Amélioration de l'alignement de propriétés d'ontologies grâce aux plongements et à l'extension d'alignement

Guilherme Sousa[1], Rinaldo Lima[2], Cassia Trojahn[1]

[1] Institut de Recherche en Informatique de Toulouse, Toulouse, France
[2] Universidade Rural de Pernambuco, Recife, Brazil

Agenda

- Motivation
- Définition du problème
- L’existant
- Approche proposée
  - Calcul de similarité
  - Extension de l'alignement
- Évaluation
- Conclusions et travaux futurs

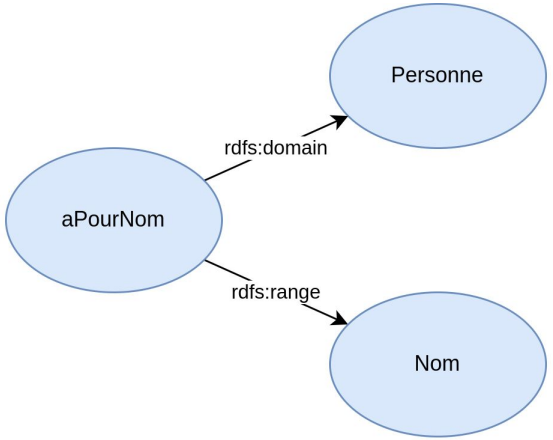
Motivation

- L'objectif du processus d'alignement d'ontologies est de trouver des correspondances entre les entités (classes et proprietés) de différentes ontologies.
- L'une des principales tâches de ce processus consiste à trouver des correspondances entre **propriétés**.
- Les approches d'alignement de propriétés de schémas de graphes de connaissances restent en retrait par rapport à la mise en correspondance des classes.

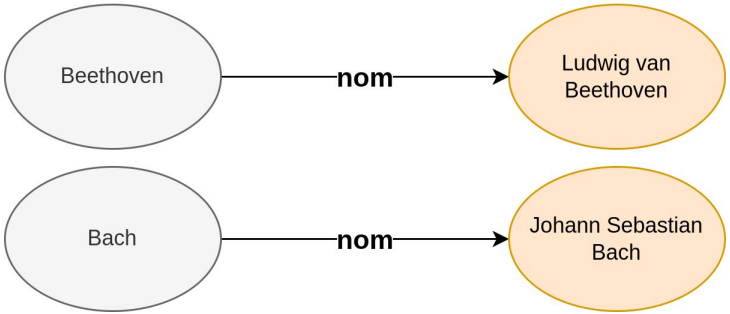
Motivation

- Les propriétés impliquent souvent une variation plus importante dans leur terminologie que les classes :
  - variation du verbe
  - mots fonctionnels
  - synonymes

A)



B)

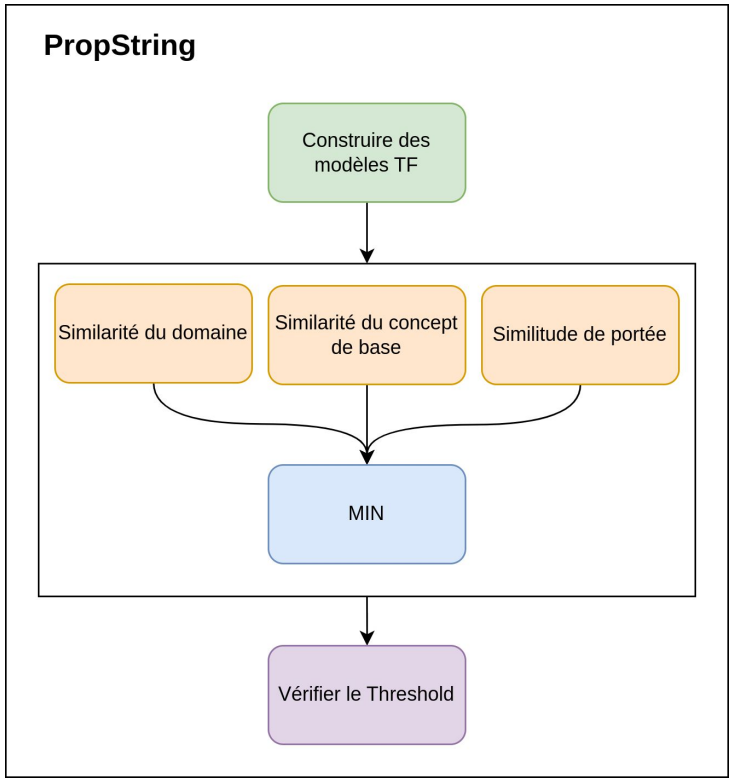


Définition du problème

- Recherche du meilleur ensemble de correspondances de propriétés **A** étant donné les ontologies en entrée O_1 et O_2.
- Les propriétés sont définies comme toutes entités **S** satisfaisant le prédicat P(S): $\exists D, \exists R$, domain(S, D) $\wedge$ range(S, R).
- Etant donné la fonction de similarité des propriétés **Sim**, trouver l'ensemble de correspondances **A** = {(p1, p2) $\in O_1 \times O_2$ | Sim(p1, p2) > t}.

L'existant : PropString

- Proposé par (Cheatham, 2014)
- Basé uniquement sur des mesures lexicales (TF-IDF)
- Une correspondance entre les deux propriétés est envisagée si la similarité minimale entre le domaine, la portée et le concept de base dépasse le seuil.



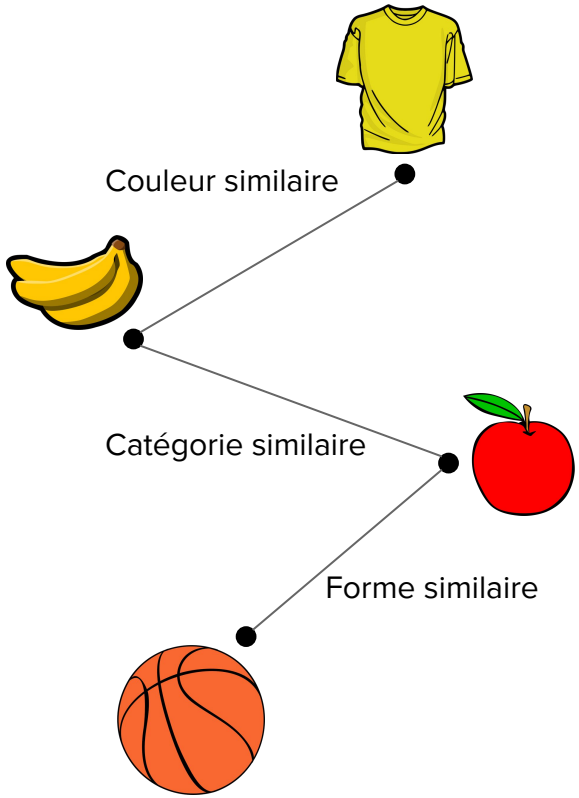
Cheatham, M., Hitzler, P.: The properties of property alignment. In: Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), October 20, 2014. CEUR Workshop Proceedings, vol. 1317, pp. 13–24. CEUR-WS.org (2014)

Proposition : PropMatch

- Nous étendons le système PropString
    - En utilisant de **plongements** pré-entraînés et
    - En exploitant la notion **d’extension de l'alignement**
- Implémenté en Python à partir de l’implémentation Java originale de PropString
- Disponible sur **https ://gitlab.irit.fr/melodi/ontology-matching/propmatch**

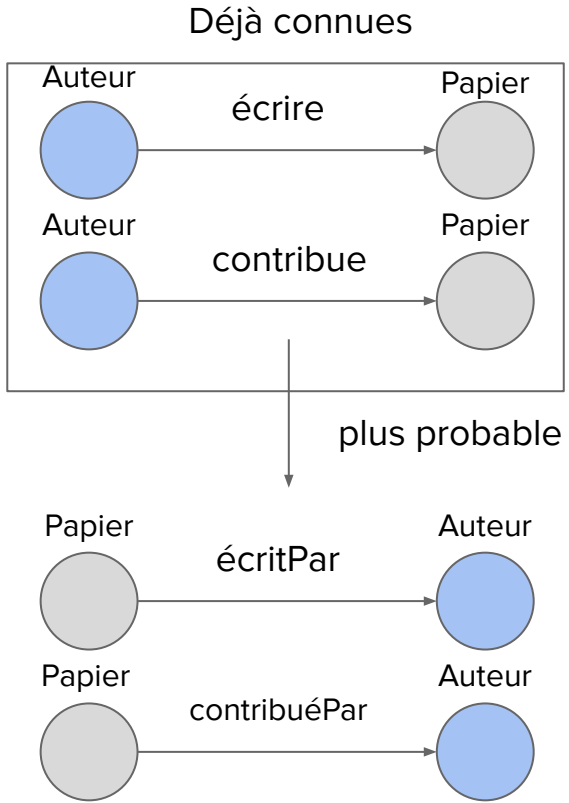
Proposition : plongements

- Représentation vectorielle d'objets, tels que des mots, des phrases ou des images.
- Ces vecteurs sont conçus pour capturer les relations sémantiques et contextuelles entre objets.



Proposition : extension d'alignement

- Processus d'utilisation des informations provenant de correspondances déjà connues pour découvrir de nouvelles correspondances.

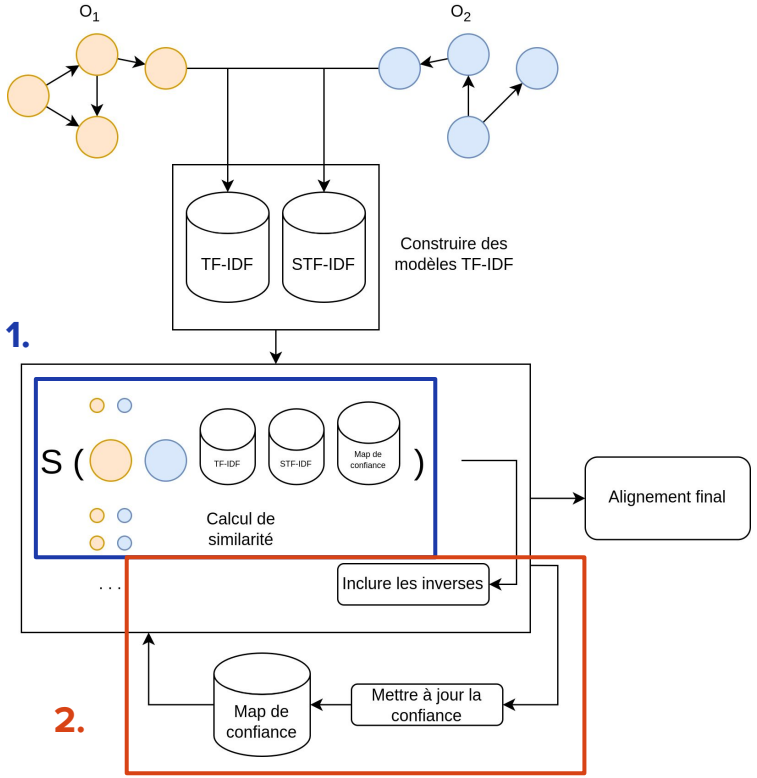


Architecture PropMath

Deux composants majeurs :

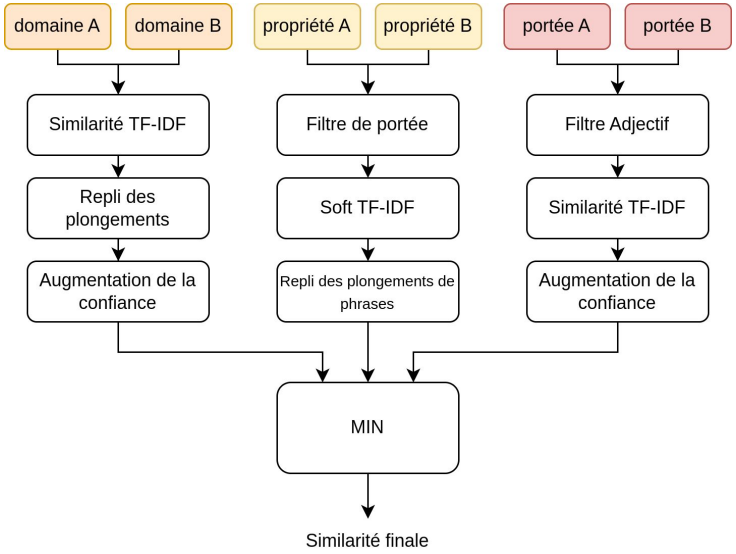
1. Calcul de similarité à l'aide de modèles de plongements.

2. Extension d'alignement



Calcul de la similarité

- Parallèlement à TF-IDF, deux modèles de plongements sont utilisés :
  - **plongements de mots** Finnish Internet Parsebank utilisé dans la **similarité de domaine**.
  - **plongements de sentences** all-MiniLM-L6-v2 du Hugging Face utilisé pour la **similarité de portée, domaine, propriété**.



Extension de l'alignement

- Implémenté en utilisation **une carte de confiance** pour effectuer un renforcement de similarité et addition d’inverses de propriétés.
- La carte de confiance met en correspondance une paire de domaines à une valeur et les domaines qu'elle associe sont mis à jour de manière itérative.
- Cette carte est utilisée pour le calcul de **similarité de domaine et de portée.**

Carte de confiance

ontologie 1:

(Paper, hasTitle, Title),

(Author, writes, Paper)

ontologie 2:

(Contribution, hasTitle, Title),

(Author, contributes, Contribution)

* Les similitudes dans le domaine et la portée sont différentes car les plongements de mots ne s'appliquent qu'aux domaines.

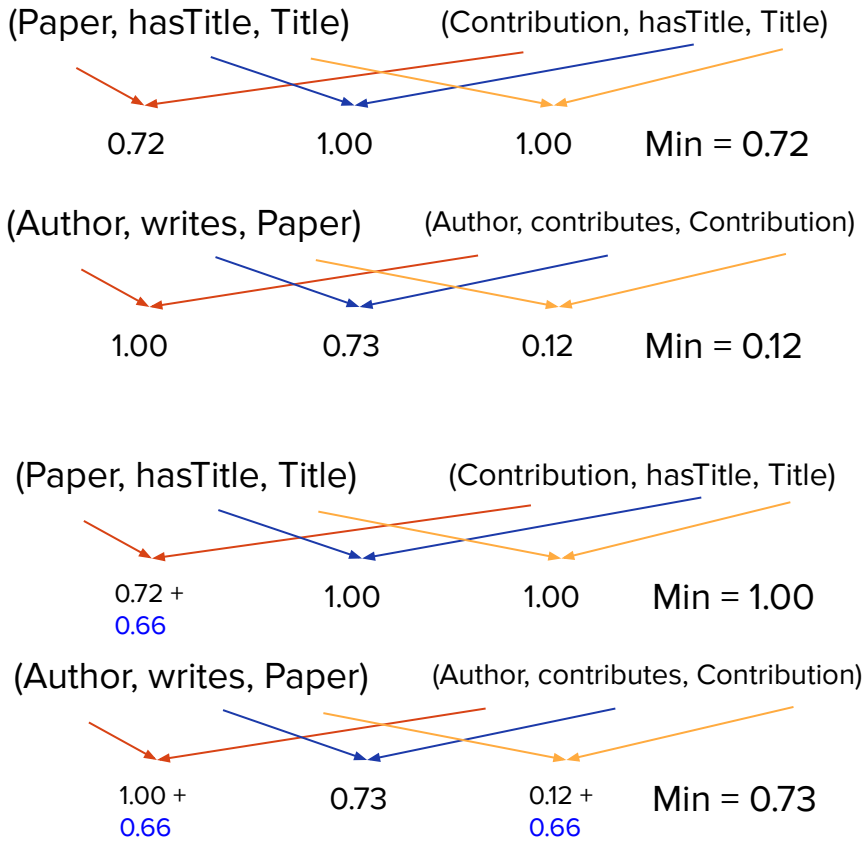
Carte de confiance

Itération 1

Carte de confiance

(Paper, Contribution) = 0.66
(Author, Author) = 0.66

Itération 2



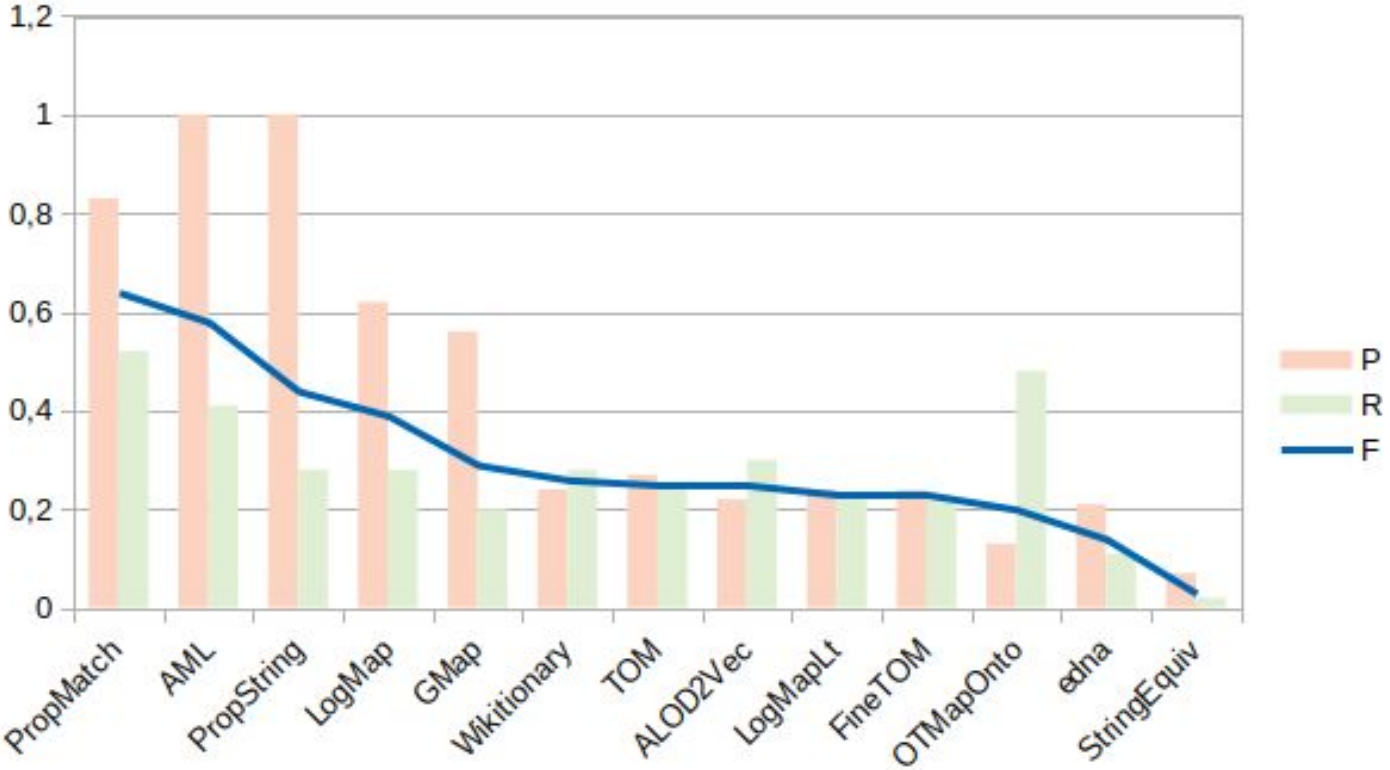
Évaluation

- Deux jeux de données d'OAEI 2021
  - **Conférence** : l'évaluation ra1-M2 (alignements de référence entre propriétés)
    - 21 paires d'alignements entre 7 ontologies
  - **Knowledge Graph**
    - 8 graphes de connaissances sont alignés en 5 paires
- Les propriétés sont représentées différemment dans ces deux jeux de données.

Progression des performances

Description	P	R	F
PropString	**1.00**	.28	.44
Utilisation des plongements dans la similarité des domaines	.84	.35	.49
Addition d'inverses	.81	.39	.53
Similitude de plongements des phrases appliquée aux étiquettes des propriétés	.68	.41	.51
Filtre des mots répétés et les adjectifs dans les étiquettes des portées	.71	.48	.57
similarité des domaines pour les propriétés précédemment alignées	.73	.52	.61
Limitation de la cardinalité (1-1)	**.83**	**.52**	**.64**

Résultats en Conférence



Résultats en Knowledge Graph

Paire	mcu-marvel			malpha-mbeta			malpha-stexpand			starwars-swg			starwars-swtor		
	P	R	F-m	P	R	F-m	P	R	F-m	P	R	F-m	P	R	F-m
AMD	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
ATMatcher	.91	**.91**	**.91**	.98	**.92**	**.95**	.95	**.95**	**.95**	**1.0**	**1.0**	**1.0**	**1.0**	**.98**	**.99**
BaselineLabel	**1.0**	.36	.53	**1.0**	.34	.51	**.97**	.68	.80	**1.0**	**1.0**	**1.0**	**1.0**	**.98**	**.99**
KGMatcher	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
LogMap	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
LSMatch	.82	.82	.82	.62	.58	.60	.62	.61	.62	.72	.65	.68	.88	.79	.83
Matcha	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
PropMatch	.13	.36	.19	.19	.45	.26	.16	.32	.21	.11	.35	.16	.21	.52	.30

Discussion

- PropMatch a obtenu les meilleurs résultats en Conférence
- Pour Knowledge Graph, PropMatch est capable de trouver des alignements alors que des systèmes très connus tels que LogMap et AMD n'en ont trouvé aucun.
- PropMatch est capable de gérer différentes représentations de propriétés.

Conclusions et travaux futurs

- L'utilisation des plongements peut aider à augmenter le rappel du système.
- Cependant, des recherches supplémentaires sont encore nécessaires pour étendre l'utilisation des plongement à tous les calculs de similarité dans le système (pas seulement en repli de TF-IDF)
- Des modèles plus robustes (graphe) doivent être utilisés afin d'être appliqués directement aux constructions complexes présentes dans les domaines et portée de propriété.

Merci de votre attention !

L'existant : PropString

- Proposé par (Cheatham, 2014)
- Basé uniquement sur des mesure lexicales
- TF-IDF pour mesurer la similarité lexicale entre les **domaines** (rdfs:domain) et entre les **portées** (rdfs:range).
- Pour les étiquettes (rdfs:label) de propriété, la similarité du '**concept de base**' est utilisée.
    - premier verbe d'au moins quatre caractères ou un nom avec des adjectifs dans les étiquettes de propriété. (hasName -> Name)
    - TF-IDF avec JaroWinkler est utilisée comme métrique de similarité
- Une correspondance entre les **deux propriétés** est envisagée si la **similarité minimale entre le domaine, la portée et le concept de base dépasse le seuil.**