Rapport de stage
Utilisation des cardinalités dans les résumés linguistiques de données

Projet Badins,
Université de Rennes 1
ENSSAT
Marcq François

Tuteurs : Daniel Rocacher
Ludovic Liétard

Résumé

Le travail de ce stage consiste en l'élaboration d'un modèle de résumés de données faisant intervenir des étiquettes linguistiques définies par des ensembles flous. Ce travail entre dans le cadre du stage de fin d'étude de la formation ENSSAT et du master de recherche de l'école doctorale Matisse de l'université de Rennes 1. Le modèle élaboré permet de déterminer complètement les données résumées, et une attention particulière est portée au calcul de la cardinalité des ensembles résumés regroupant des données. De plus, le modèle a été élaboré dans l'objectif de changer la granularité des données afin de gagner en précision ou de généraliser les informations selon les besoins.

Mots-clés : Résumés de données, ensembles flous, étiquettes linguistiques, cardinalités floues, changement de granularité.

Abstract

The work for this training course reposes on the build of a model for data summarization with linguistic labels define by fuzzy sets. This work comes as the end-studies training course for the ENSSAT's training and for the research master degree. The elaborate model allowed to completely determine the summarized data, and a particular attention is set on the computation of the cardinality of summarized sets gathering data. Moreover, the model was elaborated with an objective that is the possibility of changing the granularity of data in order to have best precision or to generalize the information according to needs.

Keywords: Data summarizations, fuzzy sets, linguistic labels, fuzzy cardinalities, granularity changing.

Table des matières

Introduction

Dans le cadre de la formation d'ingénieur ENSSAT (Ecole Nationale Supérieure des Sciences Appliquées et de Technologie) et du double diplôme de master recherche en informatique préparé conjointement avec l'université de Rennes 1, le stage de fin d'étude est réalisé au sein du Laboratoire Lannionnais d'Informatique (LLI). Ce stage s'intègre dans les activités de recherche menées par le projet Badins. Les principaux domaines de recherche de ce projet sont : l'interrogation flexible et la prise en compte de données mal connues dans le cadre des bases de données relationnelles.

L'étude se place dans le cadre de l'accès à des grandes masses de données où se pose le problème de l'accès à l'information dans un temps acceptable. Dans certains cas, il est possible d'atténuer l'influence de la quantité de données sur les temps de réponse des traitements et des algorithmes. Les traitements concernés peuvent se satisfaire d'approximations, généralement de moindre qualité que des résultats non approximatifs, mais obtenues en des temps plus acceptables. C'est dans ce domaine que se place les résumés de données.

Un autre problème auquel sont confrontés les traitements informatiques est lié à la difficulté de modéliser des représentations « humaines » ou « naturelles » qui font appel à des concepts souvent graduels. Par exemple, ce principe de nuancement est présent quand on considère la couleur grise. Le passage de la couleur noir à la couleur blanche se fait en passant par tout un ensemble de nuance de gris. Cette gradualité ne peut pas être reflétée par la dichotomie des systèmes binaires. Les nuances et autres graduations sont ainsi habituellement occultées, donnant lieu à un effet de seuil. L'objectif est de pouvoir considérer qu'un élément ne passe pas d'un état à l'autre de manière brutale, il peut se situer dans un état transitoire entre la non-appartenance et l'appartenance. Cette notion de graduation est intéressante à mettre en place dans la notion de résumé de données.

Une telle notion de gradualité est prise en compte par la théorie des ensembles flous. Elle est utilisée par toutes les méthodes de résumés qui ont été analysées lors de l'étude bibliographique. Elle permet de décrire les données en termes « linguistiques » car issus du langage naturel. Les résumés produits sont alors appelés des « résumés linguistiques ». L'étude bibliographique qui a débuté ce stage a permis de mettre en évidence l'intérêt de la notion de cardinalité dans les système de résumés [Kac99, RY97]. Nous avons également montré que le modèle SaintEtiQ, développé dans le cadre du LINA (Laboratoire d'Informatique de l'université de Nantes Atlantique), basé sur des hiérarchies de résumés, ne tirait pas partie de gradualité introduite par les étiquettes linguistiques utilisé et que ce modèle ne tenait pas compte d'aspect quantitatifs sur les données regroupés. Notre objectif dans cette étude est donc d'imaginer, à partir du modèle *SaintEtiQ*, un nouveau modèle permettant de prendre en compte à la fois les notions de gradualité et de cardinalité. Pour ce faire, nous nous appuyons sur le concept de nombre graduel qui correspond à la cardinalité d'un ensemble flou [Roc05].

C'est donc autour de ce sujet que s'organise le travail réalisé au cours de ce stage, dont ce rapport est une synthèse. Dans un premier temps une présentation des éléments fondateurs du travail est faite avec une description du modèle *SaintEtiQ* ainsi qu'une

définition des nombres graduels. Puis dans une deuxième partie, partant du modèle *SaintEtiQ*, un nouveau modèle de résumés est étudié afin de prendre en considération gradualité et quantité pour synthétiser des données. Ensuite, dans une troisième partie les évolutions qui peuvent être appliquées au nouveau modèle proposé sont analysées. Dans une quatrième partie les premiers éléments liés à la réalisation technique de ce modèle sont abordés. En particulier la structure de données que nous allons utiliser pour tester le modèle est décrite. Dans une cinquième partie nous analysons les problèmes de ce modèle et nous envisageons de nouvelles perspectives d'étude.

1 Eléments de base de l'étude

Le modèle SaintEtiQ est présenté dans ce chapitre. Les informations contenues dans le modèle ainsi que les questions auxquelles il peut répondre sont aussi évoquées. Dans une seconde partie sont présentés les nombres graduels qui vont servir de base à l'utilisation de cardinalité dans le modèle qui sera proposé dans les chapitres suivants.

1.1 Le modèle SaintEtiQ

Le modèle SaintEtiQ est une méthode de construction de résumés de données composées de couples attribut/valeur. Il se distingue par deux particularités : le souci d'intelligibilité des résumés construits du fait de l'utilisation d'étiquettes linguistiques, et la génération de résumés décrivant les données à des niveaux d'abstraction différents.

1.1.1 Construction du résumé

Afin de réaliser le processus de résumé, deux types d'entrées sont nécessaires : les données à traiter et les données désignées par « connaissance de domaine », donnant une indication sur la manière de les résumer. Les données à résumer sont considérées au sens des bases de données relationnelles. A ce titre, elles sont organisées en enregistrements ou tuples qui suivent le schéma d'une relation définie sur un ensemble d'attributs $A = A_1, A_2, ..., A_n$. Chaque attribut A_i est défini sur un domaine D_{Ai}, qui peut être numérique ou symbolique. Ainsi, un tuple *t* est formé d'une suite de valeurs suivant l'ordre prédéfini des attributs A_i. Il est noté :

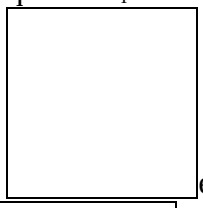
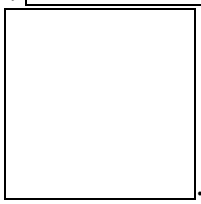
$$<t.A_1 , t. A_2 , ..., t.A_n >.$$

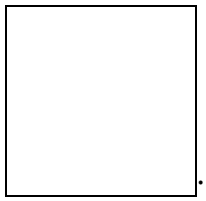
Une autre contrainte sur les données est leur complétude : toutes les valeurs d'attributs doivent être présentes. Pour un tuple *t* d'une relation R, la valeur de $t.A_i$ est nécessairement connue, élémentaire, précise et certaine. Les données incomplètes, incertaines ou mal connues ne sont donc pas traitées par le modèle.

Les *connaissances de domaine* régissent l'interprétation qui sera faite des valeurs d'attributs dans la constitution des résumés. Elles sont constituées essentiellement de variables linguistiques définies sur les domaines d'attributs de la relation résumée. Elles sont données par l'utilisateur ou un expert, ceci afin de définir un langage de description des données dont la sémantique est la plus proche de l'utilisateur. Ainsi, les connaissances de domaine fournissent le vocabulaire d'expression des résumés.

Les variables linguistiques, dans le sens introduit par Zadeh en 1975, permettent de décrire les valeurs d'un domaine d'attribut grâce à des caractérisations floues. Si A est un attribut et D_A son domaine, on écrit habituellement « t.A = x » avec x une valeur de domaine D_A. En utilisant une variable linguistique, la valeur de t.A du tuple t sur l'attribut A n'est plus une valeur spécifique x : on écrit « t.A = d » ou « t.A est d » avec d un descripteur linguistique issu de la variable linguistique de l'attribut A.

Formellement, une variable linguistique est représentée par un triplet (A, D_A, D) avec D_A le domaine de l'attribut A et D = { d_i, 1 ? i ? n}, un ensemble de sous-ensembles flous de D_A. Chaque terme d_i est muni d'une fonction d'appartenance f_{di} définie sur D_A et à valeurs dans l'intervalle [0, 1]. Un sous-ensemble flou d_i peut également être une partie ordinaire du domaine D_A, c'est-à-dire un ensemble classique (sa fonction d'appartenance n'est plus graduelle mais binaire). On peut dire que D définit une partition de D_A lorsque les d_i vérifient les propriétés :

- de couverture :  et
- de contraste :  .

La partition D est dite « forte » si on a en plus  .

Sur la figure 1.1, extraite de [VRUM06], l'attribut épaisseur instancie A sur le domaine D_A = [0,15~mm, 50~mm], avec l'ensemble des descripteurs D = {fin, mince, moyen, épais, grand}. Les éléments de D sont appelés des termes, des descripteurs ou des étiquettes linguistiques.

FIG. 1.1 – Partition de l'épaisseur

D'une manière générale, les connaissances des domaines permettent de faire la correspondance entre les valeurs de domaines d'attribut et le vocabulaire d'expression des résumés de données. Un domaine D_A est alors réécrit par l'ensemble, noté D_A^+, des termes qui lui sont applicables.

Le processus de résumé est incrémental : chaque tuple est pris en compte l'un après l'autre. Il est donc possible de considérer uniquement le traitement d'un seul et unique tuple. L'architecture proposée dans [Ras01] distingue deux étapes pour la génération d'un résumé : la phase de réécriture et celle de résumé.

Le service de réécriture réalise une abstraction des données grâce aux connaissances de domaine. On obtient une représentation des données sous forme de sous-ensembles flous, qui est homogène quelle que soit la nature du domaine d'attribut. Le principe de l'abstraction est de négliger les différences non significatives entre les valeurs d'attributs. Ainsi les valeurs 45 mm et 48mm, désignant l'épaisseur d'une plaque de métal, qui, bien que distinctes, seront décrites par la même étiquette « grand » (voir figure 1.1).

Il s'agit de trouver toutes les réécritures possibles du tuple t, d'après les connaissances de domaine.

L'opération de réécriture d'un tuple peut conduire ce dernier à être réécrit plusieurs fois avec des descripteurs différents pour un même attribut. C'est ce que l'on peut voir avec figure 1.1. Lorsque l'on considère un tuple pour lequel l'attribut épaisseur prend la valeur 8, ce dernier sera donc réécrit avec le descripteur « mince » mais aussi avec « moyen ». Ainsi on génère plusieurs candidats pour le résumé ; on note φ(t) l'ensemble des tuples candidats engendrés par la réécriture de t.

La phase de résumé réalise quant-à elle une classification des données en résumés et la structuration des résumés en hiérarchie. Une relation d'ordre partiel est à la base de l'organisation en arbre. Il s'agit lors de cette étape de prendre en compte les tuples candidats obtenus suite à la réécriture, et de les incorporer dans l'arbre de hiérarchie. Un tuple candidat *ct* suit un chemin de la racine z_0 pour atteindre un résumé z_f dont la description est identique à celle de *ct*. Cependant, le parcours de ct dans l'arbre est découvert au fil de sa progression grâce à des opérateurs d'apprentissage. Le rôle de ces opérateurs est de modeler la hiérarchie de résumés en fonction des données déjà incorporées (de la forme courante de la hiérarchie).

A chaque n?ud z traversé par *ct*, le système décide de l'opérateur qui fournit le meilleur résultat en fonction des critères d'évaluation définis par l'application. Par cette opération, le prochain n?ud sur le chemin menant à la feuille z_f est déterminé : c'est celui sur lequel la dispersion des données (après l'incorporation de *ct*) sera minimale. De même, une mesure de spécificité détermine si z est suffisamment spécifique pour prétendre au statut de feuille. Si tel est le cas, l'incorporation de *ct* est terminée. Autrement, l'opération recommence avec le nouveau n?ud.

Pour réaliser l'ensemble de ces opérations, G. Raschia propose dans sa thèse [Ras01] l'utilisation des quatre opérateurs suivant :

- initialiser : crée un nouveau résumé z_* ne contenant que *ct*.
- affecter : permet de matérialiser le passage de *ct* dans un résumé z.
- fusionner : crée un parent commun à deux résumés z_1 et z_2 lorsque ces derniers sont jugés trop spécifiques.
- scinder : inverse de fusionner, supprime un résumé z jugé trop général.

1.1.2 Exemple de construction de résumé

Le tableau 1.1 présente les données qui vont servir à présenter le fonctionnement de *SaintEtiQ*. Cet exemple est extrait de [VRUM06] et n'est qu'une partie d'une base de données relationnelle. Chaque tuple de la base se réécrit selon les étiquettes linguistiques qui sont définies comme le montre la figure 1.2. Avec ces étiquettes linguistiques, il est possible de réécrire les tuples. On obtient les réécritures présentent dans le tableau 1.2. On peut constater que certains tuples ont été réécrits plusieurs fois, c'est le cas de CuSn12 qui est réécrit deux fois.

matériaux	épaisseur	dureté	température
UZ40	10	38	900
CuSn12	8	40	850
CuAsO5	12	44	896
Fe	10	35	1530
Ni	5	35	1453

TAB. 1.1 – Extrait de la table des matériaux

FIG. 1.2 – Partition des domaines

matériaux	épaisseur	µ	dureté	µ	température	µ	ref :
T1	moyenne	0,7	tendre	1,0	modérée	0,85	UZ40
T2	moyenne	0,35	tendre	0,9	modérée	1,0	CuSn12
T3	mince	0,35	tendre	0,9	modérée	1,0	CuSn12
T4	moyenne	1,0	tendre	0,4	modérée	0,9	CuAsO5
T5	moyenne	1,0	dure	0,4	modérée	0,9	CuAsO5
T6	moyenne	0,7	tendre	1,0	normale	0,85	Fe
T7	mince	1,0	tendre	1,0	normale	0,96	Ni

TAB. 1.2 – Extrait de la table des matériaux après réécriture

Ainsi, avec ces informations, il est possible de regrouper les données. L'application des méthodes présentées précédemment permet de réaliser les regroupements suivants : T1, T2 et T4 vont former un même groupe, noté Z3, qui portera les information : {1,0/moyenne, 1,0/tendre, 1,0/modérée}. Le tuple T6 est ensuite regroupé avec ce résumé afin de former un nouveau résumé qui est noté Z1 qui porte les étiquettes : {1,0/moyenne, 1,0/tendre, 1,0/modérée + 0,85/normale}. D'un autre côté les tuples T3, T5, T7, vont quant à eux être regroupés en un résumé noté Z2 dont les informations sont : {1,0/mince + 1,0/moyenne, 1,0/tendre + 0,4/dure, 1,0/modérée + 0,96/normale}. Les deux résumés Z1 et Z2 se regroupent en un résumé général Z0 sur les données. Les informations contenues dans ce résumé sont {1,0/mince + 1,0/moyenne, 1,0/tendre + 0,4/dure, 1,0/modérée + 1,0/normale}. La figure 1.3 montre comment les données s'organise sur cet exemple pour former un arbre hiérarchique de résumé.

FIG. 1.3 – Arbre hiérarchique des résumés

Toutefois, il est à noter qu'avec ce modèle et la méthode employée pour construire le résumé, il n'est pas possible de faire intervenir la mesure selon laquelle un attribut s'exprime en fonction des connaissances du domaine. Du fait de la perte d'information relative au degré d'appartenance d'une valeur d'attribut à une étiquette linguistique, il devient impossible de poser certaines questions à la hiérarchie de résumé. Par exemple, il est impossible d'interroger la hiérarchie pour savoir combien il y a de tuples présents dans un résumé avec un degré supérieur à un seuil α. Nous verrons par la suite que les nombres graduels permettent de répondre à ce genre de question.

1.1.3 Interrogation d'un résumé

Ce modèle permet de répondre à plusieurs questions simples. Dans SaintEtiQ, les questions portent sur l'ensemble des attributs n'intervenant pas dans les critères de sélection. Ainsi si on note X l'ensemble des critères de sélection et Y le complémentaire de X sur l'ensemble de la relation, on obtient la formulation d'une question de la manière suivante :

Comment sont sur Y les individus qui sont x sur X ?

Avec l'exemple précédent, si on fait une recherche avec comme critère la dureté, les informations que l'on va obtenir porteront sur la température et l'épaisseur. Ainsi si la recherche des caractéristiques des matériaux durs, se traduit par la question : *comment sont la température de fusion et l'épaisseur des matériaux durs ?* Il s'agit en fait de décrire les matériaux par leur température de fusion et leur épaisseur en fonction du critère de dureté qui a été fixé.

Le langage d'interrogation utilisé est un langage de requête proche de SQL. Dans ce langage une requête commence par le mot clef « DESCRIBE ». La forme de la requête est pour le reste très proche d'une sélection traditionnelle en SQL et une requête prend la structure est la suivante :

```
DESCRIBE [<table>] ON <liste_d_attribut> WHERE <condition>.
```

Si l'on reprend la question précédente on obtient : DESCRIBE ON température, épaisseur WHERE dureté IN (dure)

Le fait de préciser les attributs dont les caractéristiques sont recherchés permet de rendre des réponses plus concises.

Pour évaluer une requête, le système parcourt l'arbre de la hiérarchie de résumés à partir de sa racine. A chaque n?ud de la hiérarchie un processus est lancé afin de déterminer s'il faut continuer l'exploration des sous-arbres ou arrêter l'exploration. Voici les différents cas de figure qui peuvent être rencontré lors de l'exploration :

- **Cas 1, arrêt de l'exploration par défaut de réponse :** c'est le cas où il existe une clause (ou plus) pour laquelle le résumé ne satisfait pas le critère de sélection. En reprenant l'exemple précédent on retrouve ce phénomène avec le résumé Z1 et la question comment sont les matériaux d'épaisseur fine ou mince car Z1.épaisseur = {1.0/moyenne}.
- **Cas 2, arrêt de l'exploration par satisfaction de la requête :** pour chaque attribut présent dans la requête, le résumé étudié présente uniquement les mêmes critères. Pour chaque attribut l'ensemble des étiquettes présentes dans le résumé est inclus dans l'ensemble des étiquettes de la requête.
- **Cas 3, poursuite de l'exploration :** si pour au moins un des attributs de la requête, il existe des étiquettes dans le résumé qui soit en plus de celle de la requête alors il est possible qu'il y ait des éléments à exclure de ce résumé pour répondre à la question.

Les requêtes qui sont réalisables avec SaintEtiQ sont très spécifiques, simples et limitées. En particulier, il est à noter que ce modèle ne permet pas l'interrogation sur des

quantités d'éléments présents dans un résumésur une étiquette linguistique. Il ne permet pas la résolution de question tel que « *Combien de matériaux sont d'épaisseur fine et de fureté tendre ? »* ou « *Est-ce que la plupart des matériaux durs ont une température de fusion haute? »*. Ces deux questions font intervenir des cardinalité et des proportions ce qui n'est pas réalisable avec SaintEtiQ. Pour prendre en compte les cardinalités plusieurs solutions sont envisageables. Pour le modèle que nous proposerons par la suite nous avons choisi d'utiliser les nombres graduels.

○ *1.2 Les nombres graduels*

Pour représenter des cardinalités d'ensembles flous, plusieurs pistes ont été explorées. La première qui a été étudiée est le sigma-count [RY97]. Le principe repose sur le calcul de la somme des degrés d'appartenance de chaque élément à l'ensemble. Toutefois, cette manière de représenter les cardinalités pose des problèmes puisqu'il n'est pas possible de différencier le fait que peu d'éléments appartiennent fortement à un ensemble et le fait que beaucoup n'appartiennent que faiblement à un ensemble. Par exemple, la cardinalité pour un élément présent au degré 1 est la même que pour dix éléments, tous présents dans un même ensemble, au degré 0,1.

Pour palier à ce problème, d'autres représentations de la cardinalité ont été proposées. Nous allons nous intéresser ici à la méthode proposé par D. Rocacher dans [Roc05]. Elle consiste à représenter la cardinalité d'un ensemble flou par un nombre graduel qui est basé sur la définition de la cardinalité floue définie par :

$$\forall n \in \mathbb{N}, \mu_{card(E)}(n) = \sup \{ \alpha / card(E_\alpha) \geq n \}$$

avec α un degré compris entre 0 et 1 (coupe) et E un ensemble flou. Cela permet d'écrire, pour un ensemble flou A = {1/x1, 1/x2, 0.8/x3, 0.7/x4, 0.7/x5, 0.5/x6}, que card(A) = {1/0, 1/1, 1/2, 0.8/3, 0.7/4, 0.7/5, 0.5/6}. On est sur que A contient totalement deux éléments. 0,8 représente dans quelle mesure A contient 3 élémenets. Cette représentation de la cardinalité permet de connaître pour chaque coupure de niveau (α-coupe) le nombre d'éléments de l'ensemble A qui ont un degré supérieur à α.

Définition : une α-coupe ou coupe de niveau α l'ensemble des éléments d'un ensemble flou pour lesquels la valeur du degré d'appartenance est supérieure ou égale à α. α est donc défini sur l'intervalle [0,1].

Les nombres graduels peuvent être représentés de manière schématique par un graphe représentant les appartenances des éléments à l'ensemble en considérant leur degré. Cette représentation permet de visualiser les données en faisant ressortir les points les plus importants de l'ensemble.

FIG. 1.4 - Représentation d'un nombre graduel

Les nombres graduels sont définis par extension de l'ensemble des entiers naturels. Cet ensemble est noté N_f afin de rappeler les correspondances avec l'ensemble N. Afin de pouvoir effectuer des calculs avec ces nombres, il est nécessaire de définir

des opérateurs tels que l'addition ou la multiplication.

Pour chaque α-coupe, le comportement du nombre graduel est le même que celui d'un entier naturel. Ainsi pour réaliser l'addition de deux nombres graduels il suffit d'effectuer des additions pour chaque α-coupe. Si on considère les nombres A = {1/3, 0.7/4, 0.4/6} et B = {1/0, 0.9/1, 0.7/3, 0.5/4}, on commence par effectuer le calcul pour la coupe de niveau 1 on obtient 3 provenant uniquement de A, puis on considère la coupe de niveau 0.9. Pour le degré 0.9 on a trois éléments provenant de A (puisqu'ils sont présents à un degré supérieur) et un élément provenant de B ce qui nous fait un total de 4 pour la coupe de degré 0.9. Au final une fois le calcul complètement effectué, on obtient le résultat : {1/3, 0.9/4, 0.7/7, 0.5/8, 0.4/10}.

Une telle approche a été étendu au nombre relatif Z_f et au rationnel Q_f et leurs opérateurs (addition, soustraction, division, multiplication) ont été définis. De tels ensembles sont utilisés pour définir des quantificateurs absolus (*au moins 3*, *environ 5*), relatifs (*la plupart*, *environ la moitié*, *presque tous*, *peu de*) et pour comparer des quantités.

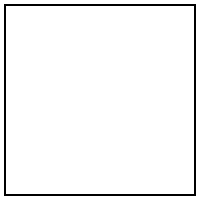
● 2 Vers une représentation de résumés avec cardinalités

Nous avons vu dans la partie précédente que le modèle SaintEtiQ ne permettait pas de répondre à certaines questions qui paraissent pourtant intéressantes à résoudre. Ce chapitre vise à définir un nouveau modèle permettant de prendre en considération des notions de quantité et de gradualité. Ainsi, nous allons chercher à développer une structure capable de résumer des données et qui puisse permettre l'interrogation facile de cette dernière. La structure doit être capable de répondre à des questions portant sur le nombre de tuples présents dans un résumé ou sur les proportions avec lesquels les tuples sont présent dans ce résumé. Pour mieux répondre aux objectifs que nous nous sommes fixés, il est indispensable que nous fixions les limites de contraintes et d'hypothèses que nous devons remplir. Puis nous verrons quelle est l'organisation des données que nous envisageons de réaliser.

○ *2.1 Hypothèse de travail*

Il est tout d'abord nécessaire de définir des règles à appliquer afin de résoudre le problème de recherche du nombre d'éléments présents dans un résumé et de savoir dans quelle mesure ils sont représentatifs du résumé. Il nous a donc fallu poser des hypothèses de travail qui soient raisonnables et pas trop générales afin de pouvoir valider les résultats théoriques que nous obtenons.

Pour commencer, nous avons considéré une hypothèse qui est reprise généralement dans les différents modèles de résumé de données. Cette hypothèse consiste à admettre que les seules partitions floues acceptables sont celles pour lesquelles la coupure entre deux sous-espaces se fait pour un degré d'appartenance de 1/2 sur chacun des deux sous-espaces avec une réécriture sur au maximum deux étiquettes linguistiques. Cette hypothèse a été utilisée par exemple dans [BDPP00] et peut se réduire à l'expression d'une formule :



Dans cette formule μ_i désigne la fonction d'appartenance d'un élément au ième étiquette, qui est une fonction à valeur dans [0, 1]. Cette hypothèse permet de pouvoir calculer le degré sur un sous-espace voisin quand on sait qu'il y appartient partiellement.

L'exemple de la figure 2.1 permet d'illustrer cette contrainte. Les ensembles qui sont voisins (A et B, B et C, C et D, D et E) ont des intersections non nulles. Mais les deux ensembles formant les extrémités de l'espace d'origine, ici A et E, ne peuvent pas avoir de parties en commun.

○ *2.2 Organisation des données*

L'analyse des propositions qui ont été faites par d'autres équipes de recherche sur des données floues [Kac99, RY97, BDPP00] a permis de constater qu'aucune de celles qui ont été présentées ne nous permettait de prendre en compte des

FIG. 2.1 – Division de l'espace en sous-ensembles

requêtes quantifiées. Nous avons donc été amené à proposer une nouvelle structure.

● 2.2.1 Principe de la construction de résumés sur une dimension

Une structure arborescente, dont chaque feuille est caractérisée par une étiquette linguistique, est envisagée pour regrouper les données. Les ensembles de la partition floue ayant une intersection non vide sont toujours deux ensembles contigus. En conséquent, un tuple se réécrit, au plus deux fois et avec des étiquettes consécutives. Les seuls regroupements d'étiquettes intéressants sont donc ceux de deux étiquettes consécutives. C'est pourquoi, il peut être utile de réaliser une arborescence des unions de sous-ensembles voisins. Avec le regroupement en union et l'enregistrement des cardinalités floues dans chacun des n?uds, il est possible de retrouver les caractéristiques de l'intersection de deux sous-ensembles. La figure 2.2 présente le résultat de la construction de cet arbre d'unions pour un ensemble à cinq sous-ensembles.

FIG. 2.2 - Regroupement de cinq sous-ensembles

Les lettres U représentent A□B□C, V : B□C□D et ainsi de suite, et par conséquent
Z = A□B□C□D□E.

Cette représentation en arbre n'est valable que si on ne considère qu'un seul attribut lors de la phase de résumé. Il faut ajouter une dimension pour chaque attribut supplémentaire que l'on veut prendre en compte. Par exemple, si l'on prend en compte les données concernant la taille et le poids de personnes d'une base de données, on obtient un système à trois dimensions. Ce point est présenté au paragraphe 2.2.3.

Dans la structure à une dimension représentée par la figure2.2, le degré d'appartenance sur un attribut x d'un tuple t au sous-ensemble flou définit par une étiquette est placé dans le n?ud N. Ceci permet d'avoir pour chaque n?ud l'ensemble des tuples avec leur degré sur l'attribut x pour l'étiquette linguistique associé au n?ud. Cette structure peut donc être vue comme un index puisqu'il est possible de retrouver l'ensemble des enregistrements présents dans la base de données relationnelle. Il est à noter que pour chaque niveau, les degrés de tous les tuples sont présents et pas uniquement le meilleur comme cela est le cas dans SaintEtiQ. On peut aussi savoir combien de tuples sont présents dans un résumé. Il suffit de déterminer la cardinalité floue de l'ensemble flou des tuples associé à un n?ud. Une telle cardinalité peut se représenter par un nombre graduel.

Les données présentent dans la structure permettent de réaliser des interrogations qui n'étaient pas possible avec le modèle SaintEtiQ. Avec ce modèle, on peut ainsi poser facilement des questions du type : « Combien de personnes sont de taille grande ou moyenne à un degré supérieur à 0,7 ? ». C'est avec ce genre de question que le modèle prend son intérêt. En effet, peut de modèle offre la possibilité de pouvoir sélectionner les tuples vérifiant un critère à un certain niveau de satisfaction.

L'analyse de la structure ci-dessus a permis de mettre en évidence que, avec l'utilisation de cardinalité graduelle, de nombreuses informations n'ont pas besoin d'être pré-calculées mais peuvent être calculées de manière dynamique. En effet, les cardinalités graduelles permettent de calculer la cardinalité d'une union en fonction des ensembles et de leur intersection. On obtient donc la formule :

$$card(A \cup B) = card(A) + card(B) - card(A \cap B)$$

ou A et B sont des sous-ensembles floues. La démonstration se fait par α-coupe et s'appuie sur d'une part une propriété des α-coupe : $(A \cup B)_\alpha = A_\alpha \cup B_\alpha$ et d'une propriété sur les cardinalités d'ensemble ordinaire. Il en découle :

$$card((A \cup B)_\alpha) = card(A_\alpha \cup B_\alpha)$$
$$= card(A_\alpha) + card(B_\alpha) - card(A_\alpha \cap B_\alpha)$$
$$= card(A_\alpha) + card(B_\alpha) - card((A \cap B)_\alpha).$$

L'utilisation d'autre cardinalité floue ne permet pas de vérifier cette propriété.

Ce constat permet donc de limiter les calculs, lors de la construction de l'arbre, aux lignes composées des ensembles que l'on définit à partir des étiquettes linguistiques issues du domaine de connaissance et composées des unions deux à deux des ensembles adjacents. Ainsi, le schéma de la figure 2.2, peut être limiter au calcul des ensembles : A, B, C, D, E, A∪B, B∪C, C∪D, D∪E, puisque les autres éléments peuvent être calculés à partir de ces informations.

Il faut prendre en compte une question portant sur au moins trois étiquettes voisines pour utiliser la propriété précédente, puisque dans ce cas les données sont calculées dynamiquement et non pas lors de l'ajout d'un tuple. Le meilleur exemple de question pour cette structure est : « Combien y a-t-il de personne de taille petite ou moyenne ou grande au degré au moins 0,6 ? ». En effet avec cette question on interroge sur trois étiquettes, on ne retrouve donc pas de n?ud regroupant l'ensemble de ces informations. Par contre on connait les cardinalités sur les ensembles caractérisé par les étiquettes *petit* ∪ *moyen*, *moyen* ∪ *grand* et *moyen*, le calcul de l'union est donc réalisable avec la formule précédente. Ainsi on recalcule l'ensemble petit ou moyen ou grand à partir des données de la structure et on ne conserve que celle dont le degré est supérieur à 0,6.

● 2.2.2 Exemple de représentation avec un attribut

Afin de préciser le propos, nous allons considérer l'exemple portant sur une population dont on ne s'intéresse qu'à la taille. On défini pour caractériser cette population quatre étiquettes linguistiques : *très petit*, *petit*, *moyen* et *grand*. On

considère la figure 2.3 comme la représentation de la répartition des étiquettes.

FIG. 2.3 - Répartition de la taille

Soit le tableau 2.1 qui représente les différentes personnes présentent dans la base ainsi que leur taille. A partir de ce tableau on peut en déduire les réécritures qui sont faites. Ces réécritures sont présenté dans le tableau 2.2.

Avec les données réécrites, on va pouvoir construire la structure de données en plaçant tout d'abord les tuples qui ont été réécrits dans les feuilles. Les tuples t1_a, t5_b, t7 et t8_a se retrouvent donc réunis sous l'étiquette *petit*. La figure 2.4 représente l'arbre qui a été construit par le système suite à l'organisation des données.

nom	taille (m)
Paul	1,67
Luc	1,36
Marc	1,74
Georges	1,87
Léon	1,44
Robert	1,81
Alexandre	1,55
David	1,69

TAB. 2.1 - Liste de personne avec leur taille

index	étiquette	degré	origine
t1_a	petit	0,6	Paul
t1_b	moyen	0,4	Paul
t2	très petit	1,0	Luc
t3	moyen	1,0	Marc
t4	grand	1,0	Georges
t5_a	très petit	0,2	Léon
t5_b	petit	0,8	Léon
t6_a	moyen	0,8	Robert
t6_b	grand	0,2	Robert
t7	petit	1,0	Alexandre
t8_a	petit	0,2	David
t8_b	moyen	0,8	David

TAB. 2.2 - Réécriture de la taille en fonction des étiquettes

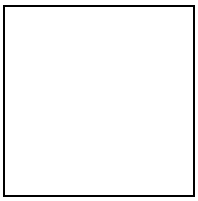


FIG. 2.4 - Arbre d'organisation des tuples réécrits

Il est à noter que l'ensemble des informations qui sont présentes dans la racine, noté 1, qui contient l'étiquette *très petit* *petit* *moyen* *grand* peut être calculé à partir des informations présentes dans les n?uds numérotés 2, 3 et 5. Il en est de même pour les n?uds 2 et 3 qui peuvent être calculés à partir des informations présentent dans les niveaux inférieurs. Cet exemple montre bien qu'il est possible de ne calculer les informations uniquement pour les feuilles et les unions de deux feuilles.

Il est possible de répondre à la question « combien de personne sont de taille très petit ou petit ou moyen ? ». Pour cela on va rechercher le n?ud contenant ces trois étiquettes et extraire le nombre de tuples qui est présent dans ce n?ud. Il faut calculer la cardinalité de ce n?ud puisqu'il n'est pas calculé lors de l'ajout des données. Sa cardinalité ce calcul à partir de celle de *très petit* *petit*, *petit* *moyen* et *petit* et on obtient que card(*très petit* *petit* *moyen*) = {1,0/3, 0,8/6, 0,6/7}. La réponse à la question est donc 3 au degré 1, 6 au degré au moins 0,8 et 7 au degré au moins 0,6.

● 2.2.3 Principe de construction de résumé de plusieurs dimensions

Au paragraphe 2.2.1 on a montré qu'on pouvait se limiter à la construction d'une structure sur deux niveaux. Grâce à cette limitation à deux niveaux des données pré-calculées lors de la construction du modèle, il est possible de donner une représentation de ce modèle pour le cas ou l'on aurait deux attributs pris en compte. Afin d'expliquer le graph, on réalise une construction progressive de ce dernier. La figure 2.5 ne considère tout d'abord que les intersections entre les étiquettes qui ont été définies dans la base de connaissance. L'ensemble repéré par le chiffre 1 correspond à l'ensemble des tuples qui s'écrivent avec les étiquettes « *petit* » pour la taille et « *moins jeune* » pour l'âge. Il permet de répondre à une question comme « combien de personnes sont petit et moins jeune ? ». La figure 2.6 présente le résultat des intersections d'une étiquette sur un attribut avec des unions de deux étiquettes sur l'autre attribut. Le n?ud 2 est carractérisé par les étiquettes « *petit* ou *très petit* » pour la taille et « *jeune* » pour l'âge. La figure 2.7 présente l'ensemble des intersections entre les unions de deux étiquettes sur chaque attribut. Le n?ud 3 représente l'ensemble des tuples qui s'ecrivent avec « *jeune* ou *très jeune* » pour l'âge et « *petit* ou *moyen* » pour la taille.

Il est possible de revenir à la vision en deux dimensions en réalisant une projection des données avec un attribut fixé. En effet, la version sur un seul attribut peut être vue comme le résultat d'une interrogation dans laquelle on précise que l'on sait la valeur du second attribut. On peut, par exemple, considérer que la figure 2.8 est la

représentation de la question : « quel est l'âge des personnes *sachant qu'elles sont petites* ? » de la figure2.7. On a réalisé ici une coupe le long de la ligne vérifiant : taille est petite. Les tuples sont enregistrés dans chacun des n?uds présents dans la structure en graphe afin de faciliter les recherches. Chaque n?ud contient l'ensemble des informations relatives aux étiquettes qui le définissent. Ainsi pour le n?ud numéroté 1 de la figure 2.7, les tuples qui se réécrivent avec les étiquettes « *moins jeune* » et « *petit* » sont présents sous forme de lien vers la base de données relationnelles une fois pour l'attribut âge et une fois pour l'attribut taille. Les informations présentes dans un n?ud sont décrites, plus en détaile dans la quatrième partie. On peut par exemple interroger le graphe afin de savoir : « combien de personne sont jeune au degré au moins 0,6 sachant qu'elle sont petites ou moyennes ? ».

FIG. 2.5 - Intersection d'étiquettes sur des attributs différents
FIG. 2.6 - Intersection d'une union avec une étiquette seul
FIG. 2.7 -Intersection d'une union avec une étiquette seul

FIG. 2.8 - Représentation de l'âge sachant que le salaire est élevé

3 Evolution du modèle

Dans ce chapitre nous nous intéressons aux évolutions applicables aux données par rapport aux changements de granularité.

3.1 Changement de granularité

Pour pouvoir changer de granularité il est nécessaire de respecter quelques contraintes. Par exemple il est nécessaire que les ensembles s'imbriquent les uns dans les autres quand on veut effectuer des changements de granularité. Un autre problème lié à ces transformations est que les degrés d'appartenance ne sont pas constants. En effet lorsqu'un ensemble est redéfini par deux sous-ensembles, les données se trouvant dans l'intersection de ces deux sous-ensembles vont avoir des degrés différents de ce qu'ils avaient au départ.

3.2 Augmentation de la précision

Avec le choix d'architecture pour la représentation des données que nous avons fait, il peut être intéressant de pouvoir modifier les ensembles qui ont été définis pour obtenir plus de précision pour un ensemble qui serait fortement représenté. Dans ce cas, il faut être capable de scinder un ensemble en plusieurs sous-ensembles sans avoir la totalité de la structure à modifier. Une des contraintes qui est à appliquer sur les sous-ensembles qui vont découler de cette augmentation de précision est que l'enveloppe de l'ensemble d'origine doit servir de support à la répartition des sous-ensembles qui en sont dérivés. Toutefois des « creux » pouvant apparaître entre les sous-ensembles, ces derniers doivent pouvoir être pris en compte comme étant une donnée d'origine et non comme un artefact lié à une perte d'information lors du changement de granularité.

Définition : Un creux est l'ensemble des éléments appartenant à l'intersection de deux sous-ensembles lors de la redéfinition des étiquettes linguistiques.

FIG. 3.1 - Découpe d'un ensemble en trois sous-ensembles

La figure 3.1 présente le résultat d'un gain de précision sur l'ensemble A qui se retrouve découpé en trois sous-ensembles notés 1, 2 et 3. On peut remarquer que sur l'intervalle [a,b] l'ensemble A et 1 sont identiques. Sur l'intervalle [c,d], il en est de même avec les ensembles A et 2, ainsi que sur l'intervalle [e,f] pour A et 3. Sur les intervalles [b,c] et [d,e] l'ensemble A n'est pas repris avec exactitude par un des sous-ensembles mais est repris partiellement par deux des sous-ensembles, ce qui permet d'identifier ces ensembles et donc de réajuster les valeurs d'appartenance pour les éléments présents dans ces intervalles afin qu'ils soient présents dans les sous-ensembles avec les bons degrés d'appartenance.

Avec le remplacement d'un ensemble F par plusieurs sous-ensembles, l'ensemble des données faisant référence à A qui a été fractionné. Ainsi il faudra, en plus de la

modification de l'ensemble lui-même en trois nouveaux ensembles, réévaluer les ensembles d'union faisant intervenir l'ensemble A. La figure3.2 représente le découpage de F en trois sous-ensembles. Sur cette figure on peut constater que les ensembles E□F

FIG. 3.2 - Evolution de l'architecture avec le partitionnement de F

F□G doivent être recalculés puisque les nouvelles unions sont E□1 et 3□G et qu'elles contiennent moins d'éléments que les précédentes. Toutefois, s'il y avait eu d'autres données à côté de E et G, ces données n'auraient pas été modifiées et par conséquent n'auraient pas eu à être recalculées. L'organisation des données permet donc de changer la granularité, afin d'obtenir plus de précision sur une partie des données, sans pour autant affecter l'ensemble de la structure. Les modifications restent localisés au voisinage de l'ensemble qui a été décomposé en sous-ensemble plus précis.

Par exemple, on a défini pour l'attribut âge une étiquette portant le terme linguistique « *jeune* ». Il est possible de la redéfinir pour donner plus de précision et de prendre en compte trois nouvelles étiquettes remplaçant *jeune*. Ainsi on peut utiliser les étiquettes « *très jeune* », « *adolescent* » et « *moins jeune* » pour réécrire l'étiquette *jeune*. La figure 3.3 montre la répartition le passage de l'étiquette *jeune* à l'ensemble composé des trois étiquettes. Pour une personne dont l'âge est de 14 ans, avec la granularité la moins précise, elle s'écrivait jeune au degré 1 alors que avec la description la plus précise, il y a deux possibilités pour exprimer son âge. Soit on considère qu'elle est *très jeune* au degré 0,33, soit on considère qu'elle est adolescente au degré 0,67. Le phénomène de non égalité des degrés entre l'ancienne étiquette et les nouvelles se produit dans cet exemple.

FIG. 3.3 - Exemple de gain de précision avec redéfinition d'étiquette

○ *3.3 Généralisation d'un résumé*

La généralisation est le principe inverse de l'augmentation de précision. Il consiste à remplacer plusieurs ensembles par un unique ensemble qui les recouvre complètement. Il n'est cependant pas possible de se contenter de remplacer les ensembles par leur union, car si nous choisissions de procéder comme ceci nous aurions des problèmes liés à la présence de degrés inférieurs comme le montre la figure 3.4. Ces *creux* se trouvant au milieu de l'ensemble n'ont aucun sens sémantique puisqu'ils signifient qu'un élément qui est présent à cet endroit de l'ensemble s'exprime dans une moindre mesure par l'étiquette alors qu'il se situe au milieu de l'ensemble sur lequel est défini l'étiquette. Normalement, ils appartiennent à ce nouvel ensemble avec le même degré. Il est donc nécessaire de trouver une méthode qui va permettre de reconnaître ces *creux* et de les réévaluer pour qu'ils ne soient plus présents dans la version généralisée du système.

Ces *creux* sont reconnaissables puisqu'ils sont situés au niveau d'une intersection entre

deux ensembles. Il est donc possible de connaître les tuples qui sont présents dans le *creux* et de revaloriser leur degré d'appartenance à 1. Cela peut se faire automatiquement du fait de la formule suivante :

$$F_{sanscreux} = F_{aveccreux} - A \sqcap B + support(A \sqcap B).$$

Ainsi tout élément présent en dehors des extrémités d'un sur-ensemble que l'on crée doit être affecté d'un degré d'appartenance égal à 1.

FIG. 3.4 - Problème d'un creux sur un ensemble

Une fois encore, il faut prendre en considération les données qui ont été modifiées et voir à quel point cela affecte le reste du modèle. Une modification dans le modèle n'a d'impact que sur les données locale et n'affecte que très faiblement la structure.

Le modèle pour lequel nous avons opté présente donc une grande flexibilité au changement de granularité et peut donc être facilement utilisable dans des cas ou l'on ne connaît pas a priori la répartition des données. En effet, il est possible partir d'un partitionnement arbitraire sur les attributs et progressivement par fusion et subdivision d'ensembles arriver à une répartition des données suivant des étiquettes qui soit plus précise lorsque de nombreux tuples sont présents et plus générale quand les données ne sont que peu présentes.

4 Implémentation du modèle

Cette section présente des éléments de mise en ?uvre du modèle présenté précédemment. Sa réalisation sera effectuée au cours de la période de stage ingénieur ENSSAT allant de juin à septembre. Toutefois, dans ce chapitre, les structures de données sont présentées afin de servir de support pour expliquer le principe d'évaluation d'une requête sur le modèle.

4.1 Organisation des données dans un n?ud

Avant de définir la manière dont sont reliés entre eux les différents n?uds, il est important de préciser quelle est la structure d'un n?ud. Afin de rester le plus fidèle possible aux données et de prendre en considération un maximum de questions, nous avons choisi d'associer pour chaque attribut qui est réécrit les degrés d'appartenance de chaque tuple à l'étiquette linguistique. Dans l'exemple de la figure 4.1, sont pris en compte deux attributs l'âge et le salaire. Les étiquettes sont : *élevé* pour le salaire et *très*

FIG. 4.1 - Organisation des données dans un n?ud

jeune pour l'âge. Pour chacun de ces attributs on retrouve la même liste de tuple (ici T1, T4, T6, T8, T9) avec leur degré d'appartenance aux étiquettes *très jeune* et *élevé*. Ainsi T1 est présent au degré 1 pour le salaire, mais n'a qu'un degré de 0,7 pour l'âge. T8 quant-à lui vérifie totalement les étiquettes sur les deux attributs qui sont considérés dans cet exemple.

Ainsi avec une structure de ce type, il est possible de d'évaluer la question : « quels sont les personne qui ont un salaire *élevé* au degré au moins 0,9 sachant qu'elles sont *très jeunes* au degré au moins 0,5 ». On peut noter qu'il n'y a pas de calcul d'un degré général pour l'ensemble des attributs pris en compte dans le n?ud. Ceci ne pose en fait aucun problème puisqu'il est toujours possible d'appliquer l'opérateur de norme sur les données présentent dans le n?ud afin d'obtenir ces informations. Dans l'exemple précédent, pour prendre en compte la conjonction de *très jeune* et *élevé*, si on applique la norme de Zadeh (min(a,b)), on obtient les valeurs suivantes : T1 : 0,7, T4 : 0,4, T6 : 0,4, T8 : 1 et T9 : 0,8. Avec ces données on en déduit le nombre graduel des personnes qui sont *très jeune* et ont un salaire *élevé* est égal à {1,0/1 + 0,8/2 + 0,7/3 + 0,4/5}. On peut donc constater que toutes les données présentes dans la structure d'un n?ud permettent de calculer les informations qui n'ont pas été explicitées.

4.2 Organisation de la structure de données

Après avoir défini la structure d'un n?ud au 4.1, nous étudions ici la structure de données qui va nous permettre de représenter le modèle défini en 2.2 et de permettre un parcours pour répondre aux questions que posera un utilisateur. Dans la structure en partant d'un n?ud, on doit être capable d'accéder aux unions qui le composent.
D'un n?ud, il est aussi nécessaire de connaître les n?uds dont il est l'union afin de

pouvoir calculer certains degrés il est nécessaire de faire des différences entre les données de l'union et celle d'un des ensembles qui composent cette union. Par exemple, si on cherche les personnes qui sont *jeune* mais pas *très jeune*, il est possible de déterminer les personnes qui composent cet ensemble en calculant la différence entre l'union de *jeune* et *rès jeune* et l'ensemble *très jeune*.

Le schéma 4.2 présente les dépendances de l'on souhaite avoir dans la structure.

FIG. 4.2 - Organisation des données du modèle

Cette représentation ne porte que sur une structure basée sur l'utilisation d'un seul attribut. Dans cette représentation les liens sont considérés comme bidirectionnels afin de faciliter le parcours du modèle. Lorsque l'on considère qu'il y a plus qu'un attribut qui est pris en compte il faut remarquer que le nombre de liens vers les n?uds qui sont voisins se trouve lui aussi augmenté. En effet on peut avoir un maximum de 2^n liens vers des n?uds réalisant des unions avec n le nombre d'attribut pris en considération pour le résumé. L'exploration du graphe se fait en partant du bas avec une construction progressive des unions nécessaires pour répondre à la question qui a été posée.

En plus, de ces liens entre les données des différents niveaux, des liens existent entre les données feuilles afin de relier un ensemble avec ses voisins. Ces liaisons ont pour but de permettre le parcours des feuilles à la recherche d'une réponse pour les questions du type : « quelles sont les personnes qui sont *jeune* et *bien payé* ? ». Sans ces liaisons, la seule manière de répondre à cette question serait un parcours en changeant de profondeur dans l'arbre et de passer alternativement d'un n?ud seul à un union pour connaitre le n?ud suivant à explorer, et ainsi savoir si il peut répondre à la question. Mais le modèle ne se contente pas de répondre seulement à ce type de question, comme nous allons le voir au chapitre suivant.

5 Les apports de ce modèle

Partant d'un modèle complètement renouvelé, il est nécessaire d'évaluer les capacités de ce dernier en terme de requête qu'il est possible de lui adresser. ceci permet de mieux cerner ce qui peut être attendu du système et ainsi prévoir ce qui sera encore à modifier par la suite.

5.1 Interrogation du modèle

L'intérêt de créer un nouveau modèle est avant tout de permettre son interrogation afin de répondre à des questions. Il convient pour cela de commencer par définir le mode de parcours de l'arbre. Comme le remplissage de la structure se fait par les feuilles, une exploration par les feuilles est envisageable. Il est aussi possible de conserver un parcours plus habituel avec une exploration dont l'origine est la racine de l'arbre. C'est cette deuxième proposition que nous avons décidé de mettre en ?uvre.

Ce type d'exploration permet de répondre facilement à des questions portant sur des alternatives dans les étiquettes employées sur un attribut. On peut ainsi poser facilement des questions du type : « Combien de personnes sont de taille grande ou moyenne à un degré supérieur à 0,7 ? ». C'est avec ce genre de question que le modèle prend son intérêt.

On se place pour le moment dans le cas d'une interrogation sur un système à une dimension. Partant de la racine, seul deux chemins s'offrent pour le parcours de l'arbre. Il est donc nécessaire de définir un choix a priori du sous-arbre qui sera exploré en premier, et un parcours gauche-droite a été retenu. Pour chaque n?ud qui est exploré, on commence par comparer les étiquettes du n?ud avec celle de la requête. Si les étiquettes qui sont dans le n?ud sont exactement les mêmes que celles de la requête, l'exploration peut s'arrêter. Dans le cas contraire, il faut explorer un des sous-arbres. On commence par explorer le sous-arbre gauche pour regarder si les étiquettes présentent dans ce sous-arbre recouvre celles présentent dans la question. Si c'est le cas l'exploration se poursuit dans ce sous-arbre, si ce n'est pas le cas il faut regarder si le sous-arbre droit peut convenir. Si l'arber est exploré, sinon il est nécessaire de fractionner la requête pour pouvoir continuer à explorer l'arbre en ne considérant qu'une partie de la requête puis par la suite en regroupant les résultats de ces différentes sous-requêtes.

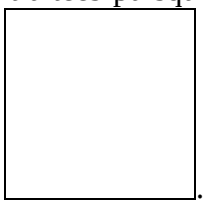
Par exemple il est nécessaire de fractionner la requête quand on pose la question suivante : « Combien de personnes sont jeune ou vieille ou très vieille ? », il n'est pas possible de trouver une union qui ne regroupe que ces étiquettes puisqu'elles ne sont pas contigus. Il va donc falloir diviser la requête en deux sous-requêtes afin de déterminer d'un côté combien de personnes sont jeune, puis combien sont vieille ou très vieille. Une fois ces deux résultats obtenus il est possible de faire la somme à partir de l'opérateur d'addition défini pour les nombres graduels.

Pour réaliser une interrogation sur deux attributs, il est possible de se rapporter au cas de l'exploration de l'arbre pour un attribut puisqu'on peut réaliser des projections suivant un attribut afin de fixer une étiquette qui est une partie de la question. Ainsi pour

répondre à la question : « combien de personne sont grande ou très grande et très jeune ou jeune ? » on va commencer par réaliser une projection sachant la taille grande ou très grande pour rechercher le n?ud contenant les étiquettes *jeune* et *très jeune*.

Pour les questions du type : « quelles sont les personnes dont la taille est petite ou moyenne? » l'exploration se fait directement au niveau des feuilles en supprimant les doublons. Ceci permet de mettre en évidence que les liens entre les feuilles réduisent considérablement le temps de recherche d'une réponse pour une question de ce type. Ainsi avec ce modèle il est possible de répondre à des questions portant sur les quantités de tuples et sur les tuples eux-mêmes qui vérifient une relation dans la base de données relationnelles de départ.

○ *5.2 Les évolutions à apporter*

Ce système ne permet cependant pas de répondre à toutes les questions. Par exemple, les questions portant sur une négation d'une étiquette ne peuvent pour le moment pas être traitées puisqu'avec des ensembles flous pour un ensemble considéré nous n'avons pas  . Il y a donc des éléments qui ne sont pas pris en compte si on se contente de retirer les éléments de A à l'ensemble du domaine. Une étude sur la possibilité de prendre en compte ces négations pourrait être entreprise par la suite.

Si on regarde la nature des données que l'on prend en compte dans ce système, on constate que tous les types de données ne sont pas gérés. Il peut s'avérer utile par la suite de permettre n'utiliser des données cycliques ou qui puissent avoir des intersections non vides entre plus de deux étiquettes. Ces modifications ajouteraient de nouvelles contraintes dont une plus particulière sur le critère de calcul des degrés d'appartenance d'un tuple dans une union. La formule

$$card(A \cup B) = card(A) + card(B) - card(A \cap B)$$

ne serait plus vérifiée pour l'union de trois ensembles. Pour trois ensembles A, B et C, on aurait donc la formule :

$$card(A \cup B \cup C) = card(A) + card(B) + card(C) - card(A \cap B) - card(A \cap C) - card(B \cap C) + card(A \cap B \cap C).$$

Cette modification pour trois ensembles est à considérer pour plus de trois ensembles. Ceci est donc un point important à prendre en compte mais ne peut pas être résolu rapidement, une analyse complète est nécessaire.

Une étude peut aussi être menée afin de déterminer quel est la meilleure stratégie à mettre en ?uvre pour réaliser les projections dans le cas d'une requête sur un graphe multidimensionnel. Dans cette étude nous avons choisi de prendre une décision en fonction du nombre d'étiquettes présente dans la question pour désigner l'attribut qui serait ciblé pour la projection. Dans certain cas il peut s'avérer que cela n'est pas une

bonne stratégie puisque les étiquettes peuvent être distinctes et ceci multiplie le nombre de plan de projection pour résoudre la requête.

Conclusion

Au cours de ce stage, une part importante du travail a été consacrée à l'analyse critique des modèles existants et notamment le modèle SaintEtiQ. On peut remarquer que cette analyse n'a pas été une perte de temps, puisqu'il s'en est dégagé que la notion de cardinalité était importante mais que SaintEtiQ ne la prend pas (ou peu) en considération.

Notre objectif a été de créer un nouveau modèle permettant de prendre en considération les cardinalités. Le modèle proposé s'appuie sur des résumés hiérarchiques multidimensionnels. Ces résumés s'organisent en treillis formé de n?uds contenant des étiquettes linguistiques et les informations relatives aux tuples qui se réécrivent suivant ces étiquettes linguistiques et contiennent des informations relatives à la cardinalité.

Le modèle proposé présente un intérêt appréciable du fait de sa modularité. En effet, il est aisé de changer la granularité sur tout ou partie des étiquettes linguistiques, dans la mesure où l'on conserve l'enveloppe des étiquettes qui ont été définies au départ. Les règles de transformation d'un ensemble en sous-ensemble et réciproquement sont suffisamment automatisables pour rendre ces mécanismes accessibles à un utilisateur.

Toutefois, la simplicité de ces transformations masque la présence d'une structure de données dont l'organisation est relativement complexe mais dont l'utilisation reste simple. Cette structure a été pensée pour fournir un maximum d'informations en limitant l'utilisation de mémoire. Elle doit permettre entre autre de poser des questions du type : « combien d'éléments sont x sur X au degré α ? » où X est un ensemble d'attribut et α bre compris entre 0 et 1. Nous envisageons d'expérimenter ce modèle sur la base de données movielens utilisée dans le cadre du projet national APMD (Accès Personnalisé à des Masses de Données); projet dans le cadre duquel cette étude s'intègre. La prolongation du stage permettra d'apporter des réponses sur la faisabilité et d'affiner les algorithmes de construction et d'interrogation du modèle.

Au vu de l'organisation de la formation dispensée pour la préparation du double diplôme de master recherche de l'université de Rennes 1 et le diplôme d'ingénieur de l'ENSSAT, le stage n'a commencé que fin mars. De ce fait, un manque de temps n'a pas permis la réalisation d'une implémentation, néanmoins des éléments concernant la manière dont va être réalisé le prototype ont été données. Ces informations sont susceptibles d'être modifiées mais présentent une indication sur la manière dont s'organisent les données.

Cette étude est une première investigation sur cette problématique de résumé de données, il reste bien entendu de nombreux point à approfondir notamment sur les heuristiques à utiliser afin d'optimiser le parcours du graphe de n?ud et sur l'évaluation de la complexité des algorithmes mis en ?uvre lorsque le nombre de dimensions pris en compte augmente. Il reste encore de nombreux éléments de ce modèle qui peuvent être étudiés.