20 ans de découverte de motifs : une étude bibliographique quantitative

Arnaud Giacometti, Dominique Haoyuan Li, Arnaud Soulet

Université François Rabelais Tours, LI
3 place Jean Jaurès
F-41029 Blois France
prenom.nom@univ-tours.fr

Résumé. Depuis deux décennies, la découverte de motifs a été l'un des champs de recherche les plus actifs de l'exploration de données. Cet article en établit une étude bibliographique quantitative en nous appuyant sur 1030 publications issues de 5 conférences internationales majeures : KDD, PKDD, PAKDD, ICDM et SDM. Nous avons d'abord mesuré depuis 2005 un sévère ralentissement de l'activité de recherche dédiée à la découverte de motifs. Puis, nous avons quantifié les principales contributions en terme de langages, de contraintes et de représentations condensées de sorte à comprendre ce ralentissement et à esquisser les directions actuelles.

1 Introduction

En 1993, Rakesh Agrawal, Tomasz Imielinski et Arun N. Swami publiaient l'un des articles fondateurs de la découverte de motifs "Mining Association Rules between Sets of Items in Large Databases" dans les actes de "the ACM SIGMOD International Conference on Management of Data" en introduisant le problème de l'extraction des règles d'association intéressantes. Formellement, ce problème consiste à énumérer toutes les règles du type $X \rightarrow I$ où X est un ensemble d'items et I un item non présent dans X telles que les probabilités $P(X \rightarrow I)$ et $P(X \rightarrow I|X)$ soient suffisamment élevées. Moins qu'une finalité nouvelle, Agrawal et al. (1993) ont surtout substitué à la traditionnelle recherche heuristique, la recherche complète et consistante. En effet, le problème de la découverte de règles de classification (où I est seulement une valeur de classe) était déjà une thématique de recherche active dans le domaine de l'intelligence artificielle mais les algorithmes n'étaient pas exhaustifs (Breiman et al., 1984),(Quinlan, 1986),(Piatetsky-Shapiro, 1991). Ces notions de complétude et consistance très importantes en bases de données justifient la publication de Agrawal et al. (1993) dans ACM SIGMOD. De même, la généralisation proposée l'année suivante par Agrawal et Srikant (1994) (où la conclusion de la règle est désormais un ensemble d'items) est publiée dans VLDB [1].

1. D'ailleurs, Rakesh Agrawal revendique son affiliation au domaine des bases de données dans Winslett (2003) :"I'm a database person, so my view of data mining has been that it is essentially a richer form of querying. We want to be able to ask richer questions than we could conveniently ask earlier."

Depuis 20 ans, la communauté de la *découverte de motifs* n'a cessé de puiser son inspiration dans l'article fondateur (Agrawal et al., 1993) comme en témoignent ses nombreuses citations : 26ème article le plus cité de l'informatique selon CiteSeer, le 5ème le plus cité dans le domaine de l'exploration de données selon Microsoft Acamedic Research, plus de 11 900 citations selon GoogleScholar [2]. Pour mieux définir ce qu'est la découverte de motifs, revenons au problème initial qui se divise en deux sous-problèmes : (1) trouver tous les motifs ensemblistes (itemsets) présents dans au moins x% des transactions et (2) générer à partir de ces motifs toutes les règles d'association intéressantes. Cette subdivision illustre les deux grandes problématiques qui ont animées la découverte de motifs pendant 20 ans : l'*extraction de motifs* et l'*utilisation des motifs*. Premièrement, l'extraction de motifs consiste à énumérer tous les motifs d'un *langage* (e.g., les itemsets ou les séquences) qui satisfont une *contrainte* (e.g., une fréquence minimale). Il est toutefois possible de se contenter d'une *représentation condensée* de ces motifs i.e., une fraction des motifs qui garantit encore la regénération totale des règles. Ces trois dimensions (i.e., langage, contrainte et représentation) illustrées dans (Mannila et Toivonen, 1997) donnent lieu à un très grand nombre de problèmes. Deuxièmement, l'utilisation des motifs consiste à combiner plusieurs motifs pour constuire des modèles plus complexes et/ou plus généraux. Typiquement, des classifieurs sont construits en regroupant des règles de classification, elles-mêmes dérivées de motifs ensemblistes.

Cet article ambitionne de faire une étude des travaux de la découverte de motifs réalisés de 1995 à 2011. Plutôt que de proposer une étude bibliographique basée sur quelques dizaines d'articles et forcément parcellaire, nous avons opté pour une vision plus globale en s'appuyant sur plusieurs centaines d'articles. Il s'agit de 1030 articles consacrés à la découverte de motifs issus des 6223 articles publiés dans les 5 conférences majeures de l'exploration de données : KDD, PKDD, PAKDD, ICDM et SDM (le protocole est détaillé en section 2). Outre la vision globale, notre corpus d'étude est suffisamment conséquent pour quantifier des phénomènes et ce dans le temps. Les traitements automatisés porteront sur les titres des publications tandis que nous nous appuierons parfois manuellement sur les résumés afin de lever certaines ambiguïtés. Dans un premier temps, nous proposons de positionner la découverte de motifs au sein de l'exploration de données (section 3). Nous désirons également mieux cerner le domaine à la lumière des 3 dimensions mentionnées ci-avant : le langage, la contrainte et la représentation (section 4). Nous souhaitons pouvoir répondre à des questions simples : quels sont les langages les plus étudiés ? les motifs maximaux ont-ils été davantage étudiés que les générateurs ? etc. Au-delà des réponses, nous nous attachons à une quantification précise de ces phénomènes.

2 Périmètre de l'étude

2.1 Frontière de l'Exploration de Données

La base de données se compose de toutes les conférences dont l'intitulé contient "Data Mining" et classées en rang A selon l'organisme Computing Research and Education Association of Australia [3]. Ces 5 conférences font largement référence dans le domaine de l'exploration de données : KDD (ACM SIGKDD conference on Knowledge Discovery and Data mining),

2. citeseerx.ist.psu.edu/stats/articles (mai 2012) ; academic.research.microsoft.com (octobre 2012) ; scholar.google.com (novembre 2012)

3. www.core.edu.au

Conférence	Début	Indexation	Nbr. de publi. KDD	Nbr. de publi. PM	ratio
KDD	1994	1995	1699	233	0,14
PKDD	1997	1997	1171	193	0,16
PAKDD	1997	1998	1191	246	0,21
ICDM	2001	2001	1447	254	0,18
SDM	2001	2002	715	104	0,15
		Total	6223	1030	0,17

TAB. 1 – *Conférences "Data Mining" de rang A et leurs publications indexées sous DBLP.*

PKDD [4] (European Conference on Principles of Data Mining and Knowledge Discovery), PAKDD (Pacific Asia Knowledge Discovery and Data Mining), ICDM (IEEE International Conference on Data Mining) et SDM (SIAM International Conference on Data Mining) [5]. Ce choix a peut-être tendance à exclure des travaux plus matures publiés en revue et à l'inverse des travaux plus prospectifs publiés en atelier. De même, ce travail manque des travaux publiés dans des conférences connexes notamment en base de données (e.g., VLDB) ou en Recherche d'Information (e.g., ICKM). Cependant, intégrer ces conférences à l'étude aurait dilué l'essence de l'exploration de données (et donc, celle de la découverte de motifs). Au final, nous estimons que les 350 publications annuelles issues des 5 conférences retenues constitue un échantillon significatif pour une étude statistique de l'ensemble de la production mondiale.

Cette étude porte sur les titres des publications indexées sous The DBLP Computer Science Bibliography [6] pour les 5 conférences. Même si le volume et les types de publication (e.g., articles longs, posters, tutoriaux, panels) varient selon la conférence et les années, la très grande majorité concerne des articles longs et courts. Par ailleurs, la première édition des conférences KDD, PAKDD et SDM n'ont pas été indexées sous DBLP. Au final, le tableau 1 récapitule pour les 5 conférences l'année de la première édition, l'année de la première indexation des publications sous DBLP et le nombre total de publications indexées. Si les traitements automatisés portent exclusivement sur les titres, la validation manuelle de ces traitements s'est appuyée sur les résumés lorsque le titre se révélait insuffisant. Le traitement de l'intégralité des articles permettrait probablement d'améliorer le filtrage automatique (à condition d'utiliser des méthodes du TAL pour circonscrire avec justesse les contributions de l'article analysé du reste).

La figure 1 (a) reporte le nombre de publications en exploration de données entre 1995 et 2011 pour chacune des conférences. Ce nombre n'a cessé de croître sur ces 17 années (excepté en 2007 et en 2010) traduisant l'essor du domaine. Cette augmentation globale s'explique à la fois par la création de nouvelles conférences (jusqu'en 2002) et l'augmentation du nombre de publications pour chacune des conférences (e.g., plus 88% de publications depuis 2002).

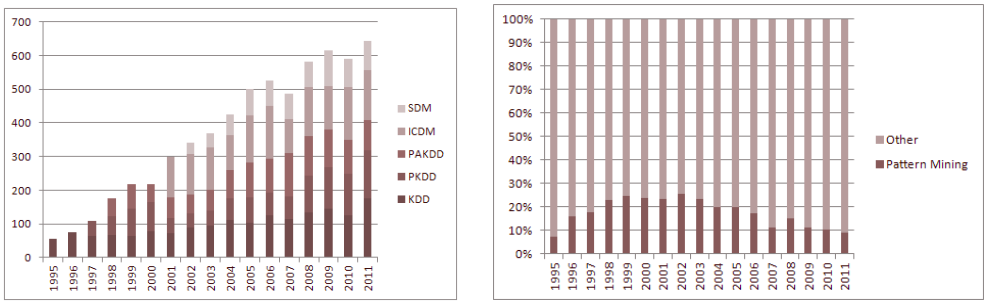
2.2 Frontière de la découverte de motifs

La difficulté majeure consiste à déterminer quelles sont les publications concernant la découverte de motifs. Nous ne nous limitons pas aux travaux dédiés à l'extraction de motifs mais

4. PKDD a été rattachée en 2001 à ECML (European Conference on Machine Learning) puis les deux conférences ont fusionnées en 2008. A partir de 2008, PKDD correspond donc à ECML/PKDD.

5. `www.kdd.org`; `www.ecmlpkdd.org`; `pakdd.togaware.com`; `www.cs.uvm.edu/~icdm`; `www.siam.org/meetings/archives.php#sdm`

6. `www.informatik.uni-trier.de/~ley/db`



(a) Nombre de publications en exploration de données (b) Proportion des publications en découverte de motifs

FIG. 1 – *Evolution du nombre de publications.*

considérons également les travaux où ces derniers sont utilisés (par exemple pour construire un modèle). Par motif, nous entendons motif *local* au sens de Hand (2002) i.e., qui ne décrit qu'une portion de la base de données. Pour cette raison, nous considérons que les arbres de décision, les réseaux bayésiens, les réseaux de neuronnes ou les machines à vecteurs de support (SVM) ne sont pas des travaux concernant l'extraction ou l'utilisation de motifs locaux.

Nous décrivons maintenant le processus semi-automatique qui par raffinement successifs, a permis de déterminer les publications concernant la découverte de motifs.

1. **Filtrage automatique :** Cette première étape a consisté à conserver tous les articles susceptibles de concerner la découverte de motifs. Nous avons retenu toutes les publications dont le titre mentionne un terme relatif à la découverte de motifs. Cette liste a été établie en nous appuyant sur les trois dimensions mentionnées en introduction :
   (a) *Langage :* pattern, item, sequence, rule, tree, graph, string
   (b) *Contrainte :* support, frequent
   (c) *Représentation condensée :* free, generator, closed, condensed, concise

   Bien sûr, cette liste de termes est élargie à leurs variations (e.g., pour le terme string, on considère substring, strings, etc). Ainsi, 1590 publications ont été retenues soit environ un quart de la base de données.
2. **Filtrage manuel :** Cette seconde étape a consisté à éliminer tous les faux positifs i.e., les publications ne concernant pas la découverte de motifs mais contenues dans les 1590 publications obtenues à l'étape 1. Pour ce faire, nous avons lu un à un chaque titre et si nécessaire, nous avons consulté le résumé de l'article voire son intégralité pour lever toute ambiguïté. 1030 publications ont été retenues par le filtrage manuel sachant que le résumé a été consulté dans 126 cas.

Il est clair que la solution de filtrage automatique a été priviligiée pour éviter de parcourir les 6223 publications. Même si le filtrage repose sur une liste de termes assez ouverte, des articles pertinents ont été manqués (i.e., des faux négatifs). Afin d'estimer ce nombre de faux négatifs, nous avons tiré au hasard 100 publications parmi les 4633 éliminées à l'étape 1. 5 de ces publications concernaient bien la découverte de motifs. Par conséquent, nous estimons

qu'environ 232 ($= 0.05 \times (6223 - 1590)$) publications relatives à la découverte de motifs ont été manquées par notre approche. Dans la suite, nous faisons l'hypothèse que les 1030 publications retenues sont un échantillon représentatif de la découverte de motifs.

Clairement notre approche repose sur une part de subjectivité à l'étape 1 dans le choix des termes retenus et à l'étape 2 dans l'élimination des faux positifs. Nous pensons néanmoins que ce protocole n'introduit pas plus de subjectivité qu'une étude bibliographique traditionnelle.

3 Décroissance de la découverte de motifs

Le tableau 1 reporte le nombre de publications relatives à la découverte de motifs par conférence et calcule quelles proportions ces publications représentent. La découverte de motifs, en correspondant à environ 1 article sur 6 de l'exploration de données, en constitue un réel sous-domaine. Parmi les 8489 auteurs qui ont contribués aux 5 conférences, 1685 d'entre eux (soit $19,85\%$) ont participé à au moins une publication en découverte de motifs. 851 auteurs (soit $10,02\%$) se sont cantonnés à des publications en découverte de motifs.

La figure 1 (b) présente l'évolution de la place de la découverte de motifs au sein de l'exploration de données. Nous remarquons une décennie florissante entre 1997 et 2006 où la part de la découverte de motifs est supérieure à sa part moyenne de 17%. Pendant 5 années, entre 1998 et 2003, 1 article sur 5 est même consacré à la découverte de motifs, l'âge d'or de la découverte de motifs en quelque sorte. Jusqu'en 1999, l'essor de la découverte de motifs est plus important que celui de l'exploration de données pourtant conséquent (cf. figure 1). Il est certain que KDD, PKDD et PAKDD étaient les conférences privilégiées pour diffuser les travaux naissants relatifs à la découverte de motifs (Piatetsky-Shapiro, 2000). A l'inverse, d'autres travaux (e.g., les arbres de décision) disposaient déjà de tribunes reconnues avec les conférences en apprentissage et intelligence artificielle. Peut-être que la reconnaissance des conférences "Data Mining" et leur attractivité accrue a pu attirer davantage de ces travaux à partir de 2002. Quoiqu'il en soit, depuis cette date, la part de la découverte de motifs a progressivement diminué jusqu'à atteindre moins de 10% en 2011 (proche des 7% de 1995). Cette baisse relative se traduit aussi par une baisse absolue avec une chute de 99 articles en 2005 à 59 articles en 2011.

Pour finir, nous introduisons un nouvel indicateur afin d'estimer le dynamisme d'un domaine via son ensemble de publications. Pour commencer, la *fraîcheur* mesure le degré de nouveauté d'une publication dans le domaine (par rapport à la période de 1995 à 2011) : $Freshness(p) = (year(p) - 1995)/16$ où $year(p)$ est l'année de la publication de p. Plus la fraîcheur d'une publication est proche de 1, plus elle est récente i.e., proche de 2011. A l'inverse, une fraîcheur égale à 0 signifie que la publication date de 1995. Nous étendons alors cet indicateur à un ensemble de publications P en en calculant la valeur moyenne : $Freshness(P) = 1/|P| \times \sum_{p \in P} Freshness(p)$. Cette métrique donne une tendance grossière sur le dynamisme d'un domaine via ses publications. Lorsque la fraîcheur d'un ensemble de publications atteint 1, cela signifie que les publications se concentrent sur les dernières années de la période 1995-2011. Par exemple, la fraîcheur des 6223 publications retenues pour l'étude est de 0,659 en raison de la hausse du nombre de publications annuelles. En comparaison, la fraîcheur des 1030 publications de la découverte de motifs est seulement de 0,590. Ce retrait traduit donc un affaiblissement de la découverte de motifs par rapport au reste de l'exploration de données. Dans la suite, nous utiliserons également la fraîcheur pour caractériser le

Langage	Mots-clés	Nombre	Proportion	$Freshness$
rule	association	335	0,33	0,455
itemset	set	320	0,31	0,638
sequence	episode, string, stream, protein, periodic, temporal	180	0,17	0,650
graph	molecular, structure, network	97	0,09	**0,725**
tree	xml	46	0,04	0,621
spatial	spatio-temporal	27	0,03	**0,694**
generic	–	17	0,02	**0,705**
relational	–	8	0,01	0,625

TAB. 2 – *Répartition des publications suivant le langage.*

dynamisme de certains langages, contraintes ou représentations condensées. Une thématique sera dynamique si sa fraîcheur excède celle de l'exploration de données (soit 0,659).

4 Typologie des publications

4.1 Les langages

Préparation des données Nous utilisons à nouveau une démarche par filtrage automatique puis correction manuelle. Tout d'abord, une liste de termes est associée à chaque type de langage comme indiqué par les deux premières colonnes du tableau 2 (sans les variantes). Aucun filtrage automatique n'est procédé pour les langages portant sur des motifs relationnels et génériques. Le langage "generic" correspond aux approches destinées à plusieurs langages (Mannila et Toivonen, 1997). Ces listes sont alors utilisées pour réaliser une première classification automatique de chaque publication sachant qu'une partie n'est pas classée (27.57% soit 284 articles). Ensuite, toutes les publications (y compris celles non-classées) sont parcourues manuellement pour une classification définitive (en utilisant le titre et le résumé). Lors de cette phase, il a été observé que par défaut le mot "pattern" réfère implicitement à "itemset" puisque 135 articles comportant ce mot correspondent aux motifs ensemblistes. La dernière colonne du tableau 2 reporte la fraîcheur des publications associées à chaque langage en mettant en gras celles avec une dynamique favorable par rapport à l'exploration de données ($> 0,659$).

Sans surprise, les règles d'association et les motifs ensemblistes à l'origine de la découverte de motifs sont les plus étudiés à hauteur d'environ 2/3 de l'ensemble. Ensuite, environ un quart concerne les séquences et les graphes. La découverte de motifs dans les données spatio-temporelles et les données relationnelles est assez marginale. Plus étonnant, nous constatons que très peu de travaux se sont attaqués à des approches génériques en terme de langage. Une explication probable est la difficulté de proposer des approches translangagières tant sur le plan théorique que sur le plan implémentation [7]. La fraîcheur élevée de ces articles (0,705) dénote cependant un intérêt plutôt récent pour ce type de travaux.

Complexification des langages Le tableau 2 montre que globalement plus un langage est complexe, moins il y a d'articles qui lui sont dédiés. Premièrement, la complexification des

7. Cependant, les articles de revue (absents de nos données) sont peut-être plus appropriés pour les approches translangagières faisant souvent la synthèse de travaux publiés précédemment pour des langages distincts.

langages entraine une marginalisation de par la difficulté intrinsèque liée à la combinatoire du problème. En effet, plus un langage est complexe, plus il est difficile d'extraire exhaustivement tous les motifs. Par exemple, avec 3 items, il est possible de former 80 séquences distinctes contre seulement 8 itemsets. Deuxièmement, cette complexification s'inscrit dans le temps comme le décrit la figure 2 : aux itemsets succèdent les séquences, puis les graphes. En fait, la capitalisation de savoir-faire diminue de langage en langage le nombre de verrous à lever. Typiquement, les méthodes d'élagage de l'espace de recherche pour les motifs ensemblistes (fondées sur l'anti-monotonie par exemple) sont transposables aux autres langages.

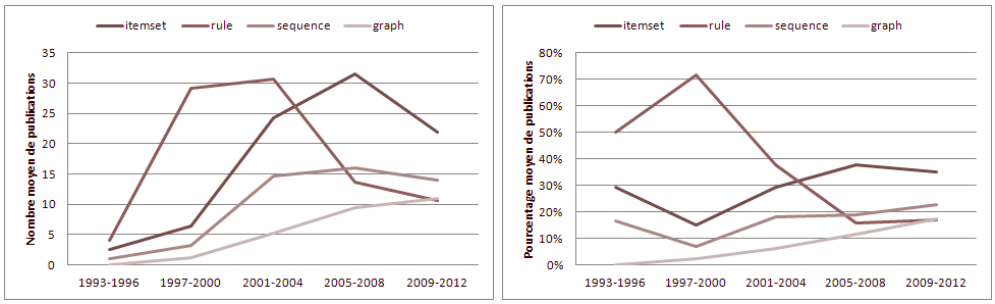


FIG. 2 – *Evolution du nombre de publications pour les principaux langages.*

Néanmoins, nous observons deux exceptions avec les arbres et les motifs ensemblistes qui sont respectivement moins étudiés que les graphes et les règles. Les arbres sont parfois simplifiés pour être traités comme une variante des séquences ou être traités comme un cas particulier des graphes. L'exception la plus notable concerne les motifs ensemblistes plus simples que les règles d'association et pourtant moins étudiés. Le fait que les règles soient particulièrement étudiées avant 2000 s'explique d'abord historiquement puisque l'extraction de règles de classification était déjà une thématique de recherche importante en intelligence artificielle, bien avant 1993. De plus, l'article fondateur de Agrawal et al. (1993) a désigné les règles d'association comme objectif ultime au détriment des itemsets considérés comme un outils intermédiaire (même si techniquement, l'obtention des itemsets est la phase la plus difficile). La faible fraîcheur des articles concernant les règles (0,455) renforce l'hypothèse de cet ancrage historique.

Limite de la complexification Tandis que le nombre de publications concernant les règles et les itemsets diminue, les deux langages les plus complexes (les séquences et les graphes) prennent une part toujours plus grande au sein de la découverte de motifs (cf. la figure 2). Par exemple, sur les 4 dernières années, la proportion d'articles consacrée aux graphes dépasse celle consacrée aux règles. Néanmoins, cette complexification semble atteindre ses limites puisqu'aucun langage successeur aux graphes (motifs relationnels ou spatio-temporels) ne semble pour l'instant prendre la relève de manière significative. Ces données ne sont peut-être pas à disposition en suffisamment grande quantité et il est possible que celles disponibles soient réduites à des langages plus simples comme les graphes. Etant donné la fraîcheur non négligeable pour le langage "spatial", cette conclusion provisoire pourrait être révisée prochainement. Enfin, à la complexification de la nature du langage pourrait succéder la complexifica-

Contrainte	Mots-clés	Nombre	Proportion	$Freshness$
regularity	frequent, support	253	0,50	0,630
contrast	emerging, discriminative	70	0,14	0,596
significant	chi-square, correlated	52	0,10	0,651
interesting	relevant	47	0,09	0,551
generic	monotone, anti-monone, constraint	41	0,08	0,637
exception	abnormal, surprising, anomaly, unexpected	30	0,06	0,554
utility		17	0,03	**0,724**

TAB. 3 – *Répartition des publications suivant la contrainte.*

tion des caractéristiques des données, nous avons observé que certains mots clés sont de plus en plus présents dans les titres : données incertaines (“uncertain” avec une fraîcheur de 0,892), données massives (“massive” avec 0,729) ou données dynamiques (“dynamic” avec 0,687).

4.2 Les contraintes

Préparation des données Nous utilisons à nouveau une démarche par filtrage automatique puis correction manuelle. Tout d’abord, une liste de termes est associée à chaque type de contraintes (cf. les deux premières colonnes du tableau 3). Ces listes sont utilisées pour réaliser une première classification automatique de chaque publication mettant en avant la notion de contrainte. Ensuite, les 510 publications classées sont parcourues manuellement pour une classification définitive en utilisant le titre et le résumé. Le terme “significant” regroupe les articles distinguant les motifs statistiquement valides des autres tandis que “interestingness” se rapporte à des approches plus variées souvent fondées sur des connaissances subjectives. A noter que le terme “generic” correspond aux approches dédiées à une classe de contraintes. Comme pour les langages (et pour les mêmes raisons), peu de publications sont consacrées aux contraintes génériques même si ces dernières sont plutôt récentes (cf. leur fraîcheur de 0,637).

L’obsession de la fréquence Globalement, la contrainte de fréquence minimale en représentant 50% des publications est de très loin la plus utilisée. En effet, de nombreuses publications s’attaque au sous-problème 1 (présenté en introduction) pour proposer une méthode plus efficace, pour l’adapter à un langage différent, pour se limiter à une représentation condensée, etc. Plus qu’une nécessité applicative, nous pensons que cette utilisation récurrente de la contrainte de fréquence découle du paradigme imposé par l’article originel de Agrawal et al. (1993).

Premièrement, remplacer la contrainte de fréquence par une autre plus sélective, c’est remettre en cause la regéneration *exhaustive* de toutes les règles intéressantes (le sous-problème 2). Hors cette exhaustivité est un fondement du paradigme qui entraine une surabondance assumée : “When we started doing data mining, we were concerned that we were generating too many rules, but the companies we worked with said, ‘this is great, this is what exactly what we want !’ ” (Winslett, 2003). Dès lors, la contrainte inspire toujours la peur de trop ou mal filtrer ce qui conduirait à perdre des motifs pertinents. Cette crainte s’illustre aussi avec l’obsession de diminuer le seuil de fréquence minimal quitte à extraire des motifs non-significatifs.

Deuxièmement, l’évaluation de l’approche dans (Agrawal et al., 1993) ne repose pas sur l’évaluation de la qualité des motifs extraits comme on valide par exemple la qualité d’un classifieur avec une validation croisée. Plus généralement, la plupart des articles d’extraction

de motifs n'évaluent pas la qualité des motifs extraits mais la rapidité de leur extraction ou la quantité de mémoire requise. De ce point de vue, améliorer le processus d'extraction de motifs revient à diminuer le coût temporel et/ou spatial, mais surtout pas à en modifier le résultat, i.e., les motifs fréquents. De plus, la contrainte de fréquence minimale, quel que soit le langage dispose de propriétés intéressantes (souvent issues de l'anti-monotonie) qui facilitent l'extraction. L'évaluation d'une méthode fondée sur une autre contrainte est alors doublement désavantageuse. De manière grossière, comme une contrainte pertinente ne satisfait en général pas la propriété d'anti-monotonie, l'algorithme d'extraction associé sera donc moins efficace que l'extraction des motifs fréquents. Par ailleurs, il est difficile de montrer que les motifs extraits suivant une nouvelle contrainte sont de meilleure qualité que ceux extraits avec la contrainte de fréquence minimale car il n'existe pas de protocole objectif de validation.

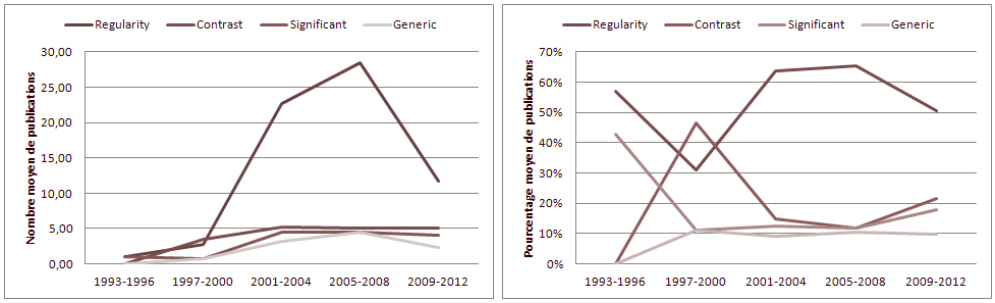


FIG. 3 – *Evolution du nombre de publications pour les principales contraintes.*

Vers plus de qualité Désormais, quelque soit le langage, l'extraction des motifs fréquents est une tâche maîtrisée. Pour cette raison, le nombre de publications mettant en avant les motifs fréquents diminue spectaculairement depuis 2005 (cf. la figure 3) expliquant en partie la décroissance de la découverte de motifs. Le défi combinatoire de la découverte de motifs cède sa place à celui de la qualité des motifs extraits. Ainsi, le recours à une contrainte pour affiner le filtrage gagne en légitimité suivant une perspective envisagée par Agrawal : "we need work to bring in some notion of 'here is my idea of what is interesting,' and pruning the generated rules based on that input." (Winslett, 2003). Cependant, la définition de ces contraintes demeure une problématique complexe. Déjà, Fayyad et al. (2003) indiquaient comme challenge de 2003 à 2013, la proposition d'une théorie générale de l'intérêt ("Interestingness"). Plus tard, Han et al. (2007) surenchérissaient "it is still not clear what kind of patterns will give us satisfactory pattern sets in both compactness and representative quality".

Malgré cette difficulté, la fraîcheur de certaines catégories et termes semble dessiner un renouveau de l'extraction de motifs sous contrainte. Même si la fraîcheur des contraintes dédiées aux contrastes est seulement de 0,596, on observe des fraîcheurs élevées de 0,893 pour "discriminative", de 0,812 pour "contrast" et de 0,736 pour "subgroup". Ce regain d'intérêt est aussi marqué pour les motifs significatifs (avec une fraîcheur de 0,651) et surtout pour les contraintes d'utilité (avec une fraîcheur de 0,724). Cette dynamique est également visible sur le graphique de droite de la figure 3. Enfin, une autre voie connexe au filtrage par seuillage

RC	Mots-clés	Nombre	Proportion	$Freshness$
closed	closure	68	0,58	**0,682**
border	maximal, minimal	24	0,20	0,599
free	generator, non-derivable, NDI	16	0,14	**0,676**
other		9	0,08	0,556

TAB. 4 – *Répartition des publications suivant le type de représentation condensée.*

est le classement des motifs en utilisant une mesure pour les ordonner comme l'illustrent les termes "ranking" et "top-k" avec une fraîcheur respective de 0,823 et 0,762.

4.3 Les représentations condensées

Préparation des données Nous utilisons à nouveau une démarche par filtrage automatique puis correction manuelle. Les deux premières colonnes du tableau 4 donne les représentations condensées et leurs termes associés. Ces listes permettent de classer automatiquement chaque publication en fonction de son titre. Ensuite, les 117 publications classées sont vérifiées manuellement pour une classification définitive en utilisant le titre et le résumé. La fraîcheur de chaque type de représentation condensée est indiquée dans la dernière colonne.

Pour rappel, l'objectif des représentations condensées est de réduire les redondances entre motifs (Calders et al., 2004). Les notions de bordures se contentent des motifs les plus spécifiques ou les plus généraux au sens de l'inclusion. Les motifs fermés et générateurs (libres ou clés) exploitent le même principe mais à fréquence égale. A noter que le terme "other" regroupe des articles se concentrant principalement sur les bases génériques de règles d'association.

Succès des représentations condensées : bordures puis motifs fermés 11,35% des publications relative à la découverte de motifs exploitent la notion de représentation condensée. Ce succès découle de leur indéniable bénéfice et de leur validation aisée (i.e., un contexte méthodologique opposé à celui des contraintes). Le concept de représentation condensée s'est rapidement imposée car il concilie la diminution du nombre de motifs et la conservation de l'exhaustivité via la regénération. De ce point de vue, elle s'inscrit parfaitement dans la perspective du sous-problème 2. En outre, les travaux sur les représentations condensées sont faciles à évaluer. D'une part, la validité de la regénération peut être formellement démontrée. D'autre part, la qualité de la condensation peut être estimée empiriquement en calculant le ratio de compression. Le plus souvent, ce gain en compression s'accompagne d'un gain de vitesse et d'une diminution des ressources mémoires consommées.

Parmi les différentes représentations, les bordures constituent la première représentation à succès à avoir été proposée (cf. figure 4) même si le nombre de travaux concernant les bordures décroît régulièrement depuis 1997-2000. Par nature, ces motifs maximaux/minimaux ont des propriétés extrêmes (e.g., fréquence très basse) et ne permettent pas d'inférer les propriétés des autres motifs (e.g., déduire la fréquence d'un motif plus général). Les motifs libres et fermés en dépassant ces limites se sont largement imposés. Finalement, le succès écrasant des fermés face aux générateurs s'explique par une conjonction de facteurs : moins nombreux, plus faciles à extraire et validité statistique plus forte (car ils véhiculent plus de régularités que les libres).

FIG. 4 – *Evolution du nombre de publications pour les représentations condensées.*

Bonne représentation mais mauvais modèle Désormais, les techniques concernant les représentations condensées sont bien maîtrisées et ce pour la plupart des langages. Le nombre de publications sur cette thématique a atteint son paroxysme entre 2005 et 2008. Seules les publications dédiées aux motifs générateurs et fermés se maintiennent dans le paysage de la découverte de motifs. Cependant, la taille des représentations condensées (exactes ou approximatives) reste trop importante pour autoriser une analyse globale des motifs. Il est alors nécessaire de mettre en œuvre d'autres mécanismes pour réduire leur taille soit en opérant un filtrage individuel des motifs (usage de contrainte) ou en opérant un filtrage collectifs des motifs (construction de modèle). Cette dernière option ressemble à l'objectif initial des représentations condensées mais en l'étendant à des motifs non-comparables (par rapport à l'inclusion). Cette direction peut être vue comme une valorisation des motifs : "to make frequent pattern mining an essential task in data mining, much research is needed to further develop pattern-based mining methods." (Han et al., 2007). Les termes "collection" et "pattern-based" avec une fraîcheur respective de 0,739 et 0,723 montre un intérêt récent pour cette piste.

5 Conclusion

Le titre "Extraction efficace des motifs fréquents fermés" serait probablement le meilleur résumé de la découverte de motifs, reprenant tous les codes de l'article originel : être consistant et exhaustif pour régénérer ; et ce le plus efficacement possible. De facto, tout ce qui concerne l'efficacité de l'extraction est désormais maîtrisé nous laissant face au principal problème : parmi l'abondance de motifs, trouver ceux qui sont réellement pertinents pour l'utilisateur. Traiter de nouveaux langages et de nouvelles sortes de données a juste déplacé ce problème ; utiliser des représentations condensées l'a à peine atténué. De nombreux travaux récents s'attaquent à ce challenge en s'appuyant sur un post-traitement pour sélectionner un sous-ensemble de motifs, sur des filtrages plus subtils ou sur des classements élaborés.

Enfin, si la flexibilité offerte par le choix du langage et de la contrainte offre un spectre d'applications a priori large, l'ancrage applicatif de la découverte de motifs reste moins fort que celui de l'exploration de données. En effet, de nombreux termes orientés application tels que "mobility", "multi-document", "patent", "multi-task", "advertising", "malware", "media", etc ; caractérisent les titres de l'exploration de données avec une fraîcheur supérieure à 0,9. A

l'inverse, la découverte de motifs se limite à moins de termes orientés application comme "social", "monitoring", "trajectory" ou "behavior", avec de surcroît des fraîcheurs plutôt réduites. Il semble ainsi qu'une valorisation plus importante de la découverte de motifs reste à conduire.

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pp. 207–216. ACM Press.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *VLDB*, pp. 487–499.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth.

Calders, T., C. Rigotti, et J.-F. Boulicaut (2004). A survey on condensed representations for frequent sets. In *Constraint-Based Mining and Inductive Databases*, Volume 3848 of *LNCS*, pp. 64–80. Springer.

Fayyad, U. M., G. Piatetsky-Shapiro, et R. Uthurusamy (2003). Summary from the kdd-03 panel : data mining : the next 10 years. *SIGKDD Explor. Newsl. 5*(2), 191–196.

Han, J., H. Cheng, D. Xin, et X. Yan (2007). Frequent pattern mining : current status and future directions. *Data Min. Knowl. Discov. 15*(1), 55–86.

Hand, D. J. (2002). Pattern detection and discovery. In *Pattern Detection and Discovery*, Volume 2447 of *LNCS*, pp. 1–12. Springer.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*(3), 241–258.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pp. 229–248. AAAI/MIT Press.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases : 10 years after. *SIGKDD Explor. Newsl. 1*(2), 59–61.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*(1), 81–106.

Winslett, M. (2003). Interview with rakesh agrawal. *SIGMOD Record 32*(3), 83–90.

Summary

For two decades, pattern discovery has been one of the most active fields in data mining. This paper provides a quantitative survey of the literature relying on 1030 publications from five major international conferences. We first measured a severe slowdown of research dedicated to pattern discovery. Then, we quantified the main contributions with respect to languages, constraints and condensed representations to outline the current directions.