Apport des ontologies pour le calcul de la similarité sémantique au sein d'un système de recommandation

LE Ngoc Luyen[1,2], Marie-Hélène ABEL[1], Philippe GOUSPILLOU[2]

[1] Université de technologie de Compiègne, CNRS, Heudiasyc (Heuristics and Diagnosis of Complex Systems), CS 60319 - 60203 Compiègne Cedex, France

[2] Vivocaz, 8 B Rue de la Gare, 002200, Mercin-et-Vaux, France

Résumé

La mesure de la parenté ou ressemblance sémantique entre les termes, les mots, ou les données textuelles joue un rôle important dans différentes applications telles que l'acquisition de connaissances, les systèmes de recommandation, et le traitement du langage naturel. Au cours des dernières années, de nombreuses ontologies ont été développées et utilisées pour structurer les connaissances au sein des systèmes d'information. Le calcul de similarité sémantique à partir d'ontologie s'est développé et selon le contexte est complété par d'autres méthodes de calcul de similarité. Dans cet article, nous proposons et appliquons une approche pour le calcul de la similarité sémantique basée sur l'ontologie au sein d'un système de recommandation.

Mots-clés

Similarité sémantique, Ontologie, Système de Recommandation, Plongement de mots

Abstract

Measurement of the semantic relatedness or likeness between terms, words, or text data plays an important role in different applications dealing with textual data such as knowledge acquisition, recommender system, and natural language processing. Over the past few years, many ontologies have been developed and used as a form of structured representation of knowledge bases for information systems. The calculation of semantic similarity from ontology has developed and depending on the context is complemented by other similarity calculation methods. In this paper, we propose and carry on an approach for the calculation of ontology-based semantic similarity using in the context of a recommender system.

Keywords

Semantic Similarity, Ontology, Recommender System

1 Introduction

Avec le développement d'Internet et du World Wide Web, les sites Web ou applications e-commerce contiennent des de données textuelles structurées, semi-structurées ou non structurées qui ne cessent d'augmenter. La recherche d'informations sur ces sources de données permet d'améliorer certaines tâches telles que la recherche, le classement. Plus précisément, le calcul de la similarité sémantique montre à quel point deux concepts, deux termes ou deux entités sont proches, sur la base de la comparaison des liens taxonomiques et des propriétés sémantiques [32].

En structurant et en organisant un ensemble de termes ou de concepts au sein d'un domaine de manière hiérarchique et en modélisant les relations entre ces ensembles de termes ou de concepts à l'aide d'un descripteur relationnel, une ontologie permet de spécifier un vocabulaire conceptuel standard pour représenter les entités du domaine [30]. Diverses applications utilisant des ontologies décrivent des termes, des entités et quantifient les relations entre eux [29, 18]. Ces dernières années, l'utilisation d'ontologies est devenue plus populaire dans les systèmes de recommandation [15, 27]. Ainsi, le calcul de similarité sémantique basé sur l'ontologie permet d'améliorer la précision des tâches d'appariement, de recherche et de classement sur des éléments ou des profils d'utilisateurs.

Une ontologie peut être représentée selon différents modèles : (1) Le modèle de représentation en triplet définit une ontologie comme un ensemble de triplets $\langle sujet, prédicat, objet \rangle$ où la relation entre le sujet et l'objet est exprimée par le prédicat. Le sujet est une ressource [1], le prédicat est une propriété d'une ressource, et l'objet identifie la valeur de la propriété de la ressource. L'objet d'un triplet peut contenir une autre ressource ou un littéral. (2) Le modèle de représentation graphique considère qu'une ontologie est un graphe orienté où les nœuds représentent les ressources ou les littéraux tandis que les arcs représentent les propriétés nommées. (3) Le modèle de représentation orienté objet définit une ontologie comme un ensemble d'objets, dans lequel les objets correspondent aux ressources et les variables d'instance de l'objet correspondent aux propriétés de ressources [8].

En considérant une ontologie comme un ensemble de triplets, les approches courantes de calcul de similarité sé-

1. Une ressource peut être une classe, un instance, un concept, un nombre, un chaîne de caractères [8]

mantique basées sur une ontologie présentent deux points faibles. Le premier point faible concerne la mesure de similarité qui se calcule soit entre objets, soit entre objets et prédicats [32, 21]. Le calcul basé sur les objets n'utilise pas les informations du sujet, alors qu'elles peuvent contenir des informations contextuelles du triplet intéressantes pour la comparaison. Le second point faible concerne la distinction du type des objets : textuels ou numériques [24]. Le calcul de similarité entre des objets numériques consiste en un simple calcul arithmétique. Le calcul de similarité entre des objets textuels est basé sur la fréquence des mots composant les objets textuels à comparer. Ce calcul ne tient pas compte de la dépendance sémantique entre ces mots. Cette dernière peut être une richesse pour la comparaison. Dans le cadre de nos travaux, nous visons le traitement de ces deux points faibles afin de définir un calcul de similarité sémantique plus précis au sein d'un système de recommandation.
Le reste de cet article est organisé comme suit. Tout d'abord, la section 2 présente des travaux de la littérature sur lesquels s'appuie notre approche. La section 3 présente nos contributions principales sur la construction du système de recommandation exploitant la mesure de similarité entre des ensembles de triplets. Avant de conclure, nous testons nos travaux dans la section 4 à partir d'un cas expérimental traitant de l' achat/vente de véhicules d'occasion. Enfin, nous concluons et présentons les perspectives.

2 Travaux de la littérature

2.1 Apport des ontologies

Dans le contexte du partage des connaissances, une ontologie est une description formelle et explicite des connaissances partagées qui consiste en un ensemble de concepts dans un domaine et les relations entre ces concepts [13]. L'utilisation des ontologies facilite le partage et la réutilisation des connaissances entre les personnes et les applications largement diffusées. L'usage des ontologies permet [2] :

- L'organisation des données : une ontologie est construite sur la base des structures naturelles de l'information en permettant de visualiser les concepts et leurs relations.
- L'amélioration de la recherche : au lieu de rechercher par mot-clé, la recherche sur les ontologies peut renvoyer des synonymes à partir des termes de la requête.
- L'intégration de données issues de différentes sources, différents langages.

Fondamentalement, une ontologie peut être représentée par le langage OWL qui permet de contraindre les faits RDF dans un domaine particulier. Un fait RDF est défini par un triplet qui est un ensemble de trois composants : un sujet, un prédicat et un objet. Intuitivement, un triplet $\langle sujet, prédicat, objet \rangle$ exprime qu'un sujet donné a une valeur donnée pour une propriété donnée[2, 23]. Une ontologie représentée en OWL possède un mécanisme d'inférence ou de raisonnement permettant de déduire les connaissances supplémentaires.
La similarité sémantique basée sur l'ontologie fait référence à la proximité de deux termes [2] au sein d'une ontologie donnée. La distance entre deux termes est une représentation vectorielle numérique de la distance entre deux termes l'un de l'autre [20]. Cela permet d'utiliser l'ontologie pour rechercher efficacement des éléments liés ou pour identifier des associations entre des termes.
L'utilisation des ontologies comme une base de connaissance devient de plus en plus populaire dans les tâches de modélisation, d'inférence des nouvelles connaissances, ou de calcul de similarité pour des systèmes de recommandation [11]. Dans la section suivante, nous rappelons les notions de base des systèmes de recommandation et précisons le rôle que peut y jouer une ontologie notamment dans certains domaines.

2.2 Système de recommandation basé sur les ontologies

Le système de recommandation (SdR) est conventionnellement défini comme une application qui tente de recommander les éléments les plus pertinents aux utilisateurs en raisonnant ou en prédisant les préférences de l'utilisateur dans un élément en fonction d'informations connexes sur les utilisateurs, les éléments, et les intéractions entre les éléments et les utilisateurs [22, 19]. En général, les techniques de recommandation peuvent être classées selon 6 principales approches : les SdRs basés sur les données démographiques, les SdRs basés sur le contenu, les SdRs basés sur le filtrage collaboratif, les SdRs basés sur la connaissance, les SdRs sensibles au contexte, et les SdRs hybrides.
Dans plusieurs domaines tels que les services financiers, les produits de luxe coûteux, l'immobilier ou les automobiles, les articles sont rarement achetés et les évaluations des utilisateurs ne sont souvent pas disponibles. De plus, la description des articles peut être complexe et il est difficile d'obtenir un ensemble raisonnable de notes reflétant l'historique des utilisateurs sur un article similaire. Par conséquent, les SdRs basés sur les données démographiques, sur le contenu, et sur le filtrage collaboratif ne sont généralement pas bien adaptés aux domaines dans lesquels les éléments possèdent les caractéristiques mentionnées. Des systèmes de recommandation basés sur les connaissances représentées au moyen d'ontologies sont alors proposés pour relever ces défis en sollicitant explicitement les besoins des utilisateurs pour ces éléments et une connaissance approfondie du domaine sous-jacent pour les mesures de similarité et le calcul des prédictions [17].
Pour améliorer la qualité de la recommandation, les calculs de similarité entre éléments ou le profil utilisateur dans un système de recommandation joue un rôle très important. Ils permettent d'établir une liste de recommandations tenant compte des préférences des utilisateurs obtenues suite aux déclarations des utilisateurs ou bien de leurs interactions. Nous détaillons dans la section suivante les mesures de similarité sémantique entre les éléments au sein d'un système de recommandation.

2. Une terme est utilisé pour exprimer un concept, un sujet, un prédicat, un objet, ou un ensemble de triplets

2.3 Mesure de similarité sémantique

Les avantages de l'utilisation des ontologies consistent en la réutilisation de la base de connaissances dans divers domaines, la traçabilité et la capacité d'utiliser le calcul et l'application à une échelle complexe et à grande échelle [26]. En fonction de la structure du contexte applicatif et de son modèle de représentation des connaissances, différentes mesures de similarité ont été proposées. En général, ces approches peuvent être classées selon quatre stratégies principales [32, 24] : (1) basée sur le chemin, (2) basée sur les caractéristiques, (3) basée sur le contenu de l'information, et (4) la stratégie hybride qui inclut des combinaisons des trois stratégies de base.

En mesurant la similarité sémantique basée sur le chemin, les ontologies peuvent être considérées comme un graphe orienté avec des nœuds et des liens, dans lequel les classes ou les instances sont interconnectées principalement au moyen de relations d'hyperonyme et d'homonyme où l'information est structurée de manière hiérarchique en utilisant la relation 'est-un' [24]. Ainsi, les similarités sémantiques sont calculées en fonction de la distance entre deux classes ou instances. De cette manière, plus le chemin est long, plus les deux classes ou instances sont sémantiquement différentes [32]. Le principal avantage de cette stratégie est la simplicité car elle nécessite un faible coût de calcul basé sur le modèle de graphe et ne nécessite pas les informations détaillées de chaque classe et instance [21]. Néanmoins, le principal inconvénient de cette stratégie concerne le degré de complétude, d'homogénéité, de couverture et de granularité des relations définies dans l'ontologie [32].

Lors de la mesure des similarités sémantiques basées sur les caractéristiques, les classes et les instances dans les ontologies sont représentées comme un ensemble de caractéristiques ontologiques [32, 24]. Les points communs entre les classes et les instances sont calculés en fonction de leur ensemble de caractéristiques ontologiques. De cette manière, l'augmentation de la différence de deux classes ou instances dépend de l'augmentation de nombreuses propriétés partagées et de la diminution des propriétés non-partagées entre elles [34]. L'évaluation de la similarité peut être réalisée en utilisant plusieurs coefficients sur les ensembles de propriétés tels que l'indice de Jaccard [16], le coefficient de Dice [10] ou l'indice de Tversky [33]. L'avantage de cette stratégie est qu'elle évalue à la fois les points communs et les différences d'ensembles de propriétés comparées qui permettent d'exploiter plus de connaissances sémantiques que l'approche basée sur le chemin. Cependant, la limitation est qu'il est nécessaire d'équilibrer la contribution de chaque propriété en décidant la standardisation et la pondération des paramètres sur chaque propriété.

En mesurant les similitudes sémantiques basées sur le contenu de l'information (CI), on utilise le contenu de l'information comme mesure de l'information en associant des probabilités d'apparition à chaque classe ou instance dans l'ontologie et en calculant le nombre d'occurrences de ces classes ou instances dans l'ontologie [32]. De cette manière, les classes ou instances peu fréquentes deviennent plus informatives que les classes ou instances fréquentes. Un inconvénient de cette stratégie est qu'elle exige des ontologies larges avec une structure taxonomique détaillée afin de bien différencier les classes.

Au-delà de la mesure des similarités sémantiques mentionnée ci-dessus, il existe un certain nombre d'approches basées sur des combinaisons des trois principales stratégies. Par exemple, Hu et al. [14] utilisent la combinaison de la stratégie basée sur les caractéristiques et la stratégie basée sur le chemin. Ils utilisent la logique de description pour représenter les caractéristiques des entités et la mesure de similarité cosinus pour calculer une similarité. De leur côté, Batet et al. [5] utilisent l'équation 1 pour calculer la similarité sémantique basée sur les caractéristiques des classes et des instances et l'approche basée sur le contenu de l'information.

$$Sim(c_1, c_2) = -log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \quad (1)$$

où $T(c_i) = \{c_j \in C \mid c_j$ est la superclasse de $c_i\}$, C contient la hiérarchie complète des concepts ou la taxonomie de l'ontologie.

Dans nos travaux, nous avons choisi de travailler sur la représentation d'une ontologie au moyen de triplets. Un triplet RDF comporte trois composants : sujet, prédicat et objet. En particulier, le sujet peut être le nom d'une classe, ou un instance. Le prédicat est le nom d'une propriété d'un classe ou d'un instance. L'objet est une valeur d'une propriété de la classe ou du instance qui peut se séparer en un littéral ou un nom d'une autre classe ou un autre instance. Le nom d'une classe, d'un instance, ou des littéraux sont exprimés via un texte pouvant comporter plusieurs mots. Afin de préparer leur traitement, ces contenus textuels sont vectorisés. Nous précisons dans la section suivante les méthodes que nous avons étudiées à cette fin.

2.4 Représentations vectorielles de mots

Une ontologie est composée de concepts et de relations. Ces éléments sont étiquetés par des textes (un ou plusieurs mots). Pour que les machines comprennent et effectuent des calculs sur ces contenus textuels, il faut les transformer en une représentation numérique en utilisant un corpus textuel [6]. La vectorisation de mots permet de représenter un mot par un vecteur à valeurs réelles et ce vecteur décrit le mieux possible le sens de ce mot dans son contexte. En général, plusieurs techniques sont proposées pour vectoriser un mot telles que celles basées sur la fréquence de mots (e,g. TF-IDF [31]) ou le sac de mots continus (CBOW) ou encore le saut de gramme (Skip-Gram) (e,g. Word2vec [25]).

Le TF-IDF [3] est une mesure statistique basée sur un corpus de documents [4]. Cette technique évalue la pertinence d'un

3. TF-IDF (Term Frequency-Inverse Document Frequency) est noté pour la Fréquence du Terme et la Fréquence Inverse du Document

4. Dans le contexte d'une ontologie, un ensemble de triplets est équivalent un document

mot par rapport à un document dans un corpus de documents. Tout d'abord, on calcule la fréquence relative d'un mot m dans un document d comme suit :

$$tf(m,d) = \frac{f(m,d)}{\sum_{m' \in d} f(m',d)} \quad (2)$$

où $f(m,d)$ dénote le nombre de fois où le mot m apparaît dans le document d, $\sum_{m' \in d} f(m',d)$ dénote le nombre total des mots dans le document d. Ensuite, on mesure la quantité d'informations fournies par le mot m dans le corpus de documents D avec la fréquence inverse du document comme suit :

$$idf(m,D) = log \frac{N}{|d \in D : m \in d|} + 1 \quad (3)$$

où N est le nombre de documents dans le corpus, $|d \in D : m \in d|$ est le nombre de documents où le mot m apparaît. Donc, la valeur de $tf.idf$ du mot m dans le document d au sein du corpus D est définie comme suit :

$$tf.idf(m,d,D) = tf(m,d) \times idf(m,D) \quad (4)$$

Une valeur $tf.idf(m,d,D)$ élevée d'un mot m dans un document d indique que ce mot est pertinent pour ce document au sein du corpus de documents D [31].
La technique de sac de mots continus, CBOW, construit la représentation vectorielle d'un mot m_i via la prédiction de son occurrence et la connaissance des mot avoisinants. Autrement dit, le saut de gramme, Skip-Gram, construit la représentation vectorielle d'un mot m_i en prédisant son contexte d'occurrence. Donc, étant donné une séquence de mots d'apprentissage $\{m_1, m_2, ..., m_T\}$ l'objectif du CBOW est de maximiser la moyenne des log-probabilités :

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} log\ p(m_t | m_{t+j}) \quad (5)$$

Tandis que l'objectif de Skip-gram est de maximiser la moyenne des log-probabilités :

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} log\ p(m_{t+j} | m_t) \quad (6)$$

où c est la taille du contexte. La formulation de Skip-gram définit $p(m_{t+j}|m_t)$ en utilisant la fonction softmax :

$$p(m_{t+j}|m_t) = \frac{exp((v'_{m_{t+j}})^T v_{m_t})}{\sum_{i=1}^{M} exp((v'_{m_i})^T v_{m_t})} \quad (7)$$

où v_{m_t} est la représentation vectorielle d'entrée du mot m_t, et $v'_{m_{t+j}}$, v'_{m_i} sont les représentation vectorielles de sortie du mot m_{t+j}, m_i. M est le nombre de mots dans le dictionnaire du corpus.
Word2vec est l'une des implémentations les plus populaires pour créer un plongement de mots en utilisant une architecture d'apprentissage automatique à l'aide d'un réseau de neurones. Il prédit les mots en fonction de leur contexte en combinant les deux techniques CBOW et Skip-gram [25, 4]. En particulier, la figure 1 illustre l'architecture de Word2vec qui comporte conventionnement trois couches : couche d'entrée, couche cachée, et couche de sortie. D'abord, un dictionnaire de mots avec la taille N est synthétisé à partir d'un corpus de textes. Ensuite, le processus d'apprentissage automatique crée et met à jour les valeurs des poids des matrices $W_{T\times N}$, $W'_{T\times N}$. Une fois l'apprentissage terminée, nous obtenons la matrice $W_{T\times N}$ pour le plongement de mots.

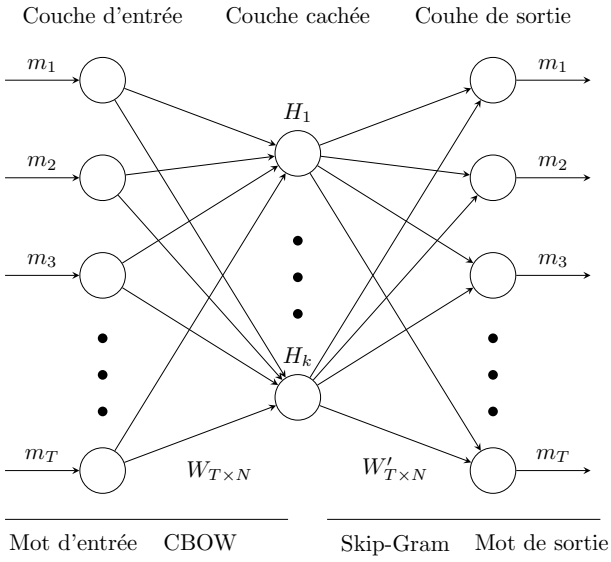

FIGURE 1 – L'architecture du modèle de Word2vec

Plusieurs plongements de mots sont créés en utilisant ce modèle pour des langues différentes [25]. Fauconnier [12], et Hadi et ses collègues [1] implémentent ce modèle à partir des textes en français. D'ailleurs, plusieurs autres travaux ont obtenu de bons résultats dans la conversion d'un mot en une représentation vectorielle tels que Fastext [7], Glove [28], le modèle Transformer avec l'implémentation du BERT [9].

Un plongement de mots entraîné avec de très grands corpus permet d'obtenir rapidement la représentation vectorielle d'un mot. Dans nos travaux, nous avons fait le choix de calculer la mesure de similarité entre deux termes textuels en tenant compte de la combinaison de CBOW et Skip-gram. La similarité entre les deux termes textuels qui se composent de mots différents peut profiter de cette forme de représentation afin de calculer la distance entre eux. Dans la section suivante, nous détaillons notre approche proposée pour mesurer de similarité au sein d'un système de recommandation.

3 Mesure de similarité au sein d'un système de recommandation

3.1 Système de recommandation pour l'achat/vente des véhicules d'occasion

Dans le cadre de nos travaux, nous nous intéressons à l'illustration de la mesure de similarité sémantique sur le système de recommandation basé sur les connaissances représentées au moyen d'ontologies dans une application e-commerce de vente/achat des véhicules d'occasion.
Les données d'un SdR basé sur la base de connaissances représentées au moyen d'ontologies se concentrent sur trois types principaux : les profils de l'utilisateur, les descriptions d'éléments ou les attributs d'éléments, et les intéractions entre les utilisateurs et les éléments. Tout d'abord, les profils d'utilisateur incluent les informations personnelles et les préférences de l'utilisateur sur les éléments de véhicule. Ils peuvent être organisés et être réécrits sous la forme des triplets formellement définis comme suit :

$$G_U = \{a_1^u, a_2^u, ..., a_n^u, \} \tag{8}$$

où a_i^u dénote le triplet $a_i^u = \langle sujet_i, prédicat_i, objet_i \rangle$. Autrement dit, le triplet a_i^u peut aussi s'exprimer comme $\langle ressource_i, propriété_i, énoncé_i \rangle$. Par exemple, "***Louis aime la voiture modèle S de Tesla***". Cette expression naturelle peut se représenter sous la forme de deux triplets différents comme $\langle Louis, aime, la\_voiture\_modèle\_s \rangle$, $\langle la\_voiture\_modèle\_s, est\_fabiquée\_par, Tesla \rangle$. Ensuite, les descriptions de véhicule peuvent également être représentées comme un graphe de connaissance. Elles peuvent être définies selon la même approche :

$$G_V = \{a_1^v, a_2^v, ..., a_n^v, \} \tag{9}$$

où a_i^v dénote le triplet $a_i^v = \langle sujet_i, prédicat_i, objet_i \rangle$ ou $a_i^v = \langle ressource_i, propriété_i, énoncé_i \rangle$. Enfin, lorsqu'un utilisateur effectue une intéraction sur des éléments de description de véhicule en donnant une note, un commentaire ou en ajoutant à une liste de favoris, on marque ces intéractions pour avoir une analyse de l'intention et du comportement de l'utilisateur afin de proposer des recommandations pertinentes. Donc, les intéractions sont définies comme une fonction à plusieurs paramètres :

$$SR : G_U \times G_V \times G_{C_1} \times ... \times G_{C_n} \rightarrow Intéraction \tag{10}$$

où G_U correspond à l'utilisateur, G_V correspond aux éléments de description de véhicule, G_{C_i}s correspond aux informations contextuelles, par exemple : objectifs, locations, temps, ressources [3]. Les ontologies sont développées pour profiler des utilisateurs et modéliser des éléments de description de véhicules [19]. Sur la base de ces ontologies, les données RDFs sont collectées et stockées dans un triplestore interrogeable au moyen de requêtes SPARQL. Des règles peuvent être définies pour déduire ou filtrer les éléments en utilisant les inférences ontologies. Dans ce cas, le SdR basé sur les connaissances comporte les quatre principales tâches suivantes :

- Recevoir et analyser les demandes des utilisateurs à partir de l'interface utilisateur.
- Construire et réaliser des requêtes sur la base de connaissance.
- Calculer des similarités sémantiques entre l'élément de véhicules, le profil utilisateur.
- Classer les éléments correspondant aux besoins de l'utilisateur.

Les mesures de similarité entre les éléments ou le profil utilisateur est une tâche importante pour générer la liste des recommandations la plus pertinente. Le travail s'effectue à partir des données RDFs qui sont organisées sous la forme de triplets $\langle sujet, prédicat, objet \rangle$. Les comparaisons entre deux triplets se limitent souvent aux objets communs ou non communs. Les informations de sujet et prédicat peuvent cependant également fournir des informations importantes sur l'objet lui-même et sa comparaison avec d'autres triplets. Dans la section suivante nous présentons comment dans notre approche nous exploitons ces deux accès à l'information pour calculer les similarités sémantiques entre les triplets d'une base de connaissances.

3.2 Mesure de similarité sémantique entre les triplets

Nous avons choisi de définir une approche hybride tenant compte de la combinaison des approches de calcul de la mesure de similarité sémantique basées sur les caractéristiques et basées sur le contenu de l'information. Le sujet, le prédicat et l'objet dans un triplet contiennent des informations importantes. Un ensemble de triplets permet d'agréger des informations provenant de triplets simples. Par conséquent, la mesure de la similarité sémantique entre ensembles de triplets doit prendre en compte tous les triplets/éléments de chaque ensemble.
La mesure de la similarité sémantique se concentre sur la comparaison de deux ensembles de triplets à partir de tous leurs éléments en les séparant en informations quantitatives et informations qualitatives. D'une part, la comparaison d'objets est réalisée en utilisant la stratégie de similarité sémantique basée sur les propriétés. D'autre part, la comparaison des sujets et des prédicats est effectuée par la stratégie de similarité sémantique basée sur le contenu de l'information.

3.3 Mesure des informations qualitatives

Les informations qualitatives font référence aux mots, aux étiquettes utilisés pour décrire les classes, les relations, et les annotations. Dans un triplet, le sujet et le prédicat expriment une information qualitative. Les objets peuvent contenir des informations qualitatives ou quantitatives. Par exemple, nous avons trois triplets suivants :
$\langle ford\_focus\_4\_2018, la\_boîte\_de\_vitesse, mécanique \rangle$
$\langle ford\_focus\_4\_2020, la\_boîte\_de\_vitesse, mécanique \rangle$
$\langle citron\_c5\_aircross, la\_boîte\_de\_vitesse, mécanique \rangle$
Tous les composants de ces trois triplets sont qualitatifs. L'information du sujet de trois triplets peut être utilisée pour contribuer à la mesure de similarité entre eux. Dans cette section, nous nous concentrons sur la mesure de la simila-

rité sémantique pour les Sujets, Prédicats et Objets Qualitatifs (SPOQ). Nous proposons la même formule pour les trois composants afin de calculer la similarité.

Soient deux SPOQs a_{s1} et a_{s2} dont les vecteurs de mots sont $M_1 = \{m_{11}, m_{12}, ..., m_{1k}\}$ et $M_2 = \{m_{21}, m_{22}, ..., m_{2l}\}$, leur similarité sémantique est définie comme suit :

$$Sim_1(a_{s1}, a_{s2}) = \frac{\sum_{i=1}^{k} \bar{S}(m_{1i}, a_{s2}) + \sum_{j=1}^{l} \bar{S}(m_{2j}, a_{s1})}{k+l} \tag{11}$$

où $\bar{S}(m, a_s)$ dénote la similarité sémantique d'un mot m et d'un SPOQ. La fonction $\bar{S}(m, a_s)$ est formellement calculée comme suit :

$$\bar{S}(m, a_s) = \max_{m_i \in M} \bar{S}(m, m_i) \tag{12}$$

où $m_i \in M = \{m_1, m_2, ..., m_k\}$ est le vecteur de mots de a_s. Chaque mot m_i est représenté par un vecteur numérique. On peut utiliser les techniques introduits dans le section 2.4. L'approche basée sur la fréquence de mots TF-IDF facilite l'obtention de la probabilité d'un mot dans un ensemble de triplets. Cependant, le principal inconvénient de cette approche est qu'elle ne peut pas capturer l'information sémantique du mot et l'ordre du mot dans l'ensemble de triplets parce qu'elle crée le vecteur basé sur la fréquence du mot dans un ensemble de triplets et la collection des ensembles de triplets. Nous proposons l'utilisation de CBOW et Skip-gram avec l'implémentation de Word2vec [25, 1] afin de surmonter cela. Nous calculons finalement la similarité entre deux mots m_i, m_j par la similarité cosinus : $\bar{S}(m_i, m_j) = \frac{m_i . m_j}{\|m_i\| \|m_j\|}$.

3.4 Mesure des informations quantitatives

Les informations quantitatives sont des informations numériques qui sont utilisées pour exprimer l'information de type nominal, ordinal, intervalle, ou ratio. Dans un triplet, l'objet utilise souvent cette forme d'information pour manifester des informations des propriétés pour les classes, concepts de l'ontologie. Par exemple, nous avons des triplets suivants :
$\langle ford\_focus\_4\_2018, a\_le\_kilométrage, 107351 \rangle$
$\langle ford\_focus\_4\_2020, a\_le\_kilométrage, 25040 \rangle$
$\langle citron\_c5\_aircross, a\_le\_kilométrage, 48369 \rangle$
Les objets de ces triplets sont des valeurs numériques. La comparaison entre chiffres s'effectue simplement par les mesures de distance. Afin de comparer deux objets différents, nous utilisons la distance euclidienne entre deux objets. Ainsi, plus la différence entre deux objets est petite, plus la similitude entre eux est grande. Soient deux objets a_{o1} et a_{o1} dont les vecteurs sont $a_{o1} = \{o_{11}, o_{12}, ..., o_{1k}\}$ et $a_{o2} = \{o_{21}, o_{22}, ..., o_{2k}\}$, leur similarité sémantique est définie comme suit :

$$Sim_2(a_{o1}, a_{o2}) = \frac{1}{1 + \sqrt{\sum_{i=0}^{k} (o_i - o_j)^2}} \tag{13}$$

3.5 Mesure des triplets

La comparaison de deux triplets $a_1 = \langle a_{s1}, a_{p1}, a_{o1} \rangle$ et $a_2 = \langle a_{s2}, a_{p2}, a_{o2} \rangle$ est effectuée en fonction du type d'information des objets dans les triplets. Si l'objet contient des informations qualitatives, la similarité sémantique entre a_1 et a_2 est définie comme suit :

$$Sim_I(a_1, a_2) = \frac{1}{N} \sum_{i \in P, \omega \in Q} \omega \times Sim_1(a_{i1}, a_{i2}) \tag{14}$$

où $P = \{s, p, o\}$ correspond aux informations de $sujet$, $predicat$, et $objet$ sous la forme de vecteur de mots. $Q = \{\alpha, \beta, \gamma\}$ est le poids respectifs pour les composants de triplet. N est le nombre de composants de triplet.

Par ailleurs, si l'objet contient des informations quantitatives, la mesure de similarité sémantique des triplets a_1 et a_2 est définie comme suit :

$$Sim_{II}(a_1, a_2) = \frac{1}{N} (\sum_{i \in P, \omega \in Q} \omega \times Sim_1(a_{i1}, a_{i2}) + \gamma \times Sim_2(a_{o1}, a_{o2})) \tag{15}$$

où $P = \{s, p\}$ correspond aux informations de $sujet$ et $predicat$ sous la forme de vecteur de mots. $Q = \{\alpha, \beta\}$ représente les poids respectifs du sujet et du prédicat. Et γ est le poids pour l'objet.

Par conséquent, la similarité sémantique de deux ensembles de triplets $G_1 = \{a_1, a_2, ..., a_g\}$ et $G_2 = \{a_1, a_2, ..., a_g\}$ est calculée sur la base de comparaison de similarité de chaque triplet simple comme suit :

$$Sim(G_1, G_2) = \frac{1}{L} (\sum_{j=0}^{L} Sim_I(a_{1j}, a_{2j})) + \frac{1}{H} (\sum_{j=0}^{H} Sim_{II}(a_{1j}, a_{2j})) \tag{16}$$

où L est le nombre de triplets qui contient les objets qualitatifs. H est le nombre de triplets qui contient les objets quantitatifs.

4 Cas expérimental

Dans cette section nous testons notre approche dans le cas d'une application d'achat/vente de véhicules. Nous mesurons ainsi la similarité sémantique entre deux ensembles de triplets représentant chacun un véhicule. Tout d'abord, la transformation d'un mot à un vecteur est réalisée en utilisant le corpus de mots entraîné qui est développé dans le travail de Hadi et al [1]. Nous avons choisi d'utiliser le modèle CBOW et Skip-gram au lieu de TF-IDF à cause des problèmes concernant la capture de l'information sémantique qui est presque impossible sur la technique de TF-IDF. La figure 2 démontre la distance très proche des mots, groupes de mots dans un même secteur en utilisant les vecteurs entraînés de Word2vec.

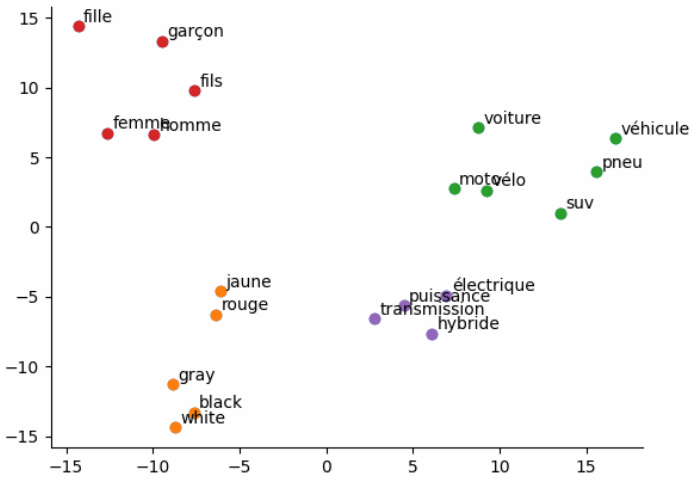

FIGURE 2 – La distance proche entre les mots, groupes qui sont vectorisés par le corpus de mots entraîné

En utilisant l'ontologie, nous pouvons reconstruire la base de connaissances d'un domaine sous une forme lisible par des machines ainsi que les humaines. À partir des ontologies des véhicules développées dans le travail [19], nous réalisons une collection des instances des classes et leurs relations afin de créer un triplestore de données RDF. La figure 3 illustre deux ensembles de triplets représentant deux voitures.

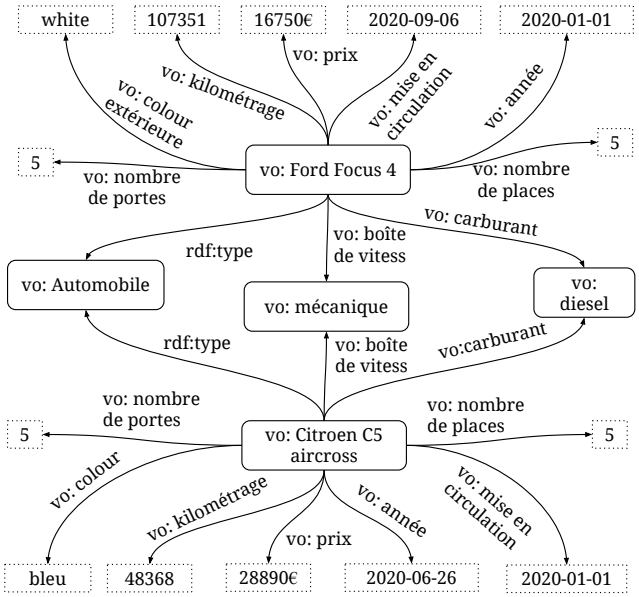

FIGURE 3 – Les données de triplets sont représentées par un graphe (*vo* note pour l'Ontologie de Véhicule)

La description de chaque véhicule est représentée par les informations textuelles et les informations numériques. Notre approche propose de séparer la mesure de la similarité en deux calculs séparés. L'un s'applique aux informations numériques parce qu'elles exigent les calculs simples pour avoir la distance ou la similarité. L'autre s'applique aux informations textuelles lorsqu'une classe où un instance d'une ontologie se composent à partir d'un groupe de mots et chaque mot a des dépendance sémantique avec les autres. En profitant des travaux du domaine de traitement du langage naturel avec les méthodes d'apprentissage profond sur de très grands corpus, les données catégorielles peuvent être représentées dans un vecteur numérique qui contient les relations du mot avec les mots récurant provenant de plusieurs documents en ligne.

		V_1	V_2	V_3	V_4	V_5
V_1	SiLi	1.0				
	N-2	1.0				
	N-1	1.0				
V_2	SiLi	0.57	1.0			
	N-2	0.50	1.0			
	N-1	**0.61**	1.0			
V_3	SiLi	0.50	0.48	1.0		
	N-2	0.48	0.46	1.0		
	N-1	**0.64**	**0.57**	1.0		
V_4	SiLi	0.54	0.49	**0.62**	1.0	
	N-2	0.49	0.47	0.50	1.0	
	N-1	**0.58**	**0.60**	0.59	1.0	
V_5	SiLi	0.54	0.46	**0.69**	0.68	1.0
	N-2	0.48	0.45	0.53	0.52	1.0
	N-1	**0.59**	**0.55**	0.60	**0.71**	1.0

TABLE 1 – La mesure de similarité entre les 5 voitures avec les trois approches différentes

Sur la base des instances collectées, nous réalisons des expérimentations et des évaluations sur trois approches suivantes :

1. N-1 : notre approche proposée principale avec l'utilisation du modèle de Word2vec [1] pour vectoriser les informations qualitatives.
2. N-2 : notre approche avec l'utilisation du modèle de TF-IDF pour vectoriser les informations qualitatives.
3. SiLi : l'approche proposée par Siying Li et ses collèges [21], cette approche hybride combine la stratégie basée sur le contenu et celle sur les caractéristiques mais ne considère que les objets et les prédicats des triplets.

La table 1 affiche les résultats de calcul de la similarité entre les 5 voitures de marques différentes. En particulier, V_1 est le "*Renault captur 2*", V_2 est la marque "*posrche taycan*", V_3 est le "*ford focus 4*", V_4 est la marque "*audi a1 sportback*", et V_5 est le "*citroen c5 aircross*". Les ensembles de triplets de ces voitures sont montrés dans l'appendice A.

En analysant les résultats obtenus et présentés dans la table 1, nous arrivons sur plusieurs conclusions. Premièrement, notre approche **N-1** donne le résultat de calcul de la similarité entre les voitures plus élevé que les autres approches dans 8 sur 10 cas de comparaisons. Toutefois, le résultat de calcul de la similarité de notre approche est moins élevé que l'approche de **SiLi** dans la comparaison de deux cas : entre les voitures V_4, V_3 et entre les voitures V_5, V_3. Deuxièmement, notre approche en utilisant le technique TF-IDF **N-2** pour la représentation vectorielle de mot a obtenu les résultats le plus bas dans tout les cas de comparaison. Cela s'explique par la capacité de capture des informations contextuelles et sémantiques de l'approche Word2vec qui est meilleur que celle de l'approche TF-IDF.
Les expérimentations montrent que notre approche **N-1** a obtenu de bons résultats pour les mesures de similarité entre les ensembles de triplets. L'utilisation du sujet dans la comparaison permet d'ajouter de l'information à la mesure de similarité d'un triplet. Aussi, la distinction contenus textuels et numériques permet d'appliquer la formule appropriée selon le type de contenu. Au final la somme des deux calculs représente la similarité mesurée. Compte tenu des cette distinction, de la prise en compte des triplets contextuels et du calcul à partir des contenus textuels enrichi des dépendances sémantiques entre les mots constituant le texte, la similarité obtenue est plus précise que celles rencontrées dans la littérature [32, 24, 21].

5 Conclusion et perspectives

La mesure de similarité sémantique sur la base de l'ontologie est une tâche importante pour proposer une liste de recommandations pertinentes à un utilisateur. Dans cet article, nous proposons une stratégie hybride qui combine la stratégie basée sur les caractéristiques et basée sur le contenu de l'information. Avec notre approche, afin de ne pas perdre d'information, les trois composants d'un triplet sont considérés dans le calcul de similarité. La distinction de type de données, textuel ou numérique, permet d'effectuer un calcul adapté et plus précis. Nous avons effectué une première expérimentation de notre approche et l'avons comparée à deux autres calculs de similarité. Les résultats obtenus montrent son intérêt. Nous devons maintenant poursuivre nos travaux et en premier lieu effectuer d'autres tests sur des corpus différents et des applications différentes. Nous devons concéder que les mots non considérés dans le corpus entraîné posent un problème. En perspective, la résolution de ce problème ainsi que la construction d'un corpus des triplets entraînés pourraient être des travaux prometteurs dans le futur.

Références

[1] Hadi Abdine, Christos Xypolopoulos, Moussa Kamal Eddine, and Michalis Vazirgiannis. Evaluation of word embeddings from large-scale french web content. 2021.

[2] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Ontologies, RDF, and OWL*, page 143–170. Cambridge University Press, 2011.

[3] Gediminas Adomavicius and Alexander Tuzhilin. *Context-Aware Recommender Systems*, pages 217–253. Springer US, Boston, MA, 2011.

[4] Oussama Ahmia, Nicolas Béchet, Pierre-François Marteau, and Alexandre Garel. Utilité d'un couplage entre word2vec et une analyse sémantique latente : expérimentation en catégorisation de données textuelles. In *Extraction et Gestion des Connaissances : Actes de la conférence EGC*, 2019.

[5] Montserrat Batet, David Sánchez, and Aida Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 44(1) :118–125, 2011.

[6] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Matthijs Douze, and Herve Jegou. Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*, 2016.

[8] Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J Carroll, and Brian McBride. Rdf 1.1 concepts and abstract syntax. *W3C recommendation*, 25(02) :1–22, 2014.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[10] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, 1945.

[11] Yu Du, Sylvie Ranwez, Nicolas Sutton-Charani, and Vincent Ranwez. Apports des ontologies aux systèmes de recommandation : état de l'art et perspectives. In *30es Journées Francophones d'Ingénierie des Connaissances, IC 2019*, pages 64–77, 2019.

[12] Jean-Philippe Fauconnier. French word embeddings, 2015.

[13] N Guarino, P Giaretta, and N Mars. Towards very large knowledge bases : Knowledge building and knowledge sharing. ontologies and knowledge bases : Towards a terminological clarification. n. *Mars. Amsterdam, IOS Press*, pages 25–32, 1995.

[14] Bo Hu, Yannis Kalfoglou, Harith Alani, David Dupplaw, Paul Lewis, and Nigel Shadbolt. Semantic metrics. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181. Springer, 2006.

[15] Mohammed E Ibrahim, Yanyan Yang, David L Ndzi, Guangguang Yang, and Murtadha Al-Maliki. Ontology-based personalized course recommendation framework. *IEEE Access*, 7 :5180–5199, 2018.

[16] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37 :547–579, 1901.

[17] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Knowledge-based recommendation*, page 81–123. Cambridge University Press, 2010.

[18] Rui Jiang, Mingxin Gan, and Peng He. Constructing a gene semantic similarity network for the inference of disease genes. In *BMC systems biology*, volume 5, pages 1–11. Springer, 2011.

[19] Ngoc Luyen Le, Marie-Hélène Abel, and Philippe Gouspillou. Towards an ontology-based recommender system for the vehicle sales area. In Luigi Troiano, Alfredo Vaccaro, Nishtha Kesswani, Irene Díaz Rodriguez, and Imene Brigui, editors, *Progresses in Artificial Intelligence & Robotics : Algorithms & Applications*, pages 126–136, Cham, 2022. Springer International Publishing.

[20] Wei-Nchih Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. In *AMIA annual symposium proceedings*, volume 2008, page 384. American Medical Informatics Association, 2008.

[21] Siying Li, Marie-Hélène Abel, and Elsa Negre. Ontology-based semantic similarity in generating context-aware collaborator recommendations. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 751–756. IEEE, 2021.

[22] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments : A survey. *Decision Support Systems*, 74 :12–32, 2015.

[23] LE Ngoc Luyen, Anne Tireau, Aravind Venkatesan, Pascal Neveu, and Pierre Larmande. Development of a knowledge system for big data : Case study to plant phenotyping data. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, pages 1–9, 2016.

[24] Rouzbeh Meymandpour and Joseph G Davis. A semantic similarity measure for linked data : An information content-based approach. *Knowledge-Based Systems*, 109 :276–293, 2016.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013.

[26] Van Nguyen. Ontologies and information systems : a literature survey. 2011.

[27] Charbel Obeid, Inaya Lahoud, Hicham El Khoury, and Pierre-Antoine Champin. Ontology-based recommender system in higher education. In *Companion Proceedings of the The Web Conference 2018*, pages 1031–1034, 2018.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7) :e1000443, 2009.

[30] M Andrea Rodriguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE transactions on knowledge and data engineering*, 15(2) :442–456, 2003.

[31] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[32] David Sánchez, Montserrat Batet, David Isern, and Aida Valls. Ontology-based semantic similarity : A new feature-based approach. *Expert systems with applications*, 39(9) :7718–7728, 2012.

[33] Amos Tversky. Features of similarity. *Psychological review*, 84(4) :327, 1977.

[34] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, 2005.

A Appendice : Ensembles de triplets des voitures utilisés dans les expérimentations

```
<vo:RC2,rdf:type,vo:Automobile>
<vo:RC2,vo:année,vo:2022-01-01>
<vo:RC2,vo:mis en circulation,vo
    :2022-04-28>
<vo:RC2,vo:contrôle technique,vo:non
    requis>
<vo:RC2,vo:kilométrage,vo:5493>
<vo:RC2,vo:carburant,vo:hybride essence é
    lectrique>
<vo:RC2,vo:boîte de vistesse,vo:
    automatique>
<vo:RC2,vo:coleur extérieure,vo:noir>
<vo:RC2,vo:nombre de portes,vo:5>
<vo:RC2,vo:nombre de places,vo:5>
<vo:RC2,vo:puissance fiscale,vo:5>
<vo:RC2,vo:puissance din,vo:93>
<vo:RC2,vo:Critique d'Air,vo:1>
<vo:RC2,vo:emission de CO2,vo:35>
<vo:RC2,vo:consommation mixte,vo:1.5>
<vo:RC2,vo:norme euro,vo:euro6>
<vo:RC2,vo:fabriquer par,vo:Renault
    occasion>
<vo:RC2,vo:type de véhicule,vo:4x4, SUV &
    Crossover occasion>
<vo:RC2,vo:location,vo:Cher>
```

```
<vo:RC2,vo:price,vo:36580>
```

Listing 1 – V_1 Renault Captur 2 (RC2)

```
<vo:PT,rdf:type,vo:Automobile>
<vo:PT,vo:année,vo:2022-01-01>
<vo:PT,vo:mis en circulation,vo
    :2022-09-10>
<vo:PT,vo:contrôle technique,vo:non requis
    >
<vo:PT,vo:kilométrage,vo:4932>
<vo:PT,vo:carburant,vo:electrique>
<vo:PT,vo:boîte de vistesse,vo:automatique
    >
<vo:PT,vo:coleur intérieure,vo:cuir ivoire
    >
<vo:PT,vo:coleur extérieure,vo:noir metal>
<vo:PT,vo:nombre de portes,vo:4>
<vo:PT,vo:nombre de places,vo:4>
<vo:PT,vo:garranty,vo:20 mois>
<vo:PT,uvso:puissance fiscale,vo:8>
<vo:PT,vo:puissance din,vo:530>
<vo:PT,vo:Critique d'Air,vo:0>
<vo:PT,vo:emission de CO2,vo:0>
<vo:PT,vo:norme euro,vo:euro6>
<vo:PT,vo:fabriquer par,vo:Porsche
    occasion>
<vo:PT,vo:type de véhicule,vo:Berline
    occasion>
<vo:PT,vo:location,vo:Rhône>
```

Listing 2 – V_2 Porsche Taycan (PT)

```
<vo:FF4,rdf:type,vo:Automobile>
<vo:FF4,vo:année,vo:2020-01-01>
<vo:FF4,vo:mis en circulation,vo
    :2020-09-06>
<vo:FF4,vo:contrôle technique,vo:non
    requis>
<vo:FF4,vo:kilométrage,vo:107351>
<vo:FF4,vo:carburant,vo:diesel>
<vo:FF4,vo:boîte de vistesse,vo:mécanique>
<vo:FF4,vo:coleur extérieure,vo:gris foncé
    >
<vo:FF4,vo:nombre de portes,vo:5>
<vo:FF4,vo:nombre de places,vo:5>
<vo:FF4,vo:garranty,vo:12 mois>
<vo:FF4,vo:puissance fiscale,vo:4>
<vo:FF4,vo:puissance din,vo:95>
<vo:FF4,vo:Critique d'Air,vo:2>
<vo:FF4,vo:emission de CO2,vo:89>
<vo:FF4,vo:consommation mixte,vo:4.5>
<vo:FF4,vo:norme euro,vo:euro6>
<vo:FF4,vo:fabriquer par,vo:Ford occasion>
<vo:FF4,vo:type de véhicule,vo:Berline
    occasion>
<vo:FF4,vo:location,vo:Loiret>
<vo:FF4,vo:price,vo:16750>
```

Listing 3 – V_3 Ford Focus 4 (FF4)

```
<vo:AA1,rdf:type,vo:Automobile>
<vo:AA1,vo:année,vo:2018-01-01>
<vo:AA1,vo:mis en circulation,vo
    :2018-09-15>
<vo:AA1,vo:contrôle technique,vo:non
    requis>
<vo:AA1,vo:kilométrage,vo:20211>
<vo:AA1,vo:carburant,vo:diesel>
<vo:AA1,vo:boîte de vistesse,vo:
    automatique>
<vo:AA1,vo:coleur extérieure,vo:bleu>
<vo:AA1,vo:coleur intérieure,vo:noir>
<vo:AA1,vo:nombre de portes,vo:5>
<vo:AA1,vo:nombre de places,vo:5>
<vo:AA1,vo:garranty,vo:12 mois>
<vo:AA1,vo:puissance fiscale,vo:4>
<vo:AA1,vo:puissance din,vo:90>
<vo:AA1,vo:Critique d'Air,vo:2>
<vo:AA1,vo:emission de CO2,vo:101>
<vo:AA1,vo:consommation mixte,vo:3.6>
<vo:AA1,vo:norme euro,vo:euro6>
<vo:AA1,vo:fabriquer par,vo:Audi occasion>
<vo:AA1,vo:type de véhicule,vo:Citadine
    occasion>
<vo:AA1,vo:location,vo:Yvelines>
<vo:AA1,vo:price,vo:23200>
```

Listing 4 – V_4 Audi A1 sportback (AA1)

```
<vo:CC5,rdf:type,vo:Automobile>
<vo:CC5,vo:année,vo:2020-01-01>
<vo:CC5,vo:mis en circulation,vo
    :2020-06-26>
<vo:CC5,vo:contrôle technique,vo:non
    requis>
<vo:CC5,vo:kilométrage,vo:48368>
<vo:CC5,vo:carburant,vo:diesel>
<vo:CC5,vo:boîte de vistesse,vo:mécanique>
<vo:CC5,vo:coleur extérieure,vo:bleu>
<vo:CC5,vo:nombre de portes,vo:5>
<vo:CC5,vo:nombre de places,vo:5>
<vo:CC5,vo:garranty,vo:12 mois>
<vo:CC5,vo:puissance fiscale,vo:6>
<vo:CC5,vo:puissance din,vo:131>
<vo:CC5,vo:Critique d'Air,vo:2>
<vo:CC5,vo:emission de CO2,vo:106>
<vo:CC5,vo:consommation mixte,vo:4.1>
<vo:CC5,vo:norme euro,vo:euro6>
<vo:CC5,vo:fabriquer par,vo:Citroen
    occasion>
<vo:CC5,vo:type de véhicule,vo:4x4, SUV &
    Crossover occasion>
<vo:CC5,vo:location,vo:Yvelines>
<vo:CC5,vo:price,vo:28890>
```

Listing 5 – V_5 Citroen C5 aircross (CC5)