$ be a formal context and (A, B) be a formal concept of $\mathbb{K}$. The stability index, σ, of (A, B) is defined as follows:

$$\sigma(A, B) = \frac{\{C \subseteq A | C' = B\}}{2^{|A|}}.$$

It is obvious that $0 \leq \sigma \leq 1$. Stability indicates how much the concept intent depends on particular objects of the concept extent. Thus, a stable intent is less sensitive to noise in object descriptions. In other words, stability measures how much the group of students depends on some of its individual members. In this respect, contexts where students are objects and students' marks on courses which they study are attributes are particularly adequate: here, formal concepts represent student's communities as groups of marks' on courses along with corresponding students. Removing a few students from the context should not change drastically the well-studied (or worse-studied) courses of a student community – "real" student communities ought to be stable in spite of noisy data. In a dual manner, we can define an extensional stability index, which indicates how a concept extent depends on particular attributes. In our domain it helps to answer the question: would the students of a given concept still belong to the same category if they stop sharing the same level of achievements on some courses?

By these means we are able to reveal stable groups of high achievers and well-studied courses as well as poor students and worse-studied subjects.

2.2 Scaling

In this paper we have dealt with student marks, so, our contexts did not contain binary attributes. Such contexts are called many-valued contexts. There is a technique to transform a many-valued formal context into a conventional single-valued context. This technique is called conceptual scaling. The main idea of conceptual scaling is to represent one many-valued attribute of the initial many-valued context by some binary attributes. There are some different kinds of scaling; we prefer nominal scaling, where the attribute values are not comparable, and ordinal scaling, where attribute values are comparable to each other.

2.3 Dynamic Mappings

Let $\mathbb{K}_1 = (G, M, I)$ and $\mathbb{K}_2 = (H, N, J)$ be two contexts describing the same class of students in two different time periods (or points). How has the situation changed between these time periods? In particular, if (A, B) is a concept of $\mathbb{K}_1$, what has happened to it in $\mathbb{K}_2$? Let's consider a concept $(C, D) \in \mathfrak{B}(\mathbb{K}_2)$; if the closure of $B \cap D$ is equal to $B \in \mathbb{K}_1$ and $D \in \mathbb{K}_2$, we say that (A, B) and (C, D) are *intensionally related* [Kuznetsov et al. 2007]. In the case of student assessment data, concepts intensionally related to (A, B) represent the evolution of the student's achievements in the disciplines B between two periods. Dually we can define the notion of *extensionally related* concepts. Two concepts $(A, B) \in \mathfrak{B}(\mathbb{K}_1)$ and $(C, D) \in \mathfrak{B}(\mathbb{K}_2)$ are said to be extensionally related if $(A \cap C)^{II} = A$ and $(A \cap C)^{JJ} = C$, where $(.)^{II}$ and $(.)^{JJ}$ are closure operators of the contexts $\mathbb{K}_1$ and $\mathbb{K}_2$ respectively.

Sometimes, in addition to main definition requirements for intensionally (dually extensionally) related concepts (A, B) and (C, D), it is useful to use such constraints as follows: $|B \cap D| \geq n\% \cdot |B|$ and $|B \cap D| \geq n\% \cdot |D|$. This helps us to reduce the number of related concepts.

3. EXPERIMENTS AND RESULTS

3.1 Data

In our study we consider two assessment datasets describing students who entered the university in 2006 and 2007 respectively. By means of intensional (extensional) relatedness we are trying to find main trends in students achievements and to understand which disciplines were the most complicated.

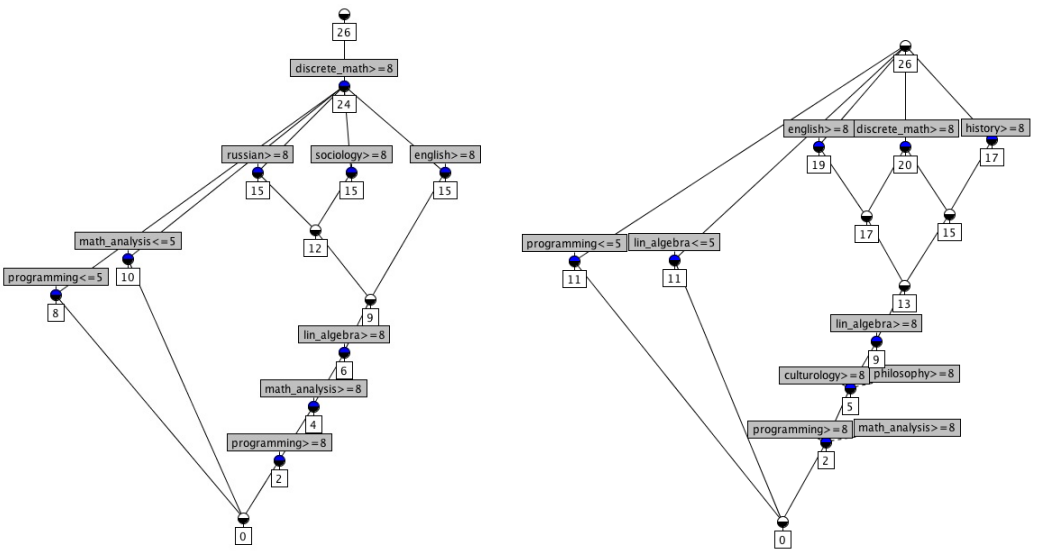
We analyse term marks of students who were studying at the department of "Applied Mathematics and Information Science" in two different academic years 2006/2007 and 2007/2008. It's worth to mention that our university maintains 10-grades system of assessment; more precisely, marks from 1 to 3 mean unsatisfactory results, 4 and 5 show satisfactory results, marks 6 and 7 indicate good results, and marks greater or equal 8 certify excellent achievements.

We compose several different contexts and conduct experiments (see subsection 3.3)

3.2 Preprocessing

We have to transform our multi-valued contexts to single-valued. To this end we introduce the following rules: for each course we consider 3 single-valued attributes, namely [course_name]$\leq$3 (a student has failed the final exam), [course_name]$\leq$5 (an exam is satisfactory passed or failed), [course_name]$\geq$8 (an exam is perfectly passed). Thereby, derived contexts have exactly $3|M|$ attributes. Scaling is a matter of interpretation, but in our case, by doing so, we are able to divide successful students and those who have some difficulties in studying. Cutting off students with good marks helps us to better sort out high achievers and mediocre students.

Fig. 1. Line diagrams of 13 most stable concepts for the first term (left) and the second term (right) of 2006/2007 academic year



3.3 Experiments

For each experiment we build a taxonomy of N most stable concepts of a certain context and then compare the taxonomies' diagrams.

3.3.1 *Entrants of 2006: comparison in terms of extensional relatedness between the first part of 2006/2007 and the second one.* Let us compare two diagrams on figure 1. They have a similar structure; at the right top of each diagram there are three incomparable nodes with 3 disciplines that are likely to be easily passed. In the first term they are Russian Language, Sociology, and English Language, in the second term – English Language, Discreet Mathematics, and History. Our approach shows us that the concept X of those who successfully passed Discreet Mathematic, Russian Language and Sociology extensionally related to the concept Y of those students who have 8 or higher mark in Discreet Mathematics, English Language, and History. At the same time one can see that some people have moved along humanities and there are students who became more successful in mathematical disciplines. Just the thing what we need to trace dynamics is related concepts by the closure of student sets intersection. Also the students from the extent of X are related to students with best results in Discreet Mathematics, English Language, and Calculus.

In the left part of the diagram there are those disciplines that cause the most difficulty. We can also derive that not all of those students who passed programming with the mark 5 or below on the second term have had any trouble with this discipline previously. Besides programming, the courses that caused the most difficulty were Analysis in the first term (let us denote corresponding concept as Z_1) and Linear Algebra in the second term (concept Z_2). A legitimate question arising from the similarity of the two diagrams would be whether or not Z_1 and Z_2 are extensionally related (connected). In fact, they are, and thus the intent of these formal concepts is comprised of mostly the same students.

3.3.2 *Entrants of 2006 and 2007: comparison in terms of intensional relatedness between the first part of 2006/2007 and 2007/2008.* The other diagrams are omitted due to a lack of space. On the whole, in the first semester of 2006 student marks were higher than in first semester of 2007. The requirements in Discrete Mathematics were made stricter: if in the former time period only two students received mark 7 or lower (7.7

percent), in the latter time period the fraction of such students rose to 94 percent. In the 2006 set we see a rather large block of "humanities oriented students"(the ones who successfully passed English, Russian, Sociology, and Discrete Mathematics (because 94 percent of "humanities oriented students" did pass Discrete Mathematics with marks 8 or higher)) that consists of 9 students. It is intensionally related with the set of students in the year 2007 who successfully passed English and Russian (17 students), humanities, Linear Algebra, Analysis, and Discrete Mathematics (3 students). Taking into account that the 2007 set contains twice as many students, it has become more challenging to pass the humanities (17 students passed English and Russian with excellent marks), and it seems more likely that there will be no "humanities oriented students" (everyone who passed the humanities are also strong math students). However there are also some positive trends. It turns out that in order to be good at programming, one does not necessarily have to be good at math or humanities, therefore, more students received excellent marks for programming.

3.3.3 *Entrants of 2006 and 2007: comparison in terms of intensional relatedness between the second part of 2006/2007 and 2007/2008.* The formal concept "poor students of the second semester second year enrolled in 2006" contains 5 students in its intent, of which 4 students subsequently dropped out, so the students enrolled in 2007 that comprise an intensionally related concept, become the most likely drop out candidates. We also note that in the 2007 there formed a group of students who had difficulty with philosophy (40% of the class), which were at the same time successful in Linear Algebra (8 %), Analysis, and English. Among those 40% none were good at History, which makes sense. It is also interesting to note some success dependencies for Linear Algebra: whereas 2006 students had to receive an excellent grade in Linear Algebra in order to receive mark 8 or above for Analysis, students enrolled in 2007, would not receive 8 for Linear Algebra unless they had successfully passed the Analysis course.

3.3.4 *Entrants of 2007: comparison in intensional terms between two student subgroups in the first part of 2007/2008.* Using a concept of intensional relatedness and diagram analysis we can make the following conclusions: in general the 272 group is more successful in studying English (52% who passed with excellent marks as opposed to 38%), which is hardly surprising – it is likely that groups were formed according to the students different level of command of the English language. Comparative analysis lead us to the idea that people who were proficient in English could afford to lose a couple of points on the math entrance exam. For that reason, Linear Algebra and Analysis grades are lower in group 271. We are not going to see a stable concept "those who passed Discrete Mathematics with mark 8 or higher" on the first diagram. Also, in 272 group we don't see students that are only good at languages (and of those there are 6). These students also have good results in Linear Algebra. The corresponding formal concepts are intensionally related. A common rule for both groups holds true: those who are good at Russian have no difficulty with Sociology. The reason being that the final Sociology exam is given in a form of an essay, which is standard practice in Russian classes. If we look on the very top level, then we note there are 5 formal concepts in both cases: "Russian$\geq$8", "Linear Algebra$\geq$8", "English$\geq$8", "Programming$\leq$5", "Analysis$\leq$5". So the student performance structure is very similar, which is what to be expected considering that we are comparing the same class (just different groups), that have common subjects, and common teaching staff. Some internal differences are of quantitative nature. According to our hypothesis, group 272 has a more solid math background and is better organized.

3.3.5 *Entrants of 2007: comparison in intensional terms between two student subgroups in the second part of 2007/2008.* Let's start top down. In general, we observe some quantitative differences. Core formal concepts (first level formal concepts): common concepts – "Analysis$\geq$8", differences – group 272 does not have a stable concept "philosophy$\leq$5", on the other hand, in group 271, there are no failing History students. In group 272 there are no clear failing Linear Algebra students. It is very noticeable that the second diagram contains two formal concepts the intents of which contain both mathematics and humanities. This indicates that either success or failure in one field correlates with success or failure in the other. So students are

either well organized and therefore are successful overall, or fail everything. In 272 diagram, there is one concept that stands out – "Linear Algebra≤5" It strongly correlates with formal concepts that deal with academic failure, and does not have anything to do with with English proficiency. So those students who did not get on well with programming, Culture, and Discrete Mathematics could not pass Linear Algebra. "English≥8" correlates with other disciplines≥8, whereas in the first diagram "English≥8" correlates with "programming≤5", "Culture≤5" and "Analysis≤5". So there is a negative correlation. It is not surprising considering that English does not present any major difficulty for 271. In general, there are no other important differences, besides the ones mentioned above.

4. CONCLUSION

In this paper we proposed to use closed sets of students and their marks to reveal some interesting patterns and implications in student assessment data, especially to trace dynamics. We restricted ourself only to a few useful techniques from FCA, but there are a lot of tools which can help an analyst to explore the assessment data. For example, to better represent and mine multi-valued numerical context we can use so called interval pattern structures [Ganter and Kuznetsov 2001]. Also, there are some alternative approaches to dynamics mappings, for example, the theory of multicontexts [Wille 1996]. We suppose that existing FCA-based techniques can be potentially useful in the other fields of EDM domain.

ACKNOWLEDGMENTS

We would like thank Jonas Poelmans, Katholieke Universiteit Leuven and Jane Ilyina, U-HSE for their help and support.

REFERENCES

Ganter, B. and Kuznetsov, S. O. 2001. Pattern structures and their projections. In *ICCS*, H. S. Delugach and G. Stumme, Eds. Lecture Notes in Computer Science Series, vol. 2120. Springer, 129–142.

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

Ignatov, D. I. and Kuznetsov, S. O. 2009. Frequent itemset mining for clustering near duplicate web documents. In *ICCS*, S. Rudolph, F. Dau, and S. O. Kuznetsov, Eds. Lecture Notes in Computer Science Series, vol. 5662. Springer, 185–200.

Kuznetsov, S. O. 2007. On stability of a formal concept. *Ann. Math. Artif. Intell. 49,* 1-4, 101–115.

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.

Kuznetsov, S. O., Obiedkov, S. A., and Roth, C. 2007. Reducing the representation complexity of lattice-based taxonomies. In *ICCS*, U. Priss, S. Polovina, and R. Hill, Eds. Lecture Notes in Computer Science Series, vol. 4604. Springer, 241–254.

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.

Wille, R. 1982. Restructuring lattice theory: an approach based on hierarchies of concepts. In *Rival, I. (ed.): Ordered Sets.* Boston, 445–470.

Wille, R. 1996. Conceptual structures of multicontexts. In *ICCS*, P. W. Eklund, G. Ellis, and G. Mann, Eds. Lecture Notes in Computer Science Series, vol. 1115. Springer, 23–39.