Second Atelier Qualité des Données et des Connaissances

DKQ 2006

en conjonction avec ECG 2006

17 Janvier 2006, Villeneuve d'Ascq, France

Après le succès du premier atelier *Qualité des Données et des Connaissances* - DKQ 2005 (*Data and Knowledge Quality*) – organisé à Paris l'année dernière en conjonction avec EGC'2005, la seconde édition de l'atelier est organisée à Villeneuve d'Ascq cette année. Cet atelier se concentre sur les méthodes, les techniques d'analyse et de nettoyage des données, les méthodologies, les approches algorithmiques et les métriques permettant de mesurer la qualité des données et la qualité des connaissances dans le processus de fouille de données et d'extraction de connaissances.

Les problèmes de qualité des données stockées dans les bases, les entrepôts ou puits de données s'étendent à tous les domaines d'application qu'elles soient gouvernementales, commerciales, industrielles ou scientifiques. 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 critique 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 liée à la qualité des données, et une dimension subjective liée aux intérêts du spécialiste.

Nous remercions chaleureusement les deux conférenciers invités : Monica Scannapieco (Univ. Roma LA SAPIENZA) et Einoshin Suzuki (Yokohama National University) ainsi que les auteurs et les membres du comité de relecture pour leur contribution au succès de l'atelier DKQ2005.

Laure Berti-Équille et Fabrice Guillet
Organisateurs de DKQ 2006

Programme de l'atelier Qualité des Données et des Connaissances - 17 janvier 2006

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

9h30-10h45: Conférence invitée - **Object Identification and Data Integration: Data Quality Issues in Database Research**, **Monica Scannapieco** (**Univ. Roma La Sapienza**, **Italy)** ..... p. 5

10h45-11h: Pause

11h - 12h: Session 1 - Qualité des données : Théorie et Pratiques Opérationnelles

- Data Profiling versus Data Quality Problems, Paulo Oliveira, Fátima Rodrigues (Polytechnic of Porto, Portugal), Pedro Henriques (University of Minho, Portugal) ..... p. 9
- Delphine Clément (Hewlett-Packard), Brigitte Laboisse (AID), Audit des données du 'Référentiel Client' Hewlett Packard..... p.16

12h30-14h: Pause déjeuner

14h-15h: Conférence invitée – **Discovering Interesting Exception Rules with Rule Pair, Einoshin Suzuki (Yokohama National University, Japan)** ..... p. 7

15h- 16h: Session 2 - Qualité et interprétation des règles d'association

- Normalisation d'une mesure probabiliste de qualité des règles d'association : étude de cas, Daniel Feno, Jean Diatta (Univ. de la Réunion), André Totohasina (Univ. d'Antsiranana – Madagascar) .....p. 25
- Aide à l'interprétation des règles d'association composées, Martine Cadot UHP/LORIA, Pascal Cuxac et Claire François (INIST-CNRS) ..... p. 31

16h-16h30: Pause

16h30-17h30 : Session 3 – Mesures d'intérêt des règles d'association

- Agrégation de mesures d'intérêt de règles d'association, Jean-Pierre Barthélemy, Angélique Legrain, Philippe Lenca (TAMCIC, ENST Bretagne), Benoît Vaillant (IUT de Vannes) ..... p. 38
- Extraction de mesures d'intérêt représentatives pour le post-traitement des règles d'association, Xuan-Hiep Huynh, Fabrice Guillet, Henri Briand, LINA, Ecole polytechnique, Univ. de Nantes) ..... p. 45

Object Identification and Data Integration: Data Quality Issues in Database Research

Dr. Monica Scannapieco
Dipartimento Informatica e Sistemistica
Univ. Roma LA SAPIENZA, Italy
monscan@dis.uniroma1.it

Abstract

Data quality is a multidisciplinary research field, where statistics, management and computer science issues are investigated since early 60's. If focusing on computer science, data mining, machine learning, and, of course, databases are the major areas that have contributed to the research on data quality. This talk will give an overview of the state of the art and of the current data quality research challenges in the database area. Two major topics will be analysed and presented, namely object identification and data integration.

Object identification is a data quality activity that, given different representations of the same real world object, has the purpose of classifying such representations as match, non-match, or possible match.

Data integration, besides solving traditional kinds of heterogeneities like schema and model conflicts, must also deal with conflicts occurring on data values provided by different sources, namely it must deal with an instance level heterogeneity. This type of heterogeneity can be caused by accuracy, completeness, currency, and consistency errors, to mention only few types of errors. When integrating data, such type of conflicts must be solved, by means of a specific instance-level conflict resolution activity, otherwise the whole integration process cannot be correctly terminated.

Bio

Dr. Monica Scannapieco obtained her Master and Ph.D. degrees in Computer Engineering from the Università di Roma LA SAPIENZA, Italy. She is currently research associate and lecturer at the Dipartimento di Informatica e Sistemistica of the same University. Her research interests include data quality models and techniques, cooperative systems for eGovernment, XML data modeling and querying.

Discovering Interesting Exception Rules with Rule Pair

Einoshin SUZUKI
Yokohama National University, Japan
esuzuki@acm.org

Abstract

In this talk, I summarize a part of our 10-year endeavor for exception rule discovery with rule pair. An exception rule, which is defined as a deviational pattern to a strong rule, exhibits unexpectedness and is sometimes extremely useful.

Previous discovery approaches for this type of knowledge can be classified into a directed approach, which obtains exception rules each of which deviates from a set of user-prespecified strong rules, and an undirected approach, which typically discovers a set of rule pairs each of which represents a pair of an exception rule and its corresponding strong rule. We call such a pair of rules a rule pair.

It has been pointed out that unexpectedness is often related to interestingness. In this sense, an undirected approach is promising since its discovery outcome is free from human prejudice and thus tends to be highly unexpected.

However, this approach is prohibitive due to extra search for strong rules as well as unreliable patterns in the output. In order to circumvent such difficulties we have proposed several methods. Our results mainly concern interestingness measure, reliability evaluation, practical application, parameter reduction, and knowledge representation.

Bio

Pr. Einoshin Suzuki is currently full professor at Department of Electrical and Computer Engineering at the Yokohama National University, Japan. His research interests include Machine Learning, Knowledge Discovery, Data Mining, and Knowledge Engineering. He has served as a PC Chair of DS 2004 and a PC vice Chair of ICDM 2004.

Data Profiling versus Data Quality Problems

Paulo Oliveira*, Fátima Rodrigues*, Pedro Henriques**

* Institute of Engineering – Polytechnic of Porto
Rua Dr. António Bernardino de Almeida
4200-072 Porto – Portugal
{pjorge,fr}@dei.isep.ipp.pt
http://www.dei.isep.ipp.pt/~{pjorge,fr}
**University of Minho
Campus de Gualtar
4710-057 Braga – Portugal
prh@di.uminho.pt
http://www.di.uminho.pt/~prh

Abstract. Data suffers from several quality problems. Before starting a data-driven project, data must be assessed to check whether the required quality is assured. Data profiling is a new emerging technology whose intention is to detect the data quality problems. In this paper, we present an overview about data profiling and assess the coverage that is given to data problems. The assessment is based on our proposed taxonomy of data quality problems.

1 Introduction

Nowadays, the importance and usefulness of data is recognized by everyone. However, it is also known that usually data suffers from several quality problems (Orr, 1998). These problems affect all data-driven projects (*e.g.*, data warehouse, data mining). The quality of the input data strongly influences the quality of the results ("garbage in garbage out" principle). Therefore, before starting a new project, the quality of data must be checked and improved if necessary. The Data Quality (DQ) problems need to be detected and corrected.

Data Profiling (DP) is a recent concept raising many interests in the business field. It is defined as the application of data analysis techniques to existing data sources for the purpose of determining the actual content, structure, and quality of the data (Olson, 2003). As outputs, DP produces complete and accurate metadata by relying on the data for reverse-engineering the metadata, and identifies the rows and values that violate the proper definitions of the data.

This paper assesses the coverage that DP gives to the detection of DQ problems, having our taxonomy of problems on relational data as basis. Our purpose is to understand how far DP is able to go in detecting the DQ problems. We also expect to provide some valuable guidelines for further research and enhancement of DP tools.

This paper is organized as follows. Section 2 presents our taxonomy of DQ problems. Section 3 presents an overview about DP. In Section 4 the coverage given to DQ problem by DP is assessed. Finally, in Section 5, the conclusion and future work directions are described.

2 Taxonomy of data quality problems

Figure 1 presents the organization model of relational data: (*i*) data is distributed by or replicated into multiple sources (*e.g.* databases); (*ii*) a data source is composed of several relations and relationships are established among them; (*iii*) a single relation is made up of several tuples; and (*iv*) a tuple is composed by a predefined number of attributes. This model results in a hierarchy of four levels of data granularity: multiple data sources; multiple relations; single relation (multiple tuples); and attribute/tuple.

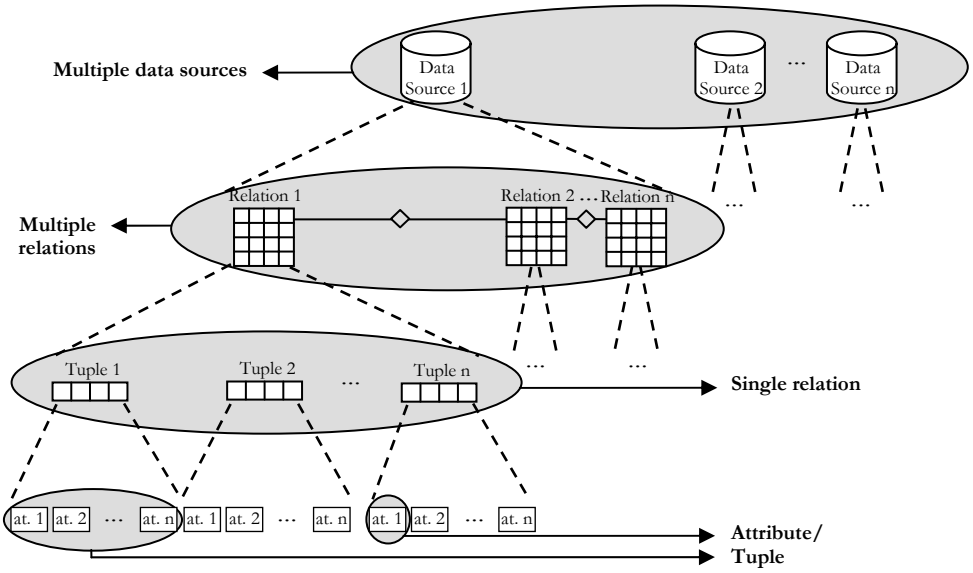

FIG. 1 – *Organization model of relational data.*

We have identified the DQ problems and created a taxonomy based on this model, as presented in detail in (Oliveira et al., 2005). Using real-world data, we thoroughly analyzed each granularity level to detect the specific DQ problems. The analysis was based on the fundamental elements of this model of data organization (*e.g.*, data types, data representation structure, relationships). The taxonomy covers the problems that affect data represented in a tabular format, *i.e.*, at least in the first normal form. It extends and complements the taxonomies proposed by (Rahm et Do, 2000) and (Müller et Freytag, 2003). All the data types are considered, except multimedia data, since it requires special manipulation. DQ problems among two or more relations are restricted just to the problems related with the instances (values) of the data. There are also other problems related with the data schemas (*e.g.*, different attribute names for the same data – synonyms; equal attribute names for different data – homonyms) (Kashyap et Sheth, 1996).

The attribute/tuple granularity level is divided into three groups of related DQ problems. These problems were found by analyzing the value(s) of: (*i*) a single attribute of a single tuple; (*ii*) a single attribute in multiple tuples (a column); and (*iii*) multiple attributes of a single tuple (a row).

Table 1 presents our taxonomy of DQ problems. The problems are organized by granularity level of occurrence.

	Attribute/Tuple			Single Relation	Multiple Relations	Multiple Sources
	Attribute	Column	Row			
Missing value	x					
Syntax violation	x					
Domain violation	x					
Invalid substring	x					
Misspelling error	x					
Ambiguous value	x					
Incorrect value	x					
Violation of business rule	x	x	x	x	x	x
Uniqueness violation		x				
Existence of synonyms		x			x	x
Violation of functional dependency			x			
Approximate duplicate tuples				x		x
Inconsistent duplicate tuples				x		x
Referential integrity violation					x	x
Incorrect reference					x	x
Heterogeneity of syntaxes					x	x
Circularity among tuples in a self-relationship					x	
Heterogeneity of measure units					x	x
Heterogeneity of representation					x	x
Existence of homonyms						x

TAB. 1 – *DQ problems by granularity level.*

A brief definition of each DQ problem included in the taxonomy is presented next:

- **Missing value** – absence of value in a mandatory attribute;
- **Syntax violation** – the value violates the predefined syntax of the attribute;
- **Domain violation –** the value violates the domain of valid values; when the attribute data type is string, this DQ problem may be further detailed:
  - **Invalid substring** – the attribute value violates the domain, but a substring of it is valid; this means that the remaining substring is invalid;
  - **Misspelling error** – the attribute value contains a misspelled error;
  - **Ambiguous value –** the attribute value is an abbreviation or acronym.

- **Incorrect value** – the domain of valid values is respected, but the attribute contains a value which is not the correct one (*e.g.*, the value is outdated as result of its dynamic nature);
- **Violation of business rule** – a given business domain rule (enterprise rule) is violated; this problem occurs at all granularity levels;
- **Uniqueness violation** – two or more tuples representing different real-world entities have the same value in a unique value attribute;
- **Existence of synonyms** – use of syntactically different values with the same meaning, within an attribute or among related attributes from multiple relations;
- **Violation of functional dependency** – an existent functional dependency among two or more attributes is violated by the values of a tuple;
- **Approximate duplicate tuples** – the same real-world entity is represented in more than one tuple; the tuples may be equal or equivalent (with minor differences);
- **Inconsistent duplicate tuples** – the same real-world entity is represented in more that one tuple, but there are inconsistencies among the values of the attributes;
- **Referential integrity violation** – there is a value in a foreign key attribute which does not exist in the related relation as a primary key value;
- **Incorrect reference** – the referential integrity is respected, but the foreign key contains a value which is not the correct one (*e.g.*, the reference is wrong as result of an erroneous data entry);
- **Heterogeneity of syntaxes** – there are different representation syntaxes among related attributes, within a data source or among data sources;
- **Circularity among tuples in a self-relationship** – there is a loop among two (direct circularity) or more (indirect circularity) related tuples in a self or reflexive-relationship (*e.g.*, suppose that a product may be a sub-product of another product; this is stored in the relation *product* with the following schema: *product*(*code*, *sub-product_code*); a circularity results from the tuples: *product*(*x*, *y*) and *product*(*y*, *x*));
- **Heterogeneity of measure units** – different measure units are used among related attributes, within a data source or among data sources;
- **Heterogeneity of representation** – different sets of values are used to code the same real-world property, within a data source or among data sources;
- **Existence of homonyms** – use of syntactically equal values with different meanings, among related attributes from multiple data sources.

3 Data profiling

Data Profiling (DP) is defined as the application of data analysis techniques to existing data sources for the purpose of determining the actual content, structure, and quality of the data (Olson, 2003). Sometimes, the terms data discovery and data auditing are also used with the same purpose. DP is different from data analysis performed for business use, done by decision support systems and data mining tools. These ones acquire information/knowledge *from* the data. Instead, DP acquires information *about* the data. It employs discovery and analytical techniques to find characteristics of the data. DP is supposed to be the first step to perform on all DQ assessment projects that intend to detect and correct the data problems. Every data-driven initiative (*e.g.*, customer relationship management deployment, data warehouse development, data mining project, data integration project) should start with DP.

A DP tool produces complete and accurate metadata by relying on the data for reverse-engineering the metadata. The other output of a DP tool is the identification of the rows and values that violate the proper definitions of data. DP tools have interactive drill-down capabilities that allow analyzing the underlying data and viewing the DQ problems instances.

The approach followed by DP tools is made up of a series of bottom-up steps, starting at the most atomic level of the data and moving progressively to higher levels of structure over the data. DQ problems are corrected at each level before moving to the next higher level. The major steps performed by a DP tool are presented in the following paragraphs.

Column property analysis. Column profiling helps to determine whether the data in a column meets the user's expectations for that data. This step analyses the values stored in each single column and infers detailed characteristics, such as: data type and size; the values or the range of values found; cardinality; null and uniqueness characteristics. This kind of analysis also includes the use of pattern matching techniques that allow checking whether the values in a column have the expected format (*e.g.*, if a field is all numeric, if it has consistent lengths). Some basic statistics about the column (*e.g.*, minimum and maximum values, mean, median, mode, standard deviation) are also produced. Statistics are useful in all types of data, especially in numeric data. Column profiling also generates frequency counts and helps to identify outliers. Frequency counts presents how values are related according to their occurrences. Outlier detection determines the data values that are remarkably different from the other values, showing the highest and lowest values.

Structure analysis. It allows dependency profiling by analyzing data across rows, comparing values in every column with values in every other column, and infers all functional dependency relationships that exist among attributes within each relation. During this process, it identifies gray-area dependencies that are true most of the time, but not all the time. Usually this is an indication of a DQ problem. It also identifies primary keys. Structure analysis also allows redundancy profiling by comparing data among tables of the same or different data sources, determining which columns have affinity to other columns, *i.e.*, overlapping or identical sets of values. It identifies columns containing the same information but with different names (synonyms), and columns that have the same name but different business meaning (homonyms). It also helps determine which columns are redundant and can be eliminated, and which are necessary to connect information among tables, *i.e.*, foreign keys needed for maintaining referential integrity.

Business rule analysis. A business rule specifies a condition that must hold true across one or more columns at any point in time, involving a single or multiple relations (*e.g.*, employees must have at least 18 years old). These rules are collected from the business specialist, converted to executable logic and then tested against the data. A business rule may provide domain checking, range checking, look-up validation or specific formulas. As result, the rows that violate the rules are identified. A robust DP tool is able to build, store and validate the organization's unique business rules.

The existing DP solutions are developed exclusively by commercial companies. At least, we were not able to find academic prototypes. There are several tools in the market (*e.g.*, DataFlux DP (DataFlux), PowerCenter DP (Informatica), i/Lytics DP (i/Lytics), Axio DP (Evoke)). The available technical information allowed us to conclude that basically they have

the same detection capacities. The algorithms used, the user's interface, and the access capabilities to different types of sources are the principle differences among these tools.

4 Data profiling coverage to data quality problems

Table 2 presents the detection support to DQ problems presently given by the analyzed DP tools. The letters A, B, C and D respectively represent the tools: DataFlux, PowerCenter, i/Lytics and Axio. A letter within a cell means that the corresponding tool is able to detect the DQ problem. An 'x' means that none of the four tools is capable to detect it.

	Attribute/Tuple			Single Relation	Multiple Relations	Multiple Sources
	Attribute	Column	Row			
Missing value	A B C D					
Syntax violation	A B C D					
Domain violation	A B C D					
Invalid substring	x					
Misspelling error	x					
Ambiguous value	A D					
Incorrect value	x					
Violation of business rule	A B C D	A B C D	A B C D	A B C D	A B C D	x
Uniqueness violation		A B C D				
Existence of synonyms		x			x	x
Violation of functional dependency			A C D			
Approximate duplicate tuples				x		x
Inconsistent duplicate tuples				x		x
Referential integrity violation					A C D	x
Incorrect reference					x	x
Heterogeneity of syntaxes					x	x
Circularity among tuples in a self-relationship					x	
Heterogeneity of measure units					x	x
Heterogeneity of representation					x	x
Existence of homonyms						x

TAB. 2 – *Coverage given by DP tools to DQ problems.*

The DQ problems that occur at the attribute/tuple granularity level are the ones which have the best coverage given by DP tools. The coverage to the problems that occur at the

level of: the relation, multiple relations and multiple data sources is insufficient. There are several problems that DP is not able to detect. None of the DQ problems that occur at the level of multiple data sources is supported.

5 Conclusion

This paper has presented an overview about DP and the coverage that this emergent technology gives to the detection of the DQ problems. For the assessment, we have used our taxonomy of DQ problems, previously proposed in the literature.

DP tools support the detection of an interesting set of problems. However, the number of problems covered is just nearly half of the problems included in the taxonomy. This means that many future developments are required for this technology to become full-featured in the detection of DQ problems. Some data cleaning tools allow the detection of specific problems currently not supported by DP tools (*e.g.*, duplicates detection). In this paper our purpose was just to assess the coverage that is presently given by DP tools to detect the DQ problems. It would definitely be important to include the same kind of support in DP tools, to achieve complete solutions that allow the detection of a larger number of DQ problems (ideally all), in a single integrated environment. For that purpose, a special attention should be given to the problems that occur at the highest granularity levels, *i.e.*, multiple relations and multiple data sources. At the level of multiple relations, the support to detect the problems is weak and it just doesn't exist at the level of multiple data sources. A DP tool should be able to deal with more than one data source at the same time to detect the DQ problems that may exist among them. The existing solutions are centered on a single source.

As future work, we intend to explore some of the research opportunities left open by the current state-of-the-art in the DP technology (*e.g.*, the DQ problems currently not covered).

References

DataFlux. http://www.dataflux.com

Evoke. http://www.evokesoft.com

Informatica. http://www.informatica.com

i/Lytics. http://www.innovativesystems.com

Kashyap, V. et A. Sheth (1996). Schematic and Semantic Similarities Between Database Objects: a Context-Based Approach. *Very Large Databases Journal*, 5(4):276–304.

Müller, H. et J.-C. Freytag, (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Technical Report HUB-IB-164*. Humboldt University, Berlin.

Oliveira, P., F. Rodrigues, and P. Henriques (2005). A Formal Definition of Data Quality Problems. In *Proceedings of the 10th Int. Conference on Information Quality*. 13-26.

Olson, J. (2003). *Data Profiling: The Accuracy Dimension*. Morgan Kaufmann Publishers.

Orr, K. (1998). Data Quality and Systems Theory. *ACM Communications*, 41(2):66-71.

Rahm, E. et H. Do (2000). Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*, 24(4):3-13.

Audit des données du 'Référentiel Client' Hewlett Packard

Delphine Clément
Data Quality Analyst
Hewlett-Packard
14 rue du Général Caunègre – 40000 MONT DE MARSAN
delphine_clement@hp.com

Brigitte Laboisse
Directeur technique
A.I.D.
4 rue Henri Le Sidaner – 78000 VERSAILLES
blaboisse@aid.fr
http://www.aid.fr

Résumé : Maîtriser, améliorer la qualité de données client dans un contexte international avec de multiples systèmes d'information est un challenge permanent auquel est confronté l'équipe Qualité de Données Client chez Hewlett Packard. Cet article présente la mise en place, les métriques définies ainsi que les raisons de leur choix , leur implémentation sur six systèmes d'information clients/prospects audités au niveau mondial. Les freins, à la fois techniques (manque de référentiels de contrôle dans certains pays), opérationnels (faire adhérer les responsables de systèmes locaux aux résultats de l'audit) sont également abordés.

1 Contexte

1.1 Organisation « qualité de données » chez Hewlett Packard

L'équipe 'Customer Data Integrity' fonctionne de manière centralisée (niveau mondial). Elle est rattachée au programme 'Customer Knowledge Management and Data Stewardship', qui lui-même dépend du groupe 'Internet and Marketing Services'. Nous fonctionnons avec des relais régionaux. 3 régions sont définies : 'EMEA' (Europe, Moyen Orient et Afrique), 'APJ' (Asie-Pacifique, Japon) et Amériques (Canada, Etats Unis, Amérique du Sud).
La mission de notre équipe Qualité centralisée est de :

- répertorier et harmoniser les différents processus qualité tels que normalisation d'adresses, dédoublonnage, …
- définir des standards qualité (indicateurs qualité, objectifs par indicateurs et par type de données, selon leur criticité métier)
- vérifier et mesurer l'adoption de ces standards au niveau des régions/pays
- aider les régions, pays et départements en général à définir des priorités et des plans d'action
- proposer des services de mesure et d'amélioration de la qualité des informations
- assortir ces services d'outils qualité (normalisation d'adresses par exemple), de ressources (équipe dédiée à la maintenance des données) et de moyens (budget qualité annuel)

1.2 Le programme 'Référentiel Client'

Ce programme est stratégique pour Hewlett Packard. Il est lui aussi rattaché à l'équipe 'Customer Knowledge Management and Data Stewardship'.

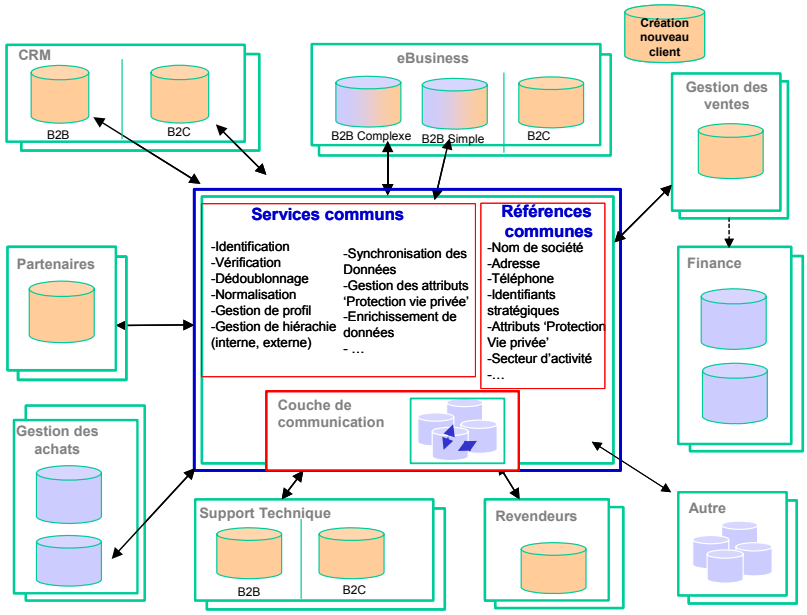

Tableau 1 : Architecture 'Référentiel Client'

'Référentiel Client' vise à dresser un inventaire de toutes les sources d'informations client au niveau Hewlett Packard monde, de leurs spécificités fonctionnelles et de leur architecture. Cet inventaire a pour but de définir en nombre limité des systèmes maîtres représentant chacun un bloc fonctionnel (Gestion des achats, Support Technique par exemple,.. présentés sur le tableau 1). Ces systèmes maîtres sont, dans leur bloc fonctionnel, la source d'information et la référence pour les autres systèmes en aval. Le programme 'Référentiel Client', comme indiqué au centre du tableau 1, apporte :

- des services : dédoublonnage, normalisation,…
- une référence en terme de dictionnaire et de contenu : nom de société, adresse,…

afin de permettre une vue centralisée et homogène des données clients chez Hewlett Packard.

1.3 L'objectif de l'audit

L'objectif est d'obtenir un baromètre qualité de l'ensemble des systèmes choisis comme maîtres dans chaque bloc fonctionnel. Comme nous l'avons dit plus haut, ces systèmes sont la référence en matière d'informations clients pour bon nombre d'autres systèmes en aval.

Il est donc nécessaire d'avoir une connaissance de la qualité de départ des références, de produire des métriques communes permettant de piloter la qualité dans le temps et d'amener les systèmes à nos objectifs par des actions correctives puis préventives ciblées.

2 Implémentation de l'audit

2.1 Le périmètre de l'audit

Parmi les différentes dimensions de la qualité de données présentées dans les ouvrages tels que Berti-Equille (1999) Huang et al. (1998), Loshin (2001) et Redman (2001), seules 2 dimensions sont analysées dans cet audit :

- La qualité intrinsèque de la donnée
- La qualité relative aux données homologues

Les autres dimensions, telles que la qualité de la représentation de la donnée par le système ou la qualité relative à l'utilisateur, seront abordées dans des étapes ultérieures.

Déclinaison de ces 2 dimensions :

- La qualité intrinsèque des données de chaque système : pour les attributs inclus dans le dictionnaire de référence tels que le nom de société, l'adresse, le téléphone,…, des métriques ont été définies pour mesurer :
  - Le taux de remplissage
  - L'exactitude
- L'unicité de la représentation du client : la société 'A.I.D. 4 rue Henri Le Sidaner à Versailles' est-elle présente une seule fois dans le système analysé ? L'unicité a été mesurée en terme de sociétés, mais également en terme de personnes.
- Le taux de recouvrement des différents systèmes : nombre de clients présents dans le système des commandes et ayant appelé le support technique par exemple. Parmi ces clients communs, combien sont reconnus en tant que tels par les identifiants communs actuellement en place ?
- La synchronisation des différents systèmes : sur les clients reconnus par des identifiants communs entre 2 systèmes, quel taux d'erreur observe-t-on ? Par rapport aux métriques 'taux de recouvrement' présentées ci-dessus qui mesurent les clients communs non liés, il s'agit là de mesurer l'exactitude des liaisons : S'agit-il de la même entité ?.
- La synchronisation des attributs de 'Protection de la vie privée' : l'objectif est de valider/quantifier les divergences éventuelles entre les systèmes. Par exemple, si Raoul Dupont a déclaré qu'il ne veut pas recevoir d'emails dans le système A, cette donnée est-elle bien répercutée dans le système B où Raoul Dupont est également enregistré ?

Le tableau 2 ci-dessous présente un exemple de publication résultant de l'audit avec les différents axes de mesure. Chaque système analysé est décliné par continent/pays et est comparé :

- aux autres systèmes (benchmark)
- sera comparé dans les prochaines mesures dans le temps avec lui-même (Major Evolutions).

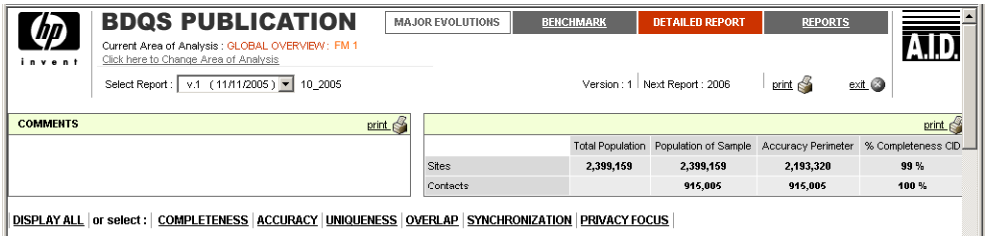

Tableau 2. Publication des résultats

2.2 Exemples de métriques mises en place

2.2.1 Pollution

Pourquoi cette métrique ?

Les taux de remplissage, dans un système d'information, sont souvent pollués par :

- Des caractères parasites, ses pseudo-remplissages 999,AAAAAA,…, des mots 'injurieux' ou pseudo-remplissage : UNKNOWN, NONE ,…

A l'origine, on peut trouver des zones obligatoires dans un logiciel qui incitent/obligent l'utilisateur à polluer par du remplissage néfaste. Il est également possible d'avoir le fruit de données provenant de questionnaires WEB, souvent plus frustres en contrôle en entrée. La base est alors polluée, avec des conséquences néfastes dans l'image de la société lorsqu'elle communique à un client/prospect.

Méthode de calcul

Des dictionnaires de mots, caractères, répétitions sont décrits par variable /pays avec un algorithme très simple de détection dans les enregistrements.

2.2.2 Contrôle des emails

Pourquoi cette métrique ?

L'objectif est de détecter, par anticipation, avant l'opération d'e-mailing où ils vont être utilisés, le nombre d'emails faux. Contrairement aux courriers ou téléphones, le fait d'envoyer un email faux a de faibles conséquences financières, en revanche l'emailing ne permettra, en aucun cas, de trouver le nouvel email de la personne : pas de facteur 'intelligent' ou de service de transfert de courrier, pas de disque avec nouveau téléphone. Donc, anticiper a réellement un sens pour gagner en temps et déclencher des actions pour qualifier en emails.

Méthode de calcul

En terme d'emails, à notre connaissance, il n'existe pas de moyen fiable actuellement pour valider l'email, si ce n'est de l'envoyer. En effet, les fonctions de demande d'existence disponibles dans les moteurs SMTP sont en général annulées par les responsables réseau. Les

algorithmes mis en place n'ont pas comme objectif de valider un email, *mais plutôt de détecter les emails faux : le périmètre restant sera supposé 'vraisemblable'*. Indicateurs mis en place :

- Détection de syntaxe incorrecte : pas de @, présence de caractères parasites,…
- Contrôle du nom de domaine : la validité du nom de domaine est faite on-line .
- Contrôle de la cohérence nom/prénom : spécifiquement en BtoB, il existe très souvent une 'charte' d'attribution des emails dans l'entreprise : début du prénom, nom,... Un certain nombre de formats ont été référencés afin de produire un indice de cohérence.

2.2.3 Contrôle de l'adresse

Pourquoi cette métrique ?

Avec le téléphone et l'email, l'adresse fait partie des éléments qui permettent de communiquer par un canal avec les clients/prospects. Avoir une adresse exacte signifie, en général :

- Une adresse qui existe : la ville, le code postal, la rue, le numéro dans la rue
- Une adresse libellée correctement : pas de fautes d'orthographe
- Une adresse aux normes postales : selon les pays, La Poste locale, afin d'améliorer la distribution du courrier en efficacité (lecture optique par exemple), a mis en place une norme de définition de l'adresse : positionnement du code postal,ville dans les lignes adresses, rue, bâtiment,.. ainsi que respect des référentiels rue/ville/code postal mis à disposition.

Les intérêts d'avoir une adresse exacte sont nombreux : moins de courriers inaboutis (coût important), tarifs postaux préférentiels, meilleure efficacité du dédoublonnage.

Méthode de calcul

Le contrôle, la conformité des adresses peuvent être contrôlés par des logiciels de normalisation d'adresse internationaux, tels que First Logic™. Dans le cas de notre projet, c'est ce dernier que nous avons utilisé. L'outil logiciel comporte plusieurs phases :

- recherche dans l'adresse des éléments clés : code postal, ville, rue, numéro dans la rue, boite postale,…
- validation par rapport aux référentiels géographiques locaux
- si la validation n'aboutit pas directement, un certain nombre d'approximations, transformations sont opérées pour corriger l'adresse et la retrouver dans les référentiels.

L'indice produit aura la valeur :

- Adresse correcte , pas de modification nécessaire
- Adresse correcte, mais après modification nécessaire
- Adresse incorrecte

2.2.4 Contrôle du téléphone

Pourquoi cette métrique ?

La présence de faux téléphones est coûteuse, pénalisante pour les centres d'appels lorsqu'ils doivent procéder à des appels sortants. En appel entrant, si le coût des 'faux' est moins élevé, il existe également : de nombreux standards téléphoniques sont équipés de la reconnaissance des numéros de téléphone entrants, afin que l'opérateur ait directement accès à la fiche du client, du prospect. Les faux téléphones diminuent l'efficacité du système. Enfin, dans les systèmes d'information qui n'utilisent pas les téléphones, l'erreur , si elle ne génère pas de coûts directs, influe sur la qualité générale de la base, en particulier sur le taux de doubles : le téléphone est un élément de comparaison important pour retrouver les doubles sites dans une base.

Méthode de calcul

La validation d'un téléphone peut être faite selon plusieurs méthodes :

- Test du téléphone directement, par un opérateur : efficace, mais très coûteux.
- Recherche dans un annuaire
- Contrôle syntaxique simple : c'est cette dernière méthode que nous avons utilisée.

Un numéro de téléphone est composé de :
Code international d'appel (00 en général) - Code téléphonique du pays (33 pour la France) - Code troncature (en général 0 : à concaténer au numéro uniquement pour les appels à l'intérieur du pays) - Code zone (correspond en général à une localisation géographique dans le pays : ensemble de départements en France) - Numéro de téléphone
Par exemple 00 – 44 – 0 – 20 - 72987174
L'algorithme mis en place se base sur des tables de référence compilées qui incluent en général :

- D'une manière quasi-exhaustive les longueurs possibles pour le code zone, le numéro de téléphone.
- De manières beaucoup plus inégales la liste des codes zones avec leur description en terme de villes par exemple.

2.2.5 Unicité

Pourquoi cette métrique ?

L'objectif est de mesurer le nombre de sociétés, de personnes présents plus d'une fois dans le système d'information sans être identifiés en tant que même entité.

Méthode de calcul

Un logiciel de dédoublonnage du marché (Proxim™ , développé par A.I.D. dans notre cas) est mis en œuvre. Le principe est le suivant : dans une zone de comparaison (Ville, E-mail, Téléphone,…), les enregistrements sont comparés sur des variables 'pertinentes' telles que :

- Pour les sociétés : nom de l'entreprise, adresse, téléphone, numéro de TVA, …
- Pour les personnes : nom, prénom, email, adresse, téléphone,…

La comparaison est faite à partir d'algorithmes tels que :

- Distance d'édition : Wood (1992)
- Présence de lettres communes : soit 2 chaînes de caractères S1 et S2, nc le nombre de lettres communes entre les 2 chaînes, l'indice de proximité Ind=nc / nombre maximum de caractères (S1,S2). Un indice dérivé Ind' consiste à diviser par le minimum.
- De nombreux autres algorithmes publics existent, Berti-Equille (2005). Les logiciels de dédoublonnage tels que Proxim™ intègrent également en général des algorithmes propriétaires.

Zone de comparaison : les comparaisons des enregistrements sont restreintes à un sous périmètre pour des raisons de performance. En effet, le nombre de comparaisons n(n-1)/2 peut prendre rapidement un peu de temps sur des volumes importants (n = nombre d'enregistrements de la population à dédoublonner, peut valoir plusieurs millions d'enregistrements). La combinaison de conditions en ET, OU entre les différentes variables, leurs indices de proximité, et également en utilisant les attributs qualité Wang (2000) des variables va permettre de décider si 2 enregistrements sont doublons ou pas.

2.3 Freins rencontrés lors de la mise en place des métriques

Mettre en place les métriques définies plus haut :

- Dans un contexte international
- Avec plusieurs systèmes d'information à mesurer

n'est pas neutre. Une illustration en est la notion de référentiels de contrôle : postaux, téléphones, prénoms, dictionnaires de mots de pollution,… La qualité de ces référentiels, leur exhaustivité sont très inégaux d'un pays à l'autre.
Nous rencontrons en général le même biais :

- Les enregistrements considérés comme corrects le sont avec un taux d'erreur proche de 0.
- Les enregistrements considérés comme erronés le sont à tort, dans une proportion très variable selon le pays, la qualité des référentiels , l'adéquation de l'algorithme.

Pour pallier ce défaut, nous utilisons la méthode suivante :
Un échantillon aléatoire est tiré par pays dans les enregistrements notés erronés. Cet échantillon est validé manuellement par des personnes locales capables :

- de téléphoner pour valider
- avoir une meilleure connaissance des prénoms
- de trouver l'information dans des référentiels locaux moins connus

Le système apprend, s'enrichit. Les résultats publiés en terme d'indicateurs sont extrapolés selon le contrôle manuel de l'échantillon.

3 Résultats

3.1 Tableau de bord

Afin de porter les résultats qualité à la connaissance de nos décideurs, nous avons positionné les systèmes et leurs résultats par métrique dans un tableau de bord, divisé en trois régions (EMEA, Amériques, AP) . Les systèmes étaient ainsi comparés les uns par rapport

aux autres mais également par rapport à un objectif de résultat. Sur chaque métrique et par région, le système pouvait être vert car égal ou supérieur à l'objectif de résultat, jaune car en deçà de l'objectif mais supérieur à 50% (85% dans le cas du taux de doublons) ou rouge car en deçà de 50% (85% dans le cas du taux de doublons). La présentation d'un tel tableau de bord, assorti de commentaires, a permis de créer une véritable prise de conscience de l'intérêt et de l'importance de la qualité des données parmi les décideurs. Nous avons donc recueilli leur soutien pour la poursuite du projet et la mesure de l'évolution.

3.2 Mise en oeuvre opérationnelle

Nous avons organisé une revue individuelle plus approfondie avec les responsables de chacun des domaines fonctionnels. Nous leur avons présenté au niveau Région/Pays le détail des métriques. Nous avons par ailleurs gagné leur confiance et leur attention en montrant des échantillons de données erronées. Nous avons essayé de comprendre ensemble quels étaient les flux ou traitements qui pouvaient expliquer tel ou tel autre résultat. Nous avons partagé avec eux nos recommandations. Ainsi, au fil de la discussion, chaque responsable de domaine a défini ses priorités d'action immédiate : remplissage et exactitude d'identifiants stratégiques, protection de la vie privée, taux de doublons, etc… Sans exception, tous les utilisateurs des systèmes ont salué l'existence et l'intérêt d'un tel projet.

Conclusion

Ce premier audit était un challenge :

- De multiples sources de données avec nécessité de prendre connaissance des règles de gestion spécifiques à chaque système .
- Plus de 100 pays à couvrir, avec un nombre important de jeux de caractères différents.
- Des délais très courts d'implémentation.
- Une attente forte de la part des responsables de domaines d'obtenir des résultats directement opérationnels.

Challenge réussi, la prochaine étape prévue est de comparer les métriques dans le temps : prochain audit en février 2006. Il est également apparu une évolution nécessaire dans les analyses : définir des segments de clients. La limite des métriques produites actuellement est qu'elles restent assez techniques tant qu'elles ne sont pas déclinées par segment client. Ce dernier correspond en général à une valeur marketing différente, à des niveaux et des moyens très hétérogènes dans la qualité de données, à des niveaux d'exigence (objectifs) spécifiques. Leur introduction permettra d'adapter les recommandations/résultats à la réalité métier de Hewlett Packard.

Mener un audit a été possible de part l'organisation de Hewlett Packard, avec une cellule qualité de données existante et rattachée à un niveau hiérarchique élevé. Cela correspond à une réelle préoccupation de la qualité de données client dans cette société, ce qui est encore, à notre connaissance, un phénomène précurseur.

Lexique

Customer Data Integrity : Intégrité des données Client

Customer Knowledge Management and Data Stewardship : Gestion de la connaissance client et services sur les données

Internet and Marketing Services : Services Internet et Marketing

B2B : Business to Business : Vente aux entreprises

B2C : Business to Consumer : Vente au grand public

Benchmark : Comparaison

Major evolutions : Evolutions majeures

Références

Berti-Equille, L. (1999), *La qualité des données et leur recommandation : modèle conceptuel, formalisation et application à la veille technologique :* Thèse de doctorat, Université de Toulon et du Var.

Berti-Equille, L. (2005), *Journées CRM & Qualité des Données au CNAM – Qualité des données multi-sources : un aperçu des techniques issues du monde académique.*

Huang, K. T., Lee, Yang W., Wang R. (1998), *Quality Information and Knowledge*. Prentice Hall

Loshin, D. (2001), *Enterprise Knowledge Management : the Data Quality Approach.* Morgan Kaufmann Publishers

Redman, T. C. (2001), *Data Quality : The Field Guide.* Digital Press

Wang, R. (2000), *Data Quality*. Kluwer Academic Publishers

Wood, D. (1992), *Data structures, algorithms and Performance.* Addison-Wesley

Summary

Measuring, improving the customer data quality level in an international environment with multiple information systems is the daily challenge of the Customer data quality team at Hewlett Packard. This article describes the metrics defined and the reasons of their choices, the implementation for an audit of six worldwide customer/prospect information systems. The brakes, technical (lack of control referentials in some countries), operational (to obtain the buy-in of the results by the data owners) are also described.

Normalisée d'une mesure probabiliste de qualité des règles d'association : étude de cas

Daniel Feno [*,**], Jean Diatta[*]
André Totohasina[**]

[*]Université de la Réunion
15-Avenue René Cassin-B.P. 7151- 97715 Saint-Denis Messag cedex 9 France
{drfeno,jean.diatta}@univ-reunion.fr,
[**]Université d'Antsiranana-BP O - 201-Antsiranana-Madagascar
totohasina@yahoo.fr

Résumé. Cet article concerne les mesures probabilistes de qualité des règles d'association. Nous donnons une condition nécessaire et suffisante pour qu'une mesure probabiliste de qualité soit normalisable. Par ailleurs, nous considérons une trentaine de mesures probabilistes de qualité proposées dans la littérature et montrons que la plupart d'entre celles qui sont normalisables ont la même normalisée.

1 Introduction

De nombreuses de mesures de qualité des règles d'association ont été proposées dans la littérature pour tenir compte de diverses caractéristiques (Piatetsky-Shapiro, 1991; Major et Mangano, 1993; Freitas, 1999; Blanchard et al., 2005). Ainsi, pour aider l'utilisateur final dans son choix, des études expérimentales ont été menées pour l'évaluation de mesures existantes (Tan et al., 2002; Lenca et al., 2004; Lallich et Teytaud, 2004). Dans cet article, nous proposons une approche analytique de comparaison de mesures probabilistes de qualité des règles (MPQ). Cette approche est fondée sur une notion de normalité introduite dans (Totohasina, 2003). Nous caractérisons les MPQ normalisables. Par ailleurs, nous montrons que la plupart des MPQ normalisables ont la même normalisée, à savoir la mesure M_{GK} introduite dans (Guillaume, 2000) et dont les propriétés mathématiques ont été étudiées dans (Totohasina, 2003; Totohasina et al., 2005). Le reste de l'article est organisé de la façon suivante. Dans la section 2, nous introduisons les MPQ et en présentons quelques exemples. La normalisation des MPQ est présentée dans la section 3. Dans la section 4, nous donnons les normalisées de plus d'une vingtaine de MPQ avant de terminer par une brève conclusion.

2 Mesures probabilistes de qualité

Dans cet article, nous nous plaçons dans le cadre d'un contexte binaire $(\mathcal{E}, \mathcal{V})$ où $\mathcal{E}$ est un ensemble fini d'entités et $\mathcal{V} = \{a_1, a_2, \cdots a_m\}$ un ensemble fini de variables boléennes

	a_1	a_2	a_3	a_4	a_5
e_1	1	0	1	1	0
e_2	0	1	1	0	1
e_3	1	1	1	0	1
e_4	0	1	0	1	1
e_5	1	1	1	0	1

TAB. 1 – *Exemple d'un contexte binaire*

définies sur $\mathcal{E}$. Les sous ensembles de $\mathcal{V}$ seront appelés *motifs*. Le tableau 1 présente un contexte binaire ayant 5 variables définies sur un ensemble de 5 entités.

Une *règle d'association* de $(\mathcal{E}, \mathcal{V})$ est un couple (A, B) de motifs, noté $A \rightarrow B$, où B est non vide. Les motifs A *et* B sont appelés respectivement la "*prémisse*" et la "*conclusion*" de la règle $A \rightarrow B$.

Etant donnés deux motifs A et B :

- A' désignera l'événement de l'espace probabilisable $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$, vérifiant le motif A, *i.e.*, $A' = \{e \in \mathcal{E} : \forall a \in A, a(e) = 1\}$;
- $\overline{A}$ désignera la négation de A, *i.e.*, $\overline{A}(e) = 1$ si et seulement si il existe $a \in A$ tel que $a(e) = 0$ ($(\overline{A})'$ est le complémentaire de A').

Dans tout ce qui suit, n désignera le cardinal de $\mathcal{E}$ et p une probabilité sur $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$, définie, pour tout événement E de $\mathcal{E}$, par $p(E) = \frac{|E|}{n}$, où $|E|$ désigne la cardinalité de E.

Une *mesure probabiliste de qualité* des règles d'association (MPQ) est une fonction μ de l'ensemble $\mathcal{P}(\mathcal{V}) \times \mathcal{P}(\mathcal{V})$ à valeurs dans $\mathbb{R}$ telle que $\mu(A \rightarrow B)$ s'exprime en fonction des probabilités $p(A'), p(B'), p(A' \cap B')$.

Exemple 1
- *Le support (Agrawal et al., 1993) défini par* $Support(A \rightarrow B) = p(A' \cap B')$ *est une MPQ symétrique (i.e.* $Support(A \rightarrow B) = Support(B \rightarrow A)$*).*
- *La confiance (Agrawal et al., 1993) définie par* $Confiance(A \rightarrow B) = p(B'/A')$ *est une MPQ non symétrique.*
- *La mesure de conviction (Brin et al., 1997) définie par* $Conviction(A \rightarrow B) = \frac{p(A').p(\overline{B}')}{p(A' \cap \overline{B}')}$ *est une MPQ implicative (i.e.* $Conviction(A \rightarrow B) = Conviction(\overline{B} \rightarrow \overline{A})$*).*
- *La mesure* M_{GK} *(Guillaume (2000)) définie par*

$$M_{GK}(A \rightarrow B) = \begin{cases} \frac{p(B'/A') - p(B')}{1 - p(B')} & si\ A\ favorise\ B\ i.e.\ p(B'/A') > p(B') \\ \frac{p(B'/A') - p(B')}{p(B')} & si\ A\ défavorise\ B\ i.e.\ p(B'/A') < p(B') \end{cases}$$

est une MPQ non symétrique implicative dans le cas où A *favorise* B*.*
- *La surprise (Aze et Kodratoff, 2002) définie par* $Surprise(A \rightarrow B) = \frac{p(A' \cap B') - p(A' \cap \overline{B}')}{p(B')}$ *est une MPQ ni symétrique ni implicative.*

Règle	Support $[0;1]$	Confiance $[0;1]$	M_{GK} $[-1;1]$	Conviction $[0;+\infty[$	Surprise $[-1;1]$
$A \rightarrow B$	0	0	-1	$\frac{4}{5}$	-1
$B \rightarrow C$	$\frac{1}{5}$	1	1	$+\infty$	1
$C \rightarrow D$	$\frac{2}{5}$	1	1	$\frac{6}{5}$	$\frac{1}{2}$
$A \rightarrow C$	$\frac{2}{5}$	$\frac{2}{3}$	$\frac{-1}{6}$	$\frac{3}{5}$	$\frac{1}{4}$

TAB. 2 – *Valeurs prises par quelques MPQ*

3 Normalisation d'une mesure de qualité

3.1 Motivation et définition

Considérons le contexte binaire du tableau 1 et prenons les motifs $A = a_1a_3$, $B = a_2a_4$, $C = a_5$, $D = a_1a_2$. Le tableau 2 présente les valeurs prises par quelques MPQ sur certaines règles de ce contexte.

On constate que les valeurs prises par ces mesures sont distribuées entre -1 et $+\infty$. De plus, certaines mesures ne prennent que des valeurs positives indépendamment du fait que la prémisse favorise ou défavorise la conclusion. L' objectif de la normalisation est alors, d'une part, de réduire et centrer les valeurs d'une MPQ sur l'intervalle $[-1, 1]$, et, d'autre part, de refléter les situations de reférence telles que l'incompatibilité (*i.e.* $A' \cap B' = \emptyset$), la répulsion $(i.e.\ p(B'/A') < p(B'))$, l'indépendance (*i.e.* $p(A' \cap B') = p(A').p(B')$), l'attraction (*i.e.* $p(B'/A') > p(B)$) et l'implication (*i.e.* $A' \subseteq B'$). D'où la définition suivante.

Une MPQ μ sera dite *normalisée* si elle vérifie les cinq conditions ci-dessous :

(1) $0 < \mu(A \rightarrow B) \leq 1$ si A favorise B $(i.e.\ p(B'/A') > p(B))$;
(2) $-1 \leq \mu(A \rightarrow B) < 0$ si A défavorise B $(i.e.\ p(B'/A') < p(B'))$;
(3) $\mu(A \rightarrow B) = 1$ si A implique B $(i.e.\ p(B'/A') = 1)$;
(4) $\mu(A \rightarrow B) = -1$ si A et B sont incompatibles $(i.e.\ p(A' \cap B') = 0)$;
(5) $\mu(A \rightarrow B) = 0$ si A et B sont indépendants $(i.e.\ p(A' \cap B') = p(A').p(B'))$.

Exemple 2 *Les MPQ suivantes sont normalisées.*

– *La mesure M_{GK} (Guillaume, 2000)*
– *La mesure de Zhang (Zhang, 2000)*

3.2 Caractérisation des MPQ normalisables

Dans ce paragraphe, nous ne nous intéressons qu'aux MPQ continues par rapport aux arguments $p(A'), p(B'), p(A' \cap B')$. Nous donnons une condition nécessaire et suffisante pour qu'une telle MPQ soit normalisable. Considérons une MPQ μ et désignons par μ_n la mesure normalisée associée si elle existe. La normalisation de μ consisterait à la centrer et la réduire pour obtenir μ_n. Soit x_f (resp. y_f) le coefficient de réduction (resp. de centrage) de μ, dans le cas où A favorise B. De façon similaire, soit x_d (resp. y_d) le coefficient de réduction (resp. de centrage) dans le cas où A défavorise B. On a donc :

$$\mu_n(A \rightarrow B) = \begin{cases} x_f.\mu(A \rightarrow B) + y_f & si \quad A \quad favorise \quad B \\ x_d.\mu(A \rightarrow B) + y_d & si \quad A \quad \text{défavorise} \quad B \end{cases} \tag{1}$$

Mesure : $\mu(A \rightarrow B)$	Expression probabiliste	Référence
Multiplicateur de cote	$\frac{p(A' \cap B').p(\overline{B'})}{p(A' \cap \overline{B'})p(B')}$	(Lallich et Teytaud, 2004)
Sebag	$\frac{p(B'/A')}{p(\overline{B}/A')}$	(Sebag et Shoenauer, 1988)
Conviction	$\frac{p(A').p(\overline{B'})}{p(A' \cap \overline{B'})}$	(Brin et al., 1997)
Odd Ratio	$\frac{p(A' \cap B').p(\overline{A'} \cap \overline{B'})}{p(\overline{A'} \cap B').p(A' \cap \overline{B'})}$	cf. (Huynh et al., 2005)
Klosgen	$\sqrt{p(A' \cap B')}(p(B'/A') - p(B'))$	cf. (Huynh et al., 2005)

TAB. 3 – *Liste de MPQ non normalisables*

Du fait de la continuité de μ, ces quatre coefficients se déterminent par passage aux limites aux situations de référence que sont l'incompatibilité, l'indépendance et l'implication logique). Posons $\mu_{imp}(A \rightarrow B)$ la limite de $\mu(A \rightarrow B)$ à l'implication, $\mu_{ind}(A \rightarrow B)$ la limite à l'indépendance et $\mu_{inc}(A \rightarrow B)$ la limite à l'incompatibilité. Ainsi, l'équation (1) est équivalente à l'équation matricielle (2) ci-dessous.

$$\begin{pmatrix} \mu_{imp}(A \rightarrow B) & 1 & 0 & 0 \\ \mu_{ind}(A \rightarrow B) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(A \rightarrow B) & 1 \\ 0 & 0 & \mu_{inc}(A \rightarrow B) & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \\ x_d \\ y_d \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \quad (2)$$

D'où la caractérisation suivante de MPQ normalisables, résultant de l'existence de solution de l'équation (2).

Théorème 1 *Une MPQ μ est normalisable si et seulement si, pour toute règle $A \rightarrow B$, les conditions suivantes sont vérifées :*

(i) les limites $\mu_{imp}(A \rightarrow B)$, $\mu_{ind}(A \rightarrow B)$ et $\mu_{inc}(A \rightarrow B)$ sont finies ,

(ii) $\mu_{imp}(A \rightarrow B) \neq \mu_{ind}(A \rightarrow B)$ et $\mu_{ind}(A \rightarrow B) \neq \mu_{inc}(A \rightarrow B)$.

Le tableau 3 présente une liste de 5 mesures non normalisables.

4 Normalisées de mesures existantes

Dans cette section, nous donnons les normalisées associées à un certain nombre de MPQ normalisables.

Voici une liste de 19 MPQ de normalisée M_{GK} : Support, Confiance (Agrawal et al., 1993), ϕ-coefficient, Ind-Imp-Lerman (Lerman et al., 1981), Cosinus, Facteur de certitude, Laplace, Kappa (cf. (Huynh et al., 2005)), Rappel, Nouveauté, Satisfaction, Fiabilité négative, Spécificité (Lavrac et al., 1999), Surprise (Aze et Kodratoff, 2002), Piatetsky-Shapiro (Piatetsky-Shapiro, 1991), Confiance-centrée (Lallich et Teytaud, 2004), Pearl (Pearl, 1988), Lift (Brin et al., 1997), Dépendance (cf. (Huynh et al., 2005)).
Le tableau 4 donne une liste de 6 MPQ normalisables dont la normalisée associée n'est pas M_{GK}.

Mesure $A \rightarrow B$	Expression	Référence
Jaccard	$\frac{p(A' \cap B')}{p(A')+p(B')-p(A' \cap B')}$	cf. (Huynh et al., 2005)
Q-Yule	$\frac{p(A' \cap B').p(\overline{A' \cap B'}-p(A' \cap \overline{B}').p(\overline{A}' \cap B')}{p(A' \cap B').p(\overline{A' \cap B'})+p(\overline{A}' \cap B').p(A' \cap \overline{B}')}$	cf. (Huynh et al., 2005)
Y-Yule	$\frac{\sqrt{p(A' \cap B').p(\overline{A' \cap B'})}-\sqrt{p(A' \cap \overline{B}').p(\overline{A}' \cap B')}}{\sqrt{p(A' \cap B').p(\overline{A' \cap B'})}+\sqrt{p(A' \cap \overline{B}').p(\overline{A}' \cap B')}}$	cf. (Huynh et al., 2005)
J-mesure	$p(A' \cap B').log(\frac{p(A' \cap B')}{p(A')p(B')})$ $+p(A' \cap \overline{B}').log(\frac{p(A' \cap \overline{B}')}{p(A').p(\overline{B}')})$	(Goodman et Smyth, 1998)
Zhang	$\frac{p(A' \cap B')-p(A').p(B')}{max\{p(A' \cap B').p(\overline{B}');p(B').p(A' \cap \overline{B}')\}}$	(Zhang, 2000)

TAB. 4 – *Liste de MPQ dont la normalisée n'est pas* M_{GK}.

5 Conclusion

Nous avons étudié une trentaine de MPQ sous l'angle de la notion de normalisation qui permet de refléter les situations de référence telles que l'incompatibilité, la répulsion, l'indépendance, l'attraction et l'implication. Nous avons caractérisé les MPQ normalisables et montré que la plupart des MPQ proposées dans la littérature ont la même normalisée, à savoir la mesure normalisée M_{GK} introduite dans (Guillaume, 2000). Ce résultat nous permet d'identifier trois classes de MPQ : (a) celles qui sont non normalisables, (b) les normalisables de normalisée M_{GK} et (c) les autres.

Remerciements

Daniel Feno remercie l'A.U.F. pour son soutien financier.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proc. of the ACM SIGMOD International Conference on Management of Data*, Volume 22, Washington, pp. 207–216. ACM press.

Aze, J. et Y. Kodratoff (2002). Evaluation de résistance au bruit de quelques mesures d'extraction de règles d'association. *Extraction de Connaissances et Apprentissage. 14*, 143–154.

Blanchard, J., F. Guillet, H. Briand, et R. Gras (2005). Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proc. of Applied stochastic Models and Data Analysis*, ENST Bretagne, France, pp. 334–344.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlation. In *Proc. of the ACM SIGMOD Conference*, pp. 265–276.

Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based System 12*, 309–315.

Goodman, R. et P. Smyth (1998). Information theoretic rule induction. In *Proc. of the ECAI-98*, pp. 357–362.

Guillaume, S. (2000). *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*. Ph. D. thesis, Universté de Nantes, France.

Huynh, X., F. Guillet, et H. Briand (2005). Une plateforme exploratoire pour la qualité des règles d'association : Apport pour l'analyse implicative. In *Proc. of Troisièmes Rencontres Internationales A.S.I.*, pp. 339–349.

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

Lavrac, N., P. Flach, et B. Zupan (1999). Rule evaluation measures : A unifying view. In G. Mineau et B. Ganter (Eds.), *Ninth international workshop on Inductive Logic Programming*, Volume 1634, pp. 174–185.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1*, 219–246.

Lerman, I., R. Gras, et H. Rostam (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires. *Math Sc. Hum 74*, 5–35.

Major, J. A. et J. Mangano (1993). Selecting among rules induced from a heuricane database. In *KDD Workshop papers, Menlo Park California*, pp. 28–41.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.

Piatetsky-Shapiro, G. (1991). Knowledge discovery in real databases. *IA Magazine 11*, 68–70.

Sebag, M. et M. Shoenauer (1988). Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of the European Knowledge Acquisition Workshop Conference*, pp. 28–1–28–20.

Tan, P., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measures for association patterns. In *Proc. of the 8th KDD Conference*, pp. 32–41.

Totohasina, A. (2003). Normalisation de mesures probabilistes de la qualité des règles. In *Proc. SFDS'03, XXXV ième Journées de Statistiques*, Volume 2, pp. 985–988.

Totohasina, A., H. Ralambondrainy, et J. Diatta (2005). Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicative. In *Proc. of TAIMA'05*, pp. 375–380.

Zhang, T. (2000). Association rules. In *PAKDD 2000*, LNAI 1805, pp. 245–256. Springer-Verlag.

Summary

This paper is concerned with probabilistic quality measures. We give a necessary and sufficient condition for a probabilistic quality measure to admit an associated normalized one. Moreover, we consider about thirty probabilistic quality measures proposed in the literature and show that most of them have the same associated normalized one.

Aide à l'interprétation des règles d'association composées.

Martine Cadot* , Pascal Cuxac**, Claire François **

* UHP/LORIA, Département Informatique, BP239, 54506 Vandoeuvre-lès-Nancy cedex
martine.cadot@loria.fr
http://www.loria.fr /~cadot/
**INIST-CNRS, 2 allée du Parc de Brabois, 54154 Vandoeuvre-lès-Nancy cedex
pascal.cuxac@inist.fr ; claire.francois@inist.fr

Résumé. L'extraction des règles d'association (RA) est une méthode qui est apparue pour les données type « tickets de caisse ». La création de nombreux indices de qualité a permis sa généralisation à d'autres types de données (Guillet 2004). Nous nous intéressons ici au problème de l'expert qui se trouve confronté à un nombre important de règles pas toujours faciles à interpréter. Les règles formées seulement de deux propriétés, une en partie gauche et une en partie droite s'interprètent aisément une fois l'indice de qualité choisi. Dans le cas de règles composées, c'est-à-dire comportant plus de deux propriétés, ces indices ne suffisent pas à aider l'expert à interpréter le lien entre ces propriétés. Nous proposons un modèle qui permet d'évaluer le gain d'information apporté par les règles de type AB→C et de sélectionner pour l'expert celles qui ajoutent du sens aux règles simples A→C et B→C. L'application de cette méthodologie dans le cadre d'une analyse d'un corpus de textes par classification montre l'aide apportée à l'expert pour l'interprétation de cette classification. Pour faciliter l'exposé, le gain d'information que nous définissons a été appliqué à des règles formées de 3 propriétés, mais il est défini pour un nombre quelconque de propriétés.

1 Introduction

Les RA ont été créées pour extraire de la connaissance à partir de données. A et B étant deux propriétés binaires, la règle d'association est un lien entre A et B noté "A→B" ou "si A alors B". Sa définition varie selon les trois principaux courants initiés par les auteurs suivants : Gras (1979) définit des règles d'implication statistique pour aider les didacticiens à trouver des relations entre les acquisitions de notions élémentaires chez les élèves d'une classe, Guigues et Duquenne (1986) se sont plutôt intéressés à une représentation ordonnée de concepts avec les implications informatives, Agrawal et Srikant (1996) ont privilégié l'extraction optimisée de règles d'association dans de grandes bases de données.

Par la suite, ces formes ont connu des extensions dans plusieurs directions. La binarité des propriétés n'est plus obligatoire, on peut maintenant faire des RA avec des propriétés numériques (Guillaume 2000, Cadot et al. 2004b). Pour éviter l'explosion du temps d'extraction des règles, due à celle de la capacité de stockage des données, des algorithmes plus performants ont été proposés (Pasquier 2000). La sémantique des règles a été affinée grâce à de

nombreux indices de qualité (Guillet 2004), ce qui aide l'utilisateur à choisir les règles les plus adaptées à ses besoins. La navigation ainsi que l'interrogation par un langage adapté ont été mises au point (Botta et al. 2002) pour faciliter l'exploration de cet ensemble de règles.

Toutefois, à notre connaissance, peu de chercheurs en fouille de données se sont intéressés au problème d'interprétation que pose la prémisse composée d'une règle. Les contributions dans ce sens se sont concentrées sur la résolution du problème par élagage des règles qui apportent de l'incohérence au jeu de règles (Cadot et al. 2004a, Zhu H. 1998). Dans cet article, nous nous concentrons essentiellement sur l'ensemble des trois règles $A \rightarrow C$, $B \rightarrow C$ et $AB \rightarrow C$, et nous convenons de garder les deux règles à prémisse simple si elles ont été sélectionnées grâce à un choix d'indices approprié, et de ne garder la règle à prémisse composée que si elle renforce les deux autres, ce renforcement étant mesuré par un indice de gain.

Le but de cet indice de gain est d'aider l'expert à analyser les résultats issus du processus de traitement des données.

2 Le gain d'information

On ne s'intéresse dans cet article qu'aux RA extraites de tableaux numériques de type « SujetsXPropriétés » comportant les valeurs des sujets aux propriétés. Les définitions sont données pour des propriétés binaires et sont ensuite étendues à des propriétés numériques. La qualité d'une règle $A \rightarrow B$ est mesurée par de nombreux indices dont les plus courants sont le ***support***, qui est le nombre d'objets vérifiant les propriétés de A et de B, c'est-à-dire de leur conjonction, appelée ***motif*** AB, et la ***confiance***, qui est le quotient de ce support et du nombre d'objets vérifiant les propriétés de A, c'est-à-dire du support de A.

L'indice de qualité que nous proposons est construit seulement sur les règles à prémisse composée. Il mesure l'apport d'une telle règle par rapport aux règles simples qui la composent. Les règles simples ne permettent pas de calculer le support de la règle composée, mais seulement l'intervalle de ses variations. Le gain d'information mesure l'importance de l'écart entre le support observé et le milieu de l'intervalle de variation qui représente une situation d'équilibre. Notre méthode est à rapprocher de l'IPEE de Blanchard et al (2005) qui mesure l'écart à l'équilibre entre la partie gauche et la partie droite de la règle. Mais notre équilibre concerne toutes les propriétés constituant la règle. Freitas (1999) a également défini un gain d'information qu'il utilise dans son indice « Attribute surprinsingness » portant sur des règles de discrimination (où le membre droit de la règle est une propriété déterminée à l'avance). Son objectif est de mesurer l'apport individuel de chaque propriété à la discrimination, alors que nous mesurons le gain d'information apporté par une propriété supplémentaire.

2.1 Définition du gain

Le principe. Pour mesurer le gain d'information d'une règle, nous nous appuyons sur les variations possibles du support du motif M obtenu en réunissant les propriétés des parties gauche et droite de cette règle. On impose à ces variations de se faire en laissant les supports des sous-motifs de M inchangés. Ainsi ce gain mesure ce que l'association de toutes les propriétés le composant apporte de plus que l'ensemble des diverses associations d'une partie de ces propriétés.

Recherche de l'intervalle de variation. Pour obtenir que le support de M augmente d'une unité, on choisit un sujet qui vérifie toutes les propriétés sauf une, et on lui rajoute cette pro-

priété. Cela a pour conséquence que le support de chaque sous-motif de M contenant cette propriété est également augmenté de une unité. On compense cette augmentation en faisant de nouveaux changements élémentaires, qui vont également devoir être compensés. Si ce processus peut se réaliser, il s'arrête nécessairement au bout de 2^(L-1) changements, L étant la longueur du motif M, c'est-à-dire son nombre de propriétés. Et le support de M peut augmenter d'autant d'unités que de réitérations possibles de ce processus. Pour le faire diminuer, on procède pareillement en faisant les changements inverses de sujets. On crée ainsi l'intervalle de variation du support de M.

Choix de la valeur du gain. L'intervalle obtenu a un centre à partir duquel le support des motifs peut augmenter ou diminuer. Nous décidons que le gain d'information correspondant aux motifs de support central est nul, et cela reste valable si l'intervalle est réduit à une valeur. Puis plus le support du motif s'éloigne de ce centre en se rapprochant des bornes de l'intervalle, plus la valeur absolue du gain augmente. Selon que le support est à droite du centre ou à gauche, le gain est positif ou négatif. Nous mesurons la valeur de ce gain en nombre de sujets dont on doit changer la valeur d'une propriété pour obtenir ce motif en partant d'un motif de support central.

La formule. Pour mesurer le gain d'information g d'une règle, nous calculons le support **s** du motif M sur lequel elle s'appuie, la longueur **L** de ce motif (le nombre de propriétés le constituant), et le centre **c** de l'intervalle décrit par le support de M. Pour valeur de gain, nous choisissons la fonction g=2^(L-1)*(s-c).

2.2 Propriétés du gain d'un motif

La condition préalable au calcul du gain de motifs est qu'ils soient construits avec des propriétés dont on connaît les valeurs pour chacun des N sujets d'un ensemble donné. Bien qu'on utilise par la suite le gain pour des règles à prémisse composée, donc construites sur des motifs de longueur au moins égale à 3, le principe de calcul du gain s'étend sans problème à des motifs de longueur 2 et la formule à des motifs de longueur 1 (notons toutefois que dans ce dernier cas g=s-N/2, ce qui fait qu'on peut obtenir un gain avec des moitiés de sujets).

- prop 1 : Le gain d'un motif M ne peut pas dépasser N/2 en valeur absolue.
- prop 2 : le gain d'un motif de longueur L est un nombre entier de fois 2^(L-2)
- prop 3 : Si a est l'amplitude de l'intervalle de variation du gain g d'un motif M de longueur L, l'intervalle de variation du gain de ses sur-motifs de longueur L+1 a une amplitude inférieure ou égale à a-2|g|. La valeur de a pour les motifs de longueur 1 est N, l'intervalle étant [-N/2 ;N/2].

Ces trois propriétés permettent de limiter le coût machine de la recherche du gain d'un motif. Avec les propriétés 1 et 2, dès que sa longueur est telle que 2^(L-2) dépasse N /2 (soit L>1+log(N)/log(2)), le gain est nul. Avec la propriété 3, chaque fois que le gain d'un motif de longueur L est différent de 0, cela réduit l'intervalle de variation des motifs qui le contiennent. Ainsi, au fur et à mesure que la longueur du motif augmente par ajout de propriétés, ses possibilités de variation diminuent ou restent constantes, ce qui limite sa valeur possible de gain. Cet effet est accentué par la propriété 2. Cela est en adéquation avec le fait que dans le cas le plus courant, une fois que l'information essentielle est apportée par quelques propriétés, au fur et à mesure qu'on ajoute de nouvelles propriétés, l'information supplémentaire qui en résulte est de plus en plus petite.

Aide à l'interprétation des règles d'association composées.

2.3 Relation du gain avec les indices de qualité des règles

Le gain de la règle est celui du motif sur lequel elle est construite. Il fait ainsi partie des indices de qualité d'une règle au même titre que le support, la confiance et tous ceux qu'on définit habituellement (Guillet 2004). Toutefois, il ne mesure pas comme les autres indices la qualité intrinsèque d'une règle, mais la valeur additionnelle d'une règle avec prémisse composée par rapport à celles avec prémisses plus simples. Dans l'application que nous en faisons, nous extrayons d'abord les motifs de longueur quelconque ayant un support suffisant, puis les règles ayant une confiance suffisante construites sur des motifs de longueur 2. L'utilisation d'un seuil de support pour l'extraction des motifs se justifie par le besoin de généraliser les règles obtenues. Le choix de la confiance pour les règles sur des motifs de longueur 2 a été justifié a posteriori par l'interprétation satisfaisante des règles obtenues. Le gain permet de sélectionner les règles de longueur 3 qui renforcent les précédentes.

2.4 Le gain des règles d'association floues

Nous avons défini précédemment des supports flous et des RA floues sur des propriétés numériques (Cadot et al. 2004b). Les supports des motifs flous restent positifs, mais leurs valeurs peuvent ne pas être entières. L'intervalle de variation du support d'un motif ne peut plus se construire en déplaçant des sujets entre les 2^L parties délimitées par les valeurs aux propriétés, car leur appartenance à une partie est floue. Pas plus qu'il ne peut se construire en remplaçant la valeur à une propriété par son complément à 1, celui-ci n'ayant de sens qu'en calcul binaire. Toutefois, comme les effectifs de chacune des 2^L parties étaient calculés précédemment à partir des supports des motifs, le calcul peut toujours se faire pareillement, la différence est que les valeurs obtenues ne sont plus entières, ce sont des effectifs flous. Cela ne gêne aucunement le calcul du gain dont il importe peu qu'il soit ou non entier. Nous avions trois propriétés permettant de limiter le coût machine du calcul du gain. La deuxième propriété due au caractère entier des effectifs disparaît, mais les deux autres restent, et il en résulte une perte d'efficacité de l'algorithme du calcul du gain sur un ensemble de motifs flous. Cette perte est compensée par le choix d'un codage flou plutôt que binaire qui permet de ne pas multiplier les propriétés habituellement binarisées avec plusieurs seuils.

Nous avons ainsi défini un gain pour les règles d'association floue qui prolonge celui que nous venons de définir pour les RA classiques.

3 Application

Notre objectif est d'appliquer des RA floues sur des résultats de classifications, afin d'aider l'expert à analyser les résultats. Les règles à prémisses composées permettent de visualiser les classes qui peuvent fusionner entre deux classifications différentes ou au sein de la même classification.

3.1 Méthode de classification

Les classifications sur lesquelles nous travaillons ont été obtenues en utilisant la plateforme Stanalyst® (Polanco et al. 2001) qui permet de traiter des corpus bibliographiques et inclut la méthode des K-means Axiales comme méthode de classification (Lelu 1993).

Cette méthode est basée sur le principe de classification par centres mobiles, plus connue sous le nom de K-means (Forgy 1965), mais elle réalise une analyse factorielle sphérique sur chaque classe ; les classes sont donc matérialisées par des demi-axes représentatifs des éléments. L'utilisation de ces axes permet de quantifier l'appartenance d'un élément à une classe. De plus, au lieu d'affecter l'élément à la seule classe où sa valeur est la plus grande, on l'affecte également aux classes pour lesquelles cette valeur dépasse un certain seuil. Cet algorithme, paramétré par le nombre maximal de classes désiré et le seuil des coordonnées des éléments et descripteurs sur les axes, permet donc de construire des classes recouvrantes où les individus et descripteurs (documents et mots-clés) sont ordonnés selon un degré de ressemblance au type idéal de la classe.

Le corpus traité est constitué de 3203 notices bibliographiques extraites de la base PASCAL sur le thème de la géotechnique, publiées en 2001 et 2002 et indexées manuellement. Nous avons calculé quatre classifications avec la méthode des K-means axiales en paramétrant 20, 30, 40, 50 classes. Dans la suite de l'article elles sont nommées respectivement C20, C30, C40, C50.

3.2 Règles obtenues

Si nous calculons toutes les RA à prémisse composée nous avons 1548 règles. Le gain calculé permet de filtrer ces résultats : la variation du gain de 5 à 30 permet de passer de 90% (1395) de règles à 7% (105 règles). Pour faciliter l'analyse des résultats nous ne considèrerons dans ce qui suit que les règles où les deux membres de la prémisse appartiennent à la même classification. Avec un gain supérieur à 30 on a les 12 règles résumées dans le tableau suivant (S :support, C :confiance, G : Gain) :

N°	Règle	S	C	G
R1	C20 Barrage, C20 Eau souterraine → C40 Pollution	216,3	0,76	30,56
R2	C20 Inélasticité, C20 Relation σ ε (σ : contrainte, ε : déformation) → C50 Mécanique rupture	25,19	0,52	42,16
R3	C20 Inélasticité, C20 Relation σ ε → C30 Résistance compression	20,7	0,43	31,66
R4	C30 Barrage, C30 Eau souterraine → C40 Pollution	16,13	0,73	30,2
R5	C30 Relation σ ε, C30 Essai sol → C20 Résistance cisaillement	21,37	0,88	37,08
R6	C30 Résistance compression, C30 Mécanique rupture → C20 Inélasticité	18,42	0,91	33,12
R7	C40 Essai sol, C40 Relation σ ε → C20 Résistance cisaillement	19,83	0,86	33,06
R8	C50 Mécanique rupture, C50 Relation σ ε → C20 Inélasticité	19,63	0,91	35,32
R9	C50 Conductivité hydraulique, C50 Pollution → C20 Eau souterraine	23,73	0,97	46,22
R10	C50 Conductivité hydraulique, C50 Pollution → C30 Eau souterraine	23,66	0,97	45,94
R11	C50 Pression pores, C50 Champ pétrole → C20 Inélasticité	16,83	0,91	30,04
R12	C50 Relation σ ε, C50 Essai sol → C20 Résistance cisaillement	18,17	0,91	32,54

TAB. 1 – *Règles avec prémisse composée de deux classes de la même classification et avec g>30 (S=support, C=confiance, G=gain).*

Analysons par exemple la règle R11, constituée des règles simples suivantes :

C50 Pression Pores → C20 Inélasticité
C50 Champ pétrole → C20 Inélasticité

A première vue l'intitulé "Champ pétrole" peut paraître surprenant. L'analyse des données qui sont regroupées dans ces classes (titre des articles, résumés, indexation) permet de comprendre cette règle et de la valider. En effet la classe "Champ pétrole" est essentiellement consacrée aux roches magasins et aux distributions des contraintes dans ces roches. La classe "Inélasticité" est dominée par des aspects liés à l'élastoplasticité et à l'analyse des champs de contraintes. Cette règle est alors plus lisible puisqu'elle lie des articles parlant de la pression de pores (donc de roches poreuses plus ou moins saturées) et des articles sur la distribution des contraintes dans des roches magasin (roches poreuses plus ou moins saturées) avec des articles sur les champs de contraintes dans le domaine élastoplastique.

Cependant, certaines de ces règles, comme par exemple R1, sont difficilement interprétables. Cela est peut-être dû au fait que le gain est gonflé artificiellement par des effets d'effectifs. Ce phénomène est bien connu en statistique et nécessite qu'on sélectionne non seulement les effets les plus importants mais aussi les plus significatifs. En effet un score élevé peut être dû au hasard. La construction d'un test permettant d'établir la significativité du gain est en cours afin de les éliminer.

L'application de cette méthodologie nous a donc permis de "filtrer" les règles à prémisse composée pour ne garder que celles porteuses de valeur ajoutée.

4 Conclusion et perspectives

Le gain que nous proposons combine les avantages des indices de qualité des RA, et de l'élagage du jeu de RA. Il garde les règles simples, construites sur deux propriétés qui ont été extraites à l'aide d'un indice de qualité choisi pour sa valeur sémantique, et sont donc aisément interprétables. Les autres règles, qui ne sont gardées que si leur gain est suffisant, voire significatif renforcent l'information tirée des premières. Au final, l'ensemble des règles obtenu est de taille réduite et sans incohérence. Le gain proposé s'étend sans problème aux RA quantitatives codées de façon floue.

Nous avons vu que l'efficacité du gain doit être renforcée par un test qui en assure la significativité. D'autre part, le gain que nous avons utilisé mesure la valeur ajoutée de la règle AB→C par rapport aux règles A→C et B→C en fixant tous les sous-motifs de ABC. Il faudrait peut-être autoriser la variation du motif AB.

Références

Agrawal, R. Srikant, H, (1994) *Fast algorithms for mining association rules in large databases*, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

Blanchard, J., Guillet, F., Briand, H., Gras, R.(2005) IPEE : Indice Probabiliste d'Ecart à l'Equilibre. pour l'évaluation de la qualité des règles. DKQ 2005 Paris, Atelier EGC 2005 pp. 26-34

Botta, M., Boulicaut J.-F., Masson C., Meo R.(2002). A Comparison between Query Languages for the Extraction of Association Rules. *DaWaK* 2002, p. 1-10

Cadot, M., di Martino, J., Napoli, A. (2004a). Réduction d'un jeu de RA par des méta-règles issues de la logique de "sens commun". *EGC'2004*. (Clermont-Ferrand, France). RNTI, 2004. p.353.

Cadot, M., Napoli, A. (2004b) RA et codage flou des données. *SFC'04.* (Bordeaux). p.130-133.

Forgy E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications, *Biometrics*, vol. 21, n° 3, p. 768.

Freitas, A.A. (1999). On rule interestingness measures. Knowledge-Based Systems 12, p. 309-315.

Guigues J.L., Duquenne V. (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum*. n°95, pp. 5-18

Guillaume S. (2000) *Traitement des données volumineuses, mesures et algorithmes d'extraction de RA et règles ordinales*, Thèse Nantes, 2000.

Guillet F. (2004) Mesure de qualité des connaissances en ECD, *Cours donné lors des journées de la conférence EGC 2004,* Clermont-ferrand, 20 janvier 2004.

Gras R., (1979) *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques* , Thèse, Rennes I, 1979.

Lelu A. (1993). *Modèles neuronaux pour l'analyse de données documentaires et textuelles.* Paris, Thèse de l'Université de Paris VI, 238 pages.

Morineau, A., Nakache, J.-P., Krzyzanowski, C. (1996) *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris 1996.

Pasquier N. (2000), *Data Mining : Algorithmes d'Extraction et de Réduction des RA dans les Bases de Données*, Thèse, Clermont-Ferrand II, 2000

Polanco X., François C., Royauté J., Besagni D (2001). STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, *CSCI 2001*, Sydney, Australia, Proceedings Vol 2, pp. 871 – 873.

Zhu H., (1998) *On-Line Analytical Mining of Association Rules*, Thèse, Simon Fraser University, 1998

Summary

When a simple rule of type A→C or B→C becomes a rule of type AB→C, the two premises A and B are a new premise AB. The fusion of these premises in only one is not a problem in binary logic, which is the foundation of the association rules, but the interpretation of the rule obtained is a semantic problem. We propose a model to evaluate the profit of information brought by rule AB→C and to select for the expert those which reinforce sufficiently the semantics of simple rules A→C and B→C. The application of this methodology to the analysis of a corpus by classification enables us to help the expert to interpret this classification.

Agrégation de mesures d'intérêt de règles d'association

Jean-Pierre Barthélemy*, Angélique Legrain*
Philippe Lenca*, Benoît Vaillant**,*

*GET – ENST Bretagne – Laboratoire TAMCIC, UMR 2872 CNRS
prenom.nom@enst-bretagne.fr
**IUT de Vannes, Département STID
benoit.vaillant@univ-ubs.fr

Résumé. L'un des principaux problèmes posés par l'extraction de règles d'association est l'évaluation de la qualité des règles produites par les algorithmes de type APRIORI. De nombreuses mesures ont été définies afin de pouvoir classer les règles dites *intéressantes*. Très hétérogènes, elles produisent des classements forts variés. C'est pourquoi, plutôt que de privilégier une mesure il paraît intéressant de tenir compte des différentes informations apportées par les mesures. Ainsi, nous avons adopté une nouvelle approche : l'agrégation à l'aide de relations valuées permettant de mesurer le degré d'intensité de préférence d'une règle sur une autre. Elles permettent d'une part de retranscrire la nature numérique des mesures, et d'autre part de réduire les problèmes d'incomparabilité entre les mesures.

Nous avons étudié différents opérateurs d'agrégation. Dans cet article, nous illustrons les résultats obtenus à l'aide d'un exemple jouet en utilisant le plus simple opérateur d'agrégation.

1 Considérations basiques

1.1 Règles d'association

Dans ce papier, nous nous restreignons au cas où les données sont des objets décrits par q attributs binaires : l'objet i satisfait la propriété x (codée en 1), ou non (codée en 0). On note $N = \{1, ..., n\}$ l'ensemble des objets et $Q = \{a, b, ...\}$ l'ensemble des propriétés.

Nous considérons des règles d'association $\mathtt{A} \rightarrow \mathtt{B}$ telles que définies par Agrawal et al. (1993) : si un sous-ensemble de N possède significativement les caractéristiques de l'ensemble $A \subseteq Q$, alors il possédera significativement les caractéristiques de l'ensemble $B \subseteq Q$. Une règle d'association est un 2-uplet (A,B) de sous-ensembles de Q tels que $A \cap B = \emptyset$.

Les algorithmes de type APRIORI (Agrawal et al., 1993) permettant de calculer les règles d'association produisent en général un nombre considérable de règles. Il est ainsi nécessaire de filtrer les règles en fonction de leur intérêt. Nous considérons dans cette étude les mesures d'intérêt dites objectives et basées sur des opérations de comptage dans les données, *i.e.* sur les grandeurs n, n_A, n_B et n_{AB}, où n_A (resp. n_B, n_{AB}) représente le nombre d'objets vérifiant toutes les propriétés de $\mathtt{A}$ (resp. $\mathtt{B}$, $\mathtt{A} \cup \mathtt{B}$), selon les notations classiques.

1.2 Mesures de qualité

Le support, $p_{AB} = n_{AB}/n$, et la confiance, $p_{AB}/p_A = n_{AB}/n_A$ sont utilisés comme premier filtre pour extraire un ensemble $\mathcal{R}$ de règles avec les algorithmes de type APRIORI. Il est nécessaire d'utiliser dans un second temps d'autres mesures de qualité. En effet, on considère que le support et la confiance n'ont que peu de bonnes propriétés pour ordonner un ensemble de règles si on les compare à d'autres mesures (Piatetsky-Shapiro, 1991; Tan et al., 2002; Lenca et al., 2003). Nous avons sélectionné et étudié une vingtaine de mesures (Lenca et al., 2003). Dans Lenca et al. (2004), nous proposons une aide à la sélection d'un ensemble de mesures selon les données et les attentes de l'utilisateur. On met en évidence que le choix d'une mesure dépend particulièrement des attentes de ce dernier.

Les mesures constituent un paysage très hétérogène : on peut observer d'importantes variations entre les formules (les mesures ne traduisent pas les mêmes caractéristiques des règles), et de grandes différences dans les co-domaines ([0, 1], [0, +∞[,]-∞, 1], bornes fonction de n_A, n_B, et/ou $n_{AB}\ldots$). Ainsi certaines règles peuvent être très bien classées par une mesure, et mal classées par une autre. La comparaison des préordres totaux induits par les mesures sur une base de règles permet de mettre en évidence cette observation (Vaillant et al., 2004).

1.3 L'agrégation pour faire face à l'hétérogénéité des mesures

Il se pose ainsi naturellement la question suivante : quelle est ou quelles sont les meilleures règles étant donné nos mesures de qualité et les données à traiter ? On peut tenter d'y répondre selon deux voies principales :

(i) position dictatoriale : choisir une mesure préférée et ne pas tenir compte des autres
(ii) position consensuelle : trouver un consensus entre les mesures.

Dans cet article, nous suivons cette seconde piste pour laquelle trois voies apparaissent naturellement :

a) l'agrégation directe des mesures en une seule, en utilisant une sorte de moyenne généralisée. Une difficulté est alors posée par la diversité des échelles de mesures. Comment agréger des mesures dont les co-domaines sont [0,1], [0,+∞[,]-∞,1] ?
b) l'agrégation des rangs induits par les différentes mesures en un rang unique. Les classements ne tiennent alors pas compte des différences dans les évaluations. De plus, une agrégation ordinale implique des problèmes "logiques" périlleux (se référer par exemple au théorème d'Arrow (Arrow, 1951))
c) l'agrégation de relations valuées. C'est la voie que nous explorons dans cette étude. Les classements sont des préordres totaux que nous généralisons sous forme de relations valuées.

Pour cela, nous rappelons qu'une relation valuée (parfois appelée " relation floue ") sur un ensemble S est une transformation R de $S \times S$ dans l'intervalle unité [0,1]. Les relations valuées permettent d'échapper aux effets d'échelles en préservant les différences d'échelles puisqu'elles tiennent compte de telles différences. L'agrégation de relations valuées a été étudiée très précisément par Fodor et Roubens (1994). Elles peuvent être affectées de nombreuses propriétés, parmi lesquelles nous retenons quelques formes particulières de transitivité.

2 Construction d'une relation valuée sur un ensemble de mesures

L'un des avantages de la modélisation de relations valuées est la retranscription des évaluations numériques qu'elle permet. Pour définir des relations valuées il va donc falloir poser certaines conditions. Pour une mesure μ et deux règles r_i et $r_{i'}$, on dira par exemple que $R(r_i, r_{i'})$ est négligeable si $\mu(r_i)$ est faiblement supérieur à $\mu(r_{i'})$. Cette faible différence peut être définie de deux façons : soit par le quotient des évaluations, soit par leur différence.

Nous utilisons trois types de transitivité, la transitivité faible ($\{R(s,t) \geq \frac{1}{2}$ et $R(t,u) \geq \frac{1}{2}\} \Rightarrow R(s,u) \geq \frac{1}{2}$), la min-transitivité ($R(s,u) \geq \min\{R(s,t), R(t,u)\}$) et la max-$\Delta$-transitivité ($R(s,u) \geq \max\{0, R(s,t) + R(t,u) - 1\}$) car elles permettent la préservation de certaines propriétés après agrégation.

Soit $\mathcal{R} = \{r_1, \ldots, r_k\}$ un ensemble de règles, et $\mu_1, \ldots, \mu_m$ les mesures sélectionnées. Chaque mesure μ_j induit une relation valuée R_j sur $\mathcal{R}$. L'idée générale est que $R_j(r_i, r_{i'})$ correspond à une différence normalisée entre les valeurs prises par la mesure μ_j sur les règles $r_i, r_{i'}$ et doit permettre de modéliser un système de préférences sur l'ensemble des règles.

Nous présentons ci-dessous, une des relations valuées que nous avons étudiée et proposée par Brans et Mareschal (1994). C'est une variante de la différence linéaire permettant de lisser les transitions entre "non préférence" et "faible préférence" ainsi qu'entre la "préférence faible" et la "préférence forte" (le paramètre σ_j représente un seuil entre les "préférences faibles" et les "préférences fortes" –point d'inflexion de la courbe) :

$$R_j(r_i, r_{i'}) = \begin{cases} 1 - exp\big(-\frac{(\mu_j(r_i) - \mu_j(r_{i'}))^2}{2\sigma_j^2}\big) & \text{si } \mu_j(r_i) - \mu_j(r_{i'}) > 0 \\ 0 & \text{sinon} \end{cases} \quad (1)$$

3 Agrégation

3.1 Généralités

Un opérateur d'agrégation est une fonction C de $\cup_{m \geq 1}[0,1]^m$ dans [0,1] non décroissante de chaque composant, idempotent, et qui satisfait $C(0, \ldots, 0) = 0$ et $C(1, \ldots, 1) = 1$. Si R^* est un m-uplet $(R_1, \ldots, R_m)$ de relations valuées, un opérateur d'agrégation C va produire une relation de consensus de R^*, notée $C(R^*) : C(R^*)(r_i, r_{i'}) = C(R_1(r_i, r_{i'}), \ldots, R_m(r_i, r_{i'}))$.

De nombreux opérateurs d'agrégation aux propriétés différentes ont été étudiés dans la littérature (cf. Fodor et Roubens (1994)) : moyennes généralisées, opérateurs OWA, intégrales de Choquet et de Sugeno, maximum et minimums pondérés, etc. Nous avons choisi de nous concentrer sur les plus simples d'entre-eux, les moyennes généralisées :

$$M(u_1, \ldots, u_m) = f^{-1}\Big(\sum_{1 \leq j \leq m} w_j f(u_j)\Big)$$

où f est une fonction monotone continue, f^{-1} sa réciproque, et les w_j des poids non négatifs. Ainsi on distingue certains cas particuliers : la moyenne arithmétique pondérée (WMean, $f(u) = u$) la moyenne géométrique pondérée (WGeom, $f(u) = \log(u)$) la moyenne d'ordre α (RPM, $f(u) = u^\alpha, \alpha \in \mathbb{R}^*$), et la moyenne harmonique pondérée (WHarm, $f(u) = 1/u$). Dans cet article, nous présentons des résultats obtenus avec RPM et WMean.

3.2 Comportement vis à vis de la transitivité et de certaines propriétés

Saminger et al. (2002); Peneva et Popchev (2003) ont étudié les propriétés préservées par ces opérateurs. Seules celles qui jouent un rôle dans la modélisation des préférences nous intéressent dans le cadre de ce travail. Nous avons par conséquent éliminé différentes formes de dissimilarités. Nous nous concentrons principalement sur la transitivité, qui est l'une des propriétés qui se rapproche le plus des attentes de l'utilisateur. Elle garantit que si un ensemble de règles est mieux évalué qu'un autre, alors cet ordre sera préservé par la procédure d'agrégation.

4 Résultats expérimentaux

Le tableau 1 contient le rangement de 21 règles pour les 20 mesures que nous avons étudiées (les notations utilisées sont des abréviations du nom des mesures, SUP pour le support, IIE pour l'intensité d'implication entropique, etc., voir Lenca et al. (2004)). Les données "jouet" et le mode de calcul de ces règles sont présentés dans (Barthélemy et al., 2006), ainsi que les évaluations numériques des règles pour chaque mesure.

La relation valuée retenue pour exprimer la préférence individuelle d'une règle sur une autre pour une mesure donnée nécessite de fixer le paramètre σ_j (cf. formule 1). Pour ce faire, nous avons choisi de tenir compte des différences d'évaluations, et utilisons la valeur pour un quantile donné. Par exemple, la valeur prise à un quantile de 0% correspond à la plus petite valeur absolue de la différence (qui est évidemment 0, puisque la différence des évaluations d'une règle sur elle-même est nulle). La valeur prise par un quantile de 100% est la plus grande valeur absolue de la différence d'évaluation entre toute paire de règles et un quantile de 50% amène à la valeur médiane de l'absolue des différences. Dans notre exemple, le quantile est fixé à 60% de toutes les différences absolues, et c'est donc une valeur élevée. Nous expliquons plus loin pourquoi nous avons choisi une telle valeur. Les poids choisis sont tous égaux à $1/20$. D'autres poids ont été proposés dans (Legrain, 2004) grâce à une modélisation de préférences d'utilisateurs (Lenca et al., 2004).

Une fois ces paramètres fixés, la procédure d'agrégation produit une matrice carrée de valeurs entre 0 et 1, chaque valeur représentant la préférence agrégée d'une règle sur une autre. Pour obtenir des résultats lisibles (i.e. binaires), on utilise un seuil de coupe λ et on compare l'indice d'agrégation à λ. Les valeurs plus faibles que λ sont considérées comme égales à 0 et les valeurs supérieures égales à 1. On représente alors les préférences par des graphes (cf. figures 1 et 2, pour RPM et WMean) où les arcs entre deux sommets représentent la préférence d'une règle sur une autre, la flèche pointant sur la règle qui est préférée. On considère qu'aucune règle n'est préférée à elle-même. Il apparaît que les règles qui sont bien classées par toutes les mesures (comme r_{18}) restent bien classées, et celles qui étaient mal classées (comme r_{19}) restent mal classées. Le cas des règles plus controversées, comme r_2 nécessite d'approfondir la méthode. Notons que WHarm et WGeom produisent un effet de véto (Barthélemy et al., 2006). En effet, quand une mesure a la même valeur pour deux règles, l'utilisation du logarithme et de la fonction inverse produit des valeurs "infinies", ce qui fait que toute autre différence n'est pas prise en compte. Ceci explique qu'un grand nombre de règles resteront incomparables en terme d'agrégation de préférences avec de tels agrégateurs. Le fait que SUP divise les règles en seulement deux groupes joue aussi un rôle sur la valeur que nous avons choisie pour fixer σ_j : si nous avions choisi une valeur plus faible que 60%, alors la valeur

TAB. 1 – *Ordonnancement induit par les mesures sur une base de règles*

measure	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6	rank 7	rank 8	rank 9	rank 10	rank 11
LIFT	r_{18}	$r_1\ r_2\ r_{17}$	r_{21}	$r_6\ r_7$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
CONFCEN	$r_1\ r_{17}$	r_{18}	r_2	r_{21}	$r_3\ r_6\ r_{16}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
CONF	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
SUP	$r_5\ r_6\ r_7\ r_8\ r_{11}\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_1\ r_2\ r_3\ r_4\ r_9\ r_{10}\ r_{16}\ r_{17}\ r_{18}\ r_{19}\ r_{20}\ r_{21}$									
IIE	$r_1\ r_{17}$	$r_3\ r_{16}$	r_{18}	$r_6\ r_{21}$	r_{20}	r_2	r_7	$r_4\ r_5\ r_8\ r_9\ r_{10}\ r_{11}\ r_{12}\ r_{13}\ r_{14}\ r_{15}\ r_{19}$			
INTIMP	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
IQC	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	r_{20}	$r_3\ r_{16}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
CONV	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	$r_6\ r_{21}$	r_2	r_{20}	r_7	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
GI	r_{18}	$r_1\ r_2\ r_{17}$	r_{21}	$r_6\ r_7$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
-INDIMP	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
IPD	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
LAP	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
LOE	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	$r_6\ r_{21}$	r_2	r_{20}	r_7	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
MC	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
PS	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$			
R	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	$r_3\ r_{16}$	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
SEB	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
MOCO	r_{18}	$r_1\ r_{17}$	r_6	$r_7\ r_{21}$	$r_3\ r_8\ r_{11}\ r_{16}$	$r_2\ r_{12}\ r_{13}\ r_{20}$	$r_5\ r_{14}\ r_{15}\ r_{19}$	$r_4\ r_9\ r_{10}$			
TEC	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
ZHANG	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	

du quantile aurait été nulle. Un moyen de palier à une telle situation pourrait être d'ajouter un léger degré de bruit dans les valeurs prises par les mesures. Cette stratégie, connue comme le "jittering", est souvent utilisée dans des outils de visualisation, afin de représenter un nombre important de points, autrement confondus. La représentation graphique a pour but d'illustrer les résultats sur un exemple jouet. Les résultats obtenus sur une base de données réelles sont présentés dans (Legrain, 2004), mais les représentations graphiques nécessitent alors des méthodes plus élaborées. Cette possibilité a notamment été explorée en fouille visuelle de règles d'associations par Lehn et al. (1999) et Blanchard et al. (2003). Notons cependant que l'exploration visuelle n'est pas requise pour sélectionner les règles, c'est un moyen confortable et centré sur l'expert.

5 Conclusion et perspectives

Dans cette étude, nous nous sommes intéressés à la construction d'opérateurs d'agrégation de relations valuées définies à partir des évaluations des mesures d'intérêt de règles d'association. Les attentes de l'utilisateur contraignent les choix possibles. De telles contraintes peuvent

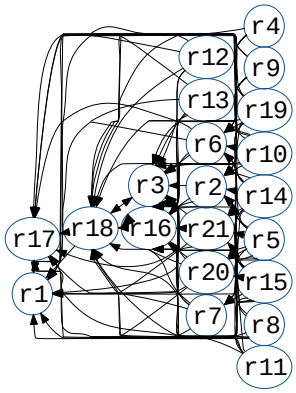

FIG. 1 – *RPM* ($\alpha = 2$)

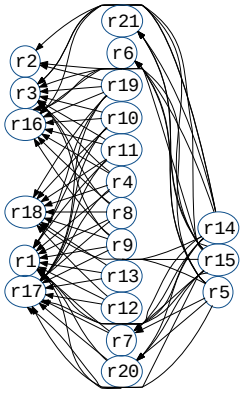

FIG. 2 – *WMean*

être explicitées mathématiquement. Nous nous sommes particulièrement intéressés à la propriété de transitivité. Parmi les opérateurs d'agrégation classiques, peu d'entre eux respectent de telles contraintes, excepté la moyenne généralisée. Un exemple jouet illustre notre approche. Des expériences prometteuses effectuées sur de plus grandes bases de données laissent entrevoir des résultats visuels intéressants. Bien qu'il persiste certains conflits entre le traitement de grandes bases de règles et leur représentation intelligible sous forme de graphe, nous pensons qu'une ouverture possible serait d'autoriser des zooms sur certaines régions significatives.

Une perspective d'étude intéressante est d'élargir le choix des opérateurs d'agrégation à des opérateurs de type compromis, comme l'intégrale de Choquet ou de Sugeno. Le classement final d'une règle pourrait aussi être consolidé si une coalition importante de mesures lui était favorable. Le choix du paramètre σ_j est déterminant. Nous l'avons fixé à l'aide d'une approche expérimentale basée sur une méthode de quantile. D'autres possibilités peuvent être envisagées, par exemple nous proposons de le fixer à l'aide d'un expert des données. Dans tous les cas, une étude de la robustesse de la solution proposée doit être menée.

Références

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

Arrow, K. J. (1951). *Social Choice and Individual Values*. Cowles Foundations and Wiley.

Barthélemy, J. P., A. Legrain, P. Lenca, et B. Vaillant (2006). Aggregation of valued relations applied to association rule interestingness measures. In *MDAI'06 (To be published)*, LNAI. Springer-Verlag.

Blanchard, J., F. Guillet, et H. Briand (2003). A user-driven and quality-oriented visualization for mining association rules. In *Third IEEE ICDM*, pp. 493–496.

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

Fodor, J. et M. Roubens (1994). *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Publishers.

Legrain, A. (2004). Agrégation de mesures de qualité de règles d'association, Rapport de DEA MIASH. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne.

Lehn, R., F. Guillet, P. Kuntz, H. Briand, et J. Philippé (1999). Felix : An interactive rule mining interface in a KDD process. In P. Lenca (Ed.), *HCP'99*, pp. 169–174.

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

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2004). A multicriteria decision aid for interestingness measure selection. Technical Report LUSSI-TR-2004-01-EN, ENST Bretagne.

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

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

Peneva, V. et I. Popchev (2003). Properties of the aggregation operators related with fuzzy relations. *Fuzzy Sets and Systems, numéro 139*, 615–633.

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

Saminger, S., R. Mesiar, et U. Bodenhofer (2002). Domination of aggregation operators and preservation of transitivity. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10*(Suppl.), 11–35.

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

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *Discovery Science*, Volume 3245 of *LNAI*, pp. 290–297. Springer-Verlag.

Summary

One of the concerns of knowledge discovery in databases is the evaluation of association rules'interestingness. Many interestingness measures have been introduced in order to rank such rules according to their interest. These measures are heterogeneous and the ranking of rules may differ largely. This is why, rather than privileging one measure it appears interesting to take account the various information brought by the measures. Thus we explore a new approach: the aggregation of valued relations that represent the intensity of preference of a rule over another. The aim in using such tools is to take into account the numerical nature of measures and reduce incomparability issues.

We studied several aggregation operators. In this contribution we discuss results obtained on a toy example using the simplest of them.

Extraction de mesures d'intérêt représentatives pour le post-traitement 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}@univ-nantes.fr

Résumé. Cet article s'intéresse au calcul d'une base réduite de mesures d'intérêt pour le post-traitement des règles d'association. Ces mesures, appelées mesures d'intérêt représentatives, sont calculées à l'aide d'une classification par médioïdes. L'intérêt de cette approche est de pouvoir offrir à un utilisateur un ensemble réduit de mesures les mieux adaptées à la nature de ses données, et ainsi de lui faciliter la validation des meilleures règles. Nous dressons un état de l'art sur le post-traitement et les travaux relatifs aux mesures d'intérêt. Puis notre approche est appliquée sur un jeu d'essai comportant 40 mesures d'intérêt et environ 120000 règles. Sur cette expérimentation, nous avons pu montrer l'intérêt de notre approche et proposer de réduire le nombre de mesures à seize mesures d'intérêt représentatives.

1 Introduction

Depuis les travaux précurseurs (Agrawal et al., 1993) (Agrawal et Srikant, 1994) en extraction de connaissances dans les données, la validation des règles d'association demeure un verrou scientifique pénalisant l'usage de cette technique.

En effet, bien que le modèle des règles d'association ait l'avantage de permettre une extraction non supervisée de règles exprimant des tendances implicatives dans les données, il a malheureusement l'inconvénient de produire de très grandes quantités de règles. L'étape finale de validation de ces règles étant laissée à l'utilisateur (un décideur ou un analyste), celui-ci se trouve alors confronté à une tâche généralement inextricable : comment repérer les règles intéressantes parmi plusieurs milliers, voire plusieurs millions de règles ?

Il s'avère donc indispensable d'aider l'utilisateur par la mise en oeuvre d'une phase de post-traitement des règles produites. Ce post-traitement a pour objectif de réduire le volume des règles, en présélectionnant un nombre réduit de règles potentiellement intéressantes pour l'utilisateur, au sens de ses critères de préférence et en fonction de la structure des données.

A cette fin, cinq approches complémentaires de post-traitement sont proposées dans la littérature : par représentations graphiques, par recherche interactive de résumés, par réduction des redondances, par aide à la décision multicritères, et par mesures d'intérêt. La première consiste à utiliser des représentations graphiques afin d'améliorer la lisibilité des résultats (Blanchard et al., 2003). La seconde consiste à considérer l'ensemble de règles comme une base de données sur laquelle l'utilisateur extrait des sous ensembles de règles par des requêtes intégrant des

contraintes (Blanchard et al., 2003) (Blanchard et al., 2005a) (Blanchard et al., 2005b), (Hilderman et Hamilton, 2001), (Boulicaut et al., 1998). La troisième consiste à opérer une réduction très significative (généralement 10%) en ne conservant que les règles non redondantes au sens logique (Lehn et al., 2004). La quatrième consiste à concevoir la sélection des meilleures règles comme un problème d'optimisation en aide à la décision multicritères (Lenca et al., 2004). La dernière consiste à valuer les règles par des indicateurs numériques : les mesures d'intérêt.

Notre travail s'insère dans le cadre de cette dernière approche par mesures d'intérêt. Depuis l'introduction du support et de la confiance (Agrawal et Srikant, 1994), de nombreuses mesures ont été proposées dans la littérature (Bayardo-Jr. et Agrawal, 1999) (Hilderman et Hamilton, 2001) (Tan et al., 2004). Cette abondance de mesures d'intérêt pose un second problème : comment aider l'utilisateur à choisir les mesures les mieux adaptées à ses buts et à ses données, afin de détecter in fine les meilleures règles ?

Dans cet article, nous proposons une technique d'analyse de données pour calculer les meilleures mesures d'intérêt objectives sur un jeu de règles. Pour cela, nous utilisons une classification par médioïdes (PAM : partitioning around medoids (Kaufman et Rousseeuw, 1990)) avec comme métrique l'indice de corrélation linéaire, afin de partitionner quarante mesures d'intérêt en k clusters. Le sous-ensemble de mesures d'intérêts représentatives du jeu de données étudié est alors constitué par les k médioïdes obtenus.

L'article est organisé en cinq parties. Après une synthèse des travaux relatifs aux mesure d'intérêt des règles d'association, nous présentons notre approche de calcul des mesures représentatives en employant la technique d'analyse de données PAM. Puis, nous décrivons l'ensemble de données et les mesures d'intérêt étudiées. Enfin, nous expérimentons notre méthode sur la base de données mushroom (Newman et al., 1998) et discutons les résultats obtenus.

2 Travaux relatifs sur la qualité des connaissances

Dans les travaux précurseurs sur les règles d'association (Agrawal et al., 1993) (Agrawal et Srikant, 1994), deux premières mesures statistiques[1] sont introduites : le support et la confiance. Celles-ci sont bien adaptées aux contraintes algorithmiques (cf Apriori), mais ne sont pas suffisantes pour capturer l'intérêt des règles pour l'utilisateur. Afin de contourner cette limite, de nombreuses mesures d'intérêt complémentaires ont été proposées dans la littérature. Freitas (1999) distingue deux types de mesures d'intérêts : les mesures subjectives, et les mesures objectives. Les mesures *subjectives* dépendent des buts, connaissances, croyances de l'utilisateur et sont combinées à des algorithmes supervisés spécifiques afin de comparer les règles extraites avec ce que l'utilisateur connaît ou souhaite (Padmanabhan et Tuzhilin, 1998) (Liu et al., 1999). Ainsi, les mesures subjectives proposent de capturer la nouveauté (novelty) ou l'inattendu (unexpectedness) d'une règle par rapport aux connaissances/croyances de l'utilisateur. Les mesures *objectives*, quand à elles, sont des indices statistiques qui évaluent la contingence d'une règle dans les données.

Dans la littérature, de nombreuses synthèses traitent des mesures d'intérêt selon deux aspects différents : la définition de l'ensemble des principes constitutifs d'une bonne mesure

[1] Lallich et al. (2005) ont proposé une autre façon de distinguer une mesure d'intérêt objective par sa "nature" *statistique* ou *descriptive*. Les mesures descriptives ne varient pas avec l'expansion de cardinalité. Bien au contraire, les mesures statistiques varient avec l'expansion de cardinalité. Ici, nous disons simplement les mesures statistiques dans tous ces deux cas.

d'intérêt, et leur comparaison par classification en fonction de critères théoriques ou expérimentaux sur des données.

Dans la perspective d'établir les principes d'une bonne mesure d'intérêt, Piatetsky-Shapiro (1991) a présenté une nouvelle mesure d'intérêt, appelé Rule-Interest, et propose trois principes fondamentaux pour une mesure sur une règle $a \rightarrow b$: (P1) valeur 0 quand a et b sont indépendants, (P2) croissant avec $a \wedge b$, (P3) décroissant avec a ou b. Hilderman et Hamilton (2001) ont proposé cinq principes : minimum value, maximum value, skewness, permutation invariance, transfer. Tan et al. (2004) ont défini cinq principes d'intérêt : symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas (1999) a proposé un principe de "surprise" d'attribut. Gras et al. (2004) ont proposé un ensemble de dix critères constitutifs d'une bonne mesure d'intérêt.

Parmi ces synthèses, certaines abordent également la comparaison des mesures d'intérêt à partir de leur classification. Bayardo-Jr. et Agrawal (1999) ont conclu que les meilleures règles selon toutes les mesures d'intérêt qu'ils étudient doivent résider le long d'une frontière de support/confiance. Kononenco (1995) a utilisé onze mesures pour estimer la qualité des attributs à valeurs multiples, et montre que les valeurs des mesures : information-gain, j-mesure, gini-index, et relevance tendent à augmenter linéairement avec le nombre de valeurs d'un attribut. Zhao et Karypis ont proposé un algorithme d'optimisation de huit critères et montrent qu'une partie des critères suffit pour extraire les meilleures règles. Gavrilov et al. (1999) ont étudié la similitude des mesures afin de les classer. Hilderman et Hamilton (2001) ont proposé cinq principes pour ranger des résumés de bases de données en employant seize mesures de diversité et montrent que : (1) six mesures ont satisfait cinq des principes proposés, (2) neuf des mesures restantes ont satisfait au moins un des principes proposés. En étudiant vingt et une mesures, Tan et al. (2004) ont montré qu'aucune mesure n'est adaptée à tous les cas et que la corrélation des mesures augmente avec la diminution du support. Par une méthode d'aide à la décision multicritère intégrant huit critères, (Vaillant et al., 2004) (Lenca et al., 2004) ont extrait un pré-ordre sur vingt mesures et identifient quatre clusters de mesures. (Carvalho et al., 2005) (Carvalho et al., 2003) ont évalué onze mesures d'intérêt objectives afin de les ranger en fonction de leur intérêt effectif pour un décideur. Choi et al. (2005) ont utilisé une approche d'aide à la décision multi-critère pour trouver les meilleures règles d'association. Lallich et Teytaud (2004) ont utilisé quinze mesures et ont proposé des critères pour les évaluer. (Blanchard et al., 2005a) (Blanchard et al., 2005b) ont classé dix-huit mesures objectives en quatre groupes selon trois critères : indépendance, équilibre, et caractère descriptif ou statistique. Huynh et al. (2005b) ont proposé une approche de classification par graphes de corrélation qui permet d'identifier onze clusters sur trente-quatre mesures d'intérêt.

Enfin, deux outils d'expérimentation sont disponibles : HERBS (Vaillant et al., 2003) et ARQAT (Huynh et al., 2005a).

3 Calculs des mesures représentatives

3.1 Formalisation des données et de la distance corrélative entre des mesures d'intérêt

Soient $R(D) = \{r_1, r_2, ..., r_p\}$ un ensemble de p règles d'association extraitent d'un ensemble de données D. Chaque règle $a \rightarrow b$ est décrite par ses deux itemsets (a, b) et ses cardinalités $(n, n_a, n_b, n_{a\bar{b}})$. Soit $M = \{m_1, m_2, ..., m_q\}$ un ensemble de q mesures d'intérêt. Chaque mesure est calculée par une fonction numérique sur les cardinalités d'une règle : $m(a \rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$. Précisément, n est le nombre total d'enregistrements de D, n_a (resp. n_b) le nombre d'enregistrements de D satisfaisant a (resp. b), et $n_{a\bar{b}}$ le nombre d'enregistrements satisfaisant $a \wedge \bar{b}$ (les exemples négatifs).

Chaque règle étant valuée par q mesures, nous disposons en entrée de notre problème d'une matrice numérique $((m_{ij}))$ de dimensions $q \times p$ dont chaque élément $m_{ij} = m_i(r_j)$ correspond à la valeur de la mesure m_i sur la règle r_j, avec $i = 1..q$ et $j = 1..p$.

Nous calculons ensuite $((d_{ij}))$ la matrice carrée de dimension q des distances entre mesures d'intérêt, à l'aide d'une distance corrélative :

$$d_{ij} = 1 - |\eta_{ij}|$$

où η_{ij} est le coéfficient de corrélation linéaire (Saporta, 1990) entre deux mesures m_i et m_j :

$$\eta_{ij} = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}}$$

3.2 Calcul des mesures représentatives avec PAM

Nous partons de la matrice $((d_{ij}))$ des distances entre mesures d'intérêts, afin de partitionner l'ensemble des mesures en k clusters, puis d'extraire le meilleur représentant de chaque cluster. Pour cela, nous avons choisi d'utiliser la méthode PAM (Partitioning Around Medoids (Kaufman et Rousseeuw, 1990)). Cette méthode de classification non supervisée, variante des k-moyennes (MacQueen, 1967), à l'avantage d'être plus robuste que cette dernière.

De plus, la méthode PAM que nous utilisons (PAM dans R[2], (Kaufman et Rousseeuw, 1990)), à le grand avantage de permettre la visualisation graphique de la classification obtenue selon les deux premiers axes principaux d'une ACP.

La partition en clusters est basée sur la notion de médioïde. Un médioïde m_i^* d'un cluster ψ_i est définit comme l'élément de ce cluster le plus proche de tous les autres[3].

Partant d'un ensemble de k médioïdes initiaux, la méthode procède en plusieurs itérations jusqu'à obtenir la stabilisation des médioïdes. A chaque itération, chaque objet est classés dans le cluster du médioïde le plus proche, puis le médioïde de chaque cluster est recalculé.

Nous rappelons quelques notions définies dans Kaufman et Rousseeuw (1990).

- La qualité de la classification sur les k médioïdes à l'aide d'une fonction, appelée *fonction objective*, définie par : $\sum_{j=1}^{q} min_{i=1...,k} d(j, m_i^*)$, où q est le nombre d'objets à classer.

[2] http ://www.r-project.org/

[3] En calculant la distance moyenne dans notre approche.

- Un cluster ψ est dit isolé si tout objets du cluster est toujours plus proche d'un second objet de son cluster que de tout autre objet des autres clusters : $\forall i, j \in \psi, \forall h \notin \psi, max(d_{ij}) < min(d_{ih})$
- Le diamètre d'un cluster ψ est la plus grande distance entre deux objets du même cluster : $max(d_{ij}), i, j \in \psi$
- La séparation d'un cluster ψ est définie comme la distance minimale entre un objet d'un cluster et un objet d'un autre cluster : $min(d_{ij}), i \in \psi, j \notin \psi$
- La distance moyenne à un médioïde j d'un cluster ψ est donnée par : $\frac{\sum_{i \in \psi}(d_{ij})}{|\psi|}$
- La distance maximale à un médioïde j d'un cluster ψ est donnée par : $max(d_{ij}), i \in \psi$

4 Expérimentation

4.1 Description des données, des règles, et des mesures étudiées

Nous expérimentons notre approche sur la base de données catégoriques mushroom (D) issue du dépôt d'Irvine (Newman et al., 1998). Nous avons ensuite calculé l'ensemble de règles d'association R sur ces données à l'aide de l'algorithme Apriori (Agrawal et Srikant, 1994). Le tableau Tab. 1 récapitule les caractéristiques principales des données étudiées.

Ensemble de données	Article	Transactions	Longueur moyenne des transactions	Nombre de règles (seuil de support)	Jeu de règles
D_1	118	8416	22	123228 (12%)	R_1

TAB. 1 – *Description des données.*

Cet ensemble de 123228 règles est évalué par la matrice $((m_{ij}))$ sur quarante mesures d'intérêt. La plupart des références détaillées sur ces mesures peuvent être trouvées dans les synthèses (Hilderman et Hamilton, 2001) (Tan et al., 2004). Nous utilisons trente-quatre mesures d'intérêt dont les formules peuvent être trouvées dans Huynh et al. (2005b),les 6 mesures restantes étant définies dans le tableau Tab. 2.

4.2 Clusters de mesures d'intérêt obtenues

La classification obtenue est illustrée par trois figures. La première (Fig. 1) présente les résultats sous une forme graphique en projetant les mesures selon les deux axes principaux obtenus par une analyse en composantes principales. Chaque mesure y est représentée par un symbole, et chaque cluster par un numéro. Bien que déformée par la projection, cette visualisation s'avère très utile pour avoir une vue synthétique de la classification, mais aussi pour valider le choix du nombre k de clusters, ici fixé à seize (Ce choix de seize clusters est issu de travaux préliminaires présentés dans Huynh et al. (2005c)).

Le tableau Tab. 3 récapitule les clusters obtenus et leur médioïde. Le tableau Tab. 4 détaille les caractérisques numériques de chaque cluster.

On peut y observer que les clusters 1 (Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Ganascia, Laplace) et 2 (Causal Confirm,

II	$1 - \sum_{k=max(0,n_a-n_b)}^{n_{a\bar{b}}} \frac{C_{n_b}^{n_a-k} C_{n_{\bar{b}}}^{k}}{C_n^{n_a}}$
IPEE	$1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} C_{n_a}^{k}$
F-measure	$\frac{2 \times recall \times precision}{precision + recall}$ avec $recall = 1 - \frac{n_{a\bar{b}}}{n_a}$ et $precision = \frac{n_a - n_{a\bar{b}}}{n_b}$
Ganascia	$1 - 2\frac{n_{a\bar{b}}}{n_a}$
Mutual Information	$\frac{\frac{n_a - n_{a\bar{b}}}{n} log(\frac{n(n_a - n_{a\bar{b}})}{n_a n_b}) + \frac{n_{a\bar{b}}}{n} log(\frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}) + \frac{n_{\bar{a}b}}{n} log(\frac{n n_{\bar{a}b}}{n_{\bar{a}} n_b}) + \frac{n_{\bar{a}\bar{b}}}{n} log(\frac{n n_{\bar{a}\bar{b}}}{n_{\bar{a}} n_{\bar{b}}})}{min(-(\frac{n_a}{n} log(\frac{n_a}{n}) + \frac{n_{\bar{a}}}{n} log(\frac{n_{\bar{a}}}{n})), -(\frac{n_b}{n} log(\frac{n_b}{n}) + \frac{n_{\bar{b}}}{n} log(\frac{n_{\bar{b}}}{n})))}$
Odd Multiplier	$\frac{(n_a - n_{a\bar{b}}) n_{\bar{b}}}{n_b n_{a\bar{b}}}$

TAB. 2 – *Formules des mesures d'intérêt supplémentaires.*

Descriptive Confirm, Example & Contra-Example, Least Contradiction) sont les plus grands et sont peu séparés des autres puisqu'ils ont les plus faibles valeurs de séparation (colonne 6, Tab. 4). Les mesures de ces 2 clusters offrent donc des points de vues très proches sur les règles. Cette proximité se confirme en observant les mesures constitutives des ces deux clusters, puisqu'on y trouve les mesures dérivées de la confiance. De plus, si l'on observe leur diamètre (colonne 5, Tab. 4), on remarque que le cluster 1 est plus petit que le cluster 2. Les mesures du cluster 1 offrent donc un point de vue plus similaire sur les règles que celles du cluster 2.

Les clusters 5 (Conviction), 14 (Sebag-Schoenaueur) et 15 (Support) sont constitués d'une seule mesure et disposent des plus grandes valeurs de séparation. Ils offrent donc des points de vues très différents sur les règles étudiées.

Les mesures représentatives sont choisies à partir des médioïdes, puisque ceux-ci ont la caractéristique d'être les plus proches des mesures de leur cluster et les plus éloignés des mesures des autres clusters. Les médioïdes permettent donc d'isoler une base réduite de mesures pour aider l'utilisateur à repérer les meilleures règles. Cette base contient l'ensemble des mesures les plus différentes, tout en conservant la richesse d'interprétation de tous les clusters.

5 Conclusion

Pour comprendre le comportement des mesures d'intérêt sur des ensembles de données spécifiques, nous avons étudié et comparé les diverses mesures d'intérêt décrites dans la littérature pour trouver une base minimale de meilleures mesures, appelées mesures représentatives.

Nous proposons de calculer les mesures représentatives à l'aide d'une classification par médioïdes (PAM). Cette technique nous permet non seulement d'obtenir les mesures représentatives, mais aussi une représentation graphique pour évaluer la classification.

En appliquant cette methode sur un jeu d'essai de 123228 règles, nous avons pu classer un ensemble de quarante mesures en seize clusters, et ainsi proposer de reduire le nombre de mesures à seize mesures représentatives.

Une interprétation de la classification effectuée est également proposée.

Nous envisageons d'améliorer notre classification, en utilisant une distance plus adaptée à nos données que celle calculée à partir du coefficient de corrélation linéaire.

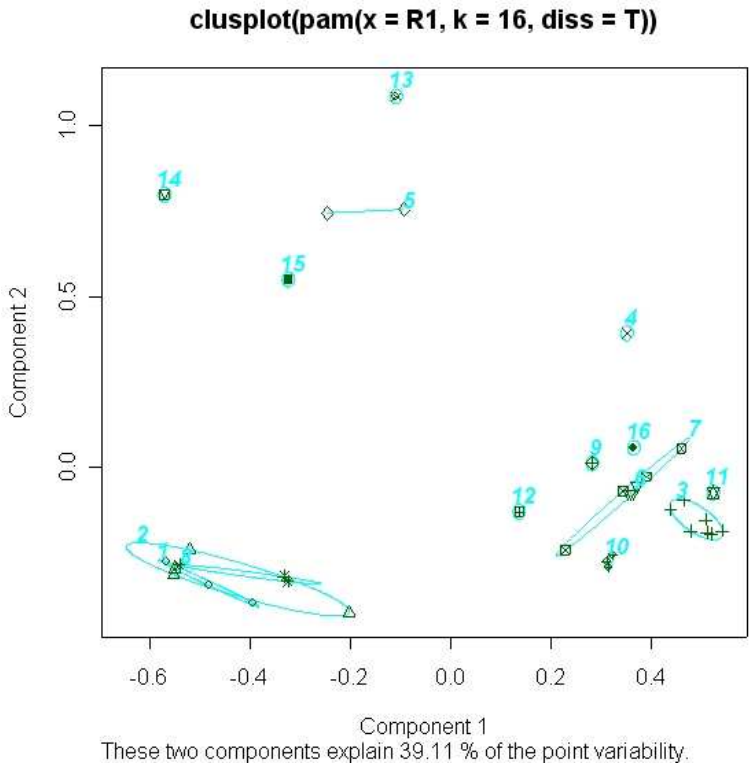

FIG. 1 – *Mesures et clusters projetés sur les deux axes principaux d'une ACP. Les mesures sont présentées par les symboles et les clusters sont présentées par les numéros.*

Références

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

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, 487–499.

Bayardo-Jr., R. J. et R. Agrawal (1999). Mining the most interestingness rules. *KDD'99, Proceedings of the 5th ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining*, 1145–154.

Blanchard, J., F. Guillet, et H. Briand (2003). Exploratory visualization for association rule rummaging. *MDM/KDD2003, Proceedings of the 4th International Workshop on Multimedia Data Mining, in conjunction with KDD'03*, 107–114.

Blanchard, J., F. Guillet, R. Gras, et H. Briand (2005a). Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. *ASMDA'05, Proceedings of the 11th*

Cluster	Nom des mesures d'intérêt	Mesure représentative
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Ganascia, Laplace	Causal Confirmed-Confidence
2	Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction	Example & Contra-Example
3	Causal Support,Kappa, Lerman, Phi-Coefficient, Rule Interest, Yule's Q, Yule's Y	Phi-Coefficient
4	Collective Strength	Collective Strength
5	Conviction, Odd Multiplier	Conviction
6	Cosine, F-measure, Jaccard	F-measure
7	Dependency, Gini-index, J-measure, Mutual Information	J-measure
8	EII, EII 2, IPEE	EII 2
9	II	II
10	Klosgen,Pavillon, Putative Causal Dependency	Klosgen
11	Lift	Lift
12	Loevinger	Loevinger
13	Odds Ratio	Odds Ratio
14	Sebag & Schoenauer	Sebag & Schoenauer
15	Support	Support
16	TIC	TIC

TAB. 3 – *Clusters des mesures obtenues.*

International Symposium on Applied Stochastic Models and Data Analysis, 191–200.

Blanchard, J., F. Guillet, R. Gras, et H. Briand (2005b). Using information-theoretic measures to assess association rule interestingness. *ICDM'05, Proceedings of the 5th IEEE International Conference on Data Mining*.

Blanchard, J., P. Kuntz, F. Guillet, et R. Gras (2003). Implication intensity: from the basic statistical definition to the entropic version (chap. 8). *Statistical Data Mining and Knowledge Discovery*, 475–493.

Boulicaut, J.-F., M. Klemettinen, et H. Mannila (1998). Querying inductive databases: a case study on the mine rule operator. *PKDD'98, Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery LNAI 1510*, 194–202.

Carvalho, D. R., A. A. Freitas, et N. F. F. Ebecken (2003). A critical review of rule surprisingness measures. *Proceedings of Data Mining IV - International Confeference on Data Mining*, 545–556.

Carvalho, D. R., A. A. Freitas, et N. F. F. Ebecken (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. *PKDD'05, the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*.

Choi, D. H., B. S. Ahn, et S. H. Kim (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. *ESA'05, Expert Sytems with Applications*, 1–12.

Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems 12(5-6)*, 309–315.

Gavrilov, M., D. Anguelov, P. Indyk, et R. Motwani (1999). Mining the stock market: which

N°	Taille	Distance maximale	Distance moyenne	Diamètre	Séparation
1	6	0.01671715	0.01279840	0.06292596	0.08627858
2	4	0.11566636	0.05082643	0.12651349	0.08627858
3	7	0.12441484	0.04555060	0.25912385	0.05987481
4	1	0.00000000	0.00000000	0.00000000	0.29374682
5	2	0.07081004	0.03540502	0.07081004	0.35996290
6	3	0.02809333	0.01194562	0.03109132	0.13513006
7	4	0.06636787	0.03827268	0.14087517	0.08086622
8	3	0.10709402	0.04553855	0.12428384	0.13022033
9	1	0.00000000	0.00000000	0.00000000	0.20684687
10	3	0.04239764	0.01994790	0.05368524	0.06515732
11	1	0.00000000	0.00000000	0.00000000	0.05987481
12	1	0.00000000	0.00000000	0.00000000	0.11371135
13	1	0.00000000	0.00000000	0.00000000	0.33950398
14	1	0.00000000	0.00000000	0.00000000	0.35996290
15	1	0.00000000	0.00000000	0.00000000	0.42697344
16	1	0.00000000	0.00000000	0.00000000	0.19199948

TAB. 4 – *Les informations supplémentaires obtenues sur les clusters de* R_1.

measure is best? *KDD'00, Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*, 487–496.

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

Hilderman, R. J. et H. J. Hamilton (2001). *Knowledge Discovery and Measures of Interestingness*. Kluwer Academic Publishers.

Huynh, H.-X., F. Guillet, et H. Briand (2005a). ARQAT: an exploratory analysis tool for interestingness measures. *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, 334–344.

Huynh, H.-X., F. Guillet, et H. Briand (2005b). Clustering interestingness measures with positive correlation. *ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems*, 248–253.

Huynh, H.-X., F. Guillet, et H. Briand (2005c). A data analysis approach for evaluating the behavior of interestingness measures. *DS'05, the 8th International Conference on Discovery Science LNAI 3735*, 330–337.

Kaufman, L. et P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kononenco, I. (1995). On biases in estimating multi-valued attributes. *IJCAI'95, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1034–1040.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de l'intérêt des règles d'association. *Mesures de Qualité pour la Fouille de Données RNTI-E-1*, 193–217.

Lallich, S., B. Vaillant, et P. Lenca (2005). Parametrised measures for the evaluation of asso-

ciation rules interestingness. *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, 220–229.

Lehn, R., F. Guillet, et H. Briand (2004). Qualité d'un ensemble de règles : élimination des règles redondantes. *Mesures de Qualité pour la Fouille de Données, RNTI-E-1*, 141–167.

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

Liu, B., W. Hsu, L.-F. Mun, et H.-Y. Lee (1999). Finding interestingness patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering 11(6)*, 817–832.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observation. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Newman, D. J., S. Hettich, C. L. Blake, et C. J. Merz (1998). [uci] repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html. *University of California, Irvine, Dept. of Information and Computer Sciences*.

Padmanabhan, B. et A. Tuzhilin (1998). A belief-driven method for discovering unexpected patterns. *KDD'1998, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, 94–100.

Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, 229–248.

Saporta, G. (1990). *Probabilité, analyse des données et statistique*. Edition Technip.

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

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. *DS'04, the 7th International Conference on Discovery Science LNAI 3245*, 290–297.

Vaillant, B., P. Picouet, et P. Lenca (2003). An extensible platform for rule quality measure benchmarking. *HCP'03, Human Centered Processes*, 187–191.

Zhao, Y. et G. Karypis. Criterion functions for document clustering: experiments and analysis. Technical report, Deparment of Computer Science, University of Minnesota. TR01-40.

Summary

This paper deals with finding a minimum set of interestingness measures in the stage of post-processing of association rules. These measures, called representative measures, are calculated with the help of a medoid clustering. The main interest of this approach is to deliver a reduced set of measures that is specific and adapted to each dataset studied. This reduction also facilitates the validation of the most interesting rules. Furthermore, the approach is applied to a rule-based dataset about 120000 association rules with 40 measures. As a result, we obtain a reduced set of sixteen representative measures. The paper also summarizes the state-of-the-art post-processing and the relative works about interestingness measures.