Vers davantage de flexibilité et d'expressivité dans les hiérarchies contextuelles des entrepôts de données

Yoann Pitarch, Cécile Favre, Anne Laurent, Pascal Poncelet

Vers davantage de flexibilité et d'expressivité dans les hiérarchies contextuelles des entrepôts de données

Yoann Pitarch*, Cécile Favre**
Anne Laurent***, Pascal Poncelet***

* Dept. d'Informatique, Aalborg, Danemark
ypitarch@cs.aau.dk
**ERIC, Lyon, France
cecile.favre@univ-lyon2.fr
***LIRMM, Montpellier, France
{laurent,poncelet}@lirmm

Résumé. Les entrepôts de données sont aujourd'hui couramment utilisés pour analyser efficacement des volumes importants de données hétérogènes. Une des clés de ce succès est l'utilisation des hiérarchies qui permettent de considérer les données à différents niveaux de granularité. Au cours de travaux précédents, nous avons montré que ce processus de généralisation n'introduisait que très peu de connaissances expertes conduisant ainsi à des analyses erronées. Nous avons alors proposé une nouvelle catégorie de hiérarchie, les hiérarchies contextuelles, pour répondre à cette limite. Malheureusement, les connaissances expertes introduites devaient être spécifiées de façon rigide et ne pouvaient donc refléter toute leur réelle complexité. Dans cet article, nous étendons ces hiérarchies et les mécanismes associés pour accroître grandement leur flexibilité et leur expressivité et utilisons pour cela une approche basée sur la logique floue. Ce qui est présenté ici constitue un travail préliminaire visant à montrer l'intérêt et la faisabilité de l'idée.

1 Introduction

Les entrepôts de données (Inmon, 1996) visent à consolider des données à des fins d'analyse, en adoptant une modélisation dite *multidimensionnelle* au travers de laquelle des faits sont analysés au moyen d'indicateurs (les mesures) selon différents axes d'analyse (les dimensions). En s'appuyant sur des mécanismes d'agrégation, les outils OLAP (On Line Analytical Process) permettent de naviguer aisément le long des hiérarchies des dimensions (Malinowski et Zimányi, 2004). L'intérêt des décideurs pour ce type d'application s'avère être croissant, et ce quel que soit leur domaine d'application : marketing, suivi de production, médical... Dans cet article, nous nous proposons d'illustrer nos propos par rapport à ce dernier domaine puisque ce travail avait débuté dans le cadre d'un projet ANR [1].

A l'heure où l'accroissement du volume de données incite à proposer de nouvelles solutions de gérance de l'informatique telles que le *cloud computing* où les données s'éloignent des utilisateurs et induisent de nombreuses problématiques parmi lesquelles la sécurité, le souci des utilisateurs s'est parallèlement renforcé avec l'intensification de travaux, dans le contexte

1. Projet ANR MIDAS (ANR-07-MDCO-008)

des entrepôts de données et de l'analyse OLAP entre autres, sur la personnalisation, la recommandation et l'intégration de connaissances utilisateurs. Ainsi, dans ce travail, nous mettons en avant l'importance des connaissances et de leur expression dans une phase de structuration des données qui sont massives et complexes. Cette phase de structuration vient en amont de la phase de fouille de données qui pourra tirer partie de deux constats : 1) l'entrepôt intègre par définition des sources de données variées et constitue donc une source *propre* de données de qualité ; 2) l'intégration de connaissances que nous proposons permet une réutilisation de ces connaissances lors du processus de fouille.

Considérons le cas réel d'un entrepôt de données médicales rassemblant les paramètres vitaux (e.g., la tension artérielle...) des patients d'un service de réanimation. Afin de réaliser un suivi efficace des patients, un médecin souhaiterait par exemple connaître ceux qui ont eu une tension artérielle basse au cours de la nuit. Pouvoir formuler ce type de requête suppose l'existence d'une hiérarchie sur la tension artérielle dont le premier niveau d'agrégation serait une catégorisation de la tension artérielle (e.g., basse, normale, élevée), hiérarchie qui pourra être exploitée durant un processus de navigation dans les données. Toutefois, cette catégorisation est délicate car elle dépend à la fois de la tension artérielle mesurée mais aussi de certaines caractéristiques physiologiques (âge du patient, fumeur ou non, ...). Dès lors, une même tension peut être généralisée différemment selon le contexte d'analyse considéré. Par exemple, 13 est une tension élevée chez un nourrisson alors qu'il s'agit d'une tension normale chez un adulte. Ainsi, pour permettre cette généralisation, nous avons été amenés dans des travaux précédents (Pitarch et al., 2010a,b) à proposer un nouveau type de hiérarchies qualifiées de *contextuelles*.

La définition de ces hiérarchies permet donc d'exprimer la connaissance métier des utilisateurs pour ensuite les exploiter lors du processus de navigation et d'analyse. Il s'avère aujourd'hui que nous avons besoin d'aller au-delà en matière d'expressivité de ces connaissances en intégrant le fait que ces connaissances peuvent être plus ou moins exactes et que le contenu même de ces connaissances peut être plus ou moins précis.

En matière d'expressivité de connaissances, l'approche floue trouve un réel intérêt (Dubois, 2011). Différents travaux se sont d'ailleurs intéressés à l'intégration de la logique floue dans les entrepôts de données pour augmenter la richesse des hiérarchies entre autres (Fasel et Shahzad, 2010; Perez et al., 2007). Ainsi, dans ce travail, nous nous proposons de nous intéresser à l'intégration de la logique floue dans le cadre des hiérarchies contextuelles, franchissant un pas supplémentaire dans l'expressivité des hiérarchies, assurant une analyse encore plus pertinente lors du processus de navigation. Les résultats présentés ici constituent un travail préliminaire visant à montrer l'intérêt et la faisabilité de cette idée.

La suite de ce papier est organisée de la façon suivante. La section 2 expose notre cas d'étude sur des données médicales qui permettra d'illustrer par la suite le problème posé et la solution apportée. Dans la section 3, nous revenons sur le principe et la formalisation des hiérarchies contextuelles. Nous proposons alors dans la section 4 l'adaptation de notre solution au contexte du flou. Puis nous discutons de l'implémentation de cette solution dans la section 5. Enfin, nous concluons et indiquons les perspectives de ce travail dans la section 6.

2 Cas d'étude

Afin d'illustrer les limites des hiérarchies contextuelles actuelles et montrer comment nous les pallions dans cet article, considérons le scénario médical suivant. Pour chaque patient, le

système enregistre son âge, sa consommation moyenne quotidienne de tabac, son traitement médical éventuel et sa tension à un instant donné. Nous souhaitons déterminer si la tension doit être généralisée en *faible*, *normale* ou *élevée*. En outre, l'attribut associé à l'âge (resp. à la consommation de tabac) peut être considéré à deux niveaux de granularité différents : l'âge (resp. la quantité exacte de cigarettes fumées) et la catégorie de l'âge (resp. la catégorie de consommation de tabac). Le tableau 1 présente les informations associées au patient $P1$ et la figure 1(a) indique les extraits des hiérarchies utiles dans ce cas d'étude. Comme argumenté dans Pitarch et al. (2010b), la généralisation d'une tension artérielle est contextuelle par nature puisque de nombreux paramètres peuvent influer sur le résultat de cette généralisation. La figure 1(b) répertorie quelques connaissances expertes pour guider le processus de généralisation. Par exemple, $R4$ indique qu'une tension supérieure à 14 est élevée pour un fumeur régulier.

En analysant conjointement le tableau 1 et les figures 1(a)-1(b), plusieurs observations peuvent être faites :

1. Les connaissances ne sont pas nécessairement exprimées sur l'ensemble des attributs impactant la généralisation d'une tension. Par exemple, $R4$ spécifie des conditions sur les attributs $CatConsoTabac$ et $Tension$ (TA) alors que $R5$ spécifie des conditions sur les attributs $Traitement$ et $Tension$ (TA).
2. Les connaissances $R5$ et $R6$ sont toutes deux adaptées donc applicables à $P1$ mais diffèrent quant au résultat de la généralisation de la tension.
3. La connaissance $R5$ est plus précise que la connaissance $R6$, *i.e.*, elle définit un plus grand nombre de conditions. Il apparaît donc plus raisonnable de généraliser en prenant en compte $R5$.
4. Bien que ne s'appliquant pas strictement à $P1$, les conditions des connaissances $R2$ et $R3$ sont *presque* réunies pour le patient $P1$. Typiquement, il aurait fallu que $P1$ consomme seulement une seule cigarette supplémentaire par jour pour que la connaissance $R3$ puisse être appliquée.

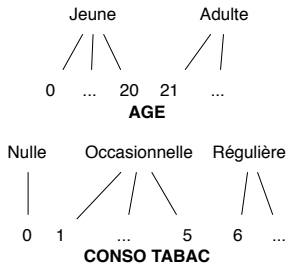
IdP	Age	CatAge	ConsoTabac	CatConsoTabac	Traitement	Tension (TA)	CatTension
P1	22	Adulte	5	Occasionnelle	Hypotenseur	13	?
...	...	...	...	...	...	...	...

TAB. 1 – *Patient exemple*

Nous montrons dans la prochaine section que les hiérarchies contextuelles, telles que nous les avons définies précédemment, n'intègrent que partiellement ces remarques et proposons ensuite une solution à ce problème.

3 Les hiérarchies contextuelles

Nous rappelons d'abord les définitions associées aux hiérarchies contextuelles classiques puis les solutions proposées pour les mettre en œuvre. Cette section s'achève par une discussion autour des limites de ce modèle en matière de flexibilité et d'expressivité.



(a) Hiérarchies

IdC	Connaissance
R1	Jeune + Nulle ou occasionnelle + TA $\in [8;12] \rightarrow$ Normale
R2	Jeune + Nulle ou occasionnelle + $TA > 12 \rightarrow$ Elevée
R3	Régulière + $TA \in [9;14] \rightarrow$ Normale
R4	Régulière + $TA > 14 \rightarrow$ Elevée
R5	Hypotenseur + $TA > 12 \rightarrow$ Elevée
R6	$TA \in [8;13] \rightarrow$ Normale
...	...

(b) Connaissances expertes

FIG. 1 – *Connaissances utiles pour la généralisation contextuelle*

3.1 Définitions

Les hiérarchies contextuelles ont été proposées dans le cadre des entrepôts de données. Le formalisme introduit redéfinit donc les concepts bien connus de dimensions, tables de faits, attributs de dimensions et mesures. Par souci de simplicité et sans perte de généralité, les définitions présentées dans cette section s'abstraient de ce cadre formel et supposent simplement l'existence d'un ensemble d'identifiants déterminant fonctionnellement un ensemble d'attributs.

Nous notons $Id = \{id_1, \ldots, id_n\}$ l'ensemble des n identifiants et $\mathcal{A} = \{A_1, \ldots, A_t\}$ l'ensemble des t attributs. Nous supposons que A_t est l'attribut à généraliser contextuellement. Pour chaque attribut A_i, $i = 1 \ldots t$, son domaine de définition est noté $Dom(A_i)$. Chaque attribut A_i est équipé d'une hiérarchie et peut donc être observé selon plusieurs granularités $A_i = A_i^0, \ldots, A_i^{M_i}$. De telles hiérarchies sont dites *simples*. Par convention, A_i^0 désigne la granularité la plus fine et $A_i^{M_i}$ désigne la granularité la plus élevée. Nous notons $a_i^j \in Dom(A_i^j)$ si la valeur a_i^j représente une instance du niveau A_i^j. Nous notons $\mathcal{A}^*$ l'ensemble $\mathcal{A}^* = \{A_1^0, A_1^1, \ldots, A_t^{M_t}\}$. Le schéma de T, la table relationnelle manipulée, est donc $T = (Id, A_1^0, A_1^1, \ldots, A_t^{M_t})$ tel que $Id \rightarrow A_1^0 \times A_1^1 \times \ldots \times, A_t^{M_t}$. Un n-uplet t de T est de la forme $t = (id, a_1^0, a_1^1, \ldots, a_t^0, \ldots, a_t^{M_t})$ tel que a_t^0 est la valeur à généraliser contextuellement et $a_t^1, \ldots, a_t^{M_t}$ en sont les généralisations contextuelles. Le tableau 1 présente un extrait de la table relationnelle utilisée dans notre cas d'étude.

Pour un attribut donné, le passage entre un niveau et sa granularité supérieure directe est permis par les *chemins d'agrégation* définis comme suit.

Définition 1 *(Chemin d'agrégation) Soient A_i^j et A_i^{j+1}, $j = 0, \ldots, M_i - 1$, deux niveaux de granularité de l'attributs A_i. Un* chemin d'agrégation*, noté G, entre A_i^j et A_i^{j+1}, est défini par $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ avec $\mathrm{C} = \{A_i^j, \ldots\}$ tel que (1) A_i^j est appelé* attribut source*, (2) $A_i^{j+1} \notin \mathrm{C}$ et (3) $\mathrm{C} \subseteq \mathcal{A}$ et A_i^{j+1} dépend fonctionnellement de* C*.*

Définition 2 *(Chemin d'agrégation contextuel et classique) Soit* $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ *un chemin d'agrégation.* G *est appelé un* chemin d'agrégation contextuel *si* $|\mathrm{C}| > 1$. *Sinon* G *est appelé un* chemin d'agrégation classique *(ou simplement un chemin d'agrégation quand cela est sans ambiguïté).*

Définition 3 *(Attribut contextualisant et contextualisé) Soit* $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ *un chemin d'agrégation contextuel. L'attribut* A_i^{j+1} *est dit* contextualisé *par les attributs de* C. *Les attributs de* C *sont appelés attributs* contextualisants.

Définition 4 *(Contexte) Soit* $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ *un chemin d'agrégation contextuel. Le triplet* $c = (A_i^j, \mathrm{C}, A_i^{j+1})$ *est appelé le* contexte *de l'attribut* A_i^{j+1} *et nous notons* $\mathcal{C} = \{c_1, \ldots, c_k\}$ *l'ensemble des contextes.*

Définition 5 *(Instance d'un chemin d'agrégation) Soit* $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ *un chemin d'agrégation. Une instance de* G *, notée* $g = a \overset{\mathrm{IC}}{\longmapsto} a'$*, est telle que :*

1. $a \in Dom(A_i^j)$ *et* $a' \in Dom(A_i^{j+1})$
2. $\mathrm{IC} = \{(A_l^m, \alpha) \mid A_l^m \in \mathrm{C}$ *et* $\alpha \subseteq Dom(A_l^m)\}$
3. $\exists (A_l^m, \alpha)$ *tel que* $A_i^j = A_l^m$ *et* $a \in \alpha$
4. $\nexists A_l^m$ *tel que* $A_l^m \in \mathrm{C}$ *et* $(A_l^m, \alpha) \notin \mathrm{IC}$

Définition 6 *(Instance de contexte) Soit* $g = a \overset{\mathrm{IC}}{\longmapsto} a'$ *une instance du chemin d'agrégation* $A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$*. La paire* $c^l = (\mathrm{IC}, a')$ *est appelée une* instance *du contexte* c *de* A_i^{j+1} *et nous notons* $\mathcal{IC} = \{c_1^1, \ldots, c_k^{l_k}\}$ *l'ensemble des instances de contextes.*

Exemple 1 *Considérons que la généralisation du niveau* Tension *au niveau* CatTension *dépend uniquement des attributs* CatConsoTabac *et* Tension*. Dès lors le chemin d'agrégation contextuel associé est noté* $G = Tension \overset{\mathrm{C}}{\mapsto} CatTension$ *avec* $\mathrm{C} = \{CatConsoTabac, Tension\}$ *et le contexte correspondant est noté* $c = (Tension, \mathrm{C}, CatTension)$*. Dans ce contexte,* $CatConsoTabac$ *et* $Tension$ *sont les attributs contextualisants et* $CatTension$ *est l'attribut contextualisé. La connaissance associée à* $R4$ *est représentée par l'instance de contexte* $c^l = (\mathrm{IC}, \mathrm{Elevée})$ *avec* $\mathrm{IC} = \{(CatConsoTabac, \{\mathrm{Régulière}\}), (Tension, \{x > 14\})\}$.

3.2 Mise en œuvre

3.2.1 Stockage de la connaissance

Dans Pitarch et al. (2010b), nous proposons de représenter les connaissances expertes (structures et instances de contexte) via une base de données relationnelles externe, notée $\mathcal{B}_{Know.}$ afin de garantir la généricité du stockage des différents contextes présents dans l'entrepôt. Cette base de données est seulement composée de deux relations que nous décrivons brièvement.

La relation `MTC` (`M`éta-`T`able des `C`onnaissances) permet de stocker les contextes, *i.e.*, quels sont les attributs contextualisants et contextualisés pour chaque contexte, et possède la structure suivante :

- **Contexte** désigne l'identifiant du contexte ;
- **Attribut** désigne un attribut intervenant dans le contexte ;
- **Type** définit le rôle joué par *Attribut* dans le contexte. *Type* vaut "*Contexte*" si *Attribut* est un attribut *contextualisant* dans ce contexte et vaut "*Résultat*" si *Attribut* est l'attribut contextualisé.

La relation `TC` (`T`able des `C`onnaissances) permet de stocker quelles sont les instances des différents contextes, *i.e.*, la connaissance experte, et possède la structure suivante :

- **Contexte** désigne l'identifiant du contexte concerné ;
- **Instance_Contexte** identifie l'instance du contexte ;
- **Attribut** désigne un attribut intervenant dans le contexte ;
- **Valeur** représente l'ensemble de valeurs d'*Attribut* concerné dans l'instance en question.

3.2.2 Généralisation contextuelle

Par manque de place, le fonctionnement de `ROLL_UP_CTX`, l'opérateur de généralisation contextuelle, est présenté succinctement [2]. D'abord, il faut identifier dans la table `MTC` quel est le contexte en jeu et quels sont les attributs contextualisants. Une recherche dans la table `TC` permet ensuite d'identifier l'unique instance concernée dans cette généralisation. Dès lors, il suffit de retourner la valeur de l'attribut contextualisé correspondant à cette instance.

3.3 Discussion

Les hiérarchies contextuelles permettent la prise en compte de connaissances expertes dans le processus de généralisation et représentent donc une solution originale et intéressante dans de nombreux domaines d'applications où l'on doit donner du sens à des attributs numériques. Malgré tout, ces hiérarchies contextuelles souffrent encore de quelques limites. Précisément, dans le modèle actuel, il ne doit exister qu'une (contrainte d'existence) et une seule (contrainte d'unicité) instance de contexte pour une généralisation contextuelle. Nous détaillons ci-dessous les arguments qui motivent de relâcher quelque peu ces contraintes.

Tout d'abord, la contrainte d'existence implique que l'ensemble des connaissances soit complet au sens des instances de T. Sans cette garantie, la tension artérielle d'un patient pourrait ne pas être généralisée par exemple. Pourtant, malgré l'existence d'outils pour vérifier la complétion d'un ensemble de règles, ceux-ci montrent leur limite quand il s'agit de traiter de nombreux attributs (Ligęza (2006)). Ce point pose problème dans la mesure où il est impossible de contrôler le nombre d'attributs pouvant impacter sur une généralisation.

Ensuite, la contrainte d'unicité garantit certes la consistance de la connaissance experte mais est en pratique très difficile à obtenir. Plusieurs arguments étayent cette affirmation. D'abord, bien que les attributs contextualisants représentent l'ensemble des facteurs pouvant potentiellement influer sur une généralisation, rien ne garantit que l'impact conjoint de ces attributs ait été étudié. Par exemple, il n'existe pas de connaissances portant sur tous les attributs contextualisants dans notre cas d'étude. Ensuite, l'appartenance d'un n-uplet à une instance de contexte est parfois délicate à définir strictement. Par exemple, comme nous l'avons remar-

2. Les lecteurs intéressés sont invités à se référer à Pitarch (2011) pour plus de détails.

qué plus tôt, les conditions de $R3$ sont presque réunies pour le patient $P1$. Pour pallier cette rigidité, l'utilisation des hiérarchies floues apparaît très prometteuse.

Relâcher les contraintes d'existence et d'unicité augmenterait considérablement la flexibilité et l'expressivité des hiérarchies contextuelles mais soulève de nouvelles problématiques quant au stockage de la connaissance et au processus de généralisation. La prochaine section détaille ces problématiques et décrit les modifications apportées au modèle.

4 Hiérarchies contextuelles généralisées

Pour pallier les limites précédemment mentionnées, nous étendons le modèle actuel de deux manières. D'abord, nous introduisons le concept d'*instance de contexte généralisée* qui relâche la contrainte d'unicité, autorise les instances de contexte à ne pas inclure systématiquement l'ensemble des attributs contextualisants et permet la définition d'un ordre partiel au sein des instances d'un contexte donné. Nous étudions ensuite les conséquences de l'*utilisation des hiérarchies floues* pour les attributs contextualisants et définissons alors un degré d'appartenance d'un n-uplet à une instance de contexte.

4.1 Instance de contexte généralisée

Nous l'avons vu, garantir la complétude au sens des instances de la base peut se révéler délicat. En outre, cette complétude peut très bien être assurée à un certains temps t et ne plus l'être au temps $t+1$ lorsque de nouvelles instances sont insérées dans la table. Garantir une complétude universelle est alors préférable. Une solution pour garantir cette complétude universelle est d'autoriser plusieurs degrés de précision dans les instances de contexte. Par exemple, l'insertion de connaissances très générales, telles que $R6$, garantirait qu'une tension sera toujours généralisée. Ces connaissances pourraient être complétées par des connaissances bien plus précises telles que $R1$ et $R2$. Intuitivement, le processus de généralisation favoriserait alors l'instance adéquate la plus précise. Nous formalisons ci-dessous ce concept d'instance de contexte généralisée et introduisons une relation d'ordre partiel entre les instances.

Définition 7 *(Instance généralisée d'un chemin d'agrégation contextuel) Soit* $G = A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$ *un chemin d'agrégation contextuel. Une instance généralisée de* G *, notée* $g = a \overset{\mathrm{IC}}{\longmapsto} a'$*, est telle que :*

1. $a \in Dom(A_i^j)$ *et* $a' \in Dom(A_i^{j+1})$
2. $\mathrm{IC} = \{(A_l^m, \alpha) \mid A_l^m \in \mathrm{C} \wedge \alpha \subseteq Dom(A_l^m)\}$
3. $\exists (A_l^m, \alpha) | A_i^j = A_l^m \wedge a \in \alpha$

Définition 8 *(Instance généralisée de contexte) Soit* $g = a \overset{\mathrm{IC}}{\longmapsto} a'$ *une instance généralisée du chemin d'agrégation* $A_i^j \overset{\mathrm{C}}{\mapsto} A_i^{j+1}$*. La paire* $c^l = (\mathrm{IC}, a')$ *est appelée une* instance *généralisée du contexte* c *de* A_i^{j+1} *et nous notons* $\mathcal{IC} = \{c_1^1, \ldots, c_k^{l_k}\}$ *l'ensemble des instances généralisées de contextes.*

Définition 9 *(Précision d'une instance généralisée de contexte) Soit* $c^l = (\mathrm{IC}, a)$ *une instance généralisée du contexte* $c = (A_i^j, \mathrm{C}, A_i^{j+1})$. *La précision de* c^l, *notée* $Prec(c^l)$, *est définie comme* $Prec(c^l) = |\mathrm{IC}|$.

Nous notons $Attrib(g)$ les attributs contextualisants présents dans l'instance g.

Exemple 2 $c^l = \big(\{(\textit{Tension}, [8; 13])\}, \textit{Normale}\big)$ *est l'instance généralisée du contexte* $c = (Tension, \{Tension, CatAge, CatConsoTabac, Traitement\}, CatTension)$ *qui représente la connaissance* $R6$. *De plus, nous avons* $Attrib(c^l) = \{Tension\}$ *et* $Prec(c^l) = 1$.

Remarquons que, dans le cadre formel initial, la précision d'une instance de contexte est toujours le nombre d'attributs contextualisants. Désormais, des instances pouvant être plus ou moins générales, il est naturel de définir un ordre partiel pour comparer le niveau de précision de deux instances.

Définition 10 *(Ordre partiel sur les instances d'un contexte) Soient* $c^l = (\mathrm{IC}, a)$ *et* $c^{l'} = (\mathrm{IC}', a')$ *deux instances du contexte* $c = (A_i^j, \mathrm{C}, A_i^{j+1})$. c^l *spécialise* $c^{l'}$ *(resp.* $c^{l'}$ *généralise* c^l*), noté* $c^l \prec c^{l'}$ *(resp.* $c^{l'} \succ c^l$*) , si :*

(1) $Attr(c^{l'}) \subseteq Attr(c^l)$ *;*

(2) $\alpha \subseteq \alpha'$ *avec* $(A_k^m, \alpha) \in \mathrm{IC}$, $(A_k^m, \alpha') \in \mathrm{IC}'$ *et* $A_k^m \in (Attr(c^{l'}) \cap Attr(c^l)) - \{A_i^j\}$.

Exemple 3 *Soient* $c^1 = \big(\{(\textit{Tension}, [8; 13])\}, \textit{Normale}\big)$, $c^2 = \big(\{(\textit{Tension}, [8; 13]), (\textit{CatAge}, \{Jeune, Adulte\})\}, \textit{Normale}\big)$ *et* $c^3 = \big(\{(\textit{Tension}, [9; 12]), (\textit{CatAge}, \{Jeune\})\}, \textit{Normale}\big)$ *trois instances de contexte. Nous avons* $c^1 \prec c^2$, $c^2 \prec c^3$ *et* $c^1 \prec c^3$ [3].

4.2 Prise en compte de hiérarchies floues

Introduits par Zadeh (1965), les sous-ensembles flous permettent de modéliser la représentation humaine des connaissances et améliorent ainsi les performances des systèmes de décision qui utilisent cette modélisation. Plus particulièrement, les hiérarchies floues présentées par Laurent (2002) permettent de modéliser l'appartenance d'un élément à plusieurs généralisations avec des degrés de confiance propres à chacune d'entre elles. L'utilisation de ces hiérarchies apparaît alors particulièrement adaptée pour représenter le fait qu'une instance *correspond presque* à un n-uplet.

Etant donné un ensemble de référence $Dom(A_i^j)$, avec $j = 1, \ldots, M_i$, un sous-ensemble flou A de $Dom(A_i^j)$ est alors défini par une fonction d'appartenance f_A qui associe à chaque élément a du niveau $Dom(A_i^0)$ le degré $f_A(a)$, compris entre 0 et 1, reflétant une gradualité dans son appartenance à A. Par exemple, $f_{CatAge=Adulte}(22) = 0,7$ indique que 22 appartient, à un degré de $0,7$ au sous-ensemble $\{Adulte\}$.

L'utilisation des hiérarchies floues nous contraint à redéfinir le contenu de la table T pour stocker les divers degrés de confiance. Un n-uplet $t = (id, a_1^0, a_1^1, \ldots, a_t^0, \ldots, a_t^{M_t})$ de T sera désormais stocké sous la forme $t' = (id', e_1^0, e_1^1, \ldots, e_t^0, \ldots, e_t^{M_t})$ tels que $id = id'$, $e_i^0 = a_i^0$ pour $i = 1, \ldots, t$ et $e_i^j = \{(\alpha_i^j, f_{\alpha_i^j}(e_i^0)) \mid \alpha_i^j \in Dom(A_i^j)$ *et* $f_{\alpha_i^j}(e_i^0) \neq 0\}$ pour $i = 1, \ldots, t-1$ et $j = 1, \ldots, M_i$. Le tableau 2 illustre comment les hiérarchies floues sur les

IdP	Age	CatAge	ConsoTabac	CatConsoTabac	Traitement	Tension	CatTension
P1	22	$f_{CatAge=Jeune}(22) = 0,25$ $f_{CatAge=Adulte}(22) = 0,7$	5	$f_{CCT=occ.}(5) = 0,3$ $f_{CCT=reg.}(5) = 0,6$	Hypotenseur	13	?

TAB. 2 – *Patient exemple avec intégration des hiérarchies floues*

attributs relatifs à l'âge et à la consommation de tabac sont intégrées dans la table relationnelle de notre cas d'étude.

Dès lors, un n-uplet flou de T peut être associé à plusieurs instances de contexte avec des degrés d'adéquation distincts. Nous nous intéressons maintenant à la définition de ce degré d'adéquation entre une instance et un n-uplet. Intuitivement, le principe est identique au cas strict : il s'agit de vérifier séparément l'adéquation du n-uplet à chaque condition d'une instance puis de les combiner.

Etudions d'abord comment calculer le degré d'adéquation d'un n-uplet à une condition d'une instance. Supposons que l'on souhaite calculer à quel point le patient $P1$ valide le critère $(SubCatAge, \{Jeune, Adulte\})$. Ici, il s'agit de mesurer le degré d'appartenance de l'élément 22 au sous-ensemble flou $\{Jeune, Adulte\}$ ou, autrement dit, à l'union des sous-ensembles flous $\{Jeune\}$ et $\{Adulte\}$. L'opérateur de t-conorme, noté $\perp$, est l'extension floue de l'opérateur d'union. Nous l'utilisons pour définir l'adéquation d'un n-uplet à un critère.

Définition 11 *(Adéquation d'une instance à une condition) Soient* $cond = (A_i^j, \alpha)$ *une condition (avec* $j = 0, \ldots, M_i$ *et* $\alpha = \{\alpha_1, \ldots, \alpha_k\}$*) et* $t' = (id', e_1^0, e_1^1, \ldots, e_t^0, \ldots, e_t^{M_t})$ *un n-uplet de* T*. L'adéquation de* t' *à* $cond$*, notée* $\mu_{unit}(t', cond)$*, est définie comme :*

$$\mu_{unit}(t', cond) = \perp\Big(f_{\alpha_1}(e_i^0), \ldots, f_{\alpha_k}(e_i^0)\Big)$$

Remarquons qu'il existe là aussi dans la littérature plusieurs fonctions pour définir la t-conorme. Une étude approfondie de la fonction la plus appropriée serait pertinente. Néanmoins, nous utilisons à nouveau la fonction proposée par Zadeh, à savoir la fonction max.

Exemple 4 *Soient* t *le n-uplet correspondant au patient* $P1$ *et* $cond = (CatAge, \{Jeune, Adulte\})$*, un critère. Nous avons* $\mu_{unit}(t, cond) = max(f_{CatAge=Jeune}; f_{CatAge=Adulte}) = 0,7$.

Etudions maintenant comment combiner ces adéquations locales pour calculer le degré d'adéquation globale d'un n-uplet à une instance de contexte, *i.e.*, à une conjonction de conditions. Supposons que l'on souhaite déterminer à quel point le patient $P1$ valide l'instance globale. Ici, il s'agit de mesurer le degré d'appartenance de $P1$ à toutes les conditions de $R1$ ou, autrement dit, à l'intersection des sous-ensembles flous représentés que sont chaque condition. L'opérateur de t-norme, noté $\top$, est l'extension floue de l'opérateur d'intersection. Nous l'utilisons pour définir l'adéquation d'un n-uplet à une instance.

Définition 12 *(Adéquation d'une instance à un n-uplet) Soient* $c^l = (\mathrm{IC}, a)$*, une instance du contexte* $c = (A_i^j, \mathrm{C}, A_i^{j+1})$ *et* $t' = (Id, e_1^0, e_1^1, \ldots, e_t^0, \ldots, e_t^{M_t})$*, un n-uplet de* T*. L'adéquation de* t' *à* c^l*, notée* $\mu(t', c^l)$*, est définie comme :*

$$\mu(t', c^l) = \top\Big(\mu_{unit}\big(t', (A_k^m, \alpha)\big) \text{ tel que } A_k^m \in Attrib(c^l)\Big)$$

3. Des investigations supplémentaires sont nécessaires pour prouver la transitivité de $\prec$ dans le cadre général.

Remarquons qu'il existe dans la littérature plusieurs fonctions pour définir la t-norme. Une étude approfondie de la fonction la plus appropriée serait pertinente. Néanmoins, nous utilisons ici la fonction proposée par Zadeh, la fonction min.

Exemple 5 *Soit t le n-uplet correspondant au patient* $P1$ *et* $c^l = \big(\{(\textit{Tension}, [8; 13]), (\textit{CatAge}, \{Jeune, Adulte\})\}, \textit{Normale}\big)$ *une instance de contexte. Nous avons :*

$$\begin{aligned}\mu(t, c^l) &= min\Big(\mu_{unit}\big(t, (\textit{Tension}, [8; 13])\big); \\ &\qquad \mu_{unit}\big(t, (\textit{CatAge}, \{Jeune, Adulte\})\big)\Big) \\ &= min(0,7\ ;\ 1) \\ &= 0,7\end{aligned}$$

5 Discussion autour de la mise en œuvre

Le cadre formel étant étendu, nous discutons maintenant ses implications sur deux aspects essentiels : la représentation des connaissances et la généralisation contextuelle.

5.1 Stockage

Un premier point très intéressant à soulever ici est que l'extension des hiérarchies contextuelles telle que définie dans la section précédente n'impacte pas sur le mode de stockage de la connaissance. En effet, dans la mesure où la définition d'un contexte reste inchangée, *i.e.*, l'existence d'attributs contextualisants et d'un attribut contextualisé est toujours valide, la structure de la relation `MTC` ne doit subir aucune modification. En outre, la solution proposée précédemment pour stocker les instances de contexte ne contraint pas à spécifier un ensemble de valeurs pour chaque attribut contextualisant. Dès lors, la relation `TC` peut stocker des instances généralisées de contexte sans aucune modification structurelle. Enfin, la présence de hiérarchies floues est propre à la table T considérée et n'a aucune incidence sur la base de données externe.

Cependant, il est important de garder en tête que, dans un contexte analytique, le nombre de généralisations à réaliser peut être très important. Il est donc nécessaire de stocker les connaissances de sorte à ce qu'elles soient efficacement exploitables. Dans un travail précédent, nous avons choisi de stocker la connaissance dans une base externe. Cette solution offre l'avantage de la généricité et nous permet de bénéficier de techniques avancées d'indexation et d'accès aux données. Cependant, un autre mode de stockage peut faire sens : le stockage de la connaissance sous forme arborescente via un fichier XML qui serait chargé en fonction des besoins. Cette solution entraînerait certes une perte de généricité dans le stockage, *i.e.*, un arbre de connaissance par contexte, mais possède des propriétés intéressantes. D'abord, d'un point de vue algorithmique, la recherche d'information dans un arbre peut être exécutée très efficacement. De plus, cette solution minimiserait grandement le nombre d'accès à la base externe qui est souvent prohibitif dans les applications nécessitant des accès fréquents à une base. Aussi, à court terme, il s'agira de réaliser une étude comparative de ces deux solutions afin d'élire celle offrant les meilleures performances.

5.2 Extension de l'algorithme de généralisation contextuelle

L'algorithme de généralisation présenté précédemment ne peut plus être appliqué car il supposait l'existence d'une unique instance adéquate. Cette contrainte relâchée, il est désormais nécessaire d'introduire un mécanisme pour élire l'instance qui sera utilisée pour généraliser. Nous avons introduit deux méthodes pour relâcher cette contrainte : la prise en compte d'instances généralisées et la possibilité d'utiliser des hiérarchies floues pour les attributs contextualisants. Dès lors, l'ensemble des instances candidates à la généralisation peut contenir des instances plus ou moins générales et/ou plus ou moins adéquates. Une fonction de score agrégeant ces deux mesures de qualité, *i.e.*, généralité et adéquation, doit donc être définie.

Définition 13 *(Score) Soient t un n-uplet de T et c^l une instance du contexte c. Le score quantifiant à quel point c^l représente t s'écrit :*

$$Score(c^l, t) = Agr\big(Prec(c^l),\ \mu(t, c^l)\big)$$

Des investigations supplémentaires sont nécessaires pour définir précisément cette fonction de score. En effet, la généralisation contextuelle choisie sera celle associée à l'instance maximisant ce score. Cette fonction occupe donc une place cruciale dans le processus de généralisation. Nous pensons qu'elle doit être définie en accord avec les utilisateurs du système. Typiquement, certains pourraient vouloir favoriser les instances plus générales plutôt que l'adéquation ou bien des instances faisant intervenir certains attributs, *e.g.*, les instances précisant le traitement médical.

Cette fonction supposée définie, nous nous intéressons maintenant aux modifications à apporter nécessairement pour définir `ROLL_UP_CTX_GEN`, l'opérateur étendu de généralisation contextuelle. La première étape de sélection du contexte est commune aux deux opérateurs. Ensuite, il s'agit de rechercher dans `TC` les instances de contexte pouvant être appliquées au n-uplet considéré. Afin de ne pas considérer des instances redondantes, nous exploitons l'ordre partiel défini dans la section 4.1 pour éliminer les instances dont il existe déjà une spécialisation dans l'ensemble des candidats. Ensuite, la fonction de score est appliquée aux instances restantes. La généralisation retournée sera finalement celle associée à l'instance maximisant le score.

L'algorithme de généralisation introduit deux nouvelles étapes qui peuvent se révéler coûteuses. L'implantation efficace de l'opérateur `ROLL_UP_CTX_GEN` passe alors par une optimisation poussée de chaque étape de l'algorithme.

6 Conclusion

Dans cet article, nous avons montré l'intérêt d'étendre le concept de hiérarchie contextuelle dans les entrepôts de données avec la logique foue. Il s'agit de pouvoir intégrer des connaissances utilisateurs avec davantage d'expressivité et de flexibilité au sein des entrepôts de données, en vue d'améliorer le processus d'analyse.

Cette étude sur l'intégration de la logique floue dans le cadre des hiérarchies contextuelles constitue un début prometteur. Le développement du système correspondant avec une interface permettra de tester l'approche auprès d'utilisateurs décideurs. Pour parvenir à ce résultat, différentes perspectives à moyen terme devront être étudiées, dont un travail poussé sur l'intégration

des connaissances floues et leur exploitation pour la généralisation, avec le souci permanent de la performance compte-tenu du contexte d'analyse en ligne sur des données volumineuses. La richesse d'expressivité a mis en avant dans la partie discussion le rôle que pouvait avoir l'utilisateur décideur dans le choix de la fonction de scoring pour découvrir la meilleure généralisation possible. Ainsi, un travail sur l'exploitation personnalisée de l'entrepôt dans le cadre du recours à des hiérarchies contextuelles floues pourrait être intéressant.

Références

Dubois, D. (2011). The role of fuzzy sets in decision sciences : Old techniques and new directions. *Fuzzy Sets and Systems 184*(1), 3–28.

Fasel, D. et K. Shahzad (2010). A data warehouse model for integrating fuzzy concepts in meta table structures. In *ECBS'10*, pp. 100–109.

Inmon, W. H. (1996). *Building the Data Warehouse, 2nd Edition* (2 ed.). Wiley.

Laurent, A. (2002). *Bases de données multidimensionnelles floues et leur utilisation pour la fouille de données*. Ph. D. thesis, Université Paris 6.

Ligęza, A. (2006). *Logical foundations for rule-based systems*, Volume 11. Springer-Verlag New York Inc.

Malinowski, E. et E. Zimányi (2004). OLAP Hierarchies : A Conceptual Perspective. In *CAiSE'04*, Volume 3084 of *LNCS*, pp. 477–491. Springer.

Perez, D., M. J. Somodevilla, et I. H. Pineda (2007). *Fuzzy Spatial Data Warehouse : A Multidimensional Model*, pp. 3–9. IEEE Computer Society.

Pitarch, Y. (2011). *Résumé de Flots de Données : Motifs, Cubes et Hiérarchies*. Ph. D. thesis, Université Montpellier 2.

Pitarch, Y., C. Favre, A. Laurent, et P. Poncelet (2010a). Analyse flexible dans les entrepôts de données : quand les contextes s'en mêlent. In *EDA'10*, Volume B-6 of *RNTI*, pp. 191–205. Cépaduès.

Pitarch, Y., C. Favre, A. Laurent, et P. Poncelet (2010b). Context-aware generalization for cube measures. In *DOLAP'10*, pp. 99–104.

Zadeh, L. (1965). Fuzzy sets*. *Information and control 8*(3), 338–353.

Summary

Data warehouses are nowadays extensively used to perform analysis on huge volume of heterogeneous data. This success is partly due to the capacity of considering data at several granularity levels thanks to the use of hierarchies. Nevertheless, in previous work, we showed that very few amount of knowledge expert was consider in the generalization process leading to inaccurate analysis. To overcome this drawback, we introduced a new category of hierarchies, namely the contextual hierarchies. Unfortunately, in contrast to the complexity of expert knowledge that should be considered, the knowledge definition process was too rigid. In this paper, we extend these hierarchies and their related techniques to drastically increase their flexibility and expressivity. To this purpose, we present a preliminary work that consists in a fuzzy-based approach.