

Recherche d'Information Sémantique: Appariement sémantique flou de documents semi-structurés

Arnaud Renard, Sylvie Calabretto, Béatrice Rumpler

LIRIS UMR 5205 - INSA de Lyon,
7 avenue Jean Capelle,

69621 Villeurbanne cedex, France,

{arnaud.renard,sylvie.calabretto,beatrice.rumpler}@insa-lyon.fr

Résumé. La sémantique constitue un des enjeux majeurs dans l'évolution des systèmes de RI¹ en général et dans les systèmes de RI structurée en particulier. Nous nous sommes intéressés aux différentes façons de la prendre en compte afin d'améliorer notamment la gestion des hétérogénéités des collections de documents XML lorsque ces derniers sont semi-structurés. La prise en compte de la sémantique passe notamment par l'emploi de ressources sémantiques externes à la collection de documents initiale sur lesquelles il est nécessaire de disposer de mesures de similarité sémantique pour pouvoir effectuer des comparaisons entre concepts. Nous avons ensuite présenté brièvement différents systèmes de RI structurée sémantique pour terminer avec la proposition d'une mesure de pondération sémantique floue pouvant être intégrée au sein d'un système déjà existant.

Mots clés: recherche d'information, documents (semi-)structurés, XML, documents hétérogènes, appariement sémantique flou, ressource sémantique, thesaurus, ontologie.

1 Introduction

Le Web évolue selon deux grandes tendances, à savoir : une structuration de plus en plus importante, et une prise en compte de la sémantique elle aussi en pleine croissance. La sémantique consiste en l'étude de la signification des mots ainsi que les rapports de sens entre ces derniers (tels que l'homonymie, la synonymie, l'antonymie, l'hyperonymie, l'hyponymie). L'introduction de sémantique dans la RI est devenue un facteur clé dans le perfectionnement des moteurs de recherche, qu'ils soient grand public ou même spécialisés dans des domaines particuliers de la RI comme la RI semi-structurée.

Divers exemples peuvent ainsi être cités et notamment celui du moteur de recherche Google qui a récemment intégré la prise en compte de notions sémantiques au sein de son moteur de recherche. Ceci permet de combler le fossé sémantique en suggérant à l'utilisateur des requêtes apparentées à la sienne.

¹ RI : Recherche d'Information

Nous traiterons ici plus spécifiquement de la RI dans les documents XML². Dans la littérature, il existe de nombreux systèmes de RI qui ont pour but de renvoyer aux utilisateurs les informations pertinentes présentes dans une collection de documents XML. Afin de satisfaire au mieux la requête traduisant le besoin d'un utilisateur, ces systèmes tentent de maximiser les critères de spécificité et d'exhaustivité de la réponse vis-à-vis de la requête.

Un problème majeur rencontré en RI relève de l'hétérogénéité des documents présents dans les collections considérées, et notamment en ce qui concerne la RI dans les documents XML une hétérogénéité structurelle, ce qui signifie que les documents considérés ne suivent ni la même DTD³ ni le même schéma XML. Cette disparité a la plupart du temps pour origine l'absence de consensus entre les différentes sources d'informations. Les utilisateurs sont donc dans l'incapacité d'avoir des connaissances à la fois complètes et précises des structures des documents d'une collection. De fait, les utilisateurs éprouvent des difficultés à exprimer des requêtes comportant des contraintes sur la structure. Il est donc pertinent dans un premier temps de ne considérer que les requêtes sur le contenu. On se contentera dans un premier temps de gérer les hétérogénéités au niveau des termes pouvant représenter aussi bien le contenu textuel des documents que les éléments de structure.

Il est communément accepté que l'utilisation de ressources sémantiques comme des ontologies et des taxonomies de concepts améliore les performances des systèmes de RI [Rosso'04]. Il est pour cela nécessaire d'être en mesure de réaliser un appariement entre les termes des documents et les termes représentant les instances de concepts dans une ressource sémantique. Quelques systèmes tentent déjà d'utiliser de telles ressources dans le cadre de la RI (semi-)structurée. Nous essayerons donc d'améliorer un système de RI existant en réalisant un appariement sémantique flou permettant de prendre en compte des erreurs courantes, telles que les fautes de frappe, mais aussi d'autres types de fautes moins fortuites, ou encore des figures de style trompeuses. Nous souhaitons pour cela nous appuyer sur des ressources issues du Web sémantique, l'objectif étant à terme de prendre en compte aussi bien la sémantique du contenu textuel des documents que la sémantique portée par leur structure. Ainsi, lors de notre étude, l'accent sera mis sur les outils et ressources sémantiques nécessaires, et notamment une ressource pour l'instant relativement peu utilisée : DBpedia qui constitue une vaste ontologie multi-lingue et multi-domaine.

La suite de l'article est organisée de la façon suivante. Nous présenterons dans la partie 2 les différentes approches prenant en compte l'aspect sémantique dans la littérature concernant la RI structurée. Nous décrirons également dans cette partie, les outils nécessaires à la prise en compte de la sémantique, notamment les ressources sémantiques (thesaurus, ontologie, ...), ainsi qu'une méthode de calcul de la similarité entre concepts. Nous présenterons ensuite notre contribution dans la partie 3. Enfin, nous conclurons et nous exposerons les perspectives d'évolutions que nous envisageons.

² XML : eXtensible Markup Language

³ DTD : Document Type Definition

2 Outils et travaux connexes

Comme nous pourrons le constater par la suite dans les différents systèmes de RI (semi-)structurée que nous présenterons, il est nécessaire de disposer de ressources externes ainsi que de mesures permettant d'effectuer des comparaisons afin d'être en mesure de mieux prendre en compte la sémantique. Ceci amène [Bellia'08] à définir la notion de cadre sémantique qui repose sur deux notions complémentaires à savoir : la ressource sémantique (externe) et le modèle de mesure de similarité entre concepts.

2.1 Ressources sémantiques

Différents types de ressources sémantiques peuvent être distingués parmi lesquels se trouvent les thesaurus, ainsi que les ontologies. Les ressources sémantiques peuvent se distinguer selon l'étendue des connaissances qu'elles comportent en deux catégories : les ressources de domaine, et les ressources généralistes. Etant donnés les objectifs d'expérimentation future que nous nous sommes fixés (la collection de documents d'INEX⁴ nécessite de disposer de connaissances généralistes), seules les ressources généralistes seront abordées dans ce document. En effet, l'emploi de ressources sémantiques de domaine ne couvrirait pas un champ suffisamment vaste et ne permettrait de disposer que de connaissances parcelaires.

2.1.1 Taxonomie

Une taxonomie est un vocabulaire contrôlé organisé sous une forme hiérarchique simple. Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation affinant le sens d'un terme. A la différence du thesaurus qui permet de parcourir la hiérarchie de manière connexe permettant ainsi de restreindre ou de spécialiser le champ des connaissances, cela n'est pas possible avec une taxonomie.

2.1.2 Thesaurus

Un thesaurus constitue un dictionnaire (vocabulaire contrôlé) hiérarchisé. Ce vocabulaire est normalisé et présente les termes (génériques ou spécifiques à un domaine) composant le vocabulaire sous une forme standard (lemme). Les termes y sont organisés dans une hiérarchie de concepts liés par des relations. Un thesaurus peut également fournir des définitions. Les relations couramment présentes dans un thesaurus sont des relations taxonomiques (hiérarchie), d'équivalence (synonymie), d'association (proximité sémantique, proche-de, relié-à, ...). Il existe des thesaurus spécialisés dans des domaines précis tels que MeSH (domaine biomédical). Le thesaurus généraliste de référence étant Wordnet [Fellbaum'98].

Wordnet

Wordnet est un thesaurus à caractère généraliste pour la langue anglaise dont l'organisation dépend du bon sens humain. Les noms, verbes, adjectifs et adverbes

⁴ INEX : INitiative for Evaluation of XML retrieval

sont organisés en ensembles de synonymes (appelés synsets), chaque ensemble représentant un concept lexical. De plus, différentes relations lient les ensembles de synonymes.

2.1.3 Ontologie

Les ontologies permettent, de décrire les connaissances d'un domaine spécifique mais aussi de représenter des relations entre concepts complexes ainsi que des axiomes et règles absentes des thesaurus. De cette façon, il est possible de donner à la machine un meilleur niveau de « compréhension » et il devient ainsi possible d'inférer des informations ce qui permet également de maintenir une plus grande cohérence.

YAGO⁵

YAGO [Suchanek'07] est une ontologie généraliste de grande taille et extensible qui s'appuie sur les entités et relations extraites à partir de Wikipedia. Le contenu de YAGO a été extrait automatiquement de Wikipedia et unifié avec la sémantique de Wordnet avec une précision 95%. Tous les objets (villes, personnes, mêmes les URLs) sont représentés comme des entités dans le modèle de YAGO. YAGO n'utilise pas la totalité de la hiérarchie des catégories de Wikipedia mais associe les catégories feuilles à la taxonomie de Wordnet pour établir des relations de type « subClassOf ».

Ainsi, la relation « subClassOf » de YAGO est utilisée par [Demartini'08] pour obtenir les concepts sémantiques décrivant les entités de Wikipedia.

DBpedia⁶

Comme cela est expliqué dans [Auer'07], DBpedia est la résultante d'un effort de toute une communauté pour extraire des informations structurées à partir de Wikipedia et les mettre à disposition sur le Web. DBpedia présente l'avantage de permettre l'interrogation des données de Wikipedia au moyen de requêtes complexes. De plus, DBpedia permet de lier d'autres ressources présentes sur le Web à celles extraites de Wikipedia.

Une part importante des ressources de DBpedia est disponible dans 36 langues différentes. Toutefois, la version anglaise de Wikipedia étant la plus utilisée, un maximum de ressources sont disponibles dans cette langue. L'ensemble des concepts de DBpedia sont décrits par un résumé disponible dans la totalité de ces langues. Toutes les connaissances stockées par DBpedia décrivent environ 2,6 millions d'objets, incluant 213000 personnes, 328000 lieux, 57000 albums musicaux, 36000 films, 20000 entreprises. La base de connaissances totalise ainsi 274 millions de triplets RDF et représente également 609000 liens vers des images, 3150000 liens vers des pages Web externes, 4878100 liens vers des données RDF externes. Tout cela est rattaché à 415000 catégories de Wikipedia, et à 75000 catégories de l'ontologie YAGO. L'ensemble des données de DBpedia est accessible par trois moyens différents à savoir : des données liées, l'interrogation SPARQL, et des sauvegardes des fichiers RDF. La richesse des interconnexions entre DBpedia et de

⁵ YAGO : Yet Another Great Ontology (<http://www.mpi-inf.mpg.de/yago-naga/yago>)

⁶ DBpedia (<http://wiki.dbpedia.org>)

Recherche d'Information Sémantique: Appariement sémantique flou de documents semi-structurés

5

nombreuses autres ressources placent DBpedia au cœur du projet LOD⁷ du W3C⁸ (cf. Fig. 1).

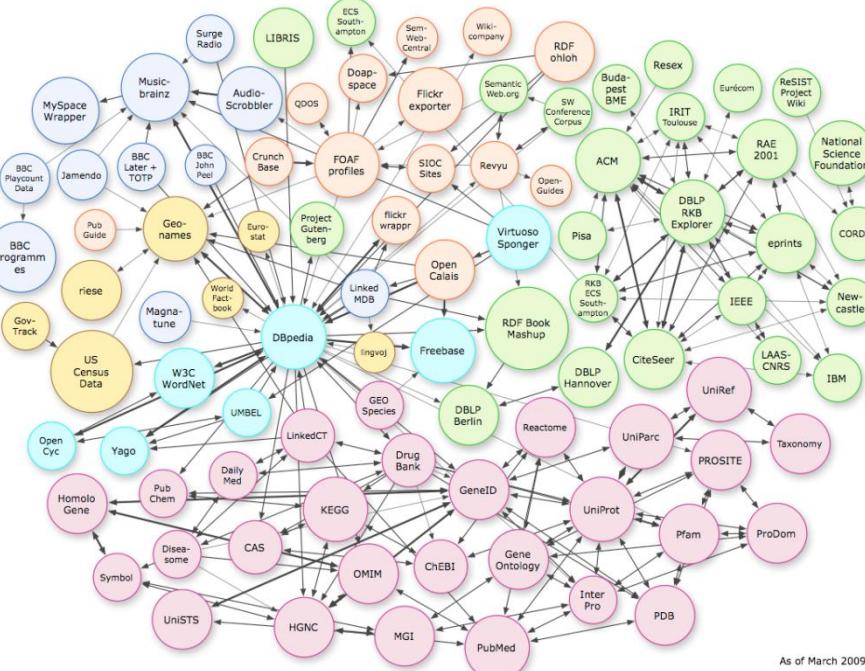


Fig. 1. LOD cloud (source: http://www4.wiwiiss.fu-berlin.de/bizer/pub/lod-datasets_2009-03-27_colored.png)

Même si relativement peu de projets de RI s'appuient sur DBpedia pour l'instant, la récente mise à disposition d'une ontologie OWL⁹ devrait faire évoluer les choses. On trouve ainsi dans les « use cases » de DBpedia que cette ressource peut être utilisée en tant que vaste ontologie multi-lingue et multi-domaine. Comparée à des ontologies de domaine qui sont en général créées par de petits groupes d'ingénieurs et qui sont très coûteuses à mettre à jour avec les évolutions du domaine, DBpedia présente plusieurs avantages dont le fait : qu'elle couvre de nombreux domaines, qu'elle représente un consensus réel de la part d'une communauté, et enfin qu'elle évolue « automatiquement » avec les changements de Wikipedia. De la même façon que YAGO a été utilisé par [Demartini'08], [Kobilarov'09] propose l'utilisation de

⁷ LOD : Linking Open Data – Il s'agit d'un projet du W3C dont le but est d'interconnecter l'ensemble des sources de données du Web pour passer d'un Web de liens à un Web de données: <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

⁸ W3C : World Wide Web Consortium (<http://www.w3.org>)

⁹ OWL : Ontology Web Language

DBpedia pour interconnecter les multiples sites spécifiques à un domaine de la BBC¹⁰.

Parmi toutes les informations contenues dans DBpedia, certains sous-ensembles sont plutôt riches d'un point de vue sémantique dans le sens où elles contiennent des informations très spécifiques. Au contraire, d'autres contiennent des métadonnées et sont plutôt pauvres. Toutefois, ces métadonnées peuvent servir à calculer des mesures de proximité entre concepts, ou de pertinence dans les résultats de recherche. Nous supposons toutefois que le calcul de ce type de mesures dans une ressource d'une telle envergure que celle-là ne pourra pas s'effectuer de la même façon que sur des ressources « classiques » comme Wordnet.

Bien que le choix d'une ressource sémantique adaptée constitue un point déterminant dans les performances d'un système de RI structurée sémantique, il est nécessaire de disposer de mesures de similarité s'appliquant sur cette ressource. En effet, la détermination du degré de similarité entre deux concepts permet de comparer les concepts entre eux. Cette comparaison est notamment utilisée dans le processus de pondération des termes, lors de l'interrogation par les concepts, mais également lors de la désambiguïsation sémantique. Un état de l'art complet de la désambiguïsation peut être trouvé dans [Navigli'09].

2.2 Mesures de similarité sémantique

Afin d'être en mesure d'évaluer la similarité sémantique des concepts (auxquels les termes des requêtes et documents sont rattachés) d'une ressource sémantique telle qu'un thesaurus sémantique ou qu'une ontologie, il faut disposer de mesures de similarités. En premier lieu, il est nécessaire d'effectuer la distinction [Zargayouna'05], entre la similarité sémantique et la proximité sémantique. En effet, cette dernière notion prend en considération tout type de relation entre les concepts et elle recouvre donc un champ plus large que la « simple » notion de similarité. Il est souligné dans ces travaux que la notion de similarité semble plus pertinente dans le cadre de la RI même si elle est plus restrictive. Il existe de multiples états de l'art concernant les mesures de similarité et de distance sémantique. Ainsi, cinq mesures de similarités/distances sémantique s'appuyant sur Wordnet (structuration des synsets avec des relations de type « est-un ») sont comparées dans [Budanitsky'01].

Deux types de mesures de similarité sémantiques peuvent être distingués. Le premier type s'appuie sur la structure de la ressource sémantique en proposant un comptage plus ou moins élaboré du nombre d'arcs séparant deux concepts. En revanche, le deuxième type de mesures s'appuie sur le contenu informationnel. Le contenu informationnel traduit la pertinence d'un concept dans la collection en tenant compte de la fréquence de son apparition dans la collection ainsi que de la fréquence d'apparition des concepts qu'il subsume. Toutefois, [Zargayouna'05] ayant montré que le premier type de mesure pouvait être aussi performant que le deuxième, seul le premier type de mesure sera abordé dans ce document. En effet, les mesures du

¹⁰ BBC : British Broadcasting Corporation (<http://www.bbc.co.uk>)

deuxième type nécessitent un apprentissage, elles sont dépendantes de la qualité de l'apprentissage et s'avèrent plus contraignantes (en particulier du fait de la difficulté à trouver un corpus d'apprentissage adapté). On peut notamment citer dans ce domaine, les travaux de [Resnik'95] qui a introduit le contenu informatif, ainsi que ceux de [Jiang'97] et [Lin'98] qui utilisent une approche mixte, et plus récemment ceux de [Formica'09]. De plus, selon [Budanitsky'06] la mesure de Jiang-Conrath permet de mieux prendre en compte les erreurs dans l'écriture des termes que l'ensemble des autres mesures.

Dans [Rada'89], il a été suggéré que la similarité au sein d'un réseau sémantique peut être calculée en se basant sur les liens taxonomiques exprimant des relations de type hyperonyme/hyponyme, et plus précisément de type « est-un ». A partir de ce constat, la similarité sémantique peut être mesurée dans une taxonomie en calculant la distance entre les concepts en suivant le plus court chemin qui les sépare. Il est toutefois mentionné dans cet article que cette méthode n'est valable que pour tous les liens de type hiérarchique (« est-un », « sorte-de », « partie-de », ...) mais qu'elle pourrait être modifiée pour prendre en considération d'autres types de liens (relations de causalité, ...).

Dans les travaux de [Wu'94], ces derniers ont mis au point une mesure de similarité entre concepts pour permettre la traduction automatique. La mesure qu'ils ont proposée est définie par rapport à la distance qui sépare deux concepts de leur plus petit ancêtre commun (le plus petit concept qui les subsume tous les deux), ainsi que la racine de la hiérarchie. La similarité entre deux concepts C_1 et C_2 se calcule par la formule :

$$Sim_{WP}(C_1, C_2) = \frac{2 * prof(C)}{dist(C, C_1) + dist(C, C_2) + 2 * prof(C)}$$

Où, C est le plus petit ancêtre commun de C_1 et C_2 (en nombre d'arcs), $prof(C)$ est le nombre d'arcs qui séparent C de la racine, et $dist(C, C_i)$ le nombre d'arcs qui séparent C_i de C .

Dans [Zargayouna'05], la mesure de similarité proposée est inspirée de celle de [Wu'94]. Le lien père-fils est ainsi privilégié par rapport aux autres liens de voisinage en adaptant la mesure de Wu-Palmer qui pénalise dans certains cas les fils d'un concept par rapport à ses frères. L'adaptation de la mesure est faite au travers de la fonction de calcul du degré de spécialisation d'un concept ($spec$) qui mesure sa distance par rapport à l'anti-racine. Cela permet ainsi de pénaliser les concepts qui ne sont pas de la même lignée.

$$Sim_{ZS}(C_1, C_2) = \frac{2 * prof(C)}{dist(C, C_1) + dist(C, C_2) + 2 * prof(C) + spec(C_1, C_2)}$$

$$spec(C_1, C_2) = prof_b(C) * dist(C, C_1) * dist(C, C_2)$$

Où, $prof_b(C)$ correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine : \perp .

Les travaux menés dans [Torjmen'08] sur la RI multimédia basée sur la structure font l'hypothèse selon laquelle la structure d'un document XML peut se rapprocher d'une ontologie. Cela permet ainsi d'utiliser une mesure dérivée de celle de Wu-Palmer [Wu'94], et de [Zargayouna'05].

Différents travaux visent à prendre en compte la sémantique dans la RI structurée, ce qui passe par l'emploi des outils et ressources que nous venons de présenter. Toutefois, la plupart des approches retrouvées dans ces travaux prennent en compte seulement la sémantique du contenu textuel des documents et pas celle de leur structure. Selon [Zargayouna'05], très peu de travaux tentaient alors de combiner ces deux aspects. Il n'y avait alors que le système XXL¹¹ qui intégrait déjà une ontologie dans le processus d'indexation.

2.3 Systèmes de RI (semi-)structurée sémantique

La situation a beaucoup évolué depuis car de plus en plus de systèmes tendent à prendre en compte la sémantique. Le langage de requête du système XXL permet l'interrogation de documents XML avec une syntaxe proche de la syntaxe SQL¹². En effet, il est basé sur les langages de requêtes tels que XML-QL et XQuery auxquels il ajoute un opérateur de similarité sémantique noté « ~ ». Cet opérateur permet d'exprimer des conditions de similarité sémantique sur les éléments ainsi que sur leur contenu textuel. L'évaluation de la requête se base sur un calcul de similarité dans une ontologie ainsi que des techniques de pondération des termes. Le moteur de recherche XXL présente une architecture s'appuyant sur 3 structures d'index [Schenkel'05] :

- Index du chemin d'élément : permet l'accès aux noeuds parents, descendants et ancêtres d'un noeud donné. Il permet de calculer la distance entre ces deux noeuds).
- Index du contenu d'élément : permet de retrouver les éléments dans lesquels un terme apparaît. La pertinence des termes est calculée par le TF-IDF¹³ (qui est dans ce cas équivalent au TF-IEF¹⁴ car l'unité d'indexation est l'élément).
- Index ontologie : permet de retrouver des mots reliés sémantiquement à un mot donné. Il calcule pour cela une similarité qui peut être restreinte à un certain type de liens. A partir de cette valeur une mesure de similarité peut être calculée entre deux concepts. Cela nécessite la désambiguisation préalable des termes pour pouvoir les rattacher aux concepts.

Bien que les deux premiers index soient relativement classiques en RI (semi-)structurée, l'indexation sémantique par une ontologie constitue une approche intéressante.

Comme cela peut être vu dans le cas du système XXL, la prise en compte de la sémantique alourdit le processus d'indexation. Même s'il est communément accepté que la prise en compte de sémantique améliore les performances des systèmes,

¹¹ XXL : fleXible XML search Language

¹² SQL : Structured Query Language

¹³ TF-IDF : Term Frequency - Inverse Document Frequency

¹⁴ TF-IEF : Term Frequency - Inverse Element Frequency

certains travaux rapportent une dégradation des performances. En effet, il peut paraître préférable pour l'indexation des documents de prendre en compte un maximum de sémantique. Toutefois, l'indexation d'une trop grande quantité de termes peut s'avérer contre productive car elle accroît le temps de traitement et réduit la précision. Il est donc nécessaire de limiter la quantité de termes indexés en filtrant les termes avec une stop-list, avec des étiquettes marquant le discours, en fonction de leur fréquence, ou via des techniques statistiques. Ces techniques n'étant pas adaptées dans le cadre de l'utilisation de ressources sémantiques, certains travaux effectuent des regroupements de concept pour n'indexer que les concepts qui les subsument. D'autres travaux plus poussés tentent de trouver des coupes optimales en utilisant des critères basés sur la théorie de l'information [Seydoux'05].

Dans les travaux de [Zargayouna'05] concernant l'indexation sémantique et ayant abouti au prototype SemIndex (dédié à l'indexation sémantique) et SemIR (dédié à la partie RI), la dimension sémantique est prise en compte tant au niveau des termes que de la structure. La mesure de similarité entre concepts définie précédemment est utilisée pour désambiguïser le sens des termes en favorisant le sens rattaché au concept qui maximise la densité du réseau sémantique. L'originalité de l'approche consiste principalement dans la mesure de similarité utilisée pour enrichir la méthode de pondération des termes.

Une première version des travaux de [Zargayouna'04] a été reprise dans [Bellia'08] pour être intégrée à la mesure de Mercier-Beigbeder qui ne tient pas compte de la sémantique [Mercier'05]. Ainsi, dans [Bellia'07] cette mesure est enrichie de façon à prendre en considération des liens de similarité latents entre documents et à l'adapter au formalisme XML. D'autres systèmes de RI structurée sémantique peuvent être cités tels que CXLEngine¹⁵ [Taha'08], système dérivé de travaux précédents ayant abouti à OOXSearch.

Dans une étude de [Van Zwol'07] sur le système de RI XSee, l'approche suivie montre que la sémantique impacte positivement les performances des systèmes qui prennent en compte la structure des documents. En effet, ils montrent qu'une structuration sémantique riche d'un document contribue à l'amélioration significative des systèmes de RI structurée. Il est donc pertinent de prendre en compte la sémantique du contenu mais également celle de la structure.

3 Contribution

3.1 Intuition concernant la pondération sémantique floue

Lors de notre étude des travaux existants, nous avons pu identifier que les travaux menés dans [Zargayouna'04, Zargayouna'05] répondent relativement bien à la problématique que nous avions définie. Nous avons donc cherché un moyen d'en

¹⁵ CXLEngine : Comprehensive XML Loosely structured search Engine

améliorer les résultats. Nous allons donc proposer une extension de la formule de pondération sémantique présentée dans [Zargayouna'04] pour que celle-ci tienne compte d'un appariement flou entre les termes présents dans les documents de la collection et les termes reflétant les concepts d'une ressource sémantique lexicalisée (thesaurus ou ontologie). En effet, cela permet de gommer les éventuelles fautes de frappe lors de la rédaction de contenu, ou même des fautes commises par méconnaissance des règles inhérentes à la langue.

3.1.1 Pondération sémantique des termes

Dans [Zargayouna'04], le poids sémantique $SemW$ du terme t au sein d'une balise b d'un document d au niveau du vecteur sémantique correspond à la somme de son poids TF-IDF¹⁶ et les poids des termes t_i qui lui sont proches sémantiquement. Les termes t_i considérés comme étant sémantiquement proches sont ceux dont la similarité est supérieure à un seuil initialisé en fonction de la similarité entre le terme t et la balise b .

$$SemW(t, b, d) = TFITDF(t, b, d) + \frac{\sum_{i=1}^n Sim_{ZS}(t, t_i) * TFITDF(t_i, b, d)}{n}$$

3.1.2 Appariement flou entre termes

Notre première idée est de repartir de la formule de pondération sémantique proposée dans [Zargayouna'04] pour l'enrichir en prenant en compte les incertitudes quant à l'écriture des termes. Afin d'intégrer cet aspect, nous nous sommes inspirés des travaux de [Tambellini'07] en termes de gestion des données incertaines.

Etant donné que nous nous basons sur une ressource sémantique lexicalisée dans laquelle les concepts sont représentés par des termes, nous pensons qu'il peut être profitable d'effectuer un appariement flou entre les termes des documents de la collection et ceux de la ressource sémantique lexicalisée.

Ainsi, selon [Tambellini'07], deux termes t_1 et t_2 s'apparent :

- Selon leur concordance, c'est-à-dire leur positionnement relatif que nous notons $Conc(t_1, t_2)$.
- Selon leur intersection, c'est-à-dire les zones communes aux deux termes que nous notons $Inter(t_1, t_2)$.

Ainsi, la valeur de concordance notée $ValConc(t_1, t_2)$ est déterminée à partir de la caractérisation des termes en fonction de relations spatiales (dérivées des relations de Allen) : « *débute* », « *pendant* », « *termine* », « *chevauche* », « *concorde* », et « *ne concorde pas* ». Chaque caractérisation est alors associée à une valeur α_i . On peut souligner que dans les travaux de [Tambellini'07] cette valeur semble être fixée empiriquement.

La valeur d'intersection suit les conditions suivantes : elle vaut 1 si les termes sont égaux ($t_1=t_2$) et sinon elle vaut $ValInter(t_1, t_2)$. La problématique de l'incertitude est présente dans de nombreux domaines, on peut notamment citer les systèmes qui permettent de déterminer si deux termes sont phonétiquement identiques tels que

¹⁶ TF-IDF : Tag Frequency – Inverse Tag and Document Frequency

l'algorithme du Soundex¹⁷ et ses dérivés, tels que Metaphone qui en corrige certaines lacunes, et Double-Metaphone qui est une version plus pointue de Metaphone. Les systèmes de correction orthographique se basent également sur la problématique de l'incertitude des données en cherchant à rapprocher deux termes ayant des lettres en commun. Ces différents algorithmes sont utilisables pour déterminer $ValInter(t_1, t_2)$. Dans le cadre de l'implémentation d'un prototype nous prévoyons d'utiliser un algorithme phonétique et notamment Double-Metaphone.

On peut ensuite à partir de ces deux valeurs déterminer une valeur d'appariement entre les termes notée $ValApp(t_1, t_2)$:

$$ValApp(t_1, t_2) = ValConc(t_1, t_2) * ValInter(t_1, t_2)$$

Un terme t_1 présent dans un document peut être considéré comme présent dans la ressource sémantique si il existe un terme t_2 dans la ressource sémantique tel que leur concordance $Conc(t_1, t_2)$ prend la valeur « *concorde* » dans l'ensemble des relations de concordance défini précédemment, et si sa valeur d'intersection $ValInter(t_1, t_2)$ vaut 1. De la même façon, si sa concordance vaut « *ne concorde pas* », alors celui-ci est clairement absent de la ressource sémantique.

A partir de là, il est possible de définir l'ensemble des approximations d'un terme de la collection de documents comme étant l'ensemble des termes représentant les concepts de la ressource sémantique qui ne sont ni absents ni exactement présents. De façon plus formelle, cela peut s'écrire :

$$\sim t_i = \{t_{RS} \in C_{RS} \mid t_i \approx t_{RS}\}$$

Où $\sim t_i$ est l'ensemble des termes t_{RS} représentant les concepts C_{RS} de la ressource sémantique RS qui sont presque égaux à un terme t_i présent dans les documents de la collection.

3.1.3 Combinaison des approches : pondération sémantique floue des termes

On dispose ainsi de tous les éléments nécessaires à la définition de notre nouvelle formule de pondération des termes dérivée de celle de [Zargayouna'04] :

$$\begin{aligned} SemW(t, b, d) = & \propto \left(TFITDF(t, b, d) + \frac{\sum_{i=1}^n Sim_{ZS}(t, t_i) * TFITDF(t_i, b, d)}{n} \right) \\ & + (1-\propto) \max_{1 \leq j \leq n'} \left(ValApp(t, t_j) * TFITDF(t_j, b, d) \right. \\ & \quad \left. + \frac{\sum_{i=1}^n Sim_{ZS}(t_j, t_i) * TFITDF(t_i, b, d)}{n} \right) \end{aligned}$$

Où la première partie de la formule correspond à la formule de pondération sémantique des termes proposée par [Zargayouna'04], alors que la seconde partie

¹⁷ Une implémentation de l'algorithme de Soundex et d'autres algorithmes dérivés de celui-ci est fournie par le projet Codec (<http://commons.apache.org/codec>) de la fondation apache.

correspond à l'approximation du terme obtenant le meilleur score selon les mêmes critères que le terme initial. De plus, n correspond à la cardinalité de l'union des $\sim t_i$ pour les termes appartenant à la balise b du document d et, $ValApp(t_1, t_2)$ correspond à la valeur d'appariement des termes qui permet d'exprimer le degré de similarité entre deux termes. Enfin, les facteurs α et $1-\alpha$ sont là pour lisser les résultats. Par défaut, on affectera une valeur médiane de 0,5 à α .

Prenons un exemple concret : admettons qu'un document contienne un terme « *dairy* » (crêmerie) alors que l'auteur aurait voulu écrire « *diary* » (agenda). Il est évident que le terme « *dairy* » se trouve hors contexte et aura donc un poids très faible par la formule de pondération de [Zargayouna'04]. Toutefois, ces deux termes « *concordent* » et leur codage phonétique anglais est identique. Leur valeur d'appariement est donc maximale et vaut 1. Ainsi, il est probable que notre extension obtiendra un poids supérieur pour l'approximation « *diary* » car celle-ci ne sera pas pénalisée par une valeur d'appariement faible et obtiendra un score TF-IDF (+ poids des termes proches sémantiquement) élevé car il correspond bien au contexte du document, et apparaîtra donc logiquement plus souvent que le terme « *dairy* ».

Nous avons ainsi présenté une formule de pondération sémantique floue des termes présents dans les documents d'une collection. Cette mesure de pondération doit ensuite être intégrée au sein d'un système de RI structuré sémantique existant (ou développé sur la base d'un système existant) pour pouvoir être évaluée.

3.2 Architecture et implémentation du prototype

Le prototype que nous souhaitons développer pour valider notre formule de pondération devra posséder de multiples structures d'index pour pouvoir accéder à une collection indexée par la structure, le contenu, et le sens. De plus, on créera un index Double-Metaphone des termes de la ressource sémantique lexicalisée afin de pouvoir faire une comparaison phonétique rapide des termes présents dans les documents de la collection avec ceux de la ressource sémantique.

Afin de développer un prototype opérationnel en mesure de valider différentes hypothèses, une étude des bibliothèques et plateformes dédiées à la RI s'est avérée nécessaire. En effet, pour être en mesure d'utiliser le prototype sur la collection de documents appartenant à la campagne d'évaluation INEX, il est souhaitable de réaliser un prototype suffisamment performant pour supporter un passage à l'échelle.

Nous avons ainsi été amenés à étudier les outils suivants : Terrier¹⁸, Zettair¹⁹, Lemur Toolkit²⁰, Dragon Toolkit²¹, Lucene²², GATE²³. Toutefois, un de ces outils a particulièrement retenu notre attention de par sa flexibilité. En effet, GATE est une plateforme d'ingénierie linguistique développée à l'Université de Sheffield. Le

¹⁸ Terrier : TERabyte RetriEveR (<http://ir.dcs.gla.ac.uk/terrier>)

¹⁹ Zettair : Zetta IR (<http://www.seg.rmit.edu.au/zettair>)

²⁰ Lemur Toolkit : (<http://www.lemurproject.org>)

²¹ Dragon Toolkit : (<http://dragon.ischool.drexel.edu>)

²² Lucene : (<http://lucene.apache.org>)

²³ GATE : General Architecture for Text Engineering (<http://gate.ac.uk>)

principe de cette plateforme est d'appliquer successivement sous forme de pipeline des ressources linguistiques, et/ou des ressources de traitement sur un document ou un corpus de documents. Cette plateforme est écrite en Java et propose nativement de nombreux modules permettant de l'interfacer avec des outils variés comme l'analyseur morphosyntaxique TreeTagger, ou même avec des ontologies, ainsi qu'un module de RI basé sur Lucene, une librairie Java dédiée à la RI.

3.3 Evaluation

Comme cela a été mentionné précédemment, notre proposition a pour objectif final la participation à la campagne d'évaluation INEX. En effet, la collection de documents d'INEX est dérivée du corpus de documents Wikipedia XML et compte ainsi 659338 articles en provenance de la version anglaise de Wikipedia, ce qui représente un volume total de 4,5Go. En moyenne un article contient 161 nœuds XML avec une profondeur moyenne des arbres de documents de 6,72. Dans cette collection, la syntaxe originale du Wiki a été convertie en XML en utilisant à la fois les balises générales de la structure logique (article, section, paragraph, title, list et item), des balises typographiques (bold, emphatic), et des balises de liaisons fréquentes. Ainsi, la prise en compte de la sémantique sur les balises présentes dans la collection de documents XML d'INEX risque d'être peu concluante car ce type de balise ne porte pas de sémantique propre. Toutefois, étant donnée la nature de cette collection de documents (dérivée de Wikipedia), il doit y avoir des fautes dans l'écriture des termes présents dans les articles (contenu textuel des documents XML), ainsi que dans les balises employées pour structurer les documents. Le type de solution proposé peut-être être vu comme un moyen de renforcer la mesure de similarité sémantique en lui permettant de comparer des termes dont le concept n'avait pas pu être identifié à cause d'une écriture erronée. Il est bien sûr ensuite nécessaire de pondérer la mesure sémantique en fonction de la proximité phonétique des termes. Nous espérons ainsi réussir à améliorer les résultats lors de nos expérimentations.

4 Conclusion et perspectives

Nous avons présenté dans cet article un état de l'art des différents outils nécessaires à la prise en compte de la sémantique dans les systèmes de RI structurée classiques. Cela passe notamment par l'emploi de ressources sémantiques externes à la collection de documents, sur lesquelles il est nécessaire de disposer de mesures de similarité sémantique pour pouvoir effectuer des comparaisons entre concepts. Nous avons ensuite présenté brièvement différents systèmes de RI structurée sémantique pour terminer avec la proposition d'une mesure de pondération sémantique pouvant être intégrée dans un système tel que celui développé par H. Zargayouna.

Nous aurions souhaité être en mesure de donner un exemple d'application de cette nouvelle formule de pondération sémantique. Toutefois, notre avancement dans le développement d'un prototype ne nous permet pas encore de pouvoir exposer des résultats issus de nos expérimentations. L'appariement flou et la pondération que nous

proposons en complément de la mesure de similarité sémantique présentée dans [Zargayouna'04] peut bien sûr s'appliquer avec des ressources sémantiques telles que Wordnet. Le but final est de pouvoir s'appuyer sur DBpedia qui représente une ressource bien plus riche. En effet, l'appariement flou présente plus d'avantages sur une ressource comme DBpedia car les termes représentant les concepts (étiquettes) ne sont pas nécessairement présents sous une forme normalisée issue d'une lemmatisation. Cet appariement flou permettra également la prise en compte de termes mal épelés dans le contenu des documents de la collection.

La première évolution à court terme consiste donc dans l'implémentation d'un prototype opérationnel en mesure de nous permettre d'évaluer notre mesure de pondération sur la collection de documents de la campagne d'évaluation INEX.

Une autre évolution potentiellement intéressante dans un avenir assez proche serait l'application du même type de traitement aux termes de la requête que celui que nous avons proposé pour la pondération des termes des documents du corpus.

5 Références bibliographiques

- [Auer'07]. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: LNCS (ed.): 6th International Semantic Web Conference (ISWC 2007), Vol. 4825, Busan, Korea (2007) 722-735
- [Bellia'07]. Bellia, Z., Vincent, N., Stamon, G., Kirchner, S.: Extension sémantique du modèle de similarité basé sur la proximité floue des termes. 7èmes journées d'Extraction et de Gestion des Connaissances (EGC 2007), Namur (2007)
- [Bellia'08]. Bellia, Z., Vincent, N., Kirchner, S., Stamon, G.: Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. 8èmes journées d'Extraction et de Gestion des Connaissances (EGC 2008), Sophia-Antipolis (2008)
- [Budanitsky'01]. Budanitsky, E., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Proceedings of the NAACL 2001 Workshop on WordNet and other lexical resources (2001)
- [Budanitsky'06]. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32 (2006) 13-47
- [Demartini'08]. Demartini, G., Firman, C.S., Iofciu, T., Nejdl, W.: Semantically Enhanced Entity Ranking. In: LNCS (ed.): 9th International Conference on Web Information Systems Engineering (WISE 2008), Vol. 5175, Auckland, New Zealand (2008) 176-188
- [Denoyer'06]. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006) 64-69
- [Fellbaum'98]. Fellbaum, C.: WordNet: An electronic lexical database. MIT press (1998)
- [Formica'09]. Formica, A.: Concept similarity by evaluating information contents and feature vectors: a combined approach. Communications of the ACM 52 (2009) 145-149
- [Jiang'97]. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings on International Conference on Research in Computational Linguistics (1997)
- [Kobilarov'09]. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Lee, R.: Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. European Semantic Web Conference (ESWC 2009), Semantic Web in Use Track, Crete (2009) (à paraître)

- [Lin'98]. Lin, D.: An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. (1998) 296-304
- [Mercier'05]. Mercier, A., Beigbeder, M.: Application de la logique floue à un modèle de recherche d'information basé sur la proximité. Actes LFA 2004 (2005) 231-237
- [Navigli'09]. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41 (2009) 1-69
- [Rada'89]. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE transactions on systems, man and cybernetics 19 (1989) 17-30
- [Resnik'95]. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995) 448-453
- [Rosso'04]. Rosso, P., Ferretti, E., Jimenez, D., Vidal, V.: Text categorization and information retrieval using wordnet senses. Proceedings of the 2nd Global Wordnet Conference (GWC 2004), Czech Republic (2004) 299-304
- [Schenkel'05]. Schenkel, R., Theobald, A., Weikum, G.: Semantic Similarity Search on Semistructured Data with the XXL Search Engine. Information Retrieval 8 (2005) 521-545
- [Seydoux'05]. Seydoux, F., Chappelier, J.-C., Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N.: Minimum Redundancy Cut in Ontologies for Semantic Indexing Progress in Artificial Intelligence. Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP 2005) (2005) 486-492
- [Suchanek'07]. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web, Vol. 6. ACM, Banff, Alberta, Canada (2007) 203-217
- [Taha'08]. Taha, K., Elmasri, R.: CXLEngine: a comprehensive XML loosely structured search engine. Proceedings of the 2008 EDBT workshop on Database technologies for handling XML information on the web, Vol. 261. ACM, Nantes, France (2008) 37-42
- [Tambellini'07]. Tambellini, C.: Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue. Vol. Thèse. Université Joseph Fourier (Grenoble 1), Grenoble (2007) 182
- [Torjmen'08]. Torjmen, M., Pinel-Sauvagnat, K., Boughanem, M.: Towards a structure-based multimedia retrieval model. Proceeding of the 1st ACM international conference on Multimedia information retrieval. ACM, Vancouver, British Columbia, Canada (2008) 350-357
- [Van Zwol'07]. Van Zwol, R., Van Loosbroek, T.: Effective Use of Semantic Structure in XML Retrieval. In: LNCS (ed.): 29th European Conference on IR Research (ECIR 2007), Vol. 4425, Rome, Italy (2007) 621
- [Wu'94]. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, Las Cruces, New Mexico (1994) 133-138
- [Zargayouna'04]. Zargayouna, H., Salotti, S.: Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004 (2004)
- [Zargayouna'05]. Zargayouna, H.: Indexation sémantique de documents XML. Vol. Thèse. Université Paris-Sud (Orsay), Paris (2005) 227