Extraction et validation par croisement des relations d'une ontologie de domaine

Lobna Karoui*, **
* Laboratoire Epitech de Recherche en Informatique Appliquée "L.E.R.I.A", 24 rue Pasteur 94270 Le Kremlin Bicêtre, France
**Supelec, Plateau de Moulon, Gif-sur-Yvette, France
Lobna.Karoui@supelec.fr

1 Introduction

Face à de grandes quantités de documents web, notre objectif est d'extraire et de valider semi-automatiquement des relations d'un domaine. Dans l'état de l'art, l'extraction des relations a été faite soit par une approche statistique, une approche linguistique ou une approche hybride. De plus, l'intérêt a été toujours porté sur un voire deux types de relations. A contrario, notre objectif est d'extraire des relations de différents types en combinant des analyses de textes et en considérant les caractéristiques des mots. Dans cet article, nous avons défini un algorithme contextuel de découverte de relations qui combine différentes analyses (lexicale, syntaxique et statistique) pour définir des processus complémentaires qui assurent l'extraction de relations variées et pertinentes. Notre algorithme établit des opérations de croisements entre analyses afin de pouvoir valider certaines relations. Les relations valides, comme celles invalides, seront présentées à l'expert du domaine mais séparément.

2 La découverte des relations

La notion de contexte. Pour l'extraction des relations, nous souhaitons trouver les mots qui sont reliés au mot étudié. Donc, nous cherchons des contextes qui contiennent ces mots reliés. Pour cela, nous avons défini différents contextes et nous les avons catégorisés en quatre types: le contexte structurel, le contexte linguistique (centré autour du verbe, globalement syntaxique et lexical), le contexte documentaire (paragraphe) et le contexte fenêtre (avec un degré de proximité). Notre approche utilise toutes ces analyses afin d'extraire de nouvelles relations (en plus de celles existantes dans la hiérarchie) et de les valider automatiquement.
L'algorithme contextuel de découverte des relations. Il applique différents types d'analyses pour extraire et évaluer les relations. Il dépend de certains paramètres comme le degré de confiance (DC), NO est le pourcentage d'occurrences de mots dans le corpus (NO) et FN est la fréquence normalisée des mots dans le corpus (FN). Ces paramètres sont utilisés lors du filtre statistique ainsi que la validation. Le DC doit être défini par l'utilisateur vu qu'il explique sa confiance en l'application. Par contre, NO et FN peuvent être définis soit par l'expert du domaine, soit par le système en les déduisant de la valeur de DC ou par défaut (valeur définie par le concepteur du système). Dans le cas où le système est utilisé pour calculer les valeurs de NO et FN, si la valeur de DC est supérieure à 50% leurs valeurs (par défaut) seront maintenues, sinon elles seront multipliées par deux. Notre algorithme catégorise quatre types de relations extraites : valides, invalides, déduites et étiquetées. Une relation valide est celle qui est récupérée après une opération de croisement entre analyses. Une relation invalide est celle qui n'a pas été retrouvé dans deux analyses.

Notre algorithme est composé de cinq étapes. Une première étape applique les différentes analyses pour extraire les relations. Une seconde étape applique un filtre interne pour éliminer les relations qui représentent les liaisons des mots à l'intérieur des classes validées. L'étape trois applique un filtre par croisement des relations résultantes des différentes analyses. Nous proposons deux types de croisements complets (qui nécessitent que la relation existe dans les deux analyses pour qu'elle soit retenue) pour la première étape de validation : un croisement au sein de l'analyse statistique. Ce croisement est fait entre les relations structurées et les relations paragraphes vu qu'une structure telle que définie dans notre démarche (contexte structurel) n'est pas systématiquement incluse dans un paragraphe. D'où l'intérêt de recueillir ces relations qui se trouvent dans les deux résultats de nos contextes de même nature ; un croisement hybride réservé pour les relations provenant de l'analyse fenêtre par proximité et celles des analyses syntaxiques et lexicales. La quatrième étape prend en compte l'ensemble des relations invalides et applique un filtre statistique. Ce dernier est fait en définissant la valeur de deux paramètres à savoir le nombre d'occurrences NO et la fréquence normalisée FN. L'étape 5 et 6 s'occupent respectivement d'établir les validations par degré de confiance et les déductions de nouvelles relations à partir de l'existant et d'étiqueter ces relations qu'elles soient valides, invalides ou déduites.

Expérimentations. Après avoir appliqué notre algorithme sur un corpus de 565 documents HTML en langue française relatif au domaine du tourisme, nous avons pu extraire: relations centrées autour du verbe (2251) ; relations globalement syntaxiques (34439) ; relations lexicales (5793) ; relations paragraphe (72476) ; relations structurelles (16966) ; relations fenêtres (206010). Par la suite, nous avons établit deux types de croisements à savoir un croisement entre les relations structurelles et paragraphes, et un second entre les relations fenêtres et lexicales. Le premier croisement nous a permis de retenir 372 relations (Hôtellerie/ hébergement, Réservation/hébergement, Camping/dormir). Quant au second croisement, nous avons pu avoir 268 relations (Catholicisme/christianisme, Ethnographie/paléontologie), sachant que dans les deux croisements nous avons supprimé certaines relations contenant des noms propres afin de minimiser le bruit. Après l'étape de filtre statistique, nous n'avons pas pu retenir des relations valides sur celles lexicales, globalement syntaxique et centrée autour du verbe vu que la relation la plus récurrente ne dépasse pas les 20 fois ; ce qui est largement loin de nos critères définis. Par contre, selon notre algorithme, pour les relations fenêtres (Activité/sport, Nautique/sport, Patrimoine/histoire, Plonger/sport) et structurelles (Casino/divertissement, Festival/musique, Vigne/vignoble), nous avons obtenu respectivement 24818 et 15257 relations validées. Pour les relations paragraphe, le résultat des validations a été négatif. Les relations qui n'ont pas été validées tout au long de notre démarche seront les relations invalides. Celles-ci seront présentées à l'expert en cas de besoin.

Summary

In this research, we focus on extracting relations among concepts in order to build a domain ontology. For this, we define a contextual relation discovery algorithm that applies different textual analyses in order to extract, deduce, label and validate the domain relations. Our algorithm is based on a rich contextual modelling that takes into account the document structure and strengthens the term co-occurrence selection, a use of the existent relations in the concept hierarchy and a stepping between the various extracted relations to facilitate the evaluation made by the domain experts. Our main perspective is using these relations for the concept hierarchy evaluation and enhancement.