Techniques d'alignement d'onologies

Sami Zghal
Université de Jendouba
sami.zghal@planet.tn, marouen.kachroudi@fst.rnu.tn

2012-2013

Plan

Problèmes d'hétérogénéité

1. Définition
2. Hétérogénéité syntaxique
3. Hétérogénéité terminologique
4. Hétérogénéité conceptuelle
5. Hétérogénéité sémiotique

Définition

Projets importants

- Divergence des domaines que peuvent couvrir les ontologies
- Divergence des formalismes requis pour leurs développement

Définition

Classification de l'hétérogénéité

- Classification se base sur l'étude du décalage sémantique et structurel qui peut exister entre les ontologies (Euzenat et Shvaiko (2007))
- Classification selon le degré l'hétérogénéité selon les niveaux d'interopérabilité sémantique (Euzenat (2001))

Définition

Types d'hétérogénéités (Euzenat et Shvaiko (2007))

- Hétérogénéité syntaxique
- Hétérogénéité terminologique
- Hétérogénéité conceptuelle
- Hétérogénéité sémiotique

Hétérogénéité syntaxique

Hétérogénéité syntaxique

- Se produit quand 2 ontologies sont décrites avec deux langages ontologiques différents
- Se manifeste lors de la comparaison d'un répertoire avec un modèle conceptuel
- Se produit aussi quand 2 ontologies sont modélisées en utilisant des formalismes différents (OWL et Frame Logic)
- Survient au niveau théorique : établir des équivalences entre les primitives de différents langages ontologiques
- Possibilité de traduire les ontologies dans différents langages ontologiques à condition de préserver la signification

Hétérogénéité terminologique

Hétérogénéité terminologique

- Se manifeste dans l'éventualité où deux entités sont référencées par deux noms différents alors qu'elles désignent le même objet
- Cause d'une telle hétérogénéité : revient à l'utilisation de différents langages naturels, ou des sous-langages techniques spécifiques à un domaine de connaissances bien déterminé
- Se manifeste aussi par l'utilisation des synonymies

Hétérogénéité conceptuelle

Hétérogénéité conceptuelle

- Appelée aussi hétérogénéité sémantique (Euzenat (2001)) ou la différence logique (Klein (2001))
- Concerne la diversité des modélisations d'un même domaine
- Utilisation de différents axiomes décrivant les concepts
- Se manifeste aussi lors de l'utilisation de concepts différents
- Klein (2001) et Visser et al. (1998) : évoquent la différence de conceptualisation et d'explication
- Différence de conceptualisation : à travers la différence entre les concepts inclus dans la modélisation
- Différence des explications : expression des concepts
- Visser et al. (1998) : classification précise de ces différences

Hétérogénéité conceptuelle

Différence de conceptualisation (Benerecetti et al. (2001))

- Différence de convergence : survient lorsque deux ontologies décrivent différentes connaissances avec le même niveau de détail pour une unique perspective
- Différence de granularité : se produit quand deux ontologies décrivent le même domaine avec une même perspective mais avec différents degrés d'expression des détails
- Différence de perspectives : se manifeste quand deux ontologies décrivent un même domaine, avec un même degré d'expression des détails mais avec des points de vue et des perspectives différents

Hétérogénéité sémiotique

Hétérogénéité sémiotique

- Appelée aussi hétérogénéité pragmatique (Euzenat et Shvaiko (2007))
- S'intéresse à la manière dont les entités ontologiques sont interprétées par leurs utilisateurs
- Entités ayant les mêmes interprétations sémantiques : peuvent être interprétées de différentes manières par l'Homme
- Différences d'interprétation : dues principalement à la diversité des contextes et des domaines d'application
- Manière de mettre en oeuvre les entités ontologiques influence leurs interprétations
- Ce type d'hétérogénéité reste difficile à détecter par la machine

Plan

Alignement d'ontologies

1. Définitions
2. Similarité

Définitions

Correspondance

Soit 2 ontologies O et O', une correspondance M entre O et O' est un quintuplet : $< id, e, e', R, n >$ tel que :

- id : un identifiant unique de l'élément de mapping
- e et e' : sont des entités de O et O'
- R : une relation (par exemple : équivalence ($\equiv$), généralisation ($\sqsupseteq$), spécialisation ($\sqsubseteq$), disjonction ($\perp$))
- n : une mesure de confiance conteunue dans une stucture mathèmatique (dans l'intervalle [0,1])

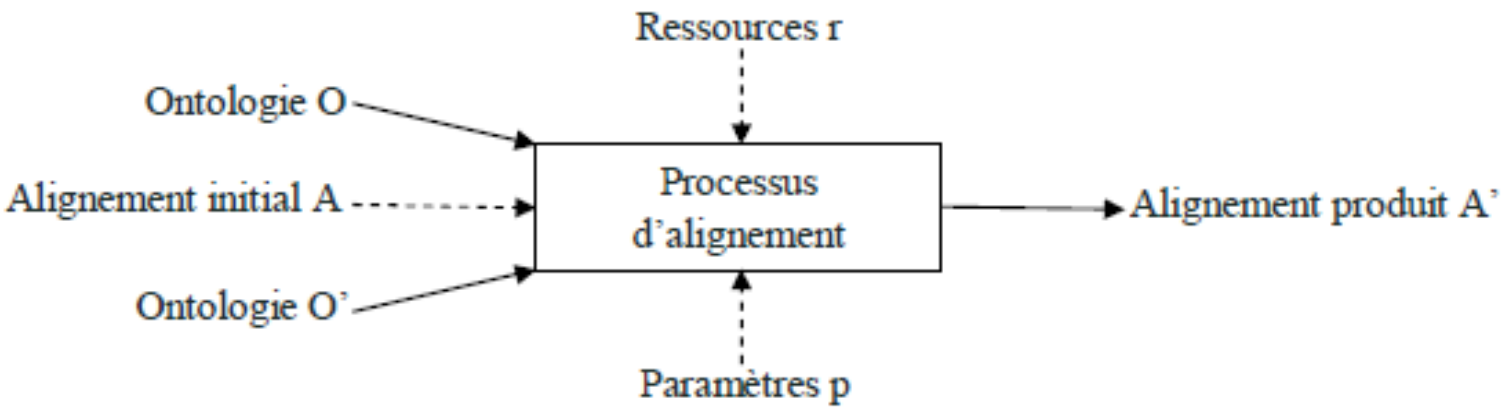
Définitions

Processus d'alignement (Shvaiko et Euzenat (2005))

- Une procédure d'alignement est une fonction notée f
- f prend en entrée :
  1. O et $O^{'}$: 2 ontologies à aligner
  2. A : un alignement initial
  3. p : un ensemble de paramètres
  4. r : un ensemble de ressources externes
- f produit comme résultat un alignement : $A^{'}$
- f est définie comme suit : $A^{'} = f(O, O^{'}, A, p, r)$
- f est définie plus simplement : $A^{'} = f(O, O^{'})$
  - Contexte : fixe
  - Alignement initial, paramètres et ressources : omis

Définitions

Schéma général d'un processus d'alignement d'ontologies



Similarité

Similarité & alignement (1/2)

- Similarité sémantique : proximité sémantique (Bach (2006))
- Calculée entre : documents, termes ou entités
- Grâce à une métrique basée : similitude de leurs significations ou de leur contenu sémantique
- Similarité : quantité qui reflète la force du rapport entre deux objets ou deux caractéristiques
- Similarité réalisée en (Bach (2006)) :
  - Définissant une similarité typologique
  - Employant des ontologies pour définir une distance entre les termes
  - Employant des moyens statistiques pour la corrélation des mots et des contextes textuels

Similarité

Similarité & alignement (2/2)

- Approches d'alignement d'ontologies : similarité sémantique considérée comme celle de la similarité typologique en mathématiques, où une valeur d'une fonction lui est associée, appelée fonction de similarité
- Fonction de la similarité peut changer selon les approches, selon les propriétés souhaitées
- Valeur de cette fonction est souvent normalisée (comprise entre 0 et 1)

Similarité

Mesure de similarité

- Mesure de similarité σ : permet de mesurer le degré de ressemblance entre deux entités
- Soit E un ensemble d'entités, la similarité entre les paires d'entités de cet ensemble est définie par la fonction $\sigma : E \times E \longrightarrow \Re$ tel que $\forall x, y, z \in E$:
  - $\sigma(x, y) \geq 0$: positivité
  - $\sigma(x, x) \geq \sigma(y, z)$: maximalité
  - $\sigma(x, y) = \sigma(y, x)$: symétrie
  - $\sigma(x, y) \leq \infty$: finitude

Similarité

Mesure de dissimilarité

- Mesure de dissimilarité δ : permet de mesurer le degré de différence entre deux entités
- Soit E un ensemble d'entités, la similarité entre les pairs d'entités de cet ensemble est définie par la fonction $\delta : E \times E \longrightarrow \Re$ tel que $\forall x, y, z \in E$:
  - $\delta(x, y) \geq 0$: positivité
  - $\delta(x, x) = 0$: minimalité
  - $\delta(x, y) = \delta(y, x)$: symétrie
  - $\delta(x, y) = \delta(y, z) \Longrightarrow \delta(x, y) = \delta(x, z)$: transitivité
  - $\delta(x, y) \leq \infty$: finitude

Similarité

Mesure de distance

- Distance mesure la dissimilarité entre deux entités : inverse de la similarité
- Distance entre deux entités est petite lorsque la valeur de la fonction de la similarité de deux entités est élevée et vice-versa
- Distance $\delta : E \times E \longrightarrow \Re$, est une fonction de dissimilarité respectant la définitivité et l'inégalité triangulaire telles que $\forall x, y, z \in E$:
  - $\delta(x, y) = 0 \Longleftrightarrow x = y$: définitivité
  - $\delta(x, y) + \delta(y, z) \geq \delta(x, z)$: inégalité triangulaire

Similarité

Normalisation

- Normalisation permet de combiner ces mesures dans des formules et d'obtenir d'autres mesures agrégées qui sont à leurs tours normalisées et peuvent être comparées
- Similarité et dissimilarité normalisée : sont notées respectivement $\bar{\sigma}$ et $\bar{\delta}$
- Ces mesures vérifient la règle de la complémentarité : $\bar{\sigma} + \bar{\delta} = 1$
- Une mesure de similarité est dite *normalisée*, si les valeurs calculées par cette mesure ne peuvent varier que dans un intervalle de 0 à 1

Plan

Classification des techniques d'alignement

1. Techniques de base d'alignement d'ontologies
2. Stratégies d'alignement
3. Évaluation d'alignement

Techniques de base d'alignement d'ontologies

1. Méthodes terminologiques
2. Méthodes structurelles
3. Méthodes extensionnelles
4. Méthodes sémantiques

Techniques de base d'alignement d'ontologies

Méthodes terminologiques

- Comparent les chaînes de caractères : déduire la similarité (ou dissimilarité) en exploitant les relations d'hyperonomie ou d'hyponymie
- Certaines méthodes se basent sur la comparaison des chaînes de caractères : méthodes syntaxiques
- Autres méthodes ont recours à une base de données lexicale (sous forme de réseau sémantique) : méthodes linguistiques

Techniques de base d'alignement d'ontologies

Méthodes syntaxiques (1/3)

- Comparent les termes ou les chaînes de caractères ou bien les textes, des entités à aligner
- Permettent de calculer la valeur de la similarité des entités textuelles
- Méthodes comparent des termes en se basant sur les caractères contenus dans ces termes
- Méthodes utilisent les distances basées sur les tokens

Techniques de base d'alignement d'ontologies

Méthodes syntaxiques (2/3)

- Méthodes se basant sur la comparaison de chaînes de caractères
- Analysent la structure de ces chaînes : calculer des mesures de similarité.
- Exploitent l'ordre des caractères dans la chaîne, le nombre d'apparitions d'une lettre dans une chaîne, etc.
- Distance de HAMMING, similarité de JACCARD, distance d'édition, distance de LEVENSHTEIN, etc.

Techniques de base d'alignement d'ontologies

Méthodes syntaxiques (3/3)

- Plusieurs mesures basées sur les tokens
- Exemple : similarité de DANG, distance de JENSEN-SHANNON, distance de FELLEGI-SUNTER, etc.
- Mesures à base des tokens sont les plus utilisées par les méthodes d'alignement

Techniques de base d'alignement d'ontologies

Méthodes linguistiques (1/4)

- Similarité entre 2 entités : représentées par des termes peut aussi être déduite en les analysant à l'aide des méthodes linguistiques
- Méthodes linguistiques : permettent de déterminer la similarité entre deux entités
- Prennent en charge les propriétés expressives et productives du langage naturel qui peuvent être intrinsèques ou extrinsèques

Techniques de base d'alignement d'ontologies

Méthodes linguistiques (2/4)

- Similarité entre 2 entités : représentées par des termes peut aussi être déduite en les analysant à l'aide des méthodes linguistiques
- Méthodes linguistiques : permettent de déterminer la similarité entre deux entités
- Prennent en charge les propriétés expressives et productives du langage naturel qui peuvent être intrinsèques ou extrinsèques

Techniques de base d'alignement d'ontologies

Méthodes linguistiques (3/4)

Propriétés intrinsèques

- Propriétés morphologiques ou syntaxiques
- Un même concept (ou entité) peut être décrit par plusieurs termes (synonymie) ou par plusieurs variantes d'un même terme
- Cherchent la forme canonique ou représentative d'un mot ou d'un terme (lemme) en exploitant ses variantes linguistiques (lexème)
- Similarité entre 2 termes est mesurée en comparant leurs lemmes

Techniques de base d'alignement d'ontologies

Méthodes linguistiques (4/4)

Propriétés extrinsèques

- Exploitent des ressources externes telles que des dictionnaires ou des vocabulaires
- Calculent la valeur de similarité entre 2 termes en employant des ressources externes
- Ressources regroupent les dictionnaires, les lexiques ou les vocabulaires
- Similarité entre deux termes est calculée en exploitant les liens sémantiques existants dans ces ressources externes
- Liens regroupent les synonymes (pour l'équivalence), des liens d'hyponymes ou d'hyperonymes (pour la subsomption)

Techniques de base d'alignement d'ontologies

Méthodes structurelles

- Déterminent la similarité entre 2 entités en fonction des informations structurelles
- Entités sont reliées entre elles par des liens sémantiques ou syntaxiques
- Liens forment ainsi une hiérarchie ou un graphe d'entités
- Méthodes structurelles se subdivisent en 2 familles :
  1. Méthodes structurelles internes
  2. Méthodes structurelles externes

Techniques de base d'alignement d'ontologies

Méthodes structurelles internes

- Aussi nommées les approches basées sur les contraintes
- Utilisent les informations contenues dans les structures internes des entités pour le calcul de la similarité
- Informations regroupent : co-domaine, cardinalité des attributs, caractéristiques des attributs (la transitivité, la symétrie, etc.)
- Exploitent la similarité entre les contraintes des 2 structures

Techniques de base d'alignement d'ontologies

Méthodes structurelles externes

- Exploitent les relations existantes entre les entités
- Relations contiennent des relations de subsomption (spécialisation ou is-a) ou de méréologie (part-whole)
- Similarité entre les entités est déterminée en fonction de leurs positions dans leurs hiérarchies
- 2 entités sont considérées similaires alors leurs voisins pourraient être à leur tour similaires

Techniques de base d'alignement d'ontologies

Méthodes extensionnelles

- Calculent la similarite entre 2 entites en fonction de leurs extensions
- Extensions representent les ensembles des instances de 2 entites
- Mesures produisent la similarite de 2 entites en fonction de la similarite entre les deux ensembles de leurs instances
- Mesures se basent sur la comparaison exacte des elements existant dans les deux ensembles
- Valtchev (1999) a propose la similarite basee sur des correspondances
- Similarite entre 2 ensembles est la similarite moyenne des elements dans l'ensemble des correspondances

Techniques de base d'alignement d'ontologies

Méthodes sémantiques

- Approche repose sur les modèles & approche de déduction
- Approches logiques : satisfiabilité propositionnelle (SAT), la SAT modale ou les logiques de descriptions
- Techniques des logiques de description (le test de subsomption) peuvent être employées
- Permettent de vérifier les relations sémantiques entre les entités telles que l'équivalence, la subsomption ou l'exclusion
- Assurent aussi la déduction de la similarité de deux entités

Stratégies d'alignement

1. Composition d'alignement
2. Agrégation de similarité
3. Calcul global de similarité
4. Méthodes d'apprentissage
5. Méthodes probabilistiques

Stratégies d'alignement

Composition d'alignement

- Permet la combinaison des résultats de deux alignements : produire un nouveau alignement
- Composition séquentielle (ou linéaire) permet d'exécuter successivement les algorithmes d'alignement
- Composition parallèle permet de combiner plusieurs résultats produits par une panoplie d'algorithmes individuels

Stratégies d'alignement

Agrégation de similarité

- Permet de calculer la similarité composée
- Calculer plusieurs similarités entre les objets des entités de chaque ontologie
- Similarités doivent être agrégées dans le but de déterminer une similarité entre les entités
- Calcul de la similarité entre deux classes nécessite une agrégation, sous forme d'une seule similarité, des similarités obtenues à partir de leurs noms, des super-classes, des instances et de leurs propriétés

Stratégies d'alignement

Calcul global de similarité

- Se réalise en prenant en considération les noeuds voisins
- Similarité peut évoquer les ontologies en entier et la valeur finale de la similarité peut dépendre de toutes les entités appartenant aux ontologies à aligner
- Circuits au niveau d'une ontologie peuvent aussi engender des dépendances circulaires (Valtchev (1999))
- calcul de la similarité dans le cas de la dépendance circulaire ne se fait pas d'une manière locale
- Résolution de ce problème se réalise par l'intermédiaire d'un calcul itératif de la distance ou de la similarité en se référant au niveau de chaque étape à la valeur de l'étape précédente

Stratégies d'alignement

Méthodes d'apprentissage

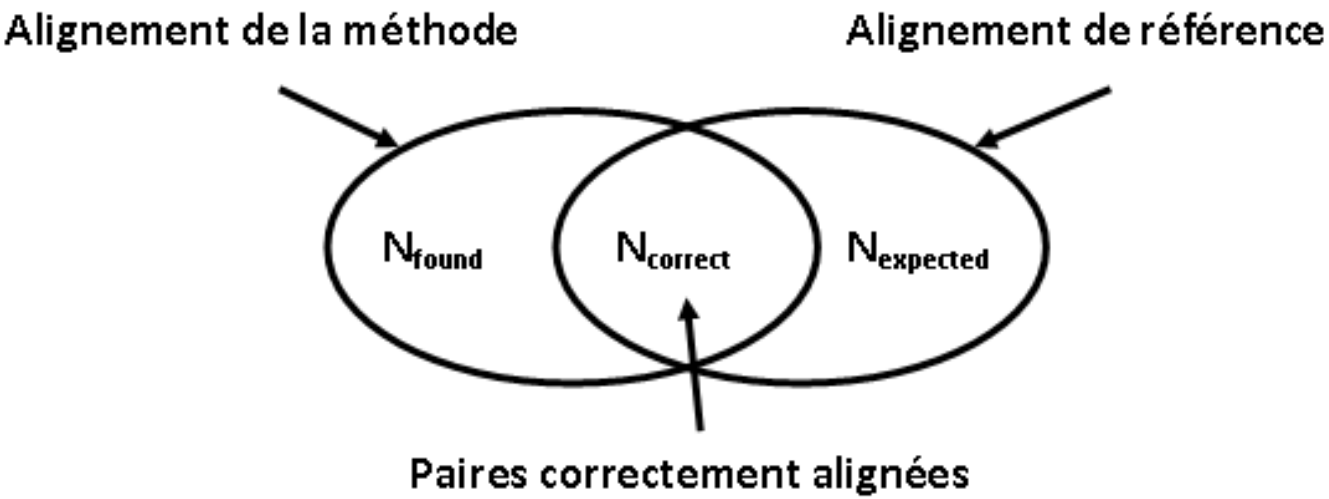
- Se basent sur le fait d'apprendre à partir des alignements corrects et des alignements incorrects
- Méthodes s'appuyant sur l'apprentissage automatique opèrent généralement en 2 phases : phase d'apprentissage et phase de classification ou d'alignement
- Première phase : données utilisées pour l'apprentissage sont générées par un alignement manuel
- Seconde phase : alignements appris sont utilisés pour l'alignement de nouvelles ontologies

Stratégies d'alignement

Méthodes probabilistes

- Méthodes probabilistes peuvent être utilisées, d'une manière similaire aux méthodes d'apprentissage, pour l'alignement d'ontologies (Euzenat et Shvaiko (2007))
- Exploitées pour la combinaison des aligneurs
- Réseaux bayesiens constituent un exemple typique des méthodes probabilistes

Évaluation d'alignement

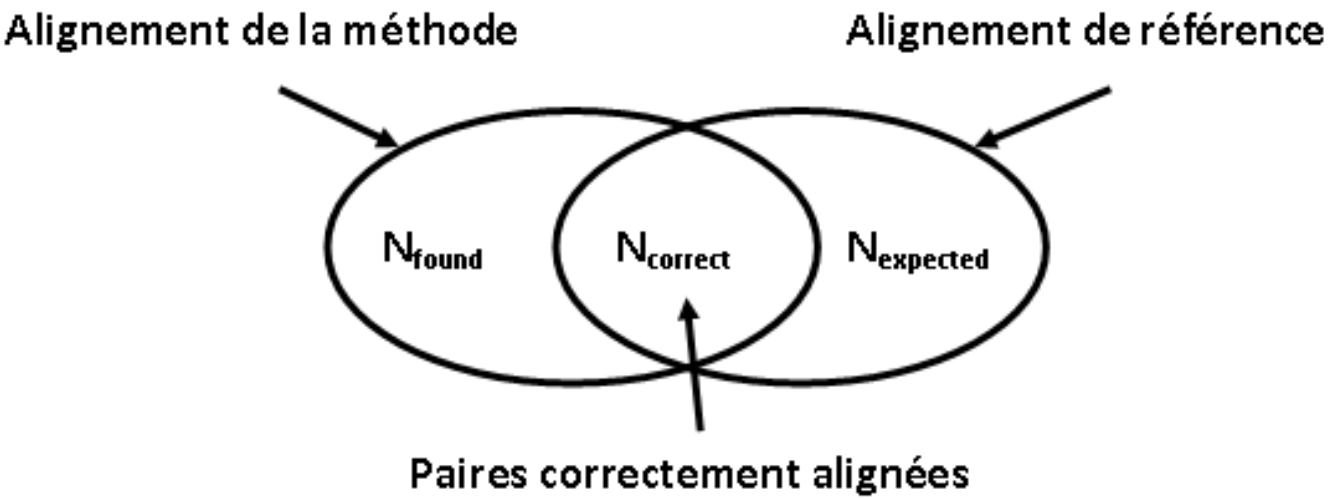


Précision

Ensemble des vraies correspondances parmi celles trouvées

$$Précision = \frac{|N_{correct}|}{|N_{found}|}$$

Évaluation d'alignement

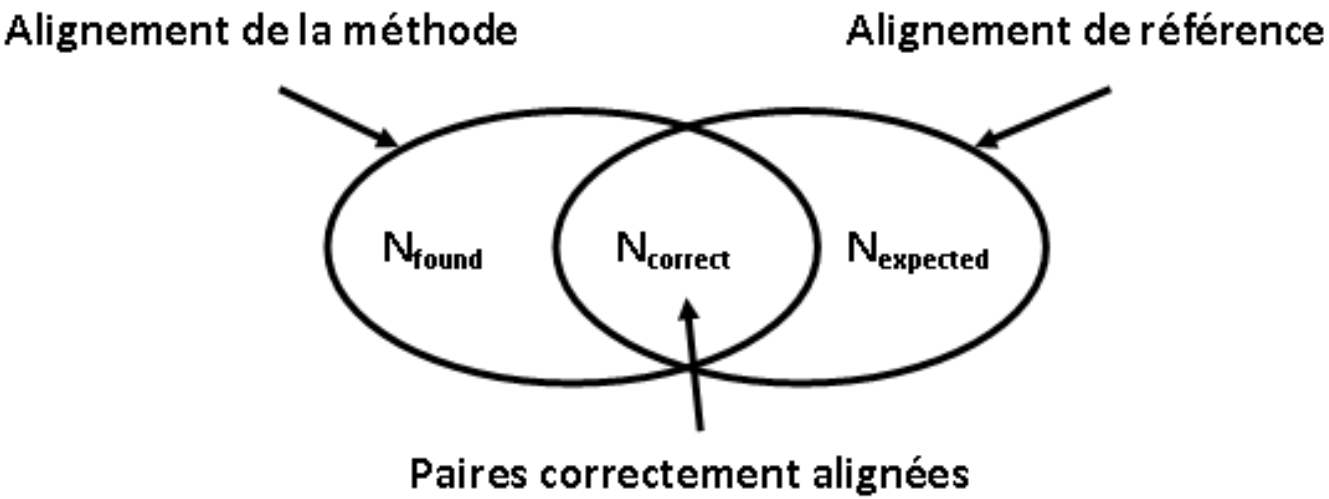


Rappel

Ensemble des vraies correspondances trouvées

$$Rappel = \frac{|N_{correct}|}{|N_{expected}|}$$

Évaluation d'alignement



Fallout

Pourcentage d'erreurs obtenues au cours du processus d'alignement

$$Fallout = \frac{|N_{found}| - |N_{correct}|}{|N_{found}|} = 1 - précision$$

Évaluation d'alignement

Expérimentation

Bases de *Benchmark* de l'OAEI (Ontology Alignment Evaluation Initiative - Campaign, http://oaei.ontologymatching.org/) :

- 2 ontologies à aligner
- Alignement de référence
- Calcul des métriques d'évaluation
- Comparaison entre les méthodes

Plan

Cahier des charges Web et étude préalable

1. Définition du projet
2. Pilotage du projet
3. Différents types de projets et les intervenants