ynov INFORMATIQUE

L'intégration des données hétérogènes
Basé sur le calcul des distances et des similarités

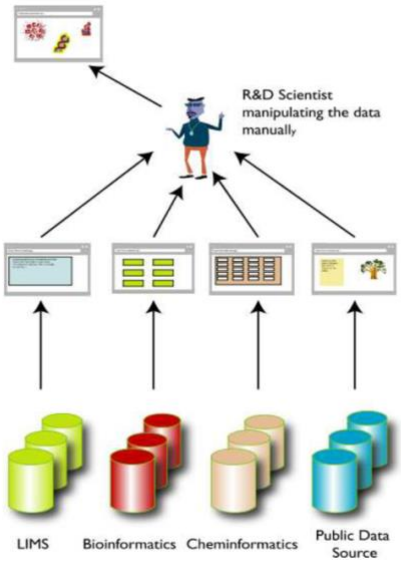
Contexte

L'entreprise *MedTrucs* est une entreprise axée sur l'utilisation des données de santé. Un de ses objectifs est d'offrir une plateforme, d'accès et d'analyses de données, permettant d'améliorer et d'innover les services de santé. La plateforme de données est régulièrement peuplée par des données issues de différentes sources.

Les sources des données sont hétérogènes et conduisent souvent à des difficultés d'intégration et d'analyse des données.

En tant qu'ingénieur de données, et en vous appuyant sur les outils mathématiques, vous êtes invité à mettre en place une solution d'intégration et d'interrogation de ces données hétérogènes.

Dans le cadre de ce tp, une source de données est une base de données relationnelle.



Définitions

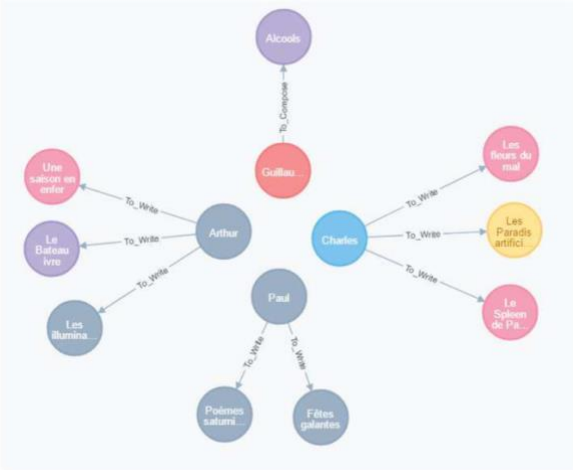
Schéma de base de données relationnelle. Une représentation structurée des données stockées dans une base de données relationnelle. Il définit la structure, l'organisation et les relations entre les tables qui composent la base de données. Le schéma de base de données décrit les entités (tables), les attributs (colonnes) de ces entités, les types de données, les contraintes d'intégrité, les clés primaires et étrangères, ainsi que les relations entre les tables. Une observation est un tuple des attributs d'une table

L'**hétérogénéité** peut être considérée selon différentes facettes :

- L'hétérogénéité structurelle, désigne le problème qu'une donnée peut être représentée par des éléments de structure variables.

- L'hétérogénéité syntaxique, désigne le problème qu'un élément de structure peut être désigné de manière variable ; par exemple, les attributs 'birth_date' et 'birth' dans les nœuds désignent toutes les deux une date de naissance d'un auteur.

- L'hétérogénéité sémantique, désigne le problème que deux éléments différents peuvent correspondre à une même donnée, ou inversement qu'un élément peut correspondre à des données variables ; par exemple, les relations 'To_Write' et 'To_Compose' ont le même sens.



Intégration des données. Processus de combinaison, de fusion ou de consolidation des données provenant de différentes sources pour fournir une vue unifiée de l'information. Cela implique de rassembler des données issues de systèmes et d'emplacements disparates, et de les transformer en un format qui peut être utilisé efficacement pour l'analyse, la génération de rapports et la prise de décisions. L'intégration des données est un élément essentiel dans le domaine de la gestion des données et est indispensable pour que les organisations puissent prendre des décisions éclairées et tirer des enseignements de leurs données.
Le processus d'intégration de données issues de différentes bases de données relationnelles aboutit à un schéma universel. Celui-ci ne signifie pas la concaténation des colonnes issues des différents schéma mais plutôt un nouveau schéma constitué sur la base de similarités entre les schémas (et donc les colonnes).

Mise en correspondance de données. Processus qui consiste à établir une connexion entre les éléments de données dans deux ou plusieurs sources de données en spécifiant comment les données dans une source correspondent ou sont liées aux données dans une autre source. L'objectif est de faciliter l'intégration des données en définissant les relations entre différents éléments de données et en indiquant comment les données doivent être transformées ou transférées d'un système à un autre.

Jeu de données

Dans ce tp, on considère trois sources de données au format tabulaire A, B et C.

Source	Noms des colonnes
A	'covid_symptoms', 'diabete_symptoms', 'smoking_symptoms', 'loc', 'age', 'weight', 'insulin', 'Pregnancies'
B	'wh', 'sd', 'medical_cat', 'medical_size', 'GR', 'am', 'Glucose', 'insu', 'BMI', 'DiabetesPedigreeFunction'

C	'cancer_symptoms', 'MPO', 'smplso', 'BloodPressure', 'insali n', 'gc', 'SkinThickness', 'HW'

Exercice 1

On souhaite effectuer l'intégration des données issues de plusieurs sources de données (A, B et C). Ces sources ne contiennent que des colonnes numériques.

Dans cet exercice, il est demandé de mettre en correspondance les colonnes de différentes sources. Soit deux colonnes $x, y \in R^n$, on dit qu'elles sont mises en correspondance si l'une des deux conditions suivantes est satisfaite :

- $d(x, y) < t \in R$
- Le produit scalaire $x.y \leq |1 - s|$ avec s $\in R$

Si une condition est respectée alors les colonnes sont considérées similaires voire identiques (si $d(x, y) = 0$ ou $x.y = 1$).

La distance d doit respecter les propriétés suivantes :

- Positivité. Pour toutes colonnes x et y, $d(x, y) >= 0$. On a $d(x, y) = 0$ si et seulement si $x = y$.
- Symétrie. Pour toutes colonnes x et y, $d(x, y) = d(x, y)$.
- Inégalité triangulaire. Pour toutes colonnes x, y et z, $d(x, z) <= d(x, y) + d(y, z)$.

A partir des sources A, B et C, identifier les paires de colonnes qui respectent la première condition. Puis effectuer la même opération en utilisant uniquement la deuxième condition.

La distance d est dans notre cas soit une distance euclidienne ou une distance de chebyshev. Les deux sont à évaluer. On fixe les seuils $t = 0.1$ et $s = 0.1$. Vous avez la possibilité de changer les valeurs des seuils.

Il est attendu pour chacune des trois solutions aboutisse à un ensemble de paires de colonnes mises en correspondance. Les trois ensembles ont-ils les mêmes paires ? Pourquoi ?

Note : les pré-traitements sont possibles sur les données pour un calcul de distance efficace.

Exercice 2

Dans cet exercice on s'intéresse toujours à la mise en correspondance des données mais dans le cadre des données massivement distribuées. Chaque fichier est découpé sur plusieurs machines d'un cluster.

Hypothèse forte :

- *On suppose que chacune des sources est distribué dans toutes les machines du cluster*
- *On suppose que dans la machine les trois sources A ont le même nombre de lignes et mêmes index (clés primaires).*

Dans cet exercice, on considère un cluster avec une architecture maître-esclave. Pour mettre en correspondance deux colonnes de deux sources différentes distribués dans un cluster, l'opération de correspondance doit s'effectuer localement dans la machine puis appliquer une opération d'agrégation de résultats dans un serveur maître pour une mise en correspondance des deux colonnes au niveau global.

Pour chacune des trois solutions (les deux distances et le produit scalaire) définir la fonction d'agrégation des résultats locaux. Pour le cas des deux distances, la fonction d'agrégation doit respecter les trois propriétés de distance telles énoncées dans l'exercice précédent. De même, pour le cas du produit scalaire, la fonction d'agrégation doit donner le même résultat que le produit scalaire appliqué sur des colonnes non distribuées.

Découper chacune des sources (A, B et C) en trois parties puis rassembler les sous-parties de A, B et C ayant les mêmes index.

Appliquer les trois fonctions d'agrégations.

Dans chacun des trois fonctions, obtenez-vous les mêmes ensembles de paires de colonnes que ceux de l'exercice précédent ?

Exercice 3 : mise en correspondance entre données catégorielles

Dans les deux exercices précédents, le processus de mise en correspondance ne porte que sur les colonnes numériques. Refaire l'exercice 1 en ne considérant que la distance d. La distance, dans le présent exercice, est appliquée sur une paire de colonnes pour une observation ayant une clé primaire i (c'est-à-dire $d(x_i, y_i)$). Nous définissons alors $d(x\ \ , y\ \ ) = \sum d(x_i, y_i)$.

La distance est soit la distance de jaro-Winkler ou la distance de levenshtein. Les deux sont à évaluer.

Exercice 4 : Mise en correspondance entre données catégorielles distribuées

Refaire l'exercice 2 en l'appliquant uniquement sur des colonnes catégorielles. Les fonctions d'agrégations doivent fournir le même résultat que si elles étaient appliquées sur les colonnes non distribuées.

Exercice 5 (Similarité structurelle et similarité de contenu)

Classiquement, la mise en correspondance est opérée en utilisant uniquement les schémas des sources de données. L'une des techniques de correspondance consiste à calculer la similarité entre noms de colonnes issues de sources différentes. Par abus de langage, on appellera cette technique similarité structurelle.

Dans les exercices précédents la mise en correspondance mène à un calcul de distances (ou produit scalaire) dans toutes les combinaisons de paires de colonnes en considérant leur

contenu. Dans le contexte de données massives (même distribuées), le processus est coûteux. On appellera cette technique similarité de contenu.

Dans le présent exercice, on définira une suite de règles de mise en correspondance :

1- Enumérer toutes les combinaisons de paires de colonnes numériques (de même pour les colonnes catégorielles) pour chaque paire de sources de données
2- Appliquer la similarité structurelle sur chaque paire.
    a. Si les colonnes de la paire sont similaires alors elles sont mises en correspondance
    b. Sinon appliquer la similarité de contenu
        i. Si les colonnes de la paire sont similaires alors elles sont mises en correspondance
        ii. Sinon elles ne sont pas mises en correspondance

La similarité structurelle est opérée en utilisant les distances catégorielles présentées dans l'exercice 3.