UMONS
Université de Mons

Faculté
des Sciences

Mémoire

Data Repairing

Projet réalisé par : Maxime Van Herzeele
Année académique : 2017-2018
Directeur de Mémoire : Jeff Wijsen
Section : 2^{me} Bloc Master en Sciences Informatiques

Faculté des Sciences • Université de Mons • Place du Parc 20 • B-7000 Mons

Remerciements

Todo : remerciement

Table des matières

Chapitre 1

Introduction

De nombreuse entreprises et institutions récoltent, conservent et utilisent un nombre importants de données. Ces données sont parfois considérées comme inexactes car elles ne respectent pas certaines règles appelées *contraintes d'intégrité*. Parfois ce sont ces contraintes d'intégrités qui sont incorrectes. Ces deux types d'erreurs sont contraignantes à l'utilisation des données.

Le *data repairing* consiste à identifier les données et les contraintes d'intégrités et à les corriger. Dans ce mémoire, nous allons analyser une technique de data repairing appelés *Modèle de réparation θ-tolérant*. Cette technique est tiré d'un article scientifique Dans un premier temps, nous allons rappelé différentes notions importantes tel que les contraintes d'intégrités. + TODO(continué l'intro au fur et à mesure des chapitres)

Chapitre 2

?

2.1 Le modèle relationnel

Il est impératif que les données soient organisées et que les liens entre les informations soient modélisés. Cela se fait par le modèle relationnel avec contraintes d'intégrités. C'est l'un des modèles les plus utilisés. Le but de ce mémoire n'est pas d'expliquer ce modèle mais nous allons rappeler quelques notions que nous allons régulièrement utilisés.

Dans le modèle relationnel, nous avons [1]

— Un alphabet A de symboles de prédicats. Chaque symbole de l'alphabet est unique

— Un set de *contraintes d'intégrités* φ qui sont des règles définissant la cohérence d'une donnée ou d'un ensemble de données de la BD. Ce sont des assertions des prédicats de l'alphabet.

Definition 1. *Soit un schéma de relation R avec comme attributs $attr(R)$ Soit un ensemble $\mathbb{P}$ de prédicat P de la forme $v_1 \phi V_2$ ou $v1\phi c$ avec $v_1, v_2 \in t_x.A$, $x \in \{\alpha, \beta\}$, $t_\alpha, t_\beta \in R$, $A \in attr(R)$, c est une constante et $\phi \in \{=, <, >, \leq, \geq, \neq\}$ est un opérateur. Une* ***contrainte de déni(denial constraint)****[2] :*

TODO

2.2 Les bases de données

Dans cette section nous présenterons les bases de données que nous utiliserons comme exemple afin d'illustrer les différentes notions que nous aborderons.

2.2.1 Exemples de bases de données

La première base de données que nous allons utilisés vient de la source principale de ce mémoire [2].

	Name	BirthDay	Cellphone Number	Year	Income	Tax
t1	Ayres	8-8-1984	322-573	2007	21k	0
t2	Ayres	5-1-1960	***-389	2007	22k	0
t3	Ayres	5-1-1960	564-389	2007	22k	0
t4	Stanley	13-8-1987	868-701	2007	23k	3k
t5	Stanley	31-7-1983	***-198	2007	24k	0
t6	Stanley	31-7-1983	930-198	2008	24k	0
t7	Dustin	2-12-1985	179-924	2008	25k	0
t8	Dustin	5-9-1980	***-870	2008	100k	21k
t9	Dustin	5-9-1980	824-870	2009	100k	21k
t10	Dustin	9-4-1984	387-215	2009	150k	40k

TABLE 2.1 – Table de l'article de référence

.

La seconde base de données que nous allons utilisés est inventée de toute pièce. C'est un exemple de base de données que l'on peut retrouver dans un service public. Les attributs de la tables sont :

— **Niss :** Le numéro national de la personne.

— **Nom :** Le nom de la personne.

— **Prénom :** Le prénom de la personne.

— **Date_naissance :** La date à laquelle la personne est née.

— **Date_décès :** La date à laquelle la personne est décédée.

— **statut_civil :** Le statut civil de la personne (exemple : célibataire, marié, décédé, divorcé,...).

	Niss	Nom	Prénom	Date_naissance	Date_décès	statut_civil
t1	14050250825	Dupont	Jean	14-05-1902	18-05-1962	décédé

TABLE 2.2 – Table Person

.

2.3

Chapitre 3

Data Repairing

3.1 Variations de contraintes d'intégrités

Insertion et délétion.

3.2 Modèle θ-tolerant

3.3 Autres modèles

3.3.1 Holistic

et d'autres

Chapitre 4

Implémentation & comparaison de modèles

Chapitre 5

Conclusion

Bibliographie

[1] Giuseppe De Giacomo Andrea Cali, Diego Calvanese and Maurizio Lenzerini. Data integration under integrity constraints. Technical report, Departement of Computer Sciences and System, University of Roma "La Sapienza", •.

[2] Han Zhu Shaoxu Song and Jianmin Wang. Constraint-variance tolerant data repairing. Technical report, Tsinghua National Laboratory of Information Science and Technology, 2016.