Restructuration de données semi-structurées : résolution par l'algorithme du volume, séquentiel et parallèle

Mourad Baïou[1], Francisco Barahona[2], Jean-Christophe Gay[3]

[1] CNRS, LIMOS, Complexe scientifique des Cézeaux, 63173 AUBIERE Cedex, FRANCE
baiou@isima.fr
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY10589, USA
barahon@us.ibm.com
[3] Université Blaise Pascal, LIMOS, Complexe scientifique des Cézeaux, 63173 AUBIERE Cedex, FRANCE
gay@isima.fr

Mots-Clés : *p-médian, algorithme du volume, données semi-structurées, programmation parallèle*

Les systèmes de gestion de bases de données (SGDB) sont en général pourvus de méthodes d'accès efficaces. Ce service est basé sur l'existence de régularités dans les données. La formulation de requêtes est grandement facilitée par la connaissance de la structure des données [4]. Les données sur le web sont des données irrégulières. Par exemple les différentes informations présentes sur les pages des membres d'un même groupe vont contenir des informations qui peuvent être communes (nom, prénom, adresse, mail, ...) ou non (surnom, photo, ...). Il existe des bases de données remplies de milliers d'enregistrements ayant des centaines de champs mis à vide par manque d'informations. Ces données qui se caractérisent par une structure qui peut être absente, irrégulière, implicite ou partielle se nomment des données semi-structurées [3, 2].

Avec l'émergence et la prise d'importance du XML, ce langage est devenu, de par ses caractéristiques, le support favori des données semi-structurées. Les différentes données présentes peuvent alors se représenter sous la forme d'un graphe dirigé labellisé. Les nœuds du graphe représentent les objets et les labels sur les arêtes portent les informations sémantiques à propos des relations entre les objets. Les nœuds pendants représentent les objets atomiques (ce sont les objets porteurs de valeurs), voir le graphe de gauche de la Figure 1.

Le problème de restructuration des données peut être ramené à un problème de p-médian. Le lien entre les deux problèmes est réalisé par le typage des objets. Le type d'un nœud c'est l'ensemble des relations qu'il possède avec ses prédécesseurs et ses successeurs. Grâce à cette notion de type on peut définir le type parfait d'une base de donnée semi-structurée. Ce type parfait consiste simplement à la juxtaposition de tous les types différents de la base. La taille de ce type parfait peut être très importante, elle peut même égaler le nombre d'objet dans la base si celle-ci est très irrégulière. On cherche alors un ensemble de types tel que la transformation des objets n'ayant pas leur type choisi soit la moins coûteuse (en information) possible. En représentant les différents types du type parfait par les nœuds d'un graphe et en assignant entre ces nœuds des arêtes ayant pour poids la distance entre eux, voir le graphe de droite de la Figure 1. Le problème se ramène à trouver p types parmi tous les types existant tout en minimisant la somme des distances entre les nœuds.

On peut alors résoudre le problème du p-médian sur ce graphe $G = (V, A)$, ce qui revient à résoudre un programme linéaire en nombres entiers. En pratique, ces problèmes sont de très grandes tailles, et

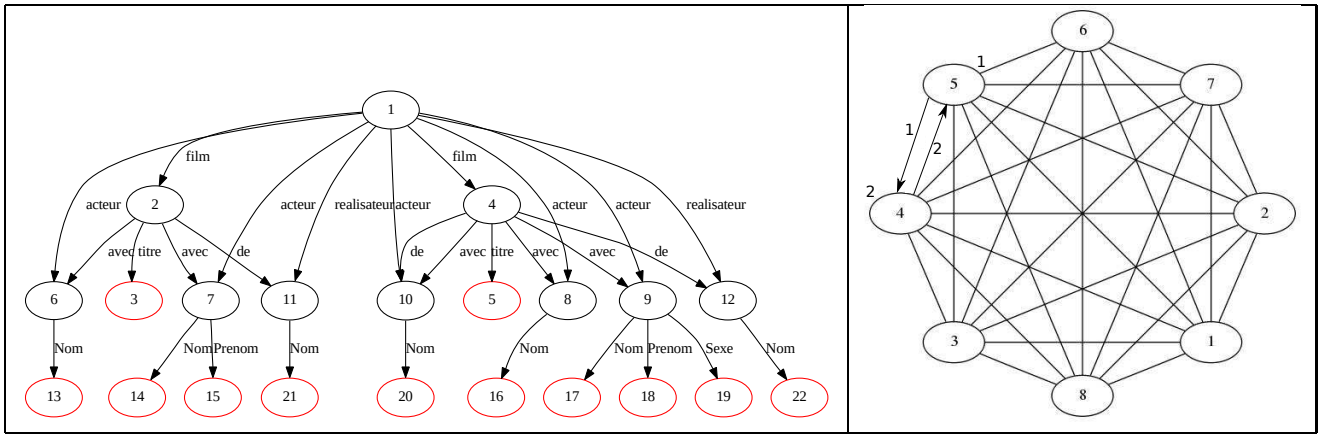

FIG. 1 – Graphe associé à un fichier XML et graphe de l'instance du p-médian associé.

donc des logiciels comme CPLEX sont impossible à utiliser, même pour résoudre la relaxation linéaire, à cause de la mémoire qu'ils requièrent. Nous montrons que l'utilisation de l'algorithme du volume [1] combiné à une heuristique, implémenté en séquentiel et en parallèlele donne des résultats très proches de l'optimum, comme l'illustre le tableau de la Figure 2. La première colonne indique le nombre, n, de sommets de l'instance, rappelons que le nombre de variables est égale à $n(n-1)$. La deuxième colonne représente le nombre de médians, p. Les colonnes 3, 4 et 5 indiquent, respectivement, la valeur de la solution duale, la valeur de la solution primale de la relaxation linéaire et la meilleure valeur de la solution entière donnée par l'heuristique. La sixième colonne donne le pourcentage de l'erreur relative de la solution entière trouvée par rapport à la valeur optimale du problème. Enfin, les deux dernières colonnes désignent le temps de résolution de l'implémentaion parallèle et sequentielle, respectivement.

n	p	L Val	Frac Val	Int Val	IGap	t para (sec)	t séq (sec)
1000	50	121325	121619	122800	1.21 %	74	50
1000	100	112959	113190	114142	1.04 %	55	52
1000	375	76218	76470	77131	1.19 %	35	46
1500	200	160440	160888	162067	1.01 %	89	121
1500	400	133821	134082	135371	1.16 %	58	117
1500	500	121113	121378	122443	1.10 %	55	126
3038	800	95305	95348	108131	13.4 %	173	585
3038	300	187516	187782	218924	16.7 %	305	624
5000	1250	772612	791040	987304	27.8 %	666	7855

FIG. 2 – Résultats expérimentaux.

Références

[1] Ranga Anbil Francisco Barahona. The volume algorithm : producing primal solution with a subgradient method. *Mathematical Programming*, pages 385–399, 2000.

[2] G. Hillbrand D. Suciu P. Bruneman, S. Davidsom. A query language and optimization techniques for unstructured data. 1997.

[3] R. Motawani S. Nestorov, S. Abiteboul. Extracting schema from semistructured data. 1997.

[4] J.D. Ullman. *Principles of Database and Knowledge-Base Systems, volumes I, II.* Computer Science Press, 1989.