Intelligent Data Mining Assistance Via Case-Based Reasoning and a Formal Ontology

Michel Charest
30000316

RÉSUMÉ

Introduction

La mise en œuvre efficace d'une activité de forage de données nécessite des connaissances pointues et des décisions appropriées sur un bon nombre de techniques spécialisées (p.ex. : le nettoyage de données, la transformation des attributs, le choix d'algorithme et de paramètres, les méthodes d'évaluations, etc.). De nos jours, il existe un grand choix de méthodes, de modèles et d'outils pour effectuer le forage de données, mais peu de soutien « intelligent » pour les non experts. Ainsi, suite à des recherches, nous avons réalisé un assistant pour le forage de données, basé sur le raisonnement à base de cas et une ontologie formelle, capable d'assister les non-spécialistes lors de leur démarche d'activités de forage de données.

Afin de demeurer efficace les preneurs de décisions ont fréquemment recourt aux techniques de forage de données pour lutter contre l'accroissement incessant d'informations suite aux opérations quotidiennes de leur entreprise. Malgré le fait que le data mining semble très prometteur pour assister à la découverte de « connaissance », l'application efficace du processus de forage de données comprend à la fois de grands défis et difficultés. Par exemple, les recherches actuelles menées sur le forage de données sont basées sur l'application de techniques très spécialisées (p.ex. : les statistiques, l'apprentissage automatique et les bases de la théorie de l'information), or la recherche portant sur des thèmes méthodologiques, stratégiques ou épistémologiques se font plutôt rare. D'autre part, sur le plan pratique très peu d'entreprises utilisent des méthodes de gestion des connaissances sur l'application pratique du forage de données (p. ex: une mémoire institutionnelle). Par ce fait, les anecdotes de réalisations fructueuses utilisant le forage de données se font rares. De plus, malgré le fait que les méthodologies fréquemment employées pour le forage (telle que la méthodologie CRISP-DM) offrent des consignes générales pour guider les utilisateurs dans leurs démarches, les non-spécialistes ont plutôt besoin de suggestions et d'explications dans des contextes précis lors de la démarche du processus. Autrement dit, il n'est pas suffisant de dire « quoi » un utilisateur doit faire, mais il est

plutôt important de dire « comment » et à quels instants on doit appliquer une telle technique, méthode ou vérification lors d'une activité de forage. Enfin, la majorité des assistants réalisés au fil des années ont focalisés uniquement à supporter le bon choix de modèle pour une activité de forage de données. Malgré que cette étape soit importante, elle ne peut assurer le succès d'une activité de forage de données. Ainsi, un assistant intelligent devrait offrir un support tout au long de l'application du processus de forage (p.ex. : assister l'analyse et la préparation des données, l'évaluation de modèles).

Objectifs du travail de recherche

Suite à une étude approfondie des problématiques de ce domaine, nous nous sommes fixés les objectifs de recherche suivants :

1) **Supporter les non-spécialistes** – Assister les analystes non experts du forage de données en considérant particulièrement leur niveau de connaissances du forage de données.

2) **La réutilisation de connaissances** – Encourager la réutilisation d'expériences antérieures de data mining sous la forme d'une base de connaissance ou de mémoire institutionnelle.

3) **Un soutien holistique** – Apporter un soutien au-delà de l'assistance au choix de modèle, mais plutôt un support qui comprend les étapes majeures telles que la préparation de données, la modélisation et l'évaluation des modèles.

4) **Utiliser des connaissances approfondies** – offrir des connaissances sous la forme de suggestions, heuristiques et réponses automatiques pour assister l'usager à prendre des décisions lors de la démarche du processus de forage de données.

En résumé, nous avons tenté de rendre le forage de données plus accessible et facile pour les non-spécialistes de ce domaine en proposant un cadre théorique, conceptuel et technologique pour la réalisation d'un assistant intelligent pour le forage de données.

Plus particulièrement, nous avons vérifié si la combinaison « synergique » d'un système de raisonnement à base de cas (RBC) et d'une ontologie formelle peut supporter de façon convenable les non experts pratiquant le forage de données.

La méthodologie utilisée

Dans un premier temps, nous avons effectué une analyse approfondie de l'état de l'art sur les assistants de forage de données, ainsi que les systèmes d'aide à la décision pertinents à celui-ci. Ceci nous a permis de bien cibler et de définir les problématiques. En fait, ceci nous a permis d'élaborer nos objectifs de recherches tels qu'ils le sont énoncés ci-dessus. Deuxièmement, nous avons effectué une analyse approfondie de l'état de l'art de plusieurs domaines sous-jacents, le forage de données et les systèmes de prise de décisions. Par exemple, nous avons enquêté sur les avancements réalisés au niveau des processus de forage de données, le méta-apprentissage et la caractérisation de données. Ensuite, nous avons examiné les divers modes de représentation de connaissances (et méthodes de raisonnement respectives) tels que le raisonnement à base de cas et les ontologies formelles basées sur la logique de description.

Particulièrement, nous avons évalué un bon ensemble de cadres et de systèmes de raisonnement à base de cas dans les milieux académiques et professionnels. Ceci nous a permis de conclure qu'il était préférable de concevoir et de réaliser notre propre système de raisonnement à base de cas. La première étape importante pour la réalisation de notre système RBC a consisté à définir une représentation d'un cas de forage de données, c'est-à-dire un ensemble de caractéristiques représentatives d'une activité de data mining. Pour ce faire, nous avons examiné attentivement le processus de forage de données CRISP-DM. Malgré que celui-ci est représenté en utilisant le langage naturel (p.ex. : anglais), nous avons pu définir une représentation d'un cas de forage comportant un ensemble de 66 caractéristiques des 5 phases principales du processus CRISP-DM (p.ex. : les besoins d'affaires, la compréhension des données, la préparation de données, la modélisation, et l'évaluation du processus). Ayant conçu une représentation abstraite d'un cas de data mining, nous avons ensuite réalisé une composante pour faire la comparaison de cas de forage de données, c'est-à-dire la

réalisation d'une mesure d'appariement globale (et les mesures de similarités locales sous-jacentes) nous permettant d'effectuer une comparaison quantitative entre deux cas de forage de données. Ainsi, un usager ayant stocké des activités de forage antérieurement dans notre base de cas est en mesure de repérer des cas « similaires » à son problème de forage de données actuel.

À cette étape de nos initiatives, ayant conçu la base de notre assistant de forage de donnée en utilisant un RBC, nous avons effectué de premiers essais. Ces essais nous ont permis de constater deux lacunes importantes à notre système :

a) Quand un usager utilise notre système pour repérer un ensemble de cas antérieurs et similaires à son problème actuel, il n'est pas évident pour l'usager de déduire quel « cas de base » est le meilleur choix pour débuter le processus d'adaptation et éventuellement résoudre son cas actuel de forage de données.
b) Ayant choisi un cas relativement similaire pour résoudre le problème actuel, il n'est pas toujours évident pour l'usager de savoir quelles informations dans ce cas de base sont utiles et pertinentes au problème actuel qu'il doit résoudre (adaptation d'un cas de base au cas de forage actuel).

Ainsi, pour résoudre le problème de sélection de cas de base reporté par le système de raisonnement à base de cas, nous avons proposé la réalisation d'une mesure supplémentaire basée sur la théorie de l'utilité. Cette nouvelle mesure sert à donner à l'usager un indice du niveau de « qualité » ou capacité de résolution d'un cas similaire. Ainsi, l'usager peut maintenant faire un choix final du cas de base à utiliser basé sur un compromis (ou équilibre) entre le niveau de similarité d'un cas et le niveau potentiel d'utilité ou d'adaptabilité de ce cas par rapport au cas de forage de données du problème à résoudre.

D'autre part, pour résoudre le second problème d'adaptation d'un cas de base, c'est-à-dire offrir des suggestions précises pour aider l'usager à modifier les caractéristiques pertinentes d'un cas similaire, nous avons constaté le besoin d'une base de connaissances supplémentaire pour offrir cette aide. En fait, au début, puisque les ontologies basées sur la logique de description offrent naturellement la représentation

de connaissance déclarative, nous avons tenté d'utiliser celles-ci (et les techniques de raisonnement de la logique de description) pour résoudre ce problème. Mais, suite à des essais, nous avons rapidement constaté qu'il était nécessaire d'ajouter une base de connaissance supplémentaire contenant des connaissances procédurales (des connaissances à base de règles). Ainsi, par la suite, nous avons effectué des enquêtes sur des cadres ontologiques offrant la possibilité de représenter à la fois des concepts déclaratif et procéduraux. Enfin, malgré que cette avenue est actuellement un sujet de recherche à ses balbutiements, nous avons réussi à intégrer des connaissances approfondies sur le data mining sous la forme de règles et concepts dans notre ontologie formelle (suite à une activité d'ingénierie et formalisation de certaines connaissances de forage de données).

Finalement, la résolution de ces problèmes fondamentaux nous a permis de mettre en œuvre un assistant intelligent pour le forage de données. Ensuite, nous avons procédé à l'évaluation de notre système tel qu'indiqué dans la section suivante.

Les résultats obtenus

Ayant à la fois défini un ensemble de cas de forage de données « noyau » afin de rendre fonctionnel notre système RBC et codé un premier ensemble de connaissances approfondies (sous la forme de concepts et règles) dans notre ontologie formelle, nous avons procédé à la réalisation de quelques activités de forages de données. Enfin, nous avons fait une évaluation comparative des suggestions fournies par notre système intelligent à ceux d'un expert humain (voir Section 5 pour plus de détails).

Conclusions

Nous avons réalisé un système intelligent pour le forage de données basées sur la synergie d'un système de raisonnement à base de cas et d'une ontologie formelle. La première composante (RBC) permet à l'usager de faire évoluer une sorte de mémoire institutionnelle de cas de forage de données (au fur et à la mesure qu'il résolut des nouveaux cas) qui permet de facilement repérer des cas similaires antérieurement résolus pour avoir un premier aperçu sur la résolution du problème actuel.

D'autre part, la composante ontologique de notre système contenant des règles et suggestions textuelles, permet d'offrir des connaissances pointues et précises à un usager lorsqu'il effectue une activité de forage (suite au raisonnement qu'offre un moteur de raisonnement à base de règles). C'est-à-dire, le système intelligent offre des suggestions précises à l'usager pendant que celui-ci procède à la résolution de son problème de forage en modifiant des caractéristiques du cas de base.

De plus, ces deux modes de représentations de connaissance complémentaires permettent aux non-spécialistes de profiter d'une assistance holistique sur l'activité de forage de données. Enfin, il est très important de mentionner que nos objectifs de recherches ont été particulièrement abordés dans la perspective d'apporter un soutien aux non-spécialistes du forage de données, c'est-à-dire les preneurs de décision au sens général du terme.