



Sujet de thèse : Analyse des trajectoires d'apprenants par IA pour l'assistance à l'enseignement de la programmation

Contexte. L'utilisation de plates-formes d'apprentissage s'est considérablement développée ces dernières années au sein des formations en informatique, en particulier pour l'enseignement des fondamentaux de l'algorithme et de la programmation. Ces plates-formes permettent de retracer finement les étapes par lesquelles passe un apprenant pour construire un programme, appelées aussi sa « trajectoire ». L'analyse de ces trajectoires ouvre ainsi des perspectives intéressantes en matière de suivi et d'aide à l'apprentissage de la programmation.

Description du sujet. L'objectif de la thèse est de développer des techniques d'Intelligences Artificielles (IA) permettant de dépasser la simple analyse syntaxique de chaque programme soumis, et de considérer non seulement la sémantique associée mais aussi son évolution dans le temps, au fil des tentatives d'un apprenant.

Pour cela, l'un des problèmes majeurs réside dans la nature complexe des programmes informatiques. En effet, les programmes sont décrits dans des langages formels décrivant des actions (potentiellement différentes en fonction du contexte d'utilisation) à faire par l'ordinateur pour réaliser une certaine fonctionnalité. Ces données sont donc complexes à analyser et diffèrent des données classiquement traitées par les méthodes d'IA (généralement des vecteurs de valeurs). Ensuite, l'analyse de séries temporelles (les « trajectoires » des apprenants pour résoudre les exercices) pour en extraire des régularités, ou des comportements surprenants, reste un problème très difficile en IA.

Considérant que les soumissions d'apprenants (et leurs trajectoires) renferment des informations cachées permettant d'identifier des situations caractéristiques du point de vue pédagogique (blocage, décrochage, mauvaises pratiques, parcours de résolution types, etc.) cette thèse explorera différentes stratégies d'analyse et de détection de ces situations cibles.

Nous nous intéresserons aux techniques d'IA basées sur la recherche de sous-groupes (« subgroup discovery ») (Atzmueller, 2015) comme moyen pour analyser ces séries temporelles de programmes. Cette famille de méthodes vise à trouver et caractériser des sous-ensembles d'objets intéressants par rapport à une classe à étudier (p.ex. la réussite à un exercice). Elle a été appliquée à de nombreux domaines d'applications (p.ex. sport, santé, réseaux sociaux). Toutefois, peu de travaux ont étudié la recherche de sous-groupes dans des données temporelles. L'un des objectifs de la thèse sera donc de développer une méthode de recherche de sous-groupes intégrant les spécificités de la dimension temporelle et des programmes.

Il s'agira également d'étudier les techniques récentes en matière d'apprentissage de représentations de programmes (ou *code embeddings*) (Wang et al., 2018 ; Alon et al., 2019 ; Cleuziou&Flouvat, 2021) et d'envisager leur exploitation pour l'analyse des trajectoires d'apprenants dans des contextes supervisés et non-supervisés. Nous envisagerons dans un deuxième temps l'apprentissage de représentations de trajectoires (ou *trajectory embeddings*) par des approches de réseaux de neurones récurrents (type LSTM, GRU, ELMO) afin de comparer ces deux stratégies de représentation.

Le volet fondamental de la thèse sera adossé à des préoccupations applicatives sur des données d'éducation visant le développement d'environnements de « Pédagogie Augmentée » orientés « Enseignant ». Il s'agira notamment d'identifier des tâches support sur lesquelles l'enseignant pourrait être assisté et de les mettre en œuvre dans un processus de Recherche & Développement s'intégrant aux outils numériques utilisés dans les formations de l'établissement.

Mots-clés. Intelligence Artificielle, Fouille de Données d'Education, Analyse de Trajectoires, Analyse de sous groupes, Apprentissage de Représentations, Réseaux Neuronaux (Récurrents, Profonds).

Profil attendu. Diplômé d'un master et/ou d'une école d'ingénieur en Mathématiques et/ou Informatique. Une expérience en Machine Learning et un intérêt pour l'analyse de données éducatives seraient un plus.

Encadrement de la thèse. La thèse se déroulera au laboratoire ISEA (Institut des Sciences Exactes et Appliquées) de l'Université de la Nouvelle-Calédonie. Elle sera dirigée par Frédéric FLOUVAT (Maître de Conférences, Habilité à Diriger des Recherches en Informatique) en collaboration avec Guillaume CLEUZIOU (Professeur des Universités en informatique) de l'Université d'Orléans.

Financement de la thèse. Bourse de l'Ecole Doctorale du Pacifique, Université de la Nouvelle-Calédonie.

Candidatures. Les candidatures (CV, notes et lettre de motivation) doivent être adressées avant le 4 juillet 2021 à Frédéric Flouvat (frederic.flouvat@unc.nc) et Guillaume Cleuziou (guillaume.cleuziou@univ-orleans.fr).

Références.

- Alon, U., M. Zilberstein, O. Levy, and E. Yahav (2019). code2vec : Learning distributed representations of code. Proceedings of the ACM on Programming Languages 3(POPL), 1–29.
- Atzmueller, M. (2015). Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(1), 35-49.
- Cleuziou, G. and F. Flouvat (2021). Learning student program embeddings using abstract execution traces, (to appear) in International Conference on Educational Data Mining (EDM'2021).
- Wang, K., R. Singh, and Z. Su (2018). Dynamic neural program embeddings for program repair. In International Conference on Learning Representations (ICLR'2018).