CAMERA-DREAM : Une étude du Web de données dans le contexte d'un projet universitaire

Patrick Giroux, Esther Nicart

CAMERA-DREAM : Une étude du Web de données dans le contexte d'un projet universitaire

Patrick GIROUX[1], Esther NICART[2]

[1] Cassidian, EADS, Parc d'Affaires des Portes B.P. 613, 27106, Val-de-Reuil Cedex
patrick.giroux@cassidian.com
[2] Université de Rouen, 1, rue Thomas Becket, 76821, Mont-Saint-Aignan Cedex
esther.hoare@etu.univ-rouen.fr

Résumé : L'enseignement dispensé dans le cadre du Master Génie Informatique et Logiciel de l'Université de Rouen inclut un projet de grande envergure qui mobilise chaque année tous les étudiants de la promotion. En 2012-2013, ce projet intitulé *CAMERA-DREAM* visait à constituer une base de connaissance consacrée au cinéma et publiable sur le Web en accès ouvert. Pour exploiter le contenu de cette base, une application de filtrage collaboratif devait être développée afin de permettre à un internaute de sélectionner des films répondant à ses goûts et à ses attentes. Pour atteindre ces différents objectifs, la modélisation d'une ontologie du cinéma et la définition d'un algorithme de calcul de distance sémantique constituaient des prérequis.

Mots-clés : Base de connaissance, ontologie du cinéma, filtrage collaboratif, distance sémantique, Recherche par similarité sémantique, triple-store, Architecture orientée Service.

1 Contexte du projet

1.1 Cadre universitaire

CAMERA-DREAM est un projet proposé aux étudiants du Master Génie Informatique Logicielle de l'Université de Rouen lors de l'année universitaire 2012-2013. Cette application a été développée dans le cadre de l'enseignement de gestion de projet informatique qui tient une place majeure dans le cursus et qui fait l'objet de travaux pratiques importants. En première année, les étudiants reçoivent un enseignement théorique basé sur les méthodes agiles et doivent réaliser un projet annuel en équipes de

5 ou 6 étudiants. Les étudiants acquièrent à cette occasion une première expérience de gestion de projet. A partir d'un sujet proposé par l'équipe enseignante, ils doivent s'organiser en équipes, spécifier précisément l'application qu'ils vont réaliser, planifier son développement, documenter sa conception et ses tests, etc. En seconde année, juste avant de partir en stage de fin d'étude, un projet de plus grande envergure est proposé sur une période de 12 semaines dont la moitié est entièrement consacrée au développement de l'application dans des conditions comparables à celles du monde industriel et dans un environnement proche de celui de l'entreprise.

1.2 Objectifs pédagogiques

Le projet représente un gros volume de travail et implique toute la promotion. Il est conduit selon une procédure inspirée de celle applicable aux marchés publics. L'objectif est de mettre les étudiants en situation aussi proche que possible de la réalité industrielle en leur demandant de conduire un projet d'envergure de la phase d'initialisation avec mise en concurrence jusqu'à la livraison d'une application fonctionnelle. L'une des difficultés majeures est de mettre en place une organisation structurée où chacun est responsabilisé sur des tâches précises. Le sujet traité doit être suffisamment complexe pour permettre la définition de lots de travaux conséquents et pouvant être alloués à différentes équipes. Le sujet doit aussi permettre aux étudiants de découvrir de nouvelles technologies et d'acquérir des compétences qui complètent ou étendent celles qu'ils ont pu acquérir dans le cadre des autres modules d'enseignement, notamment un cours sur les technologies du Web Sémantique proposé en option. Cette formation permet aux étudiants de se familiariser avec les bases technologiques, les approches méthodologiques et avec les standards du W3C qui sont exploités pendant le projet.

1.3 Organisation et processus

Le déroulement du projet suit les étapes suivantes :
- Présentation du calendrier et tirage au sort des équipes.
- Présentation générale du sujet.
- Lancement d'un appel d'offres par la MOA (maîtrise d'ouvrage, c'est à dire le client), au travers d'un cahier des clauses techniques particulières.
- Environ un mois après l'appel d'offre, remise par chaque groupe d'une réponse écrite et étayée puis soutenance orale avec réponse

aux questions ou contradictions soulevées par le client.
- Sélection par le client d'un maître d'œuvre (MOE) (l'équipe dont la proposition correspond le mieux à ses attentes), les autres équipes étant ses sous-traitants.
- Durant un mois, formalisation des contrats entre la MOE, ses sous-traitants et la MOA. La MOE définit des lots de travaux et les répartit entre les équipes.
- Lancement du projet lorsque le client considère que l'organisation et la définition des travaux ont atteint un degré de maturité suffisant.
- Phase de développement pendant 6 semaines, à temps plein.
- Soutenance finale et démonstration du système obtenu.

Pour l'année universitaire 2012-2013, vingt étudiants ont été répartis par tirage au sort en 2 équipes de 5 étudiants, 1 équipe de 6 étudiants et 1 équipe de 4 étudiants.

2 Sujet et cahier des charges

2.1 Thème proposé

Le projet vise à développer la base de connaissance *CAMERA* (Catalog of Actors and Movies Expressed as RDF Annotations) consacrée au cinéma et publiable sur le Web en accès ouvert. Cette base doit pouvoir être exploitée librement par des applications informatiques pour répondre à des besoins divers et variés. Afin de valider l'atteinte de ces objectifs, l'application *DREAM* (Discovery & Retrieval Engine for Actors & Movies) doit également être développée dans le cadre du projet. Cette application doit, entre autres fonctionnalités, permettre à un internaute de sélectionner des films répondant à ses goûts ou ses attentes.

Les résultats du projet doivent permettre de démontrer l'intérêt d'une description sémantique d'un catalogue Web de films (généralisable à différents types d'articles tels que des livres, des voyages, des produits industriels, etc.) dans l'optique de répondre au mieux aux préférences exprimées ou implicites d'un consommateur. Le système utilisé pour effectuer la démonstration est constitué par un serveur hébergeant une instance de la base de connaissance *CAMERA* et d'un client riche supportant l'application *DREAM*. Il sera désigné dans la suite de l'article sous le nom de *CAMERA-DREAM*.

2.2 Spécification fonctionnelle

CAMERA-DREAM collecte et catalogue des données cinématographiques qui sont analysées sémantiquement et annotées selon un système de métadonnées fondé sur une ontologie du cinéma. Les métadonnées issues de la phase d'annotation sont ensuite enregistrées dans la base de connaissance *CAMERA* et permettent de caractériser chaque film. Le contenu de cette base de connaissance est géré par un administrateur qui pilote le processus de collecte et de traitement des informations pour le référencement automatique des films. Il peut aussi compléter « manuellement » la base de connaissance en ajoutant des films d'origine quelconque. Le superviseur peut changer la structure du modèle (l'ontologie).

L'application *DREAM* propose une IHM destinée à tout internaute cinéphile qui veut rechercher des informations dans le catalogue.Cet utilisateur final peut interroger le catalogue en composant des requêtes complexes, se créer un profil personnel correspondant à ses goûts en valorisant les propriétés définies dans l'ontologie pour décrire son film idéal, ou donner son appréciation sur un film qu'il a visionné.

La mesure de similarité et les descriptions contenues dans *CAMERA* permettent de calculer un degré de ressemblance entre les films référencés et les profils des utilisateurs. Ainsi l'application *DREAM* peut émettre des recommandations personnalisées de films.

Ces trois rôles déterminent un certain nombre de cas d'utilisation modélisés dans le diagramme UML ci-après (Figure 1).

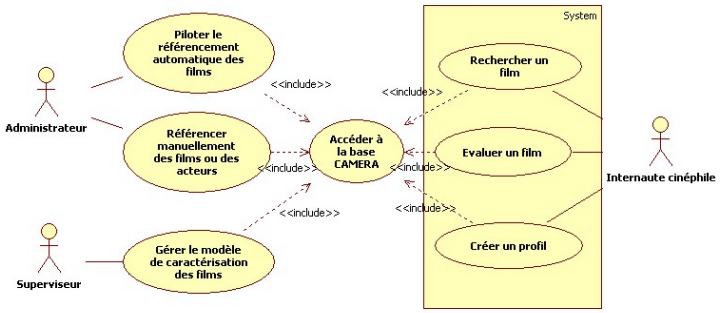

FIGURE 1 – Cas d'utilisation de *CAMERA-DREAM*

2.3 Exigences techniques

L'application est conçue selon une architecture orientée services : les différentes fonctionnalités proposées sont distribuées dans un ensemble de services indépendants ou faiblement couplés. Ces services peuvent être utilisés par des applications ou des systèmes intégrés et sont donc accessibles au travers d'une interface programmatique "publique".

Les annotations sémantiques sont formalisées en utilisant le standard RDF du W3C et la base de connaissance *CAMERA* est implémentée grâce à un triplestore qui expose un service Web conformément au Protocole SPARQL. L'application *DREAM* accède à la base *CAMERA* au travers d'une interface SparqlQuery (Figure 2).

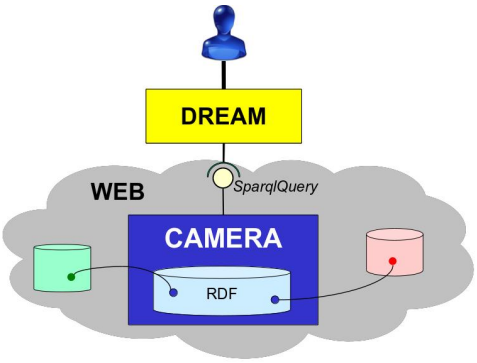

FIGURE 2 – Architecture générale du *CAMERA-DREAM*

2.4 Technologies mises en œuvre

Le développement est réalisé en Java avec l'IDE Eclipse. Le développement de l'IHM utilise le framework Play !, qui implémente les principes de JEE grâce à son serveur interne *Netty*. Les tests sont faits avec Junit 4 et SoapUI. Maven et svn sont utilisés pour la production et le contrôle du code.

Les standards mis en œuvre pour le développement de l'ontologie et du triplestore sont :

- RDF : un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées ;
- RDF / XML : l'une des syntaxes de sérialisations des éléments RDF ;

- OWL : un langage de représentation des connaissances, utilisé pour définir des ontologies web structurées en RDF ;
- SPARQL (SPARQL Protocol and RDF Query Language) : un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF dans un graphe ;
- JQuery : une bibliothèque JavaScript libre qui simplifie les commandes communes de Javascript ;
- Apache CXF : un framework open-source en langage Java, facilitant le développement de services web ;
- Apache Jena-Core : un framework open-source en langage Java, facilitant le développement des applications sémantique web ;
- Apache Jena-ARQ : Le moteur des requêtes SPARQL pour Jena ;
- JSoup : un parseur HTML ;
- Jastor : un outil permettant de générer des beans à partir d'une ontologie formalisée en OWL ;
- Fuseki : un moteur SPARQL accessible en tant que serveur via HTTP.

3 Architecture du système

3.1 Services développés

Les données sont collectées depuis des sources d'informations cinématographiques. Ces données sont annotées sémantiquement et transcrites pour les aligner avec l'ontologie de référence. Une indexation sémantique des films et des acteurs, et une indexation en texte brute sur les résumés et synopsis est faite pour capitaliser l'information dans la base *CAMERA*. (Voir Figure 3)

3.2 Sources de données utilisées

Trois sources de données cinématographiques sont utilisées :

- Allociné (Allocine, 2013) : un site web français dédié aux films de cinéma, séries, vidéos et programmes télévision,
- IMDb (IMDb, 2013) : un site web américain similaire à Allociné,
- Linked Movie Database (Linkedmdb, 2013) : une ontologie de films.

De l'information supplémentaire sur les films, les personnes et sites géographiques associés est obtenue de :

- dbpedia.fr (DBpedia, 2013) : une source des donnée structurées extraites de différents chapitres francophones de Wikipedia.
- GeoNames (GeoNames, 2013) : une base de données géographiques.

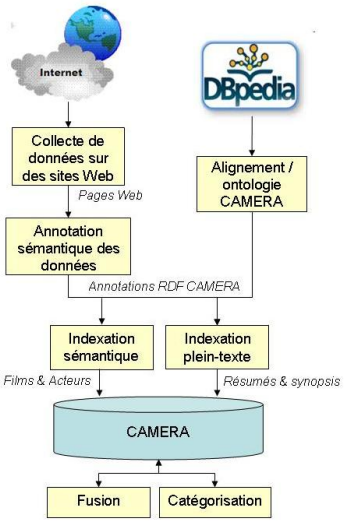

FIGURE 3 – Chaîne de Traitement

3.3 Base de connaissances CAMERA

La base de connaissances *CAMERA* est implémentée sur un triplestore (un entrepôt RDF). L'interface avec l'application *DREAM* repose sur une API Java d'Apache : Jena. Jena permet d'interroger et éditer la base par des requêtes SPARQL. Le serveur SPARQL, Fuseki, permet d'accéder à distance à l'entrepôt RDF. Il fournit différents « endPoints » SPARQL que l'on peut interroger pour rechercher, modifier, supprimer ou ajouter des données. Ceci permet aux Web Services d'accéder aux données. Trois DataSets sont implémentés, un pour les films et les personnes, un second pour les profils utilisateurs et un troisième pour la configuration du système.

3.4 Application DREAM

3.5 Cas d'utilisation et interfaces applicatives

L'interface Homme-Machine est destinée à trois utilisateurs (Figure 1)

1. l'administrateur : qui dispose des droits d'administration pour le système et la base de connaissances, et qui peut régler le seuil de la distance sémantique (voir Section 5) ;
2. le superviseur : qui s'occupe de la gestion de l'ontologie ;

3. l'internaute : un utilisateur quelconque sans droit d'administration ou de supervision ;

3.6 Prise en compte des goûts et des humeurs

Il y a deux catégories d'utilisation par un internaute :

1. De manière occasionnelle, par exemple, il peut chercher, soit à partir d'un film (*trouvez-moi les films qui ressemblent à X*), soit en fonction des caractéristiques des films (*trouvez-moi les films d'action de Spielberg*), ou même selon son humeur du moment (*trouvez-moi les films tristes*).
2. Un service personnalisé avec son propre profil, et des recommandations ciblées.

Pour construire son profil, il peut se baser sur l'ontologie (Section 4) et sur les caractéristiques des films qu'il apprécie plus particulièrement :

– les genres et sujets des films, et les émotions qu'ils évoquent.
– les personnes qui ont participé d'une façon ou d'une autre à la conception des films : les acteurs, réalisateurs, et/ou producteurs.

Il peut aussi donner une note aux films qu'il a vus. Dans ce cas, son profil est déterminé à partir des genres, sujets, émotions et personnes connectés aux films les mieux notés. Une fois les profils créés, le système calcule des groupes (clusters) d'internautes ayant des profils similaires. Ce calcul est détaillé en section 5. Ces groupes permettent à l'internaute de recevoir des alertes quand quelqu'un qui partage ses goûts donne à un film une note positive, ou d'envoyer automatiquement des alertes aux membres d'un groupe dès la sortie d'un film qui correspond aux goûts de ceux-ci. L'administrateur peut régler le seuil de proximité, qui permet de changer à volonté la taille et la constitution de ces clusters.

4 L'ontologie du cinéma

L'ontologie est la spécification d'une conceptualisation d'un domaine de connaissances cinématographiques basée sur la classification des éléments de ce domaine donné par concepts et sous-concepts et la création de tout type de relations entre les éléments et les concepts.

4.1 Portée de la modélisation

L'ontologie sert à caractériser un film selon le point de vue du public, avec des propriétés portant sur les contributeurs (acteurs, réalisateurs, producteurs, auteurs), les genres des films, les sujets traités, les émotions ressentis en regardant le film, les années de sortie en salle ou sur DVD, etc., les lieux où le film est tourné et où l'action se déroule, et sur les synopsis ou résumés.

4.2 Hiérarchies de genres et de sujets

L'ontologie de Linked Movie Database (Linkedmdb, 2013) offre de nombreux genres et sujets associés aux films, mais dans une structure plate. Pour mesurer la similarité entre deux films, deux taxonomies ont été construites dans le cadre du projet *CAMERA-DREAM*. La première de genres et la deuxième de sujets et, dans ces deux taxonomies, la proximité entre les genres d'une part et les sujets d'autre part a été étudié.

4.3 Formalisation

Pour discuter la spécification de l'ontologie avec le client, un diagramme de classes UML a été utilisé qui impose un niveau de formalisme à la fois graphique, explicite, précis et rigoureux (Figure 4). L'outil retenu pour la création et l'édition de l'ontologie est le logiciel libre Protégé, qui impose un niveau de formalisme plus formel mais moins lisible par un client.

4.4 Alignement des ontologies utilisées

L'alignement d'ontologies est le processus de découverte des correspondances entre concepts.Deux ontologies différentes peuvent avoir des termes différents pour le même concept, par exemple, le concept "cinéaste" d'une première ontologie est équivalent au concept "réalisateur" dans une seconde ontologie. Les ontologies peuvent être aussi dans des langues différentes, par exemple, l'ontologie de *CAMERA-DREAM* est exprimée en utilisant la langue anglaise, celle de (DBpedia, 2013) existe en français.

Dans *CAMERA-DREAM*, l'alignement d'ontologies consiste à établir une correspondance entre l'ontologie définie dans le cadre du projet et les modèles conceptuels sur lesquels sont basées les données collectées c'est à dire les modèles de (DBpedia, 2013), (Allocine, 2013), (IMDb, 2013), et (Linkedmdb, 2013).

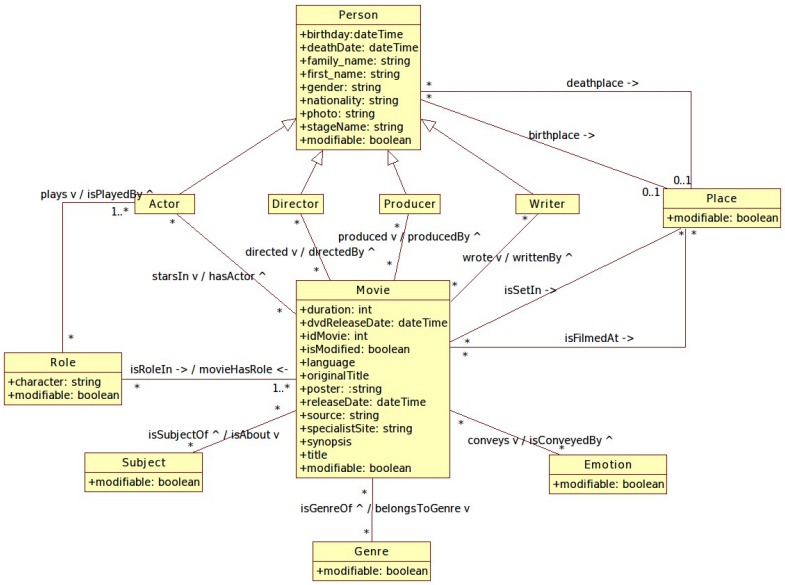

FIGURE 4 – UML de l'ontologie

Les listes de genres, sujets et émotions des sites web sont extraites, traduites et alignées « à la main » avec les genres, sujets et émotions de l'ontologie *CAMERA-DREAM*.

5 Le calcul de la distance sémantique

5.1 Principes et état de l'art

Comment déterminer si un concept, ou une chose C1 est sémantiquement plus proche d'un concept ou d'une chose C2 que d'un concept ou d'une chose C3 ? Par exemple, est-ce qu'un tournevis est plus proche d'un marteau que d'une cisaille ?

Le besoin d'une mesure de la distance entre deux concepts n'est pas nouveau (Quillian, 1968; Collins & Loftus, 1975), et la recherche sur ce sujet se poursuit encore actuellement. Pour donner quelques pistes de recherche, trois articles scientifiques sont fournis en annexe du CCTP de *CAMERA-DREAM*(Aimé *et al.*, 2011; Gandon *et al.*, 2008; Khelif *et al.*, 2008). Un point particulièrement intéressant pour ce projet est la mesure de la distance sémantique entre deux concepts définis dans une même taxo-

nomie en tenant compte de la profondeur du chemin qui les relient par des liens de subsomption (Gandon *et al.*, 2008).

5.2 Algorithmes mis en œuvre

L'un des objectifs principaux du projet est de montrer l'intérêt scientifique d'un calcul de distance sémantique dans le cadre d'une application de filtrage collaboratif. Plus spécifiquement [1], il s'agit de mesurer la distance sur une ontologie du cinéma entre :

- deux profils, pour un clustering des profils similaires ;
- deux films, pour émettre des recommandations de films similaires ;
- un profil et un film, pour émettre des recommandations personnalisées de films.

Les propriétés qui caractérisent un film sont divisées en deux groupes :

- les personnes associées à un film – les acteurs, les producteurs, les réalisateurs ;
- le contenu du film – son genre et le type de sujets qu'il aborde.

Un profil consiste en une liste de films, acteurs, réalisateurs et producteurs favoris, et de genres et sujets préférés.
La distance entre deux films ou deux profils peut être calculée de la même manière. Pour cela, on introduit la notion de ressource qui généralise ces deux concepts. Une ressource Res est défini par $Res = \{A, D, R, G, S\}$ où A est la liste des acteurs, D est la liste des producteurs, R est la liste des réalisateurs, G est la liste des genres et S est la liste des sujets associés à cette ressource.

La distance entre deux ressources, Res_1 et Res_2 est $dist(Res_1, Res_2)$ telle que :

$$dist(Res_1, Res_2) = \begin{cases} mindist & \text{si } Res_1 = Res_2, \\ P_A.(dist_P(A_1, A_2))+ & \\ P_D.(dist_P(D_1, D_2))+ & \\ P_R.(dist_P(R_1, R_2))+ & \\ P_G.(dist_T(G_1, G_2))+ & \\ P_S.(dist_T(S_1, S_2)) & \text{sinon.} \end{cases} \quad (1)$$

Le paramètre $mindist$ (par défaut 0) est réglable selon les souhaits de l'administrateur du système. P_A, P_D, P_R, P_G, P_S sont les poids donnés aux

1. À notre connaissance, ces calculs n'ont pas encore été faits.

acteurs, producteurs, réalisateurs, genres et sujets pour le calcul tels que $P_A + P_D + P_R + P_G + P_S = 1$. La formule fait intervenir deux distances différentes, $dist_P$ qui est la distance entre les listes de personnes associées à un film, et $dist_T$, la distance taxonomique.

5.2.1 La distance entre deux listes de personnes

La distance entre deux listes de personnes $dist_P(P_1, P_2)$ est calculée à partir de l'intersection de ces listes :

$$dist_P(P_1, P_2) = \begin{cases} mindist & \text{si } P_1 = P_2, \\ \frac{maxdist}{2^{P_1 \cap P_2}} & \text{sinon.} \end{cases} \quad (2)$$

où $maxdist$ est réglable selon les souhaits de l'administrateur du système.

5.2.2 La distance taxonomique entre genres et sujets

La distance entre deux listes d'éléments d'une taxonomie $dist_T(T_1, T_2)$ est calculée selon la position et la profondeur de chaque nœud dans la taxonomie (Figure 5). Intuitivement, puisqu'ils ont un détail plus fin (ils sont plus bas dans l'arbre), on dira que n_1 et n_2 sont "plus proche" que n_3 et n_4. Inversement, parce qu'ils sont dans une catégorie différente, ils peuvent être considérés comme éloignés de n_5.

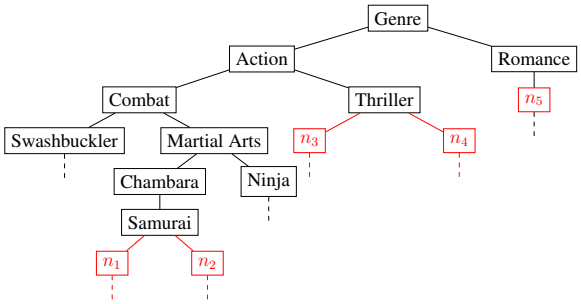

FIGURE 5 – La distance entre deux nœuds est relative à leurs profondeurs et leurs positions relatives dans le graphe taxonomique : $dist(n_1, n_2) < dist(n_3, n_4) < dist(n_4, n_5)$

On définit $LCA(n_1, n_2)$ comme l'ancêtre en commun le plus proche (Lowest Common Ancestor) des deux nœuds n_1 et n_2 (Figure 6).

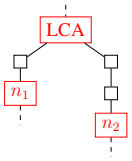

FIGURE 6 – Lowest Common Ancestor de n_1 et n_2

La distance entre deux nœuds n_1 et n_2, où $n_1 \neq n_2$ est définie comme

$$dist_T(n_1, n_2) = \sum_{0 \leq i < p_1} \frac{1}{2^{depth(LCA(n_1,n_2))+i}} + \sum_{0 \leq j < p_2} \frac{1}{2^{depth(LCA(n_1,n_2))+j}} \tag{3}$$

où $p_1 = |path(n_1, LCA(n_1, n_2))|, p_2 = |path(n_2, LCA(n_1, n_2))|$

Alors la distance entre deux listes d'éléments taxonomiques est donnée par :

$$dist_T(T_1, T_2) = \begin{cases} \frac{\sum_{n \epsilon T_1} \frac{maxdist}{2^{depth(n)}}}{|T_1|} & \text{si } T_1 = T_2, \\ \frac{\sum_{t_1 \epsilon T_1, t_2 \epsilon T_2} dist_T(t_1,t_2)}{|T_1|.|T_2|} & \text{sinon.} \end{cases} \tag{4}$$

5.3 Implémentation

La première étape consiste à formaliser les arborescences taxonomiques de l'ontologie du cinéma, et à ajouter sur chaque nœud une propriété représentant l'expression de son chemin depuis la racine. Le graphe résultant est stocké, et n'est recalculé que lorsque l'ontologie est modifiée.

Pour calculer la longueur du chemin entre deux genres, ou deux sujets, il suffit de comparer les expressions de leurs chemins stockés respectifs. La longueur du préfixe commun donne la profondeur de leur ancêtre le plus proche. La somme des longueurs des suffixes résiduels donne la longueur du chemin entre les deux nœuds.

5.4 Intérêt applicatif et transpositions envisageables

Ce projet repose sur une ontologie du cinéma, et deux mesures différentes sont proposées : l'une qui agit sur les caractéristiques fixes d'une ressource, l'autre qui agit sur des taxonomies. L'intérêt est de combiner

ces deux mesures, et de varier les poids pour ajuster la qualité des résultats. Les calculs ne sont pas spécifiques au cinéma, et on peut imaginer une application gérant n'importe quel type de ressources qui peut être décrit par des caractéristiques et des taxonomies telles que des livres, des voyages, des produits industriels, des événements etc.

6 Bilan du projet

6.1 Corpus d'évaluation

Pendant le projet, un corpus des données cinématographiques pour les films réalisés depuis 1970 (9000 films de DBPédia, environ 7000 de Allociné, entre 5000 et 7000 de IMDB et environ 2000 de LinkedMDB) a été constitué. La démonstration en-ligne (M2GIL, 2013) contient cependant un corpus réduit à 500 films.

Les résultats dépendent fortement de la justesse des catégorisations de l'ontologie, de leur répartition, et de la hauteur de la taxonomie. 679 genres et sous-genres sont identifiés (hauteur taxonomique 6) ; 176 sujets et sous-sujets (hauteur taxonomique 5).

6.2 Résultats obtenus

Jinni (Jinni, 2013) est un site web concurrent de *CAMERA-DREAM*. Il se base sur *The Movie Genôme Project* qui a pour but de catégoriser des films selon deux critères : "Experience" – l'humeur et la tonalité du contenu et "Story" – les éléments du synopsis. Pour montrer l'intérêt et la valeur ajoutée de notre approche, nous avons fait une comparaison entre les résultats retournés par *CAMERA-DREAM*, et ceux de *Jinni* (Jinni, 2013). Un exemple est donné ici pour le film *Expendables 2 : Unité Spéciale* (Figure 7, Table 1).

TABLE 1 – Comparaison des films similaires à *Expendables 2 : Unité Spéciale*

CAMERA-DREAM	*Jinni*
Le Sorcier et le Serpent Blanc	The Expendables
Bangkok Resistance	Rambo, Rambo II, Rambo First Blood
Mission Impossible	Transporter
Kill the Gringo	The Eliminator
Safe	Safe

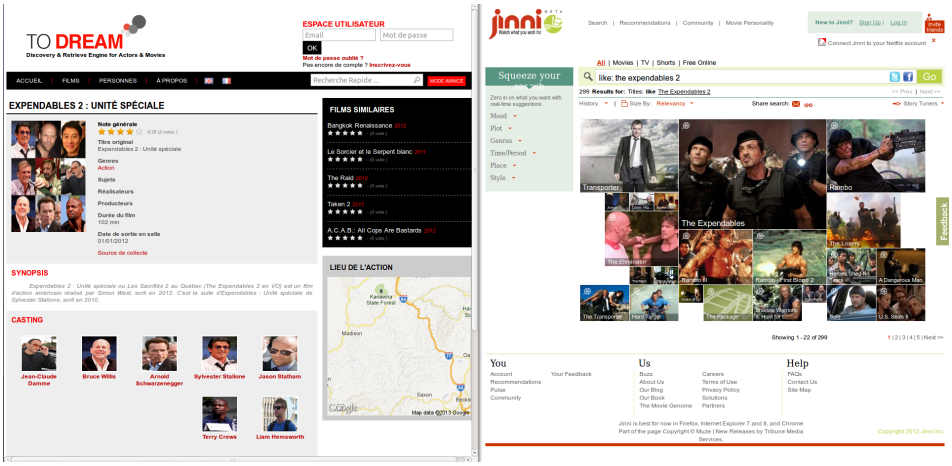

FIGURE 7 – Comparaison des recherches de films similaires à *Expendables 2 : Unité Spéciale* entre *CAMERA-DREAM* et *Jinni*

Un avantage de *CAMERA-DREAM* est que les paramètres du calcul de la distance sémantique sont tous réglables, et donc personnalisables selon les goûts de l'utilisateur de façon plus ou moins empirique en tenant compte de l'expérience. Par exemple, lors de nos expérimentations, nous avons déterminé que de meilleurs résultats étaient obtenus lorsque le poids sur les personnes associées à un film est double du poids taxonomique.

6.3 Limitations et évolutions possibles

L'utilisation de bases de données en français et en anglais nécessite une traduction des genres, sujets, et émotions. Une traduction automatique rendrait le processus plus rapide.

La modélisation des goûts de l'utilisateur pourrait être réalisée par entraînement automatique en utilisant des techniques d'apprentissage.

Les genres et sujets sont extraits à partir des étiquettes sur les sites web, et avec une indexation texte brute. Il pourrait être intéressant d'explorer la distance sémantique entre des termes pour une extraction plus complète des informations.

Il est envisageable d'ajouter plus de critères au calcul de la distance sémantique, tel que la date de sortie, l'auteur, la durée du film, les lieux où se déroule le film, les prix ou les nominations. Les taxonomies de genres et sujets peuvent être enrichies, et de nouvelles taxonomies pourraient être

ajoutées, par exemple, pour tenir compte des émotions.

Quelques duplications de données ont été relevées, par exemple quand le nom d'un acteur est épelé différemment sur deux sites. Il serait intéressant d'implémenter un calcul de similarité sur les acteurs et les descriptions des films pour les fusionner (profile matching).

Remerciements

Les auteurs impliqués dans la réalisation de ce projet tiennent à remercier Bruno PATROU, Philippe ANDARY et Florent NICART pour leurs conseils avisés sur l'implémentation du calcul de la distance sémantique.

Références

AIMÉ X., FÜRST F., KUNTZ P. & TRICHET F. (2011). Semiosem et proxsem : mesures sémiotiques de similarité et de proximité conceptuelles. In *atelier « Personnalisation du Web », 22èmes Journées francophones d'Ingénierie des Connaissances (IC'2011)*, Chambéry, France.

ALLOCINE (2013). Allociné. `http://www.allocine.fr/`.

COLLINS A. M. & LOFTUS E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, **82**(6), 407 – 428.

DBPEDIA (2013). Dbpedia website. `http://fr.dbpedia.org/`.

GANDON F., CORBY O., DIOP I. & LO M. (2008). Distances sémantiques dans des applications de gestion d'information utilisant le web sémantique. In *Proc. Workshop Mesures de similarités sémantique, EGC*, INRIA Sophia Antipolis - Méditerranée. `http://www.inria.fr/sophia/axis/egc08/atelier_5.pdf`.

GEONAMES (2013). Geonames website. `http://www.geonames.org/`.

IMDB (2013). Imdb website. `http://www.imdb.com/`.

JINNI (2013). Jinni website. `http://www.jinni.com/`.

KHELIF K., GANDON F., CORBY O. & DIENG-KUNTZ R. (2008). Using the Intension of Classes and Properties Definition in Ontologies for Word Sense Disambiguation. In *Proc. 16th International Conference on Knowledge Engineering and Knowledge Management - Knowledge Patterns, EKAW*, Acitrezza, Italy.

LINKEDMDB (2013). Linked movie database website. `http://data.linkedmdb.org/`.

M2GIL (2013). Camera-dream website. `http://www.cameradream.fr`, Master 2 Génie de l'Informatique Logicielle, Université de Rouen.

QUILLIAN M. (1968). Semantic memory. In M. MINSKY, Ed., *Semantic Information Processing*, p. 216–270. Cambridge, MA : MIT Press.