CORON

Ou comment faire parler les données…

Santé, biologie, commerce, industrie, enseignement, de nombreux domaines sont amenés à recueillir des données : résultats d'examens de santé, découverte de gènes, consommation de produits… Comment interroger ces résultats et comment faire en sorte que ces données dévoilent une logique de fonctionnement des objets étudiés ? Comment prédire que la présence de tel gène est susceptible d'engendrer tel type de maladie ? Qu'en transport ferroviaire, l'utilisation de tel mécanisme est capable d'engendrer tel type d'incident ? Qu'en matière de consommation tel comportement engendre tel type d'habitude ?
C'est pour donner aux experts les moyens de traiter et de faire parler leurs données que les chercheurs de l'équipe ORPAILLEUR du LORIA ont développé la plate-forme CORON.

- **CORON, mode d'emploi**

CORON est une plate-forme de fouille de données, composée de quatre modules. Elle est utilisée sur des bases de données binaires, c'est à dire des bases qui associent deux types de données : des objets et une série d'attributs (*L'objet 1 a pour attributs A et B, l'objet 2 a pour attributs B et C, etc…* ou plus concrètement : *le sujet 1 achète du beurre et du pain, le sujet 2 achète du beurre, du pain et de l'eau*). La première étape consiste à préparer ces données à l'aide du premier module éponyme. Dans cette phase, **CORON** exécute une opération de filtrage. En deuxième phase, on extrait des motifs fréquents, où un motif est un ensemble fini d'attributs.
Dans une troisième étape, et sur la base de ces motifs, le module **AssRuleX** (Association Rule eXtractor), recherche puis extrait des règles d'association. Ces règles sont en quelque sorte les relations cachées qui existent dans les données traitées. Ainsi explicitées, elles permettent à l'expert d'extraire des éléments de connaissance (ex. *lorsqu'un sujet achète du pain il achète également du beurre*).

Généralement, le nombre de règles obtenues est important. Il est nécessaire de faire du tri pour obtenir les plus intéressantes et les plus pertinentes. C'est le rôle du dernier module RuleMiner. Constitué de trois outils, il permet tour à tour de :
- filtrer les attributs : autrement dit éliminer ou conserver les règles qui concernent certains attributs,
- effectuer un tri par support (une règle est vraie pour un nombre donné d'objets – *4 sujets sur 10 achètent du pain*) et/ou par confiance (telle règle a tel pourcentage de chances de se produire – *75% de sujets qui achètent du pain achètent aussi du beurre*),
- colorer des règles : pour reconnaître et identifier rapidement les plus intéressantes d'entre elles.

Ce processus est complètement itératif et interactif : les résultats peuvent ainsi être affinés en répétant plusieurs fois ces différentes étapes sous le contrôle d'un analyste .

- **Les spécificités de CORON**

CORON est développé en langage Java. Il est par conséquent extrêmement portable et peut donc être exécuté sur différentes plate-formes. L'un des points fort de ce système est de faire appel à plusieurs algorithmes qui, choisis par l'utilisateur en fonction de la densité de la base de données à traiter, permettent d'obtenir plusieurs points de vue synthétiques sur les données.

Enfin, les chercheurs de l'équipe ORPAILLEUR travaillent dans une optique de mutualisation et de coopération : pour exemple, certaines fonctionnalités de CORON sont utilisées en partie par KASIMIR, un logiciel de prédiction et d'aide à la décision en cancérologie (Kasimir est aussi un système mis au point dans

l'équipe ORPAILLEUR), ou encore par le projet Galicia de l'Université de Montréal,...

CORON, concrètement...

Les domaines d'application sont très nombreux et les chercheurs de l'équipe ORPAILLEUR ont déjà testé **CORON** à travers plusieurs collaborations.

Dans le cadre d'une collaboration avec des chercheurs de l'INSERM (Institut National de la Santé Et de la Recherche Médicale), **CORON** a été directement utilisé pour extraire des règles et des profils génétiques dans le cadre du syndrome métabolique. Ce syndrome (qui se traduit notamment par de l'obésité, des prédispositions aux maladies cardiovasculaires, du diabète,...), atteint presque un quart de la population aux Etats-Unis, et est devenu un enjeu majeur de santé publique en France. A partir d'un échantillon de sujets supposés sains, dont on a traité les données comme le poids, la taille, l'âge, la pression artérielle, les habitudes alimentaires,..., les résultats obtenus ont permis d'énoncer de nouvelles hypothèses de travail pour les biologistes.

Dans un tout autre domaine, celui du transport ferroviaire, **CORON** a traité des données liées à des modèles de locomotives et des erreurs techniques constatées. Le but était de comprendre quelles connexions ou quels effets en chaîne il peut exister entre différentes erreurs. Dans ce cas, les résultats obtenus ont un but préventif.

Enfin, **CORON** peut être utilisé pour l'étude des comportements : en traitant des données sur la disposition des produits en supermarché et leur situation géographique dans le magasin par exemple, ou encore en fouille de textes comme c'est le cas dans certaines expérimentations actuelles.

CORON illustre une fois de plus l'intérêt des coopérations interdisciplinaires et laisse présager de nombreuses nouvelles collaborations.

- **Contact** - Laszlo SZATHMARY, Amedeo NAPOLI
  téléphone - (+33)3.83.59.20.45 ou (+33)3.83.59.20.68

- **Lien :** http://coron.loria.fr