FELIX: UN OUTIL INTERACTIF D'AIDE A LA FOUILLE DE CONNAISSANCES S'APPUYANT SUR L'INTENSITE D'IMPLICATION.

[1]REMI LEHN, [1]FABRICE GUILLET, [1]PASCALE KUNTZ,
[1]HENRI BRIAND, [1,2]JACQUES PHILIPPE[1]

RESUME : *Cet article décrit un outil de fouille de règles d'association dans une base de données relationnelle. Le système comporte trois composantes fortement interconnectées : une base de données qui stocke les données et des connaissances pré-calculées, un algorithme automatique qui sélectionne des sous-ensembles de données caractérisés par des conjonctions d'attributs et des règles d'association, et une interface graphique qui représente les relations entre les règles par un graphe orienté. Cette interface graphique, manipulable par l'utilisateur, lui permet de diriger sa recherche en se focalisant sur un sous-ensemble des connaissances potentiellement intéressantes, à travers des indices de qualité. En particulier, l'intensité d'implication permet de souligner, parmi ces sous-ensembles, les règles statistiquement étonnantes.*

MOTS-CLES : *Extraction de Connaissances à partir des Données (ECD), Intensité d'implication, Règles d'association, Fouille interactive.*

ABSTRACT : *This paper describes a rule mining tool. The system is composed by three tightly coupled components : a RDBMS which stores the data and the pre-computed knowledge, an automatic algorithm that selects data subsets characterized by attribute conjunctions and association rules, and a graphical interface to visualize discovered relations with directed acyclic graphs. This user-driven graphical interface allows him to focus on a subset of potentially interesting knowledge by using quality indices. More precisely, in this tool, the mining task relies on the intensity of implication, a relevant index in this context of implicative analysis.*

KEYWORDS : *Knowledge Discovery in Databases (KDD) - Intensity of implication - Association rules -Interactive mining.*

[1] [1]IRIN/IRESTE, La Chantrerie, BP60601, 44306 Nantes CEDEX 3
[2]PerformanSE SA, La Fleuriaye, Atlanpôle, BP 703, 44481 Carquefou Cedex, France

1 INTRODUCTION

L'accroissement très rapide des volumes d'information stockés dans les bases de données a conduit, lors de cette dernière décennie, au développement de l'Extraction de Connaissances à partir de Données (ECD). Ce champ interdisciplinaire a été défini de différentes manières par ses précurseurs (Frawley et al. 1992 [7]) ; d'une façon générale, il s'agit d'un processus de transformation dirigé par des modèles d'interprétation permettant de découvrir des connaissances potentiellement utiles cachées dans des données.

Cependant, à l'issu de ce processus, l'étape finale qui consiste à évaluer les résultats obtenus, se trouve la plupart du temps entièrement reportée sur l'utilisateur expert du domaine étudié.

VERS UN PILOTAGE ANTROPOCENTRE DE LA DECOUVERTE

L'utilisateur doit mener une tâche de fouille des connaissances effectivement utiles parmi celles produites, en procédant généralement par " tâtonnements successifs ". Ce besoin conduit l'ECD à devenir un processus itératif piloté par un utilisateur. Il s'avère alors nécessaire, ainsi que le soulignent différents travaux (Brachman et Anand 1996 [4]) de faire glisser le processus d'ECD vers un système d'aide à la décision dans lequel l'utilisateur tienne un rôle central.

La plupart des logiciels commercialisés, ne facilitent pas la tâche de fouille de connaissances de l'expert. En effet, le pilotage s'effectue principalement au niveau des algorithmes de découverte, à travers le choix de modèles d'interprétation ainsi que de leur paramétrage. Ce niveau de contrôle semble inadéquat lorsque l'utilisateur n'est pas un spécialiste des techniques de découverte de connaissances nécessaires au pilotage des algorithmes, mais un expert du domaine propre aux données étudiées. Cependant, les connaissances produites par les algorithmes de découverte sont généralement récapitulées dans des représentations synthétiques au graphisme abouti et facilement compréhensible pour l'utilisateur (Keim 1996 [12] pour une synthèse). Ces interfaces de visualisation des résultats constituent de notre point de vue un meilleur niveau de rétroaction sur les algorithmes de découverte puisqu'elles présentent des connaissances propres aux compétences de l'utilisateur. Malheureusement, celles-ci ne sont généralement pas dotées de capacités de manipulation qui permettraient le pilotage transparent des niveaux algorithmiques sous-jacents.

FELIX : UN OUTIL DE FOUILLE ANTROPOCENTREE DE REGLES D'ASSOCIATION

Dans cet article, nous présentons un outil de découverte de règles d'association d'associations (Agrawal et al. 1996 [1],[2]), Félix, qui comporte une interface centrée sur l'utilisateur afin de favoriser sa tâche de fouille des connaissances utiles, dans le cadre d'une analyse implicative (Gras et Lahrer 1993 [8], Gras *et al.* 2001 [9]). Nous nous référons aussi à des travaux sur le comportement de l'utilisateur menés en ECD (Bandhari 1994 [3]) et en décision (Svenson 1983 [16]), qui montrent l'importance de la focalisation (i.e. la possibilité de cibler de petits sous-ensembles d'attributs portant les règles intéressantes).

Félix associe trois composantes : une base de donnée relationnelle, un algorithme de calcul automatique de règles d'association, dont la qualité est mesurée par l'intensité d'implication ainsi que par la probabilité conditionnelle et le support, et une interface graphique interactive et dynamique.

L' interface implémente une double fonctionnalité :

- d'une part, elle sert de support visuel, sous une forme graphique adaptée aux résultats de l'algorithme de découverte automatique. Nous choisissons de présenter les règles d'association calculées sous la forme d'une graphe orienté.
- D'autre part, elle doit permettre à l'utilisateur d'interagir avec l'algorithme de découverte. Plus précisément, les actions de l'utilisateur sur le support visuel déclenchent un pilotage transparent l'algorithme de découverte qui en retour provoque une réactualisation du support.

L'utilisateur peut ainsi explorer la représentation graphique des règles d'association découvertes par l'algorithme et filtrée par leur intensité d'implication, et faire interactivement évoluer ce graphe de règles, par enrichissements successifs, afin de cibler les règles intéressantes selon son point de vue.

2 DECOUVERTE DE REGLES D'ASSOCIATION

2.1 REGLES D'ASSOCIATION

Les données présentées à l'entrée du processus d'ECD sont stockées dans une base de données relationnelle, et sont représentées par un ensemble d'objets $O=\{o_1, \ldots, o_n\}$ décrits par un ensemble fini d'attributs $A=\{a_1, \ldots, a_p\}$ à valeurs binaires {0,1}.

Nous recherchons à calculer en sortie du processus des règles d'association "intéressantes". Une règle d'association est une relation $X \rightarrow Y$ entre deux ensembles disjoints X et Y d'attributs, chacun de ces ensembles représentant une conjonction d'attributs. Ainsi la règle d'association $\{a_1, a_2, a_5\} \rightarrow \{a_3, a_6\}$ signifie que l'ensemble des objets vérifiant les attributs a_1, a_2 et a_5 ($a_1=1$, $a_2=1$, $a_5=1$) a tendance à vérifier aussi les attributs a_3 et a_6 ($a_3=1$, $a_6=1$). Ces règles ne sont pas certaines : en effet, il s'agit de quasi-implications admettant des contre-exemples. Il s'avère donc nécessaire de valider chaque règle en quantifiant la force de sa tendance implicative à l'aide de mesures de qualité. L'identification des règles intéressantes est en partie subjective : elle fait intervenir la connaissance et les préoccupations de l'utilisateur. Certains " critères objectifs ", directement calculables à partir des données, favorisent la focalisation de l'utilisateur sur des sous-ensembles des règles intéressantes.

2.2 MESURES DE LA QUALITE D'UNE REGLE

Intuitivement, une règle $X \rightarrow Y$ est significative, si elle est vérifiée par un nombre significatif d'objets de la base et est contredite par un petit nombre d'entre eux. De nombreuses mesures ont été proposées et étudiées dans la littérature (Mingers 1989 [13], Fleury *et al.* 1989 [6], Guillaume *et al.* 1998 [10]). Nous utilisons principalement trois indices de qualité complémentaires :

Support : le support d'une règle $X \rightarrow Y$ est défini par la proportion d'objets de O vérifiant le sous-ensemble d'attributs $X \cup Y$. Soit c: $2^A \rightarrow 2^O$ une fonction associant à un sous-ensemble X d'attributs, le sous-ensemble c(X) des objets vérifiant ces attributs. Le support $\rho(X \rightarrow Y)$ est le rapport $|c(X \cup Y)| / |O|$. Cette mesure de support quantifie le degré de généralité d'une règle : plus sa valeur est faible (resp. forte) plus la règle est spécifique (resp. générale)

Confiance : La confiance $\pi(X \rightarrow Y)$ de la règle $X \rightarrow Y$ est mesurée par la probabilité conditionnelle $\pi(X \rightarrow Y)=\text{proba}(Y|X)=|c(X \cup Y)|/|c(X)|$. Elle correspond à la proportion d'objets vérifiant Y parmi ceux vérifiant X. La confiance mesure donc le taux de validité d'une règle. Plus la confiance est faible (resp.forte) plus la règle est contredite (resp. valide).

La confiance est indépendante du nombre d'objets |O| et de la taille de la conclusion |c(Y)|, ainsi que d'un accroissement proportionnel de la prémisse et de la conclusion |c(X)| et |c(Y)|. De même, elle ne suffit pas à infirmer l'indépendance : la confiance peut-être élevée lorsque X et Y sont indépendants (Guillaume *et al.* 1998 [10] pour une étude plus poussée).

Afin de corriger ces effets indésirables, Gras et al. [8] ont défini une mesure appelée *intensité d'implication*.

Intensité d'implication : Soit nc $=\left|c(X)\cap c(Y)\right|$ le nombre de contre-exemples de la règle observée $X\rightarrow Y$. Soient U et V, deux sous-ensembles de O, de cardinalités respectivement égales à celles de la prémisse et de la conclusion |c(X)| et |c(Y)| de la règle observée.

Soit NC= $\left|U\cap\overline{V}\right|$ la variable aléatoire modélisant les contre-exemples. L'intensité d'implication est alors définie par : $\varphi(X\rightarrow Y)$=1- proba(NC $\leq$ nc). La qualité de la règle est d'autant meilleure que le nombre de contre-exemples observé est petit en comparaison du nombre calculé dans le cas aléatoire; c'est à dire que proba(NC$\leq$nc) est faible. Cette mesure quantifie l'étonnement statistique d'obtenir un nombre si faible de contre-exemples à la règle observée $X\rightarrow Y$.

3 ALGORITHME DE DECOUVERTE INCREMENTALE DIRIGE PAR L'UTILISATEUR

Nous utilisons, pour découvrir les règles, une version locale de l'algorithme *Apriori,* proposé par Agrawal et al. (Agrawal *et al.,* 1993, 1996, [1][2]). L'adaptation que nous en proposons est détaillée dans Kuntz *et al.* 2000 [13]. Il se décompose en deux étapes successives :

1. Découverte des attributs " fréquents ", c'est à dire, l'ensemble $L_\alpha=\{X\in 2^A | c(X)\geq\alpha\}$des conjonctions d'attributs ayant un support supérieur à un seuil α donné. Cette première étape permet de sélectionner les combinaisons d'attributs qui produiront des règles suffisamment générales.
2. Découverte locale de règles $X\rightarrow Y$ entre des attributs fréquents X et Y de L_α. De plus chaque règle est caractérisée par son support $\rho(X\rightarrow Y)$, sa confiance $\pi(X\rightarrow Y)$ et son intensité d'implication $\varphi(X\rightarrow Y)$.

Cette deuxième étape de découverte de règles locales est interactive. Grâce à ce calcul local, seules les règles requises par l'utilisateur sont calculées et l'explosion combinatoire liée au calcul global des règles dans l'algorithme apriori est ainsi évitée.

4 VISUALISATION ET INTERACTIONS

L'utilisateur visualise, à un instant *t*, les règles sous la forme d'un graphe orienté G_t dont les sommets sont des conjonctions d'attributs fréquents, et dont les arcs représentent les règles : un arc entre deux sommets X et Y représente la règle $X\rightarrow Y\backslash X$. Une solution à la difficulté inhérente d'une part au placement intelligible de ce graphe orienté et de surcroît dynamique, et d'autre part à la préservation de la carte mentale de l'utilisateur (Eades *et al.* 1991 [5]), a été proposée grâce à un algorithme génétique spécifique (pour plus de détails, consulter Guillet *et al.* 1999 [11] et Kuntz *et al.* 2000 [14]).

Il est intéressant de noter que ce graphe sert à la fois de modèle de représentation et de support de visualisation/interaction avec l'utilisateur. Ainsi, en interagissant avec la représentation visuelle du graphe, l'utilisateur déclenche un pilotage transparent des niveaux algorithmiques sous-jacents (fig. 1).

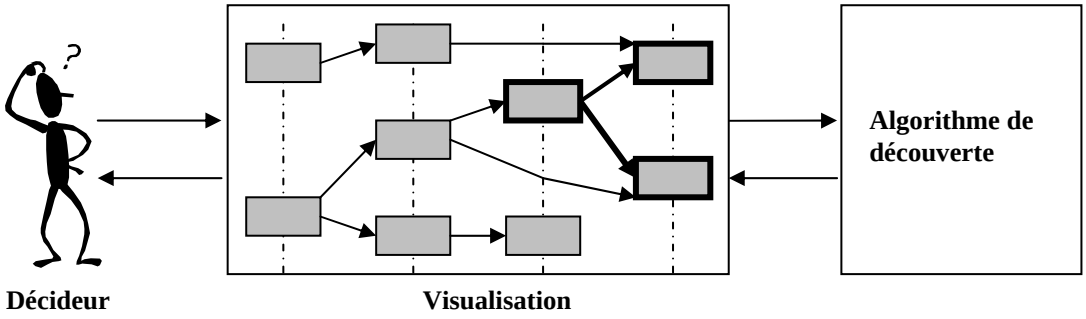

Fig. 1 : *Fouille supportée par la visualisation*

Trois opérateurs minimaux sont proposés pour passer d'un graphe de règles G_t à un graphe G_{t+1}:

- *Modification dynamique des seuils* de support (α_ρ), de confiance (α_π) et d'intensité d'implication (α_φ). L'utilisateur peut, à tout moment, faire varier ces seuils minimaux et ainsi à travers l'intensité d'implication cibler les règles les plus étonnantes.
- *Développement à droite* : rechercher, à partir d'une conjonction d'attributs X de (sommet du graphe sélectionné par l'utilisateur), toutes les conjonctions d'attributs plus spécifiques Y telles que Y soit fréquentes, $|Y|=|X|+1$, $X \subset Y$. Ceci permet de découvrir des règles plus spécifiques à partir de X.
- *Développement à gauche* : rechercher, à partir d'une conjonction d'attributs Y, toutes les conjonctions d'attributs plus générales X telles que X soit fréquentes, $|Y|=|X|+1$, $X \subset Y$. Ceci permet de découvrir des règles plus générales conduisant à Y.

Ces opérateurs déclenchés par les interactions de l'utilisateur sur le graphe entraînent sa mise à jour par apparition ou disparition de sommets et d'arcs (des règles). Ainsi, la représentation évolue-t-elle dynamiquement, et de manière incrémentale : au fil du temps l'utilisateur se construit un ensemble de plus en plus complet de règles intéressantes.

5 IMPLEMENTATION

Pour valider notre démarche, nous avons développé un prototype -*Felix*- basé sur une architecture client-serveur universelle s'appuyant sur les technologies du web. Il intègre trois composantes :

- Un système de gestion de bases de données relationnel (SGBDR) qu'il utilise non seulement pour puiser les données à explorer, mais aussi pour stocker les descriptions découvertes.
- Un gestionnaire de découverte de règles d'association, comme une application intermédiaire (middleware) elle-même connectée à un serveur WWW et pilotée par les interactions de l'utilisateur.
- Une interface graphique utilisateur (GUI) qui est un client de cette architecture. Elle comprend notamment une interface interactive de visualisation de graphe dynamique.

Nous avons employé des technologies ouvertes dont l'avantage est d'assurer la portabilité du logiciel en garantissant son déploiement sur différentes architectures matérielles; et surtout d'en permettre une plus grande évolutivité.

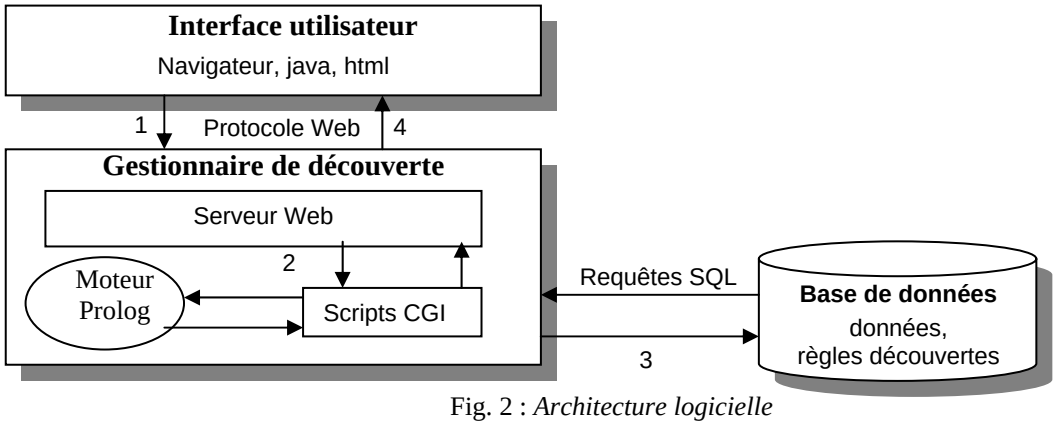

Fig. 2 : *Architecture logicielle*

Une interaction de l'utilisateur se traduit de la manière suivante : (1) l'action déclenche un événement graphique Java qui provoque l'émission d'une requête vers le gestionnaire de découverte. (2) Cette requête déclenche l'exécution d'un algorithme de découverte locale de règles d'associations. (3) L'algorithme communique avec la base de données afin d'obtenir les données nécessaires puis de stocker ses résultats (règles découvertes). (4) La visualisation est mise à jour afin d'intégrer les règles découvertes supplémentaires.

BIBLIOGRAPHIE

[1] Agrawal R., Imielinski T. and Swani A. (1993). Mining association rules between sets of items in large databases, In Proc. of ACM SIG-MOD *Conf. on Management of Data*, pages 207-216, Washington DC.

[2] Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.J. (1996). Fast discovery of association rules, *Advances in knowledge discovery and data mining*, Eds. Fayyad U.M., Piatetsky-Shapiro G. et Smyth P., pages 307-328. AAAI Press.

[3] Bhandari I. (1994). Attribute focusing: machine-assisted knowledge discovery applied to software production process control, *Knowledge Acquisition*, volume 6, pages 271-294.

[4] Brachman J.R. and Anand T. (1996). The process of knowledge discovery in databases: a human-centered approach, In *Advances in Knowledge Discovery and Data Mining*, Fayyad U.M. et al. eds, pages 37-58, AAAI Press.

[5] Eades P., Lai W., Misue K. and Sugiyama K. (1991). Preserving the mental map of a diagram. In *Proceedings of Compugraphics*, pages 24-33.

[6] Fleury L., Briand H., Philippe J., and Djeraba C. (1995). Rule evaluations for knowledge discovery in databases, In Proc. of *the 6th Int. Conf. on Database and Extert System Applications*, pages 405-414, London.

[7] Frawley W., Piatetski-Shapiro G. and Matheus C. (1992). Knowledge discovery in databases : an overview. *AI Magazine*. 14(3). 57-70.

[8] Gras R. and Lahrer A. (1993). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques, Informatique et Sciences Humaines*. 120. 5-31.

[9] Gras R., Kuntz P., Couturier R. et Guillet F.. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 69-80. Hermès Science Publication.

[10] Guillaume S., Guillet F. and Philippe J. (1998). Improving the discovery of association rules with intensity of implication, In Proc. of the *2nd Eur. Symp. on Principles of Data Mining and Knowledge Discovery*, Zytkow, J.M. and Quafafou, M. eds, Lecture Notes in Artificial Intelligence, vol. 1510, pages 318-327, Springer-Verlag.

[11] Guillet F., Kuntz P. and Lehn R. (1999). A genetic algorithm for visualizing networks of association rules, In Proc. of *12th Int. Conf. on Industrial & Engineering Appl. AI & Expert Systems*, Imam, I.F. et al. Eds, Lecture Notes in Computer Science, vol. 1611, pages 145-154, Springer-Verlag.

[12] Keim D. (1996). Databases and visualization. In Proc. *Tutorial ACM SIGMOD International conference on management of data* (SIGMOD'96), Montreal, Canada.

[13] Kuntz P., Guillet F., Lehn R. and Briand H. (2000). A User-Driven Process for Mining Association Rules. In D. Zighed, J. Komorowski and J.M. Zytkow (Eds.), Proc. of *Fourth European Conference on Principles of Data Mining and Knowledge Discovery, PKDD'2000*. Lecture Notes in Artificial Intelligence, vol. 1910, pages 483-489. Springer-Verlag.

[14] Kuntz P., Lehn R. and Briand H. (2000). Dynamic rule graph drawing by genetic search. In *Proc. of the IEEE Int. Conf. On Systems, Man & Cybernetics*, IEEE press.

[15] Mingers G. (1989). An empirical comparison of selection measures for decision-tree induction, *Machine learning*, volume 3, pages 319-342.

[16] Svenson O. (1983). Decision rules and information processing in decision making. In *Human decision making*. Bodafors: Bodaforlaget Doxa.