Apport du Web et du Web de Données pour la recherche d'attributs

Rafik Abbes — Arlind Kopliku — Karen Pinel-Sauvagnat — Nathalie Hernandez — Mohand Boughanem

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG
118 route de Narbonne, F-31062 Toulouse Cedex 9, France

rafik.abbes@irit.fr, akopliku@weboglobin.com, karen.Sauvagnat@irit.fr, nathalie.hernandez@irit.fr, mohand.boughanem@irit.fr

RÉSUMÉ. *Nous nous intéressons dans cet article aux requêtes de type entité pour lesquelles on souhaite renvoyer un ensemble d'attributs (propriétés). Ces attributs peuvent être collectés à partir de plusieurs sources et agrégés dans un seul document. Par exemple, l'entité "France" peut avoir les attributs "Langue officielle: Français", "Villes:Paris, Toulouse, Lyon, ..." et "Population:* 65350000 *(en 2012)". Un attribut peut être monovalué ou multivalué, et peut éventuellement dépendre d'autres dimensions. Pour chercher les attributs d'une entité, nous avons exploité deux sources: les tables relationnelles du Web (issues du HTML) et le Web de Données. Afin d'évaluer le potentiel de ces sources, nous avons mis en place une évaluation utilisateur. Les analyses ont montré l'utilité de combiner ces deux sources pour répondre aux requêtes de type entité.*

ABSTRACT. *In this paper, we aim at answering entity-queries by searching their relevant attributes that can be collected from several sources and aggregated into a single document. For example, the entity "France" can have attributes such as "official language: French", "Cities:Paris, Toulouse, Lyon, ..." and "Population:* 65350000 *(in 2012)". An attribute can be multivalued or monovalued, and may possibly depend on other dimensions. To search entity attributes, we used two sources: relational web tables and Linked Data. To assess the contribution of these sources, we made a user study. Analyzes showed the usefulness of combining these two sources to answer entity-queries.*

MOTS-CLÉS : *recherche d'attributs, requête de type entité, tables relationnelles du Web, Web de Données*

KEYWORDS: *attribute retrieval, entity-queries, relational web tables, Linked Data*

1 Introduction

Selon une étude récente (Pound *et al.*, 2010), plus de la moitié des requêtes du Web ciblent une entité particulière ou des entités d'une classe. Une entité est une "chose" qui peut être distinctement identifiée : *"A. Einstein"*, *"IPhone 5"*, etc. Une classe d'entités est un ensemble d'entités de même type : *"Personne"*, *"Produit"*, etc. Pour répondre à ces requêtes, une alternative à la liste traditionnelle de documents consiste à agréger dans un seul document toutes les informations (descriptions, images, vidéos, attributs, etc.) retrouvées sur l'entité.

Nous nous intéressons dans cet article aux requêtes de type entité pour lesquelles nous cherchons à décrire une entité en renvoyant tous ses attributs pertinents. Un exemple de réponse pour la requête *"IPhone 5"* est illustré dans la figure 1. Chaque attribut résultat est un couple nom-valeur(s) fournissant une information partielle sur l'entité. Il peut être monovalué (*"Fabricant : Apple"*) ou multivalué (*"Réseaux : 3G, 3G+, 4G"*) et peut éventuellement dépendre d'une autre dimension comme le temps (*"NB exemplaires vendus : 5 millions (09/2012)"*).

IPhone 5	
Fabricant	Apple
Réseaux	3G, 3G+, 4G
NB Exemplaires vendus	5 millions (09/2012)
Site internet	www.apple.com
Date de sortie	21-09-2012
...	

Figure 1. *Exemple de réponse pour la requête IPhone 5*

Le Web est la source la plus utilisée pour rechercher les attributs d'une entité (Cafarella *et al.*, 2008). Ces attributs peuvent se trouver dans les documents Web sous forme de contenu non structuré (text brut) ou structuré (tables HTML, listes, . . .).

Une autre source pouvant fournir des attributs est le Web de Données. A la différence du Web classique qui utilise des documents, le Web de données manipule des ressources identifiées par des URIs[1] (Bizer *et al.*, 2009). Il consiste en un ensemble de données conçues pour être exploitées non seulement par des humains mais aussi par des machines. Ces données sont structurées sous forme de triplets RDF (Resource Description Framework)[2] comportant chacun un sujet, un prédicat et un objet. L'interrogation de ces données se fait par le langage SPARQL[3]. Pour mettre en pratique les concepts du Web de Données, le W3C[4] a lancé le projet *Linking Open Data*[5]. Ce pro-

1. Uniform Resource Identifiers
2. http ://www.w3.org/RDF/
3. http ://www.w3.org/TR/rdf-sparql-query/
4. World Wide Web Consortium
5. http ://linkeddata.org/

jet contient 295 entrepôts de triplets RDF comme DBpedia[6], Freebase[7], LinkedMDB[8], etc.

A notre connaissance, le Web de données n'a que peu été utilisé pour la recherche d'attributs d'une entité. Dans le but d'estimer son intérêt par rapport au Web classique (et plus particulièrement aux tables HTML), nous avons mis en place une évaluation utilisateur. Nous cherchons plus précisément à répondre aux questions suivantes :

– le Web de Données permet-il d'ajouter de l'information pertinente pour répondre aux besoins utilisateur ? Est-t-il important de combiner les données du Web de Données avec celles du Web classique ?

– Si cette combinaison est utile, à quelles éventuelles problématiques devrons-nous faire face si nous devions construire un système de recherche d'information prenant en entrée des requêtes de type entité ?

Cet article est organisé comme suit. Nous présentons dans la section 2 un état de l'art sur la recherche d'attributs. La section 3 présente les sources et les données utilisées dans le cadre de notre évaluation. La section 4 décrit la procédure d'évaluation et les requêtes choisies. Nous analysons dans la section 5 les résultats obtenus et nous concluons et énonçons quelques perspectives en section 6.

2 État de l'art : Recherche d'attributs

Le Web est la source la plus utilisée pour l'extraction des attributs. La méthode proposée dans (Bellare *et al.,* 2007) utilise des patrons lexico-syntaxiques comme le patron "the x of y" pour extraire les attributs d'une entité à partir des articles Web. Dans (Pasca *et al.,* 2007), les auteurs extraient les attributs d'une classe d'entités en appliquant des patrons lexico-syntaxiques sur des logs de moteurs de recherche. Dans (**?**), la méthode proposée permet d'acquérir des attributs d'une classe d'entités à partir des documents Web en exploitant des patrons lexico-syntaxiques, des statistiques sur les termes, et les balises HTML.

D'autres méthodes (Kopliku *et al.,* 2011, Cafarella *et al.,* 2008, Chen *et al.,* 2000) exploitent les tables relationnelles du Web pour extraire les attributs des entités. Ces tables sont définies comme étant des tables HTML contenant des données relationnelles.

A notre connaissance, peu d'approches ont utilisé le Web de Données pour extraire des attributs dans un but de recherche d'information. On peut citer (Krichen *et al.,* 2011), dans laquelle une approche a été proposée pour la recherche d'attributs pertinents d'une entité ou une classe en se basant sur la base de connaissance DBpedia. Les attributs sont extraits puis triés en fonction du nombre d'entités d'une même

6. http ://dbpedia.org/
7. http ://www.freebase.com/
8. http ://www.linkedmdb.org/

classe pour laquelle ils étaient renvoyés. Cependant, la couverture de cette approche reste limitée puisqu'elle ne s'appuie que sur une seule source.

La recherche d'attributs n'est pas utile qu'en RI. Par exemple, faire des statistiques sur la corrélation des attributs des entités permet de trouver des attributs synonymes d'un schéma relationnel et peut ainsi aider les utilisateurs débutants lors de la conception d'un schéma relationnel d'une base de données (Cafarella *et al.,* 2008). Les attributs peuvent également être utilisés pour enrichir automatiquement des bases de connaissances en les alimentant par de nouveaux triplets RDF (Gerber *et al.,* 2012).

3 Données exploitées dans le cadre de l'évaluation

Dans le cadre de la recherche d'attributs d'une entité, et afin de savoir si le Web de Données permet d'ajouter de l'information pertinente par rapport aux informations du Web classique, nous avons mis en place une évaluation dans le but d'étudier les attributs provenant de ces deux sources. Dans cette section, nous présentons les méthodes utilisées pour acquérir ces attributs.

3.1 Données issues du Web classique

Étant donnée une entité, la méthode utilisée pour extraire les attributs à partir du Web se base sur les tables HTML (Kopliku *et al.,* 2011). Elle comporte 2 étapes :

– Rechercher des tables potentiellement pertinentes dans les 50 premiers documents renvoyés par le moteur de recherche *Bing*[9] en réponse à la requête entité.

– Appliquer des filtres sur les tables trouvées afin d'éliminer les tables non relationnelles (comme les tables de mise en page) et les attributs non pertinents. Ces filtres sont implémentés en se basant sur une classification utilisant des paramètres comme la dimension de la table, des statistiques sur la longueur des lignes et des colonnes, la nature des valeurs des cellules, etc.

3.2 Données issues du Web de Données

Le Projet *Linking Open Data* est un ensemble d'entrepôts répartis en 6 domaines spécifiques : *Publications*, *Geographic*, *Media*, *Gouvernment*, *Life Sciences* et *User-Generated Content* et 1 domaine générique : *Cross-Domain*. Étant donnée l'immense quantité d'information qu'ils contiennent (31 milliards de triplets RDF) et afin d'alléger la quantité d'attributs à traiter dans notre évaluation, nous nous sommes limités à certains entrepôts que nous avons sélectionnés selon certains critères. Nous avons tout d'abord éliminé les entrepôts qui n'offrent pas de terminal SPARQL en ligne ou une API pour l'accès aux informations. Nous avons également éliminé des entrepôts ne

9. http ://www.bing.com/

correspondant pas à notre tâche de recherche d'attributs, comme les entrepôts de vocabulaire WorldNet, Yago, etc. Nous avons conservé les entrepôts les plus populaires (car probablement les plus "importants") en supposant qu'un entrepôt est populaire s'il est bien référencé par les autres entrepôts. Enfin, pour alléger la quantité d'information à juger tout en essayant de diversifier les thèmes (articles, films, pays, . . .), nous nous limitons pour chacun des 7 domaines à 2 entrepôts au maximum de thèmes différents.

En appliquant ces heuristiques de sélection nous avons obtenu 11 entrepôts. Deux sont des entrepôts génériques traitant divers thèmes : *DBpedia* et *Freebase* et 9 sont des entrepôts spécifiques : *Drugbank*, *Diseasome*, *Ordnance Survery*, *Statistics data.gov.uk*, *GeoSpecies*, *World Factbook*, *DBLP*, *BBC Music* et *Linked MDB*.

Pour chaque requête entité e nous cherchons ses attributs en interrogeant ces 11 entrepôts. La figure 2, illustre le mécanisme d'interrogation de DBpedia pour chercher les attributs de l'entité *'France'*. Nous distinguons 3 étapes :

– Rechercher l'URI de l'entité e. Nous pouvons nous appuyer éventuellement sur un moteur de recherche.

– Récupérer tous les prédicats et les objets associés à l'entité e, .

– Extraire les noms d'attributs à partir de la variable $?predicat$ et leur(s) valeur(s) à partir de la variable $?objet$.

Requête SPARQL:
SELECT ?predicat ?objet
FROM <http://dbpedia.org> WHERE
{<http://dbpedia.org/resource/France> ?predicat ?objet.}

predicat	objet
http://dbpedia.org/property/**currency**	"**Euro, CFP franc**"@en
http://dbpedia.org/property/**leaderName**	http://dbpedia.org/resource/**Fran%C3%A7ois_Hollande**
http://dbpedia.org/ontology/**capital**	http://dbpedia.org/resource/**Paris**
http://dbpedia.org/ontology/**anthem**	http://dbpedia.org/resource/**La_Marseillaise**

Figure 2. *Extrait des résultats de DBpedia pour la requête entité 'France'*

4 Mise en place de l'évaluation

Nous présentons dans cette section les requêtes et le protocole d'évaluation mis en place dans le but de de connaître le potentiel des sources utilisées dans la tâche de recherche d'attributs d'une entité.

4.1 Requêtes de l'évaluation

Nous avons tout d'abord sélectionné 19 classes d'entités dont :

– 9 classes **spécifiques** qui coïncident avec les domaines des entrepôts spécifiques choisis. Ces classes sont : *Disease*, *Countries*, *Drugs*, *Films*, *Non-metropolitan counties*, *Songs*, *Articles*, *English Cities* et *Species*.

– 5 classes **génériques encyclopédiques** regroupant des entités pouvant être décrites par une encyclopédie : *National Leaders*, *Places*, *Artists*, *Companies*, *Organisations*.

– 5 classes **génériques non encyclopédiques** regroupant des entités qui ne sont pas pleinement décrites par une encyclopédie : *SLR Cameras*, *Softwares*, *Laptops*, *Mobile phones*, *Programmable calculators*.

Pour chacune de ces 19 classes, nous avons choisi aléatoirement 3 entités. Nous obtenons donc 57 requêtes entités[10]. Nous voulons savoir à travers cette classification si nos sources répondent différemment aux différents types de requêtes.

4.2 Procédure d'évaluation

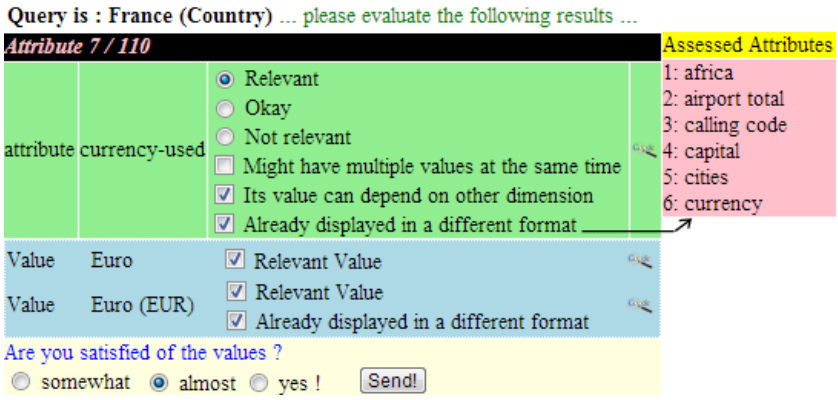

Figure 3. *Interface d'évaluation*

Nous avons mis en place une interface (figure 3) afin d'évaluer les 5783 attributs renvoyés au total par les deux sources pour les 57 requêtes. 14 volontaires ont participé à cette évaluation. Chaque requête n'a été jugée que par un seul évaluateur.

L'utilisateur juge tout d'abord la pertinence du nom d'attribut en cochant la case (*Relevant*, *Okay* ou *Not Relevant*). Si le nom d'attribut est pertinent (*Okay* ou *Relevant*), le juge précise si l'attribut est multivalué, par exemple le nom d'attribut *'Villes'* pour la requête *'France'*. Ensuite, il indique si le nom d'attribut a été déjà affiché sous une autre forme, par exemple *"currency-used"* a été déjà présenté à l'évaluateur sous

10. http ://www.irit.fr/entityEval/query.jsp

Source / Type de requêtes	Web de Données	Tables relationnelles
Spécifique (27)	50.33	23.51
Générique Encyclopédique (15)	33.60	18.46
Générique Non Encyclopédique (15)	5.13	22.06

Tableau 1. *Nombre moyen de noms d'attributs pertinents par source et type de requêtes*

une autre forme *"currency"*. Enfin, il évalue ses valeurs en cochant les valeurs pertinentes et en indiquant si une valeur a été déjà affichée sous une autre forme pour le même nom d'attribut. Si l'attribut est multivalué, il exprime sa satisfaction sur l'exhaustivité des valeurs proposées.

5 Analyses des résultats de l'évaluation

Dans cette section, nous analysons l'apport et la précision des deux sources en attributs pertinents. Puis, nous discutons de l'exhaustivité des valeurs des attributs multivalués.

5.1 Apport des sources en attributs pertinents

Nous illustrons dans le tableau 1 le nombre moyen de noms d'attributs pertinents renvoyés par source et par type de requête. Nous constatons que les tables relationnelles du Web répondent de la même manière à n'importe quel type de requêtes avec un nombre moyen d'attributs pertinents variant légèrement autour de 20. Cependant, le potentiel du Web de Données varie selon le type de requêtes : il répond très bien aux requêtes spécifiques, mais avec un nombre limité d'attributs pour les requêtes génériques non encyclopédiques. Le fait qu'une requête coïncide avec un domaine spécifique du Web de Données permet d'avoir plus d'attributs pertinents. Nous pouvons déduire que la couverture des tables relationnelles est meilleure que celle du Web de Données.

Nous voyons aussi le Web de Données répond mieux (plus d'attributs pertinents) que les tables relationnelles pour les requêtes spécifiques et génériques encyclopédiques. L'inverse est constaté pour les requêtes génériques non encyclopédiques. En effet, pour certaines classes d'entités comme *Laptop*, *Software*, etc., l'apport du Web de Données reste faible.

	Web de Données	Tables Relationnelles	Intersection
Attributs renvoyés	2859	3203	279
Attributs pertinents	1940	1223	231
Précision	67.85%	38.18%	82.79%

Tableau 2. *Précision des noms d'attributs*

5.2 Précision

5.2.1 Précision générale des noms d'attributs

Le tableau 2 illustre la précision des noms d'attributs issus des deux sources. La précision du Web de Données en noms d'attributs est meilleure que celle du Web classique. Nous pouvons expliquer ceci par le fait que les méthodes d'extraction d'attributs à partir des documents sont plus difficiles à mettre en place et génèrent beaucoup de bruit. Au contraire, l'extraction des attributs à partir du Web de Données est plus simple en se basant principalement sur le langage SPARQL.

279 noms d'attributs redondants sont renvoyés par les deux sources avec la même syntaxe. 231 noms d'attributs parmi ces 279, ont été jugés pertinents. Autrement dit, un nom d'attribut redondant a une probabilité de 83% d'être pertinent.

Les noms d'attributs peuvent être syntaxiquement différents, mais posséder le même sens, il s'agit de noms d'attributs synonymes comme *currency* et *currency-used* dans la figure 3. D'après les résultats de l'évaluation, nous avons 575 synonymies (19% des noms d'attributs pertinents renvoyés). Un système de recherche d'attributs doit détecter la redondance d'attributs d'une part pour éviter la duplication des données affichées à l'utilisateur, et d'autre part, pour l'utiliser probablement comme critère pour renforcer les scores de tri des attributs redondants.

5.2.2 Précision des noms d'attributs issus d'une ontologie

Les entrepôts du Web de Données peuvent contenir des noms d'attributs issus d'une ontologie. Notre étude sur *DBpedia* a montré que cet entrepôt a renvoyé pour les 57 requêtes 1399 noms d'attributs dont 408 sont issus d'une ontologie. La précision des noms d'attributs ontologiques est meilleur (84%) que celle des noms d'attributs non ontologiques (59%). Ce résultat pourrait être exploité dans le cas où nous voudrions trier les attributs renvoyés en donnant plus de poids pour les attributs issus d'une ontologie.

5.2.3 Précision des valeurs d'attributs

Les tables relationnelles ont renvoyé 1223 noms d'attributs pertinents dont 898 avec au moins une bonne valeur, soit une précision de 73%. Le Web de Données a restitué 1940 noms d'attributs pertinents dont 1741 noms d'attributs avec au moins

une bonne valeur, soit une précision de 89%. La précision des valeurs d'attributs est généralement bonne, mais la qualité des valeurs du Web de Données reste meilleure.

5.3 Satisfaction sur l'exhaustivité des valeurs des attributs multivalués

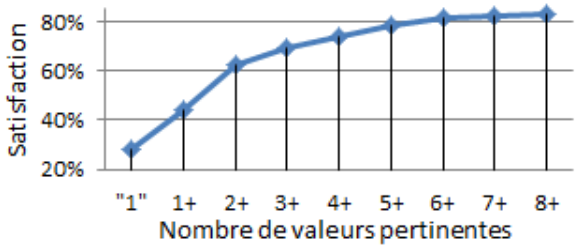

Figure 4. *Satisfaction des utilisateurs des attributs multivalués*

Pour les attributs jugés multivalués qui présentent 33% des attributs renvoyés, les évaluateurs expriment leur satisfaction des valeurs proposées. La courbe de la figure 4 illustre cette satisfaction en fonction du nombre de valeurs pertinentes proposées. Elle montre que plus les valeurs sont exhaustives, meilleure est la satisfaction.

6 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à rechercher les attributs d'une entité, à partir de deux sources : les tables relationnelles du Web et le Web de Données. Afin de savoir le potentiel de chacune de ces sources, nous avons mis en place une évaluation utilisateur. Les résultats d'évaluation ont montré que les tables relationnelles ont une bonne couverture mais une précision un peu faible alors que le Web de Données a une couverture un peu mois large pour certaines requêtes, mais sa précision est bonne. Il est utile donc de combiner ces deux sources pour la tâche de recherche d'attributs d'une requête entité.

Notre évaluation a également soulevé des problèmes qu'un système de recherche d'information devra résoudre lorsqu'il cherche et trie des attributs liés à une entité :

– Un bon système de recherche d'attributs doit détecter les redondances d'attributs, d'une part pour les éliminer lors de la présentation des résultats, et d'autre part pour s'en servir comme critère pour renforcer les scores de tri des attributs redondants.

– Afin de satisfaire au mieux l'utilisateur d'un tel système de recherche d'information, les valeurs des attributs multivalués se doivent d'être le plus exhaustives possible.

Ce travail ouvre beaucoup de perspectives intéressantes. En effet, nous souhaitons construire un système de recherche d'information permettant de répondre aux requêtes

de type entité, en tirant avantage de la large couverture du Web et de la précision des données du Web de Données. Ce système devra trier les attributs par ordre de pertinence et éviter la redondance des données.

7 Bibliographie

Bellare K., Talukdar P., Kumaran G., Pereira F., Liberman M., McCallum A., Dredze M., « Lightly-Supervised Attribute Extraction for Web Search », *Proceedings of Machine Learning for Web Search Workshop, NIPS*, 2007.

Bizer C., Heath T., Berners-Lee T., « Linked Data The Story So Far », *International Journal on Semantic Web and Information Systems*, vol. 5, n° 3, p. 1-22, 2009.

Cafarella M. J., Halevy A., Wang D. Z., Wu E., Zhang Y., « WebTables : exploring the power of tables on the web », *Proc. VLDB Endow.*, vol. 1, n° 1, p. 538-549, August, 2008.

Chen H.-H., Tsai S.-C., Tsai J.-H., « Mining tables from large scale HTML texts », *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 166-172, 2000.

Gerber D., Ngomo A.-C. N., « Extracting multilingual natural-language patterns for RDF predicates », *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management*, EKAW'12, Springer-Verlag, Berlin, Heidelberg, p. 87-96, 2012.

Kopliku A., Pinel-Sauvagnat K., Boughanem M., « Attribute retrieval from relational web tables », *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, Springer-Verlag, Berlin, Heidelberg, p. 117-128, 2011.

Krichen I., Kopliku A., Pinel-Sauvagnat K., Boughanem M., « Une approche de recherche d'attributs pertinents pour l'agrégation d'information », *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID), Lille*, http ://inforsid.irit.fr/, p. 385-400, 2011.

Pasca M., Van Durme B., « What you seek is what you get : extraction of class attributes from query logs », *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, San Francisco, CA, USA, p. 2832-2837, 2007.

Pound J., Mika P., Zaragoza H., « Ad-hoc object retrieval in the web of data », *Proceedings of the 19th international conference on World wide web*, WWW '10, ACM, New York, NY, USA, p. 771-780, 2010.