VERSUS : générateur de tableaux comparatifs à partir de bases de connaissances

Arnaud Giacometti, Béatrice Markhoff, Arnaud Soulet

Université de Tours, LIFAT, Blois
firstname.lastname@univ-tours.fr

Résumé. Les tableaux comparatifs sont utiles pour comparer des entités en dégageant leurs similarités et leurs différences non triviales. Le choix manuel des caractéristiques de comparaison reste une tâche complexe et fastidieuse. Cet article présente VERSUS qui est la première méthode automatique de génération de tableaux comparatifs à partir du Web sémantique. Pour cela, nous introduisons une mesure, nommée niveau de référence contextuel, pour évaluer si une propriété peut être une caractéristique intéressante pour comparer des entités. Cette mesure repose sur des contextes qui sont des ensembles d'entités similaires aux entités comparées. Nous montrons comment VERSUS sélectionne ces contextes et comment il évalue efficacement le niveau de référence contextuel à partir d'un point d'accès SPARQL public. Nous avons construit un benchmark à partir de Wikidata pour évaluer l'efficacité de VERSUS, avec une campagne d'évaluation manuelle des caractéristiques : la précision et le rappel sont élevés.

1 Introduction

Un tableau comparatif est un tableau à double entrée avec des entités à comparer en colonnes et les caractéristiques de comparaison en lignes. Le tableau comparatif est un outil particulièrement efficace pour la prise de décision en isolant les points communs et les différences significatives [1] entre les entités comparées. Par conséquent, cette technique analytique est populaire en science pour comparer des travaux, en culture pour comparer des oeuvres d'art ou dans le commerce pour comparer des produits ou des services. Dans ce contexte, cet article vise à automatiser entièrement le processus de génération d'un tableau comparatif d'un ensemble d'entités en interrogeant une base de connaissances disponible sur le Web sémantique telle que Wikidata (Vrandečić et Krötzsch, 2014). Par exemple, à partir d'Ada Lovelace et d'Alan Turing, nous voulons obtenir un tableau comparatif comme celui présenté par la table 1 construit automatiquement à partir de Wikidata. Au-delà des personnes, nous visons à comparer des entités telles que des lieux (pays, villes), des objets (tapisseries, statues), des institutions (universités, partis politiques), etc. Malheureusement, il n'y a pas de cadre théorique sur la conception des tableaux comparatifs pour déterminer si une caractéristique est intéressante pour comparer des entités. Cette tâche n'est pas anodine car dans 17 % des cas

1. Dans ce cadre, avoir des identifiants différents, ou des photos différentes, n'est pas considéré significatif parce que de tels traits ne sont partagés avec aucune autre entité.

un évaluateur humain ne sait pas si une caractéristique est intéressante ou non (voir la section 6). Ainsi, le principal défi est de formaliser la notion de caractéristique de comparaison *intéressante*. De plus, nous souhaitons bénéficier des immenses bases de connaissances disponibles sur le Web sémantique, ce qui pose un problème de robustesse et d'efficacité. En effet, ces bases de connaissances sont relativement fiables mais elles souffrent le plus souvent d'incomplétude (Razniewski et al., 2016). Pour cette raison, il serait souhaitable qu'une caractéristique jugée intéressante à un moment donné le reste malgré l'ajout ultérieur de faits. Par exemple, dans la table 1, déclarer la religion d'Ada Lovelace devrait maintenir l'intérêt de "religion " comme caractéristique de comparaison. De plus, plutôt que de centraliser les données, nous aimerions interroger directement les points d'accès SPARQL publics pour construire les tableaux comparatifs. Cela présente l'avantage de garantir un niveau de fraîcheur optimal et d'éviter le coût prohibitif d'une centralisation des données. Néanmoins, la politique d'usage juste de ces points d'accès, qui interrompt les requêtes trop longues, pose des problèmes d'optimisation (Soulet et Suchanek, 2019).

Caractéristiques	Ada Lovelace	Alan Turing	crl
sex or gender	female	male	0,908
spoken language	English	English	0,472
member of		Royal Society	0,205
field of work	mathematics, computing	mathematics, logic, cryptanalysis, cryptography, computer science	0,110
manner of death	natural causes	suicide	0,100
religion	?	atheism	0,015

TAB. 1 – *Exemple d'un tableau comparatif d'Ada Lovelace et d'Alan Turing*

Cet article présente VERSUS qui est la première méthode automatique pour générer des tableaux comparatifs à partir d'une base de connaissances. Notre approche centrée sur les entités conduit à plusieurs contributions. Premièrement, nous définissons une nouvelle mesure d'intérêt, appelée niveau de référence contextuel, afin de juger si une caractéristique est pertinente pour comparer des entités (section 4). Son principe est de privilégier les entités de référence dont les valeurs sont souvent utilisées par d'autres ensembles d'entités similaires, appelés contextes. Deuxièmement, nous montrons avec VERSUS (section 5) comment sélectionner les contextes et comment évaluer efficacement le niveau de référence contextuel d'une caractéristique tout en minimisant le nombre de requêtes soumises à la base de connaissances. L'idée est d'estimer les bornes inférieure et supérieure et d'interrompre le calcul dès que son intérêt est garanti ou non. Troisièmement, la section 6 évalue VERSUS sur un benchmark accessible au public, nommé *Comparison Feature Benchmark* (CFB), que nous avons développé pour évaluer la qualité des caractéristiques de comparaison. Il s'appuie sur 1000 tableaux comparatifs construits à partir de Wikidata et dont la pertinence a été évaluée manuellement. Sur ce benchmark, le niveau de référence contextuel conduit, avec une précision égale, à un meilleur rappel et une meilleure accuracy que la métrique de référence utilisée pour la génération automatique de facettes. De plus, notre évaluation optimisée nécessite beaucoup moins de requêtes.

2 Travaux relatifs

À notre connaissance, aucun travail ne s'est consacré à la construction automatique de tableaux comparatifs d'entités. La plupart des techniques qui comparent deux entités dans une

base de connaissances reposent sur une mesure de similarité (Anyanwu et al., 2005). Ces mesures sont pertinentes pour estimer la ressemblance entre deux entités, mais elles ne donnent pas explicitement les caractéristiques de comparaison (Tversky, 1977). Dans cette direction, Petrova et al. (2017) construisent des chemins dans les graphes de connaissances entre deux entités pour identifier toutes les similarités et toutes les différences. Malheureusement, aucune mesure d'intérêt ne filtre les différences inintéressantes et de ce fait, les attributs différents pour chaque entité mais inintéressants (comme les identifiants) sont aussi extraits. Les tâches les plus proches de la nôtre sont la génération de modèles d'infoboîte (Wu et Weld, 2008) et l'extraction de facettes (Hahn et al., 2010; Oren et al., 2006; Feddoul et al., 2019). Premièrement, une infoboîte est un ensemble de paires attribut-valeur décrivant une entité. Le choix des attributs est basé sur un patron défini pour chaque classe. Par exemple, les personnes [2] sont décrites par leur nom, leur date de naissance, leur nationalité, etc. De nombreux patrons ont été produits par les contributeurs de Wikipedia, mais des méthodes ont également été proposées pour affiner automatiquement ces patrons pour des classes plus spécifiques (Wu et Weld, 2008). Malheureusement, cette méthode orientée classes ignore les classes rares et les attributs rares. En outre, la plupart des attributs des infoboîtes décrivent les entités de manière singulière et ne sont donc pas utiles pour comparer des entités. Par exemple, l'image ou les oeuvres remarquables ne sont pas des caractéristiques qui peuvent être partagées par deux personnes.

Deuxièmement, la recherche par facette consiste à restreindre une collection d'entités en ne sélectionnant que celles avec une certaine valeur pour un attribut donné, appelée facette (Zheng et al., 2013). Une facette pertinente a fréquemment des valeurs partagées entre les entités observées. Il existe quelques méthodes d'extraction automatique de facettes. Pour une classe donnée, Hahn et al. (2010) extraient des modèles d'infoboîtes les attributs dont les valeurs sont fréquemment observées. De même, (Oren et al., 2006) mesure la qualité d'un attribut en privilégiant les attributs fréquemment utilisés dont les valeurs sont peu nombreuses et uniformément réparties. Récemment, Feddoul et al. (2019) ont proposés des mesures très similaires pour extraire les facettes mais une méthode de prétraitement regroupe les valeurs quantitatives (ignorées dans cet article) et une méthode de post-traitement filtre les facettes redondantes. Ces méthodes dérivent principalement des attributs pour un nombre limité de classes contenant un grand nombre d'entités. L'évaluation d'entités très similaires (comme Ada Lovelace et Alan Turing) repose sur des groupes comportant peu d'entités (e.g., les employés de l'université de Cambridge). Ainsi, la principale limite des méthodes d'extraction de facettes est de manquer certaines caractéristiques très spécifiques mais très pertinentes. Enfin, contrairement aux facettes utilisées pour la navigation, peu importe si une caractéristique de comparaison a beaucoup de valeurs dans la base de connaissances avec une distribution déséquilibrée.

3 Formulation du problème

Une base de connaissances d'un ensemble de relations $\mathcal{R}$ et d'un ensemble de constantes $\mathcal{E}$ (représentant des entités et des valeurs) est un ensemble de faits $\mathcal{K} \subseteq \mathcal{R} \times \mathcal{E} \times \mathcal{E}$. Nous écrivons un fait sous la forme $r(s, o) \in \mathcal{K}$, où r est une relation, s est un sujet et o est un objet. Par exemple, $\texttt{religion}(\texttt{Turing}, \texttt{atheism})$ indique que Alan Turing était athée. Pour une relation r, $r^{-1}(s, o) \in \mathcal{K}$ signifie que $r(o, s) \in \mathcal{K}$ où r^{-1} est la relation inverse de r. De

2. en.wikipedia.org/wiki/Template:Infobox_person

plus, $r_{\mathcal{K}}(s)$ (ou plus simplement $r(s)$ quand $\mathcal{K}$ est claire) est l'ensemble d'objets associés au sujet s pour la relation r dans $\mathcal{K}$. Par exemple, `field of work`(`Turing`) retourne l'ensemble {`mathematics`, `logic`, `computer science`, `cryptanalysis`, `cryptography`}.

Pour une base de connaissances $\mathcal{K}$, le tableau comparatif d'un ensemble d'entités $E \subseteq \mathcal{E}$ par un ensemble de caractéristiques $F \subseteq \mathcal{R}$ est le tableau avec $|F|$ lignes et $|E|$ colonnes où chaque cellule à l'intersection de la caractéristique f et de l'entité e contient les valeurs $f(e) = \{o \in \mathcal{E} : f(e,o) \in \mathcal{K}\}$. Par exemple, la table 1 présente le tableau comparatif des entités $E = \{$`Lovelace`, `Turing`$\}$ par les caractéristiques $F = \{$`sex or gender`, `spoken language`, $\ldots\}$. La cellule à l'intersection de `field of work` et `Turing` contient bien les valeurs `field of work`(`Turing`).

Une mesure d'intérêt $m : \mathcal{R} \times 2^{\mathcal{E}} \times 2^{(\mathcal{R}\times\mathcal{E}\times\mathcal{E})} \rightarrow [0,1]$ évalue l'intérêt $m(f, E, \mathcal{K})$ d'utiliser la relation f comme caractéristique pour comparer les entités E dans $\mathcal{K}$. Pour une base de connaissances $\mathcal{K}$, un ensemble d'entités $E \subseteq \mathcal{E}$, une mesure d'intérêt $m : \mathcal{R} \times 2^{\mathcal{E}} \times 2^{(\mathcal{R}\times\mathcal{E}\times\mathcal{E})} \rightarrow [0,1]$ et un seuil $\gamma \in [0,1]$, une caractéristique intéressante $f \in \mathcal{R}$ (au sens de m et γ) satisfait $m(f, E, \mathcal{K}) \geq \gamma$. **Pour une base de connaissances $\mathcal{K}$, un ensemble d'entités E, une mesure d'intérêt m et un seuil γ, notre objectif est d'extraire toutes les caractéristiques intéressantes $F = \{f \in \mathcal{R} : m(f, E, \mathcal{K}) \geq \gamma\}$ afin de construire le tableau comparatif de E par F.** Pour cela, nous devons relever deux défis. Le premier défi consiste à définir une mesure de l'intérêt qui estime la pertinence d'une caractéristique à partir d'une base de connaissances (voir la section 4). Le second défi est d'évaluer efficacement cette mesure en minimisant le nombre de requêtes SPARQL (voir la section 5).

4 Niveau de référence contextuel d'une caractéristique

Notion de contexte Intuitivement, pour comprendre et interpréter un tableau comparatif, une caractéristique est intéressante si les valeurs décrivant les entités comparées sont connues de l'utilisateur. En psychologie, Tversky (1977) a montré que l'utilisateur a besoin d'au moins une valeur dite de *référence* pour comparer deux valeurs. En particulier, si ces valeurs sont trop rares (ou même ne caractérisent que les entités comparées), l'utilisateur de la table a peu de chance de les connaître car il n'y a jamais été confronté. Parfois, ces valeurs sont informatives, mais elles n'aident pas à comparer les entités entre elles. Par exemple, le lieu de sépulture d'Ada Lovelace est l'église Hucknall St Mary Magdalene tandis que celui d'Alan Turing est le crématorium de Woking. Il n'y a pas de conclusion directe à tirer de cette différence. Bien entendu, cette notion de rareté dépend des entités comparées. Même s'il n'y a que peu de personnes qui sont membres de la Royal Society, cette caractéristique s'impose pour comparer deux personnes employées par l'Université de Cambridge. L'idée clé de notre mesure d'intérêt est d'évaluer la pertinence d'une caractéristique en fonction d'entités similaires aux entités comparées (par exemple, celles "employées par Cambridge" ou celles "parlant anglais" pour Ada Lovelace et Alan Turing). Nous formalisons cette intuition avec la notion de contexte :

Définition 1 (Contexte) *Pour un ensemble d'entités $E \subseteq \mathcal{E}$ et un couple relation-objet $(r, o) \in \mathcal{R} \times \mathcal{E}$ tel que $E \subseteq r^{-1}(o)$, le contexte C pour E issu de (r, o) est l'ensemble d'entités $r^{-1}(o) \setminus E$. $\mathbb{C}_E$ dénote l'ensemble de tous les contextes pour E.*

Intuitivement, un contexte C est un ensemble d'entités similaires mais différentes des entités de E par rapport à un couple relation-objet (r, o) partagé par toutes les entités de E. Pour le

tableau comparatif de la table 1, un exemple de contexte est l'ensemble des entités ayant l'anglais comme langue parlée (ici, le couple relation-objet est $(\texttt{spoken language}, \texttt{English})$). Naturellement, les classes sont propices aux contextes. Par exemple, toutes les personnes (c'est-à-dire les entités avec un couple $(\texttt{instance of}, \texttt{human})$ dans Wikidata) pourraient constituer un contexte pertinent pour Ada Lovelace et Alan Turing.

Définition de la mesure Pour les entités $E = \{e_1, \ldots, e_P\} \subseteq \mathcal{E}$, une caractéristique $f \in \mathcal{R}$ et un contexte $C \in \mathbb{C}_E$, plus les valeurs $f(e)$ d'une entité $e \in E$ décrivent les entités de C, plus la caractéristique f a une chance d'être une référence pour l'utilisateur du tableau. En termes probabilistes, l'intérêt de la caractéristique f doit croître avec la probabilité d'observer les valeurs de $f(e_i)$ dans l'ensemble des valeurs $f(s_i)$ parmi les entités similaires $s_i \in C$: $\Pr\big[f(s_i) \cap f(e_i) \neq \emptyset \;\big|\; s_i \in C\big]$. Ainsi, nous définissons le niveau de référence contextuel (ou *contextual reference level*) d'une caractéristique f de la manière suivante :

$$\begin{aligned} crl_C(f, E, \mathcal{K}) &= \Pr\big[(f(s_1) \cap f(e_1) \neq \emptyset) \vee \cdots \vee (f(s_P) \cap f(e_P) \neq \emptyset) \;\big|\; s_1 \in C, \ldots, s_P \in C\big] \\ &= \Pr\big[(\exists e_i \in E)(f(s_i) \cap f(e_i) \neq \emptyset) \;\big|\; s_i \in C\big] \end{aligned}$$

En pratique, les entités appartiennent à plusieurs contextes pertinents. Nous étendons donc la définition de $crl_C(f, E, \mathcal{K})$ à un ensemble de contextes :

Définition 2 (Niveau de référence contextuel) *Pour les entités $E = \{e_1, \ldots, e_P\} \subseteq \mathcal{E}$ et les contextes $\mathcal{C} = \{C_1, \ldots, C_K\} \subseteq \mathbb{C}_E$, le niveau de référence contextuel d'une caractéristique f est $crl_{\mathcal{C}}(f, E, \mathcal{K}) = \Pr\big[(\exists e_i \in E)(\exists k \in [1..K])(f(s_i^k) \cap f(e_i) \neq \emptyset) \;\big|\; s_i^k \in C_k\big]$.*

Notons que les entités comparées jouent un rôle très fort dans cette définition car elles limitent le choix de $\mathcal{C}$ dans l'ensemble des contextes potentiels $\mathbb{C}_E$. La quatrième colonne de la table 1 indique le niveau de référence contextuel de chaque caractéristique calculé à partir de Wikidata dans les 4 contextes suivants : $(\texttt{field of work}, \texttt{mathematics})$, $(\texttt{employer}, \texttt{Univ. of Cambridge})$, $(\texttt{occupation}, \texttt{computer scientist})$ et $(\texttt{spoken language}, \texttt{English})$. Avec la définition 2, il serait possible de calculer directement le niveau de référence contextuel d'une caractéristique avec une requête SPARQL. Cependant, cette requête statistique serait souvent trop coûteuse pour ne pas être interrompue par la politique d'usage juste des points d'accès publics SPARQL (Soulet et Suchanek, 2019). Néanmoins, dans la définition 2, comme les entités s_i^k sont choisies de manière identique et indépendante dans les différents contextes C_k, nous reformulons le niveau de référence contextuel :

Propriété 1 *Pour les entités $E \subseteq \mathcal{E}$, les contextes $\mathcal{C} \subseteq \mathbb{C}_E$ et une relation $f \in \mathcal{R}$, on a :*

$$crl_{\mathcal{C}}(f, E, \mathcal{K}) = 1 - \prod_{C \in \mathcal{C}} \prod_{e \in E} (1 - \Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big])$$

Par manque de place, nous omettons les preuves. De manière intéressante, chaque probabilité $\Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]$ peut facilement être calculée indépendamment par une requêtes SPARQL peu coûteuse. De cette manière, en pratique, le taux d'échec est inférieur à 0.5%. Par ailleurs, avec la propriété 1, il est facile de voir que le niveau de référence contextuel croît avec la probabilité $\Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]$ et que son codomaine est $[0, 1]$. Le niveau de référence contextuel est nul quand aucune entité parmi les contextes n'a une valeur en commun avec celles des entités E pour la caractéristique f (e.g., c'est le cas pour les attributs qui sont des identifiants). Inversement, $crl_{\mathcal{C}}(f, E, \mathcal{K})$ est égal à 1 dès qu'une valeur dans $f(e)$ est partagée par toutes les entités d'au moins un contexte C.

Robustesse de la mesure Nous présentons deux propriétés intéressantes que vérifie le niveau de référence contextuel pour faire face à l'incomplétude des bases de connaissances. Premièrement, le niveau de référence contextuel est monotone par rapport aux contextes :

Propriété 2 *Pour une base de connaissances $\mathcal{K}$, une caractéristique f et les entités E, on a $crl_{\mathcal{C}}(f, E, \mathcal{K}) \leq crl_{\mathcal{C}'}(f, E, \mathcal{K})$ si les deux ensembles de contextes satisfont $\mathcal{C} \subseteq \mathcal{C}' \subseteq \mathbb{C}_E$.*

Ce résultat s'explique par l'ajout de facteurs inférieurs à 1 dans le double produit de la propriété 1 lorsqu'un contexte est ajouté à $\mathcal{C}$. Il est pertinent que l'ajout d'un nouveau contexte favorise l'émergence de nouvelles caractéristiques intéressantes (par exemple, si une nouvelle relation est ajoutée à la base de connaissances). La propriété suivante va plus loin en montrant que le niveau de référence contextuel est également robuste contre l'incomplétude pour la caractéristique f évaluée :

Propriété 3 *Pour deux bases de connaissances $\mathcal{K}$ et $\mathcal{K}'$, les contextes $\mathcal{C} \subseteq \mathbb{C}_E$ et une caractéristique f telle que $f_{\mathcal{K}}(e) \subseteq f_{\mathcal{K}'}(e)$ pour tout $e \in E$, on a $crl_{\mathcal{C}}(f, E, \mathcal{K}) \leq crl_{\mathcal{C}}(f, E, \mathcal{K}')$.*

Cette propriété souligne que la valeur du niveau de référence contextuel est toujours sous-estimée lorsque certains faits manquent, c'est-à-dire que si de nouveaux faits sont ajoutés dans la base de connaissances, le niveau de référence contextuel d'une entité ne peut qu'augmenter. Pour cette raison, les résultats sont sûrs et ce qui a été jugé intéressant à un moment donné le restera dans le futur. Dans la table 1, la caractéristique `religion` a été sélectionnée malgré la valeur manquante pour Ada Lovelace. Quelle que soit la valeur qui pourrait être indiquée, cette caractéristique resterait intéressante pour crl.

5 VERSUS : extraction des caractéristiques intéressantes

Aperçu de VERSUS L'idée générale de l'algorithme 1 est d'analyser chaque relation f qui décrit au moins une entité de E pour déterminer s'il s'agit d'une caractéristique intéressante dans $\mathcal{K}$: $crl_{\mathcal{C}}(f, E, \mathcal{K}) \geq \gamma$. Pour commencer, l'ensemble F qui contiendra les caractéristiques intéressantes est initialisé avec l'ensemble vide (ligne 1) et l'ensemble $\mathcal{R}_E$ regroupe les relations qui décrivent au moins une entité de E (ligne 2). Après, chaque relation de $\mathcal{R}_E$ est traitée séparément (lignes 3-7). La ligne 4 sélectionne l'ensemble de contextes $\mathcal{C} \subseteq \mathbb{C}_E$ sans considérer la relation f (voir l'algorithme 2). Cet ensemble de contextes est immédiatement utilisé par l'algorithme 3 afin de déterminer si la relation f est une caractéristique intéressante pour les entités de E. Si c'est le cas, la ligne 5 l'ajoute à l'ensemble des caractéristiques intéressantes F. Finalement, cet ensemble est retourné à la ligne 7.

Sélection des contextes Cette étape essentielle vise à sélectionner un petit nombre de contextes pertinents parmi tous les contextes de $\mathbb{C}_E$ qui peuvent être redondants. En effet, dans le cas où un grand nombre de contextes dans $\mathcal{C}$ sont corrélés, le niveau de référence contextuel pourrait être anormalement surestimé à cause de la monotonie de la propriété 2. Par exemple, puisque tous les employés de l'Université de Cambridge sont forcément des humains, le contexte issu de $(\texttt{instance of}, \texttt{human})$ ne fournit pas d'informations supplémentaires, mais augmente la mesure de crl. Il est cependant important de garder un ensemble de contextes qui couvrent toutes les spécificités des entités similaires à E : $\bigcap \mathbb{C}_E$. Par exemple,

Algorithm 1 VERSUS : extraction des caractéristiques intéressantes au sens de crl

Input: Une base de connaissances $\mathcal{K}$, un ensemble d'entités $E \subseteq \mathcal{E}$ et un seuil γ
Output: L'ensemble des caractéristiques intéressantes $F \subseteq \mathcal{R}$

1: $F := \emptyset$
2: $\mathcal{R}_E := \{r \in \mathcal{R} : e \in E \wedge r(e,s) \in \mathcal{K}\}$
3: **for all** $f \in \mathcal{R}_E$ **do**
4: Sélectionner l'ensemble de contextes $\mathcal{C}$ pour les entités E et la relation f avec l'algorithme 2
5: **if** $crl_{\mathcal{C}}(f, E, \mathcal{K}) \geq \gamma$ (en utilisant l'algorithme 3) **then** $F := F \cup \{f\}$
6: **end for**
7: **return** F

Algorithm 2 Selection de l'ensemble de contextes

Input: Une base de connaissances $\mathcal{K}$, un ensemble d'entités $E \subseteq \mathcal{E}$ et une caractéristique $f \in \mathcal{R}$
Output: Un ensemble de contextes $\mathcal{C} \subseteq \mathbb{C}_E$

1: $\mathcal{C} := \{r^{-1}(o) \backslash E : r \in (\mathcal{R} \backslash \{f\}) \wedge (\forall e \in E)(r(e,o) \in \mathcal{K})\}$
2: Trier les contextes $\mathcal{C}$ par cardinalité croissante
3: **for all** context $C_i \in \langle C_1, \ldots, C_n \rangle$ **do**
4: **if** $\bigcap (\mathcal{C} \setminus C_i) = \bigcap \mathcal{C}$ **then** $\mathcal{C} := \mathcal{C} \setminus C_i$
5: **end for**
6: **return** $\mathcal{C}$

le contexte issu de $(\texttt{occupation}, \texttt{computer scientist})$ est important car il distingue Ada Lovelace et Alan Turing des mathématiciens de l'Université de Cambridge qui ne s'intéressaient pas à l'informatique. Pour cette raison, nous choisissons l'un des plus petits ensembles de contextes $\mathcal{C}^* \subseteq \mathbb{C}_E$ dont l'intersection caractérise le même ensemble d'entités que $\mathbb{C}_E$: $\mathcal{C}^* \in \arg\min_{\mathcal{C} \subseteq \mathbb{C}_E} \{|\mathcal{C}| : \bigcap \mathcal{C} = \bigcap \mathbb{C}_E\}$. La résolution exacte de ce problème NP-difficile nécessiterait de soumettre un grand nombre de requêtes à la base de connaissances. Nous proposons donc d'éliminer heuristiquement les contextes superflus du plus petit au plus grand.

Étant donné une base de connaissances $\mathcal{K}$, un ensemble d'entités E et une caractéristique f, l'algorithme 2 renvoie un ensemble de contextes $\mathcal{C}$. La ligne 1 construit l'ensemble des contextes $\mathbb{C}_E$ en excluant le contexte issu de la caractéristique f (i.e., $r \neq f$). Les contextes sont ensuite triés du plus petit au plus grand (ligne 2) pour favoriser la suppression des contextes trop généraux. La boucle (lignes 3-5) itère sur chaque contexte C_i en commençant par le plus petit. La ligne 4 teste si l'intersection de contextes sans C_i donne le même ensemble d'entités qu'avec C_i. Si tel est le cas, cela signifie que ce contexte ne fournit aucune spécificité et il est écarté de $\mathcal{C}$. Une fois la boucle terminée, tous les contextes redondants sont supprimés et l'ensemble des contextes $\mathcal{C}$ est renvoyé par la ligne 6. La table 2 présente les couples relation-objet (r, o) à partir desquels les contextes sont calculés en considérant Ada Lovelace et Alan Turing. Après avoir été triés par cardinalité ascendante dans Wikidata (i.e., $|r^{-1}(o) \backslash E|$), les deux contextes redondants ont été éliminés par les lignes 3-5 de l'algorithme 2. Par exemple, la restriction "instance of human" ne supprime aucune entité parmi celles appartenant à tous les autres contextes. On constate bien que cette méthode centrée sur les entités isole des contextes très spécifiques.

Évaluation efficace du niveau de référence contextuel L'évaluation naïve du niveau de référence contextuel serait coûteuse car chaque caractéristique nécessiterait de calcu-

Relation r	Objet o	$\lvert r^{-1}(o) \setminus E \rvert$
field of work	mathematics	2,018
employer	Univ. of Cambridge	3,129
occupation	computer scientist	7,943
~~described by source~~	~~Obalky knih.cz~~	47,563
spoken language	English	165,714
~~instance of~~	~~human~~	6,389,426

TAB. 2 – *Les couples relation-objets communs à Ada Lovelace et Alan Turing*

ler $|\mathcal{C} \times E|$ (resp. $|\mathcal{C}|$) requêtes pour les numérateurs (resp. dénominateurs) (voir la définition 2). Plutôt que de calculer le niveau de référence contextuel exact d'une caractéristique, l'idée est de faire un calcul partiel de cette valeur afin de déterminer uniquement si $crl_{\mathcal{C}}(f, E, \mathcal{K})$ est supérieur à γ. Il est facile de voir que le complément à 1 du niveau de référence contextuel (i.e., $1 - crl_{\mathcal{C}}(f, E, \mathcal{K})$) diminue à chaque multiplication par un facteur de la forme $\big(1 - \Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]\big)$. Avec cette observation, il est possible de dériver une borne inférieure pour le niveau de référence contextuel. Lors du calcul, lorsque cette borne inférieure dépasse le seuil γ, nous avons la garantie que $crl_{\mathcal{C}}(f, E, \mathcal{K}) \geq \gamma$. Inversement, il est possible de dériver une borne supérieure du niveau de référence contextuel en utilisant $\Pr[f(s) \cap f(e) \neq \emptyset, s \in \mathcal{E}]$ comme borne supérieure de la probabilité jointe $\Pr[f(s) \cap f(e) \neq \emptyset, s \in C]$. La propriété suivante formalise ces deux bornes :

Propriété 4 *Pour une base de connaissances $\mathcal{K}$, le niveau de référence contextuel d'une caractéristique f pour les entités E est bornée pour tout $\mathcal{S} \subseteq \mathcal{C} \times E$:*

$$
\begin{aligned}
crl_{\mathcal{C}}(f, E, \mathcal{K}) &\geq 1 - \prod_{(C,e) \in \mathcal{S}} \big(1 - \Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]\big) \\
&\leq 1 - \Bigg[\prod_{(C,e) \in \mathcal{S}} \big(1 - \Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]\big) \\
&\quad \times \underbrace{\prod_{(C,e) \in (\mathcal{C} \times E) \setminus \mathcal{S}} \left(1 - \frac{\min\{|\{s \in \mathcal{E} : f(s) \cap f(e) \neq \emptyset\}|, |C|\}}{|C|}\right)}_{\textit{facteur optimiste}} \Bigg]
\end{aligned}
$$

L'algorithme 3 bénéficie de ces bornes pour vérifier efficacement si $crl_{\mathcal{C}}(f, E, \mathcal{K}) \geq \gamma$. Plus précisément, les lignes 1 et 2 initialisent respectivement le produit p et le facteur optimiste o discuté ci-dessus en considérant tous les couples dans $\mathcal{C} \times E$. Les boucles des lignes 3 et 4 énumèrent les différentes entités $e \in E$ et les différents contextes $C \in \mathcal{C}$. A chaque itération, la ligne 5 affine le calcul de p en tenant compte de la probabilité $\Pr\big[f(s) \cap f(e) \neq \emptyset \;\big|\; s \in C\big]$ tandis que la ligne 7 met à jour o. Si le niveau de référence contextuel actuel est supérieur au seuil γ, la ligne 6 renvoie $true$ car $1 - p$ est une approximation pessimiste du niveau de référence contextuel final. Inversement, la line 8 renvoie $false$ lorsque la borne supérieure $1 - p \times o$ est inférieure à γ. Illustrons l'algorithme 3 pour la borne inférieure avec $\gamma = 0,01$. La probabilité d'avoir une entité avec une mort naturelle (comme Ada Lovelace) parmi ceux qui ont étudié les mathématiques est de $0,025$. Il est donc certain que `manner of death` est une caractéristique intéressante car son niveau de référence contextuel exact dépasse la

Algorithm 3 Calcul du niveau de référence contextuel d'une relation

Input: Une base de connaissances $\mathcal{K}$, les entités $E \subseteq \mathcal{E}$, un seuil γ, les contextes $\mathcal{C}$ et une relation f
Output: Retourne vrai si la relation f est intéressante i.e., $crl_{\mathcal{C}}(f, E, \mathcal{K}) \geq \gamma$

$p := 1$
$o := \prod_{(C,e) \in \mathcal{C} \times E} \left(1 - \frac{\min\{|\{s \in \mathcal{E} : f(s) \cap f(e) \neq \emptyset\}|, |C|\}}{|C|}\right)$
for all $e \in E$ **do**
  for all $C \in \mathcal{C}$ **do**
    $p := p \times (1 - (|\{s \in C : f(s) \cap f(e) \neq \emptyset\}|)/(|C|))$
    if $1 - p \geq \gamma$ **then return** $true$
    $o := o / \left(1 - \frac{\min\{|\{s \in \mathcal{E} : f(s) \cap f(e) \neq \emptyset\}|, |C|\}}{|C|}\right)$
    if $1 - p \times o < \gamma$ **then return** $false$
  end for
end for
return $false$

borne inférieure $1 - (1 - 0,025)$ qui est supérieure au seuil γ. Dans ce cas, cela économise l'évaluation de 7 requêtes nécessaires pour le calcul exact.

6 Expérimentations

Après avoir présenté notre benchmark, nos expérimentations visent à répondre aux deux questions suivantes : le niveau de référence contextuel isole-t-il vraiment les meilleures caractéristiques ? (Q1) et quel est le gain de l'évaluation optimisée ? (Q2). VERSUS est implémenté en Java et utilise la bibliothèque Jena pour interroger le point d'accès SPARQL public de Wikidata. Du fait du peu d'opérations effectuées côté client, les temps d'exécution correspondent essentiellement au temps de traitement des requêtes SPARQL côté serveur [3]. Le code source, l'outil d'évaluation et les résultats sont disponibles sur le site `lovelace-vs-turing.com`.

Comparison Feature Benchmark (CFB) La génération de tableaux comparatifs étant un nouveau problème, nous avons dû développer un benchmark pour évaluer la qualité des caractéristiques de comparaison, nommé *Comparison Feature Benchmark* (CFB).

Premièrement, nous tirons au hasard dans Wikidata 1000 types T_i ($i \in [1..1000]$) qui ont entre 10k et 1k instances. Cet échantillon garantit de couvrir une grande variété d'entités (personne, lieu, objets, événements, etc.) afin de refléter au mieux la diversité de Wikidata. Deuxièmement, pour chaque type T_i, nous sélectionnons les deux entités e_i^1 et e_i^2 qui ont le plus haut degré de faits entrants (i.e., maximiser $\deg(e) = |\{s \in \mathcal{E} : r(s,e) \in \mathcal{K} \wedge e \in T_i\}|$). Ce classement favorise les entités populaires du type T_i. Par exemple, les entités `Paris` (Q90) et `Londres` (Q84) sont sélectionnées pour le type `city` (Q515). Ensuite, pour chaque paire d'entités $E_i = \{e_i^1, e_i^2\}$, nous définissons l'ensemble F_i de relations r_j où $r_j \in \mathcal{R}$ où l'objet est une URI, r_j est une propriété directe de Wikidata (en utilisant le préfixe `http://www.wikidata.org/prop/direct/`) et $r_j(e_i^1)$ ou $r_j(e_i^2)$ n'est pas vide. Ainsi, F_i est l'ensemble des caractéristiques candidates pour comparer les entités dans E_i. Enfin, pour chaque paire d'entités $E_i = \{e_i^1, e_i^2\}$, nous stockons dans notre benchmark CFB tous les faits $r_j(e_i^k, o_i^k)$

3. `query.wikidata.org/`

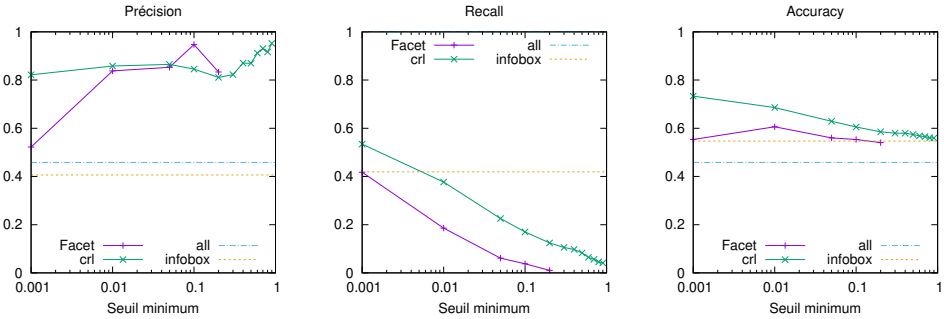

FIG. 1 – *Evaluation des mesures de précision, rappel et accuracy*

($k \in \{1, 2\}$) où $r_j \in F_i$ et o_i^k est un objet tiré aléatoirement parmi les valeurs de $r_j(e_i^k)$ (si $r_j(e_i^k)$ est l'ensemble vide, alors o_i^k est nul). Pour chaque type T_i ($i \in [1..1000]$), ce processus construit une table de comparaison avec $|F_i|$ lignes et deux colonnes pour comparer e_i^1 et e_i^2.

Deuxièmement, 1195 caractéristiques candidates (parmi les 11852 issues des tableaux, soit environ 10%) ont été tirées aléatoirement et évaluées manuellement par l'un des 6 évaluateurs. A chaque fois, on demandait si la caractéristique candidate $r_j \in F_i$ était pertinente pour comparer le couple d'entités $E_i = \{e_i^1, e_i^2\}$ (en sélectionnant dans CFB les faits $r_j(e_i^1, o_i^1)$ et $r_j(e_i^2, o_i^2)$). L'évaluateur peut répondre "Non", ce qui signifie que la caractéristique r_j n'est pas pertinente (44,9% des évaluations), "Oui" (37,9%) ou "Je ne sais pas" (17,2%). Seules 80 évaluations étaient communes dont 74 en accord ce qui donne un coefficient kappa de Cohen de 0,832 (correspondant à un accord *presque parfait* (Landis et Koch, 1977)).

Q1 : Qualité des caractéristiques extraites Tout d'abord, nous bénéficions du benchmark CFB pour comparer le niveau de référence contextuel utilisé par VERSUS (notée crl) avec la métrique utilisée pour la génération automatique de facettes (Oren et al., 2006) (notée Facet). Pour cette dernière, le type T_i des deux entités (voir ci-dessus) est utilisé pour définir la collection sur laquelle la métrique est calculée. Nous utilisons aussi deux lignes de base : la méthode all (Petrova et al., 2017) qui revient à sélectionner toutes les caractéristiques du benchmark et la méthode infobox qui sélectionne toutes les caractéristiques présentes dans au moins une des infoboîtes des entités E_i. La figure 1 trace la précision, le rappel et l'accuracy des différentes méthodes par rapport au seuil minimum. Pour les raisons évoquées lors de l'état de l'art, on observe que la précision des méthodes all et `infobox`, inférieure à 50%, est catastrophique. Ensuite, on constate que lorsque la précision de Facet est meilleure que celle de crl, le rappel de Facet est considérablement plus faible (moins de 20 caractéristiques sont extraites). Dans l'ensemble, le niveau de référence contextuel est bien meilleur que la métrique orientée facette avec une précision comparable mais un rappel et une accuracy plus élevés. Ce résultat n'est pas surprenant car, contrairement à Facet, notre méthode fait ressortir des caractéristiques spécifiques aux contextes des deux entités comparées. La précision du niveau de référence contextuel, toujours au-dessus de 76%, est élevée en comparaison avec les méthodes de base inférieures 50%. Fait intéressant, cette précision augmente avec le seuil de niveau de référence contextuel minimum (de 76% pour $\gamma = 0,0001$ à 86% pour $\gamma = 0,05$). Cela démontre la capacité de notre mesure à isoler les caractéristiques les plus pertinentes. Cependant, le rappel

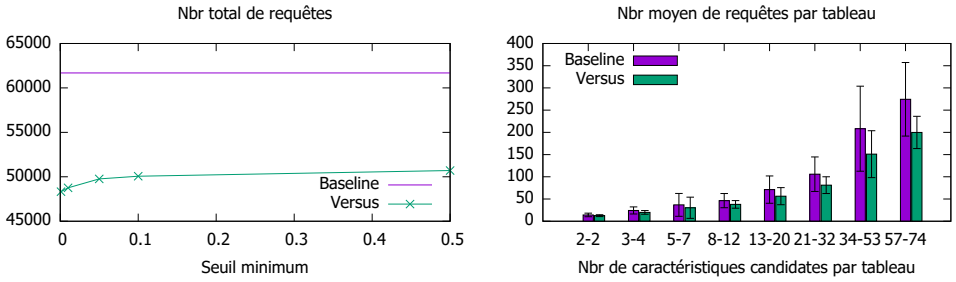

FIG. 2 – *Nombre de requêtes SPARQL exécutées sur Wikidata*

diminue très rapidement avec le seuil de niveau de référence contextuel minimum. Ceci s'explique par la diminution du nombre de caractéristiques intéressantes avec γ. Pour avoir un bon compromis, il faut fixer γ avec une valeur inférieure à 0,1.

Q2 : Efficacité de la méthode Nous évaluons maintenant le gain d'efficacité de la méthode optimisée (VERSUS bénéficiant de la propriété 4) avec une méthode naïve où la valeur exacte du niveau de référence contextuel est calculée (baseline basée sur la propriété 1). La figure 2 indique le nombre de requêtes SPARQL nécessaires pour construire les 1000 tableaux comparatifs du benchmark. La figure de gauche représente le nombre total de requêtes requises par baseline et par VERSUS par rapport au seuil de niveau de référence contextuel minimum. Il est toujours plus avantageux d'utiliser la méthode optimisée car moins de requêtes sont exécutées (environ -20% de requêtes). VERSUS est encore plus efficace pour les seuils bas (i.e. $\gamma \leq 0.1$). Pour $\gamma = 0.01$, on observe sur la figure de droite que le nombre de requêtes augmente linéairement avec le nombre de caractéristiques potentielles à tester. Ce résultat est attendu car le nombre de contextes (entre 1 et 7) est relativement indépendant du nombre de caractéristiques. Encore une fois, VERSUS s'avère toujours plus efficace.

7 Conclusion

Nous avons présenté VERSUS qui génère automatiquement un tableau comparatif d'un ensemble d'entités à partir d'une base de connaissances en interrogeant son point d'accès SPARQL public. À cette fin, nous avons introduit le niveau de référence contextuel qui évalue si une caractéristique a des valeurs pour les entités comparées qui sont suffisamment communes parmi d'autres entités similaires. Nous avons décomposé et optimisé le calcul du niveau de référence contextuel en plusieurs requêtes SPARQL à faible coût afin qu'elles satisfassent la politique d'usage juste du point d'accès public. Des expériences sur notre benchmark CFB montrent la bonne précision du niveau de référence contextuel pour isoler les caractéristiques les plus pertinentes. Il est intéressant de noter que notre approche centrée sur les entités a un rappel et une précision plus élevés qu'une méthode standard pour la détection de facette, qui repose sur des classes. De plus, grâce à notre optimisation, VERSUS économise environ 20% de requêtes par rapport à une approche naïve. Dans les travaux futurs, nous souhaiterions étu-

dier d'autres types de mesures d'intérêt plutôt basées sur l'exceptionnalité. Elles pourraient être utilisées en plus du niveau de référence contextuel pour extraire des caractéristiques inattendues.

Remerciements. Nous remercions les évaluateurs pour le temps qu'ils ont consacré pour annoter les caractéristiques. Ce travail a été en partie financé par le projet ANR-18-CE38-0009 ("SESAME").

Références

Anyanwu, K., A. Maduko, et A. Sheth (2005). SemRank : Ranking complex relationship search results on the semantic web. In World Wide Web, pp. 117–127.

Feddoul, L., S. Schindler, et F. Löffler (2019). Automatic facet generation and selection over knowledge graphs. In International Conference on Semantic Systems, pp. 310–325.

Hahn, R., C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Bürgle, H. Düwiger, et U. Scheel (2010). Faceted Wikipedia search. In Int. Conf. on Business Information Systems, pp. 1–11.

Landis, J. R. et G. G. Koch (1977). The measurement of observer agreement for categorical data. biometrics, 159–174.

Oren, E., R. Delbru, et S. Decker (2006). Extending faceted navigation for RDF data. In International Semantic Web conference, pp. 559–572. Springer.

Petrova, A., E. Sherkhonov, B. C. Grau, et I. Horrocks (2017). Entity comparison in rdf graphs. In International Semantic Web Conference, pp. 526–541. Springer.

Razniewski, S., F. Suchanek, et W. Nutt (2016). But what do we actually know ? In Proc. of the 5th Workshop on Automated Knowledge Base Construction, pp. 40–44.

Soulet, A. et F. M. Suchanek (2019). Anytime large-scale analytics of linked open data. In International Semantic Web Conference, pp. 576–592. Springer.

Tversky, A. (1977). Features of similarity. Psychological review 84(4), 327.

Vrandečić, D. et M. Krötzsch (2014). Wikidata : A free collaborative knowledgebase. Communications of the ACM 57(10), 78–85.

Wu, F. et D. S. Weld (2008). Automatically refining the Wikipedia infobox ontology. In Proc. of the 17th international conference on World Wide Web, pp. 635–644.

Zheng, B., W. Zhang, et X. F. B. Feng (2013). A survey of faceted search. Journal of Web engineering 12(1&2), 041–064.

Summary

Comparison table is useful for comparing entities for decision making. This paper presents the first automatic method for generating comparison tables from the Semantic Web. We introduce the contextual reference level to evaluate whether a feature is relevant to compare a set of entities. This measure favors the features whose values for the compared entities are reference among similar entites. We show how VERSUS efficiently evaluates this measure from a public SPARQL endpoint. The experiments show its efficiency for identifying the features deemed relevant by users with high precision and recall.