Extraction d' Information Pédagogique à l'aide de Fouilles de Données: une étude de cas

Agathe Merceron

Département Génie Informatique
Ecole Supérieure d' Ingénieurs Léonard de Vinci, PULV
92916 Paris La Défense - Cedex (France)
agathe.merceron@devinci.fr
http://aldebaran.devinci.fr/~merceron

Résumé. Les systèmes d'apprentissage qui utilisent les TIC peuvent enregistrer sous forme électronique de nombreuses données. Ces données peuvent être fouillées par des logiciels adéquats pour en retirer des informations pédagogiques. Cet article illustre cette approche en prenant pour exemple le Logic-ITA, un système d'apprentissage pour les preuves formelles en logique.

1 Introduction

L'utilisation des Technologies de l' Information et de la Communication dans les systèmes d'apprentissage permet de recueillir de nombreuses données sous forme électronique, donc traitables par logiciels.

Les fouilles de données (Han et Kamber 2001) recouvrent des techniques diverses aussi bien dans les méthodes que dans les buts. Des logiciels de fouilles de données sont de plus en plus utilisés dans les entreprises commerciales, en particulier dans les banques et dans la téléphonie mobile. Le but de cet article est de montrer un exemple de leur utilisation dans l'enseignement pour en tirer des informations à but pédagogique. Les données utilisées sont le travail d' étudiants enregistré par le Logic-ITA, un outil tuteur en ligne dans le domaine de la logique des propositions.

2 Les données du Logic-ITA

Le Logic-ITA (Merceron et Yacef 2004b) est un logiciel accessible sur le Web qui permet aux étudiants de s'exercer à faire des dérivations formelles en logique des propositions. Il est utilisé à l' Université de Sydney depuis 2001. Un exercice en dérivation formelle est composé d'un ensemble de formules : les hypothèses et la conclusion. Le but d'un exercice est de dériver la conclusion à partir des hypothèses. Pour cela, l'étudiant doit dériver de nouvelles formules, pas à pas, en utilisant des règles de logique et en les appliquant aux formules déjà dérivées, ou aux hypothèses, jusqu'à ce que la conclusion soit obtenue. Il n'y a pas nécessairement une solution unique et tout cheminement valide est accepté. Le module expert vérifie que chaque pas entré par l'étudiant est valide, et donne un message d'erreur et éventuellement une indication si le pas est incorrect. Le Logic-ITA est en libre-service et est offert aux étudiants comme une ressource complémentaire au cours en face à face. En conséquence, il n'y a ni un nombre fixe ni un ensemble fixe d'exercices faits par tous les étudiants.

Le modèle d'un apprenant enregistre toutes ses réponses, ce qui comprend tous les pas qu'il a entrés, y compris les erreurs, pour la résolution d'un exercice. Un module permet au professeur de rassembler tous ces modèles dans une base données qui peut être interrogée et

fouillée. Deux tables sont particulièrement utiles, les tables *Mistake* et *Correct_Step*. Les principaux attributs de ces tables sont montrés en figure 1, *mistake* est propre à *Mistake*.

login	login de l'étudiant		*rule*	règle de logique utilisée
qid	identité de l'exercice		*startdate*	date de début de l'exercice
mistake	nom de l'erreur		f*inishdate*	date de fin d'exercice ou 0

FIG.1 - *Principaux attributs des tables Mistake et Correct_Step.*

3 Requêtes et analyse de données symboliques

Le chargé de cours interroge la base avec de simples requêtes pour savoir quelles sont les règles de logique les plus communément utilisées par les étudiants, les règles de logique avec lesquelles les étudiants commettent le plus d'erreur, les erreurs les plus communes etc. Cela sert non seulement pour avoir une idée du travail des étudiants, mais aussi pour préparer le cours de révision avant l'examen final.

Comme l'outil est en libre-service, nous voulions répondre à la question: *comment l'outil est-il utilisé ?* Pour cela nous avons fait de l'analyse de données symboliques (Diday 2000) avec l'outil SODAS (Sodas 2003) en regroupant dans un même objet les étudiants qui ont essayé de résoudre un nombre similaire d'exercices. Afin d'enrichir la description de ces objets, nous avons défini et calculé de nouveaux attributs tels que : le nombre d'exercices commencés par étudiant, le nombre d'exercices complètement résolus par étudiant, le nombre moyen de fautes commises par exercice par étudiant, le nombre moyen de pas corrects par exercice par étudiant. Les tables 1 et 2 sont issues de cette analyse. La table 1 montre les exercices finis, donc réussis, pour les différents objets. La table 2 donne le nombre moyen de pas corrects entrés par exercice pour les différents objets.

La deuxième ligne de la table 1, avec un 1 en première colonne, représente l'objet qui regroupe tous les étudiants qui n'ont essayé de faire qu'un seul exercice. Elle se lit ainsi : parmi les étudiants qui ont essayé de faire 1 exercice, 46% n'ont pas réussi à le finir et 54% l'ont terminé. La deuxième ligne, avec un 2 en première colonne, représente l'objet composé des étudiants qui ont essayé de faire 2 exercices. Elle se lit ainsi : parmi les étudiants qui ont essayé de faire 2 exercices, 13% en ont fini aucun, 23% en ont fini un, et 65% ont fini les deux. Et ainsi de suite. La cinquième ligne dit : parmi les étudiants qui ont essayé de faire 4, 5 ou 6 exercices, 4% n'en ont fini aucun, 8% en ont fini un, 27% en ont fini 2, etc, 2% ont fini les 6.

Fin.	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	19	20	26
1	46	54																
2	13	23	65															
3	6	11	39	44														
4-6	4	8	27	19	29	10	2											
7-10	3		6	18	36	12	18	3	3									
11-15			16	16	16	21	5	5			11		5	5				
16 +			17												17	17	33	17

TAB.1 - *Objets symboliques et nombre d'exercices finis.*

La deuxième ligne de la table 2 se lit ainsi : parmi les étudiants qui ont essayé de faire un seul exercice, 46% n'ont pas réussi à entrer un seul pas correct (ceux qui n'ont pas non plus réussi à finir l'exercice), 8% ont entré 2 pas corrects, 8% ont entré 3 pas corrects, etc., 23% ont entré plus de 10 pas corrects. La quatrième ligne se lit ainsi : parmi les étudiants qui ont

essayé de faire 3 exercices, aucun n'a entré 0, 1 ou 2 pas corrects, 6% a entré en moyenne 3 pas corrects par exercice, etc.

R_cor.	0	1	2	3	4	5	6	7	8	10	10 +
1	46		8	8				8	8		23
2	3	3	6	6		10	3	10	26	23	10
3				6	11	22	11	11	17	17	6
4-6	2	2		19	13	17	27	8	6	4	2
7-10		3	3	15	9	27	33	6	3		
11-15		11	5	26	21	21		11		5	
16 +				17	17	17	33	17			

TAB.2 - *Objets symboliques et nombre de pas corrects.*

Les objets symboliques ont été choisi d'une part pour comprendre ce qui se passe quand les étudiants essaient peu d'exercices (1, 2 ou 3), puis pour suivre en gros les quartiles. Cette analyse est présentée intégralement dans (Merceron et Yacef 2004b). Elle montre clairement la tendance suivante : plus un étudiant pratique, plus il va jusqu'au bout des exercices commencés, plus il entre de pas corrects et moins il fait de fautes.

4 Associations

Les règles d'association, une autre technique des fouilles de données, sont souvent appliquées au chariots de supermarché et ont la forme suivante: a,b -> c, support 45%, confiance 60%. Cette règle se lit ainsi ainsi : si un client achète les produits a et b, alors il achète aussi le produit c. Cette association concerne 45% des clients et la confiance dans la causalité est de 60%. Nous avons appliqué ce principe aux erreurs commises par les étudiants. La question pédagogique posée est : *y-a-t-il des erreurs souvent faites ensemble par les étudiants lors de la résolution d'un exercice ?* Les associations trouvées en 2002 sont présentées en Figure 2.

Rule	*Wrong*	*Premise*		*Rule*	*Wrong*	*Premise*	*Support*	*Conf.*
		X	->	X			61.00%	79.00%
X			->			X	61.00%	82.00%
	X		->	X			65.00%	80.00%
X			->		X		65.00%	87.00%
	X		->			X	67.00%	83.00%
		X	->		X		67.00%	87.00%

FIG.2 - *Associations d'erreurs.*

L'erreur *Rule* veut dire que la règle de logique peut-être utilisée mais que l'étudiant a écrit une formule non correcte. L'erreur *Wrong* veut dire que l'étudiant s'est trompé dans les numéros des formules auxquelles la règle de logique s'applique. L'erreur *Premise* indique que l'étudiant s'est trompé en indiquant les hypothèses impliquées. La première ligne se lit

ainsi : si les étudiants font l'erreur *Rule* lors de la résolution d'un exercice, ils font aussi l'erreur *Wrong*.

Ces erreurs sont typiques de l'apprentissage des dérivations formelles. La raison d'être de la mise en service du Logic-ITA est d'aider les étudiants à acquérir plus de rigueur. Cependant, la liaison entre *Wrong* et *Rule* nous a amené à modifier le cours de façon à mieux présenter les deux sortes de règles de logique, celles qui s'appliquent à une seule formule et celles qui s'appliquent à deux formules. L'année suivante, après ces modifications apportées au cours, nous avons fait deux constatations : (i) les associations d'erreurs sont restées stables, (ii) les notes sur l'exercice de dérivation logique à l'examen final ont continué à augmenter (en moyenne). Nous en déduisons que ces associations de fautes font tout simplement partie de l'apprentissage, surtout quand il s'agit d'un outil d'entraînement comme le Logic-ITA.

5 Classification

Le but de la classification, une autre technique des fouilles de données, est de diviser une population en sous-groupes homogènes. La question pédagogique posée est : *peut-on diviser les étudiants en difficultés en sous-groupes homogènes ?* Les étudiants en difficulté sont ceux qui commencent un exercice sans aller jusqu'au bout. Nous avons donc sélectionné cette sous-population. Nous les avons classifié selon la classification par centres mobiles seule, puis combinée avec la classification ascendante hiérarchique (Merceron et Yacef 2004a). Ces deux classifications reposent sur une notion de distance entre les individus. Comme il n'y a pas d'ensembles d'exercices essayés par tous les étudiants, il n'est pas aisé de définir une distance. Nous avons finalement opté pour le nombre total de fautes. Deux étudiants ayant

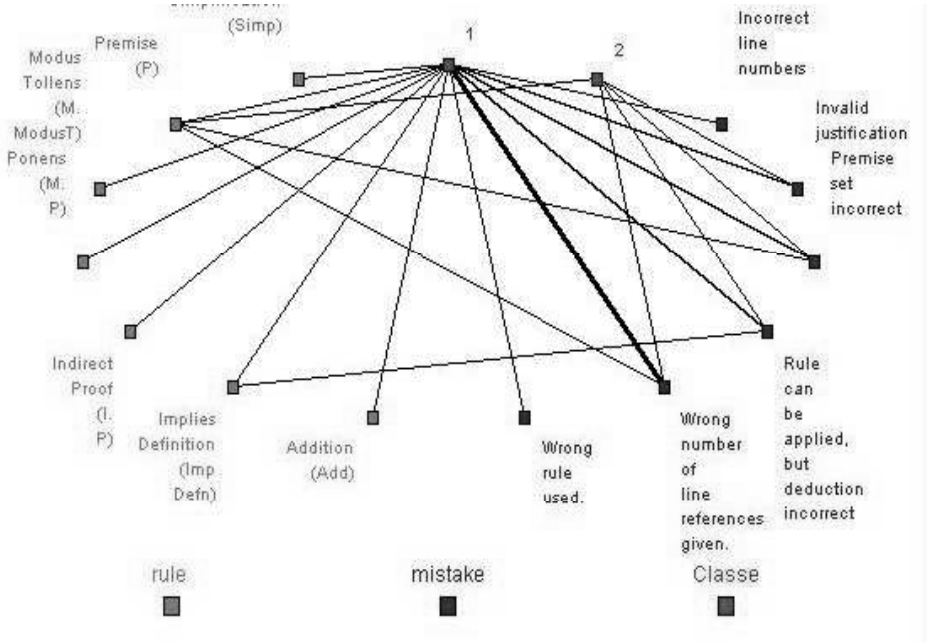

FIG.3 - *Relations entre groupes, erreurs et règles de logique utilisées.*

fait tous deux beaucoup de fautes seront donc assez proches. Différentes visualisations des groupes obtenus nous ont permis de faire une découverte intéressante.

La figure 3 montre une visualisation des groupes. Le groupe 1 représente les étudiants qui font beaucoup de fautes. Ils sont ceux qui utilisent aussi le plus de règles de logique. Un autre graphique (non montré ici) nous permet de constater qu'ils n'essaient pas plus d'exercices que les étudiants du groupe 2, ceux qui font peu de fautes. Cela suggère, qu'en partie, ils essaient les règles du menu déroulant les unes après les autres, jusqu'à temps de trouver la bonne. Un telle utilisation du système appelle une attitude différenciée du chargé de cours.

6 Conclusion

Les différentes requêtes et fouilles effectuées sur les données du Logic-ITA ont été et sont très utiles pour savoir où se trouvent les difficultés des étudiants et comment les étudiants utilisent le système. Elles nous ont amené à faire des changements dans le cours, que nous traduisons par des améliorations puisque les notes sur l'exercice de dérivation formelle à l'examen final continuent à progresser depuis l'introduction du Logic-ITA.

Il serait possible de faire des requêtes et fouilles plus précises. Par exemple, pour l'analyse de données symboliques, il serait souhaitable de définir des attributs supplémentaires comme le nombre de fautes par session de travail (en plus du nombre de fautes total), le nombre de pas corrects par session et, peut-être, faire ainsi des déductions sur la maîtrise de la matière, les dérivations formelles, au fur et à mesure que l'outil lui-même est mieux maîtrisé. Cependant, l'enregistrement actuel des données par le Logic-ITA ne permet pas de distinguer deux sessions différentes ayant lieu le même jour.

L'enregistrement de données pertinentes, leur analyse, leur fouille pour aider à mieux supporter les différents acteurs de l'enseignement est un sujet de recherche où un certain savoir-faire commence à se dégager, voire par exemple (Beck 20004). Des efforts supplémentaires, comme le projet (DPULS 2005), sont nécessaires pour que des modèles fassent consensus, soient adoptés et facilitent l' évolution des systèmes suivant les usages observés.

Références

Beck, J. (2004) ed. Proceedings of ITS2004 workshop on Analysing Student-Tutor Interaction Logs to Improve Educational Outcomes. Maceio, Brazil.

Diday, E. (2000) Analyse des données symboliques: théorie et outil pour la fouille de connaissances. In : *TSI (Technique et Science Informatiques).* Vol 19, n°1-2-3 , Janvier 2000.

DPULS (2005) http://www.noe-kaleidoscope.org/pub/patterns/index.html

Han J.W. and Kamber M. (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Merceron, A. and Yacef, K. (2004)a Clustering Students to help Evaluate Learning. In : *Proc. of the Int'l Workshop on Technology Enhanced Learning*, pp. 31-42. Toulouse, France, August 2004, Kluwer.

Merceron, A. and Yacef, K. (2004)b Mining Student Data Captured from a Web-based Tutoring Tool: Initial Exploration and Results. In : *Computational Intelligence in Web-Based Education* 4(15), 319-346.

Sodas (2003) http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm

Summary

Technology enhanced learning systems make it possible to store users' data into electronic form. These data can be mined with appropriate software to extract pedagogically relevant information. In this paper we illustrate this approach with the Logic-ITA, a web base tutoring system in the domain of formal proofs.