HAL
open science

Actes de l'Atelier Recherche d'Information SEmantique RISE 2016

J-P. Chevallet, Catherine Roussey, H. Zargayouna

To cite this version:

J-P. Chevallet, Catherine Roussey, H. Zargayouna. Actes de l'Atelier Recherche d'Information SEmantique RISE 2016. Huitième Atelier Recherche d'Information SEmantique, RISE 2016 associé à la conférence CORIA 2016 de la Semaine du Document Numérique et Recherche d'Information (SDNRI), Mar 2016, Toulouse, France. pp.28, 2016. hal-02603508

HAL Id: hal-02603508

https://hal.inrae.fr/hal-02603508

Submitted on 16 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Huitième Atelier Recherche d'Information SEmantique RISE, Toulouse 8 mars 2016

Associé à CORIA@SDNRI 2016

Actes de l'Atelier Recherche d'Information SEmantique RISE 2016

Édité par
Jean-Pierre CHEVALLET, LIG, Grenoble (France)
Catherine ROUSSEY, IRSTEA, Clermont Ferrand (France)
Haïfa ZARGAYOUNA, LIPN, Paris (France)

Huitième édition Atelier Recherche d'Information SEmantique

Atelier Recherche d'Information SEmantique RISE, Toulouse 08 mars 2016

Associé à CORIA@SDNRI 2016

1. Introduction

Nous avons le plaisir d'organiser à Toulouse la huitième édition de l'atelier Recherche d'Information SEmantique, RISE 2016, associé à la Conférence Recherche d'Information et Applications (CORIA) dans le cadre de la semaine du Document Numérique et Recherche d'Information (SDNRI). Depuis 2011 l'atelier bénéficie du soutien de l'ARIA (Association francophone de Recherche d'Information et Applications).

Le but de l'atelier est de proposer un espace d'échange autour de la synergie entre acquisition et gestion de ressources sémantiques (ontologies, terminologies, thesaurii, ...) et la Recherche d'Information. Ces thématiques sont à la croisée du Web Sémantique, de l'Ingénierie des Connaissances, du Traitement Automatique des Langues et de la Recherche d'Information.

Nous accueillons cette année trois conférenciers invités :

- Jean-Pierre Chevallet « Modéliser la correspondance RI avec une logique : liens entre logique et sémantique ».
- Brigitte Grau « Répondre à des questions, à partir de textes ou de bases de connaissances ».
- Lynda Tamine Lechani « Sur l'usage de la sémantique en recherche d'information contextuelle ».

Pour la première fois, nous organisons une session entreprise, qui permet de créer des liens entre la recherche académique et industrielle. Deux entreprises ACTIA et Mondeca, sont invitées à exposer

leurs applications à base de sémantique.

- Axel Reymonet, Ingénieur R&D chez ACTIA Automotive (www.actia.com), « Mise en place d'un cycle de RI sémantique sur un domaine technique ».
- Arnaud Cassaigne, Ingénieur R&D et Gilles Delaporte Directeur commercial chez Mondeca (www.mondeca.com), « Solution d'alerte et de détection de risques fondée sur des technologies de recherche et analyse sémantique ».

La sélection de l'atelier a permis de faire la part belle aux travaux de doctorants en leur offrant, comme chaque année, un espace d'échange autour de leurs travaux.

2. Comité de programme

- BELLOT Patrice, LSIS Avignon (France)
- BERTIN Marc, STIH Paris (France), CIRST Montreal (Canada)
- BUSCALDI Davide, LIPN, Paris (France)
- CALABRETTO Sylvie, LIRIS Lyon (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- GRAU Brigitte, ENSIIE (France)
- HERNANDEZ Nathalie, IRIT Toulouse (France)
- KAMEL Mouna, IRIT Toulouse (France)
- ROUSSEY Catherine, IRSTEA, Clermont Ferrand (France)
- SALLABERRY Christian, LIUPPA, Pau (France)
- SALOTTI Sylvie, LIPN, Paris (France)
- SCHWAB Didier , LIG-GETALP, Grenoble (France)
- TAMINE LECHANI Lynda, IRIT, Toulouse (France)
- ZARGAYOUNA Haïfa , LIPN, Paris (France)

3. Remerciements

Nous tenons à remercier les auteurs, les membres du comité de programme, les organisateurs de ce SDNRI ainsi que les participants qui assurent chaque année la bonne tenue de l'atelier.

4. Table des matières

Modéliser la correspondance RI avec une logique: liens entre logique et sémantique
Jean-Pierre Chevallet …..... 5
Répondre à des questions, à partir de textes ou de bases de connaissances
Brigitte Grau …....…..... 6
Sur l'usage de la sémantique en recherche d'information contextuelle
Lynda Tamine Lechani …..... 7
Extraction des informations d'entreprises
Armel Fotsoh Tawofaing, Annig Le Parc-Lacayrelle, Christian Sallaberry et Tanguy Moal 8
Propagation d'activation dans les graphes pour la recherche d'information sémantique
Ines Bannour, Haïfa Zargayouna et Adeline Nazarenko..... 17
Exploring Deep Learning for Query Expansion
Mohannad ALMasri, Catherine Berrut et Jean-Pierre Chevallet..... 26

Modéliser la correspondance RI avec une logique: liens entre logique et sémantique

Jean-Pierre CHEVALLET

LIG (CNRS UMR 5217 - Université de Grenoble Alpes), Grenoble, France
Jean-Pierre.Chevallet@imag.fr

Résumé : Dans le domaine de la Recherche d'Information (RI), modéliser la notion de pertinence à l'aide d'une logique est une idée ancienne, puis que dès 1971, Cooper propose de définir la pertinence par un lien logique entre un document et une requête. Ce choix impose de représenter les documents et les requêtes comme des formules logiques. Cette idée a ensuite été particulièrement étudiée dans les années 1990, avec comme point de départ une intuition de Van Rijbergen sur le lien entre déduction logique et probabilité de pertinence. Pourtant, cette approche ne semble pas s'être effectivement développée pour être utilisée dans les moteurs de recherche actuels. Alors que la logique est également un outil pour modéliser et déduire des connaissances, il semble également approprié d'utiliser les outils mathématiques offerts par la logique formelle pour modéliser et mettre en œuvre un moteur de recherche sémantique.

Dans cet exposé, nous proposerons des clés pour mieux comprendre le paradoxe historique des promesses offertes par la logique pour la modélisation de la pertinence et les relatifs échecs quant à son usage effectif dans des moteurs de recherche d'information.

Cette présentation rappelle en premier les fondements de la logique formelle. Nous mettons l'accent sur les notions de syntaxe, de niveau de langage et de sémantique. Nous proposons ensuite un bref parcours historique de la modélisation logique pour la Recherche d'Information. Nous présentons en particulier l'idée fondatrice de Rijbergen, et des propositions plus récentes comme celle de Losada et Barreiro. Nous mentionnerons les propositions de fusion entre les logiques de descriptions, et les probabilités. Enfin, nous terminons avec les moteurs de recherche sémantiques et les perspectives nouvelles de recherche qui s'offrent avec l'usage de connaissances formalisées en logique pour calculer la pertinence.

Répondre à des questions, à partir de textes ou de bases de connaissances

Brigitte GRAU

ENSIIE, LIMSI,-CNRS, France
brigitte.grau@limsi.fr

Résumé : Les systèmes de question-réponse ont pour but de fournir des réponses précises à des questions posées en langage naturel, réponses qui peuvent être trouvées dans des bases de connaissances ou dans des textes. De cette façon, l'utilisateur n'a pas à savoir comment l'information est représentée ou comment elle est recherchée. Les questions auxquelles je m'intéresserai dans cette présentation sont des questions factuelles, du type "A qui est mariée la fille de Bill Clinton" par exemple, dont on peut trouver la réponse dans une base de connaissances comme DBPedia, aussi bien que dans des textes. Dans les deux cas, l'objectif est de produire une représentation de la question de manière à pouvoir la rapprocher de la formulation de la réponse. Selon le type de ressource interrogée, les systèmes existants reposent sur des processus différents, chacun devant gérer la variabilité linguistique entre la formulation de l'utilisateur et la forme cherchée.

Cette présentation permettra de balayer l'existant pour ces deux types de recherche : les problèmes qui se posent et les réponses apportées, afin de pouvoir faire un parallèle entre les recherches dans les deux types de ressources. En effet, bien qu'à l'heure actuelle les deux problématiques soient principalement abordées séparément, l'obtention de systèmes plus performants pourrait résulter d'approches hybrides alliant leurs spécificités et leurs complémentarités. Cela permettrait d'étendre le champs des informations que l'on peut fournir et les possibilités de les retrouver effectivement. Je présenterai les quelques travaux existant dans ce domaine pour conclure sur la manière dont on peut envisager cette hybridation.

Sur l'usage de la sémantique en recherché d'information contextuelle

Lynda TAMINE LECHANI

IRIT – Université Paul Sabatier, Toulouse, France
tamine@irit.fr

Résumé : La recherche d'information contextuelle concerne l'intégration de facteurs, autres que la requête et le document, dans la modélisation du processus de sélection de l'information pertinente en réponse à un besoin exprimé par un utilisateur. Ces facteurs sont multiples : utilisateur incluant ses centres d'intérêts, sa tâche, sa localisation et son voisinage social, le domaine, le dispositif matériel etc. Lors de cette présentation :

- j'introduirai les principes et verrous de la recherche d'information contextuelle,
- je mettrai en évidence les ressources et les méthodes utilisées pour la représentation sémantique du contexte (contexte utilisateur et domaine spécifique) et son intégration dans un modèle d'appariement requête-document, et
- je présenterai une revue de résultats empiriques issus des campagnes d'évaluation TREC et CLEF permettant d'identifier les apports et limites de l'usage de la sémantique dans le domaine ainsi que les pistes de recherche futures.

Extraction des informations d'entreprises

A. Fotsoh Tawofaing[*,] — A. Le Parc Lacayrelle[**] — C. Sallaberry[**] — T. Moal[*]**

[*] *Laboratoire LUIPPA, BP-1155, 64013 PAU Université Cedex, France*
[**] *Cogniteev, 2 Rue Doyen Georges Brus, 33600 Pessac, France*

RÉSUMÉ. *Ce papier propose un service de recherche d'entreprises sur le web. Le besoin d'information est exprimé par des critères thématiques (activités et métiers de l'entreprise, produits commercialisés, ...) ainsi que des critères spatiaux (lieu). Le système comprend un premier module de filtrage de sites web d'entreprises. Un second module analyse ces sites afin d'en extraire automatiquement toutes les informations contextuelles (activités, métiers, produits, contacts, adresses, ...). Notre proposition se distingue des services similaires que l'on peut trouver sur le web par (i) la richesse des informations thématiques (catégorisation détaillée des activités, des produits, des métiers) ; (ii) la provenance web des informations ; (iii) la combinaison de critères de recherche thématique, spatial et plein texte. Un prototype est développé pour mettre en œuvre notre proposition.*

ABSTRACT. *Searching information about local businesses is not a trivial problem to address. Most of existing services are supplied with manually recorded data. Based on the observation that more and more businesses are presented on the web, we propose in this paper a new approach, which consists to extract companies targeted information (addresses, activities, jobs, products, emails, fax) from websites, to supply a local businesses search service. The retrieval information module combines thematic, spatial and full-text criteria. A prototype of this service is implemented to experiment our proposal.*

MOTS-CLÉS : *Analyse du web, Extraction d'informations spatiales, Extraction d'informations thématiques.*

KEYWORDS: *Web mining, Spatial information extraction, Thematic information extraction.*

1. Introduction

De plus en plus d'entreprises sont présentes sur le web et publient de nombreuses informations relatives à leurs activités, leurs métiers, leurs produits et leurs coordonnées (adresses, numéros de téléphone, numéros de fax, adresses mail). Partant de ce constat, le projet Cognisearch Business vise à exploiter ces données pour offrir un service de recherche d'entreprises capable de répondre à des besoins d'information du type « charpente en chêne au sud de Poyartin ». Ce type de recherche comporte une dimension thématique (charpente en chêne) et une dimension spatiale (au sud de Poyartin).

Plusieurs solutions sur le web proposent des services permettant de répondre à ce type de recherche. Nous pouvons les classer en trois catégories principales : (i) les fournisseurs de données, comme Factual[1] ou Axciom[2], qui collectent et commercialisent les informations d'entreprises. Les informations alimentant ces services proviennent généralement des «données ouvertes», du crawl du web ou même de plateformes partenaires ; (ii) les annuaires, qui sont des bases de données d'informations d'entreprises consultables en ligne. Dans cette catégorie, nous pouvons citer des services comme Google Maps, Google My Business[3], Pages Jaunes, ou même société.com (annuaire d'entreprises françaises déclarées au registre du commerce) ; (iii) les réseaux sociaux, qui sont beaucoup plus orientés partage d'informations et d'appréciations portant sur les commerces, les places et les événements dans une région. C'est le cas des services comme Yelp[4], Foursquare[5], Facebook Places[6]. Dans les catégories (ii) et (iii) les données proviennent, en général, de contributions (salariés et utilisateurs) et de fournisseurs de données. Cependant, ces solutions (i) ne permettent pas de prendre en compte toutes les relations topologiques dans la dimension spatiale ; (ii) peuvent mal interpréter la dimension thématique dans certains cas ; (iii) s'appuient en grande partie sur des données saisies manuellement. (Ahlers, 2013) propose un système d'analyse de pages web pour enrichir la base de données entreprises des Pages Jaunes. L'approche proposée consiste à croiser les données des Pages jaunes, avec celles de l'annuaire DMOZ[7] pour identifier les pages web associées à une entreprise et en extraire des informations (adresses, numéros de téléphone, adresses mail, données commerciales, données fiscales). Cependant, la proportion d'entreprises existant dans une région et qui sont référencées sur DMOZ reste faible pour le cas de la France (par exemple, on a 2580 entrées DMOZ au total pour la région d'Aquitaine alors qu'il y a plus de 250000 entreprises déclarées au registre du commerce), d'où l'enrichissement limité de la base de données.

1. http ://www.factual.com/
2. http ://www.acxiom.fr/
3. http ://www.google.com/business/
4. http ://www.yelp.fr/
5. http ://fr.foursquare.com/
6. http ://fr-fr.facebook.com/places/
7. http ://www.dmoz.org/

Notre objectif est de construire un service de recherche d'information d'entreprises qui s'appuie sur des données publiées sur le web et qui combine des critères spatiaux (prenant en compte différentes relations topologiques) et des critères thématiques avec la recherche plein texte. Notre proposition se distingue des services existants par son indépendance vis à vis des données enregistrées manuellement ; en effet, notre service est alimenté uniquement par des données publiques et des informations extraites du web. Ce service se veut également plus précis dans l'interprétation de la zone spatiale couverte par les besoins d'informations en prenant en compte les différentes relations topographiques (Vaid *et al.*, 2005) exprimées dans les requêtes. A la différence de (Ahlers, 2013), les sites web d'entreprises à partir desquels se fait l'extraction d'information sont identifiés par une heuristique à partir des données du registre du commerce. Cette démarche permet d'avoir un corpus plus exhaustif que celui construit à partir de l'annuaire DMOZ. Par ailleurs, au delà des données de localisation (adresses), nous extrayons également les métiers, produits et activités d'entreprises.

La suite de l'article est structurée comme suit. La section 2 présente l'architecture générale de notre service. La section 3 explique le processus de constitution de notre corpus de sites web. La section 4 détaille l'annotation de ces sites, tandis que la section 5 présente l'indexation des données. Le prototype développé est décrit en section 6. La section 7 montre une première évaluation. Enfin, la section 8 conclut l'article et présente les perspectives envisagées.

2. Architecture du service Cognisearch Businness

Nous proposons un service de recherche d'informations d'entreprises qui s'appuie uniquement sur des données publiées sur le Web. Nous avons défini un modèle de représentation des entités entreprise qui est composé de deux parties. La première correspond aux données d'immatriculation (de base) des entreprises (nom officiel, numéro SIRET, ...). La seconde partie est composée des données extraites des sites web des entreprises qui sont relatives à ses activités, produits et métiers. L'architecture de ce service est décrite figure 1. Une première étape consiste à identifier sur le Web les pages publiées par les entreprises (figure 1.1). Une fois le corpus de pages web constitué, une étape d'annotation (figure 1.3) permet d'en extraire automatiquement les informations relatives aux activités, aux produits, aux métiers ainsi qu'aux coordonnées de chaque entreprise (adresses, numéros de téléphones, emails, fax). Les informations annotées viennent compléter les données d'immatriculation des entreprises afin de constituer un premier index d'informations d'entreprises. En parallèle, le texte des sites web est indexé afin de permettre la recherche plein-texte (figure 1.6).

3. Filtrage de sites web

Les données d'immatriculation, telles que renseignées au registre du commerce sont issues automatiquement de la ressource societe.com, et constituent les données de

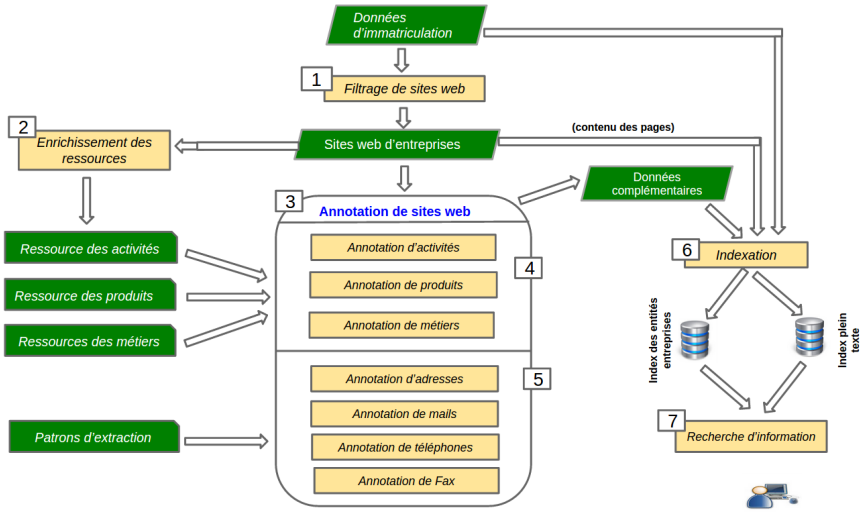


Figure 1. *Architecture de Cognisearch Business*

base d'une entité entreprise. Un traitement basé sur une heuristique combine ces données pour filtrer le web et retrouver, s'il existe, le site web associé à chaque entreprise (figure 1.1). Nous avons testé plusieurs combinaisons sur 300 compagnies françaises. Les annuaires d'entreprises et les réseaux professionnels ont été répertoriés dans une liste (appelée *liste noire*) afin d'être exclus lors de la recherche. L'algorithme retenu est le suivant. Pour chaque entreprise du registre du commerce, nous commençons par interroger le web en combinant son nom commercial (si il existe) avec sa localisation. Nous conservons, parmi les trois sites les plus pertinents, le premier qui n'est pas dans la liste noire. Si tous les sites sont dans la liste noire, nous recommençons le même principe mais en combinant cette fois le nom officiel de l'entreprise avec sa localisation. Si de nouveau, les trois premiers sites trouvés sont dans la liste noire, nous en concluons que l'entreprise n'a pas de sites web. L'ensemble des sites web ainsi obtenu, constitue le corpus d'entrée pour le processus d'annotation (figure 1.3).

4. Annotation des sites web

L'étape d'annotation (figure 1.3) permet d'extraire automatiquement des informations thématiques et spatiales relatives aux différentes entreprises. En ce qui concerne les informations spatiales, très peu de sites web utilisent les microformats ou des balises particulières pour les définir. Les adresses peuvent être situées n'importe où dans la page et les différentes informations qui les composent ne sont pas écrites forcément dans le même ordre.

4.1. *Extraction d'informations thématiques*

Afin d'annoter automatiquement les métiers, les produits et les activités contenues dans les sites web, nous avons construit, par transformation de modèles, trois ressources de type connaissance (sous forme d'ontologies au format OWL) décrivant, respectivement, les activités, les produits et les métiers d'entreprises. Ces ressources ont été construites à partir des deux hiérarchies d'organisation des activités et des produits définies par l'INSEE[8] (NAF[9] et CPF[10]) et de la hiérarchie définie par Pôle Emploi qui organise les métiers et emplois en catégories socio-professionnelles (ROME[11]). Le choix de ces ressources se justifie notamment par le niveau de finesse dans la description des différentes catégories.

L'ontologie relative aux activités ainsi construite est pauvre en vocabulaire. Nous avons de ce fait, mis en œuvre un processus pour son enrichissement de façon semi-automatique (figure 1.2). Ce processus utilise l'apprentissage pour regrouper les expressions des sites web qui sont communes à une catégorie précise d'activité. Pour cela, chaque classe de la hiérarchie des activités est associée à un ensemble de sites web d'entreprises. Un algorithme de clustering permet de mettre en évidence les expressions communes à tous les sites web d'une classe. Ces expressions constituent un vocabulaire potentiel pour enrichir la ressource. Une phase de validation par un expert intervient par la suite pour sélectionner les expressions pertinentes. L'algorithme de clustering que nous avons utilisé dans notre proposition est Latent Dirichlet Allocation (LDA) illustré dans (Blei *et al.*, 2003).

L'extraction d'informations relatives aux métiers, aux produits et aux activités d'entreprises est réalisée en utilisant les ontologies construites. Chaque syntagme nominal, trouvé dans le texte de la page web, est annoté de l'identifiant de la classe associée dans l'ontologie correspondante.

En ce qui concerne les emails, les numéros de téléphone et numéros de fax, des patrons sont utilisées pour leur extraction. Ces patrons ont été mis au point à partir de l'observation d'un échantillon de sites web. Par exemple, le patron permettant d'extraire les emails est le suivant :
$Email \rightarrow Login(\text{"@"}|\text{"}at\text{"})NomDomaine(\text{"."}|\text{"}(dot)\text{"})ExtensionDomaine$

4.2. *Extraction d'informations spatiales*

Deux approches principales peuvent être utilisées pour l'annotation automatique des adresses : une première utilise des patrons d'extraction ((Borges *et al.*, 2007), (Blohm, 2011), (Ahlers et Boll, 2008)) et une deuxième s'appuie sur les techniques d'apprentissage ((Loos et Biemann, 2008), (Taghva *et al.*, 2005)). Nous avons choisi

8. http ://www.insee.fr/fr/
9. Nomenclature Française des Activités
10. Classification des Produits Française
11. Répertoire Opérationnel des Métiers et des Emplois

Noms de champs	Exemples	Noms de champs	Exemples
CA : Complément Adresse	Résidence Rigaud	CP : Code Postal	64000
INV : Introducteur Nom Voie	Avenue	C : Commune	Pau
BP : Boite Postale	BP 1167	NV : Numéro Voie	10 ter
CS : Course Spéciale	CS 2587	D : Département	Gers
NC : Numéro Courrier	CEDEX 01	P : Pays	France
NVo : Nom Voie	Avenue de l'université		

Tableau 1. *Les différents champs pouvant composer une adresse*

une approche basée sur les patrons d'extraction. Ces approches exploitent en général des gazetiers répertoriant les noms de toutes les voies, le noms des villes, ... Dans le contexte français, une ressource complète contenant tous les noms de voie n'est pas disponible en accès libre. Une des difficultés est donc d'identifier le nom de la voie dans une adresse. En effet, il peut y avoir un ou plusieurs compléments qui peuvent être positionnés avant et/ou après le nom de la voie. Le tableau 1 répertorie les différentes informations que l'on peut trouver dans une adresse française.

L'observation d'un échantillon de 160 sites web d'entreprises a permis de définir des patrons, dont voici un extrait :

$$Adresse \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad NV? \quad NVo \quad CA?$$
$$((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad ((CP\,C) \mid (C\,CP)) \quad NC? \quad D? \quad P?$$
$$Adresse \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad ((CP\,C) \mid (C\,CP))$$
$$NC? \quad D? \quad P?$$
$$Adresse \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad NV? \quad NVo \quad CA?$$
$$((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad CP \quad NC? \quad D? \quad P?$$

Dans le premier patron, le nom de voie, le code postal et la commune sont obligatoires, et peuvent être complétés par d'autres informations («10 Rue du Maréchal Foch, 49000 Angers», «Résidence des Aubiers, 3e Étage, 14 ter Rue de la République, 64000 Pau, France»). Cette forme est la plus fréquente dans l'échantillon (75% des adresses). Pour le deuxième patron, le code postal et la commune sont obligatoires et le nom de voie est absent («Résidence Rigaud 33350 Mouliets-et-Villemartin»). Le patron 3 permet entre autre d'identifier le cas où on a un code postal et un nom de voie sans commune («10 Place de la République, F-33600»). Cette dernière forme est assez rare (moins de 4% des adresses de l'échantillon).

5. Indexation

Les annotations sont extraites des pages web pour la construction des entités finales à indexer. Ces annotations sont rajoutées aux données d'immatriculation de chaque entreprise ainsi que l'adresse du site web correspondant. Les coordonnées géolocalisées

correspondant à chaque adresse extraite sont calculées avant l'ajout. L'entité finale est stockée dans un index (figure 1.6). De plus, une opération parallèle consiste à extraire le contenu textuel des pages de chaque site web d'entreprise et à l'indexer.

Les index ainsi construits sont utilisés pour répondre à des besoins d'information qui supportent des critères d'interrogation multidimensionnels et exploitent les caractéristiques spatiales, thématiques et plein texte contenues dans les index (figure 1.7).

6. Prototype

Un premier prototype mettant en oeuvre notre approche a été développé. Il traite des données relatives aux entreprises de la région Aquitaine pour laquelle 254 000 entreprises ont été identifiées. Parmi ces entreprises, nous nous sommes intéressés uniquement à celles traitant de 6 domaines d'activités : commerce, construction, hébergement & restauration, enseignement, information & communication et activités scientifiques & techniques. Ceci a réduit la liste initiale à 115 000 entreprises. Le module de filtrage de sites web a permis d'identifier un site pour 22 000 d'entre elles. Le corpus d'analyse relatif à ces 22 000 entreprises est constitué de 550 000 pages web. Il a été constitué en utilisant l'outil Nutch[12] de la fondation Apache.

Le module d'annotation des sites web utilise la plateforme GATE[13] couplée avec le framework HADOOP[14] pour gérer la volumétrie et la scalabilité du processus. Cette étape a permis d'annoter 30 000 adresses, 44 000 labels d'activités, 12 500 labels de produits et 28 000 labels de métiers. Les entités entreprises, construites à partir des annotations extraites, ont été indexées sous Elasticsearch[15], ainsi que le texte des pages web. Les deux index ainsi construits (index des entités entreprises et index plein texte) ont une taille globale de l'ordre de 3 GB.

7. Évaluation

Nous avons fait une première évaluation du processus d'annotation pour les adresses et les activités. Etant donné qu'il n'existe pas de campagnes d'évaluation pour ce type d'entité, nous nous sommes inspirés du processus utilisé dans TREC (Voorhees et Harman, 2005). Les résultats sont résumés dans le tableau 2. Le rappel et la précision sont calculés pour chaque page, les valeurs globales de ces métriques étant calculées en faisant la moyenne arithmétique. La F_1-mesure est évaluée en faisant la moyenne harmonique des métriques globales obtenues pour l'ensemble de l'échantillon.

12. http ://nutch.apache.org/
13. http ://gate.ac.uk/
14. http ://hadoop.apache.org/
15. http ://www.elastic.co/

	Pertinentes	Annotées	Pertinentes et annotées	Précision	Rappel	F_1-mesure
Adresses	309	286	251	0,80	0,81	0,80
Activités	1131	1119	631	0,45	0,52	0,49

Tableau 2. *Évaluation du processus d'extraction d'information*

En ce qui concerne les adresses, notre échantillon est constitué de 240 pages web contenant au moins une adresse. 309 adresses ont été annotées par un expert, tandis que notre annotateur en a trouvé 286. Parmi ces 286, seules 251 sont jugées pertinentes, 13 étant incomplétement annotées. Ces résultats s'expliquent, d'une part, par le fait que nous avons considéré, dans nos patrons d'extraction, que toutes les adresses avaient obligatoirement un code postal, ce qui n'est pas le cas dans notre échantillon. D'autre part, certaines villes sont mal orthographiées.

En ce qui concerne les activités, notre échantillon est constitué de 100 pages web relatives à des entreprises travaillant dans le domaine de la toiture (charpente, couverture, zinguerie, ...). 1131 activités ont été annotées par un expert tandis que notre annotateur en a trouvé 1119. Parmi ces 1119, seules 631 sont pertinentes. Ces résultats s'expliquent par un défaut de vocabulaire dans l'ontologie des activités en dépit de son enrichissement. Par exemple, l'expert va annoter «charpente traditionnelle en chêne» alors que l'ontologie contient seulement le label «charpente en chêne».

8. Conclusion

Nous avons présenté dans cet article la première étape de la construction du service de recherche géo-localisé d'informations portant sur les entreprises. Cette étape est centrée sur l'extraction des informations sur le web et leur restructuration dans des index. Ce processus s'appuie sur plusieurs problématiques de recherche différentes, qui deviennent complémentaires dans le processus de construction des entités entreprises. Il s'agit notamment de l'apprentissage qui est utilisé pour l'enrichissement des ressources de type connaissance. De plus, le processus d'annotation de texte combine les approches basées sur les patrons d'extraction et celles exploitant les bases de connaissances. Il permet d'extraire les adresses des sites malgré les différences de formes que l'on peut trouver. Un premier prototype mettant en oeuvre notre approche a été développé. Il montre la faisabilité et l'intérêt de l'approche.

Une deuxième étape consistera à exploiter ces index pour répondre à des besoins d'information combinant les critères spatiaux, thématiques et plein texte. Une évaluation du service avec un jeu de requêtes représentatif sera également menée au terme de cette dernière étape.

9. Bibliographie

Ahlers D., « Business entity retrieval and data provision for yellow pages by local search », *Workshop of Integrating IR technologies for Professional Search*, 2013.

Ahlers D., Boll S., « Retrieving Address-based Locations from the Web », *International Workshop on Geographic Information Retrieval (GIR)*, ACM, New York, NY, USA, p. 27-34, 2008.

Blei D. M., Ng A. Y., Jordan M. I., « Latent dirichlet allocation », *Journal of machine Learning research*, vol. 3, p. 993-1022, 2003.

Blohm S., *Large-scale pattern-based information extraction from the world wide web*, KIT Scientific Publishing, 2011.

Borges K. A. V., Laender A. H. F., Medeiros C. B., Davis Jr. C. A., « Discovering Geographic Locations in Web Pages Using Urban Addresses (GIR) », *International Workshop on Geographical Information Retrieval (GIR)*, ACM, p. 31-36, 2007.

Loos B., Biemann C., « Supporting web-based address extraction with unsupervised tagging », *Data Analysis, Machine Learning and Applications*, p. 577-584, 2008.

Taghva K., Coombs J. S., Pereda R., Nartker T. A., « Address extraction using hidden markov models », *International Symposium on Electronic Imaging Science and Technology (IST/SPIE)*, p. 119-126, 2005.

Vaid S., Jones C. B., Joho H., Sanderson M., « Spatio-textual Indexing for Geographical Search on the Web », *International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, Berlin, Heidelberg, p. 218-235, 2005.

Voorhees E. M., Harman D. K., *TREC : Experiment and Evaluation in Information Retrieval*, The MIT Press, 2005.

Propagation d'activation dans les graphes pour la recherche d'information sémantique

Ines Bannour * — **Haïfa Zargayouna** * — **Adeline Nazarenko** *

* *Laboratoire d'Informatique de Paris Nord (LIPN, UMR 7030)*
Université Paris 13 – Sorbonne Paris Cité & CNRS
`Email: prenom.nom@lipn.univ-paris13.fr`

RÉSUMÉ. *Cet article présente un modèle de recherche d'information sémantique qui regroupe dans une même représentation le modèle sémantique et le modèle documentaire. La représentation en graphe que nous proposons permet des interrogations riches par termes, concepts ou même par documents. Nous appliquons sur cette représentation « unifiée » une méthode de propagation d'activation qui permet, à partir d'une requête, de retourner un ensemble de résultats ordonnés par pertinence, en propageant l'information de pertinence de proche en proche sur le graphe, depuis les éléments de la requête utilisateur.*

ABSTRACT. *This paper presents a model of semantic information retrieval which brings together in one representation the semantic and the documents' model. We propose a graph representation that allows rich queries with terms, concepts or even documents. We apply to this "unified" representation a method of spreading activation which is striggered by the elements of the user query and which enables to return a set of ordered relevant results.*

MOTS-CLÉS : *RIS, Ontologies, Graphe, Propagation d'activation*

KEYWORDS: *SIR, Ontologies, Graph, Spreading activation*

1. Introduction

L'avènement du Web, des moteurs de recherche et du Web Sémantique (Berners-Lee et Lassila, 2001) ont décuplé l'information disponible et les moyens d'accéder à cette information. Les modèles sous-jacents sont cependant hétérogènes : les moteurs de recherche reposent essentiellement sur la fréquence des mots et l'analyse de leurs distributions dans les documents ; la recherche d'information sémantique (RIS) exploite à l'inverse des connaissances sémantiques généralement consignées dans des ressources comme les ontologies ou les thesaurus.

La plupart des modèles de RIS existants sont des adaptations des modèles classiques de RI où l'exploitation de la ressource sémantique se réduit à une fonction de similarité intégrée dans la fonction de correspondance (Zargayouna *et al.*, 2015).

Des travaux récents (Castells *et al.*, 2007 ; Fernández *et al.*, 2011) proposent de combiner différents espaces d'indexation qui permettent à la fois d'exploiter au mieux les modèles sémantiques et de garder une représentation classique du modèle documentaire tels que le modèle vectoriel défini par Salton *et al.* (1975).

Le défi aujourd'hui consiste à proposer un modèle dédié à la RIS qui permette d'exploiter le modèle sémantique et le modèle documentaire d'une manière « unifiée ».

Nous proposons dans cet article une représentation en graphe qui permet d'intégrer dans une même représentation différents niveaux : le niveau conceptuel, le niveau terminologique et le niveau documentaire. Le modèle sémantique pris en compte peut être un thesaurus, un réseau terminologique ou une ontologie. Nous utilisons dans cet article le mot ontologie d'une manière générique qui regroupe ces différents types de ressources.

La construction du graphe se fait à l'aide d'un processus d'indexation et d'annotation sémantique qui permet d'établir des liens entre concepts, termes et documents. L'intérêt d'une telle représentation est de permettre de répondre à des requêtes plus complexes que juste des mots-clés. La mise en correspondance consiste en une méthode de propagation d'activation dans le graphe qui permet de raisonner de proche en proche et de rapprocher des documents, des termes et des concepts à un niveau sémantique.

L'application de la propagation d'activation en recherche d'information n'est pas récente (Preece, 1981 ; Cohen et Kjeldsen, 1987 ; Croft *et al.*, 1988 ; Salton et Buckley, 1988 ; Savoy, 1992). Crestani (1997) présente un état de l'art sur les travaux en recherche d'information qui ont proposé l'utilisation de la propagation d'activation dans des réseaux associatifs ou des réseaux sémantiques.

Plus récemment, Brouard (2013) montre qu'une propagation d'activation très simple dans un réseau associatif modélisant une couche documents et une couche termes donne des résultats équivalents au modèle vectoriel.

Dans la suite de cet article, nous décrivons la représentation en graphe, l'interrogation du graphe et le mécanisme de propagation proposé.

2. Représentation en graphe

Un modèle de recherche d'information concerne à la fois la représentation des documents et des requêtes (généralement dans le même formalisme) et une fonction de correspondance qui permet de rapprocher les deux représentations et d'attribuer un score aux documents.

Nous proposons une représentation en graphe qui permet d'intégrer dans une même représentation, le niveau conceptuel, terminologique et documentaire (voir figure1).

Le graphe est constitué d'un ensemble de nœuds et d'arcs valués et orientés $\langle N, E \subseteq N \times \mathbb{R} \times N \rangle$[1].

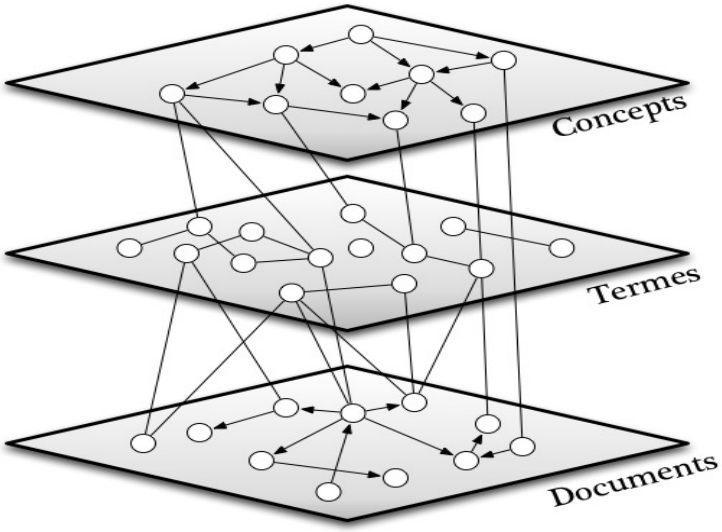


Figure 1. *Représentation en graphe des trois niveaux*

Les nœuds représentent trois types d'entités ($N = N_d \uplus N_t \uplus N_c$) :

– les *nœuds documents (N_d)* représentent les documents de la collection ;

– les *nœuds termes (N_t)* représentent le vocabulaire de la collection ainsi que les termes de l'ontologie ;

1. Dans la figure1, les arcs orientés dans les deux sens sont présentés par des arcs simples.

– les *nœuds concepts* (N_c) représentent les concepts de l'ontologie.

Les arcs traduisent les liens entre les différents nœuds. Ces arcs peuvent être pondérés, le poids d'un arc exprimant la force du lien qu'il traduit. Nous notons $w(i, j)$ le poids de l'arc allant du nœud i vers le nœud j. Les liens entre les nœuds sont :

– concept↔concept ($N_c \times \mathbb{R} \times N_c$) : ces arcs traduisent les relations ontologiques entre concepts ; il peut s'agir de relations hiérarchiques ou de rôles et le poids peut servir à distinguer la force des différents types de relations ontologiques ;

– concept↔terme ($(N_c \times \mathbb{R} \times N_t) \cup (N_t \times \mathbb{R} \times N_c)$) : ces arcs traduisent les relations entre concepts et termes qui peuvent apparaître dans le volet lexical d'une ontologie ; le poids peut éventuellement servir à distinguer le label « préféré » d'un concept (*preflabel*) par rapport aux autres termes qui lui sont associés (altlabel) ;

– concept↔document ($(N_c \times \mathbb{R} \times N_d) \cup (N_d \times \mathbb{R} \times N_c)$) : ces arcs représentent les liens de catégorisation qui peuvent être calculés lors de la phase d'annotation.

– terme↔terme ($N_t \times \mathbb{R} \times N_t$) : ces arcs représentent les liens terminologiques entre termes (ex. synonymie) ;

– terme↔document ou document-terme ($(N_t \times \mathbb{R} \times N_d) \cup (N_d \times \mathbb{R} \times N_t)$) : ces arcs représentent l'apparition d'un terme dans un document et leur poids peut naturellement traduire la fréquence d'occurrence du terme dans le document auquel il est relié ;

– document↔document ($N_d \times \mathbb{R} \times N_d$) : ces arcs représentent des relations « intertextuelles » entre documents (Mimouni *et al.*, 2014), notamment les liens de citation qui sont les plus fréquents.

Nous n'entrons pas ici dans le détail du calcul de ces poids, considérant que différents paramétrages sont possibles – depuis un graphe booléen (sans poids) à un graphe entièrement pondéré –, qu'ils reflètent différents choix de modélisation mais qu'ils sont tous compatibles avec le modèle à base de graphe que nous proposons.

3. Interrogation du graphe

La représentation unifiée de toutes ces informations sous la forme d'un graphe permet d'interroger les documents de manière plus riche que par les seuls mots-clés. On peut accéder au graphe par plusieurs points d'entrée : par les termes, les concepts ou les documents. Les nœuds pertinents renvoyés peuvent être des documents comme cela est classique mais aussi des termes, des concepts ou même une combinaison de plusieurs types de nœuds.

Il est ainsi possible de :

– poser des requêtes par des termes qui n'existent pas dans le vocabulaire de la collection, ce qui répond au problème de *term mismatch* décrit dans (Crestani, 2000) ;

– poser des requêtes par des concepts de structuration, qui ne possèdent pas forcément une dénotation lexicale dans le vocabulaire, mais qui servent à la catégorisation

du domaine et donnent un accès direct aux documents ;

– poser des requêtes par l'exemple, en soumettant comme requête un document et en recherchant les documents similaires.

Un processus de propagation d'activation permet d'intégrer à la recherche un mécanisme d'enrichissement de la requête qui prend en compte des termes synonymes, des concepts ou des documents reliés.

En effet, la fonction de correspondance que nous proposons repose un mécanisme de propagation dans le graphe : l'activation part des nœuds de la requête et se propage de proche en proche sur les nœuds voisins dans le graphe de sorte que les valeurs d'activation des nœuds du graphes traduisent, en fin de propagation, la pertinence des nœuds du graphe au regard de la requête.

4. Propagation d'activation dans le graphe

La propagation d'activation est un processus qui permet de propager une information de proche en proche sur un graphe. Cette information se représente par une valeur d'activation associée aux nœuds du graphe.

De manière générale, un algorithme de propagation d'activation dans un graphe, tel que décrit dans (Crestani, 1997), se compose d'une séquence d'étapes de propagation élémentaires et d'une condition d'arrêt. Il permet d'associer des valeurs numériques aux nœuds du graphe, les *valeurs d'activation*. Le graphe de départ est un graphe neutre, où toutes les valeurs d'activation sont nulles. L'initialisation du processus se fait par l'activation des nœuds qui figurent dans la requête. A chaque étape de propagation, le calcul de la valeur d'activation d'un nœud dépend de sa valeur précédente et des valeurs transmises par ses voisins lorsqu'ils s'activent. Le processus itératif se termine quand la condition d'arrêt est vérifiée. Le résultat est une distribution de valeurs d'activation sur l'ensemble des nœuds du graphe.

Nous détaillons les éléments clefs de cet algorithme de propagation dans ce qui suit.

4.1. *Valeurs d'activation des nœuds*

Étant donné $G = \langle N, E \rangle$, a une fonction d'affectation de valeurs d'activation sur le graphe G tel que $a : N \rightarrow \mathbb{R}^+$, nous notons $a(i)$ la valeur du nœud i dans l'affectation a.

Un algorithme de propagation peut être défini comme une suite finie de fonctions d'affectation $\pi = a_0 \ldots a_k \ldots a_n$ où a_0 représente l'affectation issue de la requête et a_n l'affectation vérifiant la condition d'arrêt.

L'activation initiale des nœuds (a_0) est fixée *a priori* avant que le processus de propagation d'activation ne soit déclenché. Ainsi, pour une tâche de RI dont la requête est

un ensemble de termes, les nœuds activés au départ peuvent correspondre aux termes de la requête, et leurs valeurs d'activation initiales dépendre de leur fréquence d'occurrence dans la requête. Le résultat de la requête est un sous-ensemble N_d de nœuds du graphe G déterminé en fonction de a_n et éventuellement restreint à un certain type de nœuds (documents, termes et/ou concepts).

4.2. *Étape de propagation*

Une étape de propagation se décompose elle-même en deux opérations : la sélection des nœuds à activer et la propagation de leur l'activation à leur voisins, qui se traduit par le calcul d'une nouvelle affectation a_k. Nous revenons dans la section 4.3 sur les règles que nous proposons pour la sélection des nœuds. Il suffit ici de considérer qu'à chaque étape de propagation, tous les nœuds actifs se déclenchent.

Le calcul d'une nouvelle affectation a_k dépend de l'affectation précédente a_{k-1} et de la structure du graphe. La fonction de propagation peut être également dépendante du type de nœud mais, pour simplifier la présentation, nous considérons ici une fonction de propagation identique pour tous les nœuds, qu'ils soient concepts, termes ou documents.

Soient un nœud i et a_{k-1} l'affectation issue de l'itération $k-1$. La valeur d'activation de i à l'itération k se définie comme suit :

$$a_k(i) = a_{k-1}(i) + \sum_{j \in pred(i) \cup actif(k-1)} a_{k-1}(j) * w(j,i) * 1/deg(j)$$

Les suites d'activation des nœuds sont croissantes. La valeur d'activation d'un nœud est augmentée à l'itération k de la somme des valeurs d'activation obtenues à l'itération précédente $(k-1)$ pour ses nœuds prédécesseurs $(pred(i))$ à condition qu'ils soient actifs, c'est-à-dire sélectionnés pour l'étape $k-1$ $(actif(k-1))$. Cette somme est pondérée par le poids des arcs reliant les prédécesseurs au nœud $(w(j,i))$. Elle est également atténuée par le degré des nœuds prédécesseurs $deg(j)$. Ce degré peut intégrer d'une manière globale tous les arcs comme il peut tenir compte des types de nœuds reliés. Il est par exemple possible de calculer le degré d'ambiguïté d'un nœud terme en prenant en compte le nombre de nœuds concepts auxquels il est relié. Entre un nœud terme et des nœuds documents, le degré permet de rendre compte du *document frequency*, le nombre de documents dans lesquels le terme apparaît.

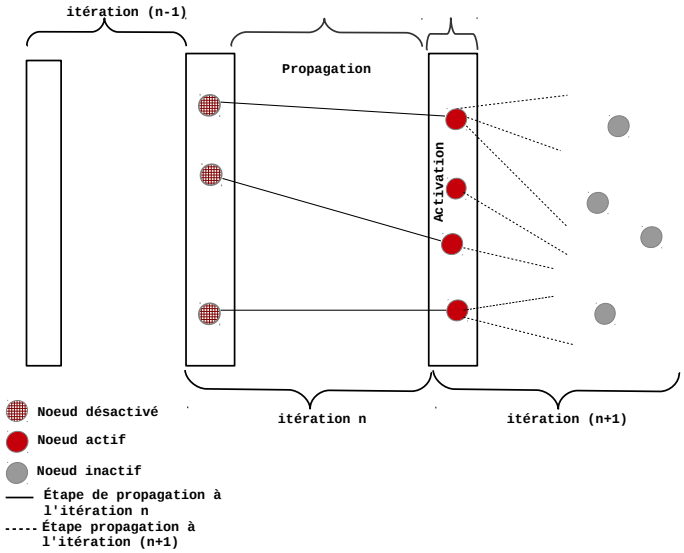


Figure 2. *Processus de propagation d'activation*

4.3. *Condition d'arrêt*

Dans le cas où le graphe contient au moins un cycle, l'itération du processus de propagation en partant d'une affectation initiale pourrait se poursuivre à l'infini en l'absence de condition d'arrêt. Il existe plusieurs façons de terminer un algorithme de propagation : on peut borner *a priori* les étapes de propagation, jouer sur la sélection des nœuds actifs ou garantir que les valeurs se stabilisent à un moment donné.

Dans ce travail nous utilisons la condition d'arrêt proposée par Rocha *et al.* (2004). Elle repose sur le contrôle de la sélection des nœuds. Tous les nœuds ayant une valeur d'activation non nulle peuvent-être sélectionnés mais un nœud ne peut être sélectionné qu'une seule fois. Ce mécanisme est contrôlé par une variable d'état associée à chaque nœud. Un nœud change d'état au cours du processus de propagation : il est *inactif* avant d'être atteint par « l'onde » de propagation, *actif* quand il est sélectionné pour se déclencher et activer ses voisins puis et *désactivé* une fois qu'il a été déclenché.

La propagation d'activation s'arrête donc quand il n'y a plus de nœuds actifs : il n'y a plus que des nœuds déjà désactivés ou des nœuds inactifs qui ne peuvent être atteint par la propagation d'activation.

4.4. *Analyse*

Le calcul d'activation dépend de la structure du graphe ($pred(i)$). Dans notre approche, tous les nœuds du graphe sont susceptibles d'être activés mais qu'un nœud ne peut se déclencher qu'une seule fois. Il s'ensuit que le calcul de la propagation est linéaire par rapport à la taille du graphe : $O(|N| * deg_{max})$, où $|N|$ est le nombre de nœuds du graphe et deg_{max} son degré maximal, c'est-à-dire le degré du nœud ayant le plus de successeurs. Le nombre d'étapes de propagation dépend de la forme du graphe : il est borné par le rayon du graphe ayant pour centre les nœuds de la requête.

5. Conclusion

Nous avons proposé une représentation unifiée en graphe du modèle sémantique et documentaire. Cette représentation permet d'exprimer des requêtes plus riches que les simples requêtes par mots-clés. Nous avons présenté une méthode de propagation d'activation dans le graphe qui permet de trouver des rapprochements entre requêtes et résultats en tenant compte des proximités sémantiques.

Les premières expérimentations ont porté sur le corpus de recettes de cuisine exploité par (Bannour et Zargayouna, 2012) et les premiers résultats sont encourageants. Des expérimentations à plus grande échelle sont en cours et vont permettre de calibrer les formules de propagation, de mettre en place les heuristiques de recherche nécessaires, d'ajuster la condition d'arrêt et d'étudier le passage à l'échelle d'un tel algorithme en utilisant la plate-forme de modélisation de graphes JUNG[2].

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

6. Bibliographie

Bannour I., Zargayouna H., « Une plate-forme open-source de recherche d'information sémantique », *COnférence en Recherche d'Information et Applications (CORIA)*, p. 167-178, 2012.

2. Java Universal Network/Graph Framework http ://jung.sourceforge.net/

Berners-Lee T., Lassila J. H. O., « The Semantic Web », *Scientific American*, 2001.

Brouard C., « Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation », *CORIA*, Neuchâtel, France, p. 1-10, April, 2013.

Castells P., Fernandez M., Vallet D., « An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval », *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n° 2, p. 261-272, 2007.

Cohen P. R., Kjeldsen R., « Information Retrieval by Constrained Spreading Activation in Semantic Networks », *Inf. Process. Manage.*, vol. 23, n° 4, p. 255-268, July, 1987.

Crestani F., « Application of Spreading Activation Techniques in Information Retrieval », *Artificial Intelligence Review*, vol. 11, n° 6, p. 453-482, 1997.

Crestani F., « Exploiting the Similarity of Non-Matching Terms at Retrieval Time », *Information Retrieval*, vol. 2, n° 1, p. 27-47, 2000.

Croft W. B., Lucia T. J., Cohen P. R., « Retrieving Documents by Plausible Inference : A Priliminary Study », *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, ACM, New York, NY, USA, p. 481-494, 1988.

Fernández M., Cantador I., López V., Vallet D., Castells P., Motta E., « Semantically enhanced Information Retrieval : an ontology-based approach », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 9, n° 4, p. 434-452, 2011.

Mimouni N., Nazarenko A., Paul È., Salotti S., « Towards Graph-based and Semantic Search in Legal Information Access Systems », *Legal Knowledge and Information Systems - JURIX 2014, Krakow, Poland*, vol. 271 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, p. 163-168, 2014.

Preece S., *A Spreading Activation Network Model for Information Retrieval*, University of Illinois at Urbana-Champaign, 1981.

Rocha C., Schwabe D., Aragao M. P., « A Hybrid Approach for Searching in the Semantic Web », *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, ACM, New York, NY, USA, p. 374-383, 2004.

Salton G., Buckley C., « On the Use of Spreading Activation Methods in Automatic Information », *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, ACM, p. 147-160, 1988.

Salton G., Wong A., Yang C. S., « A Vector Space Model for Automatic Indexing », *Commun. ACM*, vol. 18, n° 11, p. 613-620, November, 1975.

Savoy J., « Bayesian inference networks and spreading activation in hypertext systems », *Information Processing Management*, vol. 28, n° 3, p. 389 - 406, 1992.

Zargayouna H., Roussey C., Chevallet J. P., « Recherche d'information sémantique : état des lieux », *TAL (Traitement Automatique des Langues)*, vol. 56, n° 3, p. 49-73, 2015.

Exploring Deep Learning for Query Expansion

Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet

Université Grenoble Alpes
{mohannad.almasri,catherine.berrut,jean-pierre.chevallet}@imag.fr
LIG laboratory, MRIM group, Grenoble, France

User queries are usually too short to describe the information need accurately. Important terms can be missing from the query, leading to a poor coverage of the relevant documents. To solve this problem, automatic query expansion techniques leveraging on several data sources and employ different methods for finding expansion terms [2]. Selecting expansion terms is challenging and requires a framework capable of adding interesting terms to the query.

Different approaches have been proposed for selecting expansion terms. Pseudo-relevance feedback (PRF) assumes that the top-ranked documents returned for the initial query are relevant, and uses a sub set of the terms extracted from those documents for expansion. PRF has been proven to be effective in improving retrieval performance [4].

Corpus-specific approaches analyze the content of the whole document collection, and then generate correlation between each pair of terms by co-occurrence [6], mutual information [3], etc. Mutual information (MI) is a good measure to assess how much two terms are related, by analyzing the entire collection in order to extract the association between terms. For each query term, every term that has a high mutual information score with it is used to expand the user query.

Other approaches like semantic vectors and neural probabilistic language models, propose a rich term representation in order to capture the similarity between terms. In these approaches, a term is represented by a mathematical object in a high dimensional semantic space which is equipped with a metric. The metric can naturally encode similarities between the corresponding terms. A typical instantiation of these approaches is to represent each term by a vector and use a cosine or distance between term vectors in order to measure term similarity [1][7][8].

Recently, several efficient Natural Language Processing methods, based on Deep Learning, are proposed to learn high quality vector representations of terms from a large amount of unstructured text data with billions of words [5]. This high quality vector representation captures a large number of term relationships. We propose to investigate these term vector representations in query expansion. We then experimentally compare this approach with two other expansion approaches: pseudo-relevance feedback and mutual information.

First step, term vector representations are learned from a large amount of unstructured text data using Deep Learning. Several syntactic and semantic relationships are captured within these term vectors [5]. Each term is represented by a vector of a predefined dimension, a real-valued vector of a predefined dimension, 600 dimensions for example. Cosine or distance similarity could be used to

measure the strength of semantic similarity between two term vectors, and to list the top similar terms for a given query term.

Second step, we collect top similar terms for each query term using their vector representations. Then, top similar terms for each query term are used to expand the user query.

Experiments are conducted on four CLEF medical collections:

- Image2010, Image2011, Image2012: contain short documents and queries.
- Case2011: contains long documents and queries.

Table 1, shows some statistics about them, *avdl* and *avql* are average length of documents and queries, respectively. These medical collections provide a huge amount of medical text that we need in the training phase.

Table 1. Test collections statistics.

Corpus	**#d**	**#q**	**avdl**	**avql**
Image2010	77,495	16	62.12	3.81
Image2011	230,088	30	44.83	4.0
Image2012	306,530	22	47.16	3.55
Case2011	55,634	10	2594.5	19.7

We use language modeling framework to evaluate expanded queries. Two smoothing methods of language models are tested: Jelinek-Mercer and Dirichlet. We use word2vec to generate deep learning vectors [5]. The word2vec tool takes a text corpus as input and produces the term vectors as output.

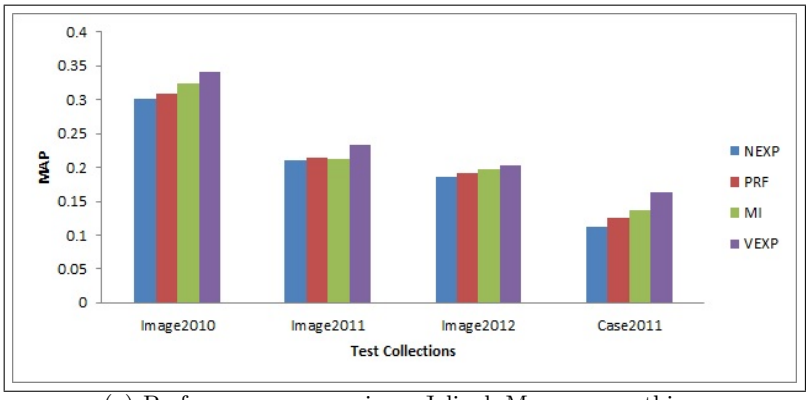
We compare three expansion methods: expansion using term vector representations (VEXP), Pseudo-relevance feedback (PRF), and Expansion using mutual information (MI). We use language models with no expansion as a baseline (NEXP).

Results are summarized in Figure 1. Experimental results show that the retrieval effectiveness can be significantly improved over the ordinary language models and pseudo-relevance feedback. We find that term vector representations, extracted using deep learning, are promising source for query expansion.

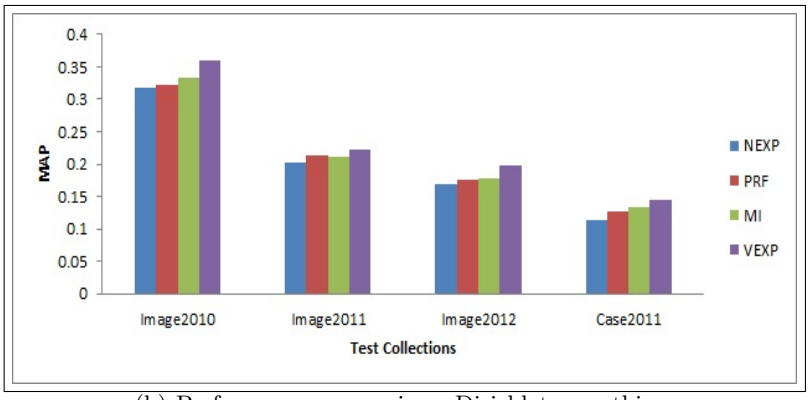
Deep learning vectors are learned from hundreds of millions of words, in contrast to PRF which is obtained from top retrieved document and MI which is calculated on the collection itself. Deep learning vectors are not only useful for collections that were used in the training phase, but also for other collections which contain similar documents.

References

1. Yoshua Bengio, Holger Schwenk, Jean-Sebastien Sencal, Frederic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. volume 194 of *Studies in Fuzziness and Soft Computing*, pages 137–186. Springer Berlin Heidelberg, 2006.



(a) Performance comparison: Jelinek-Mercer smoothing.



(b) Performance comparison: Dirichlet smoothing.

Fig. 1. Performance comparison using MAP on test collections: Image2010, Image2011, Image2012, Case2011.

2. Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012.
3. Jiani Hu, Weihong Deng, and Jun Guo. Improving retrieval performance by global analysis. In *ICPR 2006.*, pages 703–706, 2006.
4. Victor Lavrenko and W. Bruce Croft. Relevance based language models. SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.
5. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, 2013.
6. Helen J. Peat and Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *J. Am. Soc. Inf. Sci*, 1991.
7. Midori Serizawa and Ichiro Kobayashi. A study on query expansion based on topic distributions of retrieved documents. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 369–379. Springer Berlin Heidelberg, 2013.
8. D. Widdows and T. Cohen. The semantic vectors package: New algorithms and public tools for distributional semantics. In *ICSC*, pages 9–15, 2010.