Généralisation de motifs séquentiels pour la fouille de données multi-sources

Contexte général

Une grande quantité de données numériques sont créées au quotidien. Chaque jour, plus de 2,5 quintillions d'octets de données sont créés. Les données sont devenues importantes et indispensables dans notre vie quotidienne, où 90% des données mondiales ont été créées au cours des deux dernières années[1], et il est estimé que 1,7 mégaoctet de données seront créés chaque seconde pour chaque être humain sur l'année 2020[2]. Cette croissance du volume de données produites est le résultat de l'informatisation de notre société et du développement rapide des outils de collecte et de stockage de données [1]. Cela nous permet de dire que nous sommes à l'heure du numérique. Les données numériques existent dans tous les domaines de la recherche et industriel; tels que le e-commerce, la santé, la banque, l'éducation, etc.

En conséquence, des données de grands volumes peuvent être facilement collectées, et la variété de ces données est aussi importante que leur volume. En effet, les données peuvent être collectées à partir de plusieurs sources de données, et chaque source de données peut fournir un ou plusieurs types de données. Par conséquent, ces différentes données multi-sources forment des ensembles de données multi-sources et hétérogènes. Par exemple, dans le domaine du e-commerce, les données sur les clients peuvent inclure leurs données descriptives, leurs achats effectués, leurs retours sur leurs achats, etc.

Ces données sont disponibles pour être analysée. L'analyse de données (Data Analytics), est une science consistant à examiner des données brutes, dans le but de tirer des conclusions à partir des informations extraites. L'analytise de données est utilisée dans de nombreuses industries afin de permettre aux entreprises et aux organisations de prendre les meilleures décisions. Dans le domaine scientifique, elle est utilisée pour vérifier des théories ou pour réfuter des modèles existants. L'un des postes clés dans sa mise en œuvre est l'analyste de données (data analyst)[3]. Dans ce cadre, comprendre les données des clients est une première étape indispensable pour prendre des décisions, par exemple prédire les achats futurs des clients ou leur recommander des achats en fonction de leurs besoins et préférences. Il existe différentes approches pour analyser et comprendre les données et tirer des conclusions des informations qu'elles contiennent. Le choix d'une approche appropriée dépend de la nature des données ainsi que des objectifs de ces analyses et des informations que l'on souhaite obtenir.

Une des approches existantes est la découverte de connaissances dans les bases de données (KDD) qui a été définie pour la première fois par Fayyad et al., [2] comme le processus non trivial d'identification de modèles valides, nouveaux, potentiellement utiles et

1 www.sciencedaily.com récupéré le 10 décembre 2019
2 https: //www.domo.com récupéré le 10 décembre 2019
3 https://www.lebigdata.fr/definition-quest-data-analytics récupéré le 1er septembre 2020

compréhensibles dans les données. KDD est divisée en trois étapes principales: le prétraitement, la fouille et le post-traitement [1]. La première étape est le prétraitement qui filtre et prépare les données en supprimant les données non pertinentes, non fiables, redondantes ou bruyantes. La troisième étape est le post-traitement qui permet de rendre les résultats compréhensibles, puis évalue ces résultats en fonction des exigences et des objectifs. La deuxième étape est la fouille de données qui vise à extraire des connaissances à partir de grands ensembles de données. Ces grands ensembles de données peuvent être trouvés dans des bases de données de divers types. Des exemples de ces bases de données sont les bases de données relationnelles, les bases de données transactionnelles, les entrepôts de données et autres référentiels de données. Cette étape est importante pour de nombreuses tâches d'analyse, et c'est l'objectif principal de cette thèse de doctorat. La fouille de données a pour objet l'extraction de connaissances à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
Il existe trois techniques principales en fouille de données: la classification [3], le clustering [4] et la fouille de motifs [1,5]. La classification est une famille de méthodes qui permettent de regrouper des individus en classes,. Les classes de la classification ne sont pas connues au préalable, elles sont découvertes par le processus. D'une manière générale, les méthodes de classification servent à rendre homogènes des données qui ne le sont pas a priori, et ainsi permettent de traiter chaque classe avec des algorithmes sensibles aux données aberrantes. Le clustering est défini comme la création de groupes ou de clusters d'objets similaires qui sont différents de ceux d'autres clusters. Les méthodes de clustering sont particulièrement utiles pour explorer les interrelations entre les objets de l'ensemble de données, et elles représentent une façon d'attribuer des labels de classe aux objets de données lorsqu'aucune information de ce type n'est disponible ou connue [6].

La fouille de motifs vise à trouver des modèles motifs pertinents et fréquents dans les données. Elle est devenue un domaine important au fil du temps où diverses travaux dans la littérature ont été proposés dans ce domaine. Le processus de fouille de motifs ainsi que les motifs fréquents générés par ce processus sont interprétables, ce qui aide à prendre des décisions interprétables comme des recommandations. La fouille de motifs peut par exemple être utilisée dans le domaine de la santé où les chercheurs visent à analyser et à comprendre les données sur les symptômes des patients, les diagnostics effectués par les médecins selon ces symptômes et les traitements prescrits par les médecins aux patients en fonction de ces diagnostics. Ici, un motif fréquent contient certains symptômes, leur diagnostic et leur traitement. Ces motifs permettent de recommander des traitements des patients selon leurs symptômes.

La fouille de motifs séquentiels [7] est un champ de la fouille de données ayant pour but la découverte de régularités dans des données se présentant sous forme de séquences d'éléments ou d'ensembles d'éléments. La fouille de motifs séquentiels est un sujet de grand intérêt car les données séquentielles existent dans de nombreux domaines d'application. Un de ces domaines est la e-éducation où les données séquentielles peuvent représenter les séquences de ressources pédagogiques (examens, exercices, etc.) que les étudiants consultent sur leur espace numérique de travail (ENT). Un exemple de séquence de ressources pédagogiques d'un élève est: élève-143: $<R_3\ R_8\ R_{13}\ R_{27}\ R_{29}>$, élève-143 représente l'identifiant de l'élève et R_n représente l'identifiant d'une ressource pédagogique. La séquence signifie que l'étudiant a consulté les ressources R_3, puis R_8, puis R_{13}, puis R_{27} et enfin R_{29}.

Comme mentionné précédemment, un ensemble de données peut contenir plusieurs types de données provenant d'une seule source ou de plusieurs sources de données. Les données peuvent être séquentielles et représenter des séquences de transactions effectuées par les utilisateurs sur leur système; un exemple de ces données est les achats de produits effectués par les clients sur un site Web de commerce. Les données peuvent également représenter des données descriptives; telles que les données démographiques des utilisateurs telles que leur âge, leur sexe et leur adresse. Les données peuvent également représenter des données descriptives sur des éléments. Les données descriptives sur les produits pourraient être le nom du produit, la marque et la date d'expiration. Ces différentes sources de données, réunies, forment un ensemble de données complexe et hétérogène. Malgré sa complexité et son hétérogénéité, cet ensemble de données est riche et contient des informations riches. Cela rend le processus de fouille de cet ensemble de données un défi scientifique.

Problématique

Notre objectif applicatif est d'analyser et de comprendre le comportement numérique des utilisateurs. L'objectif général de cette thèse de doctorat est la fouille de motifs dans des données multi-sources. Les données à fouiller peuvent être complexes, hétérogènes et peuvent contenir différents types de données. Les données peuvent être séquentielles (ou temporelles) ou descriptives et peuvent se présenter sous différentes formes, par exemple sous forme numérique, catégorielle, textuelle ou graphique. Elles peuvent représenter plusieurs points de vue de ce comportement. En comprenant ces données, nous visons à prédire le comportement des utilisateurs et à leur fournir des recommandations de manière personnalisée. Dans ce but, nous nous concentrons sur la fouille de données séquentielles des données multi-sources afin d'extraire des motifs séquentiels fréquents des données comportementales des utilisateurs.

Si nous limitons le processus de fouille de données multi-sources à la fouille d'une seule source, naturellement séquentielle, nous considérons que les résultats peuvent être limitants pour deux raisons principales. La première raison est que les motifs représentent les données fouillées; par conséquent, lorsqu'une seule source de données parmi les multiples sources de données est fouillée, les motifs ne contiendront que des informations provenant de cette source unique; ces motifs représentent donc un point de vue unique. Évidemment, lorsque des sources supplémentaires sont disponibles, les motifs fouillés peuvent contenir des informations provenant de ces sources et seront probablement plus riches. Par exemple, lorsque nous gérons seulement la source de données séquentielle, nous ne comprendrions que le comportement numérique des utilisateurs, et nous fournirions donc des recommandations aux utilisateurs uniquement en fonction de ces informations. Cependant, ces informations peuvent ne pas être suffisantes pour fournir des recommandations fiables et bien personnalisées. L'ajout d'informations descriptives sur les utilisateurs et/ou les éléments des séquences peut aider à une meilleure compréhension du comportement numérique des utilisateurs et donc à des recommandations plus riches et plus spécifiques.

La deuxième raison est les limites des données. Même lorsque la quantité de données est grande, dans certains cas, nous pouvons avoir un problème de manque de données ou de

similarité entre certains éléments des données. Ces similarités peuvent être au niveau des éléments des séquences où certains éléments sont similaires à d'autres, et elles peuvent être au niveau des motifs où certains motifs sont similaires à d'autres. Par conséquent, certains motifs peuvent ne pas être détectés comme fréquents en raison de leur similarité car leur support a diminué. Ce problème conduit à une diminution du nombre de motifs séquentiels générés et donc à une couverture de données inférieure. Dans ce cas, les informations extraites sont limitées. Ainsi, lorsque plusieurs sources de données sont disponibles, il est important de gérer ces sources afin de limiter le problème du manque de données et d'augmenter les informations contenues dans les patterns générés.

De nombreuses techniques ont été proposées pour fouiller des données multi-sources [8, 9, 10]. Ces techniques peuvent être divisées en deux approches principales. La première approche fouille différentes sources de données de manière séparée. Un processus de fouille séparée est effectué sur chaque source de données, et les informations extraites de tous les processus de fouille sont ensuite combinées pour obtenir des informations globales qui contiennent divers types d'informations provenant de différentes sources de données. Cette approche permet d'obtenir des informations intéressantes. Cependant, le principal inconvénient de cette approche est que certaines sources de données peuvent fournir des informations utiles seulement si elles sont fouillées avec d'autres sources de données; la fouille de cette source séparément se traduira donc par des informations inutiles ou insuffisantes. Cela nous permet de comprendre que différentes sources de données sont liées, c'est-à-dire qu'elles ont différents types de relations entre elles lorsque certaines sources de données fournissent des données à d'autres sources. Cette approche conduirait à un problème de perte d'informations importantes en raison de la perte des relations existant entre les différentes sources de données.

La deuxième approche gère les données multi-sources en intégrant toutes les sources de données dans l'ensemble de données et en les fouillant dans un processus de fouille unique. Cette approche maintient les relations entre les sources de données, ce qui évite la perte de données et permet de générer des informations riches. Cependant, cette approche aboutit à une forme complexe de données qui représentent l’entrée du processus de fouille de données, et elle a également une complexité élevée, en particulier lorsqu'il existe un grand nombre de sources de données où chacune fournit différents types de données.

Compte tenu des limites des deux approches, le défi principal de notre travail est de fouiller un ensemble de données complexe formé de multiples sources de données hétérogènes et liées. Ces ensembles de données sont fouillés en gérant les relations existant entre différentes sources de données avec une complexité d'algorithme limitée afin d'extraire des informations riches et intéressantes sous la forme de motifs séquentiels contenant divers types d'informations.

Approche et Contributions

Nous proposons une approche qui gère les données multi-sources en un seul processus de fouille. Dans ce cadre, nous proposons de limiter la complexité du processus de fouille en considérant une source comme étant principale et les sources de données supplémentaires sont fouillées de manière sélective, c’est-à-dire uniquement lorsque cela est nécessaire pendant le processus. Cette approche est conçue pour être générique et ne se limite pas aux données d'un domaine ou d'une structure spécifique.

Il peut exister des relations différentes entre les sources de données. Dans ce contexte, nous introduisons deux types de relations entre les sources. Le premier type de relations existe entre la source de données séquentielles et la source de données descriptives des utilisateurs. La source de données descriptive des utilisateurs fournit des informations spécifiques à chaque utilisateur. Lorsque ces données sont fournies à la source de données séquentielles, elles permettent d'obtenir des motifs séquentiels fréquents plus précis que ceux générés par l'exploration traditionnelle de données séquentielles.

En revanche, le deuxième type de relations existe entre la source de données séquentielle et la source de données descriptives des éléments. La source de données séquentielle contient des séquences d'éléments, et la source de données descriptives des éléments fournit des données supplémentaires sur chaque élément des séquences. Lorsque ces données descriptives d'éléments sont fournies aux données séquentielles, chaque élément est fourni par des attributs descriptifs supplémentaires. Ces attributs représentent des informations plus générales que les identifiants des éléments, et ils donnent donc plus de généralité aux motifs fouillés.

L’approche que nous proposons gère les deux types de relations en se concentrant principalement sur le second type. Il tire l’avantage de la source de données descriptive des éléments afin de générer des motifs séquentiels généraux. Afin de gérer le problème de la similarité des données et de générer des motifs plus fréquents, nous définissons deux mesures de similarité : la similarité de motifs qui compare différents motifs et la similarité d'éléments qui compare différents éléments dans les motifs.

Ensuite, nous formons des motifs généraux à partir de motifs similaires, c’est-à-dire contenant des éléments similaires. Un motif général est un motif qui contient des informations plus générales que les motifs fréquents traditionnels. Enfin, nous proposons une nouvelle méthode de calcul du support de ce nouveau type de motif afin de détecter des motifs généraux fréquents. Après cette étape, nous proposons une nouvelle méthode pour calculer le support d'un motif général afin de détecter des motifs généraux fréquents.

Notre approche permet de résoudre le problème de la similarité des données et de la faible couverture des données en générant des motifs plus fréquents; de plus, les motifs généraux fréquents sont riches car ils contiennent divers types d'informations.

Notre approche dans le domaine de l’éducation

L'utilisation de la technologie pour mieux comprendre comment se déroule l'apprentissage a reçu une attention considérable ces dernières années [11]. Les données pédagogiques numériques, qu’il s’agisse des processus d’apprentissage ou d’enseignement, sont collectées des ENT et des systèmes d’information des établissements d’enseignement, et elles sont désormais disponibles pour pour une analyse pour différents objectifs.

L’analyse de l’apprentissage (Learning analytics) est la discipline qui vise à recueillir toutes les données liées à un système d’apprentissage numérique. Toutes ces informations sont collectées, mesurées, exploitées et analysées pour améliorer et optimiser le processus d’apprentissage en ligne. Pour les apprenants, il vise à améliorer le processus d'apprentissage en leur donnant la possibilité d'évaluer leurs progrès scolaires et d'améliorer leurs réalisations et leur motivation.

Ce travail est financé par le projet PIA2 e-FRAN METAL. Le projet METAL se propose de concevoir, développer et évaluer un ensemble d'outils de suivi individualisé destinés aux élèves ou aux enseignants (Learning Analytics), et des technologies innovantes pour un apprentissage personnalisé des langues à l'écrit (grammaire française) et à l'oral (prononciation de langues vivantes). Il participe ainsi à l'amélioration de la qualité de l'apprentissage et au développement de la maîtrise des langues par les élèves. Notre travail est spécifiquement dédié aux élèves des écoles, et l'objectif est de contribuer à l'amélioration de leur processus d'apprentissage.

Dans le domaine de l'éducation, les données numériques sont collectées de plusieurs sources de données qui peuvent fournir différents types de données. Cela forme un ensemble de données multi-sources et hétérogènes. Par conséquent, notre approche proposée peut être appliquée sur cet ensemble de données dans le but de fournir aux étudiants des recommandations personnalisées de ressources pédagogiques afin de les aider à améliorer leurs performances académiques grâce à l'analyse de l'apprentissage.

Notre modèle proposé a été conçu pour des objectifs éducatifs tout en étant générique. En raison d'un problème de disponibilité des données dans le projet, le modèle n'a pas encore été évalué sur des données éducatives, mais il est évalué sur des données d'un autre domaine.

Références

[1] Jiawei Han, Jian Pei, and Micheline Kamber. Data mining : concepts and techniques. Elsevier, 2011.
[2] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. Knowledge discovery
and data mining : Towards a unifying framework. In KDD, volume 96, pages 82–88, 1996.
[3] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7) :1895–1923, 1998.
[4] Anil K Jain and Richard C Dubes. Algorithms for clustering data. Englewood Cliffs : Prentice Hall, 1988, 1988.
[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In Acm sigmod record, volume 22, pages 207–216. ACM, 1993.
[6] Fedja Hadzic, Henry Tan, and Tharam S Dillon. Mining of data with complex structures, volume 333. Springer, 2010.
[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Mining sequential patterns. In icde, volume 95, pages 3–14, 1995.
[8] Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal. Multi-dimensional sequential pattern mining. In Proceedings of the tenth international conference on Information and knowledge management, pages 81–88. ACM, 2001.
[9] Elias Egho, Chedy Raïssi, Dino Ienco, Nicolas Jay, Amedeo Napoli, Pascal Poncelet, Catherine Quantin, and Maguelonne Teisseire. Healthcare trajectory mining by combining multidimensional component and itemsets. In International Workshop on New Frontiers in Mining Complex Patterns, pages 109–123. Springer, 2012.
[10] Wen-Chih Peng and Zhung-Xun Liao. Mining sequential patterns across multiple sequence databases. Data & Knowledge Engineering, 68(10) :1014–1033, 2009.
[11] Wolfgang Greller and Hendrik Drachsler. Translating learning into numbers : A generic framework for learning analytics. 2012.