M1 SID - Fouille de données
TP2 - Structuration du corpus d'informations

Y. PITARCH
Basé sur un précédent sujet créé par A. EL HADDADI et L. LAPORTE,

2014-2015

1 Extraction de l'information explicite

1.1 Traitement de structures de données

Notre principal objectif étant la réactivité, nous avons constaté qu'une majorité des logiciels du commerce utilisait un format de données propriétaire et laissait le soin aux utilisateurs d'amener leurs informations dans ce format, d'où une perte de temps considérable et souvent l'impossibilité d'y arriver faute de compétence suffisante en informatique. Nous avons donc décidé de traiter, dans la mesure du possible, les informations dans leur format natif. Il y a plusieurs avantages à cela : meilleure réactivité, mise-à-jour du corpus facilitée, conservation de l'ensemble de l'information. Mais pour s'adapter à quasiment toutes les structures, il était nécessaire d'utiliser des outils de description des formats : les métadonnées. Leur principe est le suivant :

- Trouver une technique pour différencier les documents les uns des autres (ou les unités textuelles).
- Déterminer les balises des champs sémantiques présentes dans la base, leur donner un nom et un sigle standard.
- Définir leur utilité et leur priorité.
- Déterminer d'astucieuses techniques de découpage pour extraire au mieux chaque type d'information.

Plus de 90% des cas rencontrés peuvent ainsi être traités sans aucun reformatage.

1.2 Objets définis dans les métadonnées

- Bannière de synchronisation : elle sert à détecter le changement de document à l'intérieur d'un fichier séquentiel contenant un ensemble de documents au même format. Si une ligne vide sépare toujours deux documents eux mêmes sans saut de ligne, cette séparation peut éventuellement servir de bannière de synchronisation "Vide".

- Nom standard d'un champ : c'est le nom donné, dans la langue de l'utilisateur, à un type d'information balisée, il doit être le même dans tous les descripteurs compatibles avec les analyses multi-bases (exemple : Titre, Résumé, Auteurs, Dates, Pays, ...).
- Sigle standard d'un champ : il permet de nommer de façon standard (si possible avec deux lettres) répertoires, dictionnaires et matrices (PA.ind, PA.Filtre, PA.Syn, MC-PA, AU-AU-DT,...), c'est très utile pour normaliser la structure des analyses et leur appliquer des traitements automatiques (diffusion au format html, compilation dans une base de donnée pour alimenter un portail, ...).
- Bannière du champ : c'est le nom pris par un champ de la base au moment de sa collecte (long ou court) ou à l'issue d'un éventuel reformatage (html) : TI:, Title:, TI-, Titre :, ...
- Drapeau d'utilisation : il sert à masquer certains champs inutiles à l'analyse ou à masquer temporairement des champs peu utiles. Les interfaces utilisateur des fonctions de la plate-forme n'affichent donc que les champs requis.
- Liste des séparateurs : elle définit l'ensemble des chaînes de caractères qui servent à séparer les unités sémantiques à extraire de chaque champ. Certains jokers sont prévus : le saut de ligne "\n", l'espace "b", ... Un opérateur permet aussi de ne conserver, à l'extraction, que l'élément de rang i : "ORDi", "ORD0" pour le dernier

1.3 Description de formats spécifiques

Pour chaque base structurée ou semi-structurée, il convient donc de définir son descripteur de format spécifique qui permet de l'interfacer définitivement avec notre plate-forme de traitement de l'information. La Figure 1 présente un exemple partiel de descripteur de format.

1.4 Exercice

1. Définir le descripteur pour le fichier de données suivant :

Numero: 1
Auteurs: A1, A3
Journal: Nature
Index: Acp, Afc
Date: 2009

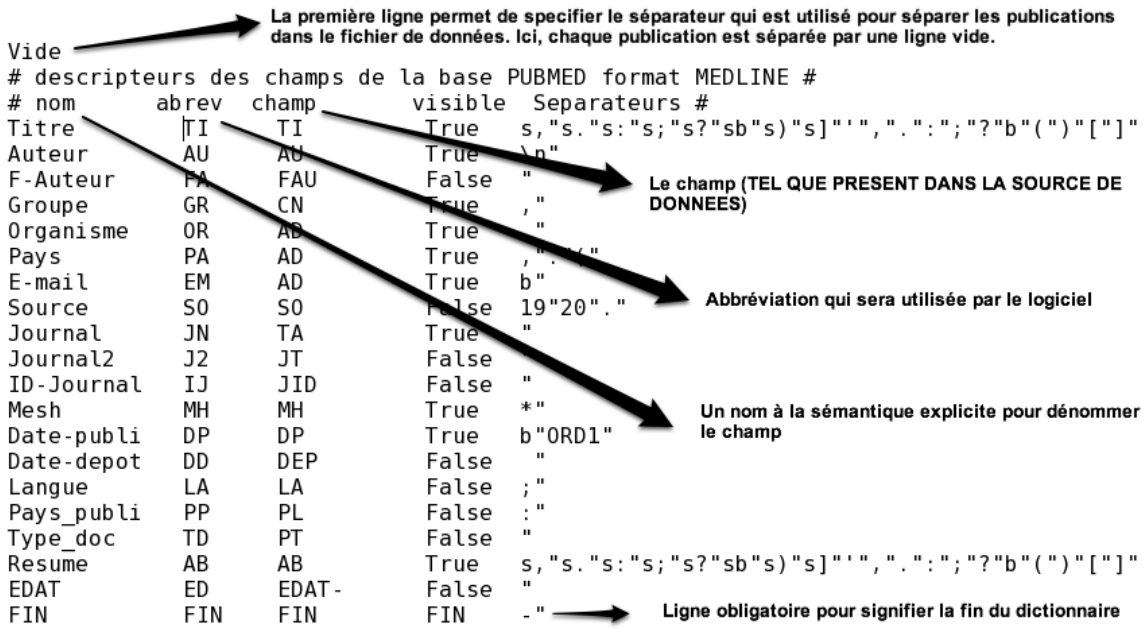

Figure 1: Exemple d'un fichier dictionnaire

2 Pris en main du logiciel Tetralogie

2.1 Connexion au serveur

2.1.1 Sous Linux ou Mac OS X

Vous trouverez ci-dessous les étapes à réaliser pour vous connecter au serveur de Tetralogie et démarrer le logiciel:

1. Ouvrer une console (ou un terminal suivant la terminologie Mac OS X)
2. Connecter vous en ssh au serveur en tapant la commande suivante : ssh -X u6@tetralogie2.irit.fr
3. Le mot de passe est : user6-sid
4. Une bonne pratique consiste à ouvrir 3 connexions ssh au serveur dans 3 terminaux (ou onglets). Un terminal permettra de lancer le logiciel; un autre servira à lancer un navigateur de fichiers; le dernier permettra l'exécution de commande UNIX pour manipuler/lister/modifier des fichiers
5. Dans un autre terminal, lancer un explorateur de fichiers, Konqueror en l'occurence, via la commande : konqueror&

2.1.2 Sous Windows

1. Lancer Xming
2. Lancer Putty avec les information de connextion suvantes : Host = tetralogie2.irit.fr, User = u6 et Password = user6-sid
3. 'Connection' → 'SSH' → X11 → 'Enable X11 forwarding'
4. Dans la console, lancer la commande : set
5. Puis ouvrir le gestionnaire de fichier via la commande : konqueror&

2.2 Première récupération de données

1. Sur PubMed, soumettez la requête "*Tobacco vaccine*"
2. Enregistrez les résultats dans un fichier texte : Send To → File → Format : MEDLINE → Create File
3. Ouvrez le fichier et observez sa structure
4. Création de votre répertoire de travail

   - Dans le répertoire *donnees*, créez un dossier personnel, avec pour nom:

   $$\langle idGroupe \rangle - \langle idListe \rangle - TP2$$

   - Transférez le fichier que vous venez de télécharger dans ce nouveau répertoire
   - Renommez votre fichier : **il doit porter le même nom que votre dossier, et avoir une extension .not** (et non .txt)
   - Dans votre dossier, créez un fichier vide 'basededonnees'. Ce sera le descripteur de vos données. **Notez que ce fichier descripteur devra toujours être nommé "basededonnees".**
5. Définissez le descripteur pour la base de données scientifique PubMed, à l'aide du fichier de notices
6. Pour tester votre descripteur, lancez tétralogie dans la console : tetralogie2&
7. Générez les fréquences absolues. Pour chaque champ, vérifiez le contenu des différents dictionnaires

2.3 Création des dictionnaires

Dans une analyse, il est recommandé d'extraire les termes de plus grand sens. Cette extraction se réalise méthodiquement après une utilisation par défaut de divers "dictionnaires" ou corpus de mots signifiant par leur sens ou leur non-sens. L'utilisation de fonctions UNIX est très commode. Les dictionnaires permettent d'extraire du même champ des informations de caractères différents (année, ville, pays...). Ils jouent un rôle important notamment dans le traitement de textes libres en typant les données. Nous allons vous indiquer les grands principes; il n'empêche que ces conseils ne sont pas limitatifs, vous pouvez les enrichir de votre propre expérience d'analyste. Nous vous proposons quelques indications pour construire des dictionnaires.

2.3.1 Mots vides

Le dictionnaire des mots vides regroupe les mots utilisés communément dans la structuration d'une langue mais n'ayant pas grand intérêt pour analyser une base documentaire. Il s'agit, par exemple, des articles, des adverbes, des pronoms ou des verbes auxiliaires. Ce dictionnaire se construit au fur et à mesure du temps et s'enrichit avec de nouveau corpus. Il est donc réutilisable au moment des opérations de comptage. Ce dictionnaire s'utilise en filtre négatif afin d'éliminer leur contenu de la base.

2.3.2 Les synonymes

Plusieurs mots peuvent avoir le même sens. Il est utile de construire par champ analysé un dictionnaire de synonyme et cela permet également de gérer les hiérarchies associées parfois aux données. Vous pouvez essayer de les construire manuellement, ce qui est par expérience assez aléatoire. L'oeil humain est moins sûr qu'un algorithme bien construit pour trier efficacement. Prenons par exemple le fichier des auteurs, les auteurs signent avec leur nom, leur(s) prénom(s), les prénoms sont tronqués plus ou moins. Il est important de trouver toute la production d'articles propres à un seul auteur. Il faut donc synonymer efficacement. Grâce au menu de synonymage, vous créez automatiquement à partir des "champ.ind" des synonymes par champ obtenu après l'opération de comptage simple. Il vous faudra relancer de nouveaux comptage en tenant compte des fichiers de synonymes. Il est préférable de faire vérifier le synonymage par un expert du domaine.

2.3.3 Instructions

1. Créez un fichier de mots vides que vous nommerez "motsVides.Filtre". Appliquer le en filtre négatif lors d'un comptage sur la dimension "Abstract". Validez le résultat en comparant le fichier des fréquences avant et après l'application du filtre négatif.
2. Créez un fichier de synonymes pour la dimension Pays à l'aide du menu. Regardez le contenu du fichier généré et modifiez le s'il est partiellement inexact ou incomplet. Vérifiez la validité de votre démarche en effectuant un comptage simple.
3. **Question complémentaire.** Si vous en avez le temps, simulez une hiérarchie sur la dimension Pays.