J-F. Boulicaut and B. Crémilleux (coordinateurs). Extraction de motifs dans des bases de données. RSTI ISI 9(3/4), Hermes-Lavoisier, 2004

Editorial

Il est chaque jour plus facile de collecter des données mais notre capacité à en extraire des informations à forte valeur ajoutée reste limitée. Pour répondre à ces opportunités, l'extraction de connaissances dans les bases de données (ECBD ou « Knowledge Discovery in Databases ») est le domaine de recherche au sein duquel coopérent statisticiens, spécialistes en bases de données et en intelligence artificielle, ou encore chercheurs en conception d'interfaces homme-machine. Ce domaine connaît une croissance spectaculaire, sous l'impulsion des organisations propriétaires de données.

L'extraction de connaissances dans les bases de données a été définie dès 1991 comme le processus non trivial d'extraction d'informations valides, nouvelles, potentiellement utiles, et compréhensibles à partir de données et l'on trouve d'autres définitions dans l'un des ouvrages fondateurs du domaine [1]. Les processus ECBD sont des processus itératifs et interactifs complexes que l'on peut résumer en 5 macro-étapes : la sélection et consolidation des sources de données brutes (ce qui relève souvent de la construction d'un entrepôt de données), la préparation de ces données pour produire des contextes d'extraction, les extractions proprement dites de motifs ou de modèles, le post-traitement et l'interprétation des motifs ou modèles calculés et enfin l'exploitation des résultats. Parmi les problématiques phares en ECBD, le challenge algorithmique de l'extraction de motifs locaux dans de grands volumes de données a mobilisé de nombreux chercheurs. Même si les motifs locaux (par exemple des régularités dans des données séquentielles ou transactionnelles) sont rarement utilisables en tant que produit final d'un processus ECBD, leurs découvertes jouent un rôle majeur tout d'abord dans la construction de résumés (ou plus généralement dans les méthodes visant à décrire de grands volumes de données comme celle des règles d'association [2]) mais aussi dans le calcul de descripteurs avec des applications en construction de modèles descriptifs (groupes ou partitions) ou prédictifs (classifieurs).

L'objectif de ce numéro thématique est de faire le point, grâce à des articles de synthèses mais aussi des contributions originales, sur l'extraction et les usages de motifs en ECBD. Les articles sélectionnés par le comité de lecture portent sur les trois principaux types de motifs locaux étudiés depuis une dizaine d'années, c'est-à-dire les dépendances ou requêtes fréquentes dans des bases de données relationnelles, les motifs ensemblistes dans des données transactionnelles et enfin, les motifs séquentiels dans des données ordonnées (ordre spatial ou temporel).

Pour mettre en perspective les différentes contributions retenues, il nous semble intéressant de resituer le problème et les solutions étudiées dans un cadre formel

adapté. C'est cet objectif qui explique la longueur inhabituelle de cet éditorial. Nous voulons aussi présenter quelques uns des problèmes ouverts qui mobilisent actuellement les chercheurs en ECBD et compléter les contributions par davantage de références aux usages de motifs.

Heikki Mannila et Hannu Toivonen ont étudié une abstraction simple de nombreux travaux en ECBD via le concept de théorie [3]. Il ne s'agissait pas seulement de spécifier des tâches d'extraction de motifs, i.e., la définition de collections à calculer, mais aussi d'étudier des algorithmes génériques efficaces pour les calculer. Considérons l'instance r d'une base de données, un langage L pour l'expression de propriétés dans les données, i.e., un langage de motifs, et un prédicat de sélection q utilisé pour dire si oui ou non, une phrase ϕ de L doit être considérée comme intéressante sur r. Une tâche d'extraction peut alors être formalisée comme le calcul de la théorie $Th(L,q,r) = \{\phi \in L \mid q(\phi,r)$ est vrai$\}$. Il faut noter qu'ici, nous exigeons que tous les motifs de L qui satisfont q soient produits : cette hypothèse de complétude des extractions est habituelle dans ce domaine de l'extraction de motifs locaux. Dans un cas général, le prédicat q est défini comme une combinaison booléenne de contraintes primitives : on parle d'ailleurs de fouille de données sous contraintes. Dans la pratique, de nombreux travaux se concentrent sur une seule contrainte primitive (comme la fréquence minimale) ou sur des conjonctions de contraintes primitives (par exemple la fréquence minimale et la présence ou l'absence de certains sous motifs dans les motifs à extraire).

Le calcul de Th(L,q,r) ne peut pratiquement jamais être réalisé par une génération systématique de toutes les phrases de L suivie d'une vérification du prédicat de sélection. La taille de L est en effet souvent exponentielle dans le nombre de variables du problème. Pour fixer les idées, considérons le cas très étudié des données transactionnelles et des ensembles dits fréquents. Soit r un ensemble de transactions composées d'items. Ces items peuvent, par exemple, correspondre à chacun des produits qui peuvent être achetés lors d'une transaction. Si l'on considère le langage de motifs des ensembles d'items et le cas réaliste de 1000 items, nous sommes devant un langage de motifs comportant environ 10^{330} éléments et il est clairement exclu de les considérer tous ! Les ensembles dits σ-fréquents sont ceux qui contiennent des items présents dans au moins σ % des transactions, σ étant un seuil fixé par l'utilisateur. Calculer les ensembles fréquents c'est exploiter un prédicat de sélection q qui impose une fréquence minimale dans r pour n'explorer qu'une infime partie de l'espace de recherche L.

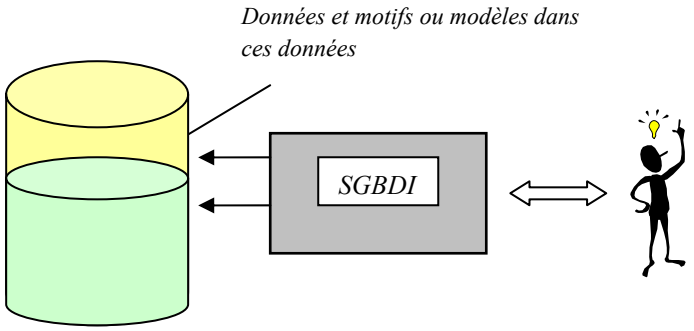
Dans de nombreux contextes, cet espace de recherche est naturellement structuré par des relations de spécialisations anti-monotones par rapport au prédicat de sélection. Ce sera, par exemple, le cas de l'inclusion ensembliste vis-à-vis d'une contrainte de fréquence minimale sur des ensembles. L'anti-monotonie signifie ici que lorsqu'une phrase ne satisfait pas q, aucune de ses spécialisations ne peut le satisfaire (et donc dans notre calcul d'ensembles fréquents, le fait que si un ensemble n'est pas fréquent, aucun de ses sur ensembles ne peut l'être). Une idée simple, bien étudié depuis la définition du cadre « learning as search » de Tom

Mitchell [4] consiste alors à optimiser le parcours de l'espace de recherche structuré en treillis en éliminant des sous-espaces dont on sait qu'ils ne peuvent contenir des phrases intéressantes. On peut ainsi travailler avec des algorithmes par niveaux efficaces pour calculer Th(L,q,r). L'algorithme générique décrit dans [3], d'ailleurs rappelé dans plusieurs des contributions à ce numéro et dont l'instance la plus célèbre est l'algorithme apriori [5], propose de calculer d'abord les phrases les plus générales par rapport à q (par exemple les singletons dans le cas du calcul des ensembles fréquents), puis d'alterner l'évaluation de candidats (par exemple, le comptage des fréquences des candidats dans les données) et la génération de nouveaux candidats (par exemple certains sur ensembles des ensembles fréquents déjà trouvés). L'algorithme s'arrête lorsqu'il n'est plus possible d'avoir de nouveaux candidats (par exemple lorsque que l'on a trouvé tous les plus grands ensembles fréquents). Dans ce contexte ECBD, l'efficacité se mesure notamment en nombre d'évaluations du prédicat de sélection car c'est ce qui demande des lectures des données et, typiquement, on peut devoir compter la fréquence d'un motif dans plusieurs centaines de milliers de transactions. Pour citer Heikki Mannila au cours de sa conférence invitée à Blois (BDA 2000) : « data mining is the art of counting » et l'ont peut considérer qu'effectivement de nombreux travaux algorithmiques visent à compter toujours moins pour « disposer des mêmes nombres ».

L'algorithme générique donné dans [3] s'applique à de nombreux problèmes d'extraction de motifs locaux et c'est le cas des divers travaux présentés dans ce numéro (dépendances et requêtes fréquentes dans des données relationnelles, motifs ensemblistes ou séquentiels). Les prédicats de sélection étudiés ont généralement des propriétés d'anti-monotonicité par rapport à des relations de spécialisation qui structurent les espaces de recherche (cas des contraintes de fréquence minimale pour tous les types de motifs mais aussi cas de l'existence de dépendances éventuellement approximées dans les bases de données). Lorsque l'anti-monotonicité n'est pas garantie, soit l'on met en place des stratégies d'évaluation ad-hoc, i.e., dédiées à des conjonctions particulières (cas typique des algorithmes d'extraction de motifs séquentiels sous des contraintes non anti-monotones), soit l'on s'intéresse à des algorithmes génériques d'extraction sous contraintes.

Il est utile de replacer l'extraction sous contraintes dans le cadre des bases de données inductives (BDI, voir notamment [6,7,8,9,10]). Ce cadre formel a été proposé pour abstraire un grand nombre de pratiques en ECBD et faciliter les nouveaux développements du domaine. L'idée est qu'il est possible de voir les processus ECBD comme des processus d'interrogation où chacune des étapes repose sur des requêtes. On va alors trouver des requêtes classiques (c'est-à-dire au sens des requêtes usuelles sur des bases de données, exprimées, par exemple, au moyen du langage SQL) mais aussi des requêtes dites inductives qui spécifient de façon déclarative les propriétés des motifs ou modèles recherchés. On pourrait donc dire que calculer une théorie pour le prédicat q c'est évaluer la requête inductive q.

La figure ci-dessous résume l'approche du cadre BDI où l'utilisateur final est en interaction avec un Système de Gestion de Bases de Données Inductif.

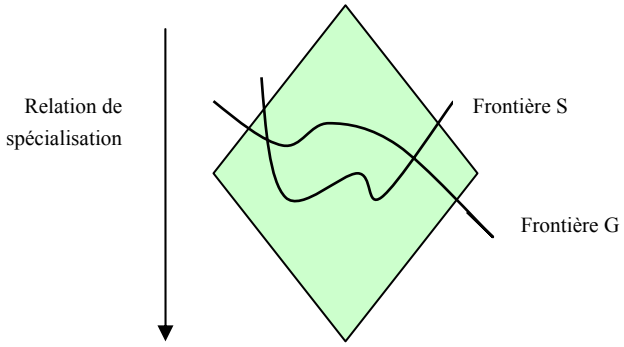


Il est entendu que si les données peuvent exister de façon extensionnelle (par exemple matérialisées dans des tables) ou intentionnelles (par exemple définies ou moyen de vues), il en est de même pour les motifs ou modèles qui tiennent dans les données.

L'évaluation des requêtes inductives demande donc l'extraction de motifs sous contraintes. Dans un processus ECBD, il faut généralement calculer plusieurs extractions et celles-ci sont souvent corrélées d'une certaine manière. L'une des problématiques majeures pour le cadre BDI concerne l'étude des relations entre les extractions et d'ailleurs plus généralement les requêtes utilisées : dans quel cas, le résultat d'une extraction antérieure peut-il être réutilisé pour une nouvelle extraction ? on retrouve ici le nouveau défi d'étudier les propriétés d'inclusion et d'équivalences entre requêtes inductives mais aussi entre requêtes classiques et requêtes inductives (voir par exemple [11] pour une application à MINE RULE, un langage dédié à la découverte de règles d'association [12]). Le Graal de la communauté BDI émergente est de rechercher une sorte d'équivalent de l'algèbre relationnelle mais avec l'ECBD en tête plutôt que l'interrogation classique de bases de données relationnelles. C'est un objectif à long terme qui a été étudiée notamment au sein du contrat européen cInQ IST-FET-2000-26469. Ce consortium cInQ a produit de nombreuses publications concernant principalement l'extraction de motifs locaux au moyen de requêtes inductives : [13] contient des versions étendues des contributions présentées aux deux premiers ateliers qu'il a organisé, i.e., DTDM 2002 associé à EDBT 2002 et KDID 2002 associé à ECML PKDD 2002. Divers tutoriels ont été consacrés au cadre BDI à l'occasion des conférences ECML PKDD 2002, IDA 2003 et EGC 2003. Le site http://www.informatik.uni-freiburg.de/ml/IDB/ présente la manifestation de clôture organisée en Mars 2004 avec des interventions d'une quarantaine de chercheurs du domaine. [14] dans [13] et [15] sont des présentations synthétiques du travail réalisé par ce consortium. Parmi les résultats les plus significatifs, on trouve l'étude d'algorithmes génériques pour traiter des cas généraux de requêtes inductives sur des motifs locaux et le travail sur les représentations condensées.

Les premiers travaux sur l'extraction de motifs sous contraintes se sont intéressés à des conjonctions de contraintes impliquant une contrainte de fréquence minimale et d'autres contraintes souvent non anti-monotones. Dès les premières études, il est apparu que l'élagage par l'exploitation de la fréquence minimale était un outil très puissant et que l'exploitation prématurée des contraintes non anti-monotones pouvait nous priver d'une partie de cette puissance [16,17,18]. Dès lors, de nombreux auteurs ont étudiés des stratégies plus ou moins ad-hoc pour différentes sélectivités des contraintes non anti-monotones. C'est le cas exemplaire des travaux sur la famille d'algorithmes SPIRIT [17] pour extraire des séquences fréquentes satisfaisant une expression régulière : 4 algorithmes sont proposés qui « poussent » plus ou moins la contrainte d'expression régulière. Des travaux pionniers comme ceux sur CAP étudient d'autres propriétés des contraintes comme la « succinctness » pour permettre un usage efficace de certaines contraintes syntaxiques non anti-monotones [19].

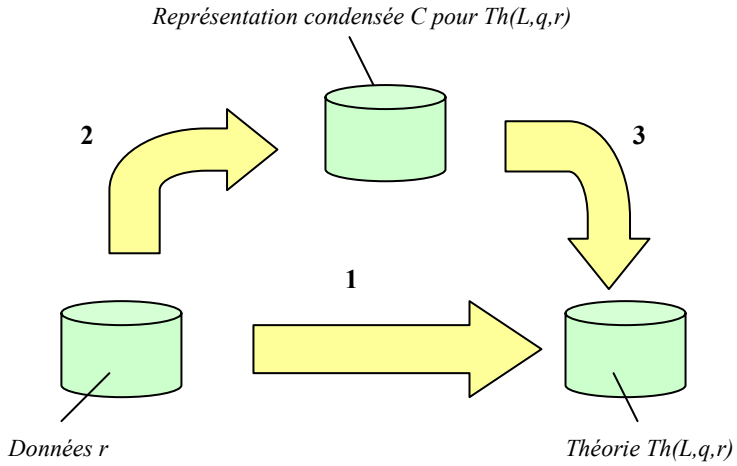
Au sein du consortium cInQ des algorithmes génériques ont été étudiés pour le cas important dans la pratique des combinaisons de contraintes anti-monotones et monotones. Une contrainte est monotone quand sa négation est anti-monotone et une contrainte est donc monotone si lorsqu'une phrase la satisfait, toutes ses spécialisations la satisfont. Dans le cas des motifs ensemblistes, une contrainte forçant une fréquence maximale ou bien l'appartenance d'un élément à l'ensemble est monotone vis-à-vis de l'inclusion ensembliste. Cette dualité entre contraintes monotones et anti-monotones se retrouve illustrée dans la figure ci-dessous.



Dans le cas des conjonctions, les contraintes anti-monotones et monotones définissent deux frontières entre lesquelles se trouvent les solutions. La frontière S correspond aux phrases maximalement spécifiques vis-à-vis de la part anti-monotone. Nous utilisons la notation S car elle correspond à l'appellation usuelle dans les espaces des versions de Mitchell [4]. Dans [3], cette frontière est appelée frontière positive par rapport à la contrainte anti-monotone utilisée. La frontière G correspond à la part monotone, c'est-à-dire aux phrases maximalement spécifiques vis-à-vis de la négation de la contrainte monotone ou si l'on préfère les phrases

minimalement générales au regard de cette part monotone. Le couple (S,G) constitue un espace des versions. On peut alors s'intéresser aux différentes façons de calculer ces frontières (par exemple, des algorithmes de parcours en largeur ou en profondeur d'abord sur l'espace de recherche) et de nombreux algorithmes génériques ont été étudiés pour traiter des conjonctions de contraintes anti-monotones et monotones [18,20,21,22,23]. Dans le cas de combinaisons booléennes (par exemple des disjonctions) de contraintes monotones et anti-monotones, il n'y a plus un mais généralement plusieurs espaces des versions pour délimiter la solution, une question qui a été étudiée notamment dans [24].

Les frontières apparaissent comme la forme la plus évidente de représentation condensée d'une théorie et donc de la solution à une requête inductive. Intuitivement, une représentation condensée pour une théorie Th(L,q,r), c'est une collection C ⊂ Th(L,q,r) telle que tous les éléments de Th(L,q,r) puissent être retrouvés efficacement à partir de C. La figure ci-dessous rend compte de ce processus où il est possible de calculer Th(L,q,r) soit directement (flèche 1) soit par calcul d'une représentation condensée (flèche 2) puis une phase de régénération (flèche 3).



En fait, ce concept est important car nous connaissons plusieurs exemples de représentations condensées pour lesquelles réaliser les phases 2 puis 3 est moins coûteux que la réalisation directe de la phase 1.

Reprenons le cas bien connu des ensembles fréquents (la théorie est la collection de tous les ensembles fréquents pour un seuil de fréquence donné) et supposons que l'on n'ait calculé ou matérialisé que la frontière S de cette théorie c'est-à-dire que les ensembles fréquents maximaux. Il est clair que cette collection est une représentation condensée de l'ensemble des ensembles fréquents et elle peut d'ailleurs être extraite par des algorithmes dédiés (par exemple [25]). Il est maintenant possible de régénérer la collection tous les ensembles fréquents en produisant tous les sous-ensembles des ensembles maximaux sans qu'il soit

nécessaire d'accéder aux données. Dans le même ordre d'idée, on peut rechercher des ensembles qui sont fréquent dans une partie des données (par exemple des transactions dites normales) et qui sont infréquents (fréquence maximale) dans une autre partie (par exemple la partie des transactions dites frauduleuses). On peut alors représenter cette collection par un espace des versions et utiliser les motifs ainsi découverts comme descripteurs dans des phases de classification supervisée. C'est ce qui a été appliqué avec succès dans l'analyse de fragments moléculaire [21,26] et l'on retrouve ce courant de recherche dans les applications des motifs dits émergents [27,28].

Ceci étant, ce qui a motivé les études majeures sur les représentations condensées est lié au contexte des extractions pour lesquelles l'utilisateur ne peut pas se contenter d'une collection de motifs mais a besoin des résultats de fonctions d'évaluation sur les motifs a priori intéressants. C'est bien sûr le contexte le plus courant lorsque l'on veut travailler avec des ensembles fréquents : non seulement nous devons connaître les motifs fréquents mais aussi leurs fréquences. Elles sont clairement indispensables aux post-traitements comme celui du calcul des règles d'association de confiance suffisante [2]. Notons aussi que le calcul de la plupart des autres mesures d'intérêt pour les règles d'association fréquentes nécessitent seulement la connaissance de la fréquence des ensembles fréquents.

Nous sommes donc souvent devant le besoin de calculer des théories étendues c'est-à-dire des collections du type $\{(\phi,e) \in L \otimes E \mid q(\phi,r)$ est vrai et $e=\zeta(\phi,r)\}$ où e est le résultat de la fonction d'évaluation ζ dans r à valeur dans E. Par exemple, e peut être la fréquence relative (i.e., un nombre compris entre 0 et 1) d'un ensemble dans la base de données r. L'enjeu dans la conception d'une représentation condensée pour une théorie étendue ThE est alors de pouvoir compter sur des sous-ensembles de ThE à partir desquels il est possible de retrouver ThE exactement (on parle de représentation exacte qui, par exemple, permettent de retrouver tous les motifs fréquents et leurs fréquences sans effectuer de nouveaux accès aux données) ou de façon approximée (on parle alors de représentations approximatives qui permettent, par exemple, de retrouver tous les motifs fréquents mais avec une erreur bornée sur les fréquences dérivées). Il est clair que les représentations condensées basées sur la seule connaissance d'une frontière comme la collection des ensembles fréquents maximaux constitue une représentation approximative de la théorie étendue des ensembles fréquents et leurs fréquences. On peut borner la fréquence d'un ensemble sous-ensemble d'un ensemble maximal σ-fréquent comme étant comprise entre 1 et σ. Autrement dit, en considérant la valeur médiane, on connaît sa fréquence à $(100-\sigma)/2$ % près ce qui est clairement inacceptable dans les applications pratiques.

Le principe général de représentation condensée a été introduit dans [29] avec plus précisément le cadre des représentation ε-adéquate ou représentations alternatives des données pour répondre à une classe de requête, par exemple des requêtes de fréquence ou encore le calcul d'agrégats sur des cubes de données. A précision égale, on privilégie les représentations les plus concises qui sont alors qualifiées de condensées et, à notre connaissance, la première étude concrète pour une application aux ensembles fréquents est dans [30].

Dans le cas des représentations condensées d'ensembles fréquents, les représentations basées sur les ensembles fermés fréquents sont désormais bien étudiées. Il existe de nombreux algorithmes de calcul des ensemble fermés fréquents (par exemple [31,32,33,34,35,36]). Ces collections peuvent être de taille très petite devant le nombre d'ensembles fréquents et les algorithmes de régénération ont été proposés. Il est ainsi possible de calculer tous les ensembles fréquents dans des jeux de données difficiles car denses et/ou fortement corrélés, ce qui était impossible avec les variantes des algorithmes dérivés de apriori. Ceux-ci demandent effectivement le comptage de la fréquence de tous les ensembles fréquents alors que les algorithmes de calcul de fermés fréquents permettent d'inférer la fréquence d'un grand nombre d'ensembles sans accès aux données. Ces techniques sont abordées dans plusieurs articles de ce numéro qui témoigne aussi d'un rapprochement prometteur entre la communauté des concepts formels (treillis de Galois) spécialisée dans le calcul d'ensembles fermés et celle de l'ECBD avec les recherches sur les représentations condensées des motifs fréquents.
Une particularité de nombreux algorithmes calculant des ensembles fermés fréquents est de s'appuyer sur la recherche de générateurs des ensembles fermés qui ont eux-mêmes été étudiés comme des représentations condensées (par exemple les ensembles libres). C'est une très bonne illustration des techniques d'extraction sous contraintes efficaces. En effet, des algorithmes comme Pascal [34] ou AcMiner [33] calculent les ensembles fréquents libres dont les fermetures donnent par définition des ensembles fermés fréquents : l'intérêt est que la contrainte d'être un ensemble libre est une contrainte anti-monotone (contrairement à la contrainte d'être un ensemble fermé) et peut donc donner lieu à élagage dans l'espace de recherche. Les travaux sur les représentations condensées exactes ont aussi généralisé le concept d'ensemble libre avec les ensembles ou-libres [37,38] ou encore les ensembles non dérivables [39]. Les représentations condensées sont plus ou moins explicitement présentes dans de nombreuses contributions à ce numéro, qu'il s'agisse de représentations intermédiaires des données pour faciliter la découverte de motifs (cas des représentations des tables pour la découverte de dépendances) ou bien des treillis de Galois considérées comme des représentations condensées de règles d'association a priori intéressantes ou encore des représentations précitées utilisées pour assurer la faisabilité des calculs d'ensembles fréquents dans des cas difficiles.

Nous pouvons maintenant détailler le contenu de ce numéro.

Deux articles font une synthèse des progrès les plus récents dans le domaine de la découverte de dépendances dans des bases de données relationnelles.

L'article de Fabien de Marchi (LIMOS, Clermont-Ferrand) est une synthèse remarquable sur l'exploitation de l'algorithme générique de Mannila et Toivonen pour la découverte de dépendances d'inclusion dans des bases relationnelles. Il montre de façon didactique où sont les difficultés (et des solutions efficaces) pour réaliser une instance opérationnelle (identification des phrases les plus générales, ici les dépendances d'inclusion unaires, génération des candidats, et enfin test de la

satisfaction des dépendances). L'article de Stéphane Lopes (PRISM, Versaille) et Noël Novelli (LaBRI, Bordeaux) traite de la découverte de dépendances fonctionnelles. Ces deux auteurs ont contribué aux développements les plus récents dans ce domaine. Il est aujourd'hui possible de calculer une couverture des dépendances fonctionnelles dans des bases opérationnelles. Ces progrès présentent un potentiel intéressant pour l'optimisation sémantique de requêtes, la maintenance et la rétro-conception de bases de données.

De très nombreuses recherches se sont concentrées sur l'extraction des ensembles fréquents dans des données transactionnelles et des dizaines de chercheurs ont conçu des algorithmes d'extraction efficaces pour, notamment, calculer des règles d'association a priori intéressantes depuis de très grands volumes de données. François Rioult (GREYC, Caen) propose une synthèse conceptuelle sur ces travaux algorithmiques. Là encore, l'algorithme générique de Mannila et Toivonen sert de support à la discussion et l'article présente plusieurs travaux sur les représentations condensées d'ensembles fréquents et leurs applications. Ces représentations condensées sont particulièrement pertinentes, qu'il s'agisse de calculer des ensembles fréquents dans des cas difficiles ou encore de dériver des motifs (typiquement des règles) a priori plus pertinents que ceux que l'on dérive des ensembles fréquents eux-mêmes.

L'article de Sadok Ben Yahia et Engelbert Mephu Nguifo (CRIL, Lens) concerne justement une synthèse des approches d'extraction de règles d'association basées sur la sémantique de la correspondance de Galois, et donc les notions d'ensembles fermés associés par les opérateurs de Galois (également appelés concepts formels). De tels liens ont été étudiés à travers les nombreuses propositions de calcul de bases ou de couvertures pour les grandes collections de règles d'association (voir notamment [40,41,42]). Un point de vue un peu différent traité dans cet article consiste à considérer la production de règles à partir des treillis de concepts.

Huauguo Fu et Engelbert Mephu Nguifo (CRIL, Lens) proposent d'ailleurs de faire le point sur divers algorithmes de génération de concepts formels et donc de calcul de tels treillis. Ce faisant, ce numéro atteste du rapprochement actuel entre la communauté qui travaille sur les treillis de concepts et leurs applications en intelligence artificielle et celle qui est spécialisée dans la recherche de motifs ensemblistes dans des données transactionnelles. On notera, par exemple, que des travaux sur l'analyse de données d'expression de gènes (recherche de groupes de synexpression et de modules de transcription) ont pu déboucher grâce par l'exploitation des propriétés formelles des treillis de Galois [43].

Deux articles traitent de l'extraction de motifs dans des données séquentielles, c'est-à-dire soit de longues séquences soit des bases de séquences. La contribution de. Florent Masseglia (INRIA Sophia-Antipolis), Magdelonne Teisseire (LIRMM, Montpellier) et Pascal Poncelet (EMA/LGI2P, Nîmes) est un état de l'art dans ce domaine. Les principales familles d'algorithmes développés ces dernières années sont considérées et quelques applications importantes sont introduites. L'article de

Marion Leleu, Nicolas Méger et Christophe Rigotti (LIRIS, Lyon) présente un travail sur l'adaptation des algorithmes d'extraction de motifs séquentiels sous contraintes dits « à listes d'occurrences » (par exemple [44]) au cas des données contenant des répétitions consécutives.

Enfin, l'article de Cheikh Talibouya Diop (LI, Blois et Université Saint-Louis, Sénégal), Arnaud Giacommetti (LI, Blois), Dominique Laurent (LICP, Cergy-Pontoise), et Nicolas Spyratos (LRI, Orsay) considère le problème important de l'extraction itérative de requêtes fréquentes. Dans le contexte de l'extraction de requêtes fréquentes dans des bases de données relationnelles, ils étudient les propriétés formelles d'inclusion et d'équivalence qui peuvent permettre des optimisations importantes.

Plusieurs questions importantes en extraction de motifs n'ont finalement été que peu traitées et, de fait, pourraient faire l'objet de nouveaux appels à contributions. De notre point de vue, trois directions de recherche sont actuellement très importantes :

- *La question de la pertinence, notamment statistique, des motifs extraits demande de nouveaux développements. Il faut par exemple s'intéresser à la nature des données réelles (numériques imprécises, manquantes, ou bruitées) et encore peu de travaux étudient l'impact de ces contextes réalistes sur les méthodes d'extraction de motifs.*
- *Le problème de l'extraction de motifs sous contraintes en présence de contraintes qui ne sont pas anti-monotones reste plutôt ouvert. Si des progrès ont été obtenus dans le cas des extractions en présence de conjonctions de contraintes anti-monotones et monotones, des cas plus difficiles existent et doivent être étudiés (contraintes basées sur des agrégats ou des mesures statistiques, contraintes d'optimisation, disjonctions de contraintes). Une voie particulièrement prometteuse concerne les stratégies d'extraction adaptatives qui adaptent les méthodes d'élagage dynamiquement au cours du calcul (voir, par exemple ExAMiner dans [23] et la proposition RE-Hackle dans [45]).*
- *Il faut mieux comprendre les multiples usages des motifs locaux dans la construction de modèles. Des résultats préliminaires ont été obtenus sur les usages des représentations condensées pour, par exemple, l'aide au calcul de partitions ou encore la construction de systèmes de classification supervisée. Nous pensons qu'il faut poursuivre ce type de travail pour non seulement valoriser le savoir-faire en matière d'extraction de motifs locaux sous contraintes mais aussi résoudre les problèmes difficiles de la construction de modèles pertinents dans des données réelles.*

Nous allons justement évoquer certains de ces usages dans le domaine de la classification. Les représentations condensées des ensembles fréquents sont un solide point de départ pour la réalisation de nombreuses tâches ECBD. [46] est un

document de synthèse qui dresse un état de l'art des principaux usages des représentations condensées d'ensembles fréquents.

Nous avons déjà évoqué le calcul de règles d'association non redondantes. Ainsi, [40] étudie le calcul de règles informatives (règles à prémisse minimale construites sur un ensemble libre et conclusion maximale construite sur sa fermeture). On peut aussi travailler avec des règles construites sur des motifs delta-libres fréquents [47], l'un des rares cas de représentation approximative des ensembles fréquents. On obtient alors des règles à prémisse minimale et presque exactes car contredites par quelques exceptions. Ces règles sont particulièrement intéressantes dans le cas de données réelles comme en biologie ou en médecine. La possibilité de travailler dans des contextes transactionnels difficiles permet de considérer le calcul d'ensembles fréquents généralisés (avec codage simultané de la présence et de l'absence d'items). On peut ainsi chercher certaines formes de règles avec des négations (voir par exemple [20,48] pour une première approche). Un motif émergent est un motif dont la fréquence varie fortement entre deux ensembles de données (ou classes), le rapport entre ces deux fréquences formant son taux de croissance. Les motifs émergents caractérisent des classes de manière quantitative et qualitative et sont une solide assise pour construire des classifieurs. L'extraction de tous les motifs émergents reste une tâche difficile mais peut bénéficier des progrès récents puisque l'émergence se trouve être la conjonction d'une contrainte anti-monotone avec une contrainte monotone.

La classification supervisée à partir de règles a été très étudiée : en considérant des règles dont la conclusion est une des valeurs de classe, on sélectionne un sous-ensemble de ces règles pour construire un classifieur efficace, une efficacité généralement évaluée par le taux d'exemples bien classés sur des échantillons tests. L'utilisation d'heuristiques pour la construction de règles de classification est bien connue : les arbres de décision, par exemple, procèdent ainsi au détriment d'une optimalité globale. A l'inverse, la technique des règles d'association assure de produire toutes les règles aux seuils fixés, et donc également les meilleures. Cependant, cette exhaustivité conduit aussi à des règles peu fiables qu'il faut alors éliminer. Les techniques visant à éviter le sur apprentissage peuvent ici être utilisées et l'on peut combiner le savoir-faire des deux communautés pour construire des classifieurs à base de règles d'associations. Pour cela, l'idée la plus simple consiste, à partir de l'extraction de règles d'association sur une base d'apprentissage, de filtrer les règles et de ne retenir que celles concluant sur la classe. Alex Freitas a montré dans [49] les limites de cette approche en discutant les différences entre règles de classification et règles d'association. Plusieurs chercheurs ont cependant étudié la sélection de règles de classification à partir de telles règles d'association. Dans CBA [50], une fois les règles concluant sur la classe extraites, on construit un classifieur à partir d'un ordre de précédence sur les règles basé sur la confiance et la fréquence, d'une sélection des règles couvrant au moins un exemple de la base d'apprentissage et d'un élagage des règles dont le déclenchement entraîne le plus d'erreurs sur l'ensemble d'apprentissage. L'utilisation d'heuristiques positionne

CBA sous un angle relevant plus de la classification. Cette stratégie a été améliorée dans [51] par l'introduction de plusieurs seuils de fréquence et la combinaison avec les arbres de décision. CMAR [52] sélectionne les règles par des mesures statistiques basées sur le $\chi 2$ et par une évaluation de leur couverture sur la base d'apprentissage, le classement d'un nouvel exemple s'effectuant par un vote. CPAR [53] repose sur un algorithme glouton et apprend les règles distinguant les exemples d'une classe par rapport à ceux des autres classes. En argumentant que des méthodes telles que CBA et CMAR élague trop, [54] propose un algorithme de classification à base d'associations à deux niveaux et un élagage paresseux. Cette approche est généralisée dans [55] par une procédure de vote. Roberto Bayardo a proposé des techniques d'élagages pour le calcul de règles de classification [56] et une contrainte mesurant l'amélioration de la confiance d'une règle de classification par rapport à la confiance de toutes les règles construites à partir des items de sa prémisse [57]. Apriori-C [58] effectue une sélection préalable d'attributs conduisant à une diminution de la complexité du calcul de règles de classification. En partant de mesures statistiques, Li et al. [59] définissent le plus petit ensemble règles de classification qui a le même pouvoir prédictif que l'ensemble des règles de classification extraites. Ces exemples montrent que la littérature sur la classification à base d'associations est abondante. Rappelons que ces approches reposent, pour l'essentiel, sur l'extraction complète des règles d'association. Or, cette étape n'est pas toujours possible dans les bases de données denses ou corrélées. Nous avons-nous même contribué à ce domaine en explorant les possibilités offertes par les règles delta-fortes produites à partir de la représentation condensée approximative des ensembles delta-libres [47]. Avec la technique décrite dans [60], les règles delta-fortes de caractérisation de classes possèdent une propriété de minimalité et d'absence de certains conflits de classification particulièrement pertinente.

D'autres grands domaines de l'ECBD comme le calcul de partitions (par exemple la classification non supervisée à base d'associations) peuvent profiter des progrès en matière d'extraction de motifs locaux, et notamment les représentations condensées d'ensembles fréquents. C'est un domaine en forte évolution et nous renvoyons le lecteur à l'état de l'art de [61].

Nous espérons que vous prendrez autant de plaisir à lire ce numéro thématique que nous en avons eu à le préparer. Pour terminer, nous souhaitons remercier ici les auteurs des contributions soumises et les membres du comité de lecture qui nous ont aidés à en composer le contenu : Dominique Laurent (Université de Cergy-Pontoise), Amédéo Napoli (LORIA, Nancy), Jean-Marc Petit (LIMOS, Clermont-Ferrand), Pascal Poncelet (Ecole des Mines d'Alès), Christophe Rigotti (LIRIS, Lyon), Céline Rouveirol (Institut Curie, Paris), Michèle Sebag (LRI, Orsay), Maguelonne Teisseire (LIRMM, Montpellier), Jean-Daniel Zucker (LIM & BIO, Paris), et Gilles Zurfluh (IRIT, Toulouse).

Jean-François Boulicaut (LIRIS, INSA Lyon)
Bruno Crémilleux (GREYC, Université de Caen)

Références

[1] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Advances in Knowledge Discovery and Data Mining, “From data mining to knowledge discovery: an overview” AAAI/MIT Press, 1996. pp. 1-36.
[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. Proceedings ACM SIGMOD 93, Washington D.C., USA, May 2003, pp. 207-216.
[3] H. Mannila et H. Toivonen. Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery, 1(3):241-258, 1997.
[4] T. M. Mitchell. Generalization as search. Artificial Intelligence, 18:203-226, 1982.
[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo. Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. pp. 307-328.
[6] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of the ACM, 39(11):58-64, November 1996.
[7] H. Mannila. Inductive databases and condensed representations for data mining. Proceedings ILPS'97, Port Jefferson, Long Island N.Y., USA, October 1997. MIT Press. Pp. 21-30.
[8] J-F. Boulicaut, M. Klemettinen, and H. Mannila. Modeling KDD processes within the inductive database framework. Proceedings DaWaK'99, Florence, I, September 1999, Springer-Verlag LNCS 1676, pp. 293-302.
[9] F. Giannotti and G. Manco. Querying inductive databases via logic-based user defined aggregates. Proceedings PKDD'99, Praha, CZ, September 1999. Springer-Verlag LNAI 1704, pp. 125-135.
[10] H. Mannila. Theoretical frameworks for data mining. SIGKDD Explorations, 1(2):30-32, 2000.
[11] E. Baralis and G. Psaila. Incremental refinement of mining queries. Proceedings DaWaK'99, Florence, I, September 1999, Springer-Verlag LNCS 1676, pp. 173-182.
[12] R. Meo, G. Psaila, and S. Ceri. An extension of SQL for mining association rules. Data Mining and Knowledge Discovery, 2(2):195-224, 1998.
[13] R. Meo, P-L. Lanzi, and M. Klemettinen Eds. Database Support for Data Mining Applications – discovering knowledge with inductive queries, Springer-Verlag LNCS 2682. July 2004.
[14] J-F. Boulicaut. Inductive databases and multiple uses of frequent itemsets: the cInQ approach. In [13], pp. 3-26.
[15] L. de Raedt. A perspective on inductive databases. SIGKDD Explorations, 4(2):66-77, 2002.
[16] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. Proceedings KDD'97, Newport Beach, CA, USA, August 1997, AAAI Press, pp. 67-73.
[17] M. M. Garofalakis, R. Rastogi, and K. Shim. SPIRIT: Sequential pattern mining with regular expression constraints. Proceedings VLDB'99, Edinburgh, UK, September 1999, Morgan Kaufmann, pp. 223-234.
[18] J-F. Boulicaut and B. Jeudy. Using constraint for itemset mining: should we prune or not? Actes BDA'00, Blois, F, Novembre 2000, pp. 221-237.
[19] R. Ng, L.V.S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. Proceedings ACM SIGMOD'98, Seattle, Washington, USA, May 1998, pp. 13-24.
[20] B. Jeudy. Optimisation de requêtes inductives: application à l'extraction sous contraintes de règles d'association. Thèse de doctorat INSA Lyon, 2002.

[21] L. De Raedt and S. Kramer. The levelwise version space algorithm and its application to molecular fragment finding. Proceedings IJCAI'01, Seattle, WA, USA, August 2001, Morgan Kaufmann, pp. 853-862.
[22] C. Bucila, J. Gehrke, D. Kifer, W. M. White. DualMiner: A dual-pruning algorithm for itemsets with constraints. Data Mining and Knowledge Discovery, 7(3): 241-272 (2003)
[23] F. Bonchi. Frequent pattern queries: languages and optimizations. Ph.D. Thesis Università di Pisa TD – 10/03, 2003.
[24] L. De Raedt, M. Jäger, S. D. Lee, and H. Mannila. A theory of inductive query answering. Proceedings IEEE ICDM'02, Maebashi, Japan, December 2002, pp. 123-130.
[25] R. Bayardo. Efficiently Mining Long Patterns from Databases. Proceedings ACM SIGMOD'98, Seattle, Washington, USA, May 1998, pp. 85-93.
[26] S. Kramer, L. De Raedt and C. Helma. Molecular feature mining in HIV data. Proceedings ACM SIGKDD'01, San Francisco, CA, USA, August 2001, pp. 136-143.
[27] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. Proceedings ACM SIGKDD'99, San Diego, CA, 1999, pp. 43-52,
[28] X. Zhang, G. Dong, and K. Ramamohanarao. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. Proceedings ACM SIGKDD'00, Boston, MA, USA, August 2000, pp. 310-314.
[29] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. Proceedings KDD'96, Portland, Oregon, USA, August 1996. AAAI Press, pp. 189-194.
[30] J-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. Proceedings PaKDD'00, Kyoto, JP, April 2000, Springer-Verlag LNAI 1805, pp. 62-73.
[31] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. Information Systems 24(1):25-46, 1999.
[32] N. Pasquier. Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données. Thèse de doctorat, Université Blaise Pascal, Aubière, France, Janvier 2000.
[33] J-F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. Proceedings PKDD'00, Lyon, F, September 2000, Springer-Verlag LNAI 1910, pp. 75-85.
[34] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. SIGKDD Explorations, 2(2):66-75, December 2000.
[35] J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. Proceedings ACM SIGMOD Workshop DMKD'00, Dallas, Texas, USA, May 2000, pp. 21-30.
[36] M. J. Zaki and C. J. Hsiao. CHARM: an efficient algorithm for closed itemset mining. Proceedings SIAM DM 2002, Arlington, VA, USA, April 2002. pp. 33-43.
[37] A. Bykowski and C. Rigotti. A condensed representation to find frequent patterns. Proceedings ACM PODS'01, Santa Barbara, CA, USA, May 2001. pp. 267-273.
[38] A. Bykowski. Condensed representations of frequent sets: application to descriptive pattern discovery. Thèse de doctorat INSA Lyon, 2002.
[39] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. Proceedings PKDD'02, Helsinki, FIN, August 2002. Springer-Verlag LNAI 2431, pp. 74-85.
[40] Y. Bastide. Data mining : algorithmes par niveaux, techniques d'implantation et applications. Thèse de doctorat, Université Blaise Pascal, Aubière, France,
[41] V. Phan Luong. Reasoning on association rules. Actes BDA'01, Agadir, Maroc, Novembre 2001, Cépadues. pp. 299-310.
[42] M. J. Zaki. Generating non-redundant association rules. Proceedings ACM SIGKDD'00, Boston, MA, USA, August 2000, pp. 34-43.

[43] F. Rioult, C. Robardet, S. Blachon, B. Crémilleux, O. Gandrillon, and J-F. Boulicaut. Mining concepts from large SAGE gene expression matrices. Proceedings KDID'03 co-located with ECML PKDD'03, Dubvronik, Croatia, 2003, pp. 107-118.
[44] M. J. Zaki. SPADE: an efficient algorithm for mining frequent sequences. Machine Learning 42(1-2):31-60, 2001.
[45] A. Albert-Lorincz and J-F. Boulicaut. Mining frequent sequential patterns under regular expressions: a highly adaptive strategy for pushing constraints. Proceedings SIAM DM 2003, San Francisco, CA, USA, May 2003. pp. 316-320.
[46] B. Crémilleux. Contribution aux usages de représentations condensées de motifs dans des processus d'extraction de connaissances dans les bases de données. Mémoire d'Habilitation à Diriger des Recherches, Université de Caen Basse Normandie, 2004.
[47] J-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. Data Mining and Knowledge Discovery, 7(1):5-22, 2003.
[48] J-F. Boulicaut, A. Bykowski, and B. Jeudy. Towards the Tractable Discovery of Association Rules with Negations. Proceedings FQAS'00, Warsaw, PL, October 2000, Physica-Verlag, Advances in Soft Computing. pp. 425-434.
[49] A. A. Freitas. Understanding the crucial differences between classification and discovery of association rules - a position paper. ACM SIGKDD Explorations, 2(1):65-69, 2000.
[50] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rules mining. Proceedings KDD'98, New York, USA, August 1998, AAAI Press, pp. 80-86.
[51] B. Liu, Y. Ma, and C. K. Wong. Improving an association rule based classifier. Proceedings PKDD'00, Lyon, F, September 2000. Springer-Verlag LNAI 1910, pp. 504-509.
[52] W. Li, J. Han, and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. Proceedings IEEE ICDM'01, San Jose, CA, USA, December 2001, pp. 369-376.
[53] X. Yin and J. Han. CPAR: classification based on predictive association rules. Proceedings SIAM DM'03, San Francisco, CA, USA, May 2003.
[54] E. Baralis and P. Garza. A lazy approach to pruning classification rules. Proceedings IEEE ICDM'02, Maebashi City, Japan, December 2002, pp. 35-42..
[55] E. Baralis and P. Garza. Majority classification by means of association rules. Proceedings PKDD'03, Dubvronik, Croatia, September 2003. Springer-Verlag LNAI 2838, pp. 35-46.
[56] R. J. Bayardo. Brute-force mining of high-confidence classification rules. Proceedings KDD'97, Newport Beach, CA, USA, August 1997. AAAI Press, pp. 123--126.
[57] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense database. Proceedings IEEE ICDE 99, Sydney, Australia, March1999, pp. 188-197.
[58] V. Jovanoski and N. Lavrac. Classification rule with Apriori-C. Proceedings EPIA'01, Porto, Portugal, 2001. pp. 44-51.
[59] J. Li, H. Shen, and R. Topor. Mining the optimal class association rule set. Knowledge-based systems, 15(7):399-405, 2002. Elsevier Science.
[60] B. Crémilleux and J-F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. Proceedings ES'02, Cambridge, UK, December 2002. pp. 33-46.
[61] N. Durand. Extraction de clusters à partir du treillis de concepts : application à la découverte de communautés d'intérêt pour améliorer l'accès à l'information. Thèse de doctorat, Université de Caen, France, 2004.