DKQ2005

Atelier Qualité des Données et des Connaissances

Organisé par Laure Berti-Équille et Fabrice Guillet

Associé à ECG 2005
18 Janvier 2005, Paris, France

Les problèmes de qualité des données stockées dans les bases ou les entrepôts de données s'étendent à tous les domaines gouvernemental, commercial, industriel et scientifique. La découverte de connaissances et la prise de décision à partir de données de qualité médiocre (c'est-à-dire contenant des erreurs, des doublons, des incohérences, des valeurs manquantes, etc.) ont des conséquences directes et significatives pour les entreprises et pour tous leurs utilisateurs. Le thème de la qualité des données et des connaissances est pour cela devenu, depuis ces dernières années, un des sujets d'intérêt émergeant à la fois dans le domaine de la recherche et dans les entreprises.

Toutes les applications dédiées à l'analyse des données (telles que la fouille de données textuelles par exemple) requièrent différentes formes de préparation des données avec de nombreuses techniques de traitement, afin que les données passées en entrée aux algorithmes de fouille se conforment à des distributions relativement « sympathiques », ne contenant pas d'incohérences, de doublons, de valeurs manquantes ou incorrectes. Seulement, entre la réalité des données disponibles et toute la machinerie permettant leur analyse, un assez vaste fossé demeure.

In fine, l'évaluation des résultats issus du processus de traitement des données, est généralement effectuée par un spécialiste (expert, analyste,...). Cette tâche est souvent très lourde, et un moyen de la faciliter consiste à aider le spécialiste en lui fournissant des critères de décision sous la forme de mesures de qualité ou d'intérêt des résultats. Ces mesures de qualité des connaissances doivent être conçues afin de combiner deux dimensions : une dimension objective liées à la qualité des données, et une dimension subjective liées aux intérêts du spécialiste. L'atelier Qualité des Données et des Connaissances - DKQ 2005 (Data and Knowledge Quality) - associé à EGC'2005 concerne les méthodes, les techniques d'analyse et de nettoyage, les méthodologies, les approches algorithmiques et les métriques de qualité des données et des connaissances permettant de comprendre, d'explorer les données, de détecter et corriger les problèmes de qualité des données et de qualité des connaissances extraites à partir des données.

Nous remercions tout particulièrement les auteurs et les membres du comité de relecture pour leur contribution au succès de l'atelier DKQ2005.

Programme de la journée du 18 janvier 2005

9h-9h30 : Accueil et ouverture par Fabrice Guillet et Laure Berti-Equille

9h30-10h30: Session 1 - Qualité des données dans les BD

10h45- 12h15 : Session 2 - Qualité des connaissances extraites sous forme de règles d'association

14 h-15h : Session 3 – Qualité et classification

15h- 16h : Session 4 - Plateformes d'évaluation de la qualité des connaissances

16h15-17h15 : Session 5 - Pratiques Opérationnelles

Organisation de l'atelier

Laure Berti-Équille, IRISA-CNRS Rennes, France

Comité de Programme Provisoire

Fabrice Guillet, IRIN, Université de Nantes, France (Président)

Ansaf Salleb, IRISA-CNRS Rennes, France

Jérôme Azé, LRI, Université de Paris-Sud, France

Mokrane Bouzeghoub, PRISM, Université de Versailles, France

Henri Briand, IRIN, Université de Nantes, France

Béatrice Duval, Université d'Angers, France

Johann-Christoph Freytag, Humboldt-Universität zu Berlin, Germany

Helena Galhardas, INESC, Lisboa, Portugal

Régis Gras, IRIN, Université de Nantes, France

Yves Kodratoff, LRI, Université de Paris-Sud, France

Pascale Kuntz, IRIN, Université de Nantes, France

Stéphane Lallich, ERIC, Université de Lyon 2, France

Ludovic Lebart, ENST-CNRS, Paris, France

Philippe Lenca, ENSTbr, Brest, France

Amedeo Napoli, LORIA, Nancy, France

Gilbert Ritschard, Université de Genève, Switzerland

Monica Scannapieco, Universita' di Roma "La Sapienza", Italy

Dan A. Simovici, University of Massachussets, Boston, U.S.

Einoshin Suzuki, Yokohama National University, Japan

Djamel Zighed, ERIC, Université de Lyon 2, France

Data Freshness Evaluation in Different Application Scenarios

Verónika Peralta, Mokrane Bouzeghoub

Laboratoire PRISM, Université de Versailles
45, avenue des Etats-Unis
78035, Versailles cedex, FRANCE
{Veronika.Peralta, Mokrane.Bouzeghoub} @prism.uvsq.fr

Abstract. Data freshness has been identified as one of the most important data quality attributes in information systems. This importance increases specially in the context of systems that integrates data of a large set of autonomous data sources. In this paper we describe a quality evaluation framework that allows evaluating the freshness of the data delivered to the user in a data integration system. Concretely, we show the practical use of the framework in different application scenarios and we discuss possible improvement actions for the data integration system in order to fulfill user freshness expectations. In order to illustrate the approach, we discuss data freshness evaluation issues with several examples.

1 Introduction

Data freshness has been identified as one of the most important attributes of data quality for data consumers (Shin 2003)(Wang et al. 1996). Specifically, the increasing need to access to information that is available in several data sources introduces the problem of choosing between alternative data providers and of combining data having different freshness values (Naumann et al. 1999). This paper deals with data freshness evaluation in the context of a Data Integration System (DIS) that integrates data from different independent data sources and provides the users with a uniform access to the data.

Data freshness represents a family of quality factors. With regard to data freshness, two factors have been proposed in the literature: *currency* that describes how *stale* is data with respect to the sources and *timeliness* that describes how *old* is data. In (Bouzeghoub et al. 2004) we analyze these factors and several metrics proposed to measure them.

In (Peralta et al. 2004), we proposed a framework for analyzing and evaluating data freshness based on a calculation dag which abstracts a workflow of integration activities. After a brief recall of this framework, this paper shows how the framework can practically be used in different application scenarios and how the data integration system can be improved in order to fulfill user expectations in terms of data freshness.

The rest of the document is organized as follows: Section 2 briefly describes the data quality evaluation framework. Section 3 discusses how to use this framework through different application scenarios. Section 4 focuses on the possible improvement actions to put on the DIS workflow to achieve user expectations. Finally, section 5 concludes with our general remarks.

2 The Data Quality Evaluation Framework

In this section we briefly describe the framework for data quality evaluation. The framework models the DIS processes and properties and evaluates the quality (particularly the freshness) of the data returned to the user.

The framework consists of: (i) a set of available data sources, (ii) a set of classes of user queries, (iii) the DIS integration processes, (iv) a set of properties describing DIS features (costs, delays, policies, strategies, constraints, etc.) and quality measures, and (v) a set of quality evaluation algorithms.

The DIS is modeled as a workflow in which the activities perform the different tasks that extract, transform and convey data to end-users. Each workflow activity takes data from sources or other activities and produces result data that can be used as input for other activities. The data produced by an activity can be materialized or not. Then, data traverses a path from sources to users where it is transformed and processed according to the system logics.

The framework represents the DIS dataflow by means of a *labeled calculation dag (LCDag)* that describes the involved activities, their inputs, outputs and properties. Formally, a LCDag is a dag G = <V, E, P, *propvalue*> defined as follows: The nodes in V are of three types: *source nodes* (with no input edges), *target nodes* (with no output edges) and *activity nodes* (with both input and output edges), which represent sources, user queries and DIS activities respectively. The edges in E represent that a node is calculated from another (data flows in the sense of the arrow). P is a set of properties describing DIS features and quality measures, and *propvalue* is a partial labeling function that assigns a property value to a node or edge of the dag. Figure 1 shows the LCDag graphical representation (nodes and edges are labeled with *property=value* pairs). The three LCDags of figure 1 are discussed in section 3.2.

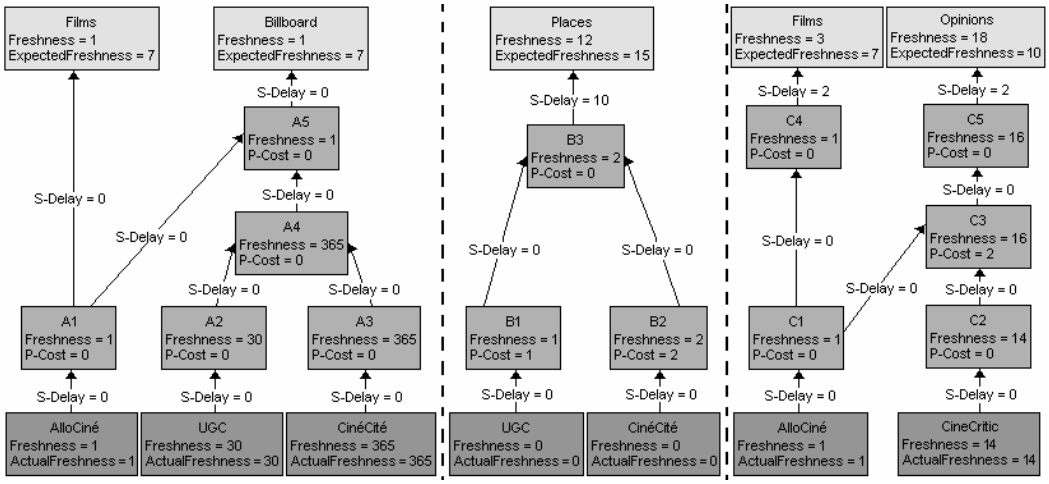


FIG. 1 – *Labeled calculation dags (source, activity and query nodes are placed at the bottom, middle and top of the dags, respectively)*

The quality evaluation is performed by evaluation algorithms. The input information needed by the evaluation algorithms is contained in the LCDag. It consists of property values, specifically, source quality values (labels of source nodes) and DIS property values (labels of activity nodes and edges). The algorithms take as input a LCDag, combine property values calculating freshness values corresponding to the data returned by queries, and return the LCDag with an additional property (corresponding to the data freshness quality factor).

3 Data Freshness Evaluation

In this section we describe the general evaluation approach. We firstly give an intuitive idea of the freshness calculation strategy and we describe a base evaluation algorithm. Then, we discuss the instantiation of the base algorithm to different application scenarios.

3.1 General Approach

The freshness of the data delivered to the user depends on *source data freshness* (the freshness of source data at extraction time) but also on the *execution delay* of the DIS processes (the amount of time from data extraction to data delivery). The execution delay is influenced not only by the processing cost of each activity but also by the delays that can exist between the executions of consecutive activities.

We briefly describe such properties as well as user expectations:

- *Processing cost*: It is the amount of time, in the worst case, that an activity needs for reading input data, executing and building result data.
- *Synchronization delay*: It is the amount of time passed between the executions of two consecutive activities.
- *Actual freshness*: It is a measure of the freshness of data in a source, which can be provided by the source or can be estimated or bounded by the DIS.
- *Expected Freshness*: It is the desired data freshness specified by the user. It measures the extent to which the freshness of the data is appropriate for the task on hand.

Our base algorithm takes into account such properties. It traverses the dag, from sources to queries (the sense of the data flow), calculating the freshness of the data produced by each node. The algorithm idea can be sketched as follows:

→ For a source node A:

Freshness(A) = getActualFreshness(A)

→ For a non-source node A, and the set of all its predecessors P:

Freshness(A) = combine {Freshness(B) +getSyncDelay(B,A) /B ∈ P} + getProcCost(A)

For source nodes, data freshness is the source actual freshness. For the other nodes, the freshness of the data produced by a node is calculated as the freshness of data at the moment of reading it (the freshness of data produced by the predecessor) plus the synchronization delay plus the processing cost. When a node reads data from several input nodes, input freshness values should be combined, for example, taking the maximum value.

We have implemented a data freshness auditing tool that implements the framework and allows evaluating the freshness of the data returned to the user in different application scenarios (Fajardo et al. 2004). Next section illustrates the approach with several examples.

3.2 Illustrating Examples

Consider three simple DIS that deal with information about cinemas and films:

- **DIS_1**: A mediation system that answers queries about films and the cinemas where they are in billboard. Typical queries are "Where can I see a film?" or "Which films are in billboard now?"
- **DIS_2**: A web portal that caches information about cinemas and the availability of places for their performances. Typical queries are "Where are available places to see a film?" or "How many places are available in a cinema?"
- **DIS_3**: A data warehousing system that stores statistic information about films, the number of persons that watch each film and their opinions. Typical questions are "Which films have the best ranking this week?" or "Which film should I watch?"

Users of DIS_1 and DIS_3 are concerned with *timeliness* but users of DIS_2 evaluate *currency*.

Figure 1 shows a simplified version of the three DIS, accessing to a small number of sources. DIS_1 extracts film information from AlloCiné (via wrapper A_1) and cinema information from UGC and CinéCité (via wrappers A_2 and A_3). Activity A_4 merges the information from both cinema sites and activity A_5 joins film and cinema information. DIS_2 extracts place information from UGC and CinéCité. Activity B_3 is the cache core, that receives user requests and asks the sources when the cache needs refreshment (invoking wrappers B_1 and B_2). DIS_3 extracts film audience statistics from AlloCiné (via wrapper C_1) and spectator's opinions from CineCritic (via wrapper C_2). Activity C_3 reconciles data from both wrappers and activities C_4 and C_5 perform aggregations and calculate statistic data.

In the LCDags of figure 1, source nodes are labeled with their *actual freshness*, target nodes are labeled with *expected freshness*, activity nodes are labeled with *processing costs* (P-Cost) and edges are labeled with *synchronization delays* (S-Delay). In next section we discuss how to obtain such values. Values are expressed in days for DIS_1 and DIS_3 but in minutes for DIS_2. Note that the "zeros" represent negligible values.

The base evaluation algorithm adds the *freshness* property to all the nodes. It first calculates freshness for source nodes, as its source actual freshness (e.g. 1, 30 and 365 days for source nodes of DIS_1). Then, the algorithm traverses the dag calculating the freshness of a node adding the predecessor freshness plus the synchronization delay plus the processing cost. For example, for activity B_2 freshness is 2 minutes (0 +0 +2 =2). When a node has several predecessors, their freshness values are combined. For example, for activity C_3 freshness is the maximum between (1 +0 +2 =3) and (14 +0 +2 =16), i.e. 16 days. However, users of DIS_1 expect to know how fresh is film information, independently to when the cinema data was last updated. DIS_1 assigns priorities to the nodes and the combination function takes the freshness value of the predecessor with highest priority (e.g. the freshness of A_5 is 1 day (1 +0 +0).

3.3 Instantiating the Framework

Data freshness is evaluated based on the source actual freshness, processing cost and synchronization delay properties. But the relevance of these properties depends on the particular scenario considered. A first remark is that its magnitude should not be considered in the absolute but compared to freshness expectations. For example, users of DIS_1 may tolerate data freshness of "7 days" so, the processing costs and synchronization delays ("some minutes") are negligible; however, users of DIS_2 require "extremely fresh" data, so, the processing costs of activities could be relevant. In addition, in the scenarios where the focus is data currency, the source actual freshness is not relevant. For example, in DIS_2, it does not matter "how old is data in the sources"; the focus is in retrieving the same data that is stored in the sources.

Another aspect is how to calculate the source actual freshness, the processing delay and the synchronization delay. Depending on the scenario, different DIS properties may influence their calculation. For example, in DIS_2 the processing cost of the wrappers is dominated by the cost of communicating with the sources. In DIS_3 and DIS_2 the materialization/caching of data introduces important synchronization delays, so the refreshment policies and frequencies are important properties to take into account. In virtual systems as DIS_1, these properties have no sense.

We propose a method for instantiating the base algorithm for adapting to the specific properties of a given scenario. The mechanism consists in overloading the following abstract functions: *getProcCost*, *getSyncDelay* and *getActualFreshness*, which calculate the respective properties according to the specific scenario. For example, for DIS_1 the processing costs and synchronization delays are negligible so the respective functions can return zero. However, source actual freshness is relevant; the *getActualFreshness* function can estimate the values from source update frequencies, for example. For DIS_3 the relevant processing costs are due to the reconciliation process (activity C_3), which may require human interaction (to solve conflicts or errors) and can last two days. Costs can be estimated using cost models or statistics or can be filled in by experts. DIS_3 materializes the data produced by C_4 and C_5; the synchronization delays with target nodes can be bounded by the refreshment frequencies. The *combine* function can also be overloaded. For example, DIS_3 takes the maximum of input nodes freshness while DIS_1 combines input values considering node priorities, as illustrated in previous section.

4 Data Freshness Enforcement

The system should provide at the query level the data freshness expected by the users. To know if user freshness expectations can be achieved by the DIS, we can calculate the freshness values for target nodes and compare them with those expected by users. If freshness expectations are not achieved, we have to improve the system design to enforce freshness or negotiate with source data providers or users to relax constraints. In this section we discuss these ideas.

4.1 Improving DIS design

Observe that for each node, it can exist a path from a source for which we add all synchronization delays and processing costs to the source actual freshness and we obtain the

freshness of the node. For example, the freshness of activity C_5 can be calculated adding source actual freshness, processing costs and synchronizations delays in the path [CineCritic,C_2,C_3,C_5]. This path is called the *critical path* and represents the bottleneck for the freshness calculation. The existence of the critical path depends on the definition of the *combine* function; taking the maximum of predecessors' freshness, the critical path always exists.

The freshness of the data delivered to the user may be improved optimizing the design and implementation of the activities in order to reduce their processing cost, or synchronizing the activities in order to reduce the delay between them. Sometimes the changes can be concentrated in the critical path that slows the system. Other times a complete reengineering of the whole system is necessary, either changing the algorithms that implement the activities, the synchronization policies, the decisions of which data to materialize or even the hardware. The synchronization of some activities implies finding the most appropriate execution frequencies for some activities respecting possible source access constraints. The main difficulty resides in the synchronization of activities having several inputs with different refreshment policies.

The auditing tool allows identifying the critical path, changing property values in order to test alternative configurations and re-executing the evaluation algorithms to see the effects of the changes. In this sense, the tool brings an aggregate value to the auditing functionalities.

4.2 Selection between alternative implementations: bottom-up propagation

If the design of the DIS cannot be improved, an alternative is negotiating with users to relax freshness expectations. The freshness values calculated using the evaluation algorithm can help users to know the freshness that the DIS can guarantee for the returned data.

Observe that the evaluation algorithm propagates freshness values from sources to queries, i.e. a bottom-up propagation (following the dataflow of the DIS).

A direct application of this bottom-up strategy is the selection between alternative implementations of the system. The DIS can offer the users several alternative processes to answer their queries and users can choose (or the DIS can choose for them) the process with the best quality. For example, even improving activities design and synchronization, the freshness expectations of the *Opinions* query cannot be achieved because of the actual freshness of the *CineCritic* source. Considering an alternative process that queries other sources can be a solution.

In this line, we have used the freshness evaluation tool within a system that automatically generates mediation queries. The tool was used for evaluating the quality of the generated queries, both in virtual and materialized scenarios, in order to select the best one for answering a given user query (Kostadinov et al. 2004).

4.3 Selection of alternative data sources: top-down propagation

Another alternative to enforce freshness is negotiating with source data providers to relax source constraints. Sometimes the system hardware can be powered to support more frequent accesses to the sources. Other times, this alternative implies demanding and eventually paying for a better service, for example, receiving data with a lower actual freshness.

Analogously to the bottom-up propagation, we can propagate freshness expectations from queries to sources (subtracting processing costs and synchronization delays). The top-down propagation algorithm is similar to the bottom-up one, but the *combine* function must consider nodes with several successors. The propagated freshness expectations can help the DIS designer to know the freshness that he must ask the source provider for.

A direct application of this top-down strategy is the selection between alternative data sources to achieve freshness expectations. For example, propagating down freshness expectations for the *Opinions* query we obtain a bound (6 days) for the actual freshness of the source providing user's opinions. This avoids considering sources as *CineCritic* that have greater actual values.

5 Conclusion

In this paper we addressed the problem of evaluating data freshness in a data integration system. We presented a quality evaluation framework and its practical use for evaluating data freshness in different application scenarios. The framework was implemented in a quality auditing tool that can be instantiated for evaluating data freshness in a concrete scenario. The tool supports the top-down and bottom-up propagation strategies in order to help the user to improve freshness.

We are now working in the development of a toolkit for implementing the instantiation in a semi-automatic way. In the future, our goal is to confront the results with user quality profiles.

References

Bouzeghoub M., Peralta V. (2004), A Framework for Analysis of Data Freshness, in Proc. of the Int. Workshop on Information Quality in Information Systems (IQIS'2004), collocated with SIGMOD'2004, France, 2004.

Fajardo F., Crispino I., Peralta V. (2004), DWE: Una Herramienta para Evaluar la Calidad de los Datos en un Sistema de Integración, in Proc. of the X Congreso Argentino de Computación (CACIC'04), Argentine, 2004.

Kostadinov D., Peralta V., Soukane A., Xue X. (2004), Système adaptatif d'aide à la génération de requêtes de médiation, short paper and demonstration, in Proc. of the 20èmes Journées de Bases de Données Avancées (BDA'2004), France, 2004.

Naumann F., Leser U. (1999), Quality-driven Integration of Heterogeneous Information Systems, Proc. of the 25th Int. Conf. on Very Large Databases (VLDB'99), Scotland, 1999.

Peralta V., Ruggia, R.; Kedad, Z.; Bouzeghoub M. (2004), A Framework for Data Quality Evaluation in a Data Integration System, Proc. of the 19º Simposio Brasileiro de Banco de Dados (SBBD'2004), Brazil, 2004.

Shin B. (2003), An exploratory Investigation of System Success Factors in Data Warehousing, Journal of the Association for Information Systems, Vol. 4(2003): 141-170, 2003.

Wang R., Strong D. (1996), Beyond accuracy: What data quality means to data consumers, Journal on Management of Information Systems, Vol. 12 (4):5-34, 1996.

Nettoyage des données XML : combien ça coûte ?

Laure Berti-Équille

Université de Rennes 1

Résumé

L'objectif de cet article est de présenter un travail en cours qui consiste à proposer, implanter et valider expérimentalement un modèle pour estimer le coût d'un processus de nettoyage de documents XML. Notre approche de calcul de coût est basée sur une méthode par calibration selon une analyse probabiliste. Pour cela, nous proposons de calculer des probabilités de pollution et au préalable de détection des différents types de pollutions. Pour valider notre modèle, nous avons choisi de polluer artificiellement une collection de données XML avec l'ensemble des types d'erreurs possibles (erreurs typographiques, ajout de doublons, de valeurs manquantes, tronquées, censurées, etc.) et d'estimer, grâce au modèle proposé, le nombre et le coût des opérations nécessaires au nettoyage des données afin de proposer des stratégies de réparation ciblées et économes. Les expérimentations en cours ne sont pas rapportées dans cet article.

1. Introduction

Le nettoyage automatique des données se décompose classiquement en trois étapes : 1) examiner les données afin de détecter les incohérences, les données manquantes, les erreurs, les doublons, etc. 2) choisir les transformations pour résoudre les problèmes, 3) et enfin, appliquer les transformations choisies au jeu de données. La plupart des outils utilisés pour le nettoyage des données par *Extraction-Transformation-Loading* (*ETL*) permettent l'extraction d'expressions régulières et structures (*patterns*) à partir des données, ainsi que leur transformation et formatage par l'application de différentes fonctions (sélection, fusion, *clustering*, etc.) dont généralement, on ignore *a priori* le coût.

Bien qu'il existe de nombreux travaux [2, 6, 12, 10], outils et prototypes (Telcordia [1], AJAX [4], Potter's Wheel [7], Arktos [9], IntelliClean [5], Tailor [3]) développés pour « nettoyer » les données relationnelles, très peu de travaux à l'exception des récents travaux de Weiss et Naumann [11], ont jusqu'ici été menés pour le nettoyage de données XML et, à notre connaissance, aucun n'a abordé l'estimation du coût d'un nettoyage de données *a fortiori* pour des données XML. C'est dans ce cadre qu'a débuté notre travail dont l'objectif est de proposer, d'implanter et valider expérimentalement un modèle de coût global permettant d'estimer combien peut coûter un processus de nettoyage sur un document XML artificiellement pollué pour les besoins de nos expériences.

La suite de l'article s'organise de la façon suivante : la section 2 propose notre démarche illustrée par un exemple simple qui énumère les différents types de pollution possibles dans un document XML. La section 3 présente plus formellement notre modèle de coût avec ses définitions préliminaires et ses paramètres. Enfin, la section 4 conclut l'article et présente brièvement nos perspectives de travail.

2. Problématique

2.1. Un typologie des problèmes de non-qualité sur des données XML

Nous avons dressé, dans un premier temps, une typologie des erreurs pouvant être introduites « en extension » dans un document XML tout en préservant sa validité « en intension » par rapport à un schéma XSD ou une DTD. Le tableau (1a) présente une synthèse des différents types de pollution pouvant être introduites au niveau du contenu d'un nœud élément ou attribut. En particulier, selon le type de contenu : textuel (#PCDATA), numérique ou composite, les pollutions sur le contenu sont des erreurs typographiques (par ajout, suppression, inversion, remplacement de (chaîne de) caractères), des ajouts de doublons, des suppressions de valeurs (ou valeurs manquantes), des valeurs tronquées ou censurées.

Type De contenu d'un noeud	**Type de pollutions introduites sur la valeur d'un noeud élément ou attribut**	
Texte	Ajout/suppression/inversion/remplacement de (chaînes de) caractères Echange dans une liste énumérée	Valeur manquante Valeur tronquée Valeur censurée Valeur dupliquée
Numérique	Echange de valeur en conformité avec les bornes l'intervalle (ex. note entre 0 et 20) Valeur arrondie	
Composite	Remplacement de valeur conforme ou non à des règles syntaxiques (ex. n° de SS ou de téléphone) ou à des contraintes « métier »	

Tableau 1a. Différents types de pollution pouvant être introduites au niveau du contenu

Le tableau (1b) présente les types de pollutions (essentiellement par échange de contenus) pouvant être introduits au niveau d'un (ou plusieurs) nœud(s) attribut ou nœud(s) élément.

Type de noeud	**Type de pollution impliquant plusieurs nœuds éléments ou attributs**
Attribut	Echange de contenu entre deux attributs de même nom appartenant à deux éléments différents Echange de contenu entre deux attributs de nom différent au sein du même élément Echange de contenu entre deux attributs de nom différent appartenant à deux éléments différents Violation de dépendances fonctionnelles ou de contraintes définies sur le contenu d'un attribut ou entre plusieurs attributs
Elément	Echange de contenu entre deux éléments de même nom à la même profondeur Echange de contenu entre deux éléments de nom différent à la même profondeur Echange de contenu entre deux éléments de même nom (ou de nom différent) à une profondeur différente

Tableau 1b. Différents types de pollution pouvant être introduites au niveau des nœuds XML

Pour cette étude, nous ne considérons pas la notion d'hyper-document et les fichiers manipulés sont pour ainsi dire « plats », c'est-à-dire sans lien hypertexte. Toutefois, à terme nous envisageons d'étendre le modèle de coût afin de prendre en compte de suivi de chemins pour le nettoyage d'une collection d'hyper-documents.

2.2. Un bref exemple illustratif

La Figure 1 illustre sur un exemple simple de document XML (clean.xml) quelques-uns des types d'erreurs pouvant être introduites artificiellement (dirty.xml) pour valider notre modèle.

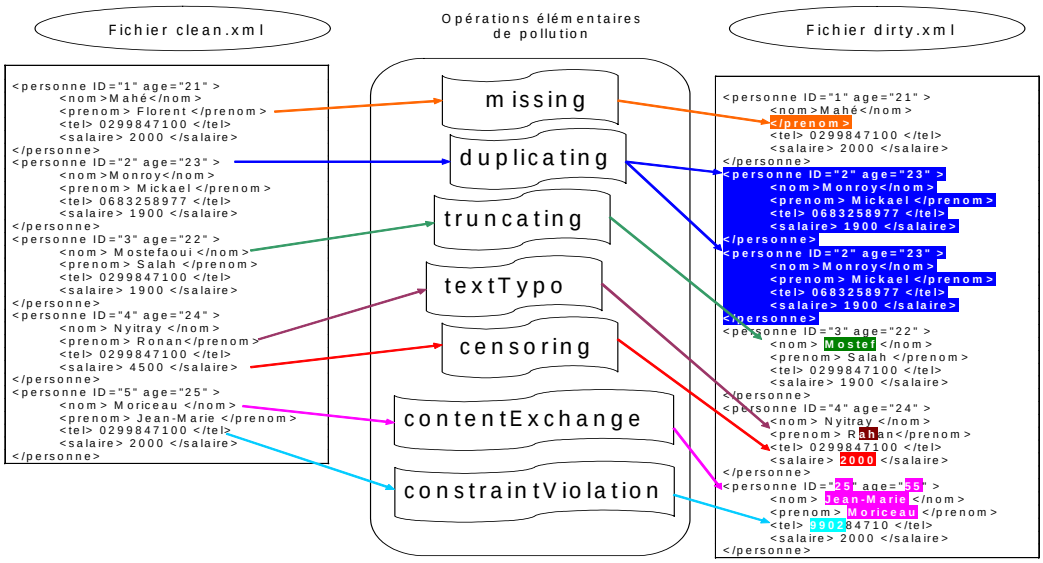


Figure 1. Exemple de pollution d'un document XML

Tout document XML est un arbre qui peut être ainsi, dès sa création, pollué par certains types de pollution qui sont (ou non) par la suite détectés. Le document XML peut être éventuellement nettoyé comme l'illustre le diagramme d'état-transition UML de la figure 2. Par la maîtrise de la pollution artificielle d'un document XML, notre approche est d'évaluer le coût probable d'un scénario de nettoyage.

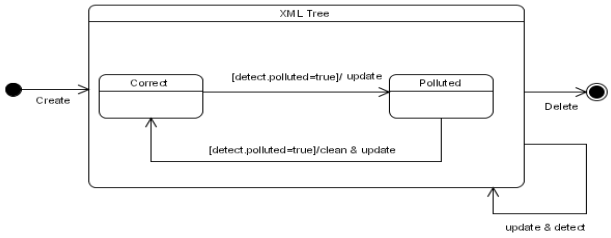


Figure 2. Diagramme d'état-transition d'un document XML au cours de son cycle de vie

2.3. Démarche

A partir d'une hypothèse assez intuitive qui est que plus la durée de vie d'un document XML est longue et ses mises à jour nombreuses, plus le risque d'introduire des erreurs à chaque mise à jour est grand, nous avons défini un taux de pollution qui prend en compte la probabilité qu'un arbre XML soit (ou non) pollué et préalablement détecté (ou non) comme tel. Ce taux de pollution, défini formellement dans la section suivante, peut être utilisé à des fins de diagnostic par un classifieur pour estimer la probabilité que le document soit correct ou bien pollué. Dans le cas d'un document détecté pollué, le modèle de coût que nous proposons permet d'estimer le coût global d'un scénario de nettoyage selon le type de pollution.

Afin d'évaluer expérimentalement notre approche, nous avons pollué artificiellement une collection de documents XML. Pour chaque nœud d'un arbre XML, nous calculons la distance de similarité existante entre ce nœud et sa version artificiellement polluée et l'identifions au taux de pollution du nœud considéré sous l'hypothèse qu'il n'y a pas d'erreur dans la classification (le nœud étant détecté pollué à juste titre). Ceci nous permet d'avoir une estimation globale du coût d'un scénario de nettoyage qui est par la suite raffiné selon le type de pollution et le coût des opérateurs de nettoyage nécessaires.

3. Modèle de coût

3.1. Définitions préliminaires

Définition 1. [Document XML]
Un document XML est défini comme un arbre T = (N, E). L'ensemble des nœuds N = NE ∪ NA ∪ NV, où NE est l'ensemble des éléments, NA l'ensemble des attributs et NV est l'ensemble des valeurs d'éléments et d'attributs de type texte ou numérique. E est l'ensemble des arcs reliant les noeuds, en particulier, l'arc (u, v) ∈ E si et seulement si v est une valeur ou un fils de u. Un élément u est un fils de l'élément v si (v,u) ∈ E. Un élément u est le parent d'un noeud v si (u,v) ∈ E. Un noeud u est un ancêtre d'un noeud v s'il existe une série d'arcs qui relient u à v.

Définition 2. [Processus de nettoyage de données XML]
Un processus de nettoyage de données XML, noté C, est un quadruplet {M, B, O, S}, tel que M est le modèle de données XML (basé sur une DTD ou un schéma XSD), B est l'unité de base d'information (c'est-à-dire le contenu d'un nœud élément ou d'un nœud attribut), O est un ensemble d'opérateurs de nettoyage qui agit sur les instances du modèles de données. Chaque opérateur accepte en entrée une à plusieurs collections d'unités de base B et produit une collection d'unités de base en sorti. S est un ensemble possible de scénarii de nettoyage, c'est-à-dire un ensemble de listes d'opérateurs à utiliser pour le nettoyage.

Définition 3. [Document XML pollué]
Soit N un ensemble fini de nœuds dont le type est donné par le modèle de données M, soit P un ensemble fini d'estampilles de pollution. Un document XML pollué est un arbre Tp = (Np,Ep), p ∈ P, où (i) l'ensemble des noeuds Np ⊆ N ; (ii) l'ensemble des arcs Ep ⊆ Np × Np × Np défini un arbre racine valide par rapport à M (c'est-à-dire satisfaisant le schéma imposé par le modèle de données entre les différents types de nœuds), ayant le quadruplet (parent, left, flag(n,p), n) qui spécifie que le nœud n a le nœud ''parent'' pour parent et le nœud ''left'' pour nœud frère à gauche. Le prédicat flag*(n,p) est vrai si le noeud n contient directement ou indirectement une (ou plusieurs) fonction(s) de pollution notée p.*

Définition 4. [Fonction de pollution]
Une fonction de pollution notée p(T, nodeNumber,[minHeight, maxHeight, N, parameters]) a pour paramètres l'arbre T, un nombre de nœuds à polluer (qui peut être un entier ou exprimé en pourcentage), et de façon optionnelle les profondeurs minimale ou maximale localisant la région où est appliquée la pollution dans l'arbre T ou encore l'ensemble des noms des nœuds ciblés (étiquettes) et un ensemble de paramètres plus spécifiques selon le type de pollution :

Définition 5. [Scénario de nettoyage] *Un scénario de nettoyage Sp consiste à appliquer un ensemble d'opérations à un ensemble de nœud pollués selon un type de pollution p afin d'obtenir un ensemble de nœuds majoritairement non pollués.*

Notre modèle de coût se base sur le calcul d'une distance de similarité entre les nœuds du document XML sain et ceux de son correspondant artificiellement pollué. Cette distance est basée sur la distance *q-gram* entre les chaînes de caractères et elle est définie comme la norme L_1 de $G_q(x)[v] - G_q(y)[v]$ telle $D_q(x,y) = \sum_{v \in \sum^q} |G_q(x)[v] - G_q(y)[v]|$, avec le *q-gram* $v = a_1 a_2 \ldots a_q \in \sum^q$ étant une chaîne de caractères dans l'alphabet fini $\sum$ de longueur *q*

et *x, y* deux chaînes de caractères quelconques dans l'alphabet $\sum$.

$G_q(x)[v]$ (resp. $G_q(y)[v]$) représente le nombre d'occurrences du *q-gram v* dans *x (*resp. *y).* La distance *q-gram* consiste donc à mesurer le nombre de caractères non communs entre les deux chaînes de caractères en prenant une fenêtre d'observation de longueur *q.*

Par exemple, prenons *q=2*, et *x = "clean"* et *y="clue"*, les 2-grams de *x* et *y* sont les suivants : *(cl,le,ea,an)* et *(cl,lu,ue),* on obtient pour *x,* $G_2(x)[cl]=1$, $G_2(x)[le]=1$, $G_2(x)[ea]=1$, $G_2(x)[an]=1$ et $G_2(x)[v]=0$ pour les autres *2-grams* de *x*. En listant les *2-grams* commençant par *cl,le,ea,an,lu,ue* dans cet ordre. On obtient un profil pour *x* tel que (1,1,1,1,0,0,…) et pour *y* tel que (1,0,0,0,1,1,0,…) et la distance $D_2(x,y)=5$.

Définition 6 [Distance entre contenu de deux noeuds XML] *La distance entre le contenu de deux nœuds n_1 et n_2 est définie telle que :*

$$contentDistance(n_1,n_2)=\begin{cases}\min\left(\dfrac{infoSize(n_1)+infoSize(n_2)}{2}, qdist(text(n_1),text(n_2))\right) & \text{pour } n_1 \text{ et } n_2 \text{ des nœuds de type texte}\\ sametag(n_1,n_2)+\sum_{C}\min\left(qdist(val(n_1,a),val(n_2,a)),attrInfo(a)\right)+c_aD & \text{pour } n_1 \text{ et } n_2 \text{ des nœuds de type élément}\\ 1 \quad \text{pour les autres cas} & \end{cases}$$

Les fonctions *infoSize* et *attrInfo* renvoient respectivement la longueur arrondie du contenu et des valeurs d'attribut d'un nœud et sont définies de la façon suivante :

$$info\ Size\ (n)=\begin{cases}\max(|text\ (n)|-c_t,1), avec\ \ n, \text{un noeud de type texte}\\ c_e+\sum_{attr\ (n,i)}\left(c_a+\max(|val\ (n,a)|-c_v,1)\right), avec\ \ n, \text{un noeud de type élément}\end{cases}$$

$$attrInfo\ (a)=\min(|val\ (n_1,a)|-c_v,1)+\min(|val\ (n_2,a)|-c_v,1)$$

c_t et c_v sont des constantes seuils permettant de diminuer la contribution de la similarité des textes courts et des valeurs d'attributs par rapport aux autres. Par exemple, une valeur d'attribut de plus courte longueur que c_v sera traitée comme ayant une longueur de 1. c_a, et c_e sont respectivement le contenu d'information correspondant au nom de l'attribut, ou au nom de l'élément. *C* est l'ensemble de tous les attributs de n_1 et n_2. *D* est le nombre d'attributs de n_1 et n_2 non présents à la fois dans les deux nœuds. *sametag(.,.)* est une fonction qui retourne c_e dans le cas où les noms d'éléments sont égaux et 0 sinon. *val(n,a)* retourne la valeur de l'attribut *a* du nœud *n*, et *text(n)* retourne le contenu textuel du nœud *n*.

3.2. Paramètres du modèle de coût pour l'estimation globale

Suite à ces définitions, nous avons établi plusieurs paramètres permettant de modéliser, dans un premier temps, le coût global d'un nettoyage.

Soient $P_p(n)$ représentant la probabilité qu'un nœud soit détecté pollué, $P_c(n)$ représentant la probabilité qu'un nœud soit détecté correct, $P_{cdp}(n)$ représentant la probabilité qu'un nœud soit détecté pollué alors qu'il est correct, $P_{pdc}(n)$ représentant la probabilité qu'un nœud soit détecté correct alors qu'il est pollué, on a l'égalité suivante illustrée par la figure 3 :

$$P_c(n) \cap P_p(n) = (P_{cdp}(n) \cap P_p(n)) \cup (P_{pdc}(n) \cap P_c(n))$$

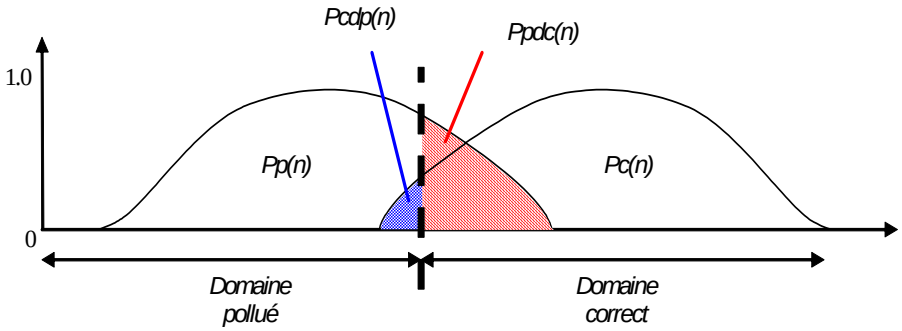


Figure 3. Partition des probabilités de détection d'un noeud XML correct ou pollué

La figure 3 représente la nature du problème dans le cas univarié. Idéalement, les aires associées au nœud *n* seront choisies telles que les distributions qu'elles représentent se ne recouvrent pas afin de ne pas être en présence d'une mauvaise classification. Mais dans le cas général comme le représente la figure 3, les distributions se chevauchent et le mécanisme de détection d'erreurs identifie les nœuds corrects comme des nœuds pollués et vice versa. On peut formuler alors le taux de pollution de type *p* pour le nœud *n* tel que :

$$Pollution\ (p,n) = E\left[\frac{\min(Cost(U), Cost(M))}{Cost(M)}\right] P_{pdc} \int_{Nclean} P_p(n)dn \quad (1a)$$

$$+ E\left[\frac{\min(Cost(U) + Cost(S_p), Cost(M))}{Cost(M)}\right] P_{pdc}\left[1 - \int_{Nclean} P_p(n)dn\right] \quad (1b)$$

$$+ E\left[\frac{\min(Cost(U), Cost(M))}{Cost(M)}\right] P_{cdp}\left[1 - \int_{Np} P_c(n)dn\right] \quad (1c)$$

$$+ E\left[\frac{\min(Cost(U) + Cost(D), Cost(M))}{Cost(M)}\right] P_{cdp} \int_{Np} P_c(n)dn \quad (1d)$$

Le terme (1a) définit les nœuds pollués qui sont malencontreusement détectés corrects et nous nous intéressons à la probabilité qu'un nœud soit classifié pollué (*Pp*) lorsque les nœuds ont toutes les caractéristiques de noeuds corrects (*Nclean*), d'où l'intégration sur cette surface. Le terme est pondéré par la probabilité qu'un nœud est détecté correct alors qu'il est en fait pollué (*Ppdc*) et par le temps de son traitement (représentant par un écart-type sur les temps de mise à jour et de maintenance). Généralement les temps de mise à jour (*U*), de détection (*D*) et de nettoyage de la pollution *p* (*Sp*) sont plus courts que la durée de vie totale du fichier et donc que le temps de sa maintenance (*M*). Le terme (1b) prend en compte les nœuds restant classifiés pollués, pondéré par la probabilité que le nœud soit détecté correct alors qu'il est pollué et par le temps de traitement (écart-type sur les temps de mise à jour, de nettoyage et de maintenance). Les termes (1c) et (1d) sont similaires. Le terme (1d) définit les nœuds corrects qui sont malencontreusement détectés pollués, pondéré par la probabilité qu'un nœud est détecté pollué alors qu'il est en fait correct (*Pcdp*) et par le temps de son traitement (écart-type sur les temps de mise à jour et de maintenance). Le terme (1d) prend en compte les nœuds restants classifiés corrects pondéré par la probabilité que le nœud soit détecté pollué alors qu'il est correct et par le temps de traitement incluant la mise à jour, la détection et la maintenance.

Le taux de pollution peut être réécrit de la façon suivante :

$$Pollution\ (p,n) = A.P_{pdc}\int_{Nclean} P_p(n)dn + B.P_{cdp}\int_{Np} P_c(n)dn + E\left[\frac{\min(Cost(U)+Cost(S_p),Cost(M))}{Cost(M)}\right]P_{pdc} + E\left[\frac{\min(Cost(U),Cost(M))}{Cost(M)}\right]P_{cdp}$$

$$A = E\left[\frac{\min(Cost(U),Cost(M))}{Cost(M)}\right] - E\left[\frac{\min(Cost(U)+Cost(S_p),Cost(M))}{Cost(M)}\right]$$

avec $$B = E\left[\frac{\min(Cost(U)+Cost(D),Cost(M))}{Cost(M)}\right] - E\left[\frac{\min(Cost(U),Cost(M))}{Cost(M)}\right]$$

Traditionnellement, dans la théorie de la classification statistique, les variables *A* et *B* représentent les coûts d'une mauvaise classification. Dans notre formule, ces quantités représentent des pourcentages de temps pendant lequel un nœud est classifié correct ou pollué. Ainsi, le taux de pollution peut être défini comme le pourcentage de temps d'une mauvaise classification avec des temps de traitement additionnels pour résoudre cette mauvaise classification.

Pour minimiser le taux de pollution *Pollution(p,n)*, le mécanisme de détection doit classifier le noeud *n* soit correct soit pollué selon la règle de décision suivante :

$$n\ est\ pollué\ \ par\ une\ \ pollution\ \ de\ type\ \ p\ si\ \frac{P_p(n).P_{pdc}(n)}{P_c(n).P_{cdp}(n)} \geq \frac{B}{A}$$

$$n\ est\ correct\ \ si\ \frac{P_p(n).P_{pdc}(n)}{P_c(n).P_{cdp}(n)} < \frac{B}{A}$$

Dans nos expériences, connaissant *Cost(U)*, *Cost(M)* et la distance entre un nœud *n* et sa version artificiellement polluée *np* est identifiée à la probabilité de pollution du nœud, et nous évaluons *A* et *B* et déterminons *Cost(Sp)* et *Cost(D)* de façon à ce que soit vérifiée l'inégalité suivante :

$$contentDistance(n,np) \geq \frac{E\left[\frac{\min(Cost(U)+Cost(D),Cost(M))}{Cost(M)}\right] - E\left[\frac{\min(Cost(U),Cost(M))}{Cost(M)}\right]}{E\left[\frac{\min(Cost(U),Cost(M))}{Cost(M)}\right] - E\left[\frac{\min(Cost(U)+Cost(S_p),Cost(M))}{Cost(M)}\right]}$$

4. Conclusion

Cet article présente succinctement un travail en cours d'expérimentation qui consiste à proposer, implanter et valider un modèle permettant d'estimer globalement le coût probable d'un processus de nettoyage de documents XML. Pour cela, nous proposons de calculer des probabilités de pollution et de détection sur les différents types d'erreurs possibles sur les données (erreurs typographiques, ajout de doublons, de valeurs manquantes, etc.) qui préservent la validité structurelle des documents XML par rapport à leur modèle de données (DTD ou XSD). Selon les résultats des expériences de pollution artificielle en cours qui permettent de corroborer notre modèle global, nous envisageons de raffiner notre approche au niveau du coût de chaque type d'opérateur de nettoyage, notamment en utilisant les historiques des scénarii de nettoyage sur des documents XML.

Références

[1] CARUSO F., COCHINWALA M., GANAPATHY U, LALK G., MISSIER P., TELCORDIA's database reconciliation and Data Quality Analysis Tool, Proc. of the Intl. Conf. VLDB, 2000.

[2] DASU T., JOHNSON T., Exploratory Data Mining and Data Cleaning, Wiley, 2003.

[3] ELFEKY M.G., VERYKIOS V.S., ELMAGARMID A.K., Tailor: A Record Linkage Toolbox, Proc. of the ICDE Conf., 2002.

[4] GALHARDAS H., FLORESCU D., SHASHA D., AND SIMON E., SAITA C., Declarative Data Cleaning: Language, model and algorithms, Proc. of the Intl. Conf. VLDB, p. 371-380, 2001.

[5] LOW W.L., LEE M.L., LING T.W., A knowledge-based approach for Duplicate Elimination in Data Cleaning, Information System, Vol. 26 (8), 2001.

[6] RAHM E., DO H., Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. 23(4): 3-13, 2000.

[7] RAMAN V., HELLERSTEIN J. M., Potter's wheel: an interactive data cleaning system, Intl. Conf. VLDB, 2001.

[8] SCANNAPIECO M., NAUMANN F. (Eds), 1st Intl. ACM SIGMOD Workshop on Information Quality in Information Systems, 2004.

[9] VASSILIADIS P., VAGENA Z., SKIADOPOULOS S., KARAYANNIDIS N., ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments. IEEE Data Eng. Bull. 23(4): 42-47, 2000.

[10] VASSILIADIS P., SIMITSIS A., GEORGANTAS P., TERROVITIS M., A Framework for the Design of ETL Scenarios. Proc. of he 15th Conf. on Advanced Information Systems Engineering (CAiSE '03), Klagenfurt, Austria, 16 - 20 June, 2003.

[11] WEIS M., NAUMANN F., Detecting Duplicate Objects in XML Documents, 1st Intl. ACM SIGMOD Workshop on Information Quality in Information Systems, 2004.

[12] WINKLER W., Data Cleaning Methods, Intl. Conf. KDD, 2003.

Extraction de Règles en Incertain par la Méthode Implicative

Régis Gras *, Raphaël Couturier[**]
Fabrice Guillet[*], Filippo Spagnolo[***]

** LINA– Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr et Fabrice.Guillet@polytech.univ-nantes.fr
** Institut Universitaire de Technologie de Belfort, BP 527, rue E. Gros, 90016 Belfort cedex : Raphael.Couturier@iut-bm.univ.fcomte.fr
*** G.R.I.M. (Gruppo di Ricerca sull'Insegnamento delle Matematiche), Department of Mathematics, University of Palermo: spagnolo@math.unipa.it.

Résumé. En relation avec des approches classiques de l'incertain de Zadeh et autres auteurs, l'analyse statistique implicative (A.S.I.) peut apparaître innovante, particulièrement pour l'opérateur d'implication. L'article montre en effet que la notion de variables à valeurs intervalles et celle de variables-intervalles sont efficaces dans la détermination de la distribution des variables et dans la recherche de règles entre variables floues. De plus, elles apportent de riches informations sur la qualité de ces règles, tout en permettant d'étudier le rôle des variables supplémentaires dans l'existence de ces règles. Cette nouvelle perspective épistémologique de l'incertain ouvre d'intéressantes perspectives d'application.
Mots-clés. Règles, hiérarchies orientées, analyse statistique implicative.

Introduction

Partant du cadre défini et formalisé par [ZADEH 79, 01], par [DUBOIS et PRADE 87], ce texte vise à étudier les proximités formelle et sémantique des cadres de l'incertain et de l'analyse statistique implicative (A.S.I.) entre variables à valeurs intervalles et variables-intervalles [GRAS 01b]. On ne rappellera pas les formalisations classiques des notions premières et de chaque opérateur de la logique floue. On s'intéressera plus particulièrement à l'opérateur **«** implication » à l'aide duquel on extrait des règles d'association.

Le texte présent s'inscrit dans le cadre des travaux initiés par R.Gras [GRAS 79] sur la méthode d'analyse de données, analyse statistique implicative (A.S.I.) qui vise à extraire et représenter, de bases de données, des règles d'association entre variables ou conjonctions de variables, du type $a \Rightarrow b$. Nous considérons celles qui croisent des sujets (ou des objets) et des variables, présentant des modalités nettes ou floues. Une règle entre deux variables ou entre conjonctions de variables est établie sur la base de la rareté statistique du nombre de ses contre-exemples, dans l'hypothèse de l'indépendance a priori des variables en jeu [GRAS 79], [LERMAN 81]. La qualité de la règle sera évidemment d'autant plus grande que ce nombre de contre-exemples sera invraisemblablement petit sous cette hypothèse, eu égard aux occurrences des variables et des instances totales.

Dans ce premier paragraphe, nous présentons la problématique. Puis, nous construisons, de façon peu classique, une distribution floue à partir de données objectives. Dans le § 3.1, nous abordons la recherche de règles d'association dans une situation « floue » en nous

appuyant auparavant sur la notion de variables modales. Enfin, dans le § 3.2, nous revenons sur la construction des règles en ramenant les variables floues à des variables-intervalles.

1 Problématique

Bien que les applications de la logique floue soient nombreuses en intelligence artificielle (par exemple en matière de diagnostic médical ou de reconnaissance des formes), plusieurs questions restent bien souvent latentes : comment obtient-on des distributions des degrés d'appartenance dans le cas de variables numériques ? Sur quelles connaissances sont-elles établies ? Sont-elles données a priori et mises à l'épreuve de la réalité ou bien sont-elles des construits ?S'il s'agit de ce dernier cas, quel processus d'extraction de connaissances à partir de données peut y conduire et quel type de règle peut-on alors extraire dans ce cadre ? Quelle signification peut-on donner à une règle associant deux sous-ensembles ou deux attributs flous ? On rejoint alors une des problématiques du data mining et de la qualité des règles.

2 Deux méthodes de construction de distributions floues par extraction de connaissances

Dans le cadre que nous retenons, les distributions des degrés d'appartenance seront le fruit de l'interaction de connaissances objectives (une vraie valeur de la variable, un attribut net ou modificateur linguistique *consensuel*) et de connaissances subjectives. Dans la littérature, les degrés sont des données. D'où proviennent-elles ?

Par ex., un échantillon d'individus étant donné on disposera **effectivement** de leur taille s (un nombre) ou des caractères ou attributs nets : « petit », « moyen » et « grand » au vu d'une décision consensuelle du type : les caractères « petit », « moyen » et « grand » seront attribués **objectivement** au regard de leur taille mesurée. Face à ces données, on pourra comparer le point de vue **subjectif** portant sur les mêmes individus qui énoncera qu'un sujet de 179 cm n'est pas petit, mais peut être considéré de taille grande ou moyenne, noncontradictoirement. Différentes méthodes pour définir la distribution des attributs visent à effectuer un processus de « fuzzification » [BERNADET, 04] : définition des classes floues pour chaque attribut, puis mise en correspondance de chaque attribut avec un degré d'appartenance à une classe floue, comme nous le voyons dans l'exemple introductif. Par ex., dans [ZADEH, 97], une méthode de discrétisation optimale est donnée. Ici, nous procèderons autrement.

2.1 Relation entre intervalles nets et attributs flous

Notre objectif, dans ce paragraphe, est de « *fuzzifier* » en quantifiant le degré d'appartenance d'un sujet à un intervalle numérique donné. Pour ce faire, la méthode de type « clustering » que nous proposons consiste tout d'abord, à partir du choix d'un indice de similarité, ici celui de la vraisemblance du lien de I.C. Lerman [LERMAN 81], d'extraire, tout d'abord, la proximité entre les attributs nets et les attributs flous.

Auparavant, selon le procédé défini dans [GRAS 01], nous choisissons de transformer l'ensemble des valeurs observées sur les sujets en sous-intervalles disjoints de variance inter-classe maximale afin de pouvoir attribuer à chaque sous-intervalles un attribut net de même désignation que celle attribuée aux attributs flous. Cette partition nette est établie par la

méthode des nuées dynamiques de [DIDAY 72]. Enfin, pour chaque classe de similarité entre attribut net et attribut flou, nous déterminons le degré d'appartenance des sujets à une classe floue à partir de la mesure normalisée de typicalité associée à chaque individu. En effet, cette typicalité, définie dans [GRAS et al. 01a], rend compte d'un degré de responsabilité dans la proximité d'attributs, soulignant l'accord entre net et flou. Ainsi, nous disposerons d'une mesure vérifiant les axiomes de Zadeh relatifs au concept de « possibilité ». Mais, son avantage par rapport à la détermination subjective classique est qu'elle est établie à l'épreuve statistique de la réalité et qu'elle varie avec la dilatation de l'ensemble des sujets.

En résumé, les données initiales sont de deux ordres :

- d'une part, des variables **objectives, consensuelles** aux valeurs numériques réparties sur des intervalles auxquels on associe respectivement un **attribut net** ,
- d'autre part, un **attribut flou** attribué **subjectivement** à chaque sujet.

Exemple : Les données portent sur 60 sujets. Leurs tailles T vraies varient de 168 et 198 cm. Appliquant l'algorithme de la variance inter-classes maximale, nous obtenons une hiérarchie entre les intervalles déterminés par l'algorithme : « Tpeti » de 168 à 174, « Tmoy » de 175 à 183, Tgran de 184 à 198. Les attributs flous sont notés respectivement TP, TM et TG.

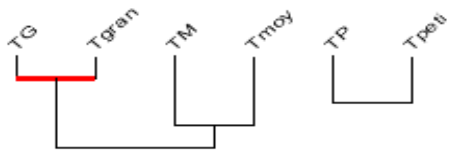


FIG. 1 – Hiérarchie des similarités entre les intervalles

On note que les attributs nets s'associent aux intervalles flous correspondants, ce que raisonnablement on pouvait attendre. Un sujet pourra donc posséder TM et TG suivant le point de vue du juge si sa taille n'est pas manifestement grande. Le logiciel CHIC [Couturier 2000], présenté en démonstration, restitue les mesures de typicalité des sujets selon les 3 classes de similarité. Rappelons qu'un individu est d'autant plus typique d'une classe qu'il a une attitude conforme à la constitution de la classe par la population de sujejets..

On observe alors en consultant les calculs donnés par CHIC que, par ex., le sujet i06 a une typicalité de 0 sur les tailles petites, 0.056 sur les tailles moyennes et 0.95 sur les tailles grandes. On peut faire de même pour les autres sujets de l 'échantillon. Ce sont ces valeurs que nous retenons comme degrés d'appartenance respectifs par rapport aux attributs flous.

Pour itérer le procédé dans le but d'affiner les distributions, il suffit, par ex, de remplacer une des modalités d'attribut ou chacune d'entre elles par 2 modalités. Ainsi, « grand » sera subdivisé en « très grand » et « assez grand », « moyen » en « pas très grand » et « pas très petit », etc. La partition de l'intervalle des tailles se fera sur la base de 6 intervalles.

« Attribuer des degrés d'appartenance à partir des contributions aux associations sur un échantillon » nous paraît atténuer l'arbitraire habituel des affectations de ces degrés.

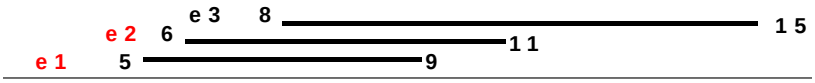
2.2 Construction de l'histogramme d'une variable-intervalle à partir des données floues des sujets

Cette fois, on dispose de la distribution des valeurs floues prises par chaque sujet d'une population sur un intervalle. On cherche à en déduire une distribution des degrés d'appartenance sur cet intervalle. L'objectif final est de définir une variable symbolique, dite

aussi variable-intervalle, qui soit l'histogramme d'un intervalle sur lequel on pourra déterminer des sous-intervalles optimaux selon le critère de la variance.

Soit f_1, f_2, …,f_n les fonctions d'appartenance respectives des n sujets à un intervalle **A**. On suppose, par analogie avec les densités, que ces fonctions sont normalisées sur **A.** Dans la majorité des cas, chaque sujet contribue de la même façon à la densité, sinon une pondération adaptée ramène à un problème analogue. Alors la fonction $f=(f_1+f_2+ \ldots.+f_n)/n$ intègre en un histogramme sur A la distribution des fonctions d'appartenance. Il suffit ensuite de discrétiser **A** en une suite de points pondérés selon f ; enfin, d'appliquer sur **A** l'algorithme de la variance selon la méthode des nuées dynamiques pour obtenir une variable-intervalle a dont on pourra étudier les relations implicatives avec les autres variables du même type.

Par ex., on donne les valeurs floues de notes obtenues sur [0 ; 20] par 3 étudiants i_1, i_2, i_3 : une correction multiple affecte à i_1, i_2 et i_3 resp . des notes : 5 à 9, 6 à 11 et 8 à 15 (*FIG. 2).*

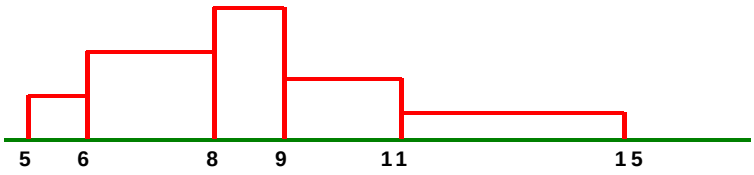


Supposant la distribution uniforme des valeurs floues, normalisées sur [0 ; 20], selon chacun des intervalles, on obtient le tableau 1 des fonctions d'appartenance : par ex. sur [0,3 ; 0,55], correspondant à l'intervalle de notes [6 ;10], f_2=1/5 sur chacun des 5 intervalles d'amplitude 0.05 et f_2= 0 ailleurs.

Individus ↓	Modalités de a				
	[0.25; 0.30]	[0.30 0.40].	[0.40; 0.45]	[0.45; 0.55]	[0.55; 0.75]
i1	1/4	2/4	1/4	0	0
i2	0	2/5	1/5	2/5	0
i3	0	0	1/7	2/7	4/7

TAB 1 – Valeurs prises par les modalités sur les 3 sujets

L'histogramme associé est *FIG 3 ;* par ex. A est discrétisable en 420(ppcm de 3,4,5,7) points



3 Règles d'association pour des variables numériques

On suppose dorénavant que les distributions des variables floues sont connues selon 2 variables observées sur les mêmes sujets : taille et poids. On veut étudier, maintenant, comme en ASI, les règles de déduction entre le prédicat taille et le prédicat poids, présentant des modalités, l'un **Taille** = {petit, moyen, grand}, l'autre **Poids** = {léger, moyen, lourd}. On dispose de données sous forme d'un tableau numérique des degrés d'appartenance aux modalités d'attributs flous, valeurs relatives à un échantillon de 20 sujets. Les 3 premiers constituent le tableau 2. L'un d'entre eux, i_1, n'est donc pas très grand et pas très lourd, l'autre i_2 assez grand et assez lourd, le dernier i_3 plutôt grand et plutôt lourd.

	taille			poids		
	petit T_1	*moyen* T_2	*grand* T_3	*léger* P_1	*moyen* P_2	*lourd* P_3
i_1	8/15	5/15	2/15	7/14	4/14	3/14
i_2	1/14	6/14	7/14	2/15	5/15	8/15
i_3	0	7/16	9/16	1/16	6/16	9/16

TAB 2 – Valeurs prises par les modalités sur les 3 sujets

3.1 Un premier traitement de variables numériques

On propose ici un traitement implicatif, selon l'A.S.I., en considérant les 6 variables tailles-poids comme des variables numériques. On obtient le graphe implicatif en utilisant l'indice de [LAGRANGE 98], réactualisé par [REGNIER et al. 04]. à partir des 20 sujets. Ainsi, les implications T3 ⇒ P3 et P1 ⇒ T1 sont valides au seuil 0.90 et signifient que les propositions grand ⇒ lourd et léger ⇒ petit, règle qui est sémantiquement contraposée de la première, sont acceptables. Une autre implication à un seuil >0.6 apparaît : $P_2 \Rightarrow T_1$.

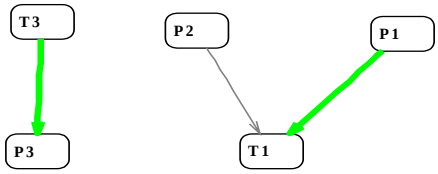


FIG. 4 – Graphe implicatif taille x poids

Ces résultats ne s'opposent pas, bien entendu, au bon sens. Les autres règles d'association confirment une meilleure adéquation à la sémantique de l'implication qu'avec les approches de Reichenbach et Lukasiewicz. On ne retrouve pas, par ex. : léger ⇒ grand.

Mais, l'approche proposée ici présente l'inconvénient de considérer que les 6 modalités des variables « taille » et « poids » sont actives dans le traitement et ne restituent pas, ainsi, les nuances de leur structure. Il semble donc intéressant, sémantiquement parlant, de revenir à la considération de modalités de variables de type intervalles où les modalités apparaissent comme sous-intervalles d'une variable-intervalle principale.

3.2 Second traitement par des variables à valeurs intervalles

Ce second traitement (cf. [GRAS et al. 01b]) va permettre de prendre en compte de façon plus fine les nuances des observations prises selon des sous-ensembles flous et de répartir leurs valeurs de façon optimale sur un intervalle numérique [0 ;1], selon une partition dont l'utilisateur définit le nombre de classes. » pour chacun de 20 sujets.

Nous disposons d'un nouveau tableau 3 donnant les distributions des 6 modalités des 2 attributs « taille » et « poids » relativement à chacun des individus et les valeurs binaires prises par 2 variables supplémentaire »Femme », « Homme ». En voici les 2 premières lignes

	Taille petite t	**Taille moyen. m**	**Taille grande T**	**Var supplé. Femme**	**Var. supplé. Homme**	**Poids léger L**	**Poids moy. o**	**Poids gran. P**
i1	0,7	0,4	0,3	1	0	0,8	0,3	0,1
i2	0,2	0,5	0,8	0	1	0,1	0,4	0,9

TAB 3 – Distributions des attributs flous « taille » et « poids »

Par ex., le sujet i_1 admet un degré d'appartenance 0.7 à la classe des sujets petits, 0.4 à celle des sujets de taille moyenne et 0.3 à la classe des sujets de grande taille. De plus (variable supplémentaire) ce sujet est une femme et la distribution de ses degrés d'appartenance aux 3 classes de poids, sont resp. 0.8, 0.3 et 0.1. Le traitement va emprunter cette fois la méthode des variables à valeurs intervalles. Comme dans le § 2, chaque modalité conduira à la construction de sous-intervalles optimaux, c'est-à-dire la détermination de sous-intervalles optimisant, du moins localement sinon globalement, l'inertie inter-classe. Utilisant ensuite CHIC de traitement de ce type de variable, on établit les règles telles que : si un sujet relève de l'intervalle t_i de la modalité t de l'attribut « taille » alors généralement il relève de l'intervalle p_j de la modalité p de l'attribut « poids ». Ainsi, si par ex., il a tendance à être plutôt petit, alors il a généralement tendance à être plutôt léger.

Les partitions en 3 sous-intervalles calculées par CHIC sont données dans tableau 3.

tailles petites: t1 de 0 à 0.1 t2 de 0.2 à 0.5 t3 de 0.6 à 1	**tailles moyennes :** m1 de 0.1 à 0.3 m2 de 0.4 à 0.6 m3 de 0.8 à 0.8	**grandes tailles :** T1 de 0 à 0.1 T2 de 0.2 à 0.5 T3 de 0.8 à 0.9
poids légers : L1 de 0.1 à 0.3 L2 de 0.5 à 0.6 L3 de 0.8 à 1	**poids moyens :** o1 de 0.2 à 0.3 o2 de 0.4 à 0.5 o3 de 0.6 à 0.7	**poids lourds :** P1 de 0 à 0.1 P2 de 0.2 à 0.4 P3 de 0.7 à 0.9

TAB 3 – Partitions optimales calculées par CHIC

Le graphe implicatif au niveau de confiance 0.90 est également donné par CHIC :

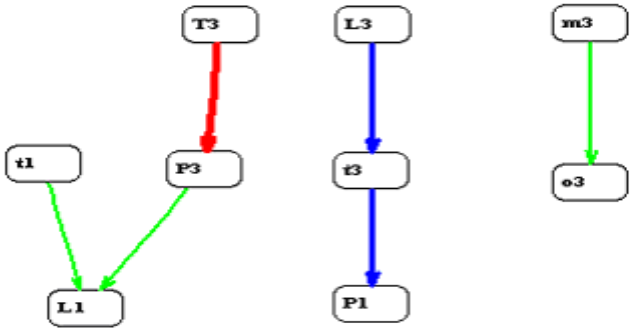


FIG. 5 – Graphe implicatif taille x poids

On voit par exemple que :

- l'individu de grande taille (T3) admet généralement un poids important (P3) et donc n'est pas considéré comme léger(L1). Ce sont les hommes qui apportent, et de très loin (risque de se tromper = 0.07), la plus importante contribution ;
- l'individu de poids plutôt léger (L3) est généralement de petite taille (t3) ;dans ce cas, ils ne sont que très rarement considérés lourds (P1). Ce sont les femmes qui sont les plus contributives à ce chemin (risque = 0.25) ;
- les deux variables t1 et L1, liées par la règle $t1 \Rightarrow L1$, correspondent à des fréquences rares. Si donc, on rencontre un sujet petit alors il est généralement léger. Le sexe Homme contribue à la formation de cette règle.

4 Conclusion

A l'aide de l'A.S.I., nous avons cherché à objectiver la notion de degré d'appartenance.

Situant le modèle d'implication entre attributs par rapport à des modèles classiques, nous avons mis en évidence par un graphe, les relations implicatives entre des modalités de variables numériques. Nous avons, semble-t-il, amélioré la formalisation de la sémantique en faisant référence à des variables-intervalles. Les règles les plus consistantes ont pu être extraites selon leur qualité. Enfin, la relation entre des variables extrinsèques et ces règles ont permis d'enrichir notre connaissance sur ces règles. Des applications à des situations réelles tenteront de valider cette nouvelle approche de l'incertain.

Remerciements à Maurice Bernadet pour sa lecture du texte et ses précieux conseils

Références

[BERNADET 96] Bernadet M., Rose G., Briand H., FIABLE and fuzzy FIABLE : two learning mechanisms based on a probabilistic evaluation of implications, Conference IPMU'96, Granada, Juillet 1996, p. 911-916

[BERNADET 04] Bernadet M., Qualité des règles et des opérateurs en découverte de connaissances floues. Mesure de qualité pour la fouille de données, Cepaduès, RNTI-E-1, p 169-192

[COUTURIER 01] Couturier R.,Traitement de l'analyse statistique implicative dans CHIC, Actes des Journées « Fouille des données par l'analyse statistique implicative », IUFM Caen, 2001, 33-50

[DIDAY 72] Diday E., Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes, Thèse d'Etat, Université de Paris VI, 1972.

[DUBOIS et PRADE 87] Dubois D. et Prade H., Théorie des possibilités. Applications à la représentation des connaissances en informatique, Masson

[GRAS 79] Gras R, Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'Etat, Rennes 1

[GRAS et al. 96] Gras R., Ag Almouloud S., Bailleul M., Lahrer A., Polo M., Ratsimba-Rajohn H. et Totohasina A., L'implication Statistique, La Pensée Sauvage, Grenoble

[GRAS et al. 01a] Gras R., Kuntz P. et Briand H., Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, Mathématiques et Sciences Humaines, n° 154-155, p 9-29, ISSN 0987 6936

[GRAS et al. 01b] Gras R., Diday E., Kuntz P. et Couturier R., Variables sur intervalles et variables-intervalles en analyse implicative, Actes du 8ème Congrès de la SFC de Pointe à Pitre, 17-21 décembre 2001, pp 166-173

[LAGRANGE 98] Lagrange J.B., Analyse implicative d'un ensemble de variables numériques , Revue de Statistique Appliquée, 1998, 71-93.

[LERMAN 81] Lerman I.C., Classification et analyse ordinale des données, Dunod, 1981.

[REGNIER et al. 04] Régnier J.C. et Gras R : Statistique de rangs et analyse statistique implicative, Revue de Statistique Appliquée, à paraître en 2005

[SPAGNOLO et al. 04] Spagnolo F., Gras R., A new approach in Zadeh's classification : fuzzy implication through statistic implication, NAFIPS 2004, 23nd Conference of the North American Fuzzy Information Processing Society, june 27-30, Banff, AB Canada

[ZADEH 79] Zadeh L.A., A Theory of Approximate Reasoning, J. Hayes, D. Michie, and L.I. Mikulich eds., Machine Intelligence 9, New York: Halstead Press, pp. 149-194.

[ZADEH 97] Zadeh L.A., Toward a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic, Fuzzy Sets and Systems 90, pp. 111-127.

{ZADEH 01]Lotfi A. et Zadeh L.A**.,** From computing with numbers to computing with words from manipulation of measurements to manipulation of perception, in Proceedings "Human and machine perception" (Thinking, deciding and acting), Edited by V. Cantoni, V. Di Gesù, A. Setti e D. Tegolo, Kluwer Academic, New York, 2001.

IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles

Julien Blanchard, Fabrice Guillet
Henri Briand, Régis Gras

LINA – FRE 2729 CNRS
Polytech'Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Résumé. La mesure de la qualité des connaissances est une étape clef d'un processus de découverte de règles d'association. Dans cet article, nous présentons un nouvel indice de qualité de règle nommé *IPEE*. Il a la particularité unique d'associer les deux caractéristiques suivantes : d'une part, il est fondé sur un modèle probabiliste, et d'autre part, il mesure un écart à l'équilibre (équirépartition des exemples et des contre-exemples). Nous étudions les propriétés de ce nouvel indice et montrons dans quelles situations il s'avère plus utile qu'un indice d'écart à l'indépendance.

1 Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d'association [Agrawal *et al.* 1993] sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives $a \rightarrow b$ où a et b sont des conjonctions d'items (variables booléennes de la forme $attribut = valeur$). Une telle règle signifie que la plupart des enregistrements qui vérifient la prémisse a dans la base de données vérifient aussi la conclusion b.

Une étape cruciale dans un processus de découverte de règles d'association est la validation des règles après leur extraction. En effet, de par leur nature non supervisée, les algorithmes de data mining peuvent produire des règles en très grande quantité et dont beaucoup sont sans intérêt. Pour aider le décideur (expert des données étudiées) à trouver des connaissances pertinentes parmi ces résultats, l'une des principales solutions consiste à évaluer et ordonner les règles par des mesures de qualité. Il en existe deux catégories : les subjectives (orientées décideur) et les objectives (orientées données). Les mesures subjectives prennent en compte les objectifs du décideur et ses connaissances *a priori* sur les données [Liu *et al.* 2000] [Padmanabhan & Tuzhilin 1999] [Silberschatz & Tuzhilin 1996]. En revanche, seuls les cardinaux liés à la contingence des données interviennent dans le calcul des mesures objectives [Tan *et al.* 2004] [Guillet 2004] [Lallich & Teytaud 2004] [Lenca *et al.* 2004]. Dans cet article, nous nous intéressons aux mesures objectives.

Nous avons montré dans [Blanchard *et al.* 2004] qu'il existe deux aspects différents mais complémentaires de la qualité des règles : l'écart à l'indépendance et l'écart

	Indice d'écart à l'équilibre	**Indice d'écart à l'indépendance**
Indice descriptif	– confiance, – indice de Sebag et Schoenauer, – taux des exemples et contre-exemples, – indice de Ganascia, – moindre-contradiction, – indice d'inclusion...	– coefficient de corrélation, – indice de Loevinger, – lift, – conviction, – TIC, – rapport de cote, – multiplicateur de cote...
Indice statistique		– intensité d'implication, – indice d'implication, – indice de vraisemblance du lien, – contribution orientée au χ^2, – rule-interest...

TAB. 1 – Classification des mesures objectives de qualité de règle

à l'équilibre. Ainsi, les mesures objectives de qualité se répartissent en deux groupes :

- les indices d'écart à l'indépendance, qui prennent une valeur fixe quand les variables a et b sont indépendantes ($n.n_{ab} = n_a n_b$)[1] ;
- les indices d'écart à l'équilibre, qui prennent une valeur fixe quand les nombres d'exemples et de contre-exemples sont égaux ($n_{ab} = n_{a\bar{b}} = \frac{1}{2} n_a$).

Les mesures objectives peuvent également être classées selon leur nature descriptive ou statistique [LALLICH & TEYTAUD 2004] [GRAS *et al.* 2004] :

- Les indices descriptifs (ou fréquentiels) sont ceux qui ne varient pas avec la dilatation des effectifs (quand tous les effectifs des données sont augmentés ou diminués selon la même proportion).
- Les indices statistiques sont ceux qui varient avec la dilatation des effectifs. Parmi eux, on trouve en particulier les mesures probabilistes, qui comparent la distribution observée des données à une distribution théorique, comme l'intensité d'implication [GRAS 1996] ou l'indice de vraisemblance du lien [LERMAN 1981].

A l'aide de ces deux critères, nous classifions les mesures objectives de qualité de règle en quatre catégories. Comme le montre le tableau 1 (voir [GUILLET 2004] pour les références bibliographiques), il n'existe aucun indice statistique qui mesure l'écart à l'équilibre. Pourtant, les indices statistiques ont l'avantage de prendre en compte la taille des phénomènes étudiés. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données. De plus, un indice statistique, lorsqu'il est fondé sur un modèle probabiliste, fait référence à une échelle de valeurs intelligible (échelle de probabilités), ce qui n'est pas le cas de beaucoup d'indices de règle. Un tel indice facilite également la fixation d'un seuil pour le filtrage des règles, puisque le complément à 1 du seuil a le sens du risque d'erreur de première espèce d'un test d'hypothèse (on choisit $\alpha \in \{0.1\%, 1\%, 5\%\}$ généralement dans un test).

Dans cet article, nous proposons un nouvel indice de qualité de règle qui mesure l'écart à l'équilibre tout en étant de nature statistique. Plus précisément, cet indice est fondé sur un modèle probabiliste et évalue la significativité statistique de l'écart à l'équilibre (là où l'intensité d'implication ou l'indice de vraisemblance du lien par exemple évaluent la significativité statistique de l'écart à l'indépendance). Dans la partie suivante nous présentons l'indice probabiliste d'écart à l'équilibre (*IPEE*), puis

[1] Les notations sont définies en partie 2.

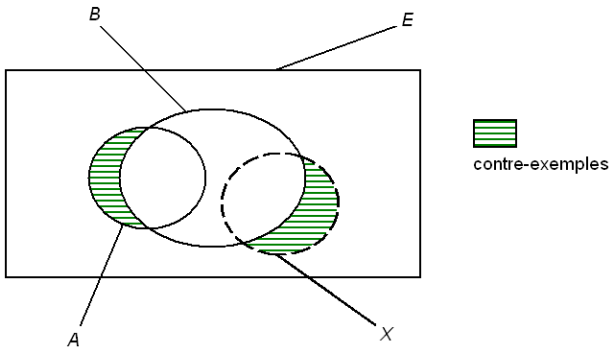


FIG. 1 – Tirage aléatoire d'un ensemble X sous hypothèse d'équiprobabilité entre les exemples et les contre-exemples

étudions en partie 3 ses propriétés. La partie 4 est consacrée à la comparaison des mesures d'écart à l'équilibre et d'écart à l'indépendance.

2 Mesure de la significativité statistique de l'écart à l'équilibre

Nous considérons un ensemble E de n objets décrits par des variables booléennes. Dans le vocabulaire des règles d'association, les objets sont des transactions enregistrées dans une base de données, les variables sont appelées des items, et les conjonctions de variables des itemsets. Etant donné un itemset a, nous notons A l'ensemble des transactions qui vérifient a, et n_a le cardinal de A. Le complémentaire de A dans E est l'ensemble $\overline{A}$ de cardinal $n_{\overline{a}}$. Une règle d'association est un couple (a, b) noté $a \rightarrow b$ où a et b sont deux itemsets qui ne possèdent pas d'item en commun. Les exemples de la règle sont les objets qui vérifient la prémisse a et la conclusion b (objets de $A \cap B$), tandis que les contre-exemples sont les objets qui vérifient a mais pas b (objets de $A \cap \overline{B}$). Dans la suite, nous appelons "variables" les itemsets.

2.1 Modèle aléatoire

Etant donnée une règle $a \rightarrow b$, nous cherchons à mesurer la significativité statistique de l'écart à l'équilibre de la règle. La configuration d'équilibre étant définie par l'équirépartition dans A des exemples $A \cap B$ et des contre-exemples $A \cap \overline{B}$, l'hypothèse de référence est l'hypothèse H_0 d'équiprobabilité entre les exemples et les contre-exemples. Associons donc à l'ensemble A un ensemble aléatoire X de cardinal n_a tiré dans E sous cette hypothèse : $P(X \cap B) = P(X \cap \overline{B})$ (voir figure 1). Le nombre de contre-exemples attendu sous H_0 est le cardinal de $X \cap \overline{B}$, noté $\left|X \cap \overline{B}\right|$. Il s'agit d'une variable aléatoire dont $n_{a\overline{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

Définition 1 L'**indice probabiliste d'écart à l'équilibre (*IPEE*)** d'une règle $a \rightarrow b$ est défini par :

$$IPEE(a \rightarrow b) = \mathrm{P}(\left|X \cap \overline{B}\right| > n_{a\overline{b}} \ \mid \mathrm{H}_0)$$

Une règle $a \rightarrow b$ est dite admissible au seuil de confiance $1-\alpha$ si $IPEE(a \rightarrow b) \geq 1-\alpha$.

IPEE quantifie donc l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\overline{b}}$ eu égard à l'hypothèse H_0. En particulier, si $IPEE(a \rightarrow b)$ est proche de 1 alors il est invraisemblable que les caractères (a *et* b) et (a *et* $\overline{b}$) soient équiprobables. Ce nouvel indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse (et α comme le risque de première espèce de ce test). Toutefois, à l'instar de l'intensité d'implication et de l'indice de vraisemblance du lien (où H_0 est l'hypothèse d'indépendance entre a et b), il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

2.2 Expression analytique

Dans le cadre d'un tirage avec remise, $\left|X \cap \overline{B}\right|$ suit une loi binomiale de paramètres $\frac{1}{2}$ (autant de chances de tirer un exemple que de tirer un contre-exemple) et n_a. *IPEE* s'écrit donc :

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\overline{b}}} \mathrm{C}_{n_a}^{k}$$

IPEE ne dépend ni de n_b, ni de n puisque l'hypothèse d'équilibre H_0 ne se définit pas à l'aide de n_b et de n (contrairement à l'hypothèse d'indépendance). Il est à noter que la significativité statistique de l'écart à l'équilibre pourrait aussi être mesurée en comparant non pas les contre-exemples mais les exemples : $\widehat{IPEE}(a \rightarrow b) = \mathrm{P}(|X \cap B| < n_{ab} \ \mid \mathrm{H}_0)$. Cependant, les distributions binomiales de paramètre $\frac{1}{2}$ étant symétriques, les deux indices sont identiques :

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \mathrm{C}_{n_a}^{n_a-K} = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \mathrm{C}_{n_a}^{K} = \widehat{IPEE}(a \rightarrow b) \text{ où } K = n_a - k$$

Quand $n_a \geq 20$, la loi binomiale peut être approximée par la loi normale de moyenne $\frac{n_a}{2}$ et d'écart-type $\sqrt{\frac{n_a}{4}}$. L'effectif centré réduit $\widetilde{n}_{a\overline{b}}$ des contre-exemples peut être être interprété comme la contribution orientée au χ^2 d'adéquation entre la distribution observée exemples/contre-exemples et la distribution uniforme : $\chi^2 = \widetilde{n}_{a\overline{b}}{}^2$. Ceci constitue une analogie forte avec l'intensité d'implication et l'indice de vraisemblance du lien, puisque dans la modélisation poissonnienne associée à ces indices, les valeurs centrées réduites de $n_{a\overline{b}}$ et n_{ab} peuvent être interprétées comme des contributions orientées au χ^2 d'indépendance entre a et de b [Lerman 1981].

3 Propriétés de la mesure *IPEE*

Le tableau 2 présente les propriétés de la mesure *IPEE*. La mesure est également représentée en fonction du nombre de contre-exemples dans la figure 2. Nous pouvons voir que :

Domaine de variation	$[0;1]$
Valeur pour les règles logiques	$1 - \frac{1}{2^{n_a}}$
Valeur pour les règles à l'équilibre	0.5
Variation avec $n_{a\overline{b}}$ pour n_a constant	$\searrow$
Variation avec n_a pour $n_{a\overline{b}}$ constant	$\nearrow$

TAB. 2 – Propriétés de la mesure *IPEE*

– *IPEE* réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle [GRAS *et al.* 2004].
– Le rejet des règles s'accélère dans une zone d'incertitude autour de l'équilibre $n_{a\overline{b}} = \frac{n_a}{2}$ (décroissance rapide).

Comme le montrent les figures 3.(a) et (b), à proportion exemples/contre-exemples constante, les valeurs prises par *IPEE* sont d'autant plus extrêmes (proches de 0 ou 1) que n_a est grand[2]. En effet, de par sa nature statistique, l'indice prend en compte la taille des phénomènes étudiés : plus n_a est grand, plus on peut avoir confiance dans le déséquilibre exemples/contre-exemples observé dans les données, et plus on peut confirmer la bonne ou la mauvaise qualité de l'écart à l'équilibre de la règle. En particulier, pour *IPEE*, la qualité d'une règle logique (règle qui ne possède pas de contre-exemples, c'est-à-dire $n_{a\overline{b}} = 0$) dépend de n_a (voir tableau 2). Ainsi, contrairement aux autres mesures d'écart à l'équilibre (voir tableau 1), *IPEE* a l'avantage de ne pas attribuer systématiquement la même valeur aux règles logiques . Ceci permet de différencier et hiérarchiser les règles logiques.

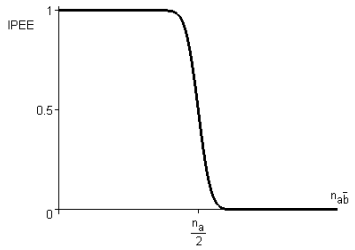


FIG. 2 – Variations de *IPEE* en fonction du nombre de contre-exemples $n_{a\overline{b}}$

En ce qui concerne les symétries, il est à noter que *IPEE* n'en possède aucune : il n'associe pas la même valeur à une règle $a \rightarrow b$ et à sa réciproque $b \rightarrow a$, à sa contraposée $\overline{b} \rightarrow \overline{a}$, ou à sa règle contraire $a \rightarrow \overline{b}$. On a toutefois la relation suivante :

$$IPEE(a \rightarrow \overline{b}) = 1 - IPEE(a \rightarrow b) - \frac{C_{n_a}^{n_{ab}}}{2^{n_a}}$$

(le dernier terme est négligeable quand n_a est grand)

[2]Quand la modélisation retenue est gaussienne, ce comportement est visible directement sur $\widetilde{n}_{a\overline{b}}$: $\widetilde{n}_{a\overline{b}} = \sqrt{n_a}(2 \times confiance - 1)$ où $confiance$ est l'indice bien connu $\frac{nab}{n_a}$ [AGRAWAL *et al.* 1993].

(a) $n_a = 200 \times \gamma$ et $n_{a\bar{b}} = 97 \times \gamma$ (b) $n_a = 200 \times \gamma$ et $n_{a\bar{b}} = 103 \times \gamma$

FIG. 3 – Variations de *IPEE* avec la dilatation des effectifs

Si la prise en compte de la taille des phénomènes étudiés fait la force des mesures de significativité statistique, ceci constitue aussi leur principale limite : elles sont peu discriminantes quand la taille des phénomènes est grande (de l'ordre de 10^4) [ELDER & PREGIBON 1996]. En effet, au regard d'effectifs importants, même des écarts triviaux peuvent s'avérer statistiquement significatifs. *IPEE* ne déroge pas à la règle : quand n_a est grand, l'indice tend à évaluer que les règles sont soit très bonnes (valeurs proches de 1), soit très mauvaises (valeurs proches de 0). Dans ce cas, pour affiner le filtrage des meilleures règles, il faut utiliser en supplément de *IPEE* une mesure descriptive, comme l'indice d'inclusion [GRAS *et al.* 2004] (voir tableau 1). En revanche, contrairement à l'intensité d'implication ou à l'indice de vraisemblance de lien, *IPEE* ne dépend pas de n. L'indice est donc autant sensible aux règles spécifiques ("pépites de connaissances") qu'aux règles générales, et a l'avantage d'être adapté à l'étude des petites bases de données comme des grandes.

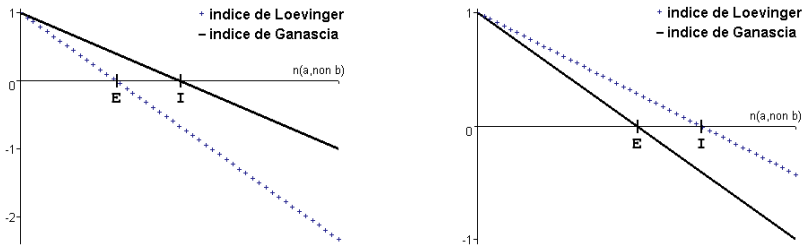
4 Mesures d'écarts à l'équilibre et à l'indépendance : une comparaison

(a) $n_b \geq \frac{n}{2}$ (b) $n_b \leq \frac{n}{2}$

FIG. 4 – Deux cas de figure possibles pour l'équilibre et l'indépendance

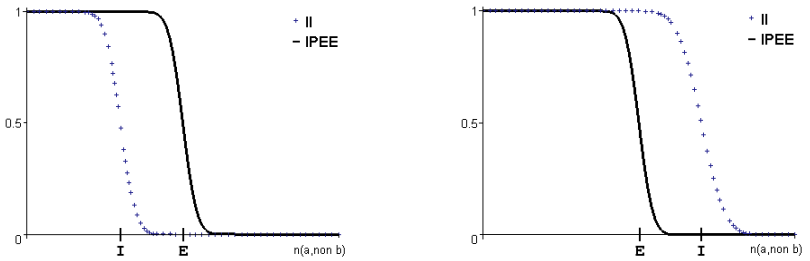
Considérons une règle dont les cardinaux associés sont $n_{a\bar{b}}$, n_a, n_b, n. En faisant varier $n_{a\bar{b}}$ avec n_a, n_b, et n fixes, on peut distinguer deux cas de figure différents pour la règle [BLANCHARD *et al.* 2004] :

– Si $n_b \geq \frac{n}{2}$ (cas 1), alors $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$, et donc la règle passe à l'indépendance avant de passer à l'équilibre quand $n_{a\bar{b}}$ augmente (figure 4.(a)).



(a) cas 1 ($n_b \geq \frac{n}{2}$) (b) cas 2 ($n_b \leq \frac{n}{2}$)

FIG. 5 – Comparaison des indices de Ganascia et Loevinger (E : équilibre, I : indépendance)



(a) cas 1 ($n_b \geq \frac{n}{2}$) (b) cas 2 ($n_b \leq \frac{n}{2}$)

FIG. 6 – Comparaison des indices *IPEE* et intensité d'implication (*II*)

– Si $n_b \leq \frac{n}{2}$ (cas 2), alors $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$, et donc la règle passe à l'équilibre avant de passer à l'indépendance quand $n_{a\bar{b}}$ augmente (figure 4.(b)).

Comparons maintenant dans ces deux cas de figure un indice d'écart à l'équilibre I_{eql} et un indice d'écart à l'indépendance I_{idp}. Afin que la comparaison soit équitable, nous supposons que les deux indices ont des comportements similaires :

– même valeur pour une règle logique,
– même valeur à l'équilibre/indépendance,
– même vitesse de décroissance en fonction des contre-exemples.

Par exemple, I_{eql} et I_{idp} peuvent être l'indice de Ganascia et l'indice de Loevinger, ou bien *IPEE* et l'intensité d'implication. Comme le montrent les figures 5 et 6, I_{idp} est plus filtrant que I_{eql} dans le cas 1, tandis que I_{eql} est plus filtrant que I_{idp} dans le cas 2. En d'autres termes, dans le cas 1 c'est I_{idp} qui contribue à rejeter les mauvaises règles, tandis que dans le cas 2 c'est I_{eql}. Ceci confirme qu'il faut considérer les mesures d'écart à l'équilibre et d'écart à l'indépendance comme complémentaires, les secondes n'étant pas systématiquement "meilleures" que les premières. En particulier, il ne faut pas négliger les indices d'écart à l'équilibre quand les réalisations des variables étudiées sont rares. Dans cette situation, en effet, pour peu que le décideur ne s'intéresse pas aux règles portant sur les non-réalisations (ce qui en général se vérifie dans la pratique),

le cas 2 est plus fréquent que le cas 1.

5 Conclusion

Dans cet article, nous avons présenté un nouvel indice de qualité de règle qui mesure l'écart à l'équilibre au regard d'un modèle probabiliste. De par sa nature statistique, cet indice a l'avantage de prendre en compte la taille des phénomènes étudiés, contrairement aux autres mesures d'écart à l'équilibre. De plus, il fait référence à une échelle de valeurs intelligible (échelle de probabilités). Notre étude montre que *IPEE* est efficace pour évaluer des règles logiques, et bien adapté à la recherche de règles spécifiques ("pépites de connaissance").

IPEE peut être vu comme l'analogue de l'intensité d'implication pour l'écart à l'équilibre. Utilisées conjointement, ces deux mesures permettent une évaluation statistique complète des règles. La suite de ce travail de recherche consistera principalement à intégrer *IPEE* à notre plate-forme de validation de règles *ARVis* [Blanchard *et al.* 2003] afin d'expérimenter le couple (*IPEE*, intensité d'implication) sur des données réelles.

Références

[Agrawal *et al.* 1993] Rakesh Agrawal, Tomasz Imielienski & Arun Swami. «Mining association rules between sets of items in large databases». Dans : «Proceedings of the 1993 ACM SIGMOD international conference on management of data», ACM Press, 1993, pages 207–216.

[Blanchard *et al.* 2003] Julien Blanchard, Fabrice Guillet & Henri Briand. «Une visualisation orientée qualité pour la fouille anthropocentrée de règles d'association». Dans : *Cahiers Romans de Sciences Cognitives*, tome 1 n° 3, 2003, pages 79–100.

[Blanchard *et al.* 2004] Julien Blanchard, Fabrice Guillet, Régis Gras & Henri Briand. «Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-2, 2004, pages 287–298. Actes des journées Extraction et Gestion des Connaissances (EGC) 2004.

[Elder & Pregibon 1996] John F. Elder & Daryl Pregibon. «A statistical perspective on knowledge discovery in databases». Dans : «Advances in knowledge discovery and data mining», , dirigé par Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth & Ramasamy Uthurusamy. AAAI/MIT Press, 1996, pages 83–113.

[Gras 1996] Régis Gras. *L'implication statistique : nouvelle méthode exploratoire de données.* La Pensée Sauvage Editions, 1996.

[Gras *et al.* 2004] Régis Gras, Raphaël Couturier, Julien Blanchard, Henri Briand, Pascale Kuntz & Philippe Peter. «Quelques critères pour une mesure de qualité de règles d'association». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 3–31. Numéro spécial Mesures de qualité pour la fouille de données.

[Guillet 2004] Fabrice Guillet. «Mesures de la qualité des connaissances en ECD», 2004. Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

[Lallich & Teytaud 2004] Stéphane Lallich & Olivier Teytaud. «Evaluation et validation de l'intérêt des règles d'association». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 193–218. Numéro spécial Mesures de qualité pour la fouille de données.

[Lenca *et al.* 2004] Philippe Lenca, Patrick Meyer, Benoît Vaillant, Philippe Picouet & Stéphane Lallich. «Evaluation et analyse multicritère des mesures de qualité des règles d'association». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 219–246. Numéro spécial Mesures de qualité pour la fouille de données.

[Lerman 1981] Israël César Lerman. *Classification et analyse ordinale des données*. Dunod, 1981.

[Liu *et al.* 2000] Bing Liu, Wynne Hsu, Shu Chen & Yiming Ma. «Analyzing the subjective interestingness of association rules». Dans : *IEEE Intelligent Systems*, tome 15 n° 5, 2000, pages 47–55.

[Padmanabhan & Tuzhilin 1999] Balaji Padmanabhan & Alexander Tuzhilin. «Unexpectedness as a measure of interestingness in knowledge discovery». Dans : *Decision Support Systems*, tome 27 n° 3, 1999, pages 303–318.

[Silberschatz & Tuzhilin 1996] Avi Silberschatz & Alexander Tuzhilin. «What makes patterns interesting in knowledge discovery systems». Dans : *IEEE Transactions on Knowledge and Data Engineering*, tome 8 n° 6, 1996, pages 970–974.

[Tan *et al.* 2004] Pang-Ning Tan, Vipin Kumar & Jaideep Srivastava. «Selecting the right objective measure for association analysis». Dans : *Information Systems*, tome 29 n° 4, 2004, pages 293–313.

Le rôle de l'utilisateur dans un processus d'extraction de règles d'association

Cyril Nortet* Ansaf Salleb**

Teddy Turmeaux* Christel Vrain*

*LIFO Rue Léonard de Vinci BP 6759 45067 Orléans Cedex 02
{Cyril.Nortet, Teddy.Turmeaux, Christel.Vrain}@lifo.univ-orleans.fr

**IRISA/INRIA, Projet Dream, Campus de Beaulieu, Rennes
Ansaf.Salleb@irisa.fr

Résumé. De nombreux travaux ont porté sur l'extraction de règles d'association. Cependant, cette tâche continue à intéresser les chercheurs en fouille de données car elle soulève encore plusieurs défis. En particulier, son utilisation en pratique reste difficile : d'une part, le nombre de règles apprises est souvent très grand et décourageant pour l'expert qui souhaite les exploiter, d'autre part, le traitement des valeurs numériques dans cette tâche est loin d'être maîtrisé, ce qui restreint sérieusement son applicabilité aux données réelles.

Nous nous intéressons dans cet article au rôle que peut jouer l'utilisateur pour pallier ces difficultés et extraire des règles de qualité. Il s'agit d'impliquer l'utilisateur dans le processus de recherche de règles d'association qui est dans ce cas interactif et guidé par des schémas de règles qu'il aurait choisis. Nous illustrons notre propos avec QuantMiner qui est un outil convivial et interactif que nous avons développé. La présence de l'expert reste indispensable durant tout le processus d'extraction de règles, pour aller à l'essentiel et visualiser les règles au plus fort potentiel.

Mots clé : Fouille de Données Interactive, Règle d'Association Quantitative, Optimisation.

1 Introduction

L'extraction de règles d'association est devenue aujourd'hui une tâche populaire en fouille de données. Elle a pour but de dégager des relations intelligibles entre des attributs dans une base de données. Une règle d'association (Agrawal et al. 1993) est une implication $\mathcal{C}_1 \Rightarrow \mathcal{C}_2$, où $\mathcal{C}_1$ et $\mathcal{C}_2$ expriment des conditions sur les attributs de la base de données. La qualité d'une règle est classiquement évaluée par un couple de mesures *support* et *confiance*, définis comme suit :

- Support($\mathcal{C}$), où $\mathcal{C}$ exprime des conditions sur les attributs, est le nombre de n-uplets (lignes de la base de données) qui satisfont $\mathcal{C}$.
- Support($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Support($\mathcal{C}_1 \wedge \mathcal{C}_2$)
- Confiance($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Support($\mathcal{C}_1 \wedge \mathcal{C}_2$)/Support($\mathcal{C}_1$)

Une règle d'association est dite *solide*, si son support et sa confiance dépassent deux seuils fixés *a priori*, MinSupp et MinConf respectivement.

De nombreux travaux se sont intéressés au problème crucial de performance que pose cette tâche (par ex. (Brin et al. 1997, Bayardo 1998, Zaki 2000, Salleb et al. 2002)) et des algorithmes de plus en plus performants sont proposés. Cependant d'autres problèmes persistent dont les deux suivants :

• **Problème 1 :** le nombre de règles apprises est souvent très grand et décourageant pour l'expert qui voudrait les exploiter. Il est donc souhaitable de privilégier la qualité des règles apprises, leur intérêt pour l'expert à la quantité de règles extraites.

• **Problème 2 :** les travaux existant ne gèrent pas bien, voire pas du tout, les valeurs numériques. Une pré-discrétisation (découpage du domaine de l'attribut numérique en intervalles) est souvent effectuée mais reste non satisfaisante. Ceci restreint sérieusement l'applicabilité de cette tâche aux données réelles qui renferment des valeurs catégoriques et numériques.

Nous montrons dans cet article qu'il est possible de répondre à ces deux problèmes en impliquant l'utilisateur durant le processus de fouille de données. Celui-ci choisit des schémas de règles d'association qui l'intéressent (certains travaux ont montré l'intérêt de restreindre la forme des règles (Srikant et al. 1997)). Un algorithme génétique permettant de découvrir de façon dynamique les intervalles des variables numériques qui optimisent le support et la confiance de chaque schéma est ensuite employé.

QuantMiner[1] est un outil interactif et convivial (Nortet et al. 2005) fondé sur cette approche du problème. Il a été utilisé sur deux applications, l'une concerne la recherche de règles d'association dans un Système d'Information Géographique (SIG) développé par le BRGM, la seconde concerne des données médicales relatives à la maladie de l'athérosclérose. Nous illustrons le processus d'extraction de règles d'association dans QuantMiner à travers cette seconde application.

2 QuantMiner

L'idée de QuantMiner (Nortet et al. 2005) est de considérer des **schémas de règles**. Un schéma de règle est une règle présentant dans ses membres gauche et droit des items catégoriques aux valeurs fixées ou non et des items numériques dont les intervalles correspondant ne sont pas encore instanciés. Puis par optimisation nous cherchons les bornes les plus adaptées pour chacun de ses intervalles, en prenant en compte la mesure du Gain (Fukuda et al. 1996) donnée par :

$$\text{Gain}(A \Rightarrow B) = \text{Supp}(AB) - \text{MinConf} * \text{Supp}(A)$$

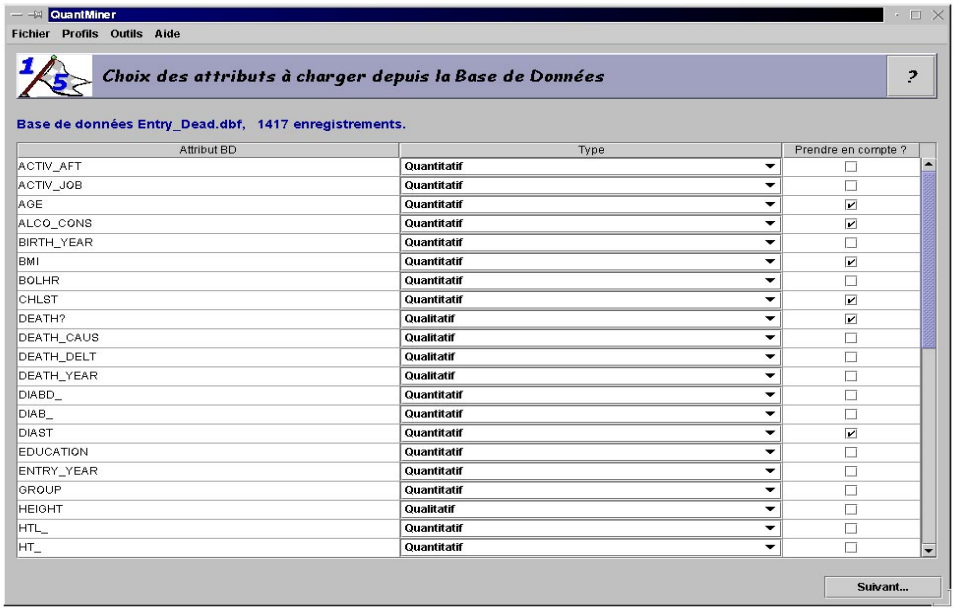
Seuls les règles aux meilleurs Gains sont gardées. Dans QuantMiner, le processus de fouille de données est interactif; l'utilisateur est dans ce cas tout *près des règles*, précisant celles qui l'intéressent. Le processus est aussi itératif dans la mesure où l'utilisateur peut sauvegarder son contexte d'extraction (schémas, méthode et paramètres) et recommencer le processus ultérieurement. Le système prend la forme d'un assistant (wizard), limité à 5 étapes. Nous illustrons les étapes de QuantMiner à travers

[1] Développé en collaboration avec le BRGM, Bureau de Recherches Géologiques et Minières (service REM/VADO)

une application réelle portant sur la maladie de l'athérosclérose (Salleb et al. 2004, Nortet et al. 2005).

L'athérosclérose est une maladie répandue et grave des artères dont les parois se durcissent provoquant une gêne considérable à la circulation du sang. Le projet STULONG[2] porte sur une étude médicale effectuée pendant 20 ans sur les facteurs de risque de cette maladie et concerne une population de patients composée de plus de 1 400 hommes adultes qui ont été classés en trois groupes : le groupe des patients normaux N, le groupe des patients à risque R et enfin le groupe P des patients présentant la pathologie. Nous nous sommes intéressés à l'extraction de règles d'association dans un but descriptif comme par exemple, décrire les patients décédés et les patients non décédés. Pour cela, nous avons construit une table Entry_Dead décrivant des patients à partir d'une table préparée par (Lucas et al. 2002)) ainsi que les tables de STULONG. Chaque patient est décrit par son poids, sa taille, son âge, ses activités physiques ainsi que sa consommation de tabac et d'alcool. On retrouve aussi d'autres informations liées aux examens cliniques effectués tels que son taux de cholestérol, ses pressions artérielles, etc. Au total 27 attributs catégoriques et 17 numériques décrivent les patients.

2.1 Étape 1 : Choix des attributs



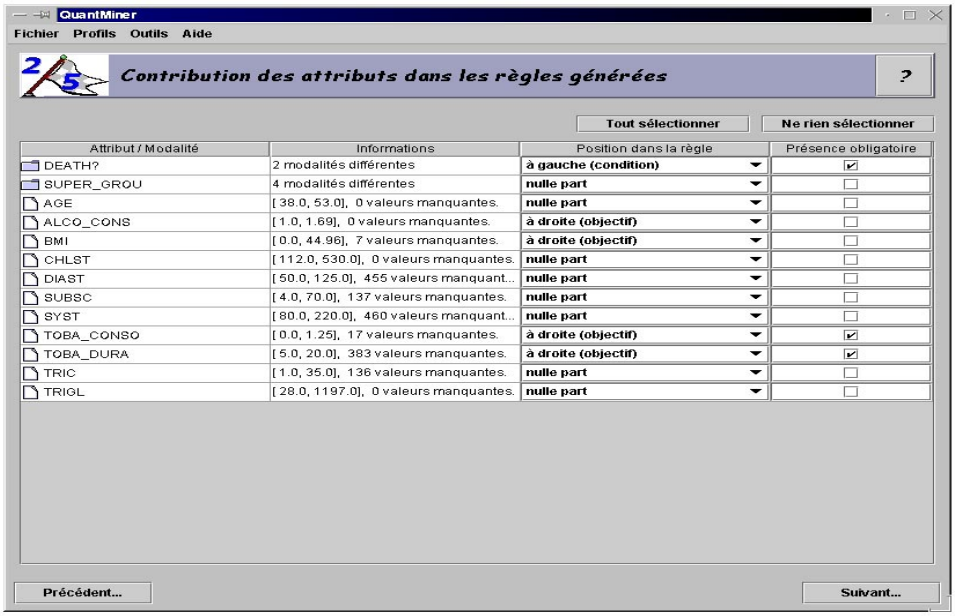
Le logiciel prend en charge des fichiers de données au format DBF (DataBase File,

[2] « The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudik, MD, ScD, with collaboration of M. Tomeckova, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvarova, DrSc). The data resource is on the web pages http://euromise.vse.cz/STULONG. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107. »

un standard dBase). Cette étape sert à sélectionner des attributs et à définir leurs types. Pour des raisons de performances lors du chargement des données en mémoire, il convient de rester raisonnable pour ne pas surcharger le système avec des attributs qu'on ne souhaite pas voir apparaître dans les règles (par exemple l'identifiant et la date de naissance). Le type de chaque attribut est détecté automatiquement, néanmoins le choix est laissé à l'utilisateur pour vérifier ou corriger si nécessaire.

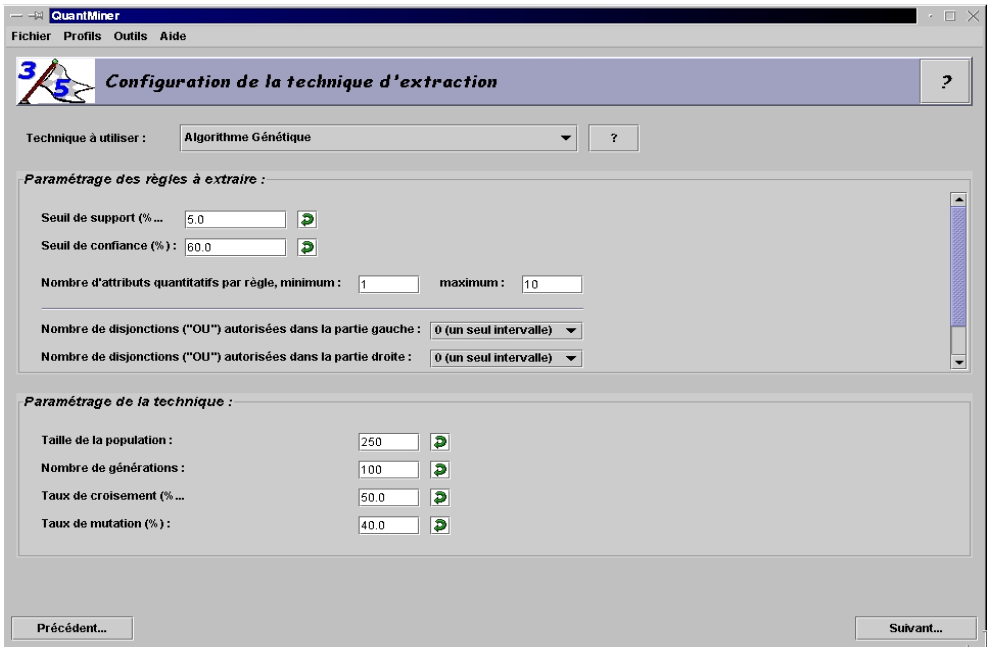
2.2 Étape 2 : Choix des schémas

Cette étape permet une répartition fine des attributs ($A_i = v_i$ ou $A_i \in [l_i, u_i]$, les valeurs l_i, u_i non instanciées, v_i instanciée ou non) aux places où on souhaite les voir apparaître dans les règles. Chacun peut être indépendamment placé à gauche de la règle (condition), à droite (objectif), ne pas y apparaître, ou apparaître impérativement dans la règle. Par exemple, si l'on souhaite travailler sur des règles portant sur l'influence de la consommation de tabac sur le décès de patients, il suffira d'indiquer que tous les attributs n'apparaissent nulle part, sauf "DEATH ?" à gauche et "TOBA_CONSO" et "TOBA_DURA" à droite, comme indiqué dans la figure ci-dessous.

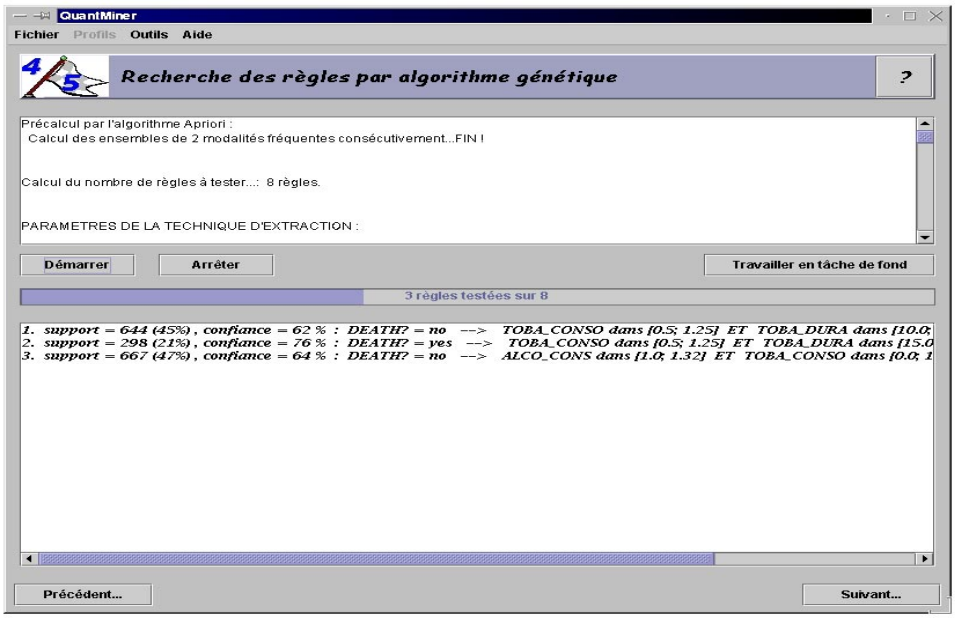


2.3 Étape 3 : Choix de la méthode

Il s'agit de choisir une technique d'optimisation et régler ses paramètres. En plus du support minimum et de la confiance minimum, les paramètres de l'algorithme génétique sont la taille de la population, le nombre de générations, les taux de mutations et de croisements. Nous avons fixé ces valeurs par défaut à 250 individus, 100 générations, 40% de mutations et 50% de croisements.



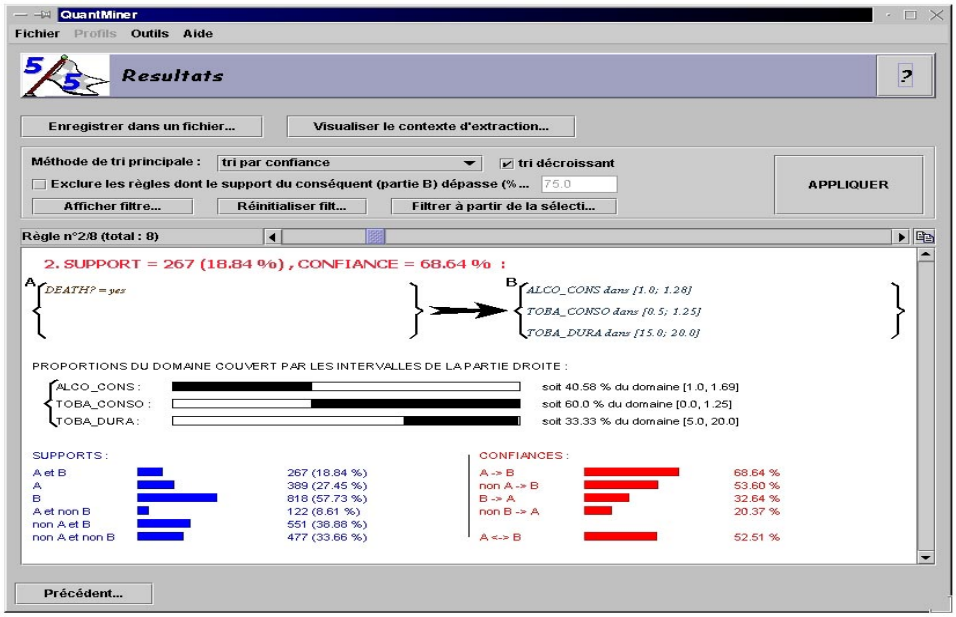
2.4 Étape 4 : Exécution



L'utilisateur lance l'exécution de l'algorithme d'optimisation choisi. Les règles apprises sont affichées sous forme *horizontale* au fur et à mesure du calcul. Le temps d'exécution dépend du nombre d'attributs choisis dans le schéma de règle. Cependant, cette étape peut être stoppée et les résultats trouvés jusqu'alors sauvegardés.

2.5 Étape 5 : Visualisation des règles

Un affichage *vertical* règle par règle est présenté à l'utilisateur. Pour rendre les règles plus exploitables, de nombreux paramètres statistiques les accompagnent dans l'affichage. Pour les attributs numériques, la proportion de l'intervalle dans la règle par rapport au domaine entier de l'attribut est présentée pour montrer la pertinence de l'intervalle optimisé par le système. L'utilisateur peut sauvegarder les résultats sous formes horizontale et verticale pour les visualiser ultérieurement.



3 Discussion et Conclusion

Quelles leçons pouvons nous tirer de l'expérience de QuantMiner ?

- Un processus d'extraction doit être aussi simple que possible avec un nombre d'étapes d'extraction et un nombre de paramètres limités.
- Une visualisation conviviale des résultats d'extraction est primordiale pour permettre à l'utilisateur de manipuler et identifier les règles intéressantes.
- Tel que nous l'avons vu, QuantMiner ne se contente pas de donner la confiance d'une règle de la forme $A \Rightarrow B$, mais il donne aussi la confiance de $\neg A \Rightarrow B$,

$B \Rightarrow A$, $\neg B \Rightarrow A$. Cependant, lors de l'application aux données médicales, il nous a semblé que ces critères devaient être affinés dans le cas de la caractérisation de plusieurs classes. Dans le cas où A représente des caractéristiques des patients à risques (exemple ci-dessous), $\neg A$ regroupe les patients bien portants et malades; il serait plus intéressant de séparer ces deux classes.

$$\texttt{GROOUP=R} \Rightarrow \left\{ \begin{array}{l} \texttt{ALCO\_CONS} \in \texttt{[1.0,1.29]} \\ \texttt{BMI} \in \texttt{[22.28,30.72]} \\ \texttt{TOBA\_CONSO} \in \texttt{[0.5,1.25]} \end{array} \right\} \begin{array}{l} \texttt{supp(A} \Rightarrow \texttt{B) = 39\%} \\ \texttt{conf(A} \Rightarrow \texttt{B) = 64\%} \\ \texttt{conf(}\neg\texttt{A} \Rightarrow \texttt{B) = 38\%} \end{array}$$

– Souvent l'expert aimerait bien découvrir dans ses données une connaissance surprenante, *exceptionnelle*, alors que les règles d'association fréquentes risquent d'être déjà connues. Par conséquent, la recherche de règles surprenantes s'oriente vers des associations ayant des supports relativement faibles mais des confiances fortes. Baisser le support n'est pas une solution car survient alors le problème du grand nombre de règles générées. Nous nous intéressons à cette problématique (Duval et al. 2004) qui nous semble prometteuse dans la voie d'extraction de connaissances de qualité.

– On peut enfin se demander jusqu'à quel point impliquer l'expert dans un processus de fouille de données et l'impact que cela entraîne sur l'intérêt des connaissances apprises. Un compromis reste sans doute à trouver en fonction des applications et du besoin de l'utilisateur final.

Les bases de données du monde réel sont de plus en plus complexes et volumineuses, manipulant des données catégoriques et numériques. Extraire des règles à partir de telles données requiert des algorithmes rapides et intelligents pouvant capturer aussi bien les *régularités* que les *exceptions* enfouies dans ces données. Ces algorithmes doivent tenir compte du besoin de l'expert et lui fournir des connaissances intéressantes; tels sont aujourd'hui les défis de l'extraction de connaissances exprimées par des règles dans les bases de données.

Remerciements Nous tenons à remercier les relecteurs pour leurs remarques et suggestions.

Références

Agrawal R., Imielinski T. et Swami A. N. (1993). Mining association rules between sets of items in large databases. Dans Buneman P. et Jajodia S., éditeurs, *Proceedings of the 1993 ACM SIGMOD*, pages 207–216, Washington, D.C.

Bayardo R. J. (1998). Efficiently mining long patterns from databases. Dans *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data*, pages 85–93, Seattle.

Brin S., Motwani R. et Silverstein C. (1997). Beyond market baskets: generalizing association rules to correlations. Dans *Proc. of ACM SIGMOD*, pages 265–276.

Duval B., Salleb A. et Vrain C. (2004). Méthodes et mesures d'intérêt pour l'extraction de règles d'exception. *Revue des Nouvelles Technologies de l'Information - Mesures de Qualité pour la Fouille de Données RNTI-E-1*, pages 119–140.

Fukuda T., Morimoto Y., Morishita S. et Tokuyama T. (1996). Data mining using two-dimensional optimized association rules: Scheme, algorithms and visualization. Dans *Proc. of the Int'l Conf. ACM SIGMOD*, pages 12–23.

Lucas N., Azé J. et Sebag M. (2002). Atherosclerosis Risk Identification and Visual Analysis. Dans *ECML/PKDD 2002 Discovery Challenge Workshop program*.

Nortet C., Salleb A., Turmeaux T. et Vrain C. (2005). Extraction de Règles d'Association Quantitatives - Application à des Données Médicales. Dans Vincent N. et Pinson S., éditeurs, *à paraître dans EGC 2005 (Cinquièmes Journées sur Extraction et Gestion des Connaissances)*. Cépaduès éditions.

Salleb A., Maazouzi Z. et Vrain C. (2002). Mining Maximal Frequent Itemsets by a Boolean Based Approach. Dans Harmelen F., éditeur, *15th European Conference on Artificial Intelligence Ecai*, pages 385–389, Lyon, France. IOS Press Amsterdam.

Salleb A., Turmeaux T., Vrain C. et Nortet C. (2004). Mining quantitative association rules in a atherosclerosis dataset. Dans *Proceedings of the PKDD Discovery Challenge 2004 (co-located with the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases)*, pages 98–103, Pisa, Italy.

Srikant R., Vu Q. et Agrawal R. (1997). Mining association rules with item constraints. Dans Heckerman D., Mannila H., Pregibon D. et Uthurusamy R., éditeurs, *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 67–73. AAAI Press.

Zaki M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

Arbre BIC optimal et taux d'erreur

Gilbert Ritschard

Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Nous reconsidérons dans cet article le critère BIC pour arbres d'induction proposé dans Ritschard et Zighed (2003, 2004) et discutons deux aspects liés à sa portée. Le premier concerne les possibilités de le calculer. Nous montrons comment il s'obtient à partir des statistiques du rapport vraisemblance utilisées pour tester l'indépendance ligne-colonne de tables de contingence. Le second point porte sur son intérêt dans une optique de classification. Nous illustrons sur l'exemple du Titanic la relation entre le BIC et le taux d'erreur en généralisation lorsqu'on regarde leur évolution selon la complexité de l'arbre. Nous esquissons un plan d'expérimentation en vue de vérifier la conjecture selon laquelle le BIC minimum assurerait en moyenne le meilleur taux d'erreur en généralisation.

1 Introduction

La qualité des arbres de classification, comme pour d'autres classifieurs, est le plus souvent établie sur la base du taux d'erreur de classement en généralisation. Si l'on examine l'évolution de ce taux en fonction de la complexité du classifieur, il est connu qu'il passe par un minimum au delà duquel on parle de sur-apprentissage (*overfitting*). Intuitivement, l'explication de ce phénomène tient au fait qu'au delà d'un certain seuil, plus on augmente la complexité, plus l'arbre devient dépendant de l'échantillon d'apprentissage utilisé, au sens où il devient de plus en plus probable que de petites perturbations de l'échantillon entraîneront des modifications des règles de classification. Lorsqu'il s'agit d'utiliser l'arbre pour la classification, il semble dès lors naturel de retenir celui qui minimise le taux en généralisation.

Mais comment s'assurer a priori que l'arbre induit sera celui qui minimisera le taux en généralisation ? On pourrait songer à partager les données disponibles pour l'induction en un échantillon d'apprentissage et un échantillon test et à exploiter le taux d'erreur sur les données test comme critère de construction de l'arbre. Ceci reviendrait cependant simplement à transformer les données test en données d'apprentissage et ne peut donc être une solution. Il s'agit de disposer d'un critère qui, tout en se calculant sur l'échantillon d'apprentissage, nous assure que le taux d'erreur sera en moyenne minimum pour tout ensemble de données supplémentaires. A défaut de pouvoir mesurer a priori le taux d'erreur en généralisation, on s'intéresse à la complexité qu'il s'agit de minimiser et l'on tentera de retenir le meilleur compromis entre qualité d'information sur données d'apprentissage et complexité.

Le critère BIC (Bayesian Information Criteria) pour arbre que nous avons introduit dans Ritschard et Zighed (2003, 2004) pour comparer la qualité de la description

des données fournies par différents arbres nous semble pouvoir être une solution de ce point de vue puisqu'il combine un critère d'ajustement (la déviance) avec une pénalisation pour la complexité (le nombre de paramètres). D'autres critères, dont la description minimale de données (Rissanen, 1983) et le message de longueur minimal, MML, (Wallace et Freeman, 1987) qui combinent également une qualité d'information et une pénalisation pour la complexité pourraient également s'avérer intéressants de ce point de vue. Le critère BIC considéré ici résulte d'une logique bayésienne (Raftery, 1995) tout comme le critère que Wehenkel (1993) utilise pour l'élagage.

Avant d'examiner le lien du BIC avec le taux d'erreur en généralisation, nous rappelons à la section 2 sa définition et en particulier celle de la déviance sur la laquelle il se fonde. Nous montrons que la déviance se déduit directement de la valeur de la statistique du rapport de vraisemblance de deux tests d'indépendance, ce qui permet en particulier à tout un chacun de calculer le BIC d'un arbre en utilisant n'importe quel logiciel statistique classique, SPSS par exemple, qui donne ces statistiques. Nous illustrons ensuite à la section 3 le lien entre le critère BIC et le taux d'erreur et discutons brièvement d'un protocole d'expérimentation en vue de vérifier la conjecture selon laquelle la minimisation du BIC assurerait la minimisation du taux moyen d'erreur en généralisation.

2 Le critère BIC pour arbre d'induction

Pour illustrer notre discussion nous utilisons les données du Titanic où il s'agit de prédire pour chaque passager s'il survit ou pas selon trois attributs, soit le sexe (F,M), l'âge (A=adulte, C=enfant) et la classe (c1, c2, c3 et c4=équipage). La figure 1 donne l'arbre induit avec la méthode Exhaustive CHAID de Answer Tree 3.1 (SPSS, 2001) en utilisant le khi-deux du rapport de vraisemblance, un seuil de signification de 5% et les contraintes minimales sur la taille des nœuds.

Avant de définir le critère BIC, nous devons expliquer la déviance que nous notons $D(m)$ pour un arbre m. Considérons pour cela la table de contingence cible dont les ℓ lignes sont définies par la variable à prédire ("survit ou pas" dans notre cas) et dont les c colonnes correspondent à l'ensemble des profils différents que l'on peut définir avec les attributs prédictifs. Dans notre cas on a 14 profils différents, soit $2 \times 2 \times 4$ moins 2 puisqu'il n'y a pas d'enfant, ni fille ni garçon parmi l'équipage. La déviance mesure la divergence entre la table cible et sa prédiction à partir de l'arbre induit. Formellement, la déviance se calcule comme suit

$$D(m) \quad = \quad -2 \sum_{i=1}^{\ell} \sum_{j=1}^{c} n_{ij} \ln \left(\frac{\hat{n}_{ij}}{n_{ij}} \right) \tag{1}$$

en considérant les termes $n_{ij} \ln(\hat{n}_{ij}/n_{ij})$ comme nuls lorsque $n_{ij} = 0$.

Les prédictions $\hat{n}_{ij}$ s'obtiennent en ventilant le total des cas avec profil j selon la distribution observée dans la feuille de l'arbre induit qui comprend le profil j. Le tableau 1 donne par exemple (sous forme transposée pour des raisons de présentation) la table cible et la table prédite avec l'arbre induit de la figure 1. Avec la formule ci-dessus on établit que la déviance vaut ici 6.93.

	table cible (n_{ij})		table prédite ($\hat{n}_{ij}$)		
	yes	no	yes	no	total
MAc1	45	88	45	88	133
MAc2	10	114	10	114	124
MAc3	59	289	67.6175	280.3825	348
MAc4	132	503	123.3825	511.6175	635
MCc1	4	0	4	0	4
MCc2	8	0	8	0	8
MCc3	10	23	10	23	33
FAc1	112	4	112.0342	3.9658	116
FAc2	66	11	67.375	9.625	77
FAc3	62	71	59.5	73.5	133
FAc4	15	2	15	2	17
FCc1	1	0	0.9658	0.0342	1
FCc2	11	0	9.625	1.375	11
FCc3	6	13	8.5	10.5	19
total	541	1118	541	1118	1659

TAB. 1 – Table cible et effectifs prédits

Le critère BIC pour un arbre induit m qui donne lieu à $q \leq c$ feuilles pour une variable à prédire avec ℓ classes est alors défini, à une constante additive près, par

$$\text{BIC}(m) \quad = \quad D(m) + p\ln(n) \tag{2}$$

où n est la taille de l'échantillon d'apprentissage, $p = (\ell-1)q+c$ le nombre de paramètres de l'arbre et $D(m)$ la déviance.

En présence d'un grand nombre d'attributs, le nombre de profils différents possibles peut évidemment rapidement devenir trop grand pour envisager un calcul manuel de

	table T_m		
	yes	no	total
MAc1	45	88	133
MAc2	10	114	124
MAc3,c4	191	792	983
MCc1	4	0	4
MCc2	8	0	8
MCc3	10	23	33
FA,Cc1	113	4	117
FA,Cc2	77	11	88
FA,Cc3	68	84	152
FAc4	15	2	17
total	541	1118	1659

TAB. 2 – Table croisant la variable à prédire avec les feuilles

Arbre BIC optimal et taux d'erreur

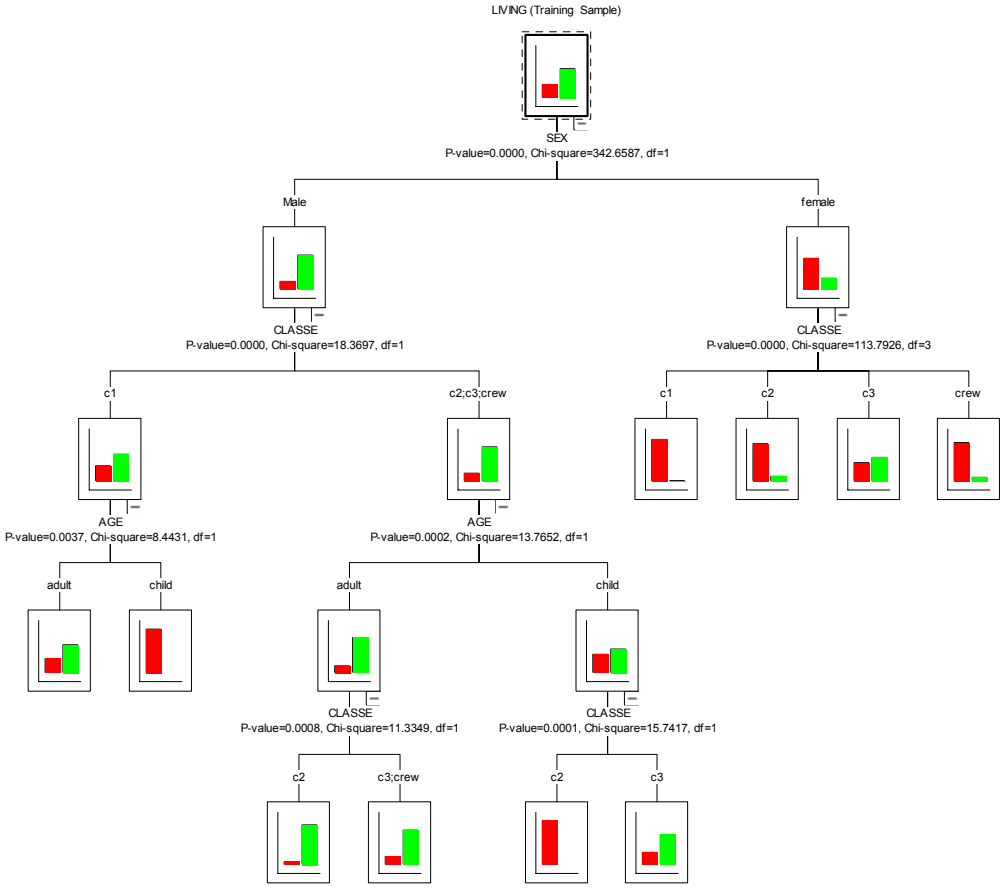


FIG. 1 – Arbre induit, échantillon d'apprentissage, $n = 1659$

la déviance. On peut dans ce cas exploiter les logiciels qui calculent la statistique du rapport de vraisemblance pour le test d'indépendance sur table de contingence. Notons T la table cible et T_m la table de contingence qui croise la variable à prédire avec les feuilles de l'arbre induit. Pour notre exemple, on trouve T au tableau 1 tandis que T_m est donné au tableau 2. En raison des propriétés d'additivité de la déviance, on a $D(m_0) = D(m) + D(m_0|m)$ où m_0 représente le nœud initial qui correspond à l'indépendance. La divergence $D(m_0)$ entre ce nœud initial et la table cible est précisément la statistique du rapport de vraisemblance pour le test d'indépendance sur le tableau T. De même la divergence $D(m_0|m)$ entre m_0 et m est le khi-deux du rapport de vraisemblance pour le test d'indépendance sur la table T_m.

La statistique du rapport de vraisemblance vaut respectivement 531.04 pour T et 524.11 pour T_m. On trouve donc $D(m) = 531.04 - 524.11 = 6.93$ ce qui correspond à la valeur trouvée précédemment. On en déduit la valeur du BIC, soit $\text{BIC}(m) = 6.93 + 24 \ln(1659) = 184.87$.

Notons que le BIC est défini à une constante additive près. Comme à chaque fois que le nombre de paramètres augmente d'une unité le nombre d de degrés de liberté associé à la déviance diminue d'une unité, on peut également considérer la définition $\text{BIC}(m) = D(m) - d\ln(n)$. C'est la définition que nous avons utilisée au tableau 3 où nous avons cependant encore ajouté 100 pour éviter les valeurs négatives.

3 BIC et le taux d'erreur en généralisation

On se propose à présent de discuter le lien entre le critère BIC et le taux d'erreur en généralisation. Notre conjecture est que l'arbre BIC optimal devrait plus ou moins assurer le plus petit taux d'erreur en généralisation.

Nous avons calculé la valeur du critère BIC et le taux d'erreur pour différents arbres obtenus en élaguant successivement l'arbre saturé de la figure 3. L'échantillon d'apprentissage comprend 1659 cas et l'échantillon test 542. On notera que sur cet exemple très simple, le taux d'erreur en généralisation, bien que supérieur au taux d'erreur en substitution, ne remonte pas pour les arbres les plus complexes. Le tableau 3 et la figure 2 font apparaître que le BIC correspond au modèle le plus simple pour lequel on a le taux d'erreur minimal. Ceci semble plutôt conforter notre conjecture.

Nous n'avons ici bien évidemment considéré qu'un seul ensemble test qui ne saurait suffire à démontrer notre conjecture. Notre intention est de procéder à une expérimentation plus complète. Le protocole envisagé est de postuler successivement plusieurs structures et de générer pour chacune d'entre elle un échantillon d'apprentissage et un ensemble de disons 100 échantillons tests. Comme dans l'exemple ci-dessus, nous considérerons plusieurs arbres de complexité variable pour lesquels nous calculerons le critère BIC. Chaque arbre sera ensuite appliqué sur chacun des échantillons test, et nous compareront la moyenne des taux d'erreur obtenus avec le critère BIC. Plus précisément nous comparerons l'évolution avec la complexité de ces deux indicateurs.

						taux d'erreur	
regroup.	model	p	-2LL	d	BIC	test	apprent.
	saturated	14	0	0	100	0.221	0.206
A,C | F,c1	m1	13	0.07	1	92.66	0.221	0.206
A,C | F,c3	m2	12	1.63	2	86.81	0.221	0.206
c3,c4 | M,A	m3	11	3.78	3	81.54	0.221	0.206
A,C | F,c2	m4	10	6.93	4	77.28	0.221	0.206
A,C | M,c1	m5	9	15.37	5	78.30	0.223	0.208
c2,c3c4 | M,A	m6	8	26.71	6	82.23	0.223	0.208
c2,c3 | M,C	m7	7	42.45	7	90.55	0.229	0.213
A,C | M,c2c3c4	m8	6	56.22	8	96.90	0.229	0.213
c1,c2c3c4 | M	m9	5	74.59	9	107.86	0.229	0.213
c1,c2,c3,c4 | F	m10	2	188.38	12	199.41	0.229	0.222
tout	indep	1	531.04	13	534.66	0.314	0.326

Tab. 3 – Qualité des modèles successifs

Arbre BIC optimal et taux d'erreur

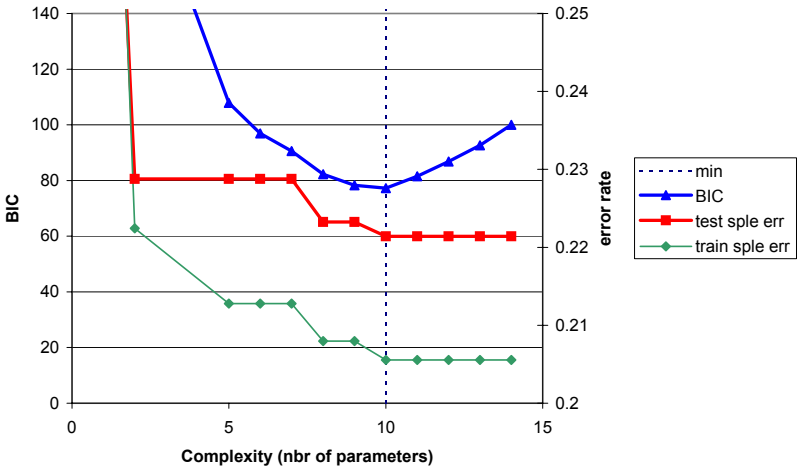


FIG. 2 – BIC sur échantillon d'apprentissage et taux d'erreur

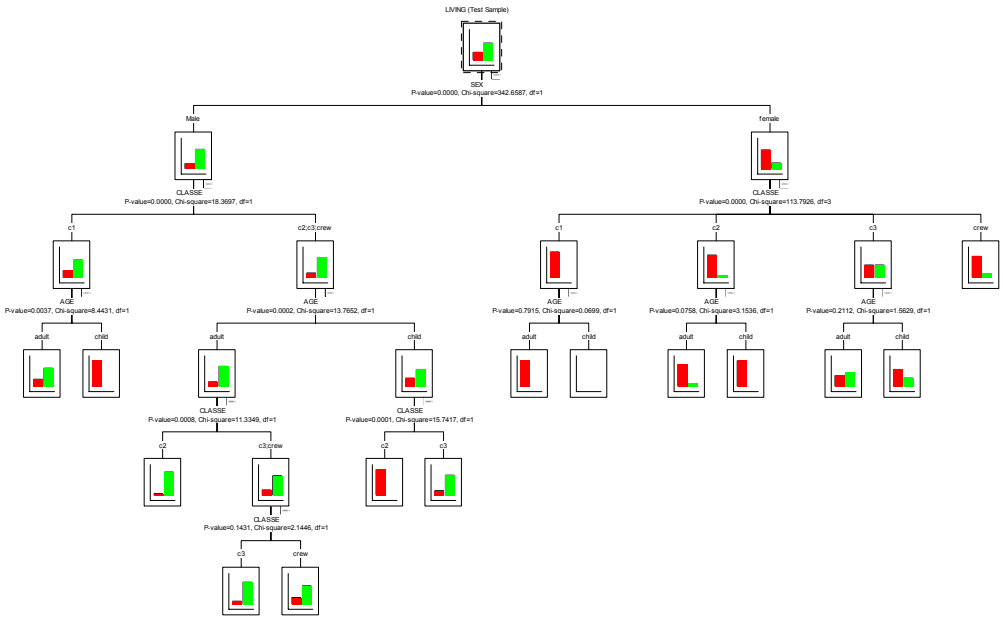


FIG. 3 – Arbre saturé, échantillon test, $n = 542$

4 Conclusion

Le critère BIC pour arbres a été introduit dans Ritschard et Zighed (2003, 2004) comme critère pour déterminer l'arbre le plus adéquat du point de vue de la description de données. Nous pensons cependant que ce critère peut également s'avérer utile dans

une optique de classification. L'illustration considérée semble confirmer notre conjecture selon laquelle le critère BIC devrait permettre de déterminer l'arbre le mieux adapté à une utilisation prédictive en dehors de l'échantillon d'apprentissage. Une expérimentation complète s'avère cependant nécessaire pour donner une assise empirique mieux fondée à cette conjecture.

Références

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC : The American Sociological Association.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics 11*(2), 416–431.

Ritschard, G. et D. A. Zighed (2003). Goodness-of-fit measures for induction trees. In N. Zhong, Z. Ras, S. Tsumo, et E. Suzuki (Eds.), *Foundations of Intelligent Systems, ISMIS03*, Volume LNAI 2871, pp. 57–64. Berlin : Springer.

Ritschard, G. et D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information E-1*, 45–67.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago : SPSS Inc.

Wallace, C. S. et P. R. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B (Methodological) 49*(3), 240–265.

Wehenkel, L. (1993). Decision tree pruning using an additive information quality measure. In B. Bouchon-Meunier, L. Valverde, et R. Yager (Eds.), *Uncertainty in Intelligent Systems*, pp. 397–411. Amsterdam : Elsevier - North Holland.

Summary

We discuss two aspects related to the scope of the BIC index for induction trees proposed in Ritschard et Zighed (2003, 2004). The first point is about how to compute it. We show that the BIC can easily be derived from the Likelihood Ratio Chi-square statistics used for testing the row-column independence of contingency tables. The second aspect is related to its interest for classification purposes. We illustrate, by means of the Titanic example, the expected link between the BIC and the generalization error rate in terms of their evolution with respect to the tree complexity. Finally, we sketch an experiment design for checking empirically the conjecture that the minimal BIC ensures on average the best generalization error rate.

Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication

Jérôme David*, Fabrice Guillet*,Vincent Philippé**, Henri Briand*, Régis Gras*

*LINA - Ecole Polytechnique de l'université de Nantes
La Chantrerie - BP 50609 - 44306 Nantes cedex 3
jerome.david,fabrice.guillet,regis.gras,henri.briand@polytech.univ-nantes.fr,
http ://www.sciences.univ-nantes.fr/lina/recherche/LEC/EDC/
**PerformanSe SAS - Atlanpole - La Fleuriaye - 44470 Carquefou
vincent.philippe@performanse.fr
http ://www.performanse.fr

Résumé. Dans le cadre d'une validation d'expertise textuelle contenue dans un test de compétences comportementales informatisé, nous proposons une méthode visant à extraire des sous-ensembles de termes caractéristiques utilisés pour décrire des traits de comportements. Notre approche consiste, après une phase de traitement automatique du langage (extraction de candidats termes), à évaluer les associations possibles entre termes et étiquettes qui structurent le corpus en s'appuyant sur la théorie de l'implication statistique.

1 Introduction.

Les documents sous forme de textes représentent aujourd'hui, dans notre société, des quantités d'information colossales. Afin d'accéder rapidement et de manière pertinente aux informations textuelles, des systèmes d'indexation performants permettent d'associer à un document un ensemble de caractères.

D'un autre coté, l'Extraction de Connaissances à partir de Textes (ECT) ou text-mining, vise à extraire des connaissances pertinentes, contenues dans des données textuelles, à l'aide des modèles utilisés en Extraction des Connaissances dans les Données (Kodratoff (2000)). Parmi les modèles utilisés en ECT, la découverte de règles d'associations entre termes contenus dans les textes est souvent utilisée (Maedche et Staab (2000), Janetzko *et al.* (2004), Roche (2003)).

La découverte de règles d'association (Agrawal *et al.* (1993)), consiste à trouver dans des bases de données, des tendances implicatives $a \Rightarrow b$ entre attributs booléens caractérisées par deux mesures : le support et la confiance. Parmi les indices alternatifs de qualité proposés dans la litérature (Tan *et al.* (2004), Guillet (2004), Lenca *et al.* (2004)), nous nous intéressons à la mesure
d'intensité d'implication définie par R. Gras (Gras (1979), Gras et al. (1996)) et son extension entropique (Gras *et al.* (2001)).

Cependant avant d'utiliser les techniques d'ECD, les données linguistiques doivent subir une phase de Traitement Automatique du Langage (TAL), dont le but est d'obtenir

à partir d'un texte, la liste des termes qu'il contient. De nombreuses approches sont proposées : approches statistiques (Salem (1986)), approches linguistiques (David et Plante (1990), Bourigault et Fabre (2000), Jacquemin (1997)), ou mixtes qui combinent les deux approches précédentes (Smadja (1993), Daille (1994)).

La démarche que nous proposons dans cet article s'inscrit à l'intersection des domaines de la recherche d'information et du text-mining. En effet, nous proposons une méthode d'étude et de validation d'une indexation par des profils psychologiques de documents traitant de bilans de compétences comportementales dans le cadre de la théorie de l'implication statistique. L'objectif de notre étude est d'associer à chaque caractère psychologique du profil, une classe de termes.

Nous présentons tout d'abord, les données et la problématique à partir desquelles nous avons construit notre approche. Ensuite, nous faisons un rappel sur l'intensité d'implication. Dans la deuxième partie, nous expliquons notre démarche d'évaluation des tendances implicatives à partir desquelles nous formons des groupes de termes associés aux caractères psychologiques. Finalement, nous présentons et analysons les résultats obtenus sur la base de textes étudiée.

2 Méthodologie.

2.1 Description des données analysées et de la problématique.

La base textuelle indexée à partir de laquelle nous avons conçu notre méthode est extraite du logiciel PerformanSe-DIALECHO, qui est un questionnaire de personnalité informatisé largement utilisé dans le domaine de la gestion des ressources humaines. Cet outil permet, à l'issue d'un QCM composé de 70 questions, de positionner la personne évaluée sur un profil psychologique composé de 10 caractères ayant chacun 3 modalités possibles (appelés caractères) et de lui restituer un bilan de compétences sous forme textuelle. Des caractères possibles sont, par exemple, l'extraversion (EXT+), introversion (EXT-), l'anxiété (ANX+) ou encore la détente (ANX-). Un bilan de compétence est ensuite généré à partir d'une base de textes, balisée par des règles de décision (composées de conjonction de caractères), écrites par le psychologue-concepteur de l'outil.

Dans le cadre de la validation de l'expertise textuelle contenue dans ce logiciel, l'expert veut vérifier l'adéquation du vocabulaire qu'il a utilisé pour décrire un ensemble de caractères du profil psychologique.

Notre problème consiste, à associer des termes (extraits du corpus de texte) à un (ou plusieurs) caractères dont la sémantique a été définie par l'expert. Dans notre cas, ces caractères indexent également la base de textes générant le bilan de compétences.

La base de textes analysée est composée de 12805 paragraphes. Chaque paragraphe, est indexé par un ou plusieurs caractères. La méthodologie que nous suivons est calquée sur le processus classique d'extraction des connaissances dans les données. En effet, une premiére phase de traitement et d'indexation terminologique (opération 1, figure 1), nous permet d'obtenir une représentation sous forme d'une base de données où chaque texte est représenté par les termes qui le composent. Une des particularités de notre méthode, consiste à ajouter à la représentation des textes les caractères issus

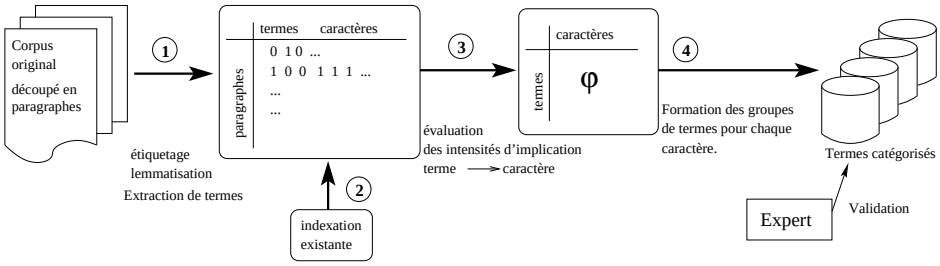


FIG. 1 – Chaîne de traitements.

d'une autre indexation existante (opération 2, figure 1). Ensuite, notre processus de fouille de données permet de former des modèles par association de termes venant de l'indexation terminologique à chaque caractère issu de l'indexation externe (opérations 3 et 4, figure 1).

2.2 Rappels sur l'analyse implicative.

Les règles d'association $(a \Rightarrow b)$ sont des tendances implicatives admettant des contre-exemples $(a\, et\, \bar{b})$ dont la qualité est évaluée par les indices de support et de confiance. Nous voulons évaluer des règles qui peuvent avoir un faible support mais qui restent toutefois significatives. Comme le mentionne Y. Kodratoff (Kodratoff (2001)), les règles les plus pertinentes sont souvent les plus rares. La confiance présente elle aussi des défauts comme le fait de ne pas rejeter l'indépendance (Blanchard *et al.* (2004)).

La mesure que nous avons retenue, l'intensité d'implication définie par R. Gras (Gras (1979)), quantifie l'étonnement que l'on peut avoir face à un nombre invraisemblablement petit de contre-exemples $card(A \cap \bar{B})$, où A (resp. B) sont les tuples de la relation r vérifiant a (resp. b). L'implication est évaluée sous l'hypothèse d'indépendance des variables a et b.

Pour modéliser une règle d'association, R. Gras propose, à l'instar de I.C. Lerman (Lerman (1981)), pour quantifier la similarité de comparer le nombre de contre-exemples $card(A \cap \bar{B})$ d'une règle $a \Rightarrow b$ par rapport à la variable aléatoire $card(X \cap \bar{Y})$ où X et Y sont deux parties (choisies de manière aléatoire et indépendante) d'un ensemble E de même cardinal que la relation r étudiée.

$$\varphi(a \Rightarrow b) = 1 - Pr\left[card(X \cap \bar{Y}) \leq card(A \cap \bar{B})\right]$$

Différents modèles sont possibles pour la variable aléatoire $card(X \cap \bar{Y})$: loi Hypergéométique, loi de Poisson, loi Binomiale ou Normale. Dans le cadre de notre étude, nous nous intéressons à des cas rares par rapport la taille de la population étudiée, ainsi nous choisissons de modéliser la variable aléatoire $card(X \cap \bar{Y})$ par une loi de Poisson.

id_doc	conscience professionnelle	sens de la méthode	preuve de creativité	attrait de la nouveauté
d1	1	1	0	0
d2	0	0	1	1
id_doc	Extraversion	Extraversion moyenne	Rigueur	Dynamisme intellectuel
d1	0	1	1	0
d2	1	0	0	1

TAB. 1 – Extrait de la table $\mathcal{D}$ représentant les paragraphes.

3 Regroupement des termes les plus représentatifs d'un caractère.

3.1 Principes de l'étude.

Nous pouvons maintenant définir la base de textes étudiée par un triplet $B = (D, T, C)$ où $D = \{d_1, .., d_m\}$ dénote l'ensemble des paragraphes, $T = \{t_1, ..., t_n\}$ l'ensemble des termes et $C = \{c_1, ..., c_y\}$ l'ensemble des caractères. Nous représentons les paragraphes par la table relationnelle $\mathcal{D}$, où chaque n-uplet représente les valeurs prises par un paragraphe d_x sur l'ensemble des attributs $A = C \cup T$. Pour un paragraphe d_x donné, un attribut a prend comme valeur 1 si l'attribut a qualifie le paragraphe d_x, 0 sinon.

L'exemple suivant (table 1) présente les valeurs prises par les paragraphes "d1" et "d2" sur l'ensemble des termes, "conscience professionnelle", "sens de la méthode", "preuve de creativité", "attrait de la nouveauté"et l'ensemble des caractères, "Extraversion", "Extraversion moyenne", "Rigueur", "Dynamisme intellectuel" :

Prenons le cas où l'expert s'intéresse aux groupes de termes décrivant au mieux un seul caractère. Nous évaluons à partir de la table $\mathcal{D}$, pour chaque terme $t_i \in T$ et pour chaque caractère $c_j \in C$, l'implication $t_i \Rightarrow c_j$ signifiant : "Si un paragraphe contient le terme t_i alors ce paragraphe est destiné à un individu possédant (entre autres) le caractère c_j". Contrairement aux recherches classiques de règles d'association, nous évaluons toutes les associations binaires possibles entre deux ensembles d'attributs disjoints. Ainsi, nous n'utilisons pas le support contrairement à l'algorithme Apriori (Agrawal et Srikant (1994)). Afin d'évaluer les associations, nous définissons donc, la matrice $\mathcal{M}_\varphi$ d'ordre $n \times m$ croisant les n termes et les m caractères où chaque valeur $\varphi_{ij} = \begin{cases} \varphi(t_i \Rightarrow c_j) \, si \, \varphi(t_i \Rightarrow c_j) \geq 0 \\ 0 \, sinon \end{cases}$.

La table 2 donne pour quelques termes ("conscience professionnelle", "preuve de créativité", ...) leur intensité d'implication envers les caractères ("Extraversion", "Extraversion moyenne", "Rigueur", "Dynamisme intellectuel").

Finalement, l'ensemble de termes les plus représentatifs d'un caractère c_x au seuil φ_{seuil} est défini de la manière suivante : $T_x = \{t_y \mid \varphi(t_y \Rightarrow c_x) \geq \varphi_{seuil}\}$.

$t \Rightarrow c$	Extraversion	Extraversion moyenne	Rigueur	Dynamisme intellectuel
conscience professionnelle	0.0	0.63	0.99	0.0
sens de la méthode	0.77	0.0	0.92	0.0
preuve de créativité	0.0	0.0	0.0	0.94
attrait de la nouveauté	0.0	0.0	0.0	0.94
domaine de la communication	0.0	0.0	0.86	0.86

TAB. 2 – Extrait de la matrice $\mathcal{M}_\varphi$ d'intensités d'implication.

Le choix du seuil φ_{seuil} est délicat car il dépend de la base de textes étudiée. Nous proposons donc fixer dans un premier temps $\varphi_{seuil} > 0,5$, car c'est le seuil à partir duquel une règle commence à être significative. Ensuite l'expert pourra l'augmenter jusqu'à satisfaire son critère de sélection : par exemple, le nombre de termes par classes.

Prenons l'exemple de la table 2 : en choisissant $\varphi_{seuil} > 0,5$, nous obtenons pour le caractère "Rigueur", la classe de termes {"conscience professionnelle", "sens de la méthode", "domaine de la communication"}. De la même manière, la classe représentative du caractère "Dynamisme intellectuel" sera constituée des termes {"preuve de créativité", "attrait de la nouveauté", "domaine de la communication"}. Nous pouvons noter que les classes ainsi formées admettent une intersection non nulle : un terme peut appartenir à plusieurs classes.

3.2 Résultats.

Pour chacun des trente caractères auxquels nous voulions associer les termes les plus descriptifs, l'expert-auteur des textes a évalué la pertinence des ensembles ainsi créés. Chaque classe de termes associée à un caractère a été scindée en 2 groupes par l'expert : les termes en adéquation avec le caractère et les autres. La précision est déduite de ce classement comme la proportion de termes bien classés. Le tableau 3 donne pour quelques classes de termes associés à un caractère, la précision pour une sélection des règles ayant $\varphi_{seuil} > 0.5$.

Nous pouvons observer dans ces résultats, qu'il y a des caractères pour lesquels, la précision est mauvaise (en particulier les caractères "E-", "E+" et "E0"). Les mauvais résultats sur certains ensembles de termes sont dus à la manière dont l'expert a rédigé son corpus : en effet, des caractères comme "E-", "E+" et "E0" sont trés peu décrits dans les textes mais servent à nuancer la description des autres caractères étudiés. Cependant, nous avons obtenu de très bon résultats sur un bon nombre de caractères. En effet, nous avons 8 caractères pour lesquels la recherche est de bonne précision (supérieure ou égale à 90%) contre 3 caractères pour lequels les résultats sont mauvais (précision inférieure ou égale à 50%).

Classe	Précision
Rigueur (CON+)	1
Combativité (P+)	0.9
Anxiété (N+)	0.9
Dynamisme intellectuel (CLV+)	0.9
Affirmation (EST+)	0.9
Remise en cause (EST-)	0.9
Motivation de pouvoir (LED+)	0.9
Motivation de protection (LED-)	0.9
Détente (N-)	0.8
Improvisation (CON-)	0.8

Classe	Précision
Motivation d'appartenance (AFL+)	0.8
Conciliation (P-)	0.7
Motivation d'indépendance (AFL-)	0.7
Anxiété moyenne (N0)	0.6
Conformisme intellectuel (CLV-)	0.6
Introversion (E-)	0.5
Extraversion (E+)	0.4
Extraversion moyenne (E0)	0

TAB. 3 – Précisions des regroupements données par l'expert.

4 Conclusion

Nous avons présenté une approche visant à étudier et valider l'adéquation entre des termes contenus dans une base de textes et l'ensemble des caractères psychologiques indexant les paragraphes du corpus textuel. Cette méthode, divisée en deux phases (extraction et sélection des termes, formation de groupes de termes par association des termes aux caractères) permet d'obtenir pour chacun des caractères étudiés une classe de termes significatifs. L'originalité de notre approche réside dans le fait qu'elle permet de créer des rapprochements entre une indexation quelconque d'une base de textes (automatique/manuelle, ontologique...) et des termes extraits du corpus. Cela permet donc, à un expert du domaine, d'étudier et de valider la sémantique voire d'enrichir une indexation d'une base de textes.

Un prototype a été développé et appliqué au jeu de données présenté dans l'article. Les résultats sont encourageants : en effet nous avons obtenu une bonne précision moyenne des regroupements de termes, et ces derniers ont permis à l'expert d'adapter son discours en fonction du type d'individu concerné.

Actuellement, nous ne nous intéressons qu'à des caractères non structurés c'est-à-dire que l'on ne prend pas en compte les relations qui peuvent exister entre les différents caractères ou entre les termes eux-mêmes. Nous comptons donc étendre notre approche afin qu'elle puisse s'appliquer à des ontologies en prenant en compte cette dimension structurelle.

Références

R. Agrawal, T. Imielinski, et A.N. Swami. Mining association rules between sets of items in large databases. In Buneman P. et Jajodia S., editors, *Proceedings of the*

1993 ACM SIGMOD ICMD, pages 207–216, 1993.

R. Agrawal et R. Srikant. Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke, et C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.

J. Blanchard, F. Guillet, R. Gras, et H. Briand. Mesurer la qualité des règles et de leur contraposées avec le taux informationnel tic. *RNTI E-2 Extraction et gestion des connaissances*, 1 :287–298, 2004.

D. Bourigault et C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, 25 :131–151, 2000.

B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat, University Paris 7, 1994.

S. David et P. Plante. De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *ICO*, 2(3) :140–155, 1990.

R. Gras, P. Kuntz, R. Couturier, et F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *ECA Extraction et Gestion de Connaissances*, 1(1–2) :69–80, 2001.

R. Gras et al. *L'implication statistique, une nouvelle méthode exploratoire de données.* La pensée sauvage, 1996.

R. Gras. Contribution à l'étude experimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didacticques mathématiques, 1979. Thèse d'Etat, Université de Rennes.

F. Guillet. Mesure de la qualité des connaissances en ecd. In *Tuturiels de la 4ème Conf. Francophone d'extraction et gestion des connaissances*, pages 1–60, Clermond-Ferrand, 2004.

C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes, 1997. Mémoire d'HDR, IRIN - Université de Nantes.

D. Janetzko, H. Cherfi, R. Kennke, A. Napoli, et Y. Toussaint. Knowledge-based selection of association rules for text mining. In *ECAI'04*, pages 485–489. IOS Press, 2004.

Y. Kodratoff. Datamining and textmining. In *EGC'2000*, pages 6–9, 2000.

Y. Kodratoff. On the induction of interesting rules. *Noesis*, XXVI :103–124, 2001.

P. Lenca, P. Meyer, et B. Vaillant. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1 Mesures de qualité pour la fouille de données*, pages 219–246, 2004.

I.C. Lerman. *Classification et analyse ordinale des données.* Dunod, Paris, 1981.

A. Maedche et S. Staab. Semi-automatic engineering of ontologies from text. In KSI, editor, *the 12th Internationnal Conference SEKE*, 2000.

M. Roche. L'extraction paramétrée de la terminologie du domaine. *RSTI Extraction et Gestion des Connaissances*, 17 :295–306, 2003.

A. Salem. Segments répétés et analyse statistique des données textuelles.étude quantitative à propos du père duchesne de hébert. *Histoire et Mesure*, 1(2) :5–28, 1986.

F. Smadja. Retrieving collocations from text : Xtract. *Computational linguistics*, 19 :143–177, 1993.

P.N Tan, V. Kumar, et J. Srivastava. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4) :293–313, 2004.

Summary

In order to validate a textual base contained in a behavioural skill testing software, we suggest a methodology which can extract subsets of caracteristic terms used to describe personality traits. Our approach permits, after an automatic language processing task, to evaluate the association rules between terms and descriptors (personality traits) which structure the corpus with the help of the theory of the statistic implication. In this way, we suggest to study the inclusions between groups of terms with the cohesitive hierarchy.

ARQAT : une plateforme d'analyse exploratoire pour la qualité des règles d'association

Xuan-Hiep Huynh*, Fabrice Guillet*, Henri Briand*
*LINA CNRS FRE 2729 - Ecole polytechnique de l'université de Nantes
La Chantrerie, BP 50609, 44306 Nantes Cedex 3, France
{xuan-hiep.huynh, fabrice.guillet, henri.briand}@polytech.univ-nantes.fr

Résumé. Le choix de mesures d'intérêt pour la validation des règles d'association constitue un défi important dans le contexte de l'évaluation de la qualité en fouille de données. De nombreuses mesures d'intérêt sont disponibles dans la littérature, et de nombreux auteurs ont discuté et comparé leurs propriétés dans ce but. Mais, comme l'intérêt dépend à la fois de la structure des données et des buts de l'utilisateur (décideur, analyste), certaines mesures peuvent s'avérer pertinentes dans un contexte donné, et ne plus l'être dans un autre. Par conséquent, il est nécessaire de concevoir de nouvelles approches contextuelles pour guider l'utilisateur dans son choix. Dans cet article, nous proposons un outil original ARQAT afin d'étudier le comportement spécifique de 34 mesures d'intérêt dans le contexte d'un jeu de règles, selon une approche résolument exploratoire mettant en avant l'interactivité et les représentations graphiques. L'outil ARQAT implémente 14 vues graphiques complémentaires structurées en 5 tâches d'analyses. Une partie de ces vues est décrite et illustrée sur un même jeu de 120000 règles issues de la base mushroom (MLrepository), afin de montrer l'intérêt de cet outil exploratoire et de la complémentarité de ses vues.

1 Introduction

L'étude et la conception de mesures d'intérêt (MI) adaptées aux règles d'association constitue un important défi pour l'évaluation de la qualité des connaissances en ECD. Les règles d'association [Agrawal *et al.*, 1993] proposent un modèle non supervisé pour la découverte de tendances implicatives dans les données. Malheureusement, en phase de validation, l'utilisateur (expert des données, ou analyste) se trouve confronté à un problème majeur : une grande quantité de règles parmi lesquelles il doit isoler les meilleures en fonction de ses préférences. Une manière de réduire le coût cognitif de cette tâche consiste à le guider à l'aide de mesures d'intérêt adaptées à la fois à ses préférences et à la structure des données étudiées.

Les travaux précurseurs sur les règles d'association [Agrawal *et al.*, 1993] [Agrawal and Srikant, 1994] proposent l'utilisation de 2 mesures statistiques : le support et la confiance. Ce couple de mesures dispose de vertus algorithmiques accélératrices, mais n'est pas suffisant pour capter l'intérêt des règles. Afin de compenser cette limite, de nombreuses mesures complémentaires ont été proposées dans la littérature. Étant donné que l'intérêt dépend à la fois des préférences de l'utilisateur et des données,

les MI peuvent être dissociées en 2 groupes [Freitas, 1999] : les mesures objectives et les mesures subjectives. Les mesures subjectives dépendent essentiellement des buts, connaissances, croyances de l'utilisateur qui doivent être préalablement recueillis. Elles sont associées à des algorithmes supervisés ad hoc [Padmanabhan and Tuzhilin, 1998] [Liu *et al.*, 1999] permettant de n'extraire que les règles conformes ou au contraire en contradiction avec les croyances de l'utilisateur, et ainsi d'orienter la notion d'intérêt vers la nouveauté (novelty) ou l'inattendu (unexpectedness). Les mesures objectives, quant à elles, sont des mesures statistiques s'appuyant sur la structure des données ou plus exactement la fréquence des combinaisons fréquentes d'attributs (itemsets). De nombreux travaux de synthèse récapitulent et comparent leurs définitions et leurs propriétés (voir [Bayardo and Agrawal, 1999], [Hilderman and Hamilton, 2001], [Tan *et al.*, 2002], [Tan *et al.*, 2004], [Piatetsky-Shapiro, 1991], [Lenca *et al.*, 2004], [Guillet, 2004]). Ces synthèses traitent deux problèmes fondamentaux et complémentaires afin d'aider l'utilisateur à repérer les meilleures règles : la caractérisation des principes sous-jacents à une "bonne" MI, et l'étude comparative de leur comportement sur des simulations et des jeux d'essai. Dans cette optique, [Vaillant *et al.*, 2003] proposent un premier outil d'expérimentation : HERBS.

Dans cet article, nous présentons une nouvelle approche et une plateforme d'implémentation ARQAT (Association Rule Quality Analysis Tool) afin d'étudier le comportement spécifique des MI sur le jeu de données de l'utilisateur et selon une perspective d'analyse exploratoire.

Plus précisément, ARQAT est une boite à outil conçue pour aider graphiquement l'utilisateur analyste à repérer dans ses données les meilleures mesures et au final les meilleures règles.

Dans une première partie nous présentons la structure et les principes de la plateforme ARQAT. Puis dans les parties suivantes nous ciblons la présentation sur 3 tâches complémentaires munies de représentations graphiques : les statistiques élémentaires sur le jeu de règles, l'analyse de corrélation, et enfin l'aide au choix des meilleures règles. Chacune de ces fonctionnalités est illustrée sur le même jeu de 120000 règles (issu de la base Mushroom MLrepository), afin de montrer l'intérêt de l'approche exploratoire s'appuyant sur un ensemble de vues complémentaires que nous proposons.

2 Principes de la plateforme ARQAT

ARQAT inclut 34 mesures objectives issues des articles de synthèse précédents. Nous complétons cette liste avec 3 mesures : l'Intensité d'Implication (II) [Gras, 1996] [Guillaume *et al.*, 1998], sa version entropique (EII) [Gras *et al.*, 2001] [Blanchard *et al.*, 2003], et la mesure de taux informationnel modulé par la contraposée (TIC) [Blanchard *et al.*, 2004] (cf Annexe 1 pour une liste récapitulative).

ARQAT (Fig. 1) implémente 14 vues graphiques complémentaires qui sont structurées en 5 groupes selon la tâche offerte.

Les données d'entrée sont constituées d'un ensemble R de règles d'association extrait d'un jeu de données initial D, où la description de chaque règle $a \Rightarrow b$ est complétée par ses contingences $(n, n_a, n_b, n_{a\bar{b}})$ dans D. Plus précisément, n est le nombre total d'enregistrements de D, n_a (resp. n_b) le nombre d'enregistrements de D satisfaisant a

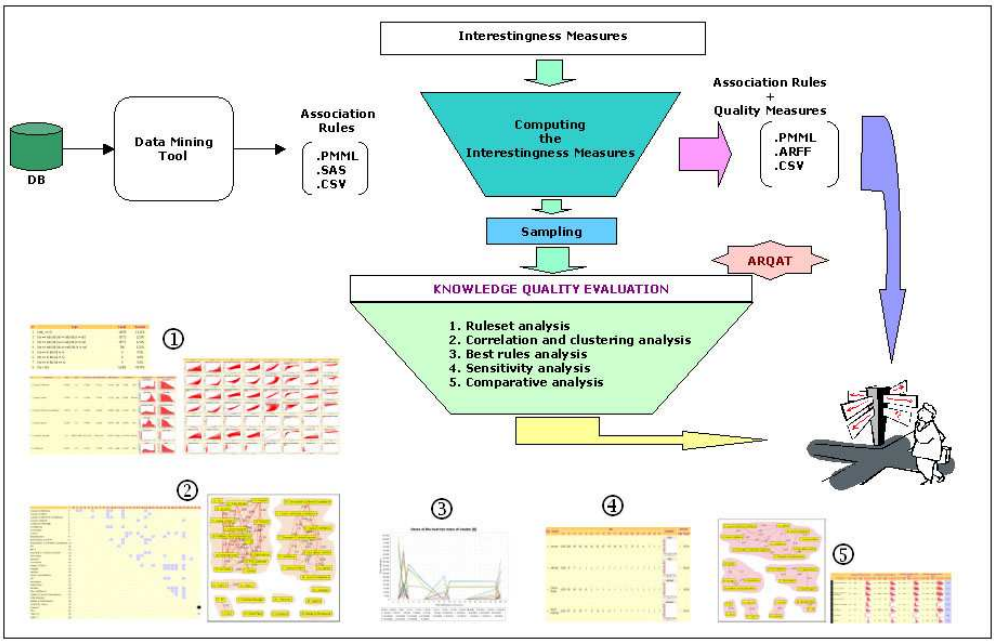


FIG. 1 – Structure d'ARQAT.

(resp. b), et $n_{a\bar{b}}$ le nombre d'enregistrements satisfaisant $a \wedge \bar{b}$ (les contre-exemples).

Dans une étape préliminaire, l'ensemble de règles R est traité afin de calculer les valeurs des MI pour chaque règle, puis les corrélations entre chaque paire de mesure. Les résultats sont stockés dans deux tables : la table des mesures (R×I) dont les lignes corespondent aux règles et les colonnes aux valeurs des mesures, et la matrice de corrélation (I×I) entre les mesures. Lors de cette étape, l'ensemble de règles R peut aussi être échantillonné afin de cibler l'étude sur un sous-ensemble de règles.

La seconde étape est ensuite interactive, l'utilisateur mène l'exploration graphique des résultats. Il s'appuie pour cela sur la structuration en 5 groupes de vues orientées tâche. Le premier groupe (1 dans Fig. 1) est dédié à la visualisation de statistiques élémentaires afin de mieux appréhender la structure de la table R×I. Le deuxième groupe (2) est orienté vers la visualisation de la table des corrélations entre mesures I×I et leur classification afin de repérer les meilleures mesures. Le troisième groupe (3) cible l'extraction des meilleures règles. Le quatrième groupe (4) permet une étude de la sélectivité des MI. Enfin, un dernier groupe offre la possibilité de mener une étude comparative des résultats obtenus sur plusieurs ensembles de règles.

Dans la suite de cet article, nous décrivons les trois premiers groupes et les illustrons sur un même jeu de règles : 120000 règles d'association extraites par un algorithme Apriori (support 10%) de la base mushroom [Blake and Merz, 1998].

N°	Type	Count	Percent
1	(nab_ == 0)	16158	13.11%
2	(na == nab) && (nb != nab) && (n != nb)	15772	12.8%
3	(nb == nab) && (na != nab) && (n != na)	15772	12.8%
4	(na == nab) && (nb == nab) && (n != na)	386	0.31%
5	(na == n) && (nb != n)	0	0.0%
6	(nb == n) && (na != n)	0	0.0%
7	(na == n) && (nb == n)	0	0.0%
8	(na > nb)	61355	49.79%

FIG. 2 – Quelques caractéristiques des règles issues de la base mushroom.

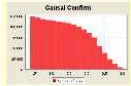
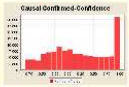
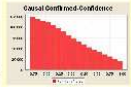
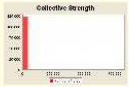
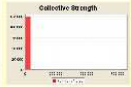
N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Inverse cumulative
0	Causal Confidence	0.0614	1.0	0.6246	0.2712	-0.2033	Left	-1.0001	Flat		
1	Causal Confirm	-1.5951	1.0	0.1084	0.5744	-1.1203	Left	0.8805	Peaked		
2	Causal Confirmed-Confidence	-0.8169	1.0	0.2018	0.5448	0.0286	Right	-1.2036	Flat		
3	Causal Support	0.1198	1.0	0.5542	0.2051	-0.0995	Left	-0.5643	Flat		
4	Collective Strength	0.0	785856.8441	2328.2323	7944.9114	30.5993	Right	1932.4436	Peaked		
5	Confidence	0.1217	1.0	0.5772	0.2799	0.194	Right	-1.3159	Flat		

FIG. 3 – Distributions de quelques mesures sur la base mushroom.

3 Statistiques sur le jeu de règles

Ce premier groupe offre trois vues graphiques résumant les caractéristiques statistiques du jeu de règles étudié. La première vue (Fig. 2) récapitule la *distribution des contingences* sous-jacente aux règles, et facilite la détection des cas limites. Par exemple, la première ligne décrit le nombre de règles "logiques" (i.e. sans contre-exemple $n_{a\bar{b}} = 0$, ou encore avec une confiance à 100%).

La deuxième vue, *distribution des mesures* (Fig. 3), présente l'histogramme de chaque mesure, en le complétant de divers indicateurs (minimum, maximum, écart-type, ...). On peut par exemple y observer que la Confiance (ligne 5) possède une distribution très irrégulière et un grand nombre de règles logiques (à 100%) ; alors que la mesure Causal Confirm (ligne 2) montre une distribution très différente.

En complément, la troisième vue (Fig. 4) montre les *distributions croisées* des couples de mesures, récapitulées dans une représentation graphique matricielle très utile pour visualiser la forme de la liaison existant entre deux mesures. Par exemple, la Fig. 4 permet d'observer 4 différentes formes de non liaison : Rule Interest vs Yule's

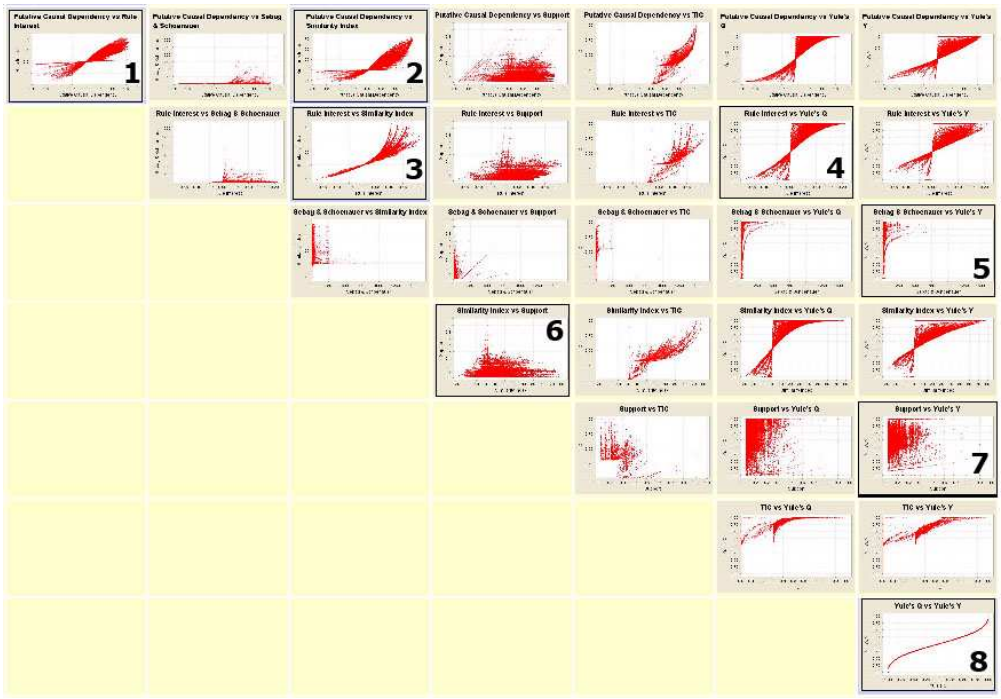


FIG. 4 – Matrice de quelques distributions croisées sur la base mushroom.

Q (4), Sebag & Schoenauer vs Yule's Y (5), Similarity Index vs Support (6), et Yule's Y vs Support (7), cette dernière révélant une forte indépendance. A l'opposé, les cellules Putative Causal Dependency vs Rule Interest (1), Putative Causal Dependency vs Similarity Index (2), Rule Interest vs Similarity Index (3), et Yule's Q vs Yule's Y (8), montrent des formes de liaison forte dont la dernière révèle une dépendance fonctionnelle.

4 Analyse de corrélation

Ce deuxième groupe de vues est orienté vers la tâche d'analyse des corrélations (matrice I × I) entre mesures et leur partitionnement en groupes corrélés, afin d'orienter l'utilisateur vers les mesures les mieux adaptées à ses besoins spécifiques au jeu de règles étudié. Les valeurs de corrélation sont calculées à titre provisoire selon la formule du coefficient de corrélation linéaire de Pearson. Les résultats sont présentés sous deux formes graphiques. La première (Fig. 5) est une visualisation élémentaire de la matrice de corrélation sous la forme d'une *matrice de niveau de gris*, où chaque valeur de corrélation est codée par un niveau de gris. Par exemple (Fig. 5), la cellule noire met en évidence une non corrélation significative entre les mesures Yule's Y et Support, et les 74 cellules grises correspondent à des corrélations significatives.

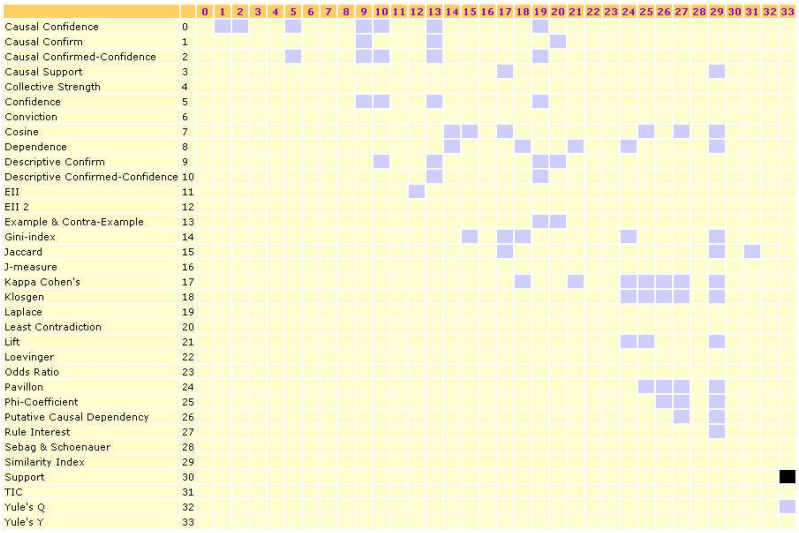


FIG. 5 – Matrice de correlation codée par niveau de gris sur la base mushroom.

La deuxième représentation envisagée, beaucoup plus expressive, est un *graphe de corrélation* (Fig. 6). Comme les graphes constituent un excellent outil d'investigation des structures complexes, nous les utilisons afin de représenter la matrice de corrélation sous la forme d'un graphe non-orienté et valué. Chaque sommet correspond à une MI, et une arête est associée à la valeur du coefficient de corrélation entre les deux sommets reliés. Nous y ajoutons une possibilité de seuillage par une valeur d'arête minimale τ (resp. maximale θ) afin de ne retenir que le sous graphe partiel CG+ (resp. CG0) des corrélations significativement élévées (resp. significativement faibles).

Ces deux sous-graphes partiels peuvent ensuite être traités afin d'être découpés en classes de mesures, chaque classe correspondant ici à une partie connexe maximale (nous orienterons ultérieurement vers un partitionnement en cliques). Dans CG+ chaque classe rassemble des mesures significativement corrélées proposant donc un point de vue proche sur les règles, alors que dans CG0 chaque groupe est révélateur de points de vue différents.

Ainsi, chaque jeu de règle produira un couple de graphes CG0 et CG+ différent, grâce auxquels l'utilisateur pourra observer rapidement la structure des MI, et valider graphiquement son choix des meilleurs indices. Par exemple, Fig. 6, CG+ fait apparaître 11 parties connexes qui peuvent aider à choisir une base réduite de 11 mesures, parmi les 34 utilisées, composée du meilleur représentant de chaque classe, afin de simplifier la validation des règles. Autre exemple, sur CG0 on voit une partie connexe composée des deux mesures Support and Yule's Y significativement non corrélées. Ce phénomène avait déjà été révélé par la matrice des niveau de gris (Fig. 5), et peut aussi être recoupé avec les distributions croisées de la Fig. 4 cellule (7). Un dernier exemple, sur le graphe CG+ apparaît une classe triviale associant les 2 mesures Yule's Q et Yule's Y comme fortement corrélées ; ce qui se retrouve sur la figure (Fig. 4 cellule (8)) montrant une dépendance fonctionnelle.

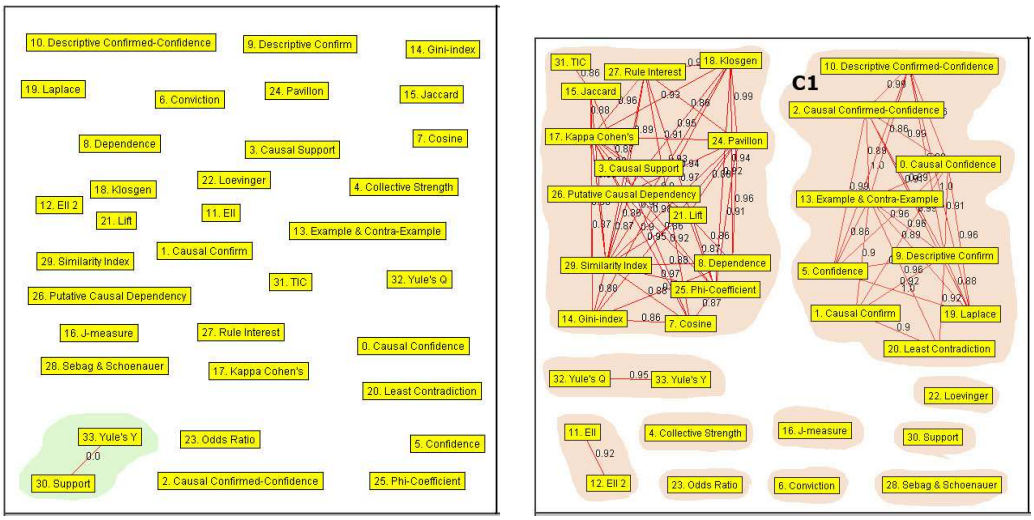


FIG. 6 – graphes de corrélation CG0 et CG+ sur la base mushroom (classes indiquées sur fond grisé).

Ces deux exemples illustrent l'intérêt d'utiliser conjointement les différentes vues complémentaires d'ARQAT. Ici, la matrice des distributions croisées (Fig. 4) permet d'évaluer la nature des liens corrélatifs portés dans les graphes CG0 et CG+, afin de contourner les limites du coefficient de corrélation linéaire.

5 Analyse des meilleures règles

Deux vues ont été spécifiquement implémentées pour guider l'utilisateur vers l'interprétation des meilleures règles. La première vue (Fig. 7) rassemble les *n meilleures règles* d'une classe (partie connexe du graphe de corrélation), celles qui sont bien évaluées par au moins une mesure de cette classe. Les règles choisies y sont visualisées par une *représentation en coordonnées parallèles* (Fig. 8) qui permet une interprétation rapide des mesures de chaque règle et de leur variation. Alternativement, les règles peuvent aussi y être représentée par leur rang (ordre de classement de la règle selon la valeur de la mesure). Par exemple, les figures 7 et 8 présentent les dix meilleures règles de la classe C1 (voir Fig. 6) selon une analyse de rang. On y observe des points de concentration sur les rangs faibles (notés 1, 2, 3 sur Fig. 8) correspondant respectivement aux trois mesures Confidence(5), Descriptive Confirmed-Confidence(10), et Example & Contra-Example(13) qui permettent une évaluation discriminante des meilleures règles. Cette observation se retrouve dans les colonnes 5, 10, 13 de la Fig. 7.

Measure Order		0	1	2	5	9	10	13	19	20	Rule's presentation
21	R107560	1	19121	1	1	41	1	1	8	5388	BROAD FREE ONE ==>veil_color=WHITE
22	R107562	1	18997	1	1	41	1	1	8	5361	BROAD ONE veil_color=WHITE ==>FREE
23	R107594	1	8972	1	1	18	1	1	3	2574	CLOSE FREE ONE ==>veil_color=WHITE
24	R107596	1	8914	1	1	18	1	1	3	2564	CLOSE ONE veil_color=WHITE ==>FREE
25	R122275	1	13800	1	1	32	1	1	5	3977	BROAD FREE ==>veil_color=WHITE
26	R122283	1	18299	1	1	38	1	1	6	5145	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
27	R122285	1	18179	1	1	38	1	1	6	5134	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
28	R122296	1	20903	1	1	55	1	1	10	6193	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
29	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	FREE ==>ONE veil_color=WHITE

FIG. 7 – Récapitulatif des 10 meilleures règles de la classe C1 et de leurs rangs sur la base mushroom (extrait).

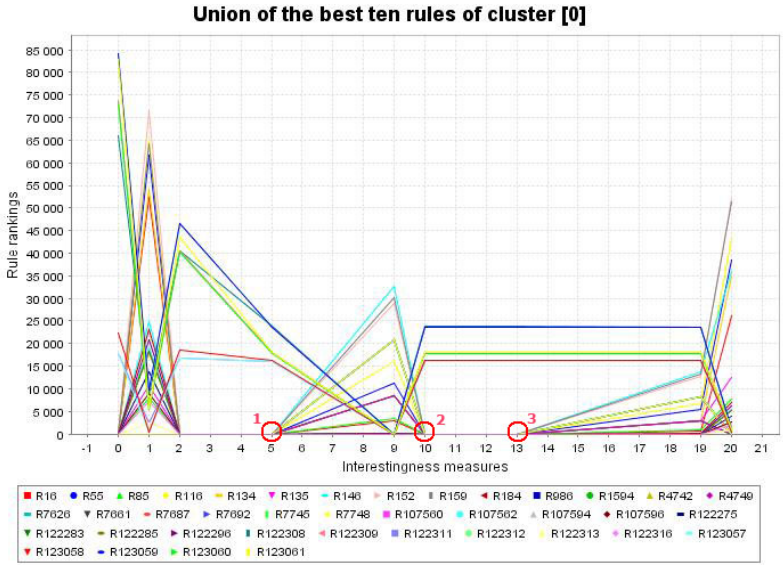


FIG. 8 – Coordonnées parallèles des rangs des 10 meilleures règles de la classe C1 sur la base mushroom.

6 Conclusion

Afin d'ouvrir les possibilités d'analyse expérimentales du comportement des mesures d'intérêt sur un jeu de règle spécifique, nous avons conçu et implémenté un outil informatique spécifique, ARQAT, suivant une approche résolument exploratoire, dont nous avons décrit la structure, une partie des 14 vues graphiques, et 3 des 5 tâches offertes.

D'un point de vue technique, ARQAT est écrit en Java, offre une interface de visualisation interactive portable à travers un navigateur web et implémente pour l'instant 34 mesures. Afin de faciliter les échanges avec les logiciels externes, ARQAT supporte 3 formats de fichiers standard pour importer/exporter un jeu de règles : PMML (XML data-mining standard), CSV (Excel et SAS) and ARFF (format WEKA). ARQAT sera téléchargeable gratuitement sur la toile à partir de l'adresse www.polytech.univ-nantes.fr/arqat.

Dans cet article, nous avons souhaité montrer sur des illustrations l'intérêt de notre approche exploratoire, où l'organisation en tâches, l'usage intensif de représentations graphiques, et de leur complémentarité, améliore et facilite l'analyse des mesures d'intérêt par la communauté scientifique.

ARQAT constitue un premier pas vers une plateforme plus complète dediée à l'évaluation de la qualité en fouille de données. Nous allons poursuivre nos travaux selon deux directions. En premier lieu, nous souhaitons améliorer l'analyse de corrélation en quittant le coefficient de corrélation linéaire pour adopter un coefficient plus performant. La deuxième perspective concerne l'amélioration de la classification des mesures en utilisant un opérateur d'agrégation dirigé par les préférences de l'utilisateur afin d'améliorer sa prise de décision pour les sélection des meilleures mesures.

Références

[Agrawal and Srikant, 1994] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[Bayardo and Agrawal, 1999] Jr.R.J. Bayardo and R. Agrawal. Mining the most interestingness rules. In *Proceedings of the Fifth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*, pages 145–154, 1999.

[Blake and Merz, 1998] C.L. Blake and C.J. Merz. *UCI Repository of machine learning databases, http ://www.ics.uci.edu/∼mlearn/MLRepository.html.* University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Blanchard *et al.*, 2003] J. Blanchard, P. Kuntz, F. Guillet, and Gras R. Implication intensity : from the basic statistical definition to the entropic version. In *Statistical Data Mining and Knowledge Discovery*, pages 475–493, 2003.

[Blanchard *et al.*, 2004] J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposés avec le taux informationnel tic. In *Revue Nationale des Technologies de l'Information (RNTI)*, pages 287–298, 2004.

[Freitas, 1999] A.A. Freitas. On rule interestingness measures. In *Knowledge-Based Systems*, pages 309–315, 1999.

[Gras *et al.*, 2001] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *Extraction des Connaissances et Apprentissage (ECA)*, pages 69–80, 2001.

[Gras, 1996] R. Gras. *L'implication statistique - Nouvelle méthode exploratoire de données.* La pensée sauvage édition, 1996.

[Guillaume *et al.*, 1998] S. Guillaume, F. Guillet, and J. Philippé. Improving the discovery of association rules with intensity of implication. In Lecture Notes in Compuper Science, editor, *Proceedings of 2nd European Symp. on Principles of Data Mining and Knowledge Discovery, PKDD'98*, pages 318–327, 1998.

[Guillet, 2004] F. Guillet. Mesures de la qualité des connaissances en ecd. In *Actes des tutoriels, 4ème Conférence francophone Extraction et Gestion des Connaissances (EGC'2004), http ://www.isima.fr/ egc2004/*, pages 1–60, 2004.

[Hilderman and Hamilton, 2001] R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interestingness*. Kluwer Academic Publishers, 2001.

[Lenca *et al.*, 2004] P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. In *Revue des Nouvelles Technologies de l'Information - Mesures de Qualité pour la Fouille de Données, RNTI-E-1*, pages 219–246, 2004.

[Liu *et al.*, 1999] B. Liu, W. Hsu, L. Mun, and H. Lee. Finding interestingness patterns using user expectations. In *IEEE Transactions on knowledge and data mining 11(1999)*, pages 817–832, 1999.

[Padmanabhan and Tuzhilin, 1998] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th international conference on knowledge discovery and data mining*, pages 94–100, 1998.

[Piatetsky-Shapiro, 1991] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248, 1991.

[Tan *et al.*, 2002] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 32–41, 2002.

[Tan *et al.*, 2004] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. In *Information Systems 29(4)*, pages 293–313, 2004.

[Vaillant *et al.*, 2003] B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In R. Bisdorff, editor, *Human Centered Processes (HCP'2003)*, pages 187–191, 2003.

Annexe 1 : Mesures d'intérêts utilisées

N°	Interestingness Measure	$f(n, n_a, n_b, n_{a\bar{b}})$
0	Causal Confidence	$1 - \frac{1}{2}(\frac{1}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$
1	Causal Confirm	$\frac{n_a + n_{\bar{b}} - 4n_{a\bar{b}}}{n}$
2	Causal Confirmed-Confidence	$1 - \frac{1}{2}(\frac{3}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$
3	Causal Support	$\frac{n_a + n_{\bar{b}} - 2n_{a\bar{b}}}{n}$
4	Collective Strength	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})(n_a n_{\bar{b}} + n_b n_{\bar{a}})}{(n_a n_b + n_{\bar{a}} n_{\bar{b}})(n_b - n_a + 2n_{a\bar{b}})}$
5	Confidence	$1 - \frac{n_{a\bar{b}}}{n_a}$
6	Conviction	$\frac{n_a n_{\bar{b}}}{n n_{a\bar{b}}}$
7	Cosine	$\frac{n_a - n_{a\bar{b}}}{\sqrt{n_a n_b}}$
8	Dependence	$\lvert \frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a} \rvert$
9	Descriptive Confirm	$\frac{n_a - 2n_{a\bar{b}}}{n}$
10	Descriptive Confirmed-Confidence	$1 - 2\frac{n_{a\bar{b}}}{n_a}$
11	EII ($\alpha = 1$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
12	EII ($\alpha = 2$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
13	Example & Contra-Example	$1 - \frac{n_{a\bar{b}}}{n_a - n_{a\bar{b}}}$
14	Gini-index	$\frac{(n_a - n_{a\bar{b}})^2 + n_{a\bar{b}}^2}{n n_a} + \frac{(n_b - n_a + n_{a\bar{b}})^2 + (n_{\bar{b}} - n_{a\bar{b}})^2}{n n_{\bar{a}}} - \frac{n_b^2}{n^2} - \frac{n_{\bar{b}}^2}{n^2}$
15	Jaccard	$\frac{n_a - n_{a\bar{b}}}{n_b + n_{a\bar{b}}}$
16	J-measure	$\frac{n_a - n_{a\bar{b}}}{n} log_2 \frac{n(n_a - n_{a\bar{b}})}{n_a n_b} + \frac{n_{a\bar{b}}}{n} log_2 \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$
17	Kappa Cohen's	$\frac{2(n_a n_{\bar{b}} - n n_{a\bar{b}})}{n_a n_{\bar{b}} + n_{\bar{a}} n_b}$
18	Klosgen	$\sqrt{\frac{n_a - n_{a\bar{b}}}{n}}(\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a})$
19	Laplace	$\frac{n_a + 1 - n_{a\bar{b}}}{n_a + 2}$
20	Least Contradiction	$\frac{n_a - 2n_{a\bar{b}}}{n_b}$
22	Lift	$\frac{n(n_a - n_{a\bar{b}})}{n_a n_b}$
23	Loevinger	$1 - \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$
24	Odds Ratio	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})}{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}$
25	Pavillon	$\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a}$
27	Phi-Coefficient	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$
28	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_a - 3n_b}{2n} - (\frac{3}{2n_a} + \frac{2}{n_{\bar{b}}})n_{a\bar{b}}$
26	Rule Interest	$\frac{1}{n}(\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}})$
29	Sebag & Schoenauer	$1 - \frac{n_{a\bar{b}}}{n_a - n_{a\bar{b}}}$
21	Similarity Index	$\frac{n_a - n_{a\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$
30	Support	$\frac{n_a - n_{a\bar{b}}}{n}$
31	TIC	$\sqrt{TI(a \to b) \times TI(\bar{b} \to a)}$
32	Yule's Q	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{n_a n_{\bar{b}} + (n_b - n_{\bar{b}} - 2n_a) n_{a\bar{b}} + 2n_{a\bar{b}}^2}$
33	Yule's Y	$\frac{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} - \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} + \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}$

Mesurer l'intérêt des règles d'association

Benoît Vaillant*, Patrick Meyer***, Elie Prudhomme**,
Stéphane Lallich**, Philippe Lenca*, Sébastien Bigaret*

*GET ENST Bretagne / Département LUSSI – CNRS UMR 2872
Technopôle de Brest Iroise - CS 83818, 29238 Brest Cedex, France
{*prenom.nom*}@enst-bretagne.fr
**Laboratoire ERIC - Université Lumière - Lyon 2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France
lallich@univ-lyon2.fr
***Service de Mathématiques Appliquées, Faculté de Droit,
d'Economie et de Finance, Université du Luxembourg,
162a, avenue de la Faïencerie, L-1511 Luxembourg
patrick.meyer@uni.lu

Résumé. A l'occasion de l'action spécifique GaFoDonnées (2002), le laboratoire LUSSI, ENST Bretagne et le Laboratoire ERIC, Université Lyon 2, ont engagé une collaboration sur le thème de l'intérêt des règles d'association. Cet article présente les travaux ainsi réalisés. Une vingtaine de mesures ont été retenues, sur la base d'un critère d'éligibilité. Différentes propriétés sont d'abord proposées qui fondent une étude formelle des mesures. Cette étude formelle se double d'une étude de comportement, grâce à Herbs, une plate-forme développée pour expérimenter les mesures sur des bases de règles. Il est alors possible de confronter la typologie formelle des règles et la typologie expérimentale associée à leur comportement sur différentes bases. Une fois transformées en critères, ces propriétés fondent une méthode d'assistance au choix de l'utilisateur. Le problème de la validation est enfin abordé, où l'on présente une méthode de contrôle du risque multiple adaptée au problème.

1 Introduction

Nous nous intéressons aux mesures relatives à l'intérêt des règles d'association $\mathtt{A} \rightarrow \mathtt{B}$ telles que définies dans (Agrawal *et al.*, 1993) : dans une base de données transactionnelles, $\mathtt{A} \rightarrow \mathtt{B}$ signifie que si les articles qui constituent A sont dans *le panier d'une ménagère*, alors le plus souvent les articles qui constituent B le sont aussi. Les algorithmes de type Apriori (fondé sur le support et la confiance) ont tendance à produire un grand nombre de règles pas toujours intéressantes du point de vue de l'utilisateur. Les mesures d'intérêt jouent alors un rôle essentiel en permettant de pré-filtrer les règles extraites. Après nous être intéressés séparément à ce problème (Teytaud et Lallich, 2001), (Vaillant, 2002), le groupe de travail GafoQualité [1] nous a donné l'occasion de développer en commun nos recherches. On trouvera

[1] Groupe de travail sur les Mesures de Qualité, animé par Fabrice Guillet, de l'Action Spécifique STIC Fouille de Bases de Données (GaFoDonnées), animée par Rosine Cicchetti et Michèle Sebag

dans (Briand *et al.*, 2004) différents articles issus des travaux menés dans GafoQualité (qualité des données, des règles d'association, des arbres de décision, etc.). Cet article présente une synthèse de nos travaux sur la qualité des règles d'association.

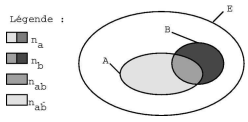
Nous nous plaçons en phase de *post-analyse*. Ainsi n'abordons nous ni les problèmes liés à la qualité des données, étudiés notamment par (Berti-Equille, 2004), ni ceux posés par l'extraction des règles (Pasquier, 2000). Les données et les règles issues du processus d'extraction sont des *entrées*.

Différentes voies ont été explorées. Ainsi, nous définissons des mesures et proposons des propriétés souhaitables section 2. La section 3 concerne le développement de la plateforme expérimentale Herbs. La section 4 est relative au développement d'une aide à la sélection de bonnes mesures. Les deux typologies des mesures, l'une fondée sur une approche expérimentale, l'autre sur une approche formelle sont mises en regard section 5. Enfin, la section 6 s'intéresse à la validation des règles.

2 Mesures et propriétés

Soit $n = |E|$, le nombre total d'enregistrements
Pour $\mathtt{A} \rightarrow \mathtt{B}$, on note :
$n_a = |\mathtt{A}|$, le nombre d'enregistrements vérifiant $\mathtt{A}$.
$n_b = |\mathtt{B}|$, le nombre d'enregistrements vérifiant $\mathtt{B}$.
$n_{ab} = |\mathtt{A} \cap \mathtt{B}|$, le nombre d'exemples de la règle.
$n_{a\bar{b}} = |\mathtt{A} \cap \bar{\mathtt{B}}|$, le nombre de contre-exemples à la règle.
$\mathtt{A} \rightarrow \mathtt{B}$ est évaluée à l'aide de mesures généralement monotones décroissantes en fonction de $n_{a\bar{b}} = n_a - n_{ab}$. $\mathtt{A} \rightarrow \mathtt{B}$ est jugée intéressante selon la mesure μ lorsque $\mu(\mathtt{A} \rightarrow \mathtt{B}) \geq \alpha$, α devant être fixé par l'utilisateur.

Pour $\mathtt{X} \subset \mathtt{E}$, on remplace n_X/n par p_X lorsque l'on considère les fréquences relatives plutôt que les fréquences absolues.



$A \backslash B$	O	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	1

Si l'on fixe les caractéristiques marginales du tableau (n, n_a et n_b ou p_a et p_b), il suffit de connaître une cellule du tableau pour reconstruire les autres.

Fig. 1 – Notations.

Notre premier travail a été de recenser les mesures de l'intérêt des règles d'association et de mettre en évidence leurs propriétés, tant dans une perspective d'analyse formelle (Teytaud et Lallich, 2001), (Lallich, 2002), que d'expérimentation (Vaillant, 2002). Ces propriétés ont ensuite été écrites de façon opérationnelle pour servir de base à nos travaux d'aide à la décision (Lenca *et al.*, 2002, Lenca *et al.*, 2004).

Les règles d'association $\mathtt{A} \rightarrow \mathtt{B}$ se focalisent sur les coprésences en traitant les itemsets de façon non symétrique, une mesure doit impérativement distinguer $A \rightarrow B$ de $A \rightarrow \overline{B}$ (Lallich et Teytaud, 2004) et une règle $\mathtt{A} \rightarrow \mathtt{B}$ doit être distinguée de l'implication $A \Rightarrow B$ et de l'équivalence $A \Leftrightarrow B$. Ainsi qu'à la différence de (Tan *et al.*, 2002), nous nous sommes limités aux mesures qui sont décroissantes avec $n_{a\overline{b}}$ (resp. croissantes avec n_{ab}), les effectifs marginaux étant fixés, excluant d'emblée les mesures comme le χ^2, le r^2 de Pearson ou la mesure de Pearl.

Nous avons retenu 20 mesures, listées tableau 1. A côté du support p_{ab} et de la confiance $p_{b|a}$, une première catégorie de mesures rassemble des transformées affines de

la confiance qui ont pour but de la comparer à p_b. Cette comparaison se fait le plus souvent en centrant la confiance sur p_b avec différents coefficients d'échelle (confiance centrée, coefficient de corrélation, indice d'implication, mesures de Piatetsky-Shapiro, Loevinger, Zhang). Elle peut aussi se faire en divisant la confiance par p_b (lift ou taux de liaison). D'autres mesures sont des transformées monotones croissantes de la confiance, ainsi la mesure de Sebag-Schoenauer, le taux d'exemples et de contre-exemples. Certaines mesures privilégient les contre-exemples, ainsi la conviction $\frac{p_{\overline{b}|a}}{p_{\overline{b}}}$ et l'indice d'implication. Ce dernier est à la base de différents indices probabilistes comme l'intensité d'implication, l'intensité d'implication entropique et l'indice probabiliste discriminant. En outre, nous avons analysé les classes d'équivalences issues de la relation d'équivalence "classer comme" définie sur les couples de mesure, bon nombre de mesures étant des transformées monotones croissantes les unes des autres (*e.g* la mesure de Sebag et la confiance, ou la mesure de Loevinger et la conviction).

Pour analyser formellement ces mesures, puis les évaluer dans une perspective d'aide à la décision, nous proposons 8 propriétés (en gras, ci-dessous). L'antécédent et le conséquent d'une règle n'ayant pas le même rôle, il est souhaitable qu'une mesure évalue de façon différente les règles $A \rightarrow B$ et $B \rightarrow A$ (**dissymétrie**). Pour une même proportion d'exemples p_{ab}, une règle est d'autant plus intéressante que p_b est faible (**décroissance avec** p_b). Les comparaisons sont plus faciles lorsque les mesures ont une **valeur fixe en cas de règle logique**, ainsi qu'en cas d'**indépendance.** Certains auteurs, tel (Gras *et al.*, 2004), privilégient des mesures concaves lors de l'apparition des premiers contre-exemples, d'autres peuvent préférer une décroissance convexe plus brutale, ou simplement linéaire (**courbure à l'origine**). La mesure doit-elle prendre en compte le nombre total de transactions n (mesure statistique) ou non (descriptive) ? Les règles statistiques intuitivement plus fondées, ont l'inconvénient de perdre leur pouvoir discriminant dès que n est grand (**prise en compte de** n). Face à la multitude de règles évaluées, il est important de pouvoir facilement fixer le seuil à partir duquel on considère que les règles ont un réel intérêt sans avoir à les classer (**fixation du seuil**). On peut se référer à la probabilité critique de la valeur observée de la mesure sous l'hypothèse d'indépendance (ou *p-value*). Celle-ci ne doit pas être interprétée comme un risque statistique compte tenu de la multitude de tests effectués, mais comme un paramètre de contrôle, sauf à contrôler effectivement le risque multiple avec le critère UAFWER (voir section 6).

3 Herbs : une plate-forme d'expérimentation

Dans le cadre de la recherche de règles d'association, il est classique d'effectuer une opération de filtrage au moyen de mesures de qualité. Des logiciels permettant l'extraction de telles règles proposent ainsi d'utiliser le lift (IBM, 1996), en plus du support et de la confiance. D'autres proposent le coefficient de corrélation linéaire. Plus récemment, l'outil Felix (Lehn, 2000) intègre l'intensité d'implication et sa version entropique. Mais à notre connaissance, les outils disponibles ne proposent qu'un sous-ensemble réduit de mesures, et qui plus est leur intégration est à but fonctionnel et non afin d'en étudier les comportements.

Nous avons développé Herbs (Vaillant, 2002), qui intègre 20 mesures de qualité

Mesure	Abréviation et référence	Définition
support	SUP (Agrawal *et al.*, 1993)	$\frac{n_a - n_{a\bar{b}}}{n}$
confiance	CONF (Agrawal *et al.*, 1993)	$1 - \frac{n_{a\bar{b}}}{n_a}$
coefficient de corrélation linéaire	R (Pearson, 1896)	$\frac{nn_{ab} - n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}} . n_{\bar{b}}}}$
confiance centrée	CONFCEN	$\frac{nn_{ab} - n_a n_b}{nn_a}$
conviction	CONV (Brin *et al.*, 1997b)	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$
Piatetsky-Shapiro	PS (Piatetsky-Shapiro, 1991)	$\frac{1}{n}(\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}})$
Loevinger	LOE (Loevinger, 1947)	$1 - \frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$
gain informationnel	GI (Church et Hanks, 1990)	$\log(\frac{nn_{ab}}{n_a n_b})$
Sebag-Schoenauer	SEB (Sebag et Schoenauer, 1988)	$\frac{n_a - n_{a\bar{b}}}{n_{a\bar{b}}}$
lift	LIFT (Brin *et al.*, 1997a)	$\frac{nn_{ab}}{n_a n_b}$
Laplace	LAP (Good, 1965)	$\frac{n_{ab}+1}{n_a+2}$
moindre contradiction	MOCO (Azé et Kodratoff, 2002)	$\frac{n_{ab} - n_{a\bar{b}}}{n_b}$
multiplicateur de cotes	MC (Lallich et Teytaud, 2004)	$\frac{(n_a - n_{a\bar{b}})n_{\bar{b}}}{n_b n_{a\bar{b}}}$
taux d'exemples et de contre-exemples	TEC	$\frac{n_a - 2n_{a\bar{b}}}{n_a - n_{a\bar{b}}}$
indice de qualité de Cohen	IQC (Cohen, 1960)	$2\frac{nn_a - nn_{a\bar{b}} - n_a n_b}{nn_a + nn_b - 2n_a n_b}$
Zhang	ZHANG (Terano *et al.*, 2000)	$\frac{nn_{ab} - n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$
indice d'implication	-INDIMP (Lerman *et al.*, 1981)	$\frac{nn_{a\bar{b}} - n_a n_{\bar{b}}}{\sqrt{nn_a n_{\bar{b}}}}$
intensité d'implication	INTIMP (Gras *et al.*, 1996)	$P\left[poisson\left(\frac{n_a n_{\bar{b}}}{n}\right) \geq n_{a\bar{b}}\right]$
intensité d'implication entropique	IIE (Gras *et al.*, 2001)	$\left\{[(1 - h_1(\frac{n_{a\bar{b}}}{n})^2) \times (1 - h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}{}_{\text{INTIMP}}\right\}^{1/2}$
indice probabiliste discriminant	IPD (Lerman et Azé, 2003)	$P\left[\mathcal{N}(0,1) >_{\text{INDIMP}}{}^{CR/\mathcal{B}}\right]$

- $h_1(t) = -(1 - \frac{n.t}{n_a})\log_2(1 - \frac{n.t}{n_a}) - \frac{n.t}{n_a}\log_2(\frac{n.t}{n_a})$ si $t \in [0, n_a/2\ n[$; $h_1(t) = 1$ sinon
- $h_2(t) = -(1 - \frac{n.t}{n_{\bar{b}}})\log_2(1 - \frac{n.t}{n_{\bar{b}}}) - \frac{n.t}{n_{\bar{b}}}\log_2(\frac{n.t}{n_{\bar{b}}})$ si $t \in [0, n_{\bar{b}}/2\ n[$; $h_2(t) = 1$ sinon
- *poisson* correspond à la loi de distribution de Poisson
- $\mathcal{N}(0,1)$ correspond à la fonction de distribution de la loi normale centrée réduite
- INDIMP $^{CR/\mathcal{B}}$ correspond à INDIMP, centré réduit (CR) pour une base de règle $\mathcal{B}$

TAB. 1 – Mesures étudiées

dont nous étudions par ailleurs les propriétés formelles. Il est possible d'importer des règles aux formats de sortie C4.5 (`http://www.rulequest.com/Personal`) et APRIORI (`http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/apriori`). Afin d'évaluer ces règles, il est possible d'importer des données au format csv. D'autres formats d'échanges sont envisagés, par exemple avec WEKA (`http://www.cs.waikato.ac.nz/~ml/weka`) et TANAGRA (`http://chirouble.univ-lyon2.fr/~ricco/tanagra`).

A partir de couples de bases de cas et de règles compatibles (*i.e.* portant sur les mêmes attributs), HERBS permet d'effectuer divers types d'analyse :

Etude d'une mesure Plusieurs traitements nous semblent intéressants afin de caractériser le comportement d'une mesure donnée :

– Evaluation des objets que la mesure va avoir à traiter au moyen de quelques grandeurs : nombre de cas et de règles, taux de couverture, indice de recouvrement, et nombre de règles "particulières" (logiques, sans exemples, et ne passant pas l'hypothèse d'indépendance).

- Sélection de l'ensemble des N meilleures règles selon la mesure donnée.
- Tracé de la distribution des valeurs prises par la mesure.

Comparaison de mesures Afin de comparer le comportement expérimental de mesures, trois voies ont été développées :

- L'extraction de l'ensemble des règles classées k fois parmi les N meilleures par p mesures.
- La comparaison des préordres induits par deux mesures.
- Le tracé des distributions croisées des valeurs de deux mesures.

Les résultats présentés dans la figure 2 sont extraits de (Vaillant *et al.*, 2004). Plusieurs jeux de données disponibles depuis le site de l'UCI (`ftp.ics.uci.edu/`) ont été utilisés afin de générer des règles. Pour une base de règles donnée, chaque mesure induit un préordre sur l'ensemble de règles. Afin d'étudier les similarités entre mesures, nous avons calculé un coefficient d'accord entre préordres (*cf.* figure 2 et (Lenca *et al.*, 2004) pour les détails de calcul).

Après une transformation linéaire de la valeur du coefficient d'accord afin d'obtenir des valeurs entre 0 et 1, et un réarrangement de l'ordre des lignes et des colonnes afin de mieux mettre en évidence les structures de blocs, on obtient une classification expérimentale des mesures de qualité, à mettre en regard avec la classification obtenue à partir de propriétés formelles.

4 Assistance au choix des mesures

Les mesures d'intérêt possèdent des propriétés diverses (Lallich et Teytaud, 2002) et l'ensemble des n meilleures règles résultant d'un préfiltrage d'une base de $m > n$ règles peut varier grandement selon la mesure utilisée (Vaillant, 2002). Ainsi, lorsque l'utilisateur est confronté à la sélection du sous-ensemble des n meilleures règles il est aussi confronté au choix des mesures d'intérêt à appliquer : choisir les *bonnes* règles c'est aussi choisir les *bonnes* mesures (Lenca *et al.*, 2002), (Lenca *et al.*, 2003b).

Ce choix doit être guidé par les préférences et les objectifs du principal intéressé, l'utilisateur expert des données. L'utilisateur est au cœur du processus et les travaux l'impliquant fortement sont à notre avis fort prometteurs, par exemple (Poulet, 1999), (Lehn *et al.*, 1999) et (Blanchard *et al.*, 2004). Partant des huit propriétés présentées section 2 nous avons identifié celles reposant sur les préférences de l'utilisateur et celles plus normatives afin de définir des critères de décision sur les mesures. Différentes méthodes d'*aide multi-critères à la décision* ont été appliquées (Lenca *et al.*, 2003b), (Lenca *et al.*, 2003a) afin d'obtenir, selon la méthode, un sous-ensemble de *bonnes* mesures ou un classement des mesures. Dans (Lenca *et al.*, 2004) nous précisons six éléments définissant le contexte et à prendre en compte pour l'assistance au choix des mesures : l'ensemble de données, l'ensemble de règles, l'ensemble de mesures, l'ensemble de propriétés des mesures, l'ensemble de préférences de l'utilisateur, l'ensemble de critères de décision. Nous y présentons une étude détaillée de deux scénarios utilisateur (tolérance –**Sc1**– ou non –**Sc2**– de l'apparition de contre-exemples dans les règles) et des classements des mesures selon ces scénarios. Le tableau 2 donne le classement des mesures pour ces deux scénarios avec des poids égaux pour les critères, obtenus avec la méthode PROMETHEE (Brans et Mareschal, 1994).

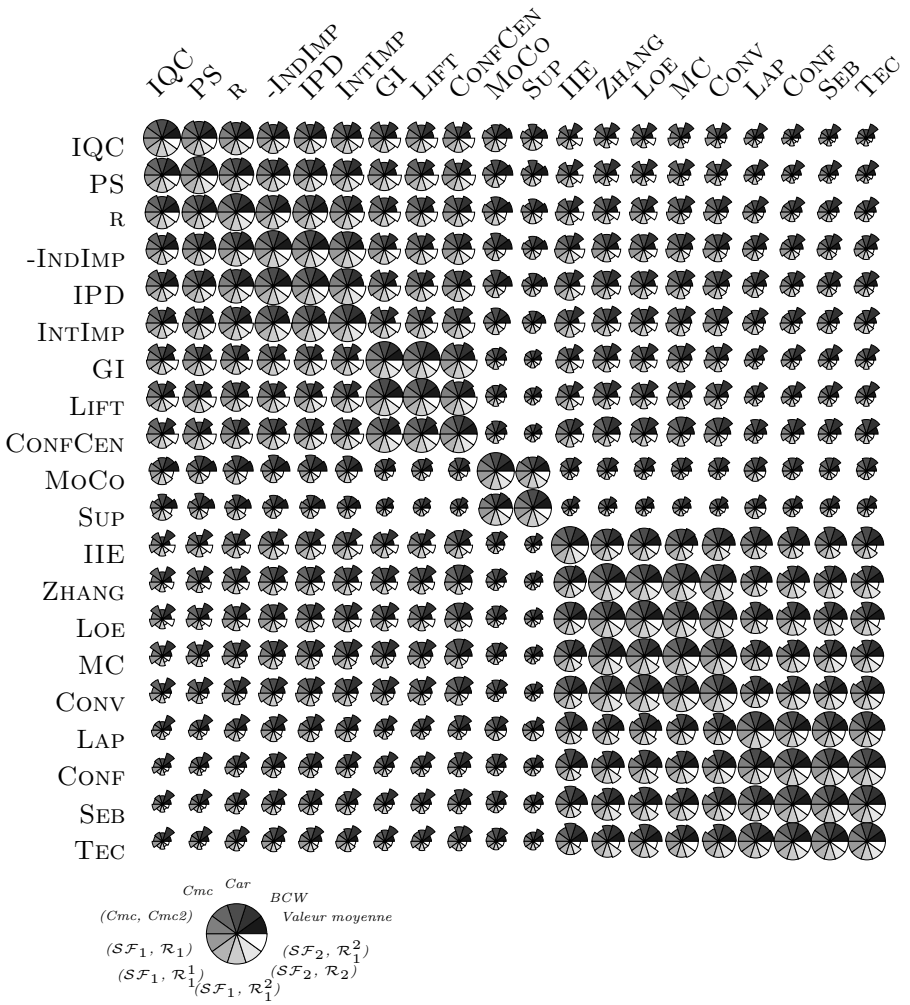


Fig. 2 – Comparaison de préordres

5 Typologies formelle et expérimentale des mesures

Nous avons extrait un sous-ensemble pertinent de 6 propriétés formelles, détaillées dans (Lenca *et al.*, 2003a). Ce sous-ensemble de propriétés nous a permis de construire une matrice de décision évaluant les mesures. A partir de ces évaluations, on peut ainsi construire une matrice de distance entre les mesures. En appliquant une classification ascendante hiérarchique avec le critère de Ward on distingue quatre classes principales : {PS, IQC, GI, ConfCen, Lift, R, -IndImp, IPD}, {IntImp, IIE, Loe, Zhang, MC, Conv}, {Conf, Seb, Tec}, et {Lap, Sup, MoCo}.

Le tableau 3 compare les deux approches. Du point de vue expérimental, seule la classe 3 présente de forts désaccords avec les résultats formels.

Rang :	1	2	3	4	5	6	7
Sc1 :	IntImp	Loe	MC	ConfCen	Conv	-IndImp,IPD	
Sc2 :	MC	Conv	Loe	ConfCen	IntImp	-IndImp, IPD	
Rang :	8	9	10	11	12	13	14
Sc1 :	IIE,Zhang		PS	Tec	Conf	GI	r, Lift
Sc2 :	PS	Seb	Conf	r, Lift		MoCo	IIE
Rang :	15	16	17	18	19	20	
Sc1 :		MoCo	Seb	IQC	Sup	Lap	
Sc2 :	Zhang	IQC	Tec	Sup	GI	Lap	

Tab. 2 – Rangements totaux pour les scénarios **Sc1** et **Sc2**.

Formelle \ Expérimentale	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	PS, IQC, GI ConfCen, Lift, r -IndImp, IPD			
Classe 2	IntImp	IIE	Loe, Zhang, MC, Conv	
Classe 3			Conf, Seb, Tec	
Classe 4			Lap	MoCo, Sup

Tab. 3 – Comparaison des classes entre l'approche formelle et expérimentale

6 Validation de règles

Le plus souvent, les *transactions* à partir desquelles sont extraites les règles d'association ne sont qu'un échantillon d'une population plus vaste. Au terme de la procédure d'extraction et d'évaluation des règles, on dispose d'une multitude de règles décrites par différentes mesures, au minimum le support et la confiance, ainsi qu'une mesure de l'intérêt de la règle. Différentes questions se posent classiquement : ces mesures, en particulier le support et la confiance, dépassent-elles significativement pour toutes les règles le seuil requis ? Telle mesure dépasse-t-elle significativement une valeur fixée ? La confiance $p_{b/a}$ d'une règle $A \rightarrow B$ est-elle significativement supérieure à sa fréquence *a priori* ? Dans ce dernier cas, il s'agit de tester l'hypothèse d'indépendance (H_0) du conséquent B et de l'antécédent A contre une hypothèse de dépendance positive (H_1), afin de ne retenir que les règles statistiquement significatives. On doit donc pratiquer une multitude de tests, ce qui pose le problème du contrôle du risque multiple. Par exemple, si l'on effectue le test d'indépendance de A et B pour 10000 règles successivement, en fixant à 0.05 le niveau du risque de 1^{re} espèce $\alpha = P(\textit{décider } H_1/H_0)$, alors même qu'aucune règle ne serait pertinente, on sélectionne quand même 500 règles en moyenne.

Dans le but de contrôler le risque multiple, nous avons d'abord utilisé des outils de la théorie de l'apprentissage statistique, fondés pour l'essentiel sur la dimension de Vapnik, afin de proposer des bornes uniformes non asymptotiques pour toutes les règles et toutes les mesures considérées (Teytaud et Lallich, 2001). Par la suite (Lallich et Teytaud, 2004), nous avons proposé *BS*, un algorithme fondé sur le bootstrap qui contrôle le risque de 1^{re} espèce sur l'ensemble des tests. Ces méthodes assurent que le risque de faire la moindre fausse découverte soit égal à un seuil fixé, mais elles ont l'inconvénient d'exprimer un point de vue très sévère sur les erreurs, ce qui les rend

peu puissantes et les amène à manquer plus souvent de vraies découvertes.

Pour remédier à ce défaut, nous avons choisi de contrôler non pas le risque, mais le nombre V de fausses découvertes, suivant les procédures de sélection de gènes dévelopées en biostatistique. Nous avons proposé (Lallich *et al.*, 2004) un critère original, *User Adjusted Family Wise Error Rate,* $UAFWER = \Pr(V > V_0)$, que nous contrôlons au risque δ grâce à une procédure fondée sur le bootstrap. Ce critère est plus tolérant au sens où il assure au risque δ d'avoir au maximum V_0 fausses découvertes. Appliquée à différentes bases de règles, cette procédure a permis d'éliminer jusqu'à 50% de règles non significatives. La méthode proposée a le double avantage de gérer la dépendance entre les différentes règles, grâce au bootstrap, et de ne pas nécessiter la connaissance de la loi de la statistique de test sous H_0, exigeant seulement une valeur fixe sous H_0.

7 Conclusion

Le principe de notre démarche commune est de mettre en regard deux approches complémentaires du problème de la mesure de l'intérêt des règles d'association, l'une formelle, l'autre empirique. Dans le cadre de l'approche formelle, nous proposons un certain nombre de propriétés opérationnelles qui permettent d'évaluer les mesures. Pour mener à bien l'approche empirique, nous avons développé HERBS, une plateforme d'expérimentation des mesures sur des bases de règles et nous nous sommes donnés de contrôler la validation statistique des règles retenues. Nous avons aussi fait le lien avec l'utilisateur en lui donnant le moyen de choisir la mesure qui correspond le mieux à ses préférences en termes de propriétés. D'autres propriétés intéressantes sont à l'étude et devraient être intégrées dans notre plateforme, ainsi que dans le module d'aide à la décision.

Références

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

Azé J. et Kodratoff Y., (2002). Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. *EGC 2002*, 1(4) :143–154.

Berti-Equille L., (2004). La qualité des données comme condition vers la qualité des connaissances : un état de l'art. *RNTI-E-1*, pp 95–118.

Blanchard J., Guillet F. et Briand H., (2004). Une visualisation orientée qualité pour la fouille anthropocentrée de règles d'association. *Cahiers Romans de Sciences Cognitives – In Cognito*, 1(3) :79–100.

Brans J.P. et Mareschal B., (1994). The PROMETHEE-GAIA decision support system for multicriteria investigations. *Investigation Operativa*, 4(2) :102–117.

Briand H., Sebag M., Gras R. et Guillet F. (éditeurs), (2004). Mesures de qualité pour la fouille de données. *RNTI-E-1.*

Brin Sergey, Motwani Rajeev et Silverstein Craig, (1997). Beyond market baskets : generalizing association rules to correlations. In *ACM SIGMOD/PODS'97*, pp 265–276.

Brin Sergey, Motwani Rajeev, Ullman Jeffrey D. et Tsur Shalom, (1997). Dynamic itemset counting and implication rules for market basket data. In Peckham Joan, editor, *ACM SIGMOD 1997 Int. Conf. on Management of Data*, pp 255–264.

Church Kenneth Ward et Hanks Patrick, (1990). Word association norms, mutual information an lexicography. *Computational Linguistics*, 16(1) :22–29.

Cohen J., (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20 :37–46.

Good Irving John. The estimation of probabilities : An essay on modern bayesian methods. The MIT Press, Cambridge, MA, (1965).

Gras R., Ag. Almouloud S., Bailleuil M., Larher A., Polo M., Ratsimba-Rajohn H. et Totohasina A., (1996). *L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application à la Didactique, Travaux et Thèses*. La Pensée Sauvage.

Gras R., Kuntz P., Couturier R. et Guillet F., (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *EGC 2001*, 1(1-2) :69–80.

Gras R., Couturier R., Blanchard J., Briand H., Kuntz P. et Peter P., (2004). Quelques critères pour une mesure de qualité de règles d'association. *RNTI-E-1*, pp 3–31.

Lallich S., Prudhomme E. et Teytaud O., (2004). Contrôle du risque multiple en sélection de règles d'association significatives. *RNTI-E-2*, 2 :305–316.

Lallich S. et Teytaud O., (2002). Évaluation et validation de l'intérêt des règles d'association. Rapport de recherche pour le groupe de travail GAFOQUALITÉ de l'action spécifique STIC fouille de bases de données, E.R.I.C., Université Lyon 2.

Lallich S. et Teytaud O., (2004). Évaluation et validation de l'intérêt des règles d'association. *RNTI-E-1*, pp 193–217.

Lallich S., (2002). Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches – Université Lyon 2.

Lehn R., Guillet F., Kuntz P., Briand H. et Philippé J., (1999). Felix : An interactive rule mining interface in a kdd process. In Lenca P., editor, *Proceedings of the Human Centered Processes Conference*, pp 169–174, Brest, France.

Lehn R., (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. Thèse de Doctorat, Université de Nantes.

Lenca P., Meyer P., Vaillant B. et Picouet P., (2002). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances - modélisation des préférences de l'utilisateur. Rapport de recherche pour le groupe de travail GAFOQUALITÉ de l'action spécifique STIC fouille de bases de données, Département IASC, ENST Bretagne.

Lenca P., Meyer P., Picouet P., Vaillant B. et Lallich S., (2003a). Critères d'évaluation des mesures de qualité en ECD. *Entreposage et Fouille de données*, (1) :123–134.

Lenca P., Meyer P., Vaillant B. et Picouet P., (2003b). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. *RSTI-RIA (EGC 2003)*, 1(17) :271–282.

Lenca P., Meyer P., Vaillant B., Picouet P. et S. Lallich, (2004). Évaluation et analyse multicritère des mesures de qualité des règles d'association. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1) :219–246.

Lerman I.C., Gras R. et Rostam H., (1981). Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75) :5–35, 5–47.

Lerman I.C. et Azé J., (2003). Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *EGC 2003*, 1(17) :247–262.

Loevinger J., (1947). A systemic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4).

IBM, (1996). *IBM Intelligent Miner User's Guide, Version 1 Release 1, SH12-6213-00.*

Pasquier N., (2000). *Data Mining : Algorithmes d'extraction et de réduction des règles d'association dans les bases de données.* Thèse de Doctorat, Université Blaise Pascal - Clermont-Ferrand II.

Pearson Karl, (1896). Mathematical contributions to the theory of evolution. regression, heredity and panmixia. *Philosophical Trans. of the Royal Society*, A.

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

Poulet F., (1999). Visualization in data-mining and knowledge discovery. In Lenca P., editor, *Proceedings of the Human Centered Processes Conference*, pp 183–191, Brest, France.

Sebag M. et Schoenauer M. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In Boose J., Gaines B. et Linster M., editors, *EKAW'88*, pp 28–1 – 28–20. (1988).

Tan P.-N., Kumar V. et Srivastava J., (2002). Selecting the right interestingness measure for association patterns. In *Eighth ACM SIGKDD Int. Conf. on KDD*, pp 32–41.

Terano Takao, Liu Huan et Chen Arbee L. P., editors. *Association Rules*, volume 1805 of *Lecture Notes in Computer Science*. Springer, April 2000.

Teytaud O. et Lallich S., (2001). Bornes uniformes en extraction de règles d'association. In *Conférence d'Apprentissage, CAp'01*, pp 133–148.

Vaillant B., Lenca P. et Lallich S., (2004). A clustering of interestingness measures. In *Discovery Science*, volume 3245 of *Lecture Notes in Artificial Intelligence*, pp 290–297. Springer-Verlag.

Vaillant B., (2002). Evaluation de connaissances : le problème du choix d'une mesure de qualité en ECD. Rapport de DEA, ENST Bretagne.

Mise en place d'un plan d'Assurance Qualité du Dossier Patient

Mireille Cosquer*, Myriam Turluche**, Béatrice Le Vu*, Bernard Asselain***, Alain Livartowski*
Institut Curie, Service d'Information Médicale*, Direction Qualité**,
Service de Biostatistiques***, 25 Rue d'Ulm, 75005 Paris
mireille.cosquer@curie.net

Résumé
Cet état des lieux présente la démarche engagée sur la problématique de la qualité des données du Dossier Patient dans un Centre de Lutte Contre le Cancer. En préliminaire à la mise en place d'un système d'assurance qualité du dossier patient, les différentes outils de contrôles spécifiques existants seront abordés, avec quelques résultats d'évaluation issus de ces outils. L'objectif est de formaliser à court terme, par un groupe de travail, une procédure décrivant l'organisation à instaurer, les méthodes d'évaluation à définir, et les actions correctives à mettre en place compte tenu des résultats de ces évaluations. Il concerne un volet organisationnel, mais aussi méthodologique. Nous nous limitons ici au champ d'application de la qualité des données du dossier patient. Ne sera donc pas abordée la qualité des données « dérivées » du Dossier Patient, ni celle des données de gestion.

1 Introduction

Le dossier patient a considérablement évolué ces dernières années dans son contenant (support informatisé), mais aussi dans son contenu (réglementation). Sa fiabilité, semble aller de soi, du fait de la nature même des informations renseignées dans un dossier médical présentant une importance majeure pour la prise en charge d'un patient. Néanmoins, un système de contrôle qualité est nécessaire du fait de la complexité inhérente à l'univers médical. Pour preuve, dans le cadre de l'accréditation, l'évaluation de la qualité du dossier patient apparaît dans le manuel d'accréditation des établissements de santé de l'ANAES (Agence Nationale d'Accréditation et d'Evaluation des Soins) sous deux références. La première insiste sur la nécessité de protocoler cette évaluation, de communiquer sur les actions d'évaluation, de proposer et de mettre en oeuvre des actions correctrices. La seconde nous rappelle qu'un plan d'amélioration de la qualité du système d'information, aux priorités hiérarchisées et auquel participent les professionnels utilisateurs, doit être mis en place.

2 Quelle qualité du dossier patient ?

2.1 Concept de la qualité des données

2.1.1 Qualité ou non qualité ?

Le modèle de la Roue de Deming correspond à un cycle de la maîtrise de la qualité, s'appliquant au processus de management de l'information.

La première difficulté à laquelle on se trouve confronté avec le concept de qualité est l'absence de consensus sur sa définition, du fait de son caractère subjectif. A l'inverse, la non

qualité semblerait plus aisée à définir par la nécessité du contrôle et la gestion de la non qualité (l'exemple de la Base des Evènements Indésirables en est une illustration).

Par ailleurs, aujourd'hui, une réelle prise de conscience de l'enjeu de la qualité des données s'opère dans les systèmes d'information rendant nécessaire de garantir la qualité des données.

2.1.2 La mesure de la qualité

La qualité n'a pas de définition consensuelle, mais on la mesure. Citons les différentes dimensions les plus fréquentes (Berti 2004) : la complétude (taux de valeurs manquantes), l'exactitude (taux de valeurs correctes), la fraîcheur ou l'actualité (comparaison de la date de saisie à la date courante), la cohérence (taux de données ne satisfaisant pas à un ensemble de contraintes).

2.2 Les documents de référence : contenu réglementaire et ANAES

Le dossier patient comprend trois composants : le dossier médical, le dossier de soins, le dossier administratif. Son contenu réglementaire du dossier médical concerne le recueil d'informations en consultation externe, lors de l'accueil aux urgences ou au moment de l'admission et durant le séjour hospitalier. Un second décret est relatif à l'identification du patient.

L'ANAES propose d'évaluer la qualité de la tenue du dossier patient par un référentiel commun à tous les établissements à partir d'un guide (ANAES 2003). Différentes méthodes d'amélioration de la qualité y sont proposées : audit clinique, benchmarking, analyse des processus, re engineering, résolution de problème.

3 Etat des lieux (2003-2004)

Avant de proposer un plan d'assurance qualité du DP, il a fallu définir la qualité de notre DP : Quelles données évaluer ? Comment évaluer ? Où évaluer ? Quand ? Par qui ?. Les différents axes de travail ont permis de : valider le contenu à évaluer, rechercher les outils d'évaluation, réaliser les évaluations.

Le dossier patient est le cœur de notre système d'information. Il ne s'agit pas d'un outil unique, mais de différentes applications sources communicantes. Par ailleurs, des applications « destinataires» reposent sur lui ; ainsi une donnée de mauvaise qualité à la source, peut se propager dans l'ensemble du SIH.

3.1 Etape n°1 : A la recherche des données essentielles

La volumétrie du DMI s'accroît quotidiennement, du fait qu'il trace le détail de la relation patient-hôpital (1000 documents créés/jour). Il est évident que l'on ne peut évaluer la totalité de son contenu. Il faut donc choisir le périmètre à évaluer…l'information « essentielle ».

Un ensemble de données dites essentielles, a été défini. Par essentielle, on définit une utilité pour la prise en charge du patient, mais aussi pour d'autres finalités (réglementation, PMSI, recherche clinique, pilotage, facturation, Enquête Permanente Cancer, Biostatistiques). . Il peut s'agir de données fixes (sexe, date naissance…) ou évolutives (date des dernières nouvelles) impliquant la mise en place d'un contrôle approprié (unique ou continu).

3.2 Etape n°2 : cartographie des outils de contrôles de données existants

On peut distinguer deux grands types de contrôle qualité du dossier patient : le contrôle des données « source » (contenu) ciblant la donnée intrinsèque, et le contrôle des systèmes (contenant) concernant les flux inter applications.

(1) les contrôles à la source : automatiques et humains

Lors de la mise en place du dossier médical informatisé, la qualité des dossier a davantage été appréhendée d'un point de vue informatique, à savoir celui du contenant (le système où est stocké la donnée) que du point de vue du métier, celui du contenu (l'analyse du sens de la données).

Les contrôles automatiques à la saisie de la donnée sont les suivants : présence de donnée obligatoire, contrôle de pertinence d'une donnée : bornes minimales et maximales, format prédéfini (saisie sur liste), contrôle de cohérence inter-champs (saisie bloquée pour tout enregistrement dans le DP des patients dont le statut est décédé), présence d'une valeur par défaut. Le contrôle d'exhaustivité d'une donnée représente une première étape indispensable dans le contrôle qualité. En terme d'analyse, il est bien souvent préférable d'avoir un tamis grossier exhaustif, qu'un tamis fin non exhaustif. Pour le dossier de soins, des contrôles (pertinence, cohérence) à la source ont été intégrés d'emblée lors de sa conception en collaboration avec des soignants..

Il ne faut pas oublier que certains contrôles humains, peuvent être d'une grande efficacité. Citons le contrôle par le patient lui-même lors de sa venue, ou par le soignant lors de l'accueil.

La maximisation des contrôles à la source par le système d'information évite par ailleurs des corrections a posteriori. Seule l'exactitude d'une donnée, fait appel très souvent à l'avis d'un expert nécessitant la mise en place d'audit sur dossiers.

(2) la mise en place de procédures de recueil

Dans un objectif d'harmonisation des pratiques de saisie, le recueil des données peut faire l'objet d'une procédure. On s'approche alors d'une démarche « protocolée », nécessaire dans certains cas où le système du Dossier Patient est complexe, et peut faire l'objet d'une analyse des processus. Les procédures existantes sont les suivantes : Guide des procédures de la tenue globale du dossier patient guide de recueil, Procédure d'identification du patient ; Procédure de l'enregistrement du décès dans le dossier patient.

(3) L'audit clinique

Ce type d'approche permet de calculer un taux de conformité par rapport à des critères précis à partir d'une méthodologie rigoureuse (échantillonnage, recueil, contrôle qualité, analyse, restitution) et permet des comparaisons inter-services. L'intérêt de cette mesure est sa reproductibilité dans le temps. Par contre en ce qui concerne la comparabilité inter-établissements, différents protocoles d'évaluation existent à ce jour, mais aucun n'est spécifique à la cancérologie. L'audit proposé par l'ANAES concerne davantage les établissements possédant un Dossier Patient papier, et ne tient pas compte de la spécificité de la pathologie cancéreuse.

Actuellement, sont mis en œuvre uniquement des audits « flash » ciblant une variable donnée (ou peu de variables).

(4) Les contrôles a posteriori

Des vues intégrées au DP, permettent de réaliser un auto-contrôle par les producteurs d'information, de détecter automatiquement des atypies, ou de faire l'objet d'un processus de validation. D'autres contrôles exhaustifs sont réalisés en aval sur des outils d'exploitation spécifiques ou au niveau des passerelles informatiques.

4 Mise en place d'une démarche d'évaluation de la qualité du Dossier Patient : vers un processus qualité de la donnée

4.1 Mise en place d'un groupe projet « Définition du système d'assurance qualité du dossier du patient »

Ce projet transversal est un projet à part entière, un groupe de pilotage, assisté de la cellule qualité a été constitué. Ce groupe a pour mission de définir le système d'assurance qualité du dossier du patient, c'est-à-dire de mettre en place un dispositif d'évaluation de la qualité de la tenue du dossier patient basé sur un plan d'actions annuel.

4.2 Plan annuel qualité

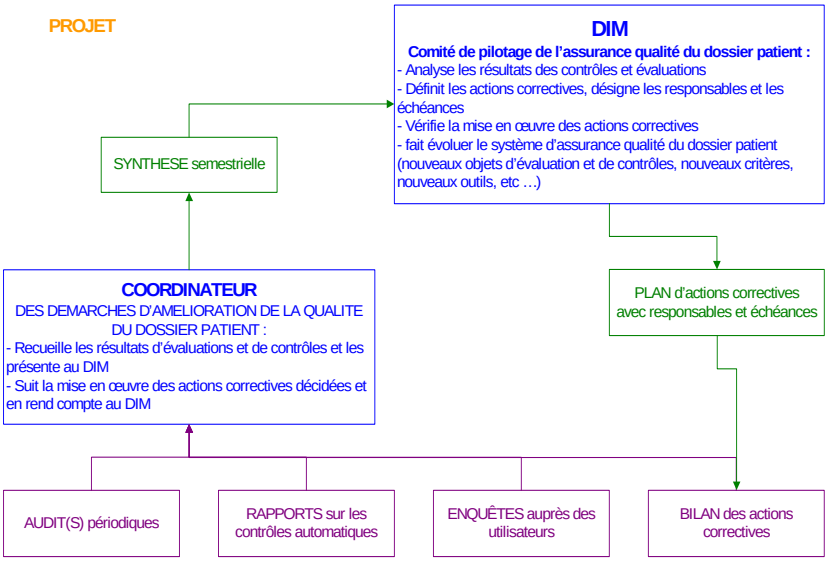
Compte tenu de notre informatisation, l'objectif est de s'orienter vers un recueil « contrôlé », et vers un audit « automatisé ».

Ce plan qualité présente différentes actions :

- Les mesures à réaliser : audits « flash » ou multi items.
- Les actions d'améliorations à engager :

1. Identité vigilance : analyse (et si besoin écriture) des procédures de recueil de l'identité, de signalement d'un doublon et de fusion de dossier.
2. Contrôles à la saisie :

- Les outils à développer : base documentaire qualité sur l'intranet, documentation des méta-données des données essentielles, mise en place d'un système d'alerte d'atypies (contrôle a posteriori automatisé selon un ensemble de règles prédéfinies), procédure de recueil des évènements indésirables (Base Notes)
- La veille : retour d'expériences des autres établissements de santé
- La communication : auprès du DIM (cf schéma ci-joint), auprès de la Commission Médicale d'Etablissement (bilan annuel), auprès des utilisateurs (bilan annuel). Les acteurs du dossier patient, représentent un volume important de professionnels (producteurs, concepteurs ou consommateurs de données), auprès desquels il est important de communiquer.
- Enquête d'opinion des utilisateurs (annuelle).



5 Conclusion

A partir d'un système d'information récent, notre objectif est d'instaurer un contrôle qualité structuré concernant le contenu du dossier du patient en termes d'exhaustivité, de fiabilité et de délai de disponibilité de l'information. Ce vaste chantier doit tenir compte également de l'évolution du système.

Le contrôle qualité est souvent un savoir faire chronophage exigeant de savoir se fixer un seuil de qualité au-delà duquel tout effort supplémentaire apparaîtrait comme un acharnement « datapeutique».

Le pilotage de la qualité de données est une démarche progressive qui va du diagnostic qualité du dossier du patient à une certification des données. Ce projet paraît important pour une bonne prise en charge du patient, et intervient par ailleurs dans un contexte de la Tarification à l'Activité où la qualité des données est primordiale.

6 Références

ANAES (2004), Manuel d'accréditation des établissements de santé – 2ème procédure d'accréditation.

Berti Equille L. (2004), La qualité des données comme condition à la qualité des connaissances : un état de l'art, RNTI, 2004.

ANAES (2003), Dossier du patient : amélioration de la qualité et de la tenue et du contenu – réglementation et recommandations, 2003.