Agrégation pour la réparation de liens

Léa Guizol

Agrégation pour la réparation de liens

Léa Guizol

LIRMM, UNIVERSITÉ DE MONTPELLIER II, CNRS, Montpellier, France
INRIA, Sophia-Antipolis, France
lea.guizol@lirmm.fr

Résumé :
Nous développons un système d'aide à la décision dans le contexte du projet Qualinca afin d'aider les bibliothécaires lors de l'ajout d'une nouvelle description de document. Nous discutons une méthode de validation des liens.
Mots-clés : système d'aide à la décision, agrégation de critères, entité résolution

1 Introduction

L'Abes, Agence bibliographique de l'enseignement supérieur, gère le Sudoc [1] (Système Universitaire de Documentation, une grande base bibliographique) depuis 2001. Le Sudoc contient environ 10 millions de descriptions de documents, ou **notices bibliographiques**, et 2,4 millions de **notices d'autorités**, des descriptions d'entités (lieux, personnes, évènements, ect.) utiles pour décrire les documents. Les notices bibliographiques sont reliées par des **liens** aux notices d'autorités identifiant des entités reliées au document décrit.

Lorsqu'un(e) bibliothécaire souhaite ajouter la description d'un livre au Sudoc, il créé une nouvelle notice bibliographique. Il renseigne les attributs de la nouvelle notice bibliographique (titres, ISBN, nombre de pages...) d'après le livre qu'il a entre les mains. Les **contributeurs** [2] sont représentés par des liens vers les notices d'autorité représentant ces personnes. Par conséquent, le(la) bibliothécaire doit chercher dans le Sudoc chaque notice d'autorité représentant l'un des contributeurs, à l'aide d'une fonction recherchant les notices d'autorités susceptibles de décrire une personne ayant une **appellation** (nom et prénom) donnée. Si il y en a plusieurs (homonymes ou orthographe proche), le(la) bibliothécaire doit décider quelle notice d'autorité représente le mieux la personne. Il(elle) peut aussi en créer une nouvelle pour représenter la personne si aucune ne convient.

Les notices d'autorités sont pauvres en informations. Par conséquent, pour décider si l'une d'elles représente la personne souhaitée, le(la) bibliothécaire doit regarder les informations contenues dans les notices bibliographiques liées à la notice d'autorité. Les erreurs de liage entre notices bibliographiques et notices d'autorité présentes dans le Sudoc favorisent donc l'ajout de nouvelles erreurs de liage.

Dans le but d'améliorer la qualité des données du Sudoc, un travail préliminaire a été présenté dans [1], où une méthodologie générale pour un système d'aide à la décision a été présentée afin de réparer les liens dans une base de connaissances bibliographiques comme celle du Sudoc. La méthode est basée sur le partitionnement des **autorités contextuelles** (objets représentant les

1. http://en.abes.fr/Sudoc/The-Sudoc-catalog
2. Personnes ayant contribué à la réalisation du document.

notices bibliographiques du point de vue d'un contributeur particulier) en fonction de **critères**. La méthode générale consiste en :

1. L'expert(e) entre une appellation A. Le système renvoie un ensemble de notices d'autorités du Sudoc susceptibles de représenter la personne désignée par l'appellation. Chaque notice bibliographique liée à une notice d'autorité sélectionnée est aussi sélectionnée.
2. Une autorité contextuelle est construite pour chaque lien entre une notice d'autorité et une notice bibliographique sélectionnées. Une autorité contextuelle correspond intuitivement à une personne, dans le contexte d'un document auquel elle a contribué.
3. Cet ensemble d'autorités contextuelles constitue le **sous-ensemble du Sudoc de l'appellation A**, noté $ses(A)$. $ses(A)$ est partitionné selon une méthode de partitionnement et un ensemble de critères. Les critères utilisés retournent des valeurs de comparaison symboliques et non pas numériques. Le but de cette étape est d'obtenir les partitions ayant le plus de "sens" selon les critères. Les partitions obtenues peuvent être comparées à la partition initiale (l'unique partition telles que toute et seulement les notices contextuelles issues d'une même notice d'autorité du Sudoc soient dans une même classe) afin de détecter des erreurs de liage dans le Sudoc et d'éventuellement les réparer.

Après avoir exploré comment les liens sont répartis dans le Sudoc et les critères implémentés (Section 2), on présente dans la section 3 un example montrant les limites de l'existant.

2 Données du Sudoc et critères

Nous avons compté le nombre de notices bibliographiques liées à chaque notice d'autorité du Sudoc représentant une personne afin d'observer la répartition des liens. Les liens sont très inéquitablement répartis entre les notices d'autorités :

– 1520285 notices d'autorités sont liées au moins 1 fois ;
– 972 notices d'autorités sont liées au moins 250 fois ;
– 113 notices d'autorités sont liées au moins 1001 fois ;

Les critères de partitions implémentés actuellement sont *domaine*, *date*, *titre*, *appellation*, *contributeurs* et *langue*. Le domaine de publication est représenté dans le Sudoc par une liste de codes de domaines. La distance entre deux codes de domaines et l'agrégation de ces distances ont été fournies par les experts. Les dates de publications sont comparées par rapport aux intervalles de temps entre elles. On utilise une distance de Levenstein adaptée pour comparer les titres. Le critère *contributeurs* donne une valeur de rapprochement en fonction du nombre de contributeurs en commun (excepté celui désigné par l'appellation). Le critère *appellation* est basée sur une fonction de comparaison fournie par les experts qui compare deux appellations (nom et prénom). Le critère *langue* donne une valeur de comparaison d'éloignement si les langues sont distinctes et qu'aucune n'est l'Anglais.

3 Approche et discussion

On considère le sous-ensemble du Sudoc représenté dans le tableau 1. On considère l'ensemble des autorités contextuelles liées à l'appellation "Sam, Harris" ($\{3,4,5,6,7,8\}$). La partition validée de façon experte est $\{\{5\},\{3,4\},\{7\},\{6\},\{8\}\}$. L'attribut "domaine" est une liste

id	titre	date	domaines	[...]	appellations
1	Le banquet	1868			“Platon”
2	Le banquet	2007			“Platon”
3	Letter to a Christian nation		[320,200]		“Harris, Sam”
4	Surat terbuka untuk bangsa kristen	2008	[200]		“Harris, Sam”
5	The philosophical basis of theism	1883	[100,200,150,100]		“Harris, Samuel”
6	Building pathology	2001	[720,690,690,690]		“Harris, Samuel Y.”
7	Aluminium alloys 2002	2002	[540]		“Harris, Sam J.”
8	Dispositifs GAA en technologie SON	2005	[620,620,530,620]		“Harrison, Samuel”

TABLE 1 – Exemple d’autorité contextuelles réelles

de codes de domaines de publication. Deux objets sont considérés par le critère *domaine* comme *proche* s’ils ont au moins un code en commun, et *eloigne* sinon. Deux objets sont considérés par le critère *date* comme étant *eloigne* si il y a plus de 59 ans entre leurs dates de publication. Donc, l’objet 5 est *eloigne* de tous les autres selon le critère *date*. Cependant, le critère *domaine* considère que les objets 3, 4 et 5 sont *proche* deux à deux parce qu’ils ont le code de domaine 200 (=religion) en commun : 3, 4 et 5 devraient être dans la même classe. Le critère *domaine* considère aussi que les objets 6, 7 et 8 sont deux à deux *eloigne* et *eloigne* des objets 3, 4 et 5. La seule meilleure partition selon les critères *domaine* et *date* est donc $\{\{5,3,4\}, \{7\}, \{6\}, \{8\}\}$. Ce n’est malheureusement pas la meilleure partition selon les experts. Nous affirmons que la raison de ce résultat insatisfaisant est due à la façon dont les valeurs de comparaison sont agrégées par de telles approches : comme des valeurs numériques.

Notre travail concerne deux sémantiques de partitionnement qui ajoutent :

– plusieurs niveaux de valeurs de rapprochement et d’éloignement ;
– pas d’interférence entre les valeurs de rapprochement et d’éloignement (par exemple, une valeur de rapprochement ne peut pas effacer une valeur d’éloignement).

Nous avons proposé deux sémantiques de partitionnement basées sur des critères à valeurs non-numériques. Ces sémantiques de partitionnement répondent aux exigences du projet dans lequel elles s’inscrivent, et en particulier au fait que nous ne souhaitions garder les valeurs de comparaison symboliques des critères le plus possible (par opposition aux techniques de partitionnement qui les numérisent pour les manipuler)[2].

Remerciements Ce travail a été soutenu par l’Agence Nationale de la Recherche (projet ANR-12-CORD-0012). Nous remercions chaleureusement Alain Gutierrez et l’ABES.

Références

[1] CROITORU M., GUIZOL L. & LECLÈRE M. (2012). On Link Validity in Bibliographic Knowledge Bases. In *IPMU’2012 : 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume Advances on Computational Intelligence, p. 380–389, Catania, Italie : Springer.

[2] GUIZOL L., CROITORU M. & LECLÈRE M. (2013). Aggregation semantics for link validity. *Proc. of SGAI-AI 2013*, p. 359–372.