DATA PREPROCESSING ACCORDING TO ACTIVITY THEORY

Jean SIMON
Université de la Réunion, Laboratoire d'Informatique et de Mathématiques (LIM), Equipe EDIM
jean.simon@univ-reunion.fr

ABSTRACT --In this paper, we propose one possible way to preprocess data according to Activity theory. Such an approach is new and particularly interesting in Educational Data Mining. In a first time, we present the methodology we have adopted and, in a second, the application of this methodology to the analysis of the traces left during five years by the preservice teachers of the Reunion Island teacher training school.

KEYWORDS Educational data mining, Activity theory, data preprocessing, teacher training

1 INTRODUCTION

This paper relies on the following approach in four points: 1.Activity theory (Engeström, 87) is frequently used to study human activity specially CSCL (Computer Supported Collaborative Learning) because this theory is particularly suitable to understand what the people do when they collaborate. 2. One major problem in data mining consists in the data preprocessing. Better those data are preprocessed, more information we obtain by their treatment. 3. The idea developed here is thus to use domain knowledge to preprocess the data *before* their treatment; the advantage of this is to permit, *after*, to use any kind of data mining algorithm. 4. One possible domain knowledge to study CSCL is Activity theory; preprocessing based on Activity theory should make it possible to get more interesting results. In a first time, we present the methodology which develops the preceding points, in a second time we show the relevance of this methodology on a concrete case.

2 METHODOLOGY

According to Activity theory (AT) (Engeström,87), in the activity, the subject pursues a goal that results in an outcome. For this, he uses tools and acts within a community. His relation, to this community is defined by rules. To achieve the goal, it may be necessary to establish a division of labor within the community. For example, in the context of hunting, there will be hunters and beaters (Kuutti, 96). The Activity theory could be represented by the activity system diagrams of Figure 1.

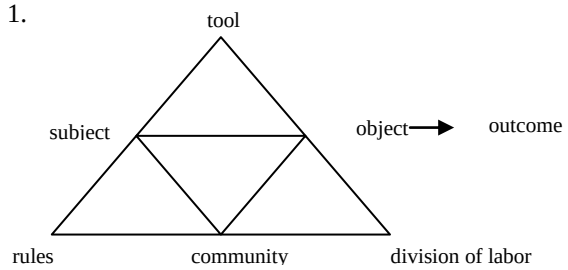

Figure 1. the triangles of AT according to Engeström(1987).

Moreover, AT considers three levels of human activity (Kuutii, 96): activity, action, and operation. We will not use them here but it could be a point to develop further. Using AT to understand what happens on a platform is common in the field of CSCL. As soon as 1996, Kuutti (1996) suggested doing it. For Stahl (2002), AT is suitable to analyze

CSCL because it proposes general structures of the broader effective context. In the same way for Halverson (2002), if the theory can't satisfy all the needs of the researcher in CSCL, it is powerful for, at least, three reasons: 1. This theory defines well its theoretical objects which are useful to manipulate data; 2. In this theory, the individual is at the center of everything; 3. Activity system diagrams highlight the processes and show both descriptive and rhetorical power. For his part, Lewis(1997) explains how AT focuses on interdependent parameters which exist in collaborative learning. He shows how each triad helps to understand what the people do on a CSCL platform. For example, how the triad Community-Subject-Object highlights the role of the trainer. These are the reasons why we will use AT further.

Educational data mining. In education, the use of groupware and learning management system grows increasingly. Most of these systems record the traces of users' activity. The study of those huge masses of data is important to understand the behavior of the users and to improve learning. It can be done by data mining. This approach consists to use algorithms (decision tree construction, rule induction, artificial neural networks, etc. (Romero & Ventura, 07) which enable to discover behavioral rules or data classification or clustering. In the context of education, we speak of Educational Data Mining (EDM). For Baker & Yacef (2009) EDM studies a variety of areas, including individual learning from educational software, computer supported collaborative learning,...

For Romero & Ventura (2007) there are four important steps in data mining: Collect data; Data preprocessing; Apply data mining; Interpret, evaluate and deploy the results. Data preprocessing is an important step because data tend to be incomplete, noisy and inconsistent (Han & Kamber,06). The data preprocessing includes the following steps (Han & Kamber, 06): Data cleaning for correcting errors; Data integration when the data come from different sources; Data selection; Data transformation. Among data transformations, Romero & Ventura (2007) propose data enrichment which consists of calculating new attributes from the existing ones. In business data mining (Talavera & Gaudioso, 04), or in medicine (Lin & Hang, 06), the domain knowledge is used for deriving those new attributes. As far as we know, this is not yet done in EDM.

Using AT for data preprocessing. So, we propose here to use domain knowledge to enrich the data during the preprocesssing step. One good candidate to represent the domain knowledge is Activity theory. We have seen that AT is used in CSCL studies, but it's also used more and more in data mining. For instance, Verbert & al (2011) see the datasets as activity streams and they reorganize their categorizations by using AT. Babic & Wagner (2007) proposed to use it to understand the activity on a platform through the three levels (activity- action - operation). Reimann &al (2009) see the groups as activity systems and the log file as containing records of these structured activities. However, none of these studies use it to preprocess data as we want to do here. What we propose is that "the preprocessing step will then consist in mapping original data to a higher-level, analysis oriented representation." (Talavera & Guadioso, 04) and, to do this, we will use AT. For example, it is possible on any platform to count the different actions: document deposits, document readings, ... it is possible then to study what are the actions, the features of the tool, which are the most used. Interpreted in terms of AT, by doing this, we consider a single node, the node "tool". If we apply to our data a first preprocessing, the user identification (Romero & Ventura, 07) ,in other words, if we associate with any action (reading, deposit, etc) the user who did it, we obtain a finer analysis. We will be able to discover that, according to the users, this is not the same features that are used, that a trainer, for example, deposits more documents than a trainee. Interpreted in terms of AT, by doing this, we studied the dyad "subject-tool".

Data useful for AT accessible on a CSCL platform. If we take the diagrams of Figure 1, we identify immediately three types of data readily available: the *tool*: it will be the traces of actions on the platform and objects on which these actions operate (e.g. document deposit, document reading); the *subject*: it will be the users registered on the platform; the *group*: every CSCL platform keeps traces of the groups created on it. Moreover, the traces of the links between these three types of data are also easily accessible. On all platforms, is recorded who are the members of a particular group (dyad: subject-group) and who did what (dyad: subject-tool). From these two dyads, it is easy to establish the third one (dyad: group-tool). The node "*objective*" is rarely the subject of a specific trace. It may be possible to find it through the name of the group or the name of the main folder that the group shares. Concerning the two last nodes "*division of labor*" and "*rules within the group*", it is exceptional to have explicit traces of them. This is not surprising because, most of the time, these two nodes are the goal of the study, what we try to understand as we shall see in the example of the third part. So they can not be the subject of a preprocessing data. Thus, we obtain figure 2:

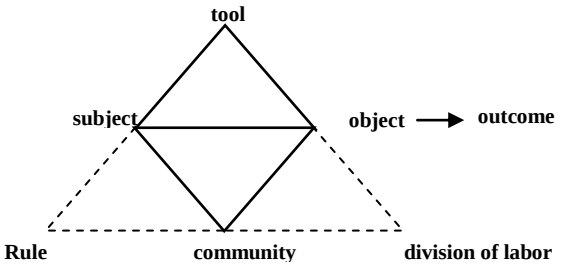

Figure 2. AT for data mining. Solid lines indicate data which are easily found among traces left on a CSCL platform and dotted lines the information that it will be necessary to infer.

Technically, it is easy to see how we can preprocess the data. In the case mentioned above, instead of studying unique data, such as, for example, "actions", we study couples “subject-action", who is doing what on the platform. In the same way, we can go on and work with 3-tuples "subject-group-action" and when we can identify the objective, 4-tuples “group-objective-subject-action” which correspond to the solid lines of figure 2. With these 4-tuples, we see it is easy to put the focus on one of the nodes, the subject, the tool, the goal, the group, without losing data interdependencies. *In fact, by this preprocessing, we inject in the data their interdependency that Lewis considers to be central in AT.* Once those 4-tuples are built, they can feed any kind of algorithms (decision tree construction, Bayesian learning,…).

3 APPLICATION

The Reunion Island teacher training school trains primary school teacher trainees (PE2s). At the end of their year of training, they should be able to teach in primary schools. PE2s have used a CSCW platform which allowed them to pool and share the work of preparation of the class. With their trainers, the platform has served various purposes: to deposit documents (“collective memory”), to improve lesson plans, to help online trainees when they are in charge of a class, to validate the C2i2e which confirms that the trainee is able to use ICT in education.

As platform, BSCW [3] was chosen essentially because users may structure as they wish spaces they have created there. For this study, we analyze the traces left on it by the PE2s during 5 years. 1167 PE2s (from 343 in 2005 to 155 in 2009) have used the platform, and left there more than 3.000.000 traces. We want to show that, according to the type of group or the type of objective, the activity is not the same. For that, we will follow the various points raised in the methodology and apply them to those traces. First, we present the results of an analysis of raw data, second, an analysis of data that have been preprocessed according to AT. In those sections, we do a simple statistical treatment; this is why, thirdly, we summarized a research exposed in (Simon & Ralambondriany, 12) where we have used data mining techniques.

Analysis of the raw data.When trainers and trainees work on BSCW they leave traces of what they do on the platform. It’s possible to connect each trace to the user who has left it (“user identification” preprocessing). For each user we were able to say how many readings he has done, how many documents he has deposited, how many folders he has shared, etc. The users were anonymized, they were just “numbers”. We obtain table 1.

Table 1. Analysis of the trace traces left by PE2s during 5 years

PE2s on the platform	over 5 years
Total number of PE2s	1167
Total number of documents shared by the PE2 on the platform	13936
Total number of documents produced by the PE2s on the platform	11308
Number of PE2 producers	840
Number of PE2s' readings	57916
Number of PE2 readers	1089

We see that 81% (11308 among 13936) of the documents are deposited by the PE2s and 19% (13936-11308=2628) by the trainers. We see also that if 93% of PE2s (1089 /1167) read documents found on the platform, only 72% of them deposit documents (840/1167). We can conclude that not all PE2s cooperate; we find the “lurker” phenomenon well-known in forums. However, if we analyze more precisely we will see that the situation is more complex.

Preprocessing data. On BSCW when a user wants to create a group, he must first create the folder in which the resource will be placed and then invite other members to share this folder. It is then possible to connect all the traces to this basic folder: the members of the group but also the objects and the events. In other words, it is rather easy to reconstitute the 3-tuples (group, subject, action on the tool). Moreover, as a name is associated with each of these folders, name often indicating the goal, it is also possible to build 4-tuples (group, goal, subject, action). We can then study the different groups or the different objectives or the different subjects or the different actions. So after we have done this preprocessing, we decide to do two distinctions: according to the composition of the group (with or without teachers : PE2s vs PE2s+trainers) and according to their goal (groups “TICE” and others).

Analysis according to the composition of the group. The 1167 PE2s have constituted more than 960 groups. 668 were groups of PE2s only and 292 groups including at least one trainer. One PE2, of course, could be in several groups. We made distinction “PE2s alone” groups (third column) and “PE2s+trainer” groups (fourth column) because we wanted to know if trainees would use the platform without being forced by the trainers. Among “PE2s+trainer” groups, we wanted also to focus on the groups whose objective was ICT for education (“TICE” groups, fifth column).

Table 2. Comparison of the groups with or without trainers

Groups without or without trainer over 5 years	all	PE2s	PE2s + trainer	*TICE*
Total number of groups	960	668	292	*68*
Total number of PE2s	1167	1167	1167	*884*
Average number of PE2s for one group	15	13	20	*13*
Average number of documents for one group	15	6	34	*41*
Average number of PE2 producers for one group	3	2	5	*11*
Average number of readings by PE2 for one group	60	29	132	*135*
Average number of PE2 readers for one group	10	8	15	*13*

As we can see the PE2s freely use the platform: the number of groups shared only by them is significantly greater than the number of groups shared with trainers (668 vs. 292). However the activity is much lower in groups shared only by PE2s than in groups shared with trainers. Thus, most figures in groups PE2s are lower than the figures in groups PE2s+trainers: fewer deposites and fewer readings. A possible explanation for this higher activity in groups PE2s+trainers is that there is an effect "teachers" that incites students to work more. For example, the trainee will be "invited" by the trainer to go and consult the documents that the trainer has deposited for him.

Analysis according to the objective. With the titles of the folders, it was very easy to isolate the groups named "TICE" (ICT for education). The folders associated with these groups were used to validate a certificate according to which the trainee is able to use ICT in Education in a correct way (C2i2e). All those groups have a trainer as member. We can therefore compare those "TICE" groups with all the groups with trainers "PE2S+trainers" in Table 2. In the "TICE" groups almost every PE2 works, 11/13 deposit (vs 5/20 for "PE2s+trainer") and 13/13 read (vs 15/20). Thus, even when the groups have the same composition, we can see that, according to the objective, activity is not the same. However, the "TICE" groups seem to reveal a strong cooperation and to be similar. We will see that it is not the case and that makes problem for the institution.

Data mining on the groups with teachers. Once the preprocessing is done, it is possible to operate more complex treatments and to use any kind of data mining algorithms. In (Simon & Ralambondriany, 2012), we have analyzed the groups with trainer of the year 2006-2007. We have done successively a Principal Component Analysis; a Ward Hierarchical Clustering and finally we use the k-means algorithm. We gathered by this way the groups with trainer in 7 clusters that we have named according to their features: C1 "groups with no activity", C2 "individualize accompaniment", C3 "weak cooperation", C4 "dissemination to a small group", C5 "dissemination to a big group", C6 "strong cooperation" and C7 "accompaniment during training course" (see Figure 3).

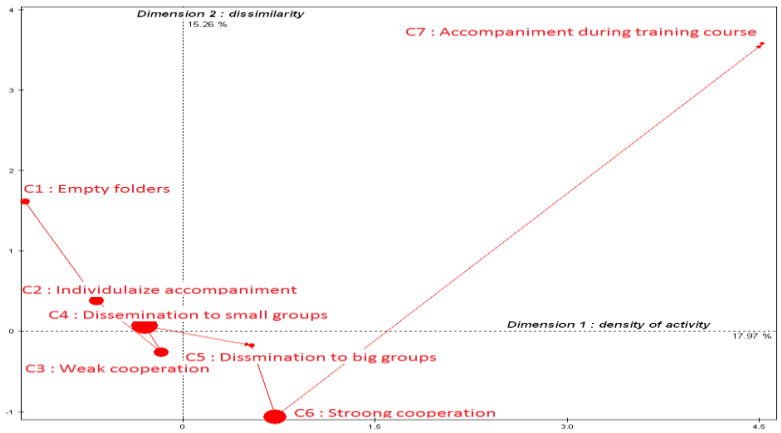

Figure 3 Clustering of the "TICE" groups of the year 2006-2007 (Simon & Ralambondriany, 12)

By this way, we were able to see that some "TICE" groups were belonging to the cluster C3 instead of C6. It was bad because all the "TICE" groups should be similar and reveal a strong cooperation. Thus, this clustering has permitted to reveal that some "TICE" groups don't follow the "rules" defined by the institution. Interpreted in the terms of AT, there was a contradiction at the node "rules". Just for information, group 7, which stands out in this figure by its high activity, was created to provide an answer, just in time and just enough, to trainees during their internship in classroom when they had to face to pupils.

4 CONCLUSION

The wealth of data mining relies on its bottom-up approach. The researcher starts from the data and expects that the machine will propose a categorization of these data. In this approach, it is assumed that, somehow, the researcher has no a priori knowledge about the data. This approach is very interesting because it can lead to orient research in an unexpected way. However, it often happens that the proposed categorization is not exploitable for the researcher (Talavera & Gaudioso, 2004). One possible way to overcome this problem is to incorporate domain knowledge in the data mining process. To represent the domain knowledge of CSCL, Activity theory is a good candidate. Incorporate AT in the data mining process has already been done in various ways (Babic & Wagner,07),(Reimann &al, 09), (Verbert K. & al, 2011) However, in those researches, in some manner, AT and data mining algorithm are tied. To avoid this, we propose to incorporate the domain knowledge in the data during the preprocessing step. By this way the choice of the data mining algorithm becomes free, it can go from simple statistics treatments to more complex algorithms (rule induction, artificial neural networks, Bayesian learning,…).

As we have seen, it is simple to do and we obtain results easily exploitable. The reason is that this kind of preprocessing is an injection of prior knowledge in the data: here, we inject in the data their interdependency. When we have applied this approach to data produced by preservice teachers with their trainers on a CSCL platform, we were able to show that according to the group in which they operate and the objective that these groups pursue the activity produced by preservice teachers is not the same.

A further step could consist to preprocess the data according to the different levels of Activity theory, activity, action, operation, but it is not sure we obtain significant results because most of the traces left on the platform fall under the "operation" level.

To conclude, as Halverson (2002) says "AT is powerful because it names and names well, but this both binds and blinds its practitioners to see things in those terms." It is possible that, by reducing the hypothesis space as we do by using AT, we avoid revealing assumptions that could be very interesting and we lose a part of the wealth of the data mining.

5 REFERENCES

Babič F., Wagner J., 2007, Modeling of knowledge creation processesbased on Activity theory; In. *Proc. of the 1st Workshop on Intelligent and Knowledge oriented Technologies*, WIKT 2006, Bratislava, Slovakia (2006), 131-134, ISBN 978-80-969202-5-9,

Baker R.S., Yacef K., 2009, The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining* 1(1), 3–17,

Bentley R., & al, 1997 ,Basic Support for Cooperative Work on the World Wide Web. *International Journal of Human Computer Studies: Special issue on Novel Applications of the WWW*, Academic Press, Cambridge, vol. 46, p. 827-846,

Engeström Y., 1987, *Learning by expanding: An Activity-Theoretical Approach to Developmental Research.* Orienta-Konsultit Oy.

Halverson C. A., 2002 Activity theory and Distributed Cognition: Or what does CSCW need to do with theories ?, *Computer Supported Cooperative Work* (CSCW), 11(1-2), 243-267,

Han J., Kamber M., 2006, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers,

Kuutti K., 1996, Activity theory as a Potential Framework for Human-Computer Interaction Research. *Context and consciousness: AT and human computer interaction*. B.A. Nardi (Ed.), MIT Press, Cambridge, MA. p. 17-44, 1996

Lewis R., 1997, An Activity theory framework to explore distributed communities. *Journal of Computer Assisted Learning*, 13: 210–218,

Lin J.H., Haug P.J. , 2006, Data preparation framework for preprocessing clinical data in data mining. In *AMIA Annual Symposium Proceedings* (Vol. 2006, p. 489). American Medical Informatics Association,.

Reimann P. & al, 2009 Using Process Mining to Identify Models of Group Decision Making in Chat Data. *Proceedings from CSCL*, 2009, Rhodes, Greece,

Romero C., Ventura S., 2007, Educational Data Mining: a Survey from 1995 to 2005. *Expert Systems with Applications*. Elsevier 1:33 135-146,

Simon J., Ralambondriny H., 2012 , Use of a CSCW platform by trainers and trainees Trace analysis: multimodal analysis vs. data mining approach, *ICLS2012, 10th International Conference of the Learning Sci*ences, Sydney Australia,

Stahl G., 2002 , Computer support for collaborative learning: Foundations for a CSCL community, *Lawrence Erlbaum Associates, Hillsdale*, NJ (2002), pp. 62–71,

Talavera L., Gaudioso E., 2004 , Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *In Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*, pp. 17–23,

Verbert K. & al, 2011, Dataset driven research for improving recommender systems for learning, *LAK '11, Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Pages 44-53, ACM New York, USA,