ÉVALUATION PAIR A PAIR PARTICIPATIVE (P³E) DE LA PLATE-FORME POEM

Pierre COLLET[1,2] (Professeur des Universités), Raaj SEEREEKISSOON[2] (Étudiant), Anna SCIUS-BERTRAND[2,3] (Étudiante), Rachel STEIN[2,4] (Étudiante), Pierre PARREND[1,2,5] (Enseignant-Chercheur)

RESUME

Cette communication décrit la procédure d'évaluation participative pair à pair P³E (Participative Peer-To-Peer Evaluation) de la plate-forme POEM (Personalised Open Education for the Masses) développée à l'Université de Strasbourg dans le cadre du *Complex Systems Digital Campus*, réseau UniTwin de l'UNESCO comportant 115 universités dans le monde. La procédure P³E offre une évaluation automatique de haut niveau car elle implique fortement les apprenants, qui doivent proposer d'eux-mêmes des questions sur les contenus proposés. Le protocole P³E peut être utilisé indépendamment de la plate-forme POEM.

MOTS CLÉS

Évaluation pair-à-pair, Système de Gestion de l'Apprentissage, Environnement Informatique pour l'Apprentissage Humain

INTRODUCTION

Dis-moi et j'oublie.
Explique-moi et je me souviens.
Implique-moi et j'apprends.

BENJAMIN FRANKLIN

Dans les temps anciens, l'éducation était personnalisée par l'intermédiaire des précepteurs. Le 19ème siècle a vu se développer un enseignement de masse avec l'école publique pour tous. Dans le primaire et le secondaire, elle permet à un professeur d'enseigner à des classes entières de 30 à 40 élèves et dans le supérieur, à des amphithéâtres pouvant comporter plusieurs centaines d'étudiants. Avec le XXIème siècle et la révolution numérique, tout un chacun peut avoir accès à la quasi-totalité des savoirs, par le biais des ordinateurs et de l'internet. Ainsi, il n'est plus besoin d'assister physiquement à un cours ou à un séminaire, ni d'être dans la même pièce pour résoudre un problème difficile avec d'autres. Depuis quelques années, des millions d'étudiants se sont inscrits dans des MOOCs (Massive Open Online Courses) proposés par les universités les plus renommées. Tout possesseur d'un ordinateur, d'une tablette voire d'un portable

[1] *Université de Strasbourg / Laboratoire ICube – France*
[2] *Complex Systems Digital Campus UNESCO UniTwin / e-Laboratory on Education*
[3] *École Pratique des Hautes Études, laboratoire LUTIN – France*
[4] *Swarthmore College in Swarthmore, Pennsylvania – USA*
[5] *ECAM Strasbourg-Europe – France*

connecté peut maintenant assister à un cours ou à un séminaire mis à disposition sur le web. Il peut faire les exercices correspondant à chaque étape d'un cours et les problèmes à la fin d'un chapitre de cours. Il peut s'attaquer à un problème vraiment difficile ensemble avec d'autres apprenants à distance dans le même réseau social. Il peut enfin, seul ou avec d'autres dans le même réseau social, proposer de nouveaux exercices et problèmes de différentes difficultés.
Malheureusement, là où des « petits » effectifs (jusqu'à quelques centaines d'étudiants ?) permettaient encore aux enseignants de corriger eux-mêmes les copies, un MOOC va pouvoir s'adresser à 100 000 participants, rendant impossible toute évaluation de la part de l'enseignant ou de l'équipe enseignante. La plate-forme POEM [1] développée à Strasbourg dans le cadre de FuturICT [2] comme un prolongement de la thèse de Grégory Valigiani dirigée par Pierre Collet [3,4] permet de mettre en œuvre une éducation numérique de masse dite 4P (Participative, Prédictive, Préventive, Personnalisée) devant à terme être utilisée par les étudiants des 115 universités de l'UniTwin CS-DC de l'UNESCO [5]. Il est absolument nécessaire qu'elle propose une évaluation pair à pair, qui est la seule possible lorsqu'il faut gérer un très grand nombre d'étudiants.
Cela dit, il est intéressant de noter que l'évaluation P^{3}E proposée ci-dessous fonctionne aussi avec à des petits groupes de 30 à 40 apprenants. L'allègement, du travail de correction de l'équipe pédagogique peut alors permettre de proposer plus d'évaluations (typiquement une évaluation après chaque cours) pour une meilleure évaluation continuée.
Actuellement le protocole P3E est expérimenté sur des cohortes d'étudiants de l'Université de Strasbourg, où les différences entre l'évaluation par les enseignants et l'évaluation automatique P3E sont mesurées par l'inégalité de Bienaymé-Chebyshev [8]. Faute de place pour les insérer dans cette communication, les résultats obtenus seront présentés lors du colloque.

L'EVALUATION PAR LES PAIRS

L'évaluation par les pairs est une solution explorée depuis longtemps [16] mais mise en œuvre plus récemment par des plates-formes de formation à distance comme Spark[6] (Self-and Peer Assessment Resource Kit) ou plus récemment dans des MOOCs [9, 15]. Elle est assez fréquemment utilisée dans le cadre d'enseignements opérationnels comme le management ou le développement logiciel [10], mais reste dirigée de l'enseignant vers l'apprenant, puisque les sujets sont actuellement toujours imposés aux étudiants et statiques. Dans les évaluations de ces enseignements techniques, elle offre une marge de déviation entre les évaluations par les pairs et les évaluations par les enseignants qui peut être inférieure à 3% [11], ce qui en fait donc une évaluation très pertinente (au moins en informatique, cadre de cette expérience). Elle est efficace également pour l'évaluation de tâches complexes comme la rédaction, en particulier lorsqu'elle est associée à un coaching approprié [12], ce qui est plus difficile à mettre en place dans le cadre de MOOCs. L'évaluation par les pairs est souvent mieux acceptée par les étudiants que l'évaluation par les enseignants. Elle leur permet de mieux progresser aussi bien en ce qui concerne le contenu du module que sur leur propre capacité à évaluer les autres.
La limitation des solutions proposées est le caractère statique des bases de questions utilisées pour l'évaluation, ce qui augmente en particulier le risque de fraude. En incluant dans l'évaluation la nécessité de poser une question (qui sera posée aux autres apprenants), l'évaluation P^{3}E répond à ce problème en transformant les apprenants en autant de producteurs de connaissances et d'analyses nouvelles, qui s'auto-alimenteront dans un cercle virtueux.
P^{3}E intègre donc dans les xMOOCs transmissifs et mieux adaptés à la transmission de savoirs spécifiques [13], des éléments dynamiques des cMOOCs [14], connectivistes fondés sur l'expérience individuelle et l'interaction entre les apprenants.

CONDITIONS NECESSAIRES A LA MISE EN ŒUVRE DE L'EVALUATION P^{3}E

6 http://spark.uts.edu.au/

L'évaluation P³E nécessite :

- Une base de contenus pédagogiques, quelle que soit leur forme (vidéo, contenu en ligne, papier, cours en présentiel, …).
- Deux bases de données contenant deux jeux de questions/réponses associées à un contenu pédagogique :
  1. Une base *archive* contenant un jeu de questions (accompagnées de leurs réponses) validées par l'équipe pédagogique. Les questions sont ouvertes ou fermées, mais aussi possiblement des exercices voire des QCM. Leur objectif est d'évaluer ce qu'ont compris les participants ayant suivi le contenu pédagogique proposé. L'archive doit comporter un minimum de 4 questions pour démarrer le processus d'auto-évaluation participatif (mais une dizaine de questions sont bienvenues). L'archive sera par la suite augmentée par l'équipe pédagogique lorsque des questions remarquables sont trouvées dans…
  2. … une base *réservoir* comportant des questions (accompagnées de leurs réponses) *proposées par les participants* dans le cadre de l'évaluation participative. Poser des questions sur un contenu est une activité pédagogique riche impliquant le participant : elle nécessite à la fois une compréhension globale des enjeux du cours et une bonne analyse des problèmes spécifiques exposés dans le cours. Cela impose une implication bien plus grande du participant que ce qui est nécessaire pour simplement répondre à une question (même ouverte) ou un QCM. De plus, cela permet de disposer d'une base de questions sans cesse renouvelée.
- Une cohorte d'une trentaine de participants au minimum est nécessaire, car c'est à partir de 30 que la loi des grands nombres commence à s'appliquer[7] (les casinos commencent même à 20 !). Il est intéressant de noter qu'il est habituel d'avoir de 20 à 30 apprenants dans une salle de classes.

FONCTIONNEMENT DE L'EVALUATION P³E

L'évaluation P³E s'effectue en trois étapes :

1. **L'apprenant répond à, et évalue trois questions (et leurs réponses)** : deux sont choisies au hasard dans l'archive et une choisie par un sélecteur stochastique [6] dans le réservoir de questions proposées (lors de l'amorçage de l'évaluation P³E, une troisième question de l'archive sera utilisée s'il n'y a pas encore de questions dans le réservoir). Ce principe est inspiré des re-captchas [7]. Après avoir répondu à la question dans une zone de texte, on demande à l'apprenant d'évaluer la pertinence/qualité de la question en la notant de 0 à 5, avec la contrainte de répartir ***exactement*** 10 points sur les 3 questions. Comme 10 n'est pas divisible par 3, cela force l'apprenant à s'impliquer en l'empêchant de mettre la même note à toutes les questions. L'évaluation de la pertinence des questions permet d'évaluer la qualité de la participation de l'étudiant ayant proposé la question du réservoir. Si les trois questions sont bonnes, le participant mettra 3 3 4 (par exemple) et toutes les questions auront plus de la moyenne, (avec un avantage à la meilleure question du lot). Si les trois questions sont mauvaises… en fait, ceci n'est pas possible car au moins deux questions proviennent de l'archive. Si une question est très mauvaise au point où il est impossible d'y répondre (question mal posée), l'étudiant lui donne 0 et une nouvelle question lui est alors proposée parmi l'archive. Mais cette possibilité ne lui sera donnée qu'une seule fois. S'il met à nouveau un 0 à une question, le nombre de points qu'il pourra distribuer passera à 7. L'apprenant ne sera pas évalué sur une question notée 0.
   Après avoir répondu aux trois questions, les réponses proposées sont montrées aux participants qui devront là aussi noter leur pertinence et leur qualité de la même manière que pour les questions. L'obtention d'un 0 à un couple question/réponse posé sera bien évidemment pénalisant, ce qui motivera donc les apprenant à faire des propositions sérieuses.
   Une case à cocher est aussi proposée pour permettre à l'apprenant de signaler que le contenu sémantique de deux questions posées est très proche. Si les questions de l'archive (validées par

[7] Cette valeur est bien connue des statisticiens. Cf. http://mathematiques.ac-bordeaux.fr/profplus/docmaths/statistiques/artigues/chapitre6.pdf par exemple.

l'équipe pédagogique) sont toutes bien distinctes, on pourra en conclure que sur les deux questions signalées comme trop proches, l'une provient d'un participant (base réservoir). On notera pour cette question du réservoir la proximité détectée avec la question de l'archive, avec pour conséquence qu'on ne posera pas à l'avenir ces deux questions dans un même jeu de 3 questions. Dans le cas où l'apprenant a signalé deux questions proches, on piochera une autre question **dans l'archive** pour laisser au participant la possibilité d'évaluer la question du réservoir en provenance d'un autre apprenant.

2. **L'apprenant pose à son tour une question** sur le contenu pédagogique **et propose une réponse**. La question est ajoutée au réservoir de questions non validées par l'équipe enseignante et fera partie des questions posées à un autre apprenant.
   La question posée par l'apprenant doit tester la compréhension et demander une réflexion sur les points les plus importants du contenu pédagogique. La réponse proposée doit être correcte aussi. Cette démarche est en elle-même d'une forte valeur pédagogique, car on ne peut poser de question pertinente et y proposer une réponse sans avoir intégré quelques éléments du contenu du cours, sachant que la pertinence de la question et de la réponse proposée seront évaluées par les autres participants dans l'étape 1. Les questions/réponses ramenant des 0 signaleront une faible implication de l'apprenant à l'origine de ces questions/réponses.
3. **L'apprenant évalue 9 réponses provenant d'autres apprenants.** L'évaluation de la qualité scientifique d'un article se fait par une procédure d'évaluation en double aveugle requérant l'évaluation de trois relecteurs de même niveau scientifique que l'auteur de l'article soumis. Ce protocole largement accepté pour établir l'état de l'art mondial en sciences est repris pour l'évaluation pair à pair entre apprenants. Si chaque réponse proposée en étape 1 doit être évaluée par trois autres apprenants, trois réponses nécessitent 9 évaluations au total. On demande donc dans cette étape à l'apprenant d'évaluer les réponses de 9 autres participants. Si, dans la phase d'amorçage du système, il n'y a pas assez de réponses pour demander 9 évaluations à un participant, celui-ci devra revenir plus tard sur la plate-forme pour terminer sa troisième étape d'évaluation du contenu pédagogique.
   Tout comme dans l'étape 1, on propose à l'apprenant-évaluateur de noter sur 5 les 9 réponses provenant d'autres apprenants ***avec exactement 30 points à distribuer***. Ainsi, on évite que les participants se donnent tous une valeur maximale de 5, ce qui invaliderait le système de notation automatique. 30/9 = 3,33… ce qui est donc la moyenne générale du groupe (qui ramenée sur 20, correspond à 13,33, ce qui est une moyenne honorable). Comme 30 n'est pas divisible par 9, l'apprenant évaluateur doit ici aussi faire un choix et l'équipe pédagogique pourra décider de donner des points de bonus aux apprenants qui auront le mieux évalué leurs collègues, en rapprochant les notes données des notes obtenues par les participants notés (réglages paramétrables). Si un participant n'a donné que des 3 et des 4, il n'aura clairement pas pris position et n'aura pas fait son travail d'évaluateur correctement. Mais il ne s'agit pas d'encourager l'apprenant évaluateur à augmenter l'écart-type de sa notation au détriment de la justesse de l'évaluation : on pourra corréler les notes données au niveau final obtenu par les participants notés. Ainsi, une absence de corrélation entre le niveau final des participants noté et les notes données par un apprenant évaluateur pourront indiquer que ce dernier n'a pas fait preuve de sérieux dans sa mission d'évaluateur.

NOTATION DES PARTICIPANTS A L'ISSUE DE L'EVALUATION P³E

On peut noter que la procédure P³E proposée mêle évaluation des connaissances de l'apprenant et participation au processus d'évaluation. On proposera donc que le système renvoie deux notes : l'une sur l'évaluation des connaissances et l'autre sur le degré d'implication de l'apprenant au processus d'évaluation participatif. L'équipe pédagogique pourra par la suite décider de pondérer les deux notes obtenues pour façonner l'évaluation finale de l'apprenant, sachant que la note d'implication dans le processus a aussi un caractère pédagogique fort : en effet, elle sanctionnera les apprenants qui auront fourni des

questions/réponses inintéressantes voire qui ne font pas sens, s'ils n'ont pas compris le contenu pédagogique proposé. Si l'équipe pédagogique donne une pondération non négligeable à la note de participation, l'apprenant aura un grand intérêt à avoir bien compris le contenu pédagogique pour ne pas soumettre de questions incohérentes, qui auront de grandes chances d'être sanctionnées par un 0 par les autres participants qui auront à répondre à ces questions.

MISE A JOUR DES BASES DE QUESTIONS

Une fois l'évaluation terminée, l'équipe pédagogique pourra enrichir l'archive avec des bonnes questions proposées par les apprenants, en y apportant des réponses validées pour préparer une future évaluation du même contenu avec une base de questions renouvelées.

CONCLUSION SUR LA PROCEDURE AUTOMATIQUE P³E PROPOSEE ET DEVELOPPEMENTS FUTURS

La procédure proposée s'appuie sur la loi des grands nombres pour fournir une évaluation pair à pair aussi juste et fiable que possible de l'assimilation de contenus pédagogiques par un très grand nombre d'étudiants. D'autre part, pour renforcer l'acquisition des connaissances, on demandera aux étudiants de participer activement au processus d'évaluation en leur demandant de poser de nouvelles questions, pour augmenter la base de questions initialement fournies par l'équipe enseignante.
De temps en temps, certains étudiants n'auront « pas de chance » alors que d'autres tomberont sur des questions faciles. La solution à ce problème réside dans le grand nombre d'interactions demandées aux participants, ce qui va ainsi réduire les cas de malchance /chance récurrente (loi des grands nombres). De même, la triche est rendue difficile par ce même nombre d'interactions et de notation complexe, car le participant doit non seulement répondre à trois questions, mais aussi évaluer la qualité des questions, évaluer la qualité des réponses, puis proposer une question, puis évaluer les réponses des autres intervenants, …). L'évaluation de chaque cours ne peut se faire de manière passive, comme ce qui peut se passer si on donne simplement à cliquer les cases d'un QCM ou même de répondre à des questions. Le protocole P³E nécessite une participation cognitive de haut niveau de la part de l'apprenant qui non seulement permet d'obtenir des évaluations de grande qualité, par l'implication requise, mais permet aussi à l'apprenant de mieux assimiler le contenu du cours, ce qui doit rester l'objectif primaire de toute activité pédagogique.
Le protocole P³E (ainsi que la plate-forme POEM pour laquelle il a été conçu) est disponible en open-source sur demande. Il a fait l'objet d'un hackathon national organisé par France Université Numérique pour en faire un plug-in à Open-edX. Dès que le plug-in sera complètement opérationnel et certifié par Open-edX, il pourra être mis en œuvre gratuitement sur toutes les plate-formes Open-edX nationales et internationales.

BIBLIOGRAPHIE

[1] Louca, J., Johnson, J., Bourgine, P., Portelli, P., Tijus, C., Scius-Bertrand, A., Lenhard, W., Escalona, M., Taramasco, C., Kohlhase, M., Cointet, J., Collet P. (2013). Poem Platform For Massive Personalized Education. Dans ECCS'13, Barcelona, Spain.
[2] Johnson, J., Buckingham, S., Willis, A., Bishop, S., Zamenopoulos, T., Swithenby, S., Mackay, R., Merali, Y., Lorincz, A., Costea, C., Bourgine, P., Louca, J., Kapenieks, A., Kelley, P., Caird, S., Deakin, R., Goldspink, C., Collet, P., Carbone, A., Helbing, D. (2012). The FuturICT education accelerator. Dans European Physical Journal - Special Topics, Springer (IF : 1.76), (p. 215-243), Volume 214, No 1, doi:10.1140/epjst/e2012-01693-0.
[3] Valigiani, G. (2006). Développement d'un paradigme d'Optimisation par Hommilière et application à

l'Enseignement Assisté par Ordinateur sur Internet. Thèse de l'Université du Littoral Côte d'Opale, 2006.
[4] Valigiani, G., Jamont, Y., Biojout, R., Lutton, E., Fonlupt, C., Collet, P. (2007). Optimisation par Hommilières de chemins pédagogiques pour un logiciel d'E-learning. Dans Techniques et Sciences Informatiques, 01/2007; 26 (p. 1245-1267).
[5] http://cs-dc.org
[6] Blickle, T., et Thiele, L. (1996). A comparison of selection schemes used in evolutionary algorithms. Dans Evolutionary. *Computing*. 4, 4 (p. 361-394). DOI=10.1162/evco.1996.4.4.361
[7] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., et Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Dans Science 321 (5895): (p. 1465-1468). doi:10.1126/science.1160379.
[8] Tchébychef, P.-L. (1867). Des valeurs moyennes. Dans Journal de Mathématiques pures et appliquées, 2e série, XII (p. 177-184). http://sites.mathdoc.fr/JMPA/PDF/JMPA_1867_2_12_A11_0.pdf
[9] Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. Research & Practice in Assessment, 8(1), 40-48.
[10] Boud, D., Cohen, R., & Sampson, J. (Eds.). (2014). Peer learning in higher education: Learning from and with each other. Routledge.
[11] Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., ... & Klemmer, S. R. (2015). Peer and self assessment in massive online classes. In Design Thinking Research (pp. 131-168). Springer International Publishing.
[12] Stanley, J. (1992). Coaching student writers to be effective peer evaluators.Journal of Second Language Writing, 1(3), 217-233.
[13] Rodriguez, O. (2013). The concept of openness behind c and x-MOOCs (Massive Open Online Courses). Open Praxis, 5(1), 67-73.
[14] Tschofen, C., & Mackness, J. (2012). Connectivism and dimensions of individual experience. The International Review of Research in Open and Distributed Learning, 13(1), 124-143.
[15] Lebrun, M. (2015). L'hybridation dans l'enseignement supérieur : vers une nouvelle culture de l'évaluation ?, dans Journal international de Recherche en Education et Formation, 1(1), pp. 65-78
[16] Tardif, J. (2006). L'évaluation des compétences : Documenter le parcours de développement. Montréal : Chenelière Éducation.