Data preprocessing using a priori knowledge

Jean Simon

Data preprocessing using a priori knowledge

Jean Simon
Université de La Réunion
IUFM de La Réunion Allée des Aigues Marines
97400 Saint Denis de La Réunion France
00262262904325
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 particularly interesting in Educational Data Mining.

Keywords

Activity theory, data preprocessing, preservice teacher

1 INTRODUCTION

This paper relies on the following approach:
1. Activity theory [1] is frequently used to study human activity specially CSCW and CSCL because this theory is particularly suitable to understand what the people do when they cooperate or collaborate.
2. One major problem in data-mining consists in the data preprocessing. Better those data are preprocessed, more information their treatment makes it possible to obtain.
3. The idea developed here is thus to rely on Activity theory to preprocess the data before their treatment. This preprocessing should make it possible to get more interesting results to study what occurs on a CSCW platform.

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.

2 METHODOLOGY

Activity theory (AT). As we can see on Figure 1, 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.

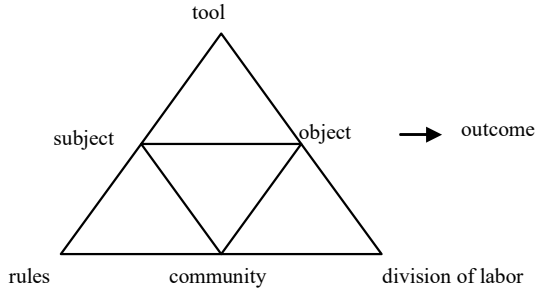

Figure 1. Triangles of AT according to Engeström [1].

Using Activity Theory to understand what happens on a platform is very common in the field of CSCL. For Halverson [2], it is powerful for, at least, three reasons: 1. this theory names well its theoretical constructs which are useful to manipulate data. 2. In this theory, the individual is at the center of everything and this is fundamental when we study learning. 3. Activity system diagrams highlight the processes and show both descriptive and rhetorical power.

Data mining. In education, the use of groupware and learning management system grows increasingly. Most of these systems record the traces of the users' activity. These traces constitute huge masses of data and permit to study the real activity of the subjects. Very broadly, the objective is to analyze what works and what does not in a given device to improve it. For Romero & al [3] there are four essential steps in data-mining: Collect data, Data preprocessing, Apply data mining, Interpret, evaluate and deploy the results. Data preprocessing is an important point for data mining because data tend to be incomplete, noisy and *inconsistent*.

Using AT to preprocess the data. What we propose is to use AT to preprocess the data and obtain a higher-level, representation. If we take the diagrams of Figure 1 we identify immediately three types readily available data:

- 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.

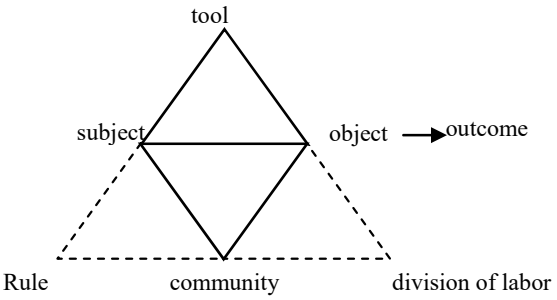

Figure 2. AT for data mining. Solid lines indicate data which are found among traces left on a CSCL platform and dotted lines information that it will be necessary to infer.

On any platform, it is recorded who are members of a particular group (dyad: subject-group) and who did what (dyad: subject-tool). From these two dyads, we can establish the third one (dyad: group-tool). If the node "objective" is rarely the subject of a specific trace, it is sometimes possible to find it through the name of the group or the name of the main folder that the group shares. The node "division of the labor" is rarely explicit and, concerning the last node, it is exceptional to have traces of the "rules within the group".

It is easy to see how technically we can preprocess the data. In the case mentioned above, instead of studying unique data, such as, for example, the "actions" alone carried out on the platform, we study couples "subject-action ", who is doing what on the platform. We can continue the process. Thus it is possible and desirable to work with 3-tuples "subject-group-action" and when you can identify the objective, 4-tuples "group- objective-subject- action" which correspond to the solid lines of figure 2. Once these 4-tuples are built, it is possible to put the focus on one of the nodes, the subject, the tool, the goal the group, without losing data interdependencies.

3 APPLICATION

The Reunion Island teacher training school trains the primary school teachers (PE2s : professeurs des écoles 2ème année). Since 2005, PE2s use 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 during 5 years: to deposit documents ("collective memory"), to improve lesson plans, to help online and at distance 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, we chose BSCW essentially because users may structure as they wish spaces they have created on it.

In BSCW, information is organized hierarchically in folders and sub-folders and is presented in the form of various documents (texts, pictures, URL ...) which are created, read, annotated, modified, restructured... So it is possible, to connect all the traces to the higher folder shared by a group and to reconstitute the 3-tuples (group, subject, action). Moreover, as a name is associated with each of these folders, name often indicating the purpose, 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.

Table 1. 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
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 PE2s'documents for one group	12	6	25	39
Average number of PE2 producers for one group	3	2	5	11
Average number of documents per producer for one group	4	3	5	4

Analysis according to the groups. We wanted to see what the actions were, depending on whether the group was composed only of PE2s or whether a teacher shared the group with them (PE2s+trainers). We made this distinction because we wanted to know if trainees would use the platform without being forced by the trainers. Here, we take into account only one action "document deposit".

The 1167 PE2s have constituted more than 960 groups. 668 were groups of PE2s only and 292 groups included at least one trainer. One PE2, of course, could be in several groups. We can see that 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, we find that the activity is much higher in groups shared with trainers than in groups shared only by PE2s in productions. There are fewer documents in the groups PE2s than in the groups PE2s+trainers (6 vs. 34). We can suppose that there is an effect "teachers" that incites students to work more.

Analysis according to the objective. With the titles of the folders, it was easy to isolate the groups named "TICE" (ICT for education). The folders associated with these groups were used to validate the C2i2e. All those groups have a trainer as member. In table 1, we have therefore compared those "TICE" groups with all the groups with trainers. As we can see, according to the objective, the actions on the platform reveal that activity is not the same. The "TICE" groups are smaller (13 members vs. 20), but in those groups, almost every PE2 works : 11/13 deposit. As there are more PE2s producers (11 vs. 5), there are also more documents in the "TICE" groups even if each PE2 producer deposits fewer documents (4 vs. 5).

4 CONCLUSION

The wealth of data mining relies on the fact that this is a bottom-up approach. The researcher starts from the data and expects that the machine will propose a categorization of these data of which he will try to find the underlying rules or, even better, that the machine itself will give rules of the categorization. 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 direct research in an unexpectedly way. However, it often happens that the proposed categorization is not exploitable for the researcher [4]. It is therefore desirable to reduce the hypothesis space that the machine is likely to return. One possible way to do this is to preprocess the data. In the context of CSCL, what we propose is to use Activity theory as a priori knowledge to do it. As we can see by this way we obtain results easily exploitable.

5 REFERENCES

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

[2] 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.

[3] Romero, C., Ventura, S., García, E. 2007 Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. *Computers and Education*, 51. pp. 368-384.

[4] 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)