Indexation et recherche conceptuelles de documents pédagogiques guidées par la structure de Wikipédia

Carlo Abi Chahine

THÈSE DE DOCTORAT
DE L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE ROUEN

Présentée par

Carlo Abi Chahine

pour obtenir le grade de
DOCTEUR EN SCIENCES
SPÉCIALITÉ INFORMATIQUE

Indexation et recherche conceptuelles de documents pédagogiques guidées par la structure de Wikipédia

Soutenue le 14 Octobre 2011, devant le jury :

Nathalie Aussenac-Gilles	Rapporteur
Yolaine Bourda	Rapporteur
Monique Grandbastien	Rapporteur
Jean-Pierre Pécuchet	Directeur de thèse
Nathalie Chaignaud	Co-encadrant
Jean-Philippe Kotowicz	Co-encadrant

Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes - EA 4108
INSA de Rouen - Avenue de l'Université - BP 8 - 76801 Saint-Étienne-du-Rouvray Cedex (France)

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE ROUEN | ENSEIGNEMENT SUPÉRIEUR ET RECHERCHE

Résumé

Cette thèse propose un système d'aide à l'indexation et à la recherche de documents pédagogiques fondé sur l'utilisation de Wikipédia.

L'outil d'aide à l'indexation permet de seconder les documentalistes dans la validation, le filtrage et la sélection des thématiques, des concepts et des mots-clés issus de l'extraction automatique d'un document. En effectuant une analyse des données textuelles d'un document, nous proposons au documentaliste une liste de descripteurs permettant de représenter et discriminer le document. Le travail du documentaliste se limite alors à une lecture rapide du document et à la sélection et suppression des descripteurs suggérés par le système pour rendre l'indexation homogène, discriminante et exhaustive.

Le corpus de documents d'étude est extrait des documents UNIT (l'une des Universités Thématiques dédiée à l'Ingénierie et aux Technologies) dont les domaines traités sont variés. Nous utilisons les données présentes dans Wikipédia pour leurs caractéristiques à la fois généralistes et semi-spécialisées, garanties par des communautés sérieuses. Les données et la structure de Wikipédia fournissent une base de connaissances hiérarchiquement semi-structurée utilisée pour extraire les descripteurs d'un document. Le modèle de représentation du document est un graphe orienté acyclique construit avec les termes du document et les relations hiérarchiques de la base de connaissances. Les nœuds de ce graphe (les titres des articles et des catégories de Wikipédia) sont appelés « concepts ».

Pour choisir les « concepts importants » du graphe permettant l'indexation, nous introduisons trois propriétés (l'« occurrence terminologique », la « généricité conceptuelle » et la « diversité conceptuelle ») à partir desquelles nous construisons une heuristique de cotation conceptuelle des concepts du graphe. Ceux-ci permettent trois opérations :

- l'extraction des termes importants du document,
- l'extraction des thématiques du document,
- et la désambiguïsation des termes polysémiques.

Ce modèle de représentation est aussi utilisé pour la recherche d'information. Pour cela, nous proposons une mesure de similarité entre deux graphes représentant deux documents (extraits ou complet) pour quatre opérations :

- la similarité inter-document pour un système de recommandation,
- la similarité intra-document pour l'analyse de la structure thématique du document,
- la similarité requêtes/documents pour un système de recherche d'information,

- et la désambiguïsation.

Après avoir réalisé une évaluation objective mettant en concurrence notre approche avec d'autres approches existantes et une évaluation subjective en présentant un prototype interactif à des documentalistes, nous sommes confortés dans l'efficacité de notre approche. L'utilisation de Wikipédia comme vocabulaire contrôlé apporte l'avantage de l'accessibilité et de l'exhaustivité pour notre analyse des documents pédagogiques UNIT. De nombreuses améliorations sont envisageables, notamment un prétraitement exploitant des approches TAL pour l'extraction des termes du document ou encore l'utilisation de méthodes d'apprentissage pour une sélection des descripteurs plus homogène avec les documents déjà indexés par les documentalistes ayant utilisé l'outil.

À Yazal.

Remerciements

Je tiens tout d'abord à remercier l'ensemble des doctorants avec qui j'ai travaillé mais surtout discuté et refait le monde à 90%. Merci Abu ! Merci Alina ! Merci Amandine ! Merci Amnir ! Merci Aurélie ! Merci Bassem ! Merci Emilien ! Merci Firas ! Merci Flo ! Merci GDD ! Merci Jérémie ! Merci Julien ! Merci Matthias ! Merci Nadine ! Merci Rémi ! Merci Polo ! Merci Ovidiu ! Merci Sourour ! Merci Xilan ! Merci Yann ! Merci Zach !

Ensuite, je voudrais remercier les membres du Litis. Sandra, Brigitte, Jean-François, Aziz, Mhamed, Stéphane, Habib, Stéphane, Sébastien, Alexandre, Nicolas, Romain, Frédéric, Laurent, Samiaa ! Merci ! Les pauses cafés ne seront pas tout à fait les mêmes sans vous.

Merci à mes encadrants, confidents, soutiens et amis : NC, JPP, JPK. Travailler avec vous a été un privilège.

Mes amis, je suis désolé d'avoir été absent ces quatre dernières années. Jad, Jad, Jana, Najla, Taher, Gaëlle, Antoine, Hassania, Lamiaa, Manu, Youssef, Soukaina, Jean-Baptiste, M. Kerkat, Benoit, Hadrien, Ludo, Loïc, Fanny, Hervé, Yacine, Caro, Arnaud, Omar, Momo, Nadir, Thierry et les autres dont je ne me rappelle plus le prénom, à très vite !

Asmaa, merci pour tout. T'es la meilleure !

A ma belle famille, Papa, Maman, Jee, Stef, Alison, Estelle, Sandro, Daadi. Je vous aime !

TABLE DES MATIERES

Problématique et contexte de recherche

Contexte de recherche

Le processus de recherche documentaire a largement évolué avec la facilitation de l'accès aux ressources grâce aux outils informatiques. Ceux-ci ont permis aux différents acteurs d'une bibliothèque (documentalistes, bibliothécaires et usagers) de simplifier leur tâche respective.

Le documentaliste et le bibliothécaire travaillent ensemble pour organiser les documents et les rendre accessibles aux usagers. Lors de la réception d'un document, le documentaliste l'analyse, le résume (cette tâche est aussi appelée condensation) et le décrit. Cette description consiste à extraire les thématiques et les notions abordées dans le document. On parle alors d'indexation du document. Ensuite, le bibliothécaire, en fonction de l'analyse du documentaliste, catalogue l'ouvrage, c'est-à-dire extrait les informations générales du document (titre, auteur, type de document, etc.). Enfin, le bibliothécaire met en place une notice pour chaque ouvrage et place ce dernier dans le rayon adéquat. La notice permet de récapituler le fruit du travail du documentaliste (indexation et condensation) et du bibliothécaire (catalogage). L'usager, quant à lui, afin de trouver un document, cherche ce qui l'intéresse dans les notices ou dans les rayons organisés.

Jadis, chaque ouvrage avait au minimum autant de notices que de mots-clés. Par exemple, si un ouvrage traitait à la fois de « mécanique des fluides » et de « thermodynamique », deux notices étaient alors créées, une pour chaque mot-clé. Pour le bibliothécaire, l'avènement du numérique a rendu possible la non duplication des notices et un gain d'espace considérable dans les bibliothèques. Pour l'usager, la recherche informatisée permet de gagner du temps en affichant immédiatement le résultat d'une requête sur un écran.

De plus, aujourd'hui, les documents eux-mêmes peuvent être numériques, c'est-à-dire que la recherche et la consultation documentaire se fait sur un même support informatique. Par conséquent, le support papier est délaissé au profit du document numérique. Les instituts et les entreprises ont désormais la possibilité de constituer leur propre bibliothèque numérique.

Le rôle d'une bibliothèque numérique est le même que celui d'une bibliothèque, à savoir fournir l'accès aux documents pour les usagers. Cependant, une première différence est la réduction des intermédiaires du métier du livre (édition, impression, publicité, transport, etc.), d'où la prolifération exponentielle des documents numériques. La seconde différence est la limitation des métiers et compétences permettant d'analyser et d'organiser l'ensemble des documents numériques. Les Systèmes de Recherche d'Information (SRI) prennent alors le relai afin de « remplacer » les documentalistes et les bibliothécaires.

L'analyse documentaire effectuée par les SRI se limite souvent à une analyse automatique prenant en compte la fréquence des mots dans les documents textuels. La lecture, l'analyse et

l'extraction de mots-clés par un expert, pourtant indispensables, ont une part négligeable, voire inexistante, lors de l'indexation automatique. Peut-on alors vraiment garantir à l'usager la pertinence des résultats d'une requête, sachant que personne n'a validé l'indexation ?

Les moteurs de recherche ont modifié notre rapport à la recherche documentaire. Les recherches en bibliothèque étaient souvent effectuées en utilisant les thématiques ou les mots-clés présents dans les notices. Avec des outils comme Google ou Yahoo, la notice et le document sont confondus. La recherche par thématique est inexistante et l'ensemble des mots des documents font office de mots-clés. La disparition progressive des documentalistes provoque également la disparition de l'extraction du contexte du document. Par exemple, si l'utilisateur cherche un document sur les « avocats », sa recherche peut mener au contexte biologique (avocat : le fruit) ou juridique (avocat : le métier). L'utilisateur devra par lui-même choisir son contexte. Les moteurs de recherche nous permettent de trouver des documents plus rapidement, mais la qualité des documents retournés est très discutable et ceci pour deux raisons. La première est la qualité intrinsèque du document. En effet, s'il y a absence de documentaliste, personne ne peut garantir la pertinence ou même la véracité d'un document. Le second problème est l'indexation du document, car si celle-ci est erronée alors les résultats du moteur de recherche le sont nécessairement. Il est donc fondamental qu'un documentaliste valide les documents et leur indexation. Vu le flux croissant de documents à indexer, leur analyse est impossible à réaliser par un unique documentaliste. Le nombre de documentalistes doit donc croitre proportionnellement au nombre de nouveaux documents.

Le documentaliste est donc bien au centre du bon fonctionnement des bibliothèques, qu'elles soient numériques ou non. Sa tâche est nécessairement amenée à évoluer et doit être assistée par l'outil informatique. Selon les besoins de l'organisation voulant mettre en place une bibliothèque numérique, il faut que celle-ci se pose la question suivante : faut-il favoriser l'indexation automatique au dépend de la fiabilité et de la pertinence des documents ? Le Tableau 1 synthétise les points forts et les points faibles des méthodes d'indexation automatiques et manuelles.

	Indexation automatique (sri)	Indexation manuelle (humain)
Documents analysés par jour	de l'ordre du million	de l'ordre de la dizaine
Contrôle, filtrage et validation	très faible	très important
Pertinence de l'extraction	variable selon les méthodes (souvent faible)	experte
Choix des mots indexant	intégralité du texte d'un document	meilleurs mots décrivant le document

Tableau 1 : Tableau de comparaison entre indexation automatique et manuelle

Comme mentionné précédemment, l'objectif de l'indexation est de classer un document dans une collection organisée mais surtout de pouvoir retrouver ce document en fonction des besoins de l'utilisateur. Dans l'hypothèse où la notice du document répond à ces besoins, le document doit répondre aux attentes de l'usager. Si l'indexation manuelle permet d'avoir une description fidèle du document grâce à l'expert, l'usager doit s'adapter à l'organisation, au langage et au vocabulaire de l'expert. Les normes, standards et bonnes pratiques suivis par les documentalistes préconisent des règles rigides pour le choix des termes descriptifs d'un document. En revanche, l'usager ne connaît pas ces règles. Par exemple, si l'usager cherche un document traitant des « équations aux dérivées partielles », il doit choisir le bon mot-clé. Dans le catalogue de la Bibliothèque Nationale de France (BNF), la vedette exacte est « équations aux dérivées partielles » (ni « équation aux dérivées partielles », ni « EDP » ne sont acceptables). L'usager doit s'adapter et trouver la bonne forme du mot-clé par lui-même. Parfois, certains outils permettent de reformuler l'intention de l'usager, lui demandant de se référer à d'autres mots-clés (Chaignaud, Delavigne, Holzem, Kotowicz, & Loisel, 2010; Loisel, 2008). Nous pensons que l'indexation automatique ne suffit pas à la recherche documentaire. En effet, l'indexation automatique classique ne permet de retrouver un document que si celui-ci contient le mot-clé recherché.

Les problèmes lexicaux et syntaxiques de l'usager ne sont pas le seul frein à la recherche d'un document, même bien indexé. Les SRI doivent être capables d'exploiter la sémantique exprimée dans l'intention (la requête) de l'utilisateur. En effet, s'il cherche un document traitant de l' « analyse en mathématique » un document traitant des « équations aux dérivées partielles » (ou des « EDP ») doit lui être suggéré même si l'indexeur (en l'occurrence, le documentaliste ou l'outil d'indexation automatique) ne l'a pas explicitement catégorisé avec le mot-clé donné.

Objectifs de la thèse

Cette thèse propose un système d'aide à l'indexation et à la recherche de documents pédagogiques fondé sur l'utilisation de Wikipedia. L'outil d'aide à l'indexation permet de seconder les documentalistes dans la validation, le filtrage et la sélection des thématiques, des notions et des mots-clés issus de l'extraction automatique d'un document. Notre outil est flexible et dynamique. La flexibilité se reflète dans la recherche de l'utilisateur qui n'a nullement besoin de s'adapter au langage ou au vocabulaire de l'indexeur. La dynamicité réside dans le fait que la notice remplie (de façon statique) par l'indexeur n'est plus l'unique base de recherche d'un document.

Cette thèse[1] s'inscrit dans le projet régional AICoTICE (Aide à l'Indexation et à la Cotation des ressources TICE) qui se propose de développer un outil d'aide à l'indexation des documents pédagogiques de l'Université Numérique Ingénierie et Technologie (UNIT). UNIT, qui rassemble des documents traitant des sciences dures et des technologies pour les élèves ingénieurs, est l'une des 7 bibliothèques numériques de l'Université Numérique Thématique (UNT) mise en place par le Ministère de l'Enseignement Supérieur et de la Recherche (UMVF pour le sport et santé, UNJF pour les sciences juridiques et politiques, etc.). Les documentalistes de UNIT rencontrent de grandes difficultés lorsqu'ils doivent indexer des documents du fait du large spectre de domaines scientifiques traités.

De façon générale, les mots-clés extraits par les documentalistes sont fortement liés aux thématiques, domaines et contexte du document. Si, pour un humain, la thématique d'un document est très rapidement identifiée (grâce à son titre ou à son résumé), il lui est cependant difficile de repérer les mots-clés. Il doit utiliser un dictionnaire spécialisé pour juger de la pertinence du mot-clé dans le contexte. Cette tâche est une perte de temps si chaque terme prête à confusion. Il est donc nécessaire de connaître la thématique avant de pouvoir extraire les mots-clés. A contrario, les mots employés dans un document sont hautement indicatifs de la thématique. Il n'est pas inenvisageable d'utiliser les mots du document pour en déduire ses thématiques et d'utiliser ces dernières pour extraire les mots-clés. Nous proposons donc de nous appuyer sur une base de connaissances qui fournisse les thématiques et leurs termes spécialisés.

Le rôle de la base de connaissances est de tenter de se « substituer » au savoir du documentaliste lors de l'indexation d'un document. Une fois les thématiques détectées, il faut choisir les « bons » mots-clés et ne pas se limiter à sélectionner tous les termes spécialisés des thématiques. Il est nécessaire de cibler très précisément le document afin d'être exhaustif dans le choix des mots-clés, en évitant au maximum le bruit dans les résultats. Il peut aussi être intéressant de sélectionner des mots-clés qui ne font pas partie du vocabulaire du ou des domaines traités. Par exemple, est-il

[1] financée par une bourse MESR et un monitorat

intéressant de choisir le mot-clé « équations aux dérivées partielles » (notion mathématique) dans un document traitant de physique ? La réponse dépend évidemment du document mais surtout de l'analyse du documentaliste.

La sélection de thématiques et de mots-clés ne suffit pas à décrire et à catégoriser un document. Par exemple, si un document, dont la thématique est la « physique », traite du « nombre de Reynolds », alors le document peut-être catégorisé dans la sous-thématique « turbulence », elle-même sous-thématique de la « mécanique des fluides ». Pour affiner la sélection des mots-clés par rapport aux thématiques, ces notions intermédiaires sont également sélectionnées ou rejetées. Il faut donc utiliser une base de connaissances représentant ces différentes notions sous forme d'une structure hiérarchique. C'est le cas des thésaurus, des ontologies ou des ressources termino-ontologiques qui représentent des concepts et des relations entre ces concepts (un concept étant une représentation d'un ensemble d'idées et de notions). Les relations entre concepts peuvent être typées. Par exemple, le concept « turbulence » est en relation avec « mécanique des fluides » par le type de relation « fait partie du domaine » ; le même type de relation lie « mécanique des fluides » et « physique ». Des inférences logiques sont possibles en utilisant ces bases de connaissances. Par exemple, la relation « fait partie du domaine » est transitive, ce qui veut dire que « turbulence » est implicitement en relation avec « physique » par la relation « fait partie du domaine ». La construction et la mise en place de telles bases de connaissances représentent un travail humain énorme et complexe. Même si l'informatique aide à la réalisation de ces bases de connaissances, des experts doivent collaborer pour valider et contrôler leur contenu.

Comme mentionné précédemment, les bases de connaissances hiérarchiques permettent de lier les concepts entre eux, c'est-à-dire qu'un concept est récursivement lié à plusieurs autres concepts. Il est alors possible de passer d'un concept à un autre en suivant les relations. En mathématique, ce type de structure peut être assimilé à un graphe. Les sommets du graphe sont les concepts et les arrêtes sont les relations. Il est possible pour chaque terme (un seul mot ou une séquence de mots) d'un document (associé à des concept de la base de connaissances) d'extraire un sous-graphe en isolant les concepts et les types de relation qui nous intéressent.

Notre outil d'aide à l'indexation utilise une ou plusieurs bases de connaissances hiérarchiques afin de détecter automatiquement les thématiques, sous-thématiques et mots-clés d'un document. Bien sûr, il n'est pas question de construire nous-même cette base de connaissances mais d'exploiter celles déjà disponibles. Nous utilisons la structure hiérarchique de la base pour réaliser :

- une **cotation** des concepts du document,
- une **sélection** des concepts importants du document,
- une **déduction** des thématiques,
- une **extraction** des mots-clés,
- une **désambigüisation** des termes polysémiques.

Les résultats (thématiques, mots-clés, concepts importants) sont ensuite soumis au documentaliste qui valide ou rejette les fruits de l'extraction.

Notre outil permet également de rechercher des documents sur Internet. Pour cela, nous proposons de traiter la requête en suivant le même processus que lors de l'indexation d'un document. Les termes de la requête sont traduits en concepts dont on extrait le graphe, les concepts sont ensuite cotés, sélectionnés, désambigüisés, etc. Nous comparons les graphes générés par la requête et ceux obtenus par chacun des documents indexés. Seuls les documents ayant une similarité proche de la requête sont retenus. Nous proposons donc une mesure de similarité entre deux graphes représentant deux documents (extraits ou complet) pour quatre opérations :

- la similarité inter-document pour un système de recommandation,
- la similarité intra-document pour l'analyse de la structure thématique du document,
- la similarité requêtes/documents pour un système de recherche d'information,
- et la désambiguïsation.

Si un utilisateur recherche un document sur l'« analyse mathématique », il est très souhaitable qu'un moteur de recherche affiche les résultats des documents classés dans cette catégorie, mais également (avec une pertinence éventuellement moindre) les documents classés uniquement dans la sous-catégorie « équations aux dérivées partielles ». Un tel système fournit des résultats intéressants, sous l'hypothèse que les documents à indexer soient des documents pédagogiques ou/et techniques. En effet, le vocabulaire employé dans ces documents semble permettre de vite cerner le contexte et d'identifier le domaine, sous réserve de connaître ce vocabulaire.

Il est question ici d'identifier le vocabulaire métier en utilisant une base de connaissances hiérarchique. Wikipedia, encyclopédie collaborative en ligne (www.wikipedia.org), utilise ce type de base. Les milliers de collaborateurs de Wikipedia ont fait émerger, de manière involontaire, un véritable réseau sémantique. En effet, chaque article de Wikipedia fait partie d'une ou plusieurs catégories et chacune de ces catégories font parties d'autres catégories, et ainsi de suite. Le choix de Wikipedia comme base de connaissances est justifié dans notre contexte :

- L'intégrité des articles et de la catégorisation est assurée par les utilisateurs de Wikipedia que nous considérons comme experts.
- Le type de relation « fait partie de la catégorie » est certes flou, mais représente bien une hiérarchie conceptuelle (relation générique/spécifique).
- Wikipedia fournit un vocabulaire contrôlé, pouvant servir à l'indexation des thématiques, sous-thématiques et mots-clés.
- Wikipedia est évolutif et dynamique, c'est-à-dire que les utilisateurs peuvent ajouter, modifier ou supprimer des articles et des catégories.

Cette thèse est organisée autour de deux parties.

Plan du manuscrit

Dans la première partie du manuscrit, nous étudions d'abord la notion fondamentale de document. Ensuite, dans un premier temps, nous introduisons les travaux existants dans le domaine de l'indexation et la recherche d'information dite classique, c'est-à-dire n'utilisant pas de base de connaissances dans leur processus. Dans un deuxième temps, nous nous intéressons à l'indexation et la recherche dite sémantique et conceptuelle, à savoir guider ces activités en utilisant une base de connaissances externes. Pour cela, nous définissons la notion de réseau sémantique, ontologie et ressource termino-ontologique ainsi que leur utilisation dans les SRI (cotation, sélection, désambigüisation etc.). Aussi, nous définissons la notion de similarité entre objets, notamment les graphes, afin de pouvoir effectuer un appariement requête/documents (socle d'un système de recherche d'information), ou éventuellement un appariement document/document (socle d'un système de recommandation). Pour finir, nous décrivons l'utilisation de Wikipedia dans les différents domaines de recherche, en particulier dans les domaines de l'indexation et la recherche d'information.

La deuxième partie de cette thèse traite de notre contribution pour l'indexation et la recherche des documents pédagogiques. Dans un premier chapitre, nous évoquons le modèle que nous utilisons pour l'indexation et la recherche documentaire. Chaque terme d'un document est détecté et contribue à la construction d'un graphe représentant ce document via une base de connaissances hiérarchique. Chaque terme du document est ainsi relié à des concepts de la base de connaissances. Le second chapitre de cette partie consiste à expliquer comment, une fois le graphe représentant le document obtenu, côter les concepts du document afin d'extraire les concepts les plus importants. Il est alors possible d'extraire les thématiques et les mots-clés associés à ces concepts. Aussi, l'extraction des concepts importants nous donne de fortes indications sur le sens correct des termes ambigus. Nous évaluons objectivement les résultats obtenus en utilisant nos approches et en exploitant Wikipedia comme base de connaissances. Nous confrontons nos résultats avec d'autres méthodes de cotation et sélection de concepts. Le troisième chapitre présente les méthodes que nous avons développées pour la recherche documentaire. La requête de l'utilisateur est transformée, à l'instar d'un document, en graphe. L'objectif est alors de trouver une métrique permettant d'évaluer la similarité entre une requête et un document en utilisant le graphe consolidé de la requête et le graphe consolidé du document et ce en considérant les choix de concepts du documentaliste lors de la phase d'indexation. L'idée est d'utiliser la même métrique pour comparer deux documents afin d'ouvrir la voie à un système de recommandation. Finalement, nous comparons nos résultats avec d'autres méthodes afin d'évaluer nos métriques. Le dernier chapitre présente un prototype permettant de concrétiser nos propositions, à savoir un outil d'aide à l'indexation et de recherche documentaire utilisant nos méthodes et opérations. Nous proposons, en effet, à l'utilisateur d'insérer un document sur une page

web, afin de l'aider à extraire les thématiques, les concepts importants et les mots-clés mais aussi d'analyser la structure thématique du document. Une autre partie du prototype permet de faire de la recherche d'information sur les documents UNIT en utilisant les métriques de similarité conceptuelle mises en place dans le chapitre précédant.

Enfin, à la lumière de nos résultats, nous dressons nos conclusions et perspectives pour la suite du projet AiCoTICE, ainsi que les leçons retenues lors de la conception et la réalisation de nos recherches.

Partie I. Etat de l'art

Introduction

Notre objectif est de concevoir un outil informatique permettant aux acteurs d'une bibliothèque numérique à vocation pédagogique de les aider lors de la phase d'indexation et de recherche de documents. L'outil ne doit pas se substituer au documentaliste et au bibliothécaire, mais plutôt les mettre pleinement à contribution pour rendre la recherche documentaire plus efficiente. Cette contribution consiste à sélectionner ou rejeter les thématiques, sous-thématiques et mots-clés suggérés par l'outil. Les choix du documentaliste permettront ensuite de catégoriser le document mais aussi d'améliorer la recherche pour l'utilisateur. L'indexation et la recherche documentaire sont les principales thématiques abordées ici.

Le pivot d'un système de recherche documentaire est le document. Il intervient lors de la phase d'indexation au moment de l'analyse documentaire, mais également dans la phase de recherche de documents, où l'utilisateur décide de la pertinence d'un document suggéré vis-à-vis de ses attentes. Il semble donc indispensable de définir la notion de document et de document électronique avant d'entreprendre une quelconque analyse sur ce type de support d'information. Dans le premier chapitre de cette partie nous traitons, en adoptant le point de vue de Marie-Francine Moens (Moens, 2000), du document numérique textuel, de ses propriétés et de ses caractéristiques. Plus particulièrement, nous passons en revue les spécificités des documents pédagogiques numériques.

Dans un deuxième chapitre, nous présentons les principales tâches d'un Système de Recherche d'Information (SRI), à savoir l'indexation et la recherche de documents. Les définitions d'indexation et de recherche documentaire y seront présentées ainsi que des techniques classiques d'indexation et de recherche et du *modus operandi* pour l'évaluation des SRI.

Nous étudions, dans un troisième chapitre, les SRI utilisant des bases externes de connaissances afin de « sémantiser » l'indexation et la recherche. Plusieurs types de base de connaissances y seront présentés ainsi que leur exploitation dans les SRI notamment pour la désambigüisation sémantique et le similarité entre documents.

Dans un quatrième chapitre, nous abordons l'utilisation d'une base de connaissances particulière : Wikipédia. Nous présentons les travaux exploitant Wikipédia pour les SRI. Nous expliquons comment extraire de Wikipédia les connaissances nécessaires à la création de bases de connaissances et leur utilisation dans différentes tâches comme l'indexation, l'annotation et la recherche d'information.

Finalement, cette partie se conclue par une discussion dans laquelle nous définissons l'ensemble des approches à adopter et/ou à adapter dans notre contexte de recherche.

Chapitre 1. Le document électronique

1.1 Introduction

La conception d'un système de recherche documentaire, notamment à vocation pédagogique, nécessite de définir la notion de document. Le terme document, provenant du latin « documentum », signifie « pièce écrite servant d'information, de preuve » ou « objet quelconque servant de preuve, de témoignage »[2]. Le document en tant que medium a vu sa définition modifiée au fil du temps. Le document était à l'origine un enregistrement d'un discours oral par le biais d'un codage, le texte. Aujourd'hui, il peut être simplement défini comme un support physique ou numérique d'information.

1.2 La communication par le texte

La communication par un texte nécessite un émetteur, celui qui envoie un message, et un destinataire, celui qui le reçoit. Afin que destinateur et destinataire se comprennent, nous devons émettre une hypothèse de connaissance mutuelle (mutual-knowledge hypothesis) (Gibbs, 1987), c'est-à-dire que l'interprétation du message faite par le destinataire et celle voulu par le destinateur peut correspondre.

Un texte est composé d'unités linguistiques ordonnées pour que ce dernier ait un sens. D'un texte émergent des attributs et des caractéristiques à deux niveaux : le niveau lexico-syntaxique (micro) et le niveau structurel (macro).

L'analyse « micro » du texte consiste à étudier ce texte d'un point de vue lexical et syntaxique. L'entité la plus petite lors d'un découpage de texte est le caractère et le phonème est la plus petite unité de son pour le discours. Ensuite vient le morphème, les composantes des mots (base, préfixe, suffixe, etc.), et les mots eux-mêmes.

Le mot est la plus petite unité linguistique. Au sens du texte, c'est une chaîne de caractères délimitée par un espace ou un caractère vide (virgule, point, etc.). Il existe plusieurs catégories de mots (verbes, noms communs, adjectifs, etc). Le mot a une signification appelée sens lexical (Ellis, 1999), c'est-à-dire qu'il se rapporte à ce qu'il symbolise, représente, dénote et connote.

Les « groupes de mots » peuvent être de différentes natures, notamment des syntagmes (en anglais : *phrases*) ou des collocations. Le syntagme est un groupement fonctionnel, souvent libre, de

[2] http://www.larousse.fr/encyclopedie/nom-commun-nom/document/44252

mots dissociables. Par exemple, le groupement de mots « la belle colombe à plumes blanches » est un syntagme composé des sous-syntagmes « la belle colombe » et « plumes blanches ». Le sens des mots est le même dans ces groupements que dans leur sens dissocié (par exemple « plumes » dénote un objet, « blanches » une couleur et « plumes blanches » un objet coloré). En revanche une collocation est un groupement de mots privilégiés ou indissociables. Par exemple, dans le syntagme « j'enseigne la physique quantique» on peut considérer « physique quantique » comme une collocation car ce groupement a un sens propre ; dans « j'aime les pommes de terre », « pomme » et « terre » ont un sens totalement différent que celui suggéré par le groupement.

Les phrases (en anglais : *sentence*) sont des séquences de mots et/ou de groupements de mots » utilisées pour informer, affirmer, infirmer, nier, suggérer, interroger, demander ou mentionner au destinataire des faits sur un sujet.

L'analyse « macro » du texte consiste à observer le texte comme une structure de segments. En effet, le texte peut être divisé en segments (partie du texte) en fonction d'un type de structure. (Moens, 2000) établit au moins quatre types de structure que nous ne détaillons pas tous dans cette partie. En fonction du type de texte et de l'intention de l'émetteur, ce dernier doit suivre, au moins, une structure afin d'assurer une bonne interprétation par le destinataire.

Une manière de segmenter le texte consiste à utiliser la *structure schématique* ou *superstructure* (Van Dijk, 1985, 1997). La structure schématique d'un document est l'agencement des segments du texte en fonction du type de texte. Qu'elle soit séquentielle ou hiérarchique, le destinateur et le destinataire ont connaissance de la structure schématique pour un type de texte. Prenons l'exemple d'un manuscrit de thèse scientifique, le destinateur (le doctorant) et le destinataire (un chercheur ou un rapporteur) partagent la connaissance de la structure schématique d'un manuscrit. Il consiste souvent d'un segment traitant du contexte et de la problématique, puis d'un état de l'art, suivi des travaux du doctorant et des résultats. Les expériences de (Dillon, 1991) sur les articles académiques, mettent en évidence l'existence d'une structure schématique partagée par le destinateur et le destinataire. Les expériences montrent que les lecteurs arrivent à prévoir, avec un haut niveau d'exactitude, la localisation d'une information dans le texte. Il est tentant de faire l'amalgame entre *structure logique* du texte (présentation des segments sous forme de chapitres, sections et paragraphes) et la structure schématique, mais ces deux manières de segmenter ne concordent pas toujours (Paice, 1991). La structure logique organise explicitement le texte alors que la structure schématique l'organise de manière floue et implicite (Van Dijk, 1997).

La *structure thématique* s'appuie sur l'organisation du ou des sujets et thèmes du texte. Les segments sont délimités en fonction des thématiques et sous-thématiques abordées. L'identification de la structure thématique nécessite de considérer le texte dans son ensemble, car il n'y a pas de moyen, a priori, de savoir quels thèmes sont abordés dans le segment suivant. Cette manière d'analyser le texte

comme un tout cohérent permet de comprendre pourquoi la structure thématique est aussi appelée *macrostructure* (Van Dijk, 1985).

L'évolution des thèmes et sous-thèmes abordés d'un segment à un autre est appelée progression thématique (Scinto, 1983). L'émetteur peut choisir de traiter du même thème dans différents segments (répétition thématique), traiter d'un rhème[3] d'un segment en thème du segment suivant (changement de thème – en anglais : *topic shift*), retraiter d'un thème précédemment abordé (retour sémantique – en anglais : *semantic return* (Allen, 1995)).

Les indices et signaux permettant d'identifier les thèmes abordés dans les segments sont dits surfaciques, c'est-à-dire que des éléments physiques du texte guident le destinataire dans la reconnaissance et l'interprétation du sujet.

Le premier signal surfacique pour l'identification du thème est l'ensemble des mots du segment décrivant ce thème. L'utilisation et la fréquence des mots dans un segment sont hautement significatives du thème abordé (Gerard Salton & McGill, 1983). Pour (Hearst & Plaunt, 1993), deux mots proches connotant un même thème dans un segment donnent une forte indication du thème abordé dans ce segment.

Le deuxième signal permettant d'identifier les thèmes d'un segment est la localisation des éléments surfaciques dans ce segment. (Kieras, 1982) démontre qu'une fois un segment défini, les éléments permettant d'identifier un thème se trouvent très souvent au début et parfois à la fin du segment.

Un troisième signal surfacique consiste à faire le parallèle entre structure thématique et logique du texte (c'est-à-dire entre segment thématique et logique). En effet, lors du passage d'un paragraphe (section ou chapitre) à un autre, les thèmes abordés peuvent être différents ; la progression logique suit la progression thématique.

Finalement, un quatrième signal surfacique est constitué des marqueurs linguistiques (en anglais : *cue words, cue phrases*) qui permettent de détecter les thèmes mais aussi la progression thématique (Kieras, 1982). Par exemple, des marqueurs tels que « pour revenir à » ou « comme mentionné dans le paragraphe XX » sont utilisés respectivement pour mettre en avant un retour sémantique ou un changement de thème, « aussi » et « de plus » pour signifier une répétition de thème.

1.3 Les documents électroniques

Le document est un moyen de communication interpersonnelle et sociale entre son rédacteur et le lecteur (Schamber, 1996). Le rédacteur utilise le document pour décrire et synthétiser ses objectifs de communication. Pour cela, il doit faire en sorte que le document réponde aux attentes du lecteur.

[3] Un rhème est une notion d'un texte qui n'est pas encore définie

Le document électronique tend à se substituer au document papier. Un document électronique est un objet informatique manipulable sur un ordinateur ou sur un appareil électronique. D'après (Schamber, 1996),

- les documents électroniques sont **facilement manipulables** par le créateur qui prend à son compte les avantages de l'outil informatique (copier/coller, contenu dynamique, etc.) ;
- les documents électroniques peuvent comporter des **liens internes ou externes** pour incorporer d'autres medias (textes, images, vidéos, etc.) ;
- les documents électroniques sont **indépendants** du support ;
- les documents électroniques sont **transportables** et **transmissibles** (par réseau) ;
- les documents électroniques sont **duplicables** ;
- la **recherche d'objet** (par exemple, un mot dans un texte) est possible au sein d'un document électronique.

C'est sur cette dernière caractéristique que reposent les SRI. La recherche d'un objet dans un document va permettre de retenir ce document si l'intention de l'utilisateur est d'obtenir un document contenant cet objet.

Dans le contexte de nos travaux de recherche, nous sommes amenés à manipuler des documents pédagogiques numériques et textuels.

(Bourda, 2001) décrit la notion d'objet (ou de ressource) pédagogique en donnant la définition IEEE[4]. Un objet pédagogique est défini comme « toute entité numérique ou non qui peut être utilisée, réutilisée ou référencée pendant des activités d'apprentissage assistées par ordinateur (enseignement – intelligent – assisté par ordinateur, environnements d'enseignement interactifs, systèmes d'enseignement à distance, environnements d'apprentissage collaboratifs) ».

Selon Bourda, les objets pédagogiques doivent répondre à certaines propriétés techniques :

- ils sont **autonomes**, c'est-à-dire leur utilisation ne dépend pas des autres objets ;
- ils sont **réutilisables**, c'est-à-dire que leur utilisation peut répondre à de multiples objectifs (les objets ne dépendent pas d'un contexte strict) ;
- ils ont d'autres propriétés : l'**agrégation** des ressources afin d'obtenir un ensemble cohérent d'activités pédagogiques, l'**indexation** afin de rechercher une ressource qui intéresse l'utilisateur, etc.

Les objets pédagogiques doivent également répondre à un besoin des utilisateurs. En effet, les ressources sont **sans superflu**, c'est-à-dire qu'à un faible niveau d'agrégation, la ressource doit fournir une information minimale (par exemple, une notion d'un cours doit avoir sa propre ressource associée).

[4] WG12: Learning Object Metadata - http://ltsc.ieee.org/wg12/

Dans la réalité, la plupart des plateformes pédagogiques ne suivent pas ces principes. En effet, les contributeurs de contenu (des enseignants en général) doivent segmenter leurs supports de cours en fonction des recommandations émises précédemment. Cette tâche nécessite une réorganisation totale de leur structure communicative, afin qu'elle coïncide avec la structure thématique. Supposons que cette segmentation soit faite, il faudra alors fournir pour chaque segment une notice descriptive du contenu.

Une manière de décrire les objets pédagogiques est l'utilisation des métadonnées, en particulier celles définies par des normes :

- Learning Object Metadata (LOM, norme IEEE) pour les documents pédagogiques ;
- LOM-FR, profil d'application LOM pour les documents pédagogiques francophones (norme AFNOR) ;
- SupLOM-FR, profil d'application LOM pour les documents pédagogiques francophones visant la formation supérieure (standard).

Le LOM et ses profils d'application fournissent un schéma de métadonnées comportant 9 catégories et un vocabulaire spécifique pour la description d'une ressource pédagogique (De La Passardière & Grandbastien, 2003). Les 9 catégories du LOM sont les suivantes :

1. Général : métadonnées décrivant les éléments primaires d'une ressource (identifiant, titre, auteur, etc.) ainsi que les mots-clés ;
2. Cycle de vie : métadonnées décrivant la version, l'état (brouillon, final, révisé, etc.) et les contributeurs (validateur pédagogique, implémenteur technique, etc.) d'une ressource ;
3. Méta-métadonnées : informations sur la notice elle-même (informations sur les indexeurs) ;
4. Technique : métadonnées décrivant techniquement une ressource, son format, les logiciels nécessaires à son utilisation (navigateur, Flash, QuickTime, etc.), sa durée pour une ressource audio ou vidéo, etc. ;
5. Pédagogique : caractéristiques pédagogiques d'une ressource, par exemple, le type (exercice, questionnaire, examen, présentation, etc.), le public cible (enseignant, apprenant, etc.) et son niveau (enseignement scolaire, licence, doctorat, etc.), la difficulté etc. ;
6. Droits : métadonnées exprimant les conditions légales d'utilisation d'une ressource ;
7. Relation : relations d'une ressource par rapport à d'autres ressources (« est un partie de », « est la base pour », « est la traduction de ») ;
8. Commentaire : information libre sur une ressource ;
9. Classification : emplacement d'une ressource dans une classification particulière.

L'un des objectifs fondamentaux du LOM est l'échange de ressources pédagogiques. (Passardière & Jarraud, 2005) explique les enjeux cruciaux de l'indexation des ressources pédagogiques d'un institut pour l'interopérabilité vis-à-vis des index d'autres instituts. L'étude

internationale menée par l'ISO-CN36 a pour vocation d'analyser les contenus des index LOM de cinq instituts et conclut que souvent seuls quelques champs sont renseignés (titre, auteur, mots-clefs libres, etc.). Aussi, lorsque d'autres champs sont renseignés l'interopérabilité est mise à mal par des entorses aux contraintes que constituent les vocabulaires du LOM. Le projet LUISA (Grandbastien, Huynh-Kim-Bang, & Monceaux, 2009) propose ainsi un framework permettant d'ajouter de la sémantique aux vocabulaires et aux termes employés pour la description des ressources pédagogiques, rendant possible l'interopérabilité des entrepôts de document. Finalement, une des préconisations issues de l'étude est l'homogénéité de l'indexation, c'est-à-dire donner un sens à son index par rapport aux autres index.

Connaître et comprendre les caractéristiques d'un document textuel est indispensable pour appréhender la recherche d'information. Nous avons montré les différents niveaux d'analyse documentaire (micro et macro) et leur importance pour l'interprétation du lecteur. Les ressources pédagogiques, selon les contributeurs de la norme LOM, induisent les activités suivantes :

- apprendre un thème ;
- échanger, collaborer et s'organiser autour d'un thème ;
- s'exercer, s'informer, se former, s'évaluer autour d'un thème.

En supposant qu'un document pédagogique réponde à ces activités, un document pédagogique textuel a nécessairement une structure thématique qui s'articule autour d'un thème principal.

La recherche documentaire commence lorsque l'on passe de la notion de document à celle de corpus de documents. Il s'agit alors de faire en sorte qu'un utilisateur retrouve un document qui l'intéresse. Les mécaniques classiques de recherche documentaire sont traités dans le chapitre suivant.

Chapitre 2. Les systèmes de recherche d'information

2.1 Introduction

Ce chapitre consiste à appréhender la notion de collection de documents textuels ou corpus, qui rassemble plusieurs documents généralement non homogènes. Cette non homogénéité se manifeste de plusieurs manières :

- différents supports (livre, magazine, CD, etc.) ;
- différents types (roman, documentaire, cours, etc.) ;
- différentes thématiques abordées ; etc.

Un usager ou un utilisateur qui cherche une information dans un corpus à deux possibilités : lire l'intégralité du corpus ou profiter d'un système de recherche mis en place en amont afin de retrouver rapidement un document. Deux notions se confrontent alors : information et document.

Dans le domaine de la recherche d'information, il y a un amalgame entre les termes « information » et « document ». En effet, d'après (*Vocabulaire de la documentation*, 1987), la recherche documentaire et la recherche d'information sont l'ensemble des « actions, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents ». Le syntagme « recherche d'information » est donc considéré comme une collocation signifiant « recherche documentaire ». A contrario, la « recherche de l'information » (Boulogne & Institut national des techniques de la documentation (Paris, France), 2004) est l'« ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes ». Dans notre contexte, nous utilisons la collocation « recherche d'information » ou RI, pour faire référence à la recherche documentaire.

L'objectif principal des SRI est d'aider un utilisateur à trouver, à retrouver ou à découvrir des documents susceptibles de l'intéresser. Son intérêt ou sa demande d'information (en anglais, *request*) s'exprime à travers une requête (en anglais, *query*). D'après (Ingwersen, 1992), la requête est la traduction de la demande d'information vers un langage compréhensible par un SRI. Parfois la demande et la requête sont confondues comme dans le cas d'une requête en langue naturel. Le plus souvent, l'utilisateur doit s'adapter au SRI afin de reformuler sa demande d'information sous forme de requête (par exemple, chercher « mathématiques » plutôt que « Je cherche des documents traitant des

mathématiques ») (Loisel, 2008; Chaignaud et al., 2010). Un SRI doit ensuite être capable de fournir à l'utilisateur des documents pertinents pour sa requête.

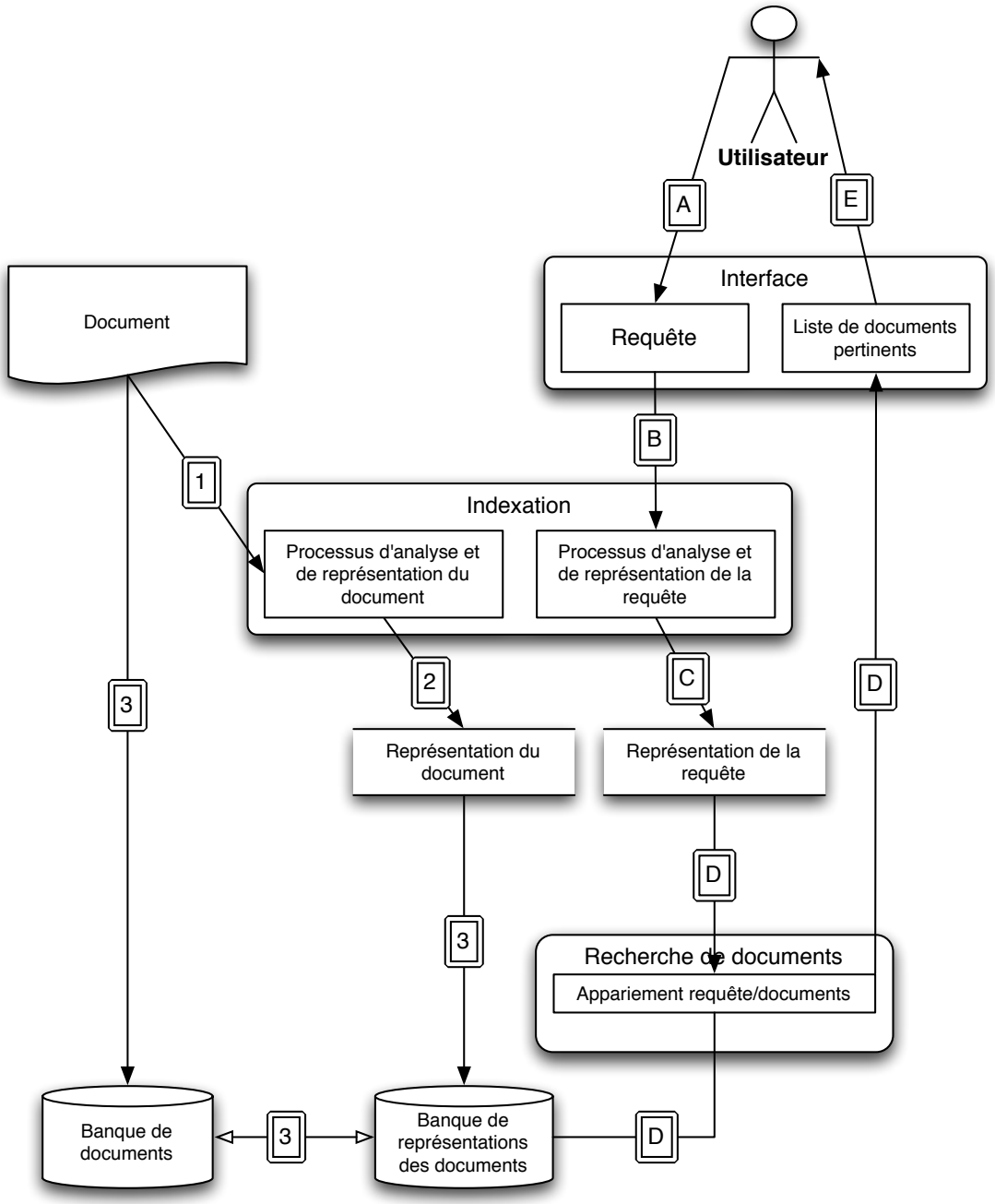


Figure 1 : Architecture générale des SRI

Afin de mettre en place un processus informatique permettant de trouver un document dans un fond documentaire par le biais d'une requête d'un utilisateur, différentes étapes sont nécessaires notamment l'indexation et la recherche documentaire. Ces deux étapes sont indépendantes et sont traitées séparément dans ce mémoire. La Figure 1 résume le processus global régissant les SRI :

- **L'indexation du document** :

  Cela consiste à lire et analyser (1) le document afin de générer une entité représentant ce document (2). Ainsi, à chaque document, est associée une représentation lors de la phase d'indexation (3). Cette représentation du document est alors stockée dans une banque de données (3) pour être mise à disposition d'un système de recherche de documents.

- **La recherche du document dans un fond documentaire :**

  Cela consiste à transformer la requête d'un utilisateur (B), c'est-à-dire à formuler explicitement la demande de l'utilisateur (A) en une entité représentant cette requête (C). Cette représentation est confrontée aux représentations des documents (D) afin de restituer les documents pertinents pour cette requête (E).

2.2 L'Indexation de documents textuels

L'indexation est le processus qui permet de créer une courte description ou caractérisation du contenu d'un document textuel sous forme de représentation suivant un modèle (Moens, 2000; G. Salton, Fox, & H. Wu, 1983). Toutefois, en informatique, l'indexation peut aussi être l'ensemble des techniques mathématiques permettant d'optimiser une recherche d'informations.

Dans le cadre des documents textuels, les éléments d'indexation sont, dans la plupart des cas, des mots, des termes ou des phrases issus ou non du document original. On parle alors de termes d'indexation ou de descripteurs qui sont les points d'entrée du document par un SRI. Avant de choisir les descripteurs d'un document, l'indexeur, humain ou machine, fait sa sélection en fonction d'un vocabulaire libre ou contrôlé.

2.3 Vocabulaire d'indexation

Lors de la conception d'un SRI, la question cruciale du vocabulaire d'indexation ou langage d'indexation (D. B. Cleveland & A. D. Cleveland, 2000) arrive en premier. Deux types de vocabulaire sont possibles : un vocabulaire libre ou un vocabulaire contrôlé. Le vocabulaire libre permet à l'indexeur de choisir des termes d'indexation sous forme de mots, de termes et de phrases quelconques en langue naturelle, qu'ils soient ou non dans le document. En revanche, le choix d'un vocabulaire contrôlé impose à l'indexeur un ensemble de descripteurs figés et préétablis, choisis en amont du processus d'indexation. En général, la collocation « terme d'indexation » s'applique pour les termes d'un vocabulaire libre ou contrôlé alors que « descripteur » s'applique uniquement à ceux d'un vocabulaire contrôlé. Dans ce mémoire, nous utiliserons uniquement le terme « descripteur », même lorsque le vocabulaire est libre. Le choix d'un langage libre ou contrôlé présente des avantages et des inconvénients.

Le vocabulaire libre (ou langage libre) consiste à utiliser comme descripteurs des termes en langue naturelle, souvent issus du texte original, et permettant de décrire son contenu (Stephen P. Harter, 1986). Lorsque les descripteurs sont dans le document, on parle d'indexation par extraction (en anglais, *extraction indexing*). Ce type d'indexation par vocabulaire libre se trouve dans les moteurs de recherche de type GOOGLE où le vocabulaire d'indexation est l'ensemble des termes des documents indexés. Le vocabulaire libre permet un très fort pouvoir d'expressivité et de flexibilité. De plus, n'étant en relation avec aucune base de connaissances, la base d'index est ainsi portable. En revanche, les descripteurs issus du langage libre peuvent s'avérer ambigus ou même trop spécifiques pour la recherche de document (l'utilisateur doit alors utiliser les bons termes pour trouver les bons documents).

Le vocabulaire contrôlé (ou langage contrôlé) consiste à utiliser des descripteurs dont l'origine est différente du document lui-même. Ces descripteurs proviennent généralement d'autorités compétentes (des experts) spécialistes du domaine traité. L'utilisation d'un vocabulaire contrôlé permet de lever les ambigüités sur les descripteurs (par exemple, il existe 2 descripteurs pour « Avocat », à savoir « Avocat_(fruit) » et « Avocat_(justice) »). Le vocabulaire contrôlé permet de catégoriser les documents en fonction des descripteurs choisis (Lancaster, 2003) et d'utiliser des techniques de filtrage (Belkin & Croft, 1992) (par exemple, chercher tous les documents traitant d'« Avocat_(fruit) » et non d'« Avocat_(justice) »). Par contre, l'utilisation d'un vocabulaire contrôlé conduit fatalement à un nombre limité de descripteurs pour décrire le document lorsque le vocabulaire n'est pas assez expressif. De plus, il faut que l'utilisateur soit familier avec le vocabulaire employé.

La catégorisation de texte est une activité de classification de texte qui répond à la question d'appartenance ou non d'un objet à une classe. L'appartenance à une classe peut-être binaire (l'objet appartient ou n'appartient pas à la classe) ou floue (l'objet appartient à la classe à un certain degré) (D. B. Cleveland & A. D. Cleveland, 2000; K Spärck Jones, 1973). Lorsque la classe en question est rattachée explicitement à un descripteur, on parle alors de catégorisation. L'attribution à un document d'un descripteur d'un vocabulaire contrôlé implique nécessairement la catégorisation du document.

2.4 Indexation manuelle, semi-automatique et automatique

L'indexeur est la personne ou le mécanisme en charge de l'indexation. Dans les moteurs de recherche sur internet, l'indexeur est le plus souvent logiciel. Il existe 3 types de processus d'indexation : l'indexation manuelle, l'indexation semi-automatique et l'indexation automatique.

L'indexation manuelle est la prise en charge de la tâche d'indexation par un être humain, dans l'idéal un expert du domaine. Même si la lecture complète d'un document est théoriquement indispensable à son indexation, en pratique un expert ne peut pas se permettre de tout lire et a une lecture sélective guidée par son expérience (connaissance du type de la structure schématique, logique

et thématique). Il est malgré tout difficile de comprendre les mécanismes cognitifs de sélection des descripteurs par un expert. En revanche, l'un des premiers réflexes d'un expert est de trouver les thèmes du document afin d'analyser le contenu (en anglais, *content analysis*) (Hutchins, 1985). Celle-ci permet de détecter les thèmes abordés dans le document de différentes manières. Par exemple, les groupes nominaux qui apparaissent souvent comme sujets de phrases ou au début d'un paragraphe donnent une bonne indication sur le thème. Une fois la lecture effectuée, l'expert sélectionne les termes importants et généralise les thèmes abordés. Finalement l'expert doit trouver une correspondance entre les termes employés et les thèmes sélectionnés avec les descripteurs (langage libre ou contrôlé). Dans le cas d'un vocabulaire libre, les choix opérés lors de la phase de sélection et de généralisation sont directement utilisés comme descripteurs. (Blair & Maron, 1985) indique cependant que l'utilisation d'un vocabulaire libre, pour l'indexation humaine, conduit à des inconsistances dans le choix des descripteurs même pour des experts expérimentés. Dans d'autres cas, les plus fréquents, les experts ont recours à un vocabulaire contrôlé pour l'indexation d'un document. (Beghtol, 1986) décrit le processus cognitif de la sélection des descripteurs qui s'appuie sur une représentation mentale des connaissances.

L'indexation semi-automatique (ou supervisée) consiste à utiliser un logiciel afin d'aider un expert dans la sélection des descripteurs. L'expert devra alors valider, rejeter ou ajouter des descripteurs suggérés par le logiciel.

L'indexation automatique (ou non supervisée) est la prise en charge de l'indexation par un logiciel. Elle a pour but d'extraire des descripteurs, permettant de décrire synthétiquement et exhaustivement le contenu du texte original. (Gerard Salton, 1989) propose un mode opératoire pour l'extraction des descripteurs d'un document textuel en langue naturelle :

1. identifier tous les mots du texte par analyse lexicale ;
2. supprimer les mots non pertinents à l'aide
    a. d'un dictionnaire de mots vides (appelé aussi anti-dictionnaire)
    b. d'un dictionnaire de mots pertinents pour un domaine donné
    c. d'une analyse fréquentielle supprimant les mots trop fréquents ;
3. réduire les mots restants à leur forme canonique (lemmatisation et lexémisation) ;
4. regrouper certains mots restants sous forme de syntagmes ;
5. remplacer les mots par les descripteurs d'un vocabulaire contrôlé ;
6. affecter à chaque descripteur un poids caractérisant quantitativement son importance.

Selon les choix de conception du logiciel permettant une indexation automatique, certaines étapes sont optionnelles. Par exemple, dans le cas d'une indexation par descripteur libre, le concepteur peut choisir de ne faire que l'étape d'identification (étape 1) et l'étape de pondération (étape 6).

L'étape d'indentification (l'étape 1) des mots se fait par analyse lexicale. Il est ici question de rechercher dans un texte l'intégralité des mots (termes ou tokens) dans ce contexte. Malgré l'aspect simpliste de cette étape, certaines difficultés peuvent survenir. Il arrive, par exemple, qu'un terme n'ait de sens qu'en utilisant une ponctuation (« etc. », « cf. », etc.). Un autre exemple est l'apostrophe « ' » considérée comme ponctuation (« l'avocat », «qu'avant », etc.) ou comme une lettre (« aujourd'hui », « prud'homme », etc.).

L'étape de suppression de termes non pertinents pour l'indexation et la recherche d'information (l'étape 2) permet d'exclure, avant même la phase de pondération, certains termes comme descripteurs potentiels. Différentes approches existent pour mettre en œuvre cette décision. (Luhn, 1957; Gerard Salton, 1989) propose de construire un anti-dictionnaire composé de mots vides, par l'étude de la fréquence des mots dans un corpus. Plus un mot est présent dans un corpus et moins ce mot est pertinent pour l'indexation. Une autre manière de procéder est d'effectuer une analyse syntaxique afin de déterminer la nature et la fonction grammaticale des mots. En fonction de celles-ci on peut décider de supprimer un mot ou, au contraire, de le retenir comme descripteur potentiel (Hoch, 1994). Par exemple, on peut décider de ne retenir que les noms (nature), les sujets des phrases (fonction) ou de supprimer les conjonctions, les adverbes et les adjectifs (nature), ainsi que les compléments d'objet (fonction). Certaines approches consistent également à supprimer systématiquement les mots trop courts (en terme de nombre de lettres) sauf s'ils sont explicitement présents dans un dictionnaire (Ballerini et al., 1996).

L'étape de lemmatisation ou de lexémisation (étape 3) consiste à mettre en correspondance un terme avec sa forme canonique (le lemme) ou son radical (le lexème). Par exemple, dans le syntagme « un couple aimant » la forme canonique de « aimant » est le verbe « aimer » et le radical est «aim-» (car toutes les formes fléchies du verbe « aimer » ont comme préfixe «aim-»). De même, le nom « aimant » (au sens de l'objet magnétique) a comme forme canonique et comme radical « aimant ». (Hafer & Weiss, 1974; Porter, 1980) proposent des algorithmes de lexémisation et (Koskenniemi, 1983) des algorithmes de lemmatisation. L'intérêt de cette étape est d'effectuer une meilleure analyse de la fréquence des mots en utilisant leur forme canonique plutôt que leurs formes fléchies. Cette étape va aussi permettre d'utiliser la forme canonique du terme pour le mettre en correspondance avec la forme canonique d'une entrée d'un dictionnaire et donc d'un descripteur potentiel.

L'étape de regroupement (étape 4) consiste à concaténer les termes pour former des syntagmes ou des collocations. L'intérêt de cette étape est double. Dans un premier temps, il s'agit de mettre en correspondance les termes concaténés avec une entrée d'un dictionnaire. Par exemple, les 3 termes « pomme », « de » et « terre » sont susceptibles d'être définitivement concaténés afin de correspondre

à une entrée « pomme de terre » dans un dictionnaire. Dans un second temps, (Smeaton, 1992) affirme que les groupes de mots sont plus révélateurs du contenu du texte que les mots simples. Une manière d'extraire les groupements syntagmatiques consiste à analyser un corpus afin de pouvoir en extraire les groupements fréquents de termes considérés alors comme syntagmes dits statistiques (G. Salton, C. Buckley, & Smith, 1990).

L'étape 5 consiste à mettre en correspondance des termes et groupements de termes avec les termes d'un vocabulaire contrôlé. Cette tâche ne consiste pas uniquement à comparer les chaines de caractères des formes canoniques des termes avec ceux des entrées d'un dictionnaire, puisque un terme peut avoir plusieurs entrées potentielles (par exemple, le terme « avocat »). Nous reviendrons sur cette étape dans le chapitre traitant de l'indexation sémantique et conceptuelle (cf. : Chapitre 3).

Enfin, l'étape 6, que nous développons dans la section suivante, consiste à pondérer les termes d'un document textuel afin de pondérer et choisir ses descripteurs.

2.4.1 Pondération des termes

La pondération des termes d'un document textuel est cruciale pour la sélection des descripteurs lors de la phase d'indexation. Pondérer un terme revient à lui affecter un score pouvant quantifier deux qualités : un score élevé pour un terme peut signifier qu'il est représentatif du contenu du texte ou alors discriminant par rapport au contenu des autres textes.

2.4.1.1 Pondération représentative et discriminante

Dire « le terme T du document D a un poids de X » ne fournit aucune information sur la manière de calculer ce poids, ni sur ses aspects représentatifs ou discriminants. Dans la démarche de conception d'un indexeur, il est nécessaire d'anticiper si le poids des termes sera plus ou moins représentatif ou discriminant. (Rijsbergen, 1979) identifie deux manières de pondérer les termes d'un document textuel : « la représentation sans discrimination » et « la discrimination sans représentation ». La représentation sans discrimination consiste à pondérer les termes en fonction de leur aptitude à caractériser l'information contenue dans le document textuel sans considérer le contenu des autres documents. A l'inverse, la discrimination sans représentation consiste à pondérer les termes d'un document en fonction de leur faculté à l'éloigner du contenu des autres documents, c'est-à-dire estimer la singularité des termes du document vis-à-vis du corpus. Dans la pratique, aucun des deux modèles n'est réellement utilisé dans les SRI, mais leurs concepteurs mettent l'accent sur les aspects représentatif ou discriminant de leur pondération. Un modèle plutôt représentatif (respectivement discriminant) consiste à affecter des scores élevés à l'ensemble des termes nécessaires à la caractérisation exhaustive (respectivement spécifique) du document.

L'indexation représentative comporte une liste de termes pondérés pour caractériser exhaustivement le document. Cette indexation est idéale pour le personnel d'une bibliothèque,

souhaitant cataloguer un document. Par contre, l'indexation représentative n'est pas la plus efficace pour la RI. Si plusieurs documents traitent d'un même sujet, il est difficile de déterminer le meilleur, celui qui répond au mieux aux attentes de l'utilisateur. Les SRI ont plutôt tendance à utiliser l'indexation discriminante, pour répondre au mieux à la requête de l'utilisateur. En revanche, une telle indexation est trop spécifique pour répondre aux attentes du personnel d'une bibliothèque.

La pondération automatique des termes d'un document textuel est étroitement liée à la distribution statistique des mots dans un texte. L'étude de la fréquence des mots dans un texte est à la base de la plupart des métriques de pondération. Nous présentons les principales métriques dans la section suivante.

2.4.1.2 Métriques de pondération

(Zipf, 1949) met en évidence l'aspect non aléatoire de la fréquence des mots dans un document. Selon lui, la fréquence des mots suit une loi de probabilité, appelé loi de Zipf. Soit f(m) la fréquence du mot m dans un corpus, une fois la fréquence de tous les mots du corpus calculée, on classe les mots en fonction de leur fréquence en ordre décroissant. Zipf conclut qu'il existe une constant C qui vérifie

$$f(m) \approx \frac{C}{\mathrm{rang}(m)}$$

Considérant cette hypothèse valide, (Luhn, 1957) propose d'analyser l'informativité d'un mot dans un document en fonction de son rang dans un corpus. Il déduit que les mots des premiers et des derniers rangs (les mots les plus fréquents et les moins fréquents dans le corpus), ne sont en général pas porteurs d'information dans le contenu textuel et donc non pertinents comme choix de descripteur (trop communs ou peu représentatifs du contenu). En général, les méthodes statistiques pour la pondération des termes prennent en considération ces hypothèses.

Il est facile de comprendre que le nombre d'occurrences d'un mot dans un texte joue un rôle essentiel dans la pondération de ce mot et donc dans le choix du descripteur associé lors de l'indexation. En revanche et comme le suggère (Luhn, 1957), l'informativité d'un terme n'est pas proportionnelle à sa fréquence. Soit tf_i^d la fréquence du terme *i* dans le document *d.* (Karen Spärck Jones, 1972) observe qu'un terme qui apparait dans un grand nombre de documents d'un corpus est en général peu discriminant (par exemple, « le », « un », « de » etc). A l'inverse, les mots qui apparaissent peu dans un document sont eux, bien entendu, très discriminants. Spärck Jones propose alors une mesure appelée fréquence inverse de document (en anglais : *inverse document frequency*) qui consiste à avoir un ordre d'idée de l'occurrence globale des termes dans un corpus. Soit un corpus *c* composé de *N* documents et dont n_i documents contiennent le terme *i*, on a :

$$idf_i^c = \log\left(\frac{N}{n_i}\right)$$

Le score idf_i^c correspond au pouvoir discriminant d'un terme. En quelque sorte la redondance d'un terme, quantifiée par le score tf_i^d, a tendance à correspondre au pouvoir représentatif du terme dans le document. Comme le laisse entendre (Rijsbergen, 1979), le choix des descripteurs prend en général leur aspect représentatif et discriminatoire. Ainsi (K Spärck Jones, 1973), propose de faire le produit pour avoir un score caractérisant le poids d'un terme dans un document appartenant à un corpus

$$tfidf^c(i,d) = tf_i^d \times idf_i^c = tf_i^d \times \log\left(\frac{N}{n_i}\right)$$

Cette manière de pondérer les termes et ses variantes est encore la plus utilisée dans les SRI. Les variantes consistent à prendre en compte de nouveaux paramètres comme, par exemple, la taille des textes. (Gerard Salton & Christopher Buckley, 1988) dressent un état de l'art précis sur l'élaboration et l'évaluation des variantes du tf-idf.

(Bookstein & Swanson, 1974) proposent une méthode probabiliste pour la pondération des termes d'un document. Ils observent que la distribution des mots-outils (mots dont le rôle sémantique est faible) dans un texte suivent une loi de probabilité proche de la loi de poisson. Cela signifie qu'un mot-outil *i* a *n* nombres d'occurrences dans un document *d* avec une probabilité *p(i,d,n)* qui s'exprime ainsi

$$p(i,d,n) = \frac{e^{-x_i} \times {x_i}^n}{n!}$$

Le paramètre x_i dépend de *i* et représente souvent la moyenne des occurrences de *i* dans les documents. L'hypothèse pour la pondération est alors la suivante : si la distribution des termes *i* suit une loi de poisson alors le mot est un mot-outil, sinon c'est un mot informatif sur le contenu d'un document. Cette hypothèse est rejetée par (S. P Harter, 1975), car si la distribution des mots-outils suit effectivement une loi de poisson, on ne peut rien dire sur la pertinence des autres mots pour une indexation. (S. P Harter, 1975) décide alors de diviser son corpus en fonction des n thèmes abordés et stipule que la distribution des mots pertinents suit une loi appelée n-poisson. Si le corpus est partitionné en deux classes C1 et C2, correspondant aux documents traitant d'un thème (C1) et aux autres documents (C2), alors la distribution des mots pertinents suit une loi 2-poisson

$$p(i,d,n) = (\pi_1)\frac{e^{-x_i^1} \times \left(x_i^1\right)^n}{n!} + (1-\pi_1)\frac{e^{-x_i^2} \times \left(x_i^2\right)^n}{n!}$$

Le paramètre π_1 est la probabilité a priori d'appartenance du document *d* à la classe C1, x_i^1 et x_i^2 les deux paramètres des lois de poisson (souvent moyennes des occurrences dans les documents des classes C1 et C2). La probabilité d'appartenance d'un document *d* à la classe C1 sachant que le terme *i* est présent exactement *n* fois est

$$p(d \in C1 \mid tf_i^d = n) = \frac{\pi_1 e^{-x_i^1} \left(x_i^1\right)^n}{\pi_1 e^{-x_i^1} \left(x_i^1\right)^n + \left(1-\pi_1\right) e^{-x_i^2} \left(x_i^2\right)^n}$$

Cette probabilité permet de décider si le terme *i* est pertinent pour l'indexation des documents de la classe C1.

2.4.2 Evaluation de l'indexation

L'évaluation de la qualité d'une indexation par un vocabulaire libre n'est pas automatisable. Le choix des descripteurs étant infini, il est difficile de décider automatiquement de leur pertinence. En revanche, les évaluations des indexations par vocabulaire libre se fait de manière extrinsèque au moment de l'évaluation du processus de recherche documentaire (Karen Spärck Jones & Galliers, 1995). Nous reviendrons sur l'évaluation de l'indexation par vocabulaire libre dans la section traitant de l'évaluation globale des SRI (cf.:2.5.3).

En revanche, pour l'indexation par vocabulaire contrôlé, il existe un mode opératoire permettant d'évaluer la qualité de cette indexation de manière intrinsèque (Karen Spärck Jones & Galliers, 1995). Nous rappelons que le choix d'un descripteur dans ce cas revient à faire de la catégorisation de documents. (Lewis, 1995) dresse un tableau de contingence des décisions d'un expert et d'un système à évaluer. Dans le Tableau 2, sur k termes sélectionnés par le système et r termes sélectionnés par l'expert, certains seront en commun (a dans le tableau).

Plusieurs critères sont évalués pour l'indexation

- rappel : R = a/(a+c)
- précision : P = a /(a+b)
- fallout : F = b/(b+d)

Le rappel permet d'évaluer les oublis du système. Un rappel égal à 1 signifie que le système a trouvé l'ensemble des descripteurs sélectionnés par l'expert. Mais il est facile d'avoir un rappel de 1 car il suffit que le système renvoie l'ensemble des descripteurs disponibles. Ainsi, le rappel doit être systématiquement accompagné de la précision qui évalue la propension qu'a le système à choisir des descripteurs inutiles. De la même façon, il est facile d'avoir une précision de 1, car il suffit de ne choisir qu'un seul bon descripteur (le plus évident). La grande difficulté est d'avoir un rappel et une précision de 1, ce qui signifie que le système est parfait. Le « fallout » est une alternative à la précision où le score idéal pour un système est 0.

	Expert est en accord	Expert est en désaccord	
Système est en accord	a	b	k=a+b
Système est en désaccord	c	d	n-k=c+b
	a+c	b+d	n

Tableau 2 : Tableau de contingence des décisions de classification

Il existe aussi des métriques hybrides permettant de quantifier la qualité du système

- taux d'erreur = (b+c)/n
- exactitude = (a+d)/n
- E-mesure = $E_\beta = 1 - \frac{(\beta^2+1)PR}{\beta^2 P + R}$
- F-mesure = $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$

Le taux d'erreur et l'exactitude sont deux métriques opposées permettant de juger respectivement le désaccord et l'accord entre l'expert et le système. Ces métriques ne sont pas souvent utilisées car n (nombre total de descripteurs) peut être très grand, biaisant ainsi le résultat. Les méthodes hybrides prioritaires sont la E-mesure ou la F-mesure. (Rijsbergen, 1979) propose la E-mesure, mesure combinée du rappel et de la précision. Le facteur $\beta \in [0;\infty[$ permet de donner la priorité au rappel ou à la précision. Le score idéal pour la E-mesure et la F-mesure (métrique opposée,) est respectivement 0 et 1.

2.5 La recherche d'information

La recherche d'information a pour objectif d'analyser une requête utilisateur afin de sélectionner et côter des documents similaires répondant mieux à cette requête. Cette notion de similarité est en réalité très complexe car elle concerne les documents entre eux mais aussi les documents et la requête. La qualité de cette mesure de similarité et la qualité de l'indexation interviennent toutes deux dans ce processus.

2.5.1 Les principaux modèles

(Baeza-Yates & Ribeiro-Neto, 1999) identifient trois grandes familles d'approche : ensembliste, algébrique et probabiliste.

2.5.1.1 L'approche ensembliste

L'approche ensembliste consiste à considérer un document comme un ensemble de descripteurs. Comme tout ensemble, les opérations d'union, d'intersection, de complément ou de différence sont applicables.

Le modèle booléen strict (Lancaster & Fayen, 1973) est le modèle de décision élémentaire pour la recherche d'information. La requête, pouvant contenir les 3 types de connecteurs logiques ET, OU et NON, est mise sous Forme Normale Disjonctive (FND) (proposition booléenne de connecteurs OU).

Ainsi, la requête est associée à une disjonction d'ensemble Q_i, avec Q = {Q_1 OU Q_2 … OU Q_n}, et la similarité entre Q et un document D est calculée ainsi :

$$sim(Q,D) = \begin{cases} 1 & si\ \exists\ Q_i, D_j |\ Q_i = D_j \\ 0 & sinon \end{cases}$$

Le modèle booléen a l'avantage d'être précis car la sélection des documents est nécessairement correcte. Il est utilisé pour la recherche de données dans les systèmes de gestion de bases de données relationnelles. En revanche, il présente l'inconvénient de ne pas pouvoir coter la pertinence des documents sélectionnés (le score étant soit 0 soit 1).

Afin de palier les inconvénients du modèle booléen, (G. Salton et al., 1983) propose le modèle booléen étendu. L'idée principale est de refuser que, pour une requête contenant 2 termes t_1 et t_2, les documents indexés par t_1 et non par t_2 soient aussi insignifiants que les documents ne traitant ni de t_1, ni de t_2. Ce modèle prend en considération les pondérations des descripteurs lors de l'indexation.

Soit $w_{1,D}$ et $w_{2,D}$ le poids normalisé des descripteurs t_1 et t_2 dans le document D, il existe alors 2 cas :

- Q = (t_1 OU t_2) alors $sim(Q,D) = \sqrt{\frac{w_{1,D}^2 + w_{2,D}^2}{2}}$
- Q = (t_1 ET t_2) alors $sim(Q,D) = 1 - \sqrt{\frac{(1-w_{1,D})^2 + (1-w_{2,D})^2}{2}}$

Comparé au modèle booléen strict, le modèle étendu propose davantage de documents en favorisant la ressemblance partielle entre requête et documents. Il met donc fin à l'appariement strict, et introduit la pondération dans son mode de calcul de la similarité. Les formules ci-dessus ne sont applicables que pour les requêtes à 2 termes. Il existe une formule générale applicable pour tout type de requête logique (G. Salton et al., 1983).

Le modèle booléen flou introduit par (Ogawa, Morita, & Kobayashi, 1991) est fondé sur la théorie des ensembles flous. L'appartenance d'un élément à un ensemble est désormais pondérée par un scalaire $\mu \in [0,1]$. Si $\mu = 0$ (respectivement $\mu = 1$) alors l'élément n'appartient pas à l'ensemble (respectivement appartient totalement à l'ensemble).

(Ogawa et al., 1991) construit une matrice de corrélation terme-terme par analyse des documents du corpus, permettant de calculer la similarité de 2 termes. La similarité de Jaccard est utilisée pour la création de la matrice, où n_1, n_2 et $n_{1,2}$ sont le nombre de documents contenant le terme t_1, t_2 et t_1 et t_2 :

$$sim(t_1,t_2) = \frac{n_{1,2}}{n_1 + n_2 - n_{1,2}}$$

L'objectif du calcul est d'estimer l'appartenance d'un document à une classe représentée par le terme. Celle-ci dépend de la corrélation entre les termes du document et t. Ainsi, le degré d'appartenance d'un document D à l'ensemble flou associé à un terme t est :

$$\mu_{t,D} = 1 - \prod_{t_i \in D} \left(1 - sim(t, t_i)\right)$$

Si t est présent dans le document alors l'appartenance est totale, si D contient un terme très similaire à t alors l'appartenance est proche de 1. Dans l'optique de la recherche d'information, nous calculons le degré d'appartenance d'un document D à une requête Q. Nous résolvons la FND de Q = {Q_1 OU Q_2 … OU Q_n} avec chaque Q_i composé des éléments ($q_{i,1}$, $q_{i,2}$… $q_{i,k}$).

$$\mu_{Q,D} = 1 - \prod_{Q_i \in Q} \left(1 - \mu_{Q_i,D}\right)$$

$$\mu_{Q_i,D} = \prod_{q_{i,j} \in Q_i} \mu_{i,j}$$

Ainsi le score $\mu_{Q,D}$ est le score de similarité pour le modèle booléen flou. Mais la similarité entre les termes peut être calculée d'une autre manière que par coefficient de Jaccard (cf. :Chapitre 3).

2.5.1.2 L'approche algébrique

L'approche algébrique utilise les structures algébriques comme mode de représentation du document et de la requête. La structure algébrique la plus fréquemment utilisée est l'espace vectoriel (G. Salton, 1971).

Le système SMART (System for the Mechanical Analysis and Retrieval of Text) (G. Salton, 1971) est le précurseur dans ce domaine. Un document est représenté par un vecteur $\vec{D}$. La dimension de ce vecteur est égale au nombre de descripteurs. Dans le cas d'un vocabulaire libre, la dimension est infinie. On peut, en revanche modéliser $\vec{D}$ dans un espace de dimension fini, à condition de ne traiter que les descripteurs libres utilisés lors de l'indexation. Ainsi, la base canonique de l'espace s'exprime de la manière suivante :

$$\begin{array}{lcccccccccc}
\overrightarrow{t_1} & = & (& 1 & 0 & \dots & 0 & 0 & \dots & 0 &) \\
\overrightarrow{t_2} & = & (& 0 & 1 & 0 & \dots & 0 & \dots & 0 &) \\
\vdots & & & & & & & & & & \\
\overrightarrow{t_i} & = & (& 0 & \dots & 0 & 1 & 0 & \dots & 0 &) \\
\vdots & & & & & & & & & & \\
\overrightarrow{t_n} & = & (& 0 & 0 & \dots & 0 & 0 & \dots & 1 &)
\end{array}$$

Le vecteur $\vec{t_i}$ représente le descripteur t_i dans l'espace de dimension n (nombre de descripteurs). Soit $w_{i,D}$ le poids du descripteur t_i dans le document D, alors, le vecteur $\vec{D}$ s'exprime ainsi :

$$\vec{D} = \sum_{i}^{n} w_{i,D} \ \vec{t_i}$$

L'ensemble des documents du corpus s'exprime sous la forme d'un vecteur. La fonction w_i renvoie le poids du terme t_i d'un document D représenté par le vecteur $\vec{D}$:

$$w_i(\vec{D}) = w_{i,D}$$

Le principe de la recherche d'information vectorielle est de représenter également la requête Q sous forme de vecteur $\vec{Q}$ afin d'appliquer des mesures de similarité vectorielle :

$$sim(D, Q) = sim(\vec{D}, \vec{Q})$$

La requête est traitée comme un document, c'est-à-dire que le traitement de la requête passe par un processus d'indexation.

Parmi les mesures de similarité vectorielles, citons :

- Le produit scalaire : $sim(\vec{D}, \vec{Q}) = \vec{D} \cdot \vec{Q} = \sum_i^n w_i(\vec{D}) \times w_i(\vec{Q})$
- La mesure cosinus : $sim(\vec{D}, \vec{Q}) = \cos(\vec{D}, \vec{Q}) = \frac{\sum_i^n w_i(\vec{D}) \times w_i(\vec{Q})}{\sqrt{\sum_i^n w_i^2(\vec{D})} \times \sqrt{\sum_i^n w_i^2(\vec{Q})}}$

Le modèle vectoriel est encore largement utilisé, notamment par le moteur de recherche Apache Lucene (Hatcher & Gospodnetic, 2004). Un premier défaut provient du fait de considérer la base canonique comme libre, c'est-à-dire que les descripteurs sont indépendants. Un autre défaut est de négliger la proximité sémantique des descripteurs. Par exemple, si un utilisateur cherche un document sur un thème précis, le système ignorera les documents n'ayant pas indexés ce terme.

Le modèle vectoriel généralisé introduit par (S. K. M. Wong, Ziarko, & P. C. N. Wong, 1985) propose de garder l'indépendance des bases canoniques du modèle vectoriel, mais rajoute l'hypothèse de non orthogonalité des bases. Pour cela, il a recours à un nouvel espace vectoriel de dimension $m=2^n$ (n nombre de descripteurs). Cette base, appelée base des « minterms », s'exprime ainsi :

$$\begin{array}{lcl} \overrightarrow{m_1} & = & (\ 1\ \ 0\ \ 0\ \ \dots\ \ 0\ ) \\ \overrightarrow{m_2} & = & (\ 0\ \ 1\ \ 0\ \ \dots\ \ 0\ ) \\ \vdots & & \\ \overrightarrow{m_m} & = & (\ 0\ \ 0\ \ 0\ \ \dots\ \ 1\ ) \quad = \overrightarrow{m_{2^n}} \end{array}$$

Pour exprimer la co-occurrence des termes dans les documents du corpus, le nouveau vecteur $\vec{t_i}$ s'exprime en fonction des combinaisons linéaires des $\overrightarrow{m_i}$ dans la nouvelle base.

(S. K. M. Wong et al., 1985) proposent un calcul des $\vec{t_i}$, vecteurs indépendants non orthogonaux. Le calcul de similarité reste le même que pour le modèle vectoriel classique. Ce modèle améliore la recherche documentaire mais présente le défaut de recalculer l'ensemble des $\vec{t_i}$ d'un nouveau document indexé.

(Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) propose une méthode de recherche d'information fondée sur la sémantique latente (LSI en anglaise : *Latent Semantic Indexing*). L'idée de LSI est, à l'inverse du modèle vectoriel généralisé, de diminuer la dimension de l'espace des descripteurs, en les regroupant par proximité sémantique. LSI construit une matrice document/terme utilisant les vecteurs $\overrightarrow{D_t}$ construits de la même manière que celui du modèle vectoriel classique. Par un procédé d'algèbre linéaire (décomposition de la matrice, extraction des valeurs propres par ordre croissant, etc.), LSI réduit drastiquement l'espace de départ, afin de regrouper les termes en fonction de leur proximité sémantique. Dans l'exemple, N documents sont indexés à l'aide de n descripteurs, LSI va réduire l'espace des descripteurs (de n descripteurs à 2) :

$$\begin{array}{c|cccccc} & \overrightarrow{D_1} & \overrightarrow{D_2} & \overrightarrow{D_3} & \overrightarrow{D_4} & & \overrightarrow{D_N} \\ \overrightarrow{t_1} & w_{1,D_1} & w_{1,D_2} & w_{1,D_3} & w_{1,D_4} & \cdots & w_{1,D_N} \\ \overrightarrow{t_2} & w_{2,D_1} & w_{2,D_2} & w_{2,D_3} & w_{2,D_4} & \cdots & w_{2,D_N} \\ \overrightarrow{t_3} & w_{3,D_1} & w_{3,D_2} & w_{3,D_3} & w_{3,D_4} & \cdots & w_{3,D_N} \\ \overrightarrow{t_4} & w_{4,D_1} & w_{4,D_2} & w_{4,D_3} & w_{4,D_4} & \cdots & w_{4,D_N} \\ & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \overrightarrow{t_n} & w_{n,D_1} & w_{n,D_2} & w_{n,D_3} & w_{n,D_4} & \cdots & w_{n,D_N} \end{array} \Rightarrow \begin{array}{c|cccccc} & \overrightarrow{D_1} & \overrightarrow{D_2} & \overrightarrow{D_3} & \overrightarrow{D_4} & & \overrightarrow{D_N} \\ \overrightarrow{C_1} & p_{1,D_1} & p_{1,D_2} & p_{1,D_3} & p_{1,D_4} & \cdots & p_{1,D_N} \\ \overrightarrow{C_2} & p_{2,D_1} & p_{2,D_2} & p_{2D_3} & p_{2,D_4} & \cdots & p_{2,D_N} \end{array}$$

Dans LSI, la requête est également représentée dans le nouvel espace. Il faut alors appliquer les métriques de similarité du modèle classique.

Les résultats obtenus sont bien meilleurs qu'avec l'utilisation du modèle classique (Dumais, 1995). En revanche, le temps de calcul de la matrice réduite est un frein considérable.

Il existe d'autres modèles algébriques permettant de retrouver des documents pertinents, notamment le modèle connexionniste (Wilkinson & Hingston, 1991).

2.5.1.3 L'approche probabiliste

Le modèle probabiliste classique, aussi appelé binary independence retrieval (BIR) (Robertson & Jones, 1976), est fondé sur la théorie des probabilités conditionnelles de Bayes. L'idée du BIR est d'estimer la probabilité que l'utilisateur trouve le document D intéressant sachant que sa requête est Q. L'ensemble des documents intéressants pour la requête Q est appelé R_Q. Le calcul de la similarité s'effectue ainsi :

$$sim(D,Q) = \frac{p\left(D \in R_Q \middle| \mathrm{D}\right)}{p\left(D \in \bar{R}_Q \middle| D\right)} = \frac{p(D \in R_Q) \times p(D|D \in R_Q)}{p(D \in \bar{R}_Q) \times p(D|D \in \bar{R}_Q)} \approx \frac{p(D|D \in R_Q)}{p(D|D \in \bar{R}_Q)}$$

$p\left(D \in R_Q \middle| D\right)$ est la probabilité que D appartienne à l'ensemble des documents intéressants pour la requête Q connaissant les descripteurs qui composent son index. La similarité est calculée en utilisant $p(t_i|D \in R_Q)$ la probabilité que t_i est un descripteur D sachant que le document D appartient à R_Q :

$$sim(D,Q) \approx \sum_{i}^{n} w_i(D) \times w_i(Q) \times \left(log \frac{p(t_i|D \in R_Q)}{1 - p(t_i|D \in R_Q)} + log \frac{1 - p(t_i|D \in \bar{R}_Q)}{p(t_i|D \in \bar{R}_Q)} \right)$$

$p(t_i|D \in R_Q)$ ne peut pas être déduite automatiquement, il faut pour cela créer une base d'apprentissage (ou mettre les utilisateurs du système à contribution) qui réponde à la question : « Ces documents répondent-ils à vos attentes ? ».

D'autres approches existent comme le modèle de la langue introduit par (Ponte & Croft, 1998). L'hypothèse est de considérer qu'un document est construit par un modèle génératif. L'objectif est ensuite de calculer la probabilité qu'une requête utilisateur puisse être générée par le même modèle. Si la probabilité est élevée, alors le document et la requête sont considérés comme similaires.

Dans les réseaux d'inférence (Turtle & Croft, 1989), les documents, les descripteurs et les termes de la requête sont des nœuds d'un graphe acyclique orienté. Des arcs joignent des documents et des descripteurs mais aussi des descripteurs et des termes de la requête. Les nœuds sont des variables aléatoires binaires et les arcs sont pondérés en fonction de la dépendance des nœuds associés. Pour un document D et une requête Q, la similarité correspond à la probabilité que les variables aléatoires associées à D et Q soient égales à 1.

2.5.2 L'expansion de requête et la réinjection de pertinence

L'expansion permet de reformuler une requête Q, en amont du calcul de similarité, afin d'augmenter le nombre de documents pertinents. Une manière de procéder est d'utiliser la réinjection de pertinence.

Soit C le corpus de document et $\{D'_i\} \subset C$ l'ensemble des documents susceptibles d'intéresser un utilisateur pour une demande, quelle est la meilleure requête Q' permettant de maximiser les D'_i? La difficulté réside dans la sélection a priori des documents intéressants. La réinjection de pertinence demande explicitement à l'utilisateur de sélectionner ou de pondérer les résultats de la recherche.

L'algorithme proposé par (Rocchio, 1971) et inclus dans SMART, permet de transformer la requête de manière incrémentale, prenant en compte les suggestions des utilisateurs ayant formulées auparavant des requêtes similaires.

(Attar & Fraenkel, 1977) proposent d'extraire l'ensemble des descripteurs des documents jugés pertinents par le système, avec l'aide éventuelle des utilisateurs. Après avoir calculé la corrélation entre les descripteurs, le système reformule la requête en fonction des termes de celle-ci et des termes qui leur sont fortement corrélés (non présents dans la requête).

D'autres approches pour l'expansion de requête utilisant des thésaurus sont présentées dans le chapitre suivant (cf. :Chapitre 3).

2.5.3 Evaluation de la recherche d'information

Après avoir mis en place un processus d'indexation et choisi un modèle de représentation associé à une mesure de similarité, la dernière étape est d'évaluer l'efficacité du SRI. Si l'évaluation de l'indexation utilisant un vocabulaire contrôlé n'est possible que par le biais d'un expert, l'efficacité d'un SRI doit être jugée par les utilisateurs qui recherchent l'information. Cette mesure est difficile à mettre en œuvre et coûteuse car, pour chaque requête, il faut demander à l'utilisateur de classer les documents.

Il existe trois niveaux d'évaluation (Baeza-Yates & Ribeiro-Neto, 1999) :

- l'évaluation fonctionnelle, qui vérifie si le SRI accomplit les tâches minimales attendues ;
- l'évaluation quantitative, qui évalue l'espace nécessaire et le temps pris pour les tâches d'indexation et de recherche d'information ;
- l'évaluation de la performance, qui évalue la pertinence des documents retournés par le système. Celle-ci peut être faite par les utilisateurs ou par un protocole d'évaluation mis en place en amont par des experts selon le paradigme de Cranfield (Cleverdon, Mills, & Keen, 1966) qui utilise 3 jeux de données : un corpus C de document D_i, un corpus Q de requête Q_i et un ensemble de paires $\{i,(Q_i,\{D_{j,i}\})\}$ qui associe à une requête Q_i donnée, l'ensemble des j documents $D_{j,i}$ pertinents pour cette requête.

Comme pour l'évaluation d'un système d'indexation par vocabulaire contrôlé, le rappel et la précision sont calculés. En général, les documents sont présentés en ordre décroissant de pertinence et donc de similarité. Dans le cas de documents ordonnés par pertinence, le rapport précision/rappel est présenté en utilisant la précision dite interpolée :

$$P'(R) = \max_{R'>R} \left(P(R')\right)$$

Pour évaluer la performance globale du SRI, ce calcul doit être fait pour toutes les requêtes à tester et il faut aussi calculer pour chaque niveau de rappel, la moyenne harmonique des précisions (Manning, Prabhakar Raghavan, & Schütze, 2008).

Une autre métrique pour évaluer la performance du système, est la précision moyenne (en anglais : *Mean Average Precision* – MAP) (Manning et al., 2008). Soit $D_{s,i,k}$, l'ensemble des k meilleurs documents ordonnés par le système pour la requête Q_i et $D_{j,i}$ l'ensemble des j documents pertinents pour la requête Q_i. On note $P(D_{s,i,k})$ la précision du système pour les k premiers documents :

$$MAP(Q) = \frac{1}{|Q|}\sum_{i}^{|Q|}\frac{1}{|D_{j,i}|}\sum_{k=1}^{j} P\left(D_{s,i,k}\right)$$

Pour l'évaluation des SRI, il existe des collections de référence, composées de documents, de requêtes, ainsi que les correspondances requête/documents attendues comme les collections TREC[5] ou les collections CLEF[6] [7].

Des mesures de performance centrées utilisateurs existent comme le ratio de couverture (en anglais : *coverage*), le ratio de nouveauté (en anglais : *novelty*), le rappel relatif et l'effort de rappel. Ces mesures prennent en compte de nouvelles données comme l'ensemble des documents pertinents pour l'utilisateur (Aouicha, 2009; Baeza-Yates & Ribeiro-Neto, 1999).

Précédemment, nous avons dit que l'évaluation automatique de l'indexation par langage libre relevait de l'évaluation globale du SRI. En effet, pour évaluer l'efficacité de l'indexation dans ce cas précis, il faut évaluer son effet sur la recherche documentaire. Il est en revanche extrêmement difficile de dissocier la contribution de l'indexation et de la mesure de similarité lors de l'évaluation.

Nous avons brièvement parcouru, dans cette partie, les méthodes d'indexation et de recherche d'information dites classiques. D'autres approches améliorant les performances des SRI existent, notamment l'analyse des liens inter-documents ou hypertextes comme le suggère les algorithmes PageRank (Page, Brin, Motwani, & Winograd, 1999) ou HITS (Kleinberg, Kumar, P. Raghavan, Rajagopalan, & Tomkins, 1999).

[5] http://trec.nist.gov/
[6] http://catalog.elra.info/
[7] http://www.clef-campaign.org/

Chapitre 3. Indexation et Recherche d'information sémantique et conceptuelle

3.1 Introduction

Comme mentionné dans le chapitre précédent, la recherche de document inclut les phases d'indexation et de recherche d'information. Les descripteurs sont extraits en analysant les documents et sont soit sélectionnés librement, soit par un vocabulaire contrôlé servant de référence à l'indexeur automatique ou humain. Pour l'indexation, le vocabulaire libre a le désavantage de ne pas fournir de référence ; le vocabulaire contrôlé, quant à lui, présente l'avantage de fixer à l'indexeur une collection précise de descripteurs possibles. L'indexeur doit maitriser ce vocabulaire afin d'anticiper l'existence d'un descripteur mais le choix de ce descripteur n'est pas trivial, dû au phénomène polysémique. La mise en relation d'une idée avec un descripteur s'appelle l'indexation conceptuelle, alors que la mise en relation d'un terme avec le meilleur descripteur s'appelle l'indexation sémantique.

Ce chapitre s'articule autour de ces deux notions. Plutôt que de sélectionner automatiquement des descripteurs en fonction de leur proximité lexicale, l'indexation sémantique et conceptuelle s'appuie sur des bases de connaissances pour le choix des descripteurs.

Dans une première partie nous présentons les différents types de bases de connaissances. Dans une seconde partie nous décrivons les méthodes permettant de calculer la similarité entre les éléments d'une base de connaissances. Nous abordons dans une troisième et quatrième partie, l'utilité des mesures de similarité pour l'indexation mais aussi pour la recherche d'information.

3.2 Bases de connaissances

La connaissance est représentée par des bases de connaissances qui peuvent être de différents types plus ou moins expressifs.

3.2.1 Dictionnaire

Le dictionnaire est la base de connaissances la moins expressive. Il consiste en une collection de termes (un seul mot ou une séquence de mots) n'ayant aucun lien entre eux. Un cas particulier de

dictionnaire est la folksonomie (Peters, 2009), dans laquelle les entrées sont collaborativement créées par une communauté. Ces entrées, appelées tags, sont notamment utilisées pour indexer et organiser le contenu des pages web de manière collaborative.

3.2.2 Réseau sémantique

(Quillian, 1968) défini le réseau sémantique comme une représenation permettant d'organiser les connaissances de la même manière que les êtres humains le ferait. Ainsi, un réseau sémantique peut être défini comme des concepts reliés à d'autres par des relations typées.

Pour (Medin, 1989), un concept (aussi appelé classe) représente une idée et inclut tout ce qui est caractéristiquement associé à elle. En d'autres termes, c'est un élément symbolique représentant un ensemble de notions et d'idées.

Par exemple, la Figure 2 représente le concept « Jus de Pomme ». Un « Jus de Pomme » est composé de « Pomme » et d' « Eau ». Dans notre exemple, « Pierre » est un « Homme » et a mangé une « Pomme » et bu du « Jus de Pomme » etc.

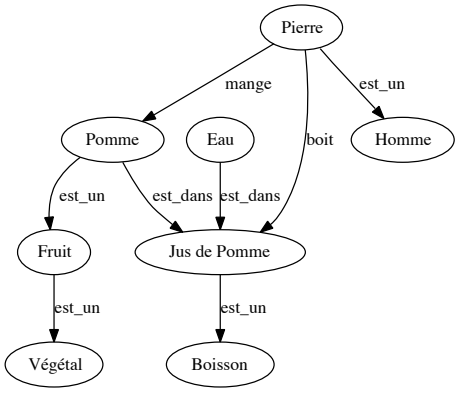


Figure 2 : Réseau sémantique centré sur le concept "Jus de Pomme"

Les réseaux sémantiques ont l'avantage d'être faciles à comprendre et à concevoir. En revanche, à ce stade, les réseaux sémantiques représentent la connaissance, mais n'établissent aucune règle pour leur utilisation dans la pratique. Afin de formaliser la connaissance en réseau de sens, il faudra attendre l'avènement des graphes conceptuels et des ontologies (cf. :3.2.5).

3.2.3 Taxinomie

La taxinomie est un réseau sémantique où l'unique relation est une relation hiérarchique, transitive et non réflexive. Souvent, la relation hiérarchique « est_un » (en anglais, « is_a ») est utilisée et on parle de classification. Les concepts sont appelés des taxons. Un exemple classique de taxinomie

est celle qui décrit les organismes vivants (Figure 3). L'avantage de l'inférence dans les taxinomies est mis à mal par le faible pouvoir d'expressivité de cette représentation.

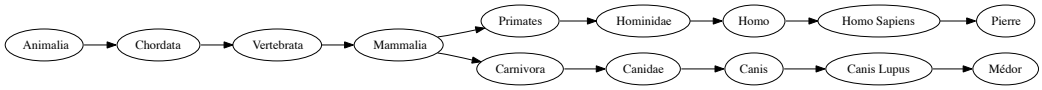


Figure 3 : Classification Biologique tirée de Wikipédia

3.2.4 Thésaurus

Les documentalistes et les bibliothécaires ont souvent recours aux thésaurus pour l'indexation de leurs documents qui concernent un seul domaine de compétence. On peut qualifier le thésaurus de réseau sémantique, où les concepts sont des descripteurs, des non-descripteurs (non utilisés pour l'indexation) et des mots vides. Les concepts sont reliés par 3 familles de relations : la relation générique/spécifique, la relation d'équivalence (synonymie en contexte) et la relation d'association (concepts proches).

La Figure 4 montre les descripteurs pour le terme « Travail pénible » issu du thésaurus de l'Organisation Internationale du Travail[8]. La relation hiérarchique TS (respectivement TG) symbolise le passage du terme générique au terme spécifique (et inversement). La relation EM signifie qu'il faut utiliser tel descripteur à la place de tel non-descripteur (la relation EP est la relation inverse). Finalement, la relation TA signifie que les termes sont associés dans le contexte. Par exemple, si un documentaliste veut indexer un document qui traite de « travaux de force », il devra, au nom de l'univocité, utiliser le descripteur « travail pénible ».

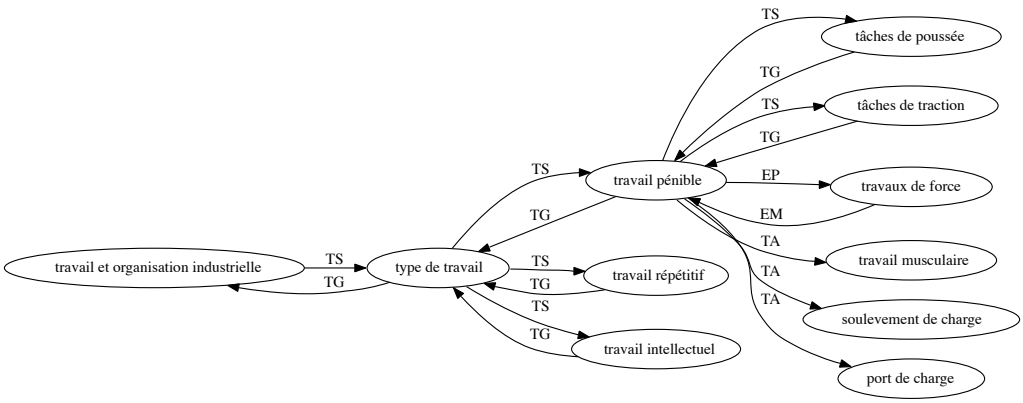


Figure 4 : Extrait du thésaurus de l'Organisation Internationale du Travail

[8] http://www.ilo.org/dyn/cisdoc/index_html?p_lang=f

La Bibliothèque nationale de France (BnF) utilise le Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié (RAMEAU) pour l'indexation de ses ouvrages. RAMEAU est un thésaurus évolutif, c'est-à-dire qu'il est mis à jour en fonction des ouvrages à cataloguer.

De plus, il existe une recommandation du World Wide Web Consortium (W3C)[9] appelé Simple Knowledge Organisation System (SKOS)[10] permettant de créer et d'organiser des connaissances sous forme de thésaurus.

3.2.5 Graphe conceptuel et ontologie

Le graphe conceptuel est un formalisme introduit par Sowa en 1976 (John F Sowa, 1976; J. F Sowa, 1983). A l'instar des réseaux sémantiques, il utilise une structure de graphe et introduit les notions de classes, de relations, d'individus et de quantifieurs. La force des graphes conceptuels réside dans leur expressivité car ils peuvent être traduits dans la logique des prédicats du premier ordre et donc permettre l'inférence.

L'ontologie, en informatique, est une représentation formelle de la connaissance introduite par Gruber en 1993 : « spécification formelle et explicite d'une conceptualisation partagée » (Gruber, 1993). (Guarino & Giaretta, 1995) essaie de clarifier le doute existant sur cette définition et propose 7 définitions possibles. De par l'ambiguïté des définitions existantes, le terme « ontologie » est devenu passe-partout pour qualifier une base de connaissances possédant des relations sémantiques. Il existe, en revanche, des notions consensuellement acceptées par la communauté de l'ingénierie des connaissances :

- concept (ou classe) et subsomption;
- relation entre concepts et subsomption des relations ;
- individu (ou instance de classe) ;
- relation entre individus ;
- restriction ;
- règle ;
- axiome (ou fait).

De plus, il existe différents langages pour exprimer des ontologies, notamment le Ontology Web Language (OWL)[11] recommandé par le W3C.

[9] http://www.w3.org
[10] http://www.w3.org/2004/02/skos
[11] http://www.w3.org/TR/owl-ref/

3.2.6 Exemple de base de connaissances : WordNet

Une des bases de connaissances de référence dans le domaine de la RI est WordNet. Elle est le fruit d'un projet initié par Miller en 1985 (Miller, 1995; Fellbaum, 1998). Abusivement, certains la considèrent comme une ontologie, mais en réalité c'est une base de données lexicales présentant des relations sémantiques. WordNet est, de notre point de vue, un réseau sémantique étendu.

Il existe dans WordNet, plusieurs types de concepts : les termes (issus du lexique) et le sens des termes (appelé *Synset*). Par exemple, le terme « wood », peut désigner l'idée de matière (1), l'idée d'un regroupement d'arbres (2), l'idée d'une famille d'instruments de musique (3) ou encore l'idée d'un accessoire de golf (4).

Plutôt que de regrouper les concepts autour de leur forme lexicale, WordNet, à travers les synsets, regroupe les concepts en fonction de leur sens en contexte. Ainsi, WordNet construit deux types de relation : 1/ entre un synset et les termes employés pour le dénoter en contexte ; 2/ entre un synset et sa définition en contexte.

D'autres types de relation sont alors introduits entre synsets :

- hyponymie : si X est hyponyme de Y alors X spécifie Y (le sens 2 de « bois » est une spécification de la « végétation ») ;
- hyperonymie : si X est un hyperonyme de Y alors X généralise Y (« végetation » est un hyperonyme du sens 2 de « bois ») ;
- méronymie : si X est un méronyme de Y alors X est composé de Y (le sens 2 de « bois » est un méronyme d' « arbre » dans le sens de la flore) ;
- holonymie : si X est un holonyme Y alors X est une partie de Y (l' « arbre » dans le sens de la flore est un holonyme de « bois »).

La Figure 5 représente une infime partie de WordNet autour des sens du mot « wood », traduit en français. Le second sens du mot bois (« Bois 2 ») peut s'exprimer avec les mots « bois » ou « forêt » dans un contexte donné (ils sont synonymes et donc interchangeables en contexte) et « Bois 2 » est composé d' « arbres ». La « Jungle » est un type de « Bois 2» ; on peut alors déduire que la « Jungle » est composée d'arbres (cette relation n'est pas présente dans WordNet). Le reste du réseau sémantique montre les hyperonymes des autres synsets. Tous les synsets sont subsumés par le synset « Entité ». Nous avons ajouté à ce réseau les hyperonymes du synset « Golf » et nous remarquons que les synsets « Bois 4 » et « Golf » n'ont en commun que le synset racine « Entité ».

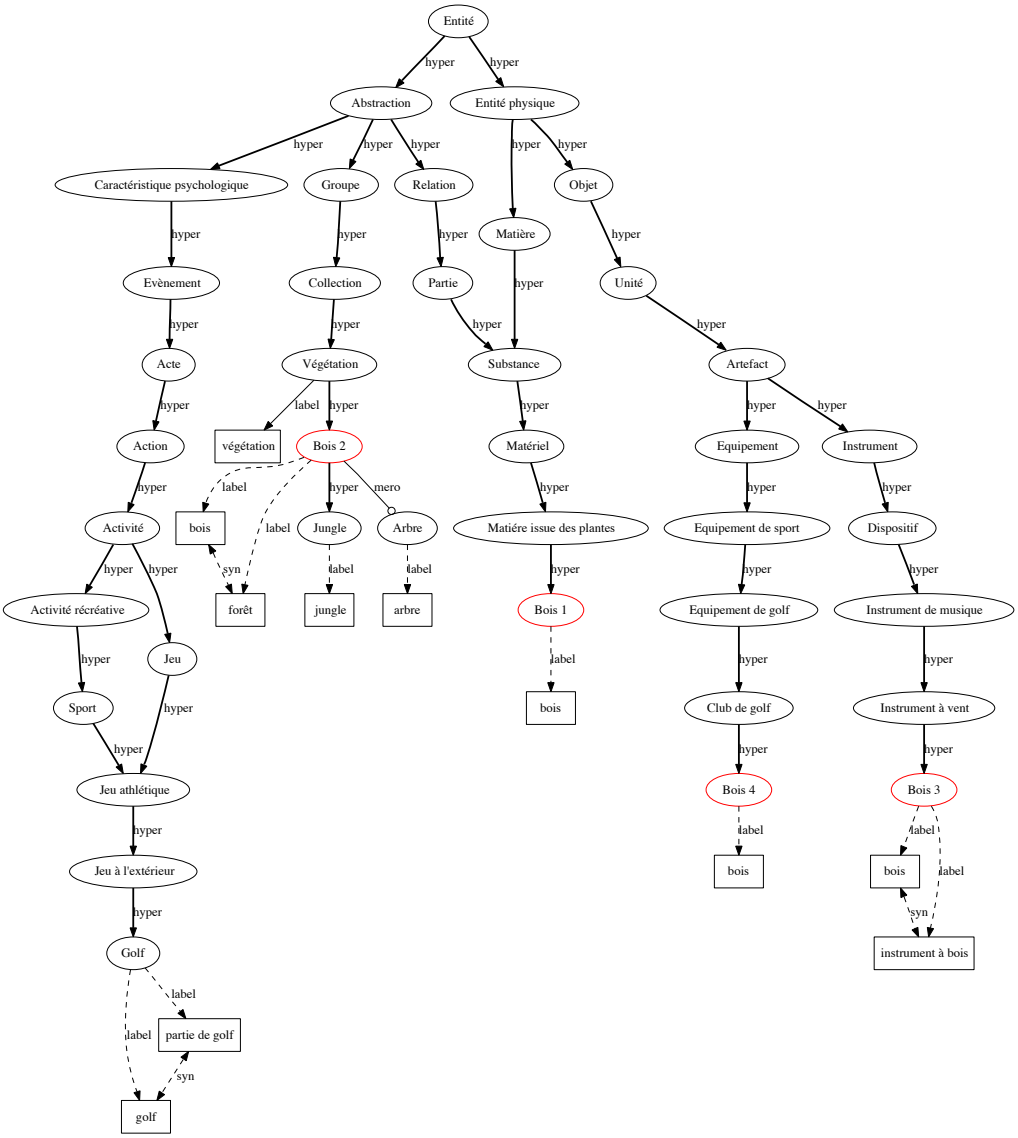


Figure 5 : Extrait traduit de WordNet

3.3 Similarité dans les bases de connaissance

La RI requiert souvent la notion de similarité entre la requête et les documents, mais aussi entre les éléments composant ces documents. Dans cette partie, nous présentons des mesures de similarité exploitant les bases de connaissances : similarité entre concepts (similarité sémantique) et similarité entre groupes de concepts issus d'une base de connaissances.

3.3.1 Similarité sémantique

La similarité sémantique consiste à quantifier le degré de ressemblance de deux concepts issus d'une base de connaissances. Elle peut être fondée sur des calculs utilisant uniquement le réseau

sémantique ou le réseau sémantique et le corpus. Elle peut aussi être fondée sur les dictionnaires et les définitions, comme la mesure de Lesk.

3.3.1.1 Mesure de Lesk

La mesure de Lesk (Lesk, 1986) s'appuie sur l'observation suivante : deux concepts sont similaires si leurs définitions sont similaires. L'idée est donc de compter le nombre d'occurrences des termes communs dans chaque définition. Si le recouvrement est de n mots consécutifs, on ajoute n^2 au score de similarité.

Le Tableau 3 montre les scores de similarité de notre exemple Figure 5. Les deux synsets les plus proches sémantiquement sont les synsets 1 et 2 (matière et forêt).

	Bois 1	Bois 2	Bois 3	Bois 4
Bois 1	36	1	0	0
Bois 2	1	36	0	0
Bois 3	0	0	16	0
Bois 4	0	0	0	144

Tableau 3 : Similarité de Lesk

En revanche, si nous voulons le synset « Bois » se rapprochant le plus de « Golf », nous obtenons une aberration car aucun mot de la définition de « Bois 4 » n'apparaît dans celle de « Golf ».

(Banerjee & Pedersen, 2003) et (Baziz, 2005) proposent de prendre en compte les définitions des hyperonymes, des hyponymes, des méronymes et des holonymes. Ainsi nous obtenons les scores suivants :

- Sim(« Golf », « Bois 1 ») = 6
- Sim(« Golf », « Bois 2 ») = 3
- Sim(« Golf », « Bois 3 ») = 12
- Sim(« Golf », « Bois 4 ») = 11

Cette mesure améliore le résultat sans pour autant être satisfaisante dans tous les cas de figure.

3.3.1.2 Mesures fondées sur le réseau sémantique

Avant d'aborder les mesures de similarité fondées sur les réseaux sémantiques, nous introduisons le concept de *Least Common Subsumer* (LCS) ou *Least Common Ancestor* (LCA). Le LCS de deux concepts est le concept hyperonyme à ces deux concepts, le plus profond. Par exemple, LCS(« Golf », « Bois 4 ») = « Entité » et LCS(« Bois 4 », « Bois 3 ») = « Artefact ». Le plus court chemin (PCC) de deux concepts est le nombre minimal d'arêtes pour aller d'un concept à un autre et, dans notre cas, sans se soucier de leur orientation. Par exemple, PCC(« Golf », « Bois 2») = 13 (en passant par « abstraction » et « jeu »). Avec ces deux notions, nous pouvons comprendre la plupart des approches pour le calcul de similarité entre deux concepts.

La première mesure est la distance de Rada (Rada, Mili, Bicknell, & Blettner, 1989). Cette distance, lorsqu'elle est appliquée sur un réseau de concept, exprime l'éloignement sémantique de deux concepts en fonction de la longueur du PCC entre ces deux concepts. Ainsi la similarité sémantique est inversement proportionnelle à cette distance.

La similarité de Wu et Palmer (Z. Wu & Palmer, 1994) est une mesure qui utilise également la notion de PPC entre les concepts mais dépend aussi de leur position dans la réseau. En effet, la mesure de Wu et Palmer a tendance à exprimer un éloignement sémantique pour des concepts proches de la racine. Ainsi pour deux concepts C1 et C2, Wu et Palmer définissent leur similarité de la manière suivante :

$$Sim_{W\&P}(C1, C2) = \frac{2 \times PCC(LCS(C1, C2), RACINE)}{PCC(LCS(C1, C2), C1) + PCC(LCS(C1, C2), C2) + 2 \times PCC(LCS(C1, C2), RACINE)}$$

RACINE est le nœud le moins profond du réseau (« Entité » dans WordNet). Selon ce calcul, deux concepts sont d'autant plus proches que leur LCS est peu profond.

- $Sim_{W\&P}$("Jeu","Sport") = $\frac{2 \times PCC(\text{"Activité","Entité"})}{PCC(\text{"Activité","Jeu"}) + PCC(\text{"Activité","Sport"}) + 2 \times PCC(\text{"Activité","Entité"})} = \frac{2 \times 6}{1+2+2 \times 6} = 0.8$
- $Sim_{W\&P}$("Groupe","Relation") = $\frac{2 \times PCC(\text{"Abstraction","Entité"})}{PCC(\text{"Abstraction","Groupe"}) + PCC(\text{"Abstraction","Relation"}) + 2 \times PCC(\text{"Abstraction","Entité"})} = \frac{2 \times 1}{1+1+2 \times 1} = 0.5$

Nous remarquons, dans cet exemple, que le PCC entre « Jeu » et « Sport » est le même que le PCC « Groupe » et « Relation » (les PCC sont égaux à 2). Cela signifie, que dans le sens de Rada, « Jeu » et « Sport » sont aussi similaire que « Groupe » et « Relation ». Or, puisque « Groupe » et « Relation » sont proches de la racine « Entité », leur similarité est plus faible au sens de Wu et Palmer.

La similarité de Leacock et Chodorow (Leacock & Chodorow, 1998) prend en compte la profondeur totale D d'une hiérarchie. Dans WordNet, la profondeur maximale est de 16 (PPC(C, « Entité ») = 16). La similarité se calcule ainsi

$$Sim_{L\&C}(C1, C2) = -\log\left(\frac{PPC(C1, C2)}{2 \times D}\right)$$

- $Sim_{L\&C}$("Jeu","Sport") $= -\log\left(\frac{3}{32}\right) = 2.36$
- $Sim_{L\&C}$("Groupe","Relation") $= -\log\left(\frac{2}{32}\right) = 2.77$

La similarité de Leacock et Chodorow est équivalente à la mesure de Rada en terme de classement. En effet, si $Sim_{L\&C}(C1, C2) > Sim_{L\&C}(C3, C4)$ alors $Sim_{Rada}(C1, C2) > Sim_{Rada}(C3, C4)$. En revanche, les scores sont pondérés différemment. Si deux concepts $C1$ et $C2$ ont un PCC de 3 et que deux autres concepts $C3$ et $C4$ ont un PCC de 6 alors, au sens de Leacock et Chodorow, $C1$ et $C2$ ne sont pas deux fois plus similaires que $C3$ et $C4$ (contrairement au sens de Rada).

3.3.1.3 Mesures fondées sur le réseau sémantique et le corpus

(Resnik, 1995) introduit la notion de contenu d'information (en anglais : *Information Content* – IC). A l'instar de Luhn, il estime qu'un terme n'est pas informatif s'il est trop fréquent. Mais contrairement à Luhn, il juge qu'un terme très peu fréquent est informatif. Il quantifie le contenu d'information d'un terme t dans un corpus en fonction de sa fréquence f(t) par la formule

$$IC(t) = -\log\left(\frac{f(t)}{\text{Nombre total de mots dans le corpus}}\right)$$

La similarité de Resnik (Resnik, 1999) calcule le contenu d'information des hyperonymes communs de deux concepts et conserve le maximum. Deux concepts se ressemblent si ce qu'ils partagent sémantiquement (appelé communialité ; en anglais, *communiality*) est porteur d'information. En pratique, il suffit donc de calculer

$$Sim_{Resnik}(C1, C2) = \text{IC}(\text{LCA}(\text{C1}, \text{C2}))$$

La similarité de Lin (Lin, 1998) calcule le rapport entre l'information contenue dans l'intersection de deux concepts (le LCS) et la somme de l'information contenue dans ces 2 concepts :

$$Sim_{Lin}(C1, C2) = \frac{2\times IC(C1 \cap C2)}{IC(C1) + IC(C2)} = \frac{2\times IC\big(LCS(C1, C2)\big)}{IC(C1) + IC(C2)}$$

3.3.2 Similarité de graphe

La section suivante traite des mesures de similarité pour calculer la ressemblance de deux collections de concepts. Soit A = $\{A_1,\ldots, A_N\}$ et B=$\{B_1,\ldots,B_M\}$ deux collections de concepts.

Les dix collections suivantes issues de WordNet (Figure 6) sont prises en exemple pour les différentes mesures présentées :

- A= {« Voiture », « Bateau »}
- B= {« Camion », « Bateau »}
- C= {« Voiture », « Camion »}
- D= {« Voiture », « Bateau », « Camion »}
- E= {« Bateau», «Réfrigérateur électrique»}
- F= {« Four micro-onde», «Convecteur»}
- G= {« Réfrigérateur électrique», «Convecteur»}
- H= {« Four micro-onde», « Réfrigérateur électrique», «Convecteur»}
- I= {« Voiture », « Réfrigérateur électrique »}
- J= {« Camion », « Four micro-onde »}

3.3.2.1 Les mesures fondées sur les sacs de mots et le modèle vectoriel

Elles considèrent les collections comme des ensembles et ne prennent donc pas en compte la hiérarchie conceptuelle :

- Similarité de Jaccard : $sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$
- Similarité de Dice : $sim(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$
- Coefficient de recouvrement : $sim(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$
- Similarité du Cosinus : $sim(\vec{A},\vec{B}) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|}$

Par exemple, en utilisant la base canonique (« Convecteur », « Réfrigérateur électrique », « Four micro-onde », « Voiture », « Camion » et « Bateau »), la représentation vectorielle de $\vec{A}$ A est (0,0,0,1,0,1).

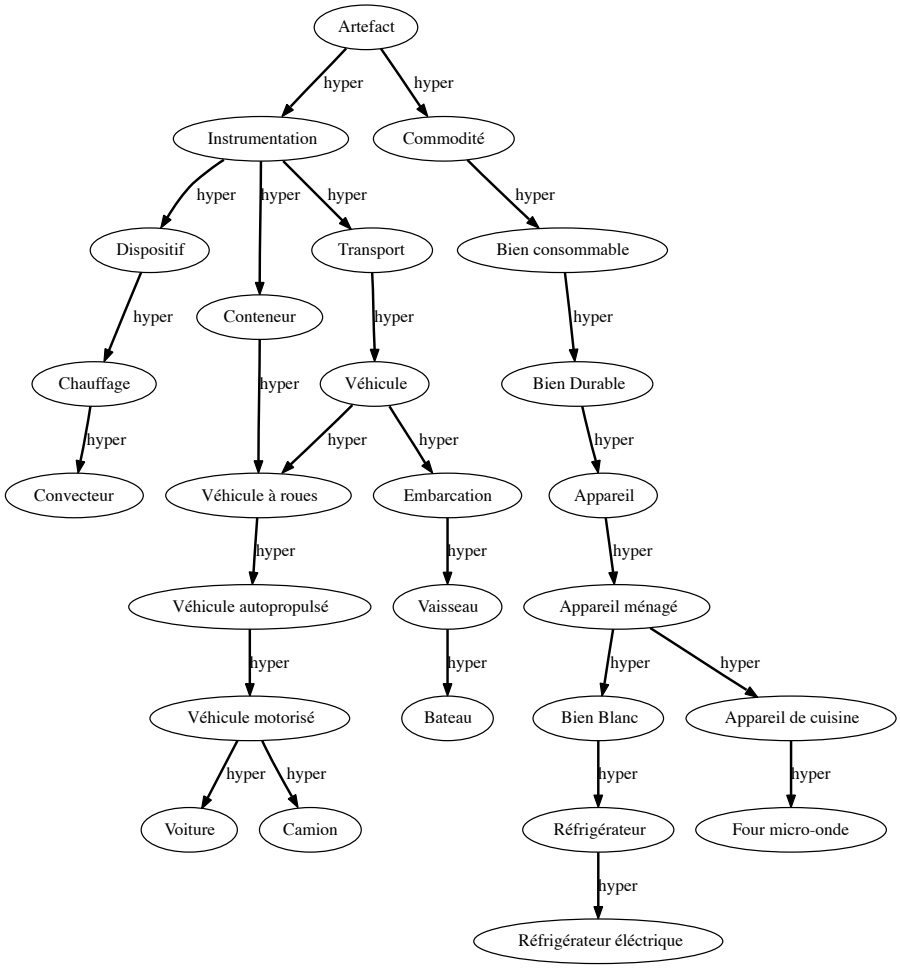


Figure 6 : Autre extrait traduit de WordNet

Le Tableau 4 (page 60) illustre les résultats obtenus en appliquant ces mesures sur quelques collections de concepts.

3.3.2.2 Mesures exploitant les relations hiérarchiques

(Ganesan, Garcia-Molina, & Widom, 2003) propose deux méthodes permettant d'estimer la ressemblance de deux collections de concepts en exploitant les relations hiérarchiques. La première méthode, appelée mesure de similarité cosinus généralisé, s'inspire du modèle vectoriel généralisé. Soit $\overrightarrow{C_1}$ et $\overrightarrow{C_2}$ les vecteurs représentant deux concepts C_1, C_2. (Ganesan et al., 2003) redéfinit le produit scalaire :

$$\overrightarrow{C_1} \cdot \overrightarrow{C_2} = \frac{2 \times profondeur(LCS(C_1, C_2))}{profondeur(C_1) + profondeur(C_2)}$$

Avec profondeur(C) = PPC(C, RACINE).

Ainsi, le calcul de similarité entre A et B suit le calcul classique :

$$\vec{A} \cdot \vec{B} = \sum_{i}^{n} \sum_{j}^{n} a_i b_j \overrightarrow{C_i} \cdot \overrightarrow{C_j}$$

Où a_i (et b_j) vaut 0 si la collection A ne contient pas C_i, sinon il vaut le poids de C_i dans la collection A. En appliquant cela sur les collections A et B de l'exemple précédent, on obtient les résultats suivants :

- $\overrightarrow{\text{Bateau}} \cdot \overrightarrow{\text{Bateau}} = 1$
- $\overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Bateau}} = \frac{2 \times 3}{6+6} = \frac{1}{2}$
- $\overrightarrow{\text{Camion}} \cdot \overrightarrow{\text{Bateau}} = \frac{2 \times 3}{6+6} = \frac{1}{2}$
- $\overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Camion}} = \frac{2 \times 5}{6+6} = \frac{5}{6}$
- $\vec{A} \cdot \vec{B} = \overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Camion}} + \overrightarrow{\text{Bateau}} \cdot \overrightarrow{\text{Bateau}} + \overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Bateau}} + \overrightarrow{\text{Camion}} \cdot \overrightarrow{\text{Bateau}} = 2.83$
- $\left|\vec{A}\right|^2 = \vec{A} \cdot \vec{A} = \overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Bateau}} + \overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Voiture}} + \overrightarrow{\text{Voiture}} \cdot \overrightarrow{\text{Bateau}} + \overrightarrow{\text{Bateau}} \cdot \overrightarrow{\text{Bateau}} = 3$
- $sim(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|} = \frac{2.83}{3} = 0.94$

Dans le Tableau 4, avec la prise en compte de la hiérarchie, la collection J {« Camion », « Four micro-onde »} est plus proche de I {« Voiture », « Réfrigérateur électrique »} que de G {« Réfrigérateur électrique », « Convecteur »} malgré un concept commun à J et G. Cela est dû au fait que « Camion » est similaire, dans la hiérarchie, à « Voiture » et que « Réfrigérateur électrique » est similaire à « Four micro-onde ».

La seconde métrique introduite par (Ganesan et al., 2003), appelée mesure de généalogie optimiste (en anglais : *Optimistic Genealogy Mesure*) prend en compte la contribution de chaque concept d'une collection au calcul de similarité.

$LCS_{A,B}(C_1)$, avec C_1 concept de A, est l'ancêtre de C_1 le plus profond présent dans B. Ensuite $sim_{A,B}(C_1)$ calcule la similarité entre C_1 et $LCS_{A,B}(C_1)$:

$$sim_{A,B}(C_1) = \frac{profondeur(LCS_{A,B}(C_1))}{profondeur(C_1)}$$

Enfin, la similarité se calcule comme suit, avec $w(C_i)$ le poids donné au concept C_i :

$$sim(A,B) = \frac{\sum_{C_i \in A} sim_{A,B}(C_i) \times w(C_i)}{\sum_{C_i \in A} w(C_i)}$$

Pour comparer les collections A et B d'une part et I et J d'autre part, issues de l'exemple, les mesures suivantes sont faites en considérant tous les poids à 1 :

- $LCS_{A,B}$(« Voiture »)= « Véhicule Motorisé »
- $LCS_{A,B}$(« Bateau »)= « Bateau »
- $sim_{A,B}$(« Voiture »)=$\frac{5}{6}$
- $sim_{A,B}$(« Bateau »)=1
- sim(A,B) = $\frac{\frac{5}{6}+1}{1+1} = 0.92$
- $LCS_{I,J}$(« Voiture »)= « Véhicule Motorisé »
- $LCS_{I,J}$(« Réfrigérateur électrique»)= « Appareil ménagé »
- $Sim_{I,J}$(« Voiture »)=$\frac{5}{6}$
- $Sim_{I,J}$(« Réfrigérateur électrique»)= $\frac{5}{8}$
- sim(I,J) = $\frac{\frac{5}{6}+\frac{5}{8}}{1+1} = 0.73$

Cette similarité n'est pas symétrique, il est alors d'usage de faire la moyenne des deux similarités. Le Tableau 4 récapitule les mesures sur l'exemple, en ne calculant pour la similarité généalogique optimiste que sim(A,B) en occultant l'aspect antisymétrique.

Collections à comparer	Jaccard	Dice	Recouvrement	Cosinus	Cosinus généralisé	Généalogie optimiste
A,B	0.33	0.5	0.5	0.5	0.94	0.92
A,C	0.33	0.5	0.5	0.5	0.85	0.75
B,C	0.33	0.5	0.5	0.5	0.85	0.75
A,D	0.66	0.8	1	0.82	0.97	1
C,D	0.66	0.8	1	0.82	0.87	1
G,J	0.33	0.5	0.5	0.5	0.6	0.65
I,J	0	0	0	0	0.75	0.73

Tableau 4 : Comparaison des mesures de similarité sur quelques collections de concepts

3.4 Indexation guidée par bases de connaissances

L'indexation sémantique et l'indexation conceptuelle permettent de pallier les manques de l'indexation classique, à savoir ne plus traiter les termes des documents comme vides de sens, mais de considérer un lien entre les termes du corpus et les connaissances extérieures à ce corpus (notamment les connaissances des destinateur et destinataire). Si un descripteur est utilisé pour indexer un document, ce descripteur discrimine le document. Ce descripteur peut être choisi indépendamment de sa fréquence dans le document et dans le corpus (comme le suggère l'indexation classique) ou même de sa présence. Toutefois, il est indispensable de connaître un domaine avant d'indexer les documents qui en traitent. L'indexation sémantique et conceptuelle prennent en compte cette problématique en proposant notamment d'utiliser les bases de connaissances comme pivot de l'indexation accompagnant le corpus.

L'indexation sémantique (en anglais : *Sense Based Indexing*) permet de trouver dans une base de connaissances, le meilleur sens correspondant à un terme d'un document. Elle est fondée sur la désambiguïsation des termes d'un document (en anglais : *Word Sense Disambiguation* – WSD). Il s'agit alors d'une indexation, non plus des termes du document, mais des termes accompagnés de leur sens. Par exemple, avec WordNet l'ensemble des descripteurs sont les labels et leur synset.

L'indexation conceptuelle consiste à indexer un document, non plus avec les termes des documents, mais avec les concepts d'une base de connaissances. L'indexation conceptuelle a pour vocation de s'affranchir de la vision purement fréquentielle ou statistique de l'indexation classique. Le document est un acte de communication entre un destinateur et un destinataire, partageant des connaissances communes, où le destinateur fournit de l'information au destinataire par le biais d'une partie de ces connaissances (Gibbs, 1987). L'objectif de l'indexation conceptuelle est d'extraire ces connaissances et de les choisir comme descripteurs du document.

L'indexation sémantique et l'indexation conceptuelle ne s'excluent pas mutuellement. Les méthodes d'indexation conceptuelle utilisent celles de l'indexation sémantique pour lever les ambiguïtés avant de choisir les meilleurs concepts décrivant le document.

3.4.1 Word Sense Disambiguation (WSD) et indexation sémantique

WSD permet d'affecter un sens issu d'une base de connaissances à un terme d'un document. (Navigli, 2009) classe les méthodes de WSD en trois familles : méthodes supervisées et semi-supervisées, méthodes non-supervisées et méthodes utilisant uniquement les bases de connaissances.

Les méthodes supervisées et semi-supervisées s'appuient sur un travail manuel a priori. Pour associer à un terme d'un document le sens approprié, les listes de décision (Rivest, 1987; Yarowsky, 1994; Agirre & Martinez, 2000) et les arbres de décision (Quinlan, 1993; Mooney, 1996) sont exploités en utilisant une base de tests. Certaines méthodes de classification ont été adaptées aux besoins de

WSD, comme les méthodes probabilistes de classification (Ng, 1997), de classification à base de cas (Daelemans, Van Den Bosch, & Zavrel, 1999) ou encore de classification à base de machines à vecteurs supports (Y. K. Lee & Ng, 2002).

Les méthodes non-supervisées tentent de désambiguïser les termes d'un document sans utiliser de base de connaissances. Elles donnent un sens aux termes en fonction des termes du voisinage. Par exemple, pour spécifier le sens du terme « Avocat_(fruit) » on se réfère à « Avocat » accompagné des termes co-occurrents (« Fruit », « Vert », «Calories», etc.). Ces méthodes sont présentées dans (Pedersen, 2006).

Les méthodes de désambiguïsation employant les bases de connaissances sont nombreuses. Cependant, l'approche la plus répandue (Dagan, L. Lee, & Pereira, 1997; Karov & Edelman, 1998; Navigli, 2009) exploite la similarité sémantique. Soit t_i un terme du document et $W_{i,n}$ une fenêtre $\{t_{i-n}, t_{i-n-1}, t_{i-n-2}, \dots, t_i, \dots, t_{i+n}\}$. Trois cas de figure sont possibles pour les termes de $W_{i,n}$:

- le terme t_i n'existe pas dans la base de connaissances, on ne le prend donc pas en compte pour WSD ;
- le terme t_i existe dans la base de connaissances mais n'a qu'un sens, on considère donc t_i désambigüisé avec cet unique sens ;
- le terme t_i existe dans la base de connaissances avec k sens associés, on extrait ses différents sens $S_{j,1}, S_{j,2}, \dots S_{j,k}$.

La sélection du meilleur sens d'un terme t_i se fait en ne conservant que le sens le plus proche aux autres termes de la fenêtre $W_{i,n}$. Cette similarité somme les similarités entre chacun des sens de t_i et le sens des autres termes qui maximise cette similarité. Finalement, on conserve le sens $\hat{S}_j$ ayant la plus forte somme :

$$\hat{S}_j = argmax_{S_{j,k}} \sum_{m=i-n}^{i+n} max_{S_{m,k'}} sim(S_j, S_m)$$

$S_{m,k'}$ sont les k' sens possibles pour le m-ième terme et la fonction sim est une mesure de similarité quelconque.

Une autre approche appelée désambigüisation par la densité conceptuelle (Agirre & Rigau, 1996) est présentée dans la section suivante.

3.4.2 Indexation conceptuelle

L'indexation conceptuelle a pour objectif d'indexer un document en utilisant les descripteurs issus d'une base de connaissances. Plus spécifique que l'indexation sémantique, l'indexation conceptuelle permet de pondérer les concepts correspondant à des termes (appartenant au document ou non) afin de sélectionner les meilleurs.

(Agirre & Rigau, 1996) proposent une méthode d'indexation sémantique pour l'indexation conceptuelle. En utilisant WordNet, ils extraient le synset le plus « fédérateur » pour la désambiguïsation. Le meilleur synset est celui qui a la plus grande densité conceptuelle définie comme suit :

$$CD(S,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendant(S)}$$

Où m est le nombre de sens des termes à désambiguïser, présents dans le sous-graphe du synset S ; descendant(S) est le nombre de synsets dans ce sous-graphe et nhyp est le nombre moyen d'hyponymes des synsets des sous-graphes (0.20 est un coefficient de lissage).

$$CD("Abstraction", 3) = \frac{1 + nhyp + nhyp^{1,15}}{descendant("Abstraction")}$$

(Woods, 1997) propose d'indexer conceptuellement les documents en utilisant un réseau ad-hoc comportant des liens syntaxiques, sémantiques et morphologiques. Le système est capable de transformer les syntagmes pour qu'ils correspondent à des concepts du réseau sémantique. Par exemple, dans la phrase « la voiture est repeinte en noir », le syntagme « repeinte en noir » va être associé au concept présent dans le réseau « changement de couleur » et sera utilisé pour l'indexation. Il s'agit bien d'indexation conceptuelle, car ce ne sont plus les termes désambiguïsés qui indexent le document, mais bien les concepts du réseau.

(Baziz, 2005) propose, dans un premier temps, d'indexer conceptuellement ces documents en utilisant WordNet. Il altère la mesure de Lesk et créé une nouvelle pondération sur les termes appelée cf-idf (*Conceptual Frequency – Inverse Document Frequency*). En faisant l'observation qu'environ 90% des termes composés (de plus de deux mots) sont monosémiques et que ces termes sont plus porteurs de sens que les mots simples (argument appuyé notamment par (Smeaton, 1992)), il décide de les avantager lors de l'indexation. Ainsi, il calcule grâce à la mesure de Lesk améliorée, les scores cumulés (voir section précédente 3.3.1.1) pour à la fois désambiguïser et sélectionner les concepts importants pour l'indexation.

3.5 Recherche d'information guidée par bases de connaissances

Dans cette section, nous dressons tout d'abord un bilan de l'impact de la désambiguïsation des termes des documents et des requêtes sur la recherche documentaire. Ensuite, nous décrivons les méthodes de recherche d'information utilisant les bases de connaissances.

3.5.1 Impact de la désambiguïsation sur la recherche d'information

(Krovetz & Croft, 1992) évaluent l'influence de la désambiguïsation sur la recherche d'information. Il ne désambiguïse que les termes des requêtes et des documents qui n'ont pas de sens dominant et ceux qui ne sont pas employés dans leur sens dominant. Le sens dominant d'un terme est défini comme le sens utilisé dans 80% des cas. Certains termes ont plusieurs sens mais n'ont pas de sens dominant comme « Avocat », contrairement au terme « France » qui a plusieurs sens (un pays, un bateau, un prénom, une chanson, etc.) et dont le sens dominant est le pays. En désambiguïsant ces termes, il obtient des améliorations de 4% à 33% par rapport à un SRI sans désambiguïsation.

(Sanderson, 1994) s'appuie sur les travaux de Krovetz et Croft pour estimer le seuil de précision requis pour que la désambiguïsation n'altère pas les performances d'un SRI utilisant une base de connaissances. Soit un corpus correctement annoté, c'est-à-dire que les termes des documents sont associés avec leur concept correspondant issu d'une base de connaissances. (Sanderson, 1994) décide d'introduire dans ce corpus un ensemble de mauvaises correspondances terme/concepts pour simuler les effets d'une mauvaise désambiguïsation. Il constate que pour un taux d'erreur de 25% (un terme sur quatre est mal désambiguïsé), le système perd beaucoup en performance. Avec un taux d'erreur de 90% de désambiguïsation, l'efficacité du SRI est relativement identique à un SRI sans désambiguïsation avec, en revanche, des meilleures performances pour les petites requêtes. Ce constat avait déjà été établi par (Krovetz & Croft, 1992). Il vaut donc mieux ne pas désambiguïser les termes, plutôt que de ne pas le faire correctement.

Finalement, (Gonzalo, Verdejo, Chugur, & Cigarran, 1998) utilisent WordNet pour démontrer que la manière de désambiguïser de Sanderson n'est pas une « vraie » mauvaise désambiguïsation. Ainsi, il propose d'indexer les documents et les requêtes avec :

- les termes non désambiguïsés ;
- les termes et leur sens unique (par exemple, le terme « Bois » dans le sens de l'instrument sera indexé par « Bois3») ;
- les synsets (par exemple, terme « Bois » dans le sens de l'instrument, sera indexé par n04598582, code du synset représentant ce concept).

Il en déduit que l'indexation par synsets et par mot-sens améliore les performances d'un SRI (respectivement 29% et 11%). La différence vient du fait que les synsets incorporent les mots-sens ainsi que leurs synonymes. Il introduit ensuite, pour l'indexation par synsets, de mauvaises désambiguïsations et déduit qu'il faut une erreur de désambiguïsation de plus de 60% pour revenir aux performances d'un SRI sans désambiguïsation. Ce résultat est en total désaccord avec les conclusions de Sanderson. Gonzalo explique cela de par la force des synsets à englober les synonymes des termes.

3.5.2 Méthodes pour la recherche d'information sémantique et conceptuelle

Les méthodes de recherche d'information sémantique et conceptuelle sont basées essentiellement sur l'expansion de la requête (reformulation avec des termes proches) ; mais d'autres approches existent.

(Voorhees, 1993) propose une méthode de désambiguïsation pour les documents et les requêtes. Il propose de construire à partir de WordNet l'ensemble des graphes appelés *hoods*. Un *hood* est le plus grand graphe comportant le concept correspondant à un sens d'un terme avec ses hyperonymes mais sans les concepts correspondant aux autres sens de ce terme. Ensuite, pour chaque terme des documents du corpus, il crée le graphe des hyperonymes en remontant jusqu'à la racine. Chaque synset visité est mémorisé dans une base de données avec le nombre de visites totales. On fait de même pour les requêtes. Pour désambiguïser un terme d'un document, le calcul suivant est fait pour chaque racine de *hood* candidat :

$$Score(S) = \frac{\#\ de\ visite\ locales}{\#\ de\ sens\ candidats\ dans\ le\ document} - \frac{\#\ de\ visite\ globales}{\#\ de\ sens\ candidats\ dans\ le\ corpus}$$

La racine du *hood* ayant le meilleur score correspond au terme désambiguïsé. Malheureusement, cette approche empire la performance d'un système, ce qui signifie que le système de recherche d'information fourni, en général, de meilleurs résultats avant cette phase de désambiguïsation. Voorhess explique cela par l'échec du système à désambiguïser correctement les requêtes courtes (ambiguë par nature).

L'approche de (Mihalcea & Moldovan, 2000), quant à elle, consiste à indexer un document avec les synsets de WordNet. Elle utilise, par exemple, des corpus pré-étiquetés pour extraire le sens des termes. Par exemple, si dans un document le terme « Avocat » est suivi de « Plaide » et qu'en analysant le corpus pré-étiqueté, le terme « Avocat » suivi du terme « Plaide » est systématiquement associé au sens « Avocat_(justice) » alors ce sens est choisi. Mihalcea propose huit manières de désambiguïser en utilisant le réseau WordNet et un corpus pré-étiqueté. Pour évaluer le SRI, elle construit 3 types de requête (une en langue naturelle et deux requêtes reformulées) :

- une requête avec les termes,
- une requête composée des termes et des synsets,
- une requête composée des termes, des synsets et des hyperonymes des synsets.

L'utilisation des synsets améliore le rappel de 16% et la précision de 4%. L'utilisation des synsets et des hyperonymes améliore le rappel de 28% mais détériore la précision de 9%. En effet, l'utilisation des hyperonymes d'un concept pour la recherche d'information retourne fatalement des documents traitant du concept mais aussi des documents traitant des hyperonymes et qui ne traite pas

du concept. En revanche, leurs expériences démontrent que l'utilisation des concepts plutôt que des termes dans un SRI améliore sensiblement sa performance.

(Baziz, Boughanem, Aussenac-Gilles, & Chrisment, 2005; Baziz, 2005) propose une méthode de recherche d'information conceptuelle dans laquelle il reformule la requête de l'utilisateur. Il se pose la question des types de relation (synonymie, hyperonymie, etc.) à prendre en compte pour l'expansion de la requête mais également la quantité d'information à ajouter dans cette expansion. Il en déduit, que contrairement à ce qu'on peut anticiper, il est préférable de n'utiliser qu'un seul synset par relation pour l'expansion de la requête. Il ajoute que la pondération des termes de la requête initiale doit être supérieure à celle des termes ajoutés lors de l'expansion. Enfin, la relation d'hyperonymie apporte le plus d'améliorations au SRI.

(Styltsvig, 2006) propose une méthode fondée sur les similarités sémantiques. En utilisant WordNet et les métriques de similarité, il reformule sa requête sans pour autant utiliser une indexation sémantique et conceptuelle des documents. Par exemple, pour la requête « Chien noir », il propose de transformer la requête en :

Q = 1(Chien ET Noir) + 0.7(Chien Et Marron)+0.68(Chien)+0.6(Chat ET Noir)+…

A l'instar de (Mihalcea & Moldovan, 2000), d'autres concepts que les concepts originaux sont séléctionnés pour enrichir la requête. Cet enrichessement augmentera forcement le rappel mais diminuera la précision

Nous avons dressé, dans ce chapitre, un aperçu des méthodes permettant d'affecter un peu de sens aux descripteurs des documents pour l'indexation mais aussi aux termes d'une requête pour la recherche des documents. Nous avons présenté des mesures de similarité ainsi que des métriques permettant de calculer la similarité entre collection de concepts. En utilisant des bases de connaissances, telles que WordNet, les résultats de l'indexation et de la recherche améliorent les performances des SRI.

Chapitre 4. Wikipédia et la recherche d'information

4.1 Introduction

Nous avons vu dans le chapitre précédent que les outils informatiques d'indexation et de recherche documentaire pouvaient être associés à une base de connaissances afin d'apporter un peu de sens aux termes des documents et des requêtes. WordNet s'impose comme la base de connaissances incontournable pour les tâches de désambiguïsation et d'expansion de la requête. Cependant, Wikipédia, issue du web 2.0 (le web collaboratif), apporte aussi de nombreux atouts.

Dans ce chapitre, nous commençons par présenter Wikipédia, son utilisation (par les contributeurs et les internautes) et sa structure. Nous décrivons ensuite les manières employées pour extraire les connaissances de Wikipédia et enfin nous présentons les travaux de recherche utilisant les connaissances extraites de Wikipédia dans les domaines connexes à la recherche d'information.

4.2 Wikipédia

Wikipédia[12] est un projet initié en 2000 par Jimmy Wales. L'objectif initial était de fournir aux usagers d'Internet un accès à une encyclopédie libre de droit et rédigée par des experts. En 18 mois, seuls 20 articles avaient été rédigés par les experts et un nouveau souffle était nécessaire. Un des contributeurs du projet propose alors d'intégrer les internautes dans la chaîne de rédaction des articles par le biais d'un wiki. Un wiki est une plateforme de travail collaborative, où les membres sont eux-mêmes fournisseurs de contenu par création ou modification des documents. Les documents possèdent des liens internes pointant vers d'autres documents du wiki (à l'instar des liens hypertextes du html). Ainsi, en janvier 2011, Wikipédia compte plus de 3,5 millions d'articles encyclopédiques en anglais et 1 million en français. Il est le 11^{ième} site internet le plus visité au monde[13].

Le contenu des articles de Wikipédia est tributaire de l'effort des internautes. L'avantage de ce processus est la rapidité de mise-à-jour de l'information et de création d'articles en fonction des évènements récents. (Lih, 2004) voit en Wikipédia un nouveau type de média journalistique entre journal et livre.

[12] http://www.wikipedia.org/
[13] En février 2011

En revanche, l'inconvénient majeur du « crowdsourcing » (faire appel à la masse pour créer du contenu) est le manque de fiabilité de l'information et le vandalisme. Concernant ce dernier, Wikipédia et ses acteurs vérifient, modifient ou suppriment les fausses informations, mais aussi limitent l'édition à certains membres de confiance. Les contributeurs courants peuvent être prévenus des modifications d'un article précis ou de la création d'un nouvel article dans une catégorie. En ce qui concerne la fiabilité de l'information, (Giles, 2005) propose de soumettre 41 articles communs à Wikipédia (en langue anglaise) et à Britannica (encyclopédie rédigée exclusivement par des experts) à un groupe d'experts. Ces derniers ont compté 168 erreurs dans Wikipédia contre 128 dans Britannica, ce qui peut être considéré comme équivalent en terme de fiabilité. Par contre, cette étude a aussi révélé que Wikipédia commettait plus d'erreurs dites d'omission (information confirmée par (Clauson, Polen, Boulos, & Dzenowagis, 2008)). (Fallis, 2008) conclut en ajoutant que les articles de Wikipédia sont plus fiables que la plupart des autres ressources du Web, et que, malgré les défauts que nous avons cités, Wikipédia présente des qualités indéniables.

Il est attendu que les contributeurs soient des experts du domaine. Même si l'opportunité s'offre à chacun des internautes de contribuer à l'encyclopédie, (Priedhorsky et al., 2007), en 2007, démontre que seul 10% des contributeurs étaient responsables de 86% du contenu, et que 0.1% (soit 4200 internautes) étaient responsables de 40% du contenu.

(Kittur, Chi, & Suh, 2009) analyse le contenu de Wikipédia et déduit que la distribution des thématiques est la suivante (version anglaise de Wikipédia en janvier 2008) :

- Culture et arts : 30 %
- Biographies et personnes : 15 %
- Géographie et lieux : 14 %
- Société et sciences sociales : 12 %
- Histoire et évènements : 11 %
- Sciences naturelles et physiques : 9 %
- Technologie et sciences appliquées : 4 %
- Religions et systèmes de croyances : 2 %
- Santé : 2 %
- Mathématiques et logique : 1 %
- Philosophie et pensée : 1 %

Wikipédia offre la possibilité d'un accès local[14] à son contenu pour notamment permettre aux utilisateurs de créer des miroirs (duplication du site original) et aux chercheurs d'analyser les données.

[14] Pour le contenu en français : http://download.wikimedia.org/frwiki/

4.3 Structure de Wikipédia

Wikipédia est fondé sur un moteur de wiki appelé MediaWiki. Ainsi, la structure de Wikipédia est celle imposée par MediaWiki, répondant aux attentes des responsables de Wikipédia. Le pivot de la structure de MediaWiki est la « page ». Une page peut être de différents types : un article (type par défaut), une catégorie, une redirection, un modèle ou autre (type défini par MediaWiki ou par l'administrateur d'un wiki MediaWiki).

4.3.1 Article

L'article est la structure où est enregistrée l'information sur un sujet. L'article est indexé par son titre et le contenu de l'article peut contenir du texte structuré (sections, sous-sections, etc.), des images, des vidéos, des enregistrements audio, des formules mathématiques et d'autres objets de type « page ». Les données textuelles sont écrites dans un langage spécifique à MediaWiki (appelé Wiki markup). Ce langage de présentation permet, à l'instar du html, de formater visuellement le texte (italique, gras, etc), de hiérarchiser la structure logique (sections, sous-sections, etc.) et de définir des liens entre « pages ».

La Figure 7, présente l'article traitant de « retour de pertinence » en langue française. Nous observons plusieurs caractéristiques :

- Le titre de l'article (1) ;
- Un titre synonyme qui pointe sur la même page (2) ;
- Des liens internes vers d'autres articles ou « pages » (3) ;
- Des références externes (4) ;
- Des liens vers des portails (5) ;
- Des liens vers des catégories (6).

Figure 7 : Exemple de l'article Wikipédia « réinjection de pertinence »

Un article peut être composé de différentes sections (non visibles dans la Figure 7) auxquelles on peut accéder directement depuis l'article.

4.3.2 Redirection

Les redirections dans MediaWiki, sont des pages de type article dont le contenu redirige directement les utilisateurs vers un autre article. Dans l'exemple de la Figure 7, nous avons recherché l'article « réinjection de pertinence ». Or, cet article nous renvoie directement à l'article « retour de pertinence ». Le lien de redirection signifie explicitement « pour avoir des information sur le sujet X, lire l'article sur le sujet Y ».

Il existe des redirections d'un article vers le même article écrit dans une autre langue.

4.3.3 Catégorie

Une catégorie est une page qui est une agrégation de plusieurs pages. Dans notre exemple de la Figure 7, « retour de pertinence » est dans les catégories « informatique théorique », « recherche d'information », « traitement automatique du langage naturel ». Pour notifier qu'une page appartient à une catégorie, il faut ajouter au Wiki markup de la page :

[[Catégorie: recherche d'information]]

Dans Wikipédia, plusieurs articles appartiennent à la catégorie « recherche d'information » qui elle-même appartient aux catégories « sciences de l'information et des bibliothèques », « application de l'informatique » et « sciences de l'information et de la communication ». De même, « recherche d'information » est un agrégat de catégories comme « base de données bibliographiques » ou « social bookmarking ». Wikipedia présente donc une hiérarchie de catégories dont la plus haute a pour titre « Accueil ».

Certaines catégories de Wikipédia sont particulières : « espace encyclopédique », « espace non-encyclopédique » et « homonymie ».

« Espace encyclopédique » agrège « articles encyclopédiques », qui elle-même contient 6 catégories : « Espace », « Temps », « Nature », « Science », « Société » et « Spiritualité ». On peut qualifier cette catégorie de super-catégorie contenant les articles apportant de la connaissance aux lecteurs.

« Espace non-encyclopédique » regroupe les catégories permettant non pas de classer les articles en fonction de leur thème, mais en fonction des pages elles-mêmes. Par exemple, si un article est en cours de rédaction, il sera classé dans la catégorie « article en travaux », lui-même dans la catégorie « page en travaux » etc. C'est une méta-catégorie, car elle informe les internautes sur la page et non son contenu.

« Homonymie » regroupe l'ensemble des articles dont le titre peut prêter à confusion (cas de polysémie). Les articles renseignent en général sur les différentes façons d'utiliser le titre de l'article. Par exemple, l'article de titre « Bois (homonymie) » de la Figure 8 (les articles ayant un titre ambigu s'intitulent Nom_article_(Domaine_traité)) informe les lecteurs des différents sens du mot « Bois ». Le premier sens renvoie l'utilisateur sur l'article « Bois ». Le deuxième sens renvoie l'utilisateur sur l'article « bois_(matériau_de_construction)). Aussi, le septième sens du mot « bois » renvoie l'utilisateur vers la section « bois » de l'article « club » dans le contexte du golf (club_(golf)#bois).

Figure 8 : Différents sens du mot « Bois » dans Wikipédia France

4.3.4 Portail

Les portails sont des types de « pages » représentant une thématique. Ils contiennent donc les articles traitant d'un même thème. La façade, visible par les utilisateurs, est générée manuellement. En

revanche la partie appelée « articles liés » permet automatiquement d'agréger les « pages » ayant le même portail.

Par exemple, il existe 12 641 articles dans le portail « portail:informatique » et 1030 portails au total[15].

4.3.5 Modèle

Un modèle (en anglais, *template*) est une page contenue dans une autre. Les internautes créent des modèles types pour homogénéiser les pages. Le contenu du modèle peut être modifiable (à l'instar d'un formulaire) ou statique.

Comme illustré par la Figure 9, pour l'utilisateur, le contenu de Wikipédia est centré sur la notion d'article. Ces articles sont classés dans des portails thématiques et des catégories, ces dernières étant également classées dans des sous-catégories.

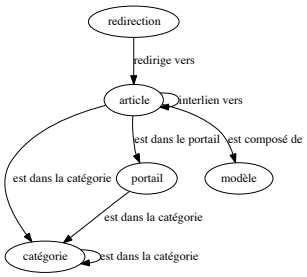


Figure 9 : Structure simplifiée de Wikipédia

Wikipédia propose également une base des résumés introductifs des articles.

L'ensemble des pages permet aux utilisateurs, et notamment aux équipes de recherche, d'extraire des bases de connaissances.

4.4 Extraction d'une base de connaissances Wikipédia

Wikipédia organise les connaissances issues des contributions des internautes en utilisant ses propres paradigmes. On peut obtenir une base de connaissances, telle que décrite dans le chapitre Chapitre 3, à partir des données de Wikipédia. Cette section traite de la création d'un dictionnaire, d'une taxinomie, d'un thésaurus et d'une ontologie utilisant les données de Wikipédia.

4.4.1 Création d'un dictionnaire

La base de connaissances la plus facile à extraire est un dictionnaire. Les titres des articles sont les entrées de ce dictionnaire ou les descripteurs d'un vocabulaire contrôlé. Le nombre d'articles de

[15] En janvier 2011

Wikipédia est de 2 millions dont 1.1 million redirigés, ce qui revient à dire que l'on peut créer un dictionnaire de 2 millions d'entrées. A cela, peut être ajouté l'ensemble des textes accompagnant les hyperliens.

4.4.2 Création d'une taxinomie

Pour extraire une taxinomie, il faut une structure permettant de passer d'un concept à un autre par une relation hiérarchique. Les catégories permettent cela : la relation « est dans la catégorie » est très clairement hiérarchique, mais elle est floue car de trop haut niveau. En effet, elle englobe les relations d'hyperonymie, de méronymie et des relations « générique/spécifique ». Par exemple, un humain est capable d'inférer que la relation hiérarchique entre « Arbre » et « Bois » est une relation méronymique et que la relation entre « Matière première végétale » et « Bois » est un lien hyponymique. La Figure 10 représente l'ensemble de relations hiérarchiques issu de Wikipédia pour l'article « Bois » avec des arcs que nous avons annotés manuellement.

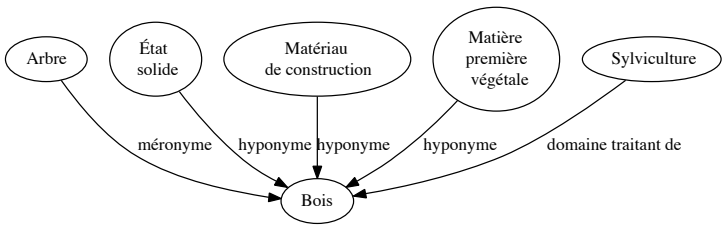


Figure 10 : Les catégories agrégeant la catégorie « Bois » avec les types de liens déduits par un humain

(Ponzetto & Strube, 2007) propose de lever certaines ambiguïtés de cette relation en appliquant les processus suivants :

1. Nettoyer le réseau de catégories sans prendre en compte les catégories « espace non-encyclopédique » ;

2. Détecter les liens « est un sous-ensemble » (en anglais, *is-refined-as*) en analysant syntaxiquement les noms des catégories de type « X de Y» et « X par Z » ;

3. Détecter les liens d'hyperonymie en utilisant des règles syntaxiques. Par exemple, considérer que « Acteurs français » est un hyponyme de « Acteurs » car « français » est un adjectif qualifiant « Acteurs ». Les « mauvaises » relations sont retirées. Par exemple, si « français » est dans la catégorie « Acteur français», la relation est retirée ;

4. Identifier les relations d'hyperonymie en détectant la présence du pluriel dans les noms de catégories. (Suchanek, Ifrim, & Weikum, 2006) observe que les articles des catégories dont le nom est au pluriel sont en relation de type « est une instance de » avec cette catégorie.

5. Utiliser des méthodes lexicales et syntaxiques issues du TAL sur un corpus afin de détecter les relations d'hyperonymies. Par exemple, « Tiger Woods et d'autres golfeurs » indication que « Tiger Woods » est un hyponyme de « golfeurs ».

6. Finalement, propager les résultats par transitivité de la relation d'hyperonymie. Si « Tiger Woods » est un « golfeurs américains» (par la règle 4) et que « golfeurs américains » est « Americains » (par la règles 4), alors on peut inférer que « Tigers Woods » est « Américains ».

Ainsi, il estime que 30% des relations de catégories de Wikipédia sont des relations d'hyperonymie.

4.4.3 Création d'un thésaurus

La création d'un thésaurus nécessite 3 types de liens : la généralisation, la synonymie et l'association. On considère que les titres des articles sont les descripteurs. Avec les catégories, Wikipédia propose déjà une structure de généralisation. Dans ce cas de figure, on peut utiliser les catégories comme un ensemble de non-descripteurs.

La synonymie aussi peut s'extraire automatiquement de Wikipédia et de deux façons. La première approche consiste, pour un article, à extraire l'ensemble des redirections. Dans l'exemple de la Figure 7, l'article « réinjection de pertinence » est une redirection vers l'article « retour de pertinence ». Nous considérons alors qu'il y a un lien de synonymie entre les deux titres. La seconde façon de procéder est plus délicate. Lorsque qu'un article affiche un lien interwiki de type (nom_article|terme_affiché), nous considérons que « terme_affiché » et « nom_article » sont synonymes, mais avec des précautions car « terme affiché » peut être très ambigu. D'après le manuel de style de Wikipédia[16], il est dit que les liens hypertextes internes doivent être clairs pour le lecteur. En théorie, il n'y a que des bons «terme_affiché », mais dans le doute, nous retirons ceux trop fréquents dans le corpus Wikipédia ou redirigeant vers un trop grand nombre d'articles.

La création des liens d'association entre concepts fait référence à la notion de similarité sémantique entre concepts de Wikipédia qui est traitée en section 4.5.1.

4.4.4 Création d'une ontologie

L'extraction d'ontologies à partir de Wikipédia est une tâche complexe. Cela nécessite

- d'extraire les concepts (ou classes) ;
- d'extraire les individus (ou instances);
- d'attribuer les individus aux bons concepts ;
- de trouver les liens d'hyponymie/hyperonymie entre concepts ;
- d'expliciter les liens entre concepts, individus/concepts et individus/individus ;
- de trouver les propriétés des concepts et des individus.

[16] http://en.wikipedia.org/wiki/Wikipedia:Only_make_links_that_are_relevant_to_the_context

L'espace des catégories est un bon espace de concepts et celui des articles un bon espace d'individus. En revanche, il existe des cas où les catégories sont des instances (par exemple la catégorie « France » est une instance de « pays d'Europe »). (Zirn, Nastase, & Strube, 2008) propose des méthodes permettant de décider si un article ou une catégorie est une classe ou une instance.

Pour trouver des nouveaux types de liens et les attribuer aux concepts et individus, nous mentionnons 3 types de méthodes.

La première méthode consiste à extraire à partir des modèles Wikipédia, des relations entre individus. Le système Kylin (F. Wu & Weld, 2008) permet d'analyser les modèles en utilisant des méthodes à base de classification semi-supervisée.

La seconde méthode, à l'instar de (Ponzetto & Strube, 2007), analyse les titres des articles et des catégories pour en déduire des relations. Par exemple, dans la catégorie « album de David Bowie », il est aisé d'extraire les relations de type « est l'auteur de » en supposant que tous les articles de cette catégorie sont bien des albums de David Bowie. Cette approche est développée par (Nastase & Strube, 2008).

La dernière approche, développée par (Ruiz-Casado, Alfonseca, & Castells, 2005a, 2005b, 2006) utilise WordNet pour l'extraction de nouvelles relations. Elle utilise le corpus textuel de Wikipédia pour extraire des nouveaux types de relations. Par exemple, si dans un article, il est écrit « X est la capitale de Y » et que « Y » est un méronyme de « X », alors la relation « est la capitale de » est une relation méronymique. Ainsi, Ruiz-Casado crée 1200 nouveaux types de relations.

Actuellement, trois ontologies issues des connaissances présentes dans Wikipédia sont disponibles en ligne : YAGO[17], DBpedia[18] et Freebase[19].

YAGO (Suchanek, Kasneci, & Weikum, 2007) fusionne les synsets de WordNet et les articles de Wikipédia. Il possède 2 millions de prédicats et 20 millions de relations avec une exactitude de 95%.

DBpédia (Auer et al., 2007; Bizer et al., 2009) contient 2.9 millions de prédicats et 190 millions de relations. Ces relations sont déduites en utilisant notamment les modèles de Wikipédia. L'importante quantité d'information s'explique par le manque de contrôle et de nettoyage de la base. En effet, dans Wikipédia plusieurs modèles différents peuvent représenter la relation. Par exemple, le modèle « Lieu de Naissance » et « Ville de Naissance » sont traités comme deux métadonnées différentes. Or, ces types de relations sont liés.

Freebase est une ontologie développée par la fondation Metaweb, agrégeant les connaissances de plusieurs bases, telle que MusicBrainz (base de connaissances musicales), WordNet, DBpédia et

[17] http://*www.mpi-inf.mpg.de/yago-naga/yago/*
[18] http://*dbpedia.org/*
[19] http://*www.freebase.com/*

Wikipédia. Il en résulte une énorme ontologie composée de 50 millions de prédicats et de 430 millions de relations.

Les ontologies créées à partir de Wikipédia permettent d'indexer et de rechercher des documents en utilisant les approches réalisées pour WordNet.

4.5 Wikipédia et les problématiques intrinsèques à la recherche d'information

L'extraction des connaissances de Wikipédia a permis aux équipes de recherche de bénéficier d'un nouveau type de données : la connaissance de masse. Autrefois, il était d'usage d'utiliser des bases de connaissances conçues par des experts. Les résultats de ces travaux étaient globalement meilleurs pour la recherche de documents, mais au prix d'un effort manuel important des experts pour la création de la base. Wikipédia a l'avantage de fournir la matière pour faire émerger la connaissance, certes floue, en un minimum d'effort.

Cette section traite des différentes approches permettant d'utiliser Wikipédia pour le calcul de similarité, la désambiguïsation, l'extraction des thématiques et des mots-clés et la recherche de documents.

4.5.1 Wikipédia et la similarité sémantique

Des travaux ont été menés pour analyser la similarité entre les concepts de Wikipédia, à savoir les catégories et les articles.

(Strube & Ponzetto, 2006) propose un système, appelé WikiRelate! qui calcule la similarité sémantique entre deux concepts de Wikipédia. Il utilise les métriques fondées sur les réseaux sémantiques pour évaluer la performance obtenue avec Wikipédia comme source de connaissances dans le calcul de similarité. Pour chaque article ou catégorie, il génère le graphe des catégories en se limitant à une profondeur de 4 (profondeur la plus adéquate pour le calcul de similarité). En évaluant le système avec deux bases de tests et la métrique de Leacock & Chodorow (Leacock & Chodorow, 1998), celui-ci est tantôt meilleur avec WordNet (86% de correspondance correcte), tantôt avec Wikipédia (49% de correspondance correcte).

L'« Analyse de la sémantique explicite » (*Explicite Semantic Analysis : ESA*) (Gabrilovich & Markovitch, 2007) étudie le corpus d'articles de Wikipédia, en faisant l'hypothèse que le titre d'un article est un concept et que l'article est sa définition. Le coefficient TF-IDF de chaque terme de l'article est calculé pour que chaque concept soit représenté par un vecteur pondéré. L'algorithme de classification de (Han & Karypis, 2000) est utilisé ; il regroupe, pour un fragment de texte donné, les concepts les plus proches. Les résultats sont très encourageants car 75% des termes similaires ont été détectés avec succès (49% pour l'approche de WikiRelate!).

Wikipedia Link-based Measure (*WLM*) (Witten & D. Milne, 2008)(David Milne, 2010) prend en compte, dans le calcul de similarité, les liens interwiki de Wikipédia. Soit a un article de Wikipédia, $r(a)$ le coefficient de rareté de a, W l'ensemble des articles Wikipédia et A l'ensemble des articles possédant un lien vers a, on a :

$$r(a) = log\left(\frac{|W|}{|A|}\right)$$

La similarité entre deux articles a et b est définie comme suit :

$$sim(a,b) = \sum_{x \in A \cap B} r(x)$$

La similarité est élevée si a et b possèdent un maximum de liens vers les mêmes articles rarement référencés.

Une seconde mesure de similarité, inspirée de la mesure de Google Normalisé (Cilibrasi & Vitanyi, 2007), est introduite :

$$sim(a,b) = \frac{log\ (max(|A|,|B|)) - log\ (|A| \cap |B|)}{log(|W|) - log\ (min(|A|,|B|))}$$

La similarité WLM est la moyenne des deux similarités et obtient 68% de correspondance correcte dans Wikipédia. L'avantage de cette méthode est qu'une fois qu'on a les coefficients de rareté, la similarité est calculée très rapidement.

Nous disposons donc de trois méthodes encourageantes pour calculer les similarités entre concepts de Wikipédia, qui utilisent trois sources de types différents : les catégories, le corpus brut et les liens interwiki avec, pour chacune, une distance spécifique.

4.5.2 Wikipédia et la désambiguïsation

(Mihalcea, 2007) utilise les liens interwiki de Wikipédia afin de répertorier l'ensemble des expressions (article|terme_X). Il construit ensuite un corpus composé des paragraphes contenant ce type de liens pour le même terme_X. Pour chaque terme_X, une mise en correspondance manuelle avec les synsets de WordNet est effectuée. Ainsi, (Mihalcea, 2007) considère que le corpus créé par cette mise en correspondance permet d'alimenter un classifieur bayésien (Ng, 1997). Le système permet alors une augmentation de 30% par rapport à la mesure de Lesk et de 44% par rapport à celle utilisant le sens dominant.

(Cucerzan, 2007) propose une désambiguïsation des entités nommées de Wikipédia. Pour cela, il considère l'ensemble des entités nommées du corpus (articles correspondants, entre autres, à des personnes, des organisations ou des lieux). Pour chaque article, il extrait les catégories auxquelles il appartient, ainsi que l'ensemble des dénominations possibles pour cet article en utilisant les liens de

type (article|terme) et les redirections. Pour chaque article représentant des entités nommées, il crée un vecteur dont les composantes sont les catégories. Pour chaque entité nommée d'un texte, il extrait l'ensemble des vecteurs candidats de cette entité donnant la liste des articles correspondants. Par exemple, si le terme « Bush » est présent dans un texte, l'ensemble des vecteurs pour les articles « George W. Bush », « George H. W. Bush », « Kate Bush », etc. est extrait. En utilisant le produit scalaire comme mesure de similarité, (Cucerzan, 2007) choisit l'ensemble des vecteurs qui maximise la somme des produits scalaires. Wikipédia permet de désambiguïser correctement 91% des entités nommées des articles d'un quotidien et 88% des entités nommées des articles de Wikipédia lui-même.

(Fogarolli, 2009) définit la notion de liens forts entre articles. Un lien entre deux articles a et b est fort si et seulement si il existe un lien dans a qui mène vers b et un lien dans b qui mène vers a. Soit D un document et T_n l'ensemble des termes non-vides de ce document. Pour un terme T_k ambigu de T_n, l'ensemble des m candidats à la désambiguïsation $C_{k,m}$est extrait et pour chaque candidat, la liste des liens forts est extraite. Le candidat gagnant est celui ayant le plus de liens forts communs à l'ensemble T_n. Sur un ensemble d'articles issus de Wikipédia, (Fogarolli, 2009) obtient une désambiguïsation de 90%.

4.5.3 Wikipédia et l'extraction des thématiques et des mots-clés

(Mihalcea & Csomai, 2007) se penchent sur l'extraction de mots-clés et leur désambiguïsation. Ils introduisent la notion de *Wikification* du texte. *Wikifier* un texte revient à trouver les mots-clés du texte et de leur affecter un hyperlien vers les bons articles. Dans le manuel de style de Wikipédia[20], il est dit d'affecter un lien à un terme uniquement s'il est important dans le contexte, ce qui pour (Mihalcea & Csomai, 2007) est la définition d'un mot-clé. Pour trouver les mots-clés d'un document, ils évaluent, entre autres, le TF-IDF et une nouvelle mesure appelée *Keyphraseness*. Le *Keyphraseness* d'un terme T prend en considération, dans le corpus Wikipédia, l'ensemble des documents $D_{\to T}$ où T est un lien et l'ensemble des documents D_T où T apparaît :

$$Keyphraseness(T) \approx \frac{\#D_{\to T}}{\#D_T}$$

Ainsi, si un grand nombre de documents contenant un lien vers T, par rapport au nombre total de documents dans lequel T apparaît, alors T est un bon candidat pour être un mot-clé. Dans un article Wikipédia, 6% des termes sont des liens ; le système est donc évalué sur le corpus de Wikipédia lui-même, afin de déterminer si leur méthode retourne les mêmes mots-clés que l'annotation manuelle des contributeurs. En choisissant les articles de qualité[21] pour l'indexation, ils obtiennent une F-mesure de 43% avec le TF-IDF et de 55% avec le *Keyphraseness*.

[20] http://en.wikipedia.org/wiki/Wikipedia:Only_make_links_that_are_relevant_to_the_context
[21] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Articles_de_qualit%C3%A9

(Medelyan, Witten, & D. Milne, 2008) utilisent une méthode de classification bayésienne pour la recherche de descripteurs, en exploitant leur propre travaux sur la similarité (Witten & D. Milne, 2008)(David Milne, 2010), mais aussi le *Keyphraseness*. Le calcul de similarité sert pour la désambiguïsation, mais aussi pour l'extraction des descripteurs candidats. Ainsi, à chaque document est associée une liste de concepts candidats désambiguïsés (titres des articles). A chacun des documents de la base de tests alimentant le classifieur, sont associés manuellement les meilleurs descripteurs ainsi que les scores TF-IDF des candidats et la mesure appelée *Total Keyphraseness* :

$$Total\ Keyphrasness(A) = \sum_{A \to T} freq(T) \times Keyphraseness(T)$$

Où A est le titre d'un article et T ses termes synonymes. Pour évaluer le système, des experts indexent manuellement 20 documents techniques et la mesure de consistance entre deux indexeurs (Rolling, 1981) est calculée :

$$Consistance = \frac{2 \times (\#descripteurs\ en\ commun)}{\#\ descripteurs\ de\ l'indexeur1 + \#descripteurs\ de\ l'indexeur2}$$

Ainsi, une consistance de 31% est obtenue avec les indexeurs humains, contre 18% en utilisant seulement TF-IDF. Cela signifie, que la mesure du *Keyphraseness* correspond mieux aux attentes d'un documentaliste pour l'extraction des descripteurs que la mesure TF-IDF.

(Janik & Kochut, 2008) utilise la base de connaissances DBpédia pour l'extraction de la thématique principale d'un document. Pour chaque concept d'un document, extrait de DBpédia, on regarde s'il est relié à un autre concept du document par une relation de DBpédia. Un graphe est ainsi construit et le sous-graphe le plus important (en terme de connexions entre termes) est gardé. Ensuite, la mesure de centralité extrait les concepts les plus importants :

$$Centralité(C_i) = \frac{1}{\sum_j pcc(C_{i,}C_j)}$$

Cette mesure favorise les concepts ayant une forte proximité avec un nombre important de concepts qui sélectionnent ainsi 10% des meilleurs concepts du graphe pour en extraire un sous-graphe représentatif du document. L'étape suivante consiste à catégoriser ce sous-graphe en utilisant un score appelé score de catégorisation thématique (Janik & Kochut, 2008) pour déterminer la meilleure catégorie décrivant le document. Le système classifie correctement 80% des documents issus du corpus CNN et 67% des documents issus du corpus Wikipédia.

(Coursey, Mihalcea, & Moen, 2009) et (Coursey & Mihalcea, 2009) utilisent une version de l'algorithme PageRank (Page et al., 1999) proposé par (Haveliwala, 2003) sur un graphe composé des articles, des liens interwiki et des catégories. La consistance de cet outil avec un indexeur humaine est de 35%, contre 31% pour (Medelyan et al., 2008).

D'autres approches existent pour la reconnaissance de thématiques, notamment les méthodes (Schönhofen, 2009) et (Syed, Finin, & Joshi, 2008), fondées sur la structure de réseau des catégories Wikipédia.

4.5.4 Wikipédia et la recherche documentaire

(Müller & Gurevych, 2009) propose d'utiliser le modèle de (Gabrilovich & Markovitch, 2007), précédemment présenté (en section 4.5.1) pour la similarité sémantique entre deux articles, pour calculer la similarité entre une requête et un document. La similarité est donnée par la formule :

$$Sim(d,q) = \frac{\sum_i^{n_d}\sum_j^{n_q} freq(t_{i,d}) \times freq(t_{j,q}) \times idf(t_{i,d}) \times idf(t_{j,q}) \times s(t_{i,d}, t_{j,q})}{(1+n_{nr}) \times (1+n_{nsm})}$$

Où n_d et n_q sont les nombres de termes uniques des documents et de la requête, $s(t_{i,d}, t_{j,q})$ le score de similarité tel que décrit dans (Gabrilovich & Markovitch, 2007), n_{nsm} le nombre de termes de la requête n'appartenant pas au document et n_{nr} le nombre de termes de la requête similaires à aucun terme du document (en fixant un seuil de similarité). Cette approche, combinée avec l'algorithme du moteur de recherche Lucene (Hatcher & Gospodnetic, 2004), permet de passer d'une précision moyenne (MAP définie dans 2.5.3) de 29% à 32%.

(Alemzadeh & Karray, 2010) propose un algorithme d'expansion de la requête, permettant de l'enrichir avec des catégories Wikipédia. Soit une requête Q composée des termes $q_1 \dots q_n$. Pour chaque terme q_i, l'ensemble des articles A_m contenant ce terme est récupéré. La masse d'un article A_j est le nombre de termes communs au titre de l'article et à la requête Q. Ensuite, on extrait l'ensemble des catégories C_k des articles A_m. Le score $w_{i,j}$ de la catégorie C_j pour le terme q_i est :

$$w_{i,j} = \begin{cases} max_p(masse\ de\ A_p) & si\ A_p\ \text{à}\ pour\ catégorie\ C_j \\ 0 & sinon \end{cases}$$

Enfin, un score pour chaque catégorie est calculé, en sommant $w_{i,j}$ pour chaque terme q_i de la requête Q. Les meilleures catégories permettent, au final, d'étendre la requête.

Pour finir, (David Milne, 2010) propose le moteur de recherche conceptuelle KORU[22]. Il identifie les concepts candidats (articles de Wikipédia) présents dans la requête (par synonymie ou par similarité) que l'utilisateur peut activer ou désactiver. Une fois les concepts choisis, une expansion de la requête est alors effectuée en utilisant l'ensemble des labels dénotant les concepts choisis.

4.5.5 Conclusion

Nous avons présenté Wikipédia ainsi que les méthodes permettant d'exploiter les connaissances en les transformant en réseau sémantique ou ontologie. Ensuite, nous avons présenté les travaux des

[22] http://www.greenstone.org/greenstone3/koru2.0/

équipes de recherche qui, depuis 2005, utilisent les bases de connaissances issues de Wikipédia pour traiter des problèmes connexes à la recherche d'information (similarité sémantique, désambiguïsation, extraction des thématiques et des mots-clés et recherche documentaire). Le Tableau 5 récapitule ces travaux.

	Utilisation					Connaissances Wikipédia utilisées			
	Similarité	WSD	Extraction thématique	Extraction de mots-clés	Recherche documentaire	Corpus Wikipédia	Liens interwiki Wikipédia	Réseau de Catégories Wikipédia	Autres
(Strube & Ponzetto, 2006)	✓							✓	
(Gabrilovich & Markovitch, 2007)	✓					✓			
(Witten & D. Milne, 2008)	✓						✓		
(Mihalcea, 2007)		✓				✓	✓		
(Cucerzan, 2007)		✓					✓	✓	
(Fogarolli, 2009)		✓					✓		
(Mihalcea & Csomai, 2007)		✓		✓		✓	✓		
(Medelyan, Witten, & D. Milne, 2008)	✓	✓	✓			✓	✓		
(Janik & Kochut, 2008)			✓						DBpedia
(Coursey, Mihalcea, & Moen, 2009)			✓				✓	✓	
(Müller & Gurevych, 2009)					✓	✓			
(Alemzadeh & Karray, 2010)					✓			✓	
(David Milne, 2010)					✓		✓		

Tableau 5 : Tableau récapitulatif des méthodes utilisant Wikipédia pour des tâches connexes à la recherche d'information

En observant ce tableau récapitulatif, nous réalisons l'importance que Wikipédia a pris dans le domaine de l'ingénierie des connaissances et de la recherche d'information en 5 ans. En général, les méthodes présentées contribuent aux calculs de la similarité, à l'extraction et à la recherche d'information, en utilisant tantôt le corpus de Wikipédia, tantôt le réseau de catégories de Wikipédia.

Discussion

Afin de concevoir un outil d'aide à l'indexation et à la recherche documentaire, nous devons, avant tout, comprendre les notions fondamentales de la recherche d'information à savoir, la pondération des termes d'un document et les mesures de similarité entre la requête et un document. La pondération permet de déterminer le sous-ensemble des termes d'un document qui le représente ou le discrimine le mieux par rapport aux documents du corpus. La similarité entre la requête et un document permet de répondre à l'intention de l'utilisateur, à savoir trouver les documents qui se rapprochent le plus de sa requête.

Les approches classiques d'indexation et de recherche d'information consistent à analyser les documents d'un corpus afin de le représenter en utilisant des modèles ensemblistes, algébriques ou probabilistes. Ces approches exploitent les statistiques sur la fréquence des termes dans un corpus (modèle booléen et vectoriel) ou les statistiques sur les documents déjà indexés (modèle probabiliste).

Le problème avec les approches de la recherche d'information classiques est que la pondération des termes est souvent plus discriminante que représentative. Nous préférons avoir une structure représentative du document permettant d'extraire directement les mots-clés en fonction d'une thématique donnée. Autrement dit, profiter d'une structure externe permettant d'informer le système sur les relations existantes entre les termes du document et ces thématiques (ou concepts).

Nous avons étudié les méthodes de recherche d'information étendue par une base de connaissances externes. Ces approches consistent à utiliser les connaissances (dans notre contexte des concepts et relations entre ces concepts). La première étape d'une indexation guidée par une base de connaissances est de mettre les termes en correspondance avec les concepts de cette base. Ensuite, le document n'est plus indexé, contrairement à l'indexation classique, par ses propres termes, mais par des concepts. Au niveau de la recherche d'information, la requête est également représentée par des concepts. La requête peut être ainsi étendue pour retourner d'autres documents en fonction de la proximité entre concepts du document et de la requête étendue. Nous avons également vu l'utilisation des mesures de similarité entre concepts pour la tâche de désambiguïsation. La tâche de désambiguïsation revient souvent à exploiter la similarité entre concepts pour déduire quel concept choisir pour un terme polysémique. Le terme est ainsi désambiguïsé en trouvant le concept candidat maximisant la similarité avec les autres concepts déjà choisis pour décrire le document.

Parmi les approches présentées, nous exploitons plusieurs résultats. L'approche de la diversité conceptuelle (Agirre & Rigau, 1996) pour la désambiguisation nous intérresse dans la mesure où elle désambigüise les termes en utilisant les concepts les plus fédérateurs d'un document en utilisant une base de connaissances. Ainsi, nous pouvons détourner cette mesure pour trouver les concepts

importants d'un document. Pour la recherche d'information, nous retenons le retour d'expérience de (Mihalcea & Moldovan, 2000) dont la conclusion est que la recherche d'information sémantique et conceptuelle améliore sensiblement les performances d'un système. En revanche, il est important qu'une requête ne soit pas exagérément étendue pour ne pas affaiblir la précision des résultats. Aussi, nous retenons les travaux de (Sanderson, 1994) qui estiment que la désambigüisation améliore un SRI, si et seulement si cette dernière est accomplie avec un succès de 90%. Finalement, nous sommes amenés à utiliser la méthode de maximisation décrite par (Navigli, 2009) pour la désambiguisation en utilisant les mesures introduites par (Rada et al., 1989), (Z. Wu & Palmer, 1994) et (Leacock & Chodorow, 1998).

Nous avons décrit la similarité entre collections de concepts. Notre objectif est d'exploiter ces mesures pour la recherche d'information en supposant qu'un document peut être représenté par une collection de concepts. Cette notion est importante car c'est sur elle que notre travail se fonde pour le calcul de similarité entre documents, pour le calcul d'appariement requête/document mais aussi pour la détection des changements de thèmes et les retours sémantiques au sein d'un même document.

Parmi les méthodes présentées, nous retenons surtout la mesure du modèle vectoriel généralisé introduit par (Ganesan et al., 2003) initalement pour accomplir du filtrage collaboratif. Nous détournons cette mesure pour nos besoins à savoir la similarité entre documents.

Nous souhaitons finalement utiliser Wikipédia comme base de connaissances pour extraire une structure de concepts permettant :

- d'identifier la thématique d'un document ;
- d'extraire les concepts importants d'un document ;
- d'extraire les termes importants d'un document ;
- de désambiguïser les termes d'un document ;
- de détecter les changements de thèmes et les retours sémantiques dans un document ;
- de rechercher un document textuel à partir d'une requête ;
- et de rechercher un document à partir d'un autre document.

Un de nos objectifs étant d'extraire les mots-clés du texte, nous sommes très intéressés par la mesure de *Keyphraseness* introduite par (Mihalcea & Csomai, 2007) dans leur système *Wikify !*

Avec ces éléments, nous sommes en mesure de construire un outil d'aide à l'indexation adjoint à un système de recherche d'information pour les documents pédagogiques. Les méthodes et les outils développés s'inspirent des méthodes d'indexation et de recherche d'information sémantique et conceptuelle associées à la base de connaissances Wikipédia.

Partie II. Contribution

Introduction

Cette partie concerne l'ensemble de nos travaux de recherche pour la conception d'un outil d'aide à l'indexation et à la recherche documentaire.

Notre objectif est de proposer aux documentalistes d'UNIT un outil permettant de les guider dans le choix des descripteurs pour l'indexation des documents pédagogiques. La phase d'indexation des documents numériques pédagogiques est fastidieuse, entre lecture complète d'un document dont le documentaliste ne connaît que partiellement le domaine et choix cohérent et consistant des descripteurs. Une fois cette lourde tâche accomplie, le document peut être catégorisé en fonction des mots-clés et des thématiques choisies. Par exemple, un cours, indexé par les documentaliste d'UNIT, traitant du « modèle relationnel »[23] est catégorisé dans le domaine UNIT « informatique » et le sous domaine « base de donnés » et les descripteurs choisis pour mots-clés sont les suivants :

conception de base de données, SGBD, modèle relationnel, domaine, produit cartésien, relation, SQL, clé étrangère, enregistrement, passage UML vers relationnel, algèbre relationnelle, système de gestion de base de données, Structured Query Language.

Pour la description de cours, d'autres champs sont présents comme le titre, l'auteur et la description (résumé). C'est une ressource pédagogique correctement indexée et cataloguée. Nous nous proposons de simuler une recherche de document afin d'évaluer rapidement et subjectivement, l'apport de l'indexation sur la recherche d'information. Pour cela, nous utilisons le moteur de recherche d'UNIT.

CAS 1 :
Intention : " Je cherche un document qui traite des clefs étrangères"
Requête : "Clefs étrangères"
Résultat : Aucun document trouvé
Le document n'est pas indexé par « clefs étrangères » mais par « clé étrangère » !

CAS 2 :
Intention : " Je cherche un document qui traite des clés primaires"
Requête : "Clé primaire"
Résultat : Aucun document trouvé
Bien que le document traite autant de « clé primaire » que de « clé étrangère », le documentaliste a choisi d'omettre le descripteur « clé primaire ».

[23] http://www.unit.eu/ori-oai-search/notice/view/unit-ori-wf-1-3107

CAS 3 :
Intention : " Je cherche un document qui traite des SGBDR ou des systèmes de gestion de bases de données relationnelles"
Requête 1 : "SGBDR"
Résultat 1 : le document recherché est placé en 6ième posistion sur 6 documents trouvés
Requête 2 : "système de gestion de bases de données relationnelles"
Résultat 2 : le document n'est pas trouvé mais 2 documents ont été trouvés

Dans la requête 1, le terme « SGBDR » ne se trouve que dans la description et est considéré moins important que les documents ayant comme mots-clés « SGBDR » ou contenant « SGBDR » dans leur titre. Dans la requête 2, le terme « système de gestion de bases de données relationnelles » n'est référencé nul part pour le document.

Plusieurs conclusions émergent de cette expérience. La première est qu'un mot-clé contribue de façon majeure à la recherche documentaire uniquement s'il n'est pas présent dans les autres champs (description, titre, etc.). La deuxième conclusion est que l'utilisateur doit s'adapter à l'orthographe des mots-clés pour retrouver les documents les utilisant et il doit connaître le vocabulaire utilisé par les documentalistes. La troisième conclusion révèle qu'un document peut à la fois être indexé par un mot-clé et par un synonyme en contexte. Les requêtes « SGBDR » et « système de gestion de bases de données relationnelles » sont a priori équivalentes, mais selon l'indexation, la recherche d'information nous propose deux ensembles de documents.

Les mots-clés sont donc importants dans le processus global de recherche d'information. Pour confirmer la pertinence d'un document, l'utilisateur peut consulter la liste de mots-clés. En revanche, les mots-clés ont un faible poids dans la recherche documentaire proprement dite. Le processus de recherche d'information ne profite pas de l'effort fourni par les documentalistes dans le processus d'indexation et de catalogage.

Notre objectif est aussi de faire en sorte que l'indexation contribue pleinement à la phase de recherche documentaire. Notre principale idée est d'étendre la requête de l'utilisateur mais également les descripteurs par rapport aux connaissances d'un expert du domaine en ayant recours aux bases de connaissances externes. Pour passer de la notion de « clé primaire » à « base de données », nous utilisons tout type de base de connaissances présentant des relations hiérarchiques permettant de passer d'un terme vers son domaine. Cette base se substitue aux connaissances des experts des domaines traités par les documents d'une bibliothèque numérique. Elle permettra d'identifier les thématiques abordées, les concepts liés à ces thématiques et les mots-clés du document.

Afin de sélectionner les descripteurs des documents pédagogiques, nous faisons deux hypothèses :

Hypothèse 1 : les termes d'un document permettent de détecter les thématiques traitées dans ce document.

Hypothèse 2 : les thématiques traitées dans un document permettent de sélectionner les descripteurs du document.

Sous ces hypothèses, en partant des termes d'un document et en nous appuyant sur une base de connaissances, nous souhaitons extraire ses thématiques. A partir de ces thématiques, nous souhaitons sélectionner des descripteurs présents dans le texte ou non, issus de la même base de connaissances.

Plusieurs difficultés découlent de cette approche :

1. détecter une thématique à partir de tous les termes d'un document (avec des termes « parasites ») ;
2. une fois la thématique identifiée, sélectionner les descripteurs du document ;
3. désambiguïser les termes du document (avant ou après la recherche de la thématique).

Pour la recherche d'information, nous utilisons les thématiques, les descripteurs mais aussi la base de connaissances pour rechercher un document répondant à une requête. Nous émettons deux nouvelles hypothèses :

Hypothèse 3 : dans le cadre des documents pédagogiques, utiliser uniquement les descripteurs d'un vocabulaire contrôlé pour la recherche d'information n'est pas pertinent.

Hypothèse 4 : dans le cadre des documents pédagogiques, pour être pertinent, un descripteur doit être accompagné d'informations décrivant ce descripteur.

Nous utilisons une base de connaissances pour l'extraction des thématiques et des descripteurs. Cette même base est utilisée pour étendre les descripteurs du document et les termes de la requête lors de la phase de recherche documentaire.

Nous proposons, dans un premier chapitre, une représentation du document en s'adossant à une base de connaissances hiérarchiques. Cela permet de décrire chaque document par une structure de données, en l'occurrence un graphe. L'idée est de conserver pour chaque terme du document, des informations issues de la base de connaissances permettant de décrire ce terme. Les graphes sont composés de nœuds qui sont des concepts et d'arêtes qui sont des relations entre concepts issues de la base. Nous introduisons les opérations que nous effectuons sur ces graphes. Nous validons cette représentation par l'utilisation du réseau de catégories Wikipédia.

Le second chapitre débute par une observation sur les graphes des documents. Leurs spécificités permettent d'identifier les concepts importants du graphe et donc du document. Ces concepts permettent de sélectionner des thématiques, ainsi les descripteurs associés. Ensuite, nous proposons de lever l'ambiguïté des termes polysémiques. Enfin, nous présentons les protocoles d'évaluation de l'extraction des thématiques, des descripteurs et de la désambigüisation.

Le troisième chapitre traite de la similarité entre graphes représentant des documents. Nous présentons des mesures de similarité permettant de comparer une requête et un document, deux documents et deux sections d'un même document. La première mesure est le socle d'un moteur de recherche, la seconde celui d'un système de recommandation et la troisième permet d'extraire la structure thématique d'un document, avec la reconnaissance de ruptures et de retours sémantiques. De plus, nous exploitons ces mesures afin de développer une nouvelle approche pour la désambiguïsation des termes polysémiques. Nous présentons, finalement, l'évaluation des métriques de similarité pour ces différents cas de figure.

Le quatrième et dernier chapitre présente un prototype pour les documentalistes d'UNIT, pour tester, de manière subjective, l'utilité d'un tel outil d'aide.

Chapitre 1. Modèle de représentation d'un document

1.1 Introduction

Dans ce chapitre, nous présentons les graphes comme représentation du contenu textuel d'un document pour effectuer plusieurs opérations comme l'extraction des thématiques et des descripteurs pour l'indexation. Ces graphes sont issus d'une base de connaissances composée de concepts hiérarchiquement liés par des relations « générique/spécifique » entre concepts.

Nous introduisons les notions de graphe d'un concept et de graphe de concepts. Ces graphes sont les représentations conceptuelles d'un concept ou de plusieurs concepts d'une base en fonction d'autres concepts de cette base. Cette famille de graphes nous permettra ensuite de construire un graphe représentant un terme. En effet, un terme peut être directement lié à un ou plusieurs concepts. Cette association terme/concepts sera possible grâce à un dictionnaire construit en amont. Finalement, nous définirons la notion de graphe de document qui est une union ou une fusion des graphes des termes d'un document. Les graphes de documents permettront de représenter conceptuellement le document pour l'extraction des descripteurs (tâche traitée dans le chapitre suivant).

Nous nous limitons, dans ce chapitre, à la présentation de la construction des graphes via une base de connaissances hiérarchiques. Dans une première section (section 1.2), nous présentons la construction d'un graphe pour un terme d'un document, en utilisant les graphes d'un concept ou les graphes de concepts et un dictionnaire de correspondance terme/concepts construit en amont. Nous donnerons des exemples de graphe de termes issus du réseau de catégories Wikipédia et un dictionnaire construit grâce au corpus Wikipédia et au réseau de catégories de Wikipédia. La construction du graphe du document est décrite dans la section 1.3. Dans la même section, nous présentons un ensemble d'opérations sur la structure de « graphe de document ». Finalement, nous validons notre modèle en utilisant notamment le réseau de catégories issues de Wikipédia dans la section 1.4.

1.2 Création d'un graphe de terme

Soit BC une base de connaissances possédant des relations hiérarchiques. Soit $C = \{C_1, C_2, .., C_n\}$ l'ensemble des n concepts présents dans BC. Chaque concept de C peut être d'un type différent TC (par exemple, un terme descripteur, un non-descripteur, un article, une catégorie, un portail, etc.). Un concept peut être lié à d'autres concepts par une relation R_k d'un certain type TR. Soit $R = \{R_1, R_2, .., R_m\}$, l'ensemble de m relations entre concepts. Par exemple, pour C_1 hyperonymes C_2, on écrit le triplet $(C_1, hyper, C_2)$ ou $C_1 \xrightarrow{hyper} C_2$.

Une base de connaissances BC est définie par :

$$BC = (C, TC, R, TR)$$

Nous définissons une sous-base de connaissances BC' comme une restriction de la base originale :

$$BC' = (C' \subseteq C, TC' \subseteq TC, R' \subseteq R, TR' \subseteq TR)$$

Par exemple, le réseau de catégories BC de Wikipédia peut être défini par :

- $BC = Wikipédia\ en\ Français$
- $TC = \{Article, Catégorie, Redirection, Portail, etc.\}$
- $TR = \{Redirection \rightarrow Article, Article \rightarrow Article, Article \rightarrow Catégorie, Catégorie \rightarrow Catégorie, etc.\}$
- $C = Toutes\ les\ pages\ Wikipédia$
- $R = Toutes\ les\ relations\ entre\ pages\ Wikipédia$

Un extrait de Wikipédia peut être une sous-base de connaissances hiérarchique BC' :

- $BC' = Sous - ensemble\ de\ Wikipédia\ en\ français$
- $TC' = \{Article, Catégorie\}$
- $TR' = \{Article \rightarrow Catégories, Catégorie \rightarrow Catégorie\}$
- $C' = Tous\ les\ articles\ et\ les\ catégories$
- $R' =$ *Les relations permettant de passer d'un article ou d'une catégorie à une catégorie supérieure*

1.2.1 Graphe d'un concept

Nous introduisons la notion de graphe d'un concept (différente de celle de graphes conceptuels (J. F Sowa, 1983; John F Sowa, 1976)) où les nœuds sont appelés concepts et les arcs appelés relations. A partir d'un concept issu de BC', le graphe est construit en remontant les relations hiérarchiques ascendantes du concept, jusqu'aux concepts racines de la base. Les opérations définies sur le type de données « graphe d'un concept » sont les suivantes :

- initialiser() : crée un graphe vide (sans concept, ni relation);
- mettreConcept(Concept C) : permet de définir le concept C comme le concept initial ;
- récupérerConcept() : retourne le concept initial ;
- ajouterConcept(Concept C) : ajoute un concept « C » au graphe, si « C » n'existe pas ;
- ajouterRelation(Relation $C_1 \overset{TR}{\rightarrow} C_2$) : ajoute une relation de type TR, entre les concepts C_1 et C_2, si la relation n'existe pas. Si les concepts existent déjà, la relation TR est créée, si un seul concept existe, alors l'autre est crée et relié au premier. Finalement si aucun des concepts n'existe alors les deux concepts sont crées et reliés.
- contient(Concept C) : retourne vrai si le concept existe dans le graphe ;
- contient(Relation $C_1 \overset{TR}{\rightarrow} C_2$) : retourne vrai si la relation existe dans le graphe ;
- récupérerRelations(Type de relation TR) : retourne toutes les relations $C_i \overset{TR}{\rightarrow} C_j$;
- récupérerConcepts(Type de concept TC) : retourne tous les concepts C_i de type TC ;
- récupérerRelationsFils(Concept C, Type de Relation TR) : retourne les relations $C \overset{TR}{\rightarrow} C_i$;
- récupérerFils(Concept C, Type de Relation TR) : retourne les concepts C_i tels que $C \overset{TR}{\rightarrow} C_i$;
- récupérerRelationsPère(Concept C, Type de Relation TR) : retourne les relations $C_i \overset{TR}{\rightarrow} C$;
- récupérerPères(Concept C, Type de Relation TR) : retourne les concepts C_i tels que $C_i \overset{TR}{\rightarrow} C$;

Nous définissons l'opération de construction du graphe de concept :

$$graphe(C, BC') = grapheC(C, BC', GrapheVide)$$

Cette opération est implémentée par l'algorithme récursif suivant :

Fonction : Graphe de concept = grapheC
Rôle : Crée le graphe de concept C utilisant la base BC'
Entrée : C: Concept, BC': Base de Connaissances
Entrée/Sortie : G : Graphe du concept C

Début
  Si $C \neq \emptyset$ *alors*
    $\{R_i\} \leftarrow$ *L'ensemble des relations*(C_i, TR_j, C) *avec* $TR_j \in TR'$ *et* $C_i \in C'$
    Pour chaque $\{R_i\}$ *faire*
      Si non $G.contient(R_i)$ *alors*
        Si non $G.contient(C_i)$ *alors*
          $G.ajouterRelation\left(C_i \overset{TR_j}{\rightarrow} C\right)$
          $grapheC(C_i, BC', G)$
        Sinon
          $G.ajouterRelation\left(C_i \overset{TR_j}{\rightarrow} C\right)$
        FinSi
      FinSi
    FinPour
  FinSi
Fin

Par exemple, construisons le graphe du concept « classification » issu de la base de connaissances Wikipédia en français. Nous ne conservons que les types de concepts et les types de relations suivants :

- $BC' = Sous - ensemble\ de\ Wikipédia\ en\ français$
- $TC' = \{Article, Catégorie\}$
- $TR' = \{Article \rightarrow Catégories, Catégorie \rightarrow Catégorie\}$
- $C' = Tous\ les\ articles\ et\ les\ catégories$
- $R' =$

  $Les\ relations\ permettant\ de\ passser\ d'un\ article\ ou\ d'une\ catégorie$ à $une\ catégorie\ supérieure$

La Figure 11 montre les étapes de l'algorithme pour la création de $graphe(« Classification »\ , BC')$.

Les catégories Wikipédia sont exclusivement représentées par des nœuds en forme d'ellipse.

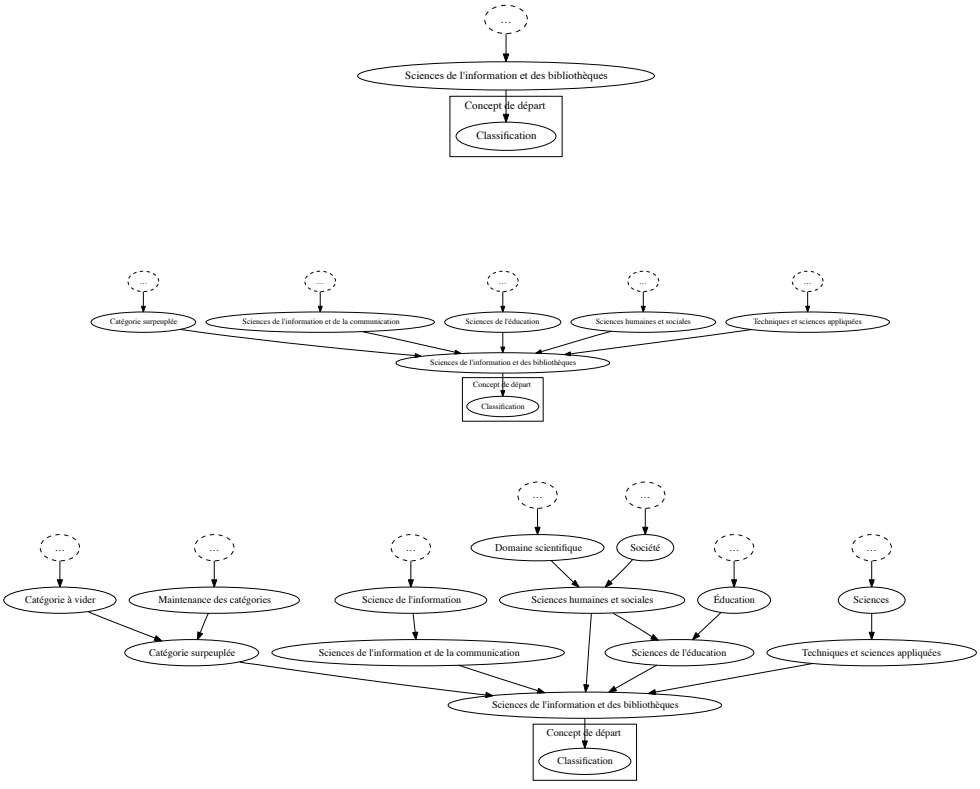


Figure 11 : Création du graphe du concept « Classification » de Wikipédia

1.2.2 Graphe de concepts

La création d'un graphe de plusieurs concepts consiste à générer les graphes représentant chacun des concepts et à les fusionner en un seul regroupant les concepts de même nom. Le type de données « graphes de concepts » possède les mêmes opérations que le « graphe de concept » à l'exception de :

- mettreConcepts(Concept C_1,...,C_k) : permet de définir les concepts initiaux ;
- récupérerConcepts() : retourne les concepts initiaux ;

Nous avons développé un algorithme permettant de générer un graphe de concepts en limitant les accès à la base de connaissances. La construction du graphe ce fait par palier. L'ensemble initial de concepts représente le premier palier de construction. Pour chaque concept de ce palier, nous extrayons les relations ascendantes. Les concepts parents ainsi générés sont les concepts du second palier. L'algorithme s'arrête lorsque toutes les relations ascendantes de tous les concepts du graphe sont récupérées :

Fonction : Graphe de concepts = grapheC
Rôle : Crée le graphe de concept de plusieurs concepts C_k via la base BC'
Entrée : C_k : $Ensemble\ de\ Concepts$, BC': $Base\ de\ Connaissances$
Sortie : G : $Graphe\ de\ concepts$

Début
  $\{C_i'\} \leftarrow \{C_k\}$ *//$\{C_i'\}$: les concepts à traiter*
  $\{NC_i'\} \leftarrow \emptyset$ *//$\{NC_i'\}$: les nouveaux concepts sources*
  Si $\{C_i'\} \neq \emptyset$ *alors*
    Pour chaque $\{C_i'\}$ *faire*
      $\{R_i\} \leftarrow$ *L'ensemble des relations*(C_i, TR_j, C_i') *avec* $TR_j \in TR'$ *et* $C_i \in C'$
      Pour chaque $\{R_i\}$ *faire*
        Si non $G.contient(R_i)$ *alors*
          Si non $G.contient(C_i)$ *alors*
            $G.ajouterArête\left(C_i \xrightarrow{TR_j} C_i'\right)$
            $\{NC_i'\} \leftarrow \{NC_i'\} + C_i$
          Sinon
            $G.ajouterArête\left(C_i \xrightarrow{TR_j} C_i'\right)$
          FinSi
        FinSi
      FinPour
    FinPour
    $grapheC(\{NC_i'\}, BC', G)$
  FinSi
Fin

L'opération de construction du graphe de concepts s'exprime de la manière suivante :

$$graphe(\{C_k\}, BC') = grapheC(\{C_k\}, BC', GrapheVide)$$

La Figure 12 montre les trois premiers paliers du graphe des concepts « classification » et « tice », $graphe\big(\{\text{« } Classification \text{ »}, \text{« } Tice \text{ »}\}, BC'\big)$.

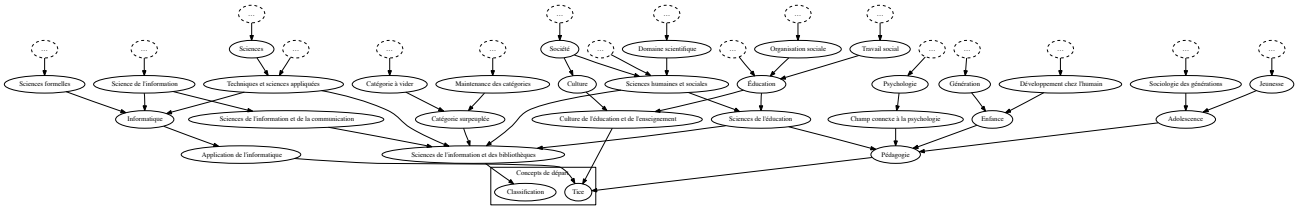


Figure 12 : Consctruction du graphe des concepts {« Classification », « Tice »}

1.2.3 Création d'un dictionnaire

Soit $Dico$ un dictionnaire composé de N entrées, $Dico = \{e_1, e_2, \dots, e_N\}$. Chaque e_k est en relation avec un ou plusieurs concepts de BC. Cette relation signifie que le terme e_k dans un texte fait référence à un concept ou plusieurs concepts de la base. Par exemple l'entrée e_1 = « Bois » est en relation avec les concepts C_i = « Bois_(golf) » et C_j = « Bois_(instrument) » issus de BC. L'opération $concepts$ est la suivante :

$$concepts(e_k, BC') = \{C_i | e_k \; est \; en \; relation \; avec \; C_i \; et \; C_i \in C'\}$$

Et l'opération inverse $entrées$:

$$entrées(C_i) = \{e_k | k, e_k \; est \; en \; relation \; avec \; C_i\}$$

L'opération $concepts(e_k, BC')$ assure le passage d'une entrée de *Dico* vers un ensemble de concepts de *BC'* (dans notre cas, un extrait de Wikipédia). *Dico* est construit comme suit :

1. Dans un premier temps, tous les titres des catégories et des articles de Wikipédia sont considérés comme des entrées e_k de *Dico* pointant vers le concept de la catégorie ou de l'article en question. Si un titre est à la fois un article et une catégorie, l'entrée e_k pointe alors vers le concept de la catégorie. Ce choix provient du fait que, le plus souvent, un article portant le même nom qu'une catégorie est en relation hiérarchique avec cette catégorie (par exemple, l'article « France » est en relation avec la catégorie « France » par la relation « $Article \rightarrow Catégorie$ »). Ainsi, il n'y pas de doublon dans les intitulés des concepts.
2. Certaines catégories de Wikipédia sont ambigües et leur titre est de la forme « Nom_article_(Domaine_traité) ». Dans ce cas, nous créons une seule entrée e_k de nom « Nom_article » et autant de concepts liés à cette entrée qu'il y a de sens pour cette catégorie. Par exemple, « Python_(serpent) » et « Python_(langage) » sont deux concepts liés à l'entrée « Python ».
3. Si la catégorie ou l'article est une redirection vers un autre article, nous créons une entrée avec la catégorie ou l'article de départ vers le concept correspondant à l'article pointé (par exemple, « algorithme » est le titre d'un article qui redirige vers l'article de titre

« algorithmique » ; de plus, il existe déjà une catégorie de nom « algorithmique ». En conséquence, nous créons une entrée e_k « algorithme » reliée à la catégorie « algorithmique »).

4. Certains articles de Wikipédia contiennent des liens internes vers d'autres articles. Ces liens sont de type (nom_article|terme_affiché). Si « terme_affiché » n'est pas une entrée du dictionnaire *Dico, alors* nous créons e_k pointant vers la catégorie ou l'article « nom_article » (par exemple, il existe dans Wikipédia le lien interne (typage_fort|fortement typé) où « fortement typé » n'est pas une entrée de *Dico* ; il n'existe pas de catégorie « typage_fort »).

Le Tableau 6 récapitule les cas possibles lors de la création du dictionnaire de correspondance entre les entrées e_k et les concepts associés $concepts(e_k, BC')$.

Entrée e_k	$concepts(e_k, BC')$	Explication
France	France, France_(Prénom), France_(Paquebot de 1912), etc,	France est à la fois un article et une catégorie. Nous gardons la catégorie. De plus « France » est un terme ambigu : 19 articles ont pour titre « France_(*)».
La France	France	Redirection vers l'article « France ». Un lien entre « La France » et la catégorie « France » est créé.
Algorithme	Algorithmique	Redirection vers l'article « algorithmique ». La catégorie « algorithmique » existe, nous créons donc le lien de « algorithme » vers la catégorie « algorithmique ».
Python	Python, Python_(langage), Python_(serpent), Python_(téléfilm), etc.	« Python » est une catégorie (cette catégorie représente le langage de programmation), il existe également 10 articles ambigus commençant par « Python ».
Fortement typé	Typage_fort	Il existe un lien (typage_fort|fortement typé). De plus, il n'existe pas d'article ou de catégorie s'intitulant « fortement typé », d'où la création de la relation de « fortement typé » vers « typage fort »

Tableau 6 : Entrées de *Dico* et les pages associées (Les catégories sont soulignées).

Avec les données de Wikipédia[24], notre dictionnaire possède :

- 166 607 catégories ;
  - 4 263 catégories ambiguës ;
  - 11 catégories redirigées ;
- 2 146 581 articles ;
  - 226426 articles ambigus ;
  - 1083744 articles redirigés ;
- 38 700 738 liens internes dont 5 939 744 différents ;
  - 3 885 275 liens différents de type (X|X) vers des articles (30 268 506 liens au total) ;
  - 1 959 254 liens différents de type (X|Y) vers des articles (8 268 804 liens au total) ;
  - 66 802 liens différents de type (X|X) vers des catégories (120 903 liens au total) ;
  - 28 413 liens différents de type (X|Y) vers des catégories (42 525 liens au total).

[24] Données Wikipédia datant du 01/02/2011 (http://dumps.wikimedia.org/frwiki/20110201/)

Il existe plus de liens internes différents de type (X|X) vers des articles que d'articles dans Wikipédia. Ce phénomène est explicable : de nombreux articles de Wikipédia possèdent des liens internes vers des articles qui n'existent pas. Ces liens apparaissent en rouge dans la version web de Wikipédia, et sont conservés pour encourager la création de nouveaux articles portant ces titres.

Suivant cette approche, nous obtenons un dictionnaire *Dico* peuplé de 4 549 392 entrées.

1.2.4 Graphe de terme

Nous souhaitons construire, à partir d'un terme t, un graphe issu de BC ou de BC' en utilisant $Dico$. La première étape consiste à faire la correspondance entre le terme t et les entrées de $Dico$. L'opération de correspondance s'écrit ainsi :

$$match(t, Dico) = \begin{cases} e_k & si\ e_k\ correspond\ \text{à}\ t \\ \emptyset & si\ \forall i, \nexists e_i\ correspondant\ \text{à}\ t \end{cases}$$

Si une entrée e_k correspond au terme t (est similaire à t), alors t est associé à e_k. La similarité peut être l'égalité stricte entre e_k et t (au sens des chaînes de caractères) ou encore l'égalité de la lemmatisation ou lexèmisation de e_k et t. Il est possible qu'aucune entrée de $Dico$ ne corresponde au terme t ; dans ce cas, l'opération retourne l'ensemble vide $\emptyset$.

S'il y a correspondance du terme t avec l'entrée e_k, alors il est possible de créer un graphe issu de BC' par $grapheTerme(t, Dico, BC')$. Cela consiste à trouver l'entrée $match(t, Dico) = e_k$, si elle existe, et d'extraire de la base l'ensemble des concepts par $concepts(e_k, BC')$ avec lesquels l'entrée est en relation (voir Figure 13). Une partie du graphe de terme est alors le graphe de concepts (avec pour concept initial le concept associé au terme) :

$$grapheTerme(t, Dico, BC') \supset graphe(concepts(match(t, Dico)), BC')$$

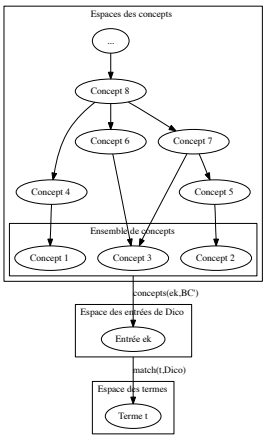


Figure 13 : Graphe de terme

Le type de données « graphe de terme » possède les opérations suivantes :

- initialiser() : pour créer un graphe vide ;
- mettreTerme() : permet de définir le terme initial ;
- récupérerTerme() : permet d'extraire le terme t du graphe ;
- ajouterEntrée(Entrée e_k) : relie le terme t à l'entré e_k ;
- ajouterConcepts(Concepts C_1,C_2…C_n) : relie les n concepts C_i à l'entrée e_k ;
- récupérerGrapheDeConcepts() : permet d'extraire le graphe des concepts.
- récupérerGrapheDeConcept(Concept C) : permet d'extraire le graphe de concept du concept C.

Pour réaliser l'alignement entre les termes du document et les entrées du *Dico* (c'est-à-dire trouver la fonction de correspondance $match(t, Dico)$), nous décrivons deux méthodes : la première utilise l'égalité stricte de chaines de caractères entre un terme et les entrées de *Dico*, et la seconde propose d'exploiter des outils de Traitement Automatique des Langues (TAL).

1.2.4.1 Alignement brut

Un document est composé de mots. Nous utilisons le terme « préfixe » comme les premières lettres d'une chaine de caractères. Par exemple, « La chambre » est un préfixe de « La chambre jaune ».

Nous disposons de 3 listes : une liste de préfixes potentiels, une liste de préfixes réels et une liste résultat dont les éléments sont les termes détectés dans le document (c'est-à-dire correspondants à une entrée du dictionnaire).

Nous initialisons la liste de préfixes potentiels par les mots du document. Pour chaque préfixe potentiel nous vérifions s'il est un préfixe réel d'une entrée du dictionnaire. Nous construisons ainsi une liste de préfixes réels. Si un préfixe réel correspond strictement à une entrée du dictionnaire, alors il est ajouté à la liste résultat. Ensuite, nous construisons une nouvelle liste de préfixes potentiels composées des préfixes réels concaténés, pour chaque préfixe, au mot suivant dans le document et nous réitérons les opérations précédentes jusqu'à ce que nous ne puissions plus compléter la liste de préfixes réels. Lorsqu'il y a reconnaissance de plusieurs entrées de *Dico*, nous ne conservons que la plus grande (en nombre de mots).

Prenons l'exemple du document « Albert Einstein a introduit la théorie de la relativité restreinte ».

Dans la première étape, les termes « Albert », « Einstein », « a », « la », « théorie » et « relativité » sont des préfixes d'entrées du dictionnaire. Les mots « introduit», « de » et « restreinte » ne le sont pas et ne sont donc plus candidats.

Dans la seconde étape, les termes « Albert Einstein », « Einstein a », « a introduit », « la théorie », « théorie de », « la relativité » et « relativité restreinte » sont analysés. Seul « Albert Einstein », « la théorie », « théorie de », « relativité restreinte » sont des préfixes. Nous rejetons les autres.

La troisième étape consiste à analyser, « Albert Einstein a», « la théorie de », « théorie de la » et nous rejetons « Albert Einstein a ».

Dans la quatrième étape, « la théorie de la », « théorie de la relativité » sont analysés et conservés. De même que « la théorie de la relativité » et « théorie de la relativité restreinte », dans la cinquième étape et « la théorie de la relativité restreinte », dans la sixième étape

Ensuite, nous ne conservons que les termes candidats ayant une entrée qui lui correspond exactement (Tableau 7) :

Etape	Liste des préfixes potentiels	Liste des préfixes réels	Termes rejetés	Liste résultat
1	« Albert » « Einstein» « a » « introduit » « la » « théorie » « de » « la » « relativité » « restreinte »	« Albert » « Einstein » « a » « la » « théorie » « la » « relativité »	« de » « introduit » « restreinte »	**« Albert »** **« Einstein »** **« théorie »** **« relativité »**
2	« Albert Einstein» « Einstein a », « a introduit » « la théorie » « théorie de » « la relativité » « relativité restreinte »	« Albert Einstein» « la théorie » « théorie de » « la relativité » « relativité restreinte »	« Einstein a », « a introduit »	« Albert » « Einstein » « théorie » « relativité » **« Albert Einstein »** **« relativité restreinte »**
3	« Albert Einstein a» « la théorie de » « théorie de la » « la relativité restreinte »	« la théorie de » « théorie de la »	« Albert Einstein a»	« Albert » « Einstein » « théorie » « relativité » « Albert Einstein » « relativité restreinte »
4	« la théorie de la » « théorie de la relativité »	« la théorie de la » « théorie de la relativité »		« Albert » « Einstein » « théorie » « relativité » « Albert Einstein » « relativité restreinte » **« théorie de la relativité »**
5	« la théorie de la relativité » « théorie de la relativité restreinte»	« la théorie de la relativité » « théorie de la relativité restreinte »		… « théorie de la relativité » **« la théorie de la relativité »** **« théorie de la relativité restreinte »**
6	« la théorie de la relativité restreinte»	« la théorie de la relativité restreinte »		… … « théorie de la relativité restreinte » **« la théorie de la relativité restreinte »**

Tableau 7 : Extraction des termes d'un texte par alignement brut

Finalement, nous ne gardons que les termes de plus grande taille (« Albert Einstein » contient « Albert » et « Einstein » que nous supprimons) et rejetons les termes vides (aucun mot vide n'a été détecté comme entrée candidate). Nous obtenons donc le résultat suivant :

« *Albert Einstein* *a introduit* *la théorie de la relativité restreinte* »

Cette approche est relativement rapide car elle permet de trouver tous les termes d'un document, en analysant un nombre réduit de termes candidats. A titre d'exemple, nous n'avons questionné *Dico* que 24 fois alors que si nous avions analysé tous les termes composés de mots connexes, nous aurions interrogé *Dico* 55 fois. Cette différence s'explique par l'exclusion immédiate de certains termes. Notre algorithme a une complexité similaire à une approche dichotomique car à chaque étape, l'espace de recherche est continuellement divisé, contrairement à l'approche systématique qui ne réduit pas cet espace. Cette approche nous donne des résultats précis mais avec un rappel, en théorie, limité. En effet, l'utilisation de l'égalité stricte de chaines de caractères ne nous permet pas de prendre en compte la grammaire, la conjugaison et la nature des mots dans les termes candidats.

1.2.4.2 Alignement utilisant des outils de TAL

Pour cette approche, nous utilisons un outil de TAL appelé FFROF (Find FROzen Forms) et développé par Paul Mycek (Mycek, 2011). Cet outil propose un alignement automatique entre les termes d'un document et un dictionnaire. Il applique le Tree Tagger[25] sur le document et sur le dictionnaire puis compare la version lexémisée d'un terme du document avec la version lexémisée d'une entrée du dictionnaire, ignorant ainsi les problèmes issus de la grammaire et de la conjugaison.

FFROF est paramétrable et peut filtrer les termes en fonction de leur nature grammaticale (adjectif, verbe, adverbe, etc.). Ainsi, nous pouvons ignorer les adverbes et les adjectifs pour l'analyse d'un document. Par exemple, pour le document « La mécanique expérimentale des fluides », FFROF vérifie l'existence de « le mécanique expérimental de le fluide ». Si le terme n'existe pas dans la version lexémisée du dictionnaire, alors il cherche le terme sans l'adjectif « expérimental ».

L'avantage de cette méthode est de s'affranchir de certains mots dans les termes. Relativement à l'algorithme précédent, nous pouvons nous attendre à un rappel supérieur.

Dans le cadre de la présentation de nos travaux, nous ne présenterons pas les résultats issus de l'alignement en utilisant l'outil FFROF. Nous sommes convaincus que cette approche améliorera l'alignement des termes avec les concepts. En revanche, nous souhaitons évaluer nos approches de manière brute pour émettre un jugement sur les aspects conceptuels de l'extraction des descripteurs. Nous ne souhaitons pas biaiser les évaluations dans la mesure où les autres approches avec lesquelles nous nous comparons ne peuvent pas bénéficier de ce prétraitement. Nous discuterons de son utilisation dans les perspectives de nos travaux (chapitre Perspectives).

[25] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

1.2.5 Exemples de graphe de terme

Dans un premier exemple, nous considérons le terme t = « réinjection de pertinence ». Le dictionnaire $Dico$, construit en amont, possède les correspondances suivantes :

$$match(\text{« réinjection de pertinence »}, Dico) = \text{« réinjection de pertinence »}$$

$$concepts(\text{« réinjection de pertinence »}, BC') = \text{« Retour de pertinence»}$$

La base BC' nous informe, quant à elle, que l'article « Retour de pertinence» est en relation avec les catégories supérieures « informatique théorique » « recherche d'information » et « traitement automatique du langage naturel » (Figure 14). Les articles Wikipédia sont représentés dans les graphes par un rectangle.

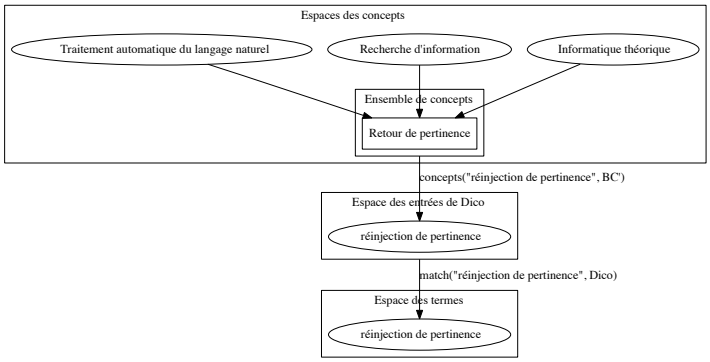


Figure 14 : extrait du graphe du terme « réinjection de pertinence »

Ensuite, l'opération $Graphe\big(\text{« } Retour\ de\ pertinence \text{ »}, BC'\big)$ est exécutée comme le montre la Figure 15.

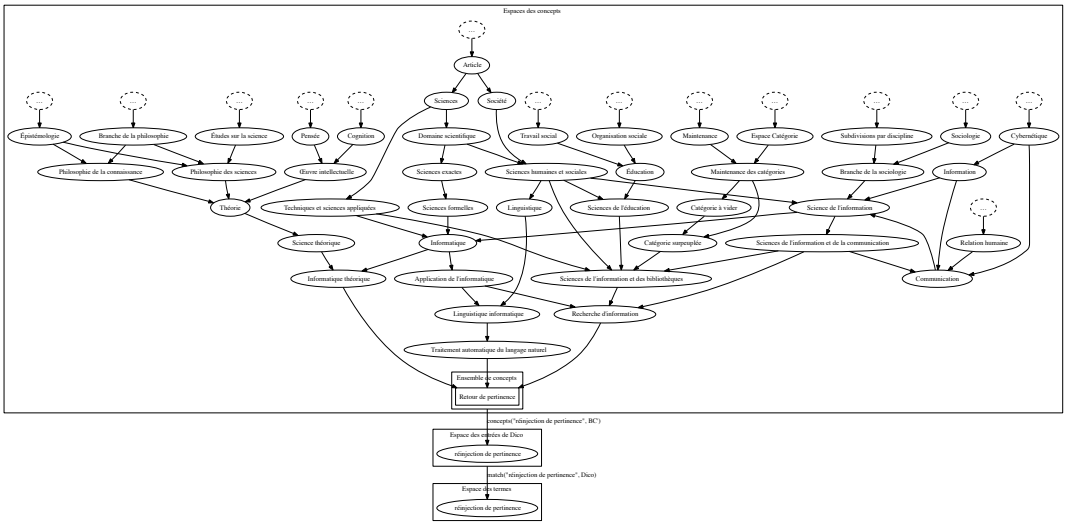


Figure 15 : graphe du terme « réinjection de pertinence » (5 niveaux de profondeur)

Pour le second exemple où le terme t= « bois », nous avons :

$$match(\text{« } bois \text{ »}, Dico) = \text{« } bois \text{ »}$$
$$concepts(\text{« bois »}, BC') = \{\text{« Bois »}, \text{« Bois_(matériau de construction) »},$$
$$\text{« Bois énergie »}, \text{« Forêt »}, \text{« Bois_(cervidé) »}, \text{« Bois_(musique) »}, \text{« Club_(golf)#Bois »}\}$$

Dans le cas où la fonction *concepts* renvoie plus d'un concept, le terme est ambigu. Cependant, dans notre contexte, nous n'avons aucune information pour décider du meilleur sens (concept). Nous construisons tout de même le graphe du terme t= « bois » sans désambiguïsation (avec tous les concepts candidats) comme le montre la Figure 16. La désambiguïsation sera effectuée ultérieurement, lorsque le système disposera de nouvelles informations pour trancher (Chapitre 2).

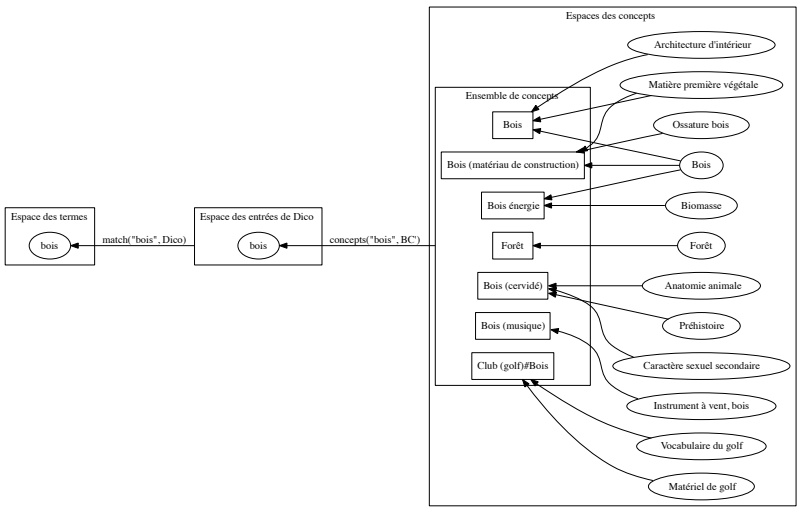


Figure 16 : Extrait du graphe du terme « bois »

1.3 Création d'un graphe représentant un document

1.3.1 Graphe d'un document

Soit D un document textuel composé d'une séquence de N mots $[m_{D,1}, m_{D,2}, \dots, m_{D,N}]$. La première étape consiste à transformer ces N mots en M termes (avec $M \leq N$ car un terme est formé d'un mot ou d'une séquence de mots) ayant un sens dans un certain contexte. Ainsi, nous introduisons la fonction $termes$, permettant d'extraire les termes d'un document D en s'appuyant sur un dictionnaire $Dico$:

$$termes(\mathrm{D}, Dico) = \left[t_{D,1}, t_{D,2}, \dots, t_{D,M}\right]$$

Nous avons proposé une méthode permettant d'extraire une liste de termes à partir d'un document (voir section 1.2.4.1). Afin d'obtenir le graphe de document $grapheDocument$ d'un document D nous procédons à une construction ascendante à partir des termes $t_{D,\mathrm{i}}$ et de leurs entrées

dans le dictionnaire. Un « graphe de document » est donc une fusion de graphes de termes à laquelle nous ajoutons une partie *Document* comme le suggère la Figure 17 et la formule suivante :

$$grapheDocument(D, Dico, BC') \supset grapheC(\{concepts(match(t_{D,i}, Dico))\}, BC', \text{GrapheVide})$$

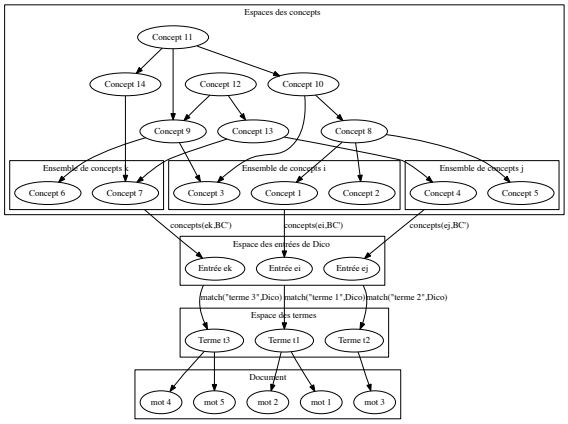


Figure 17 : représentation d'un document composé de 5 mots sous forme de Graphe

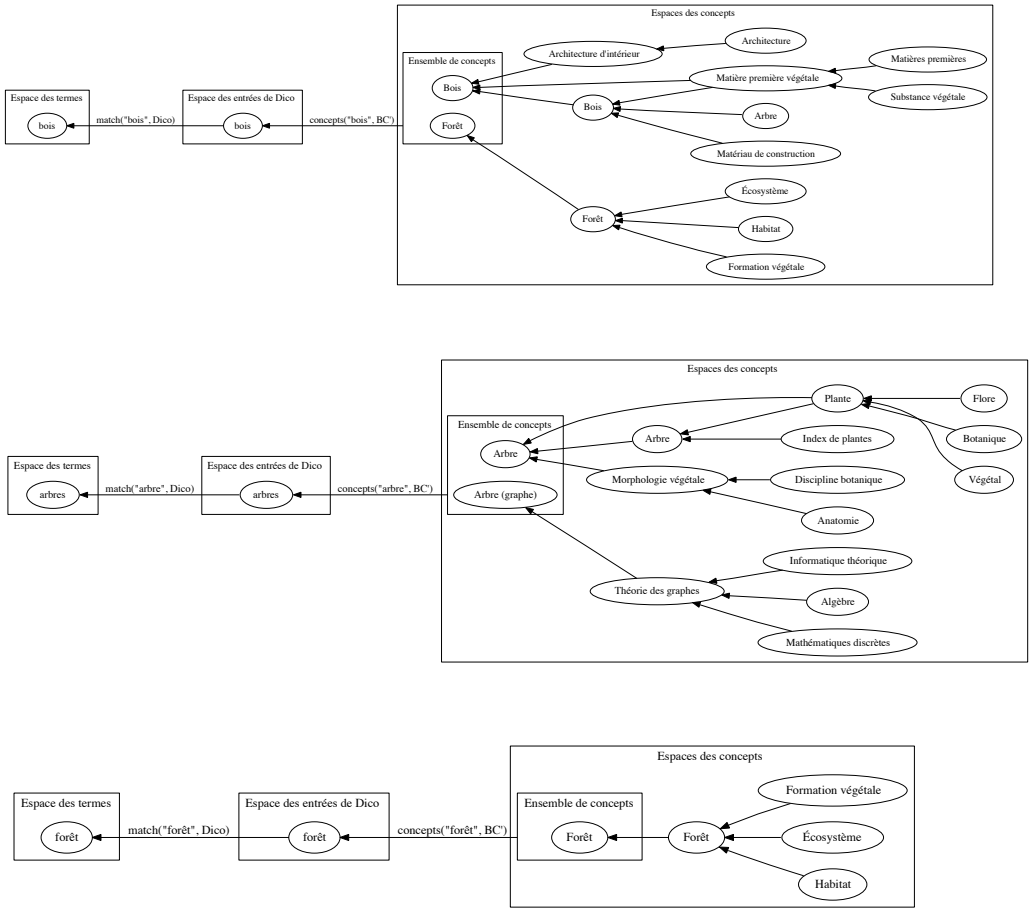


Figure 18 : Graphe des termes « bois », « Arbre » et « forêt »

Par exemple, soit le document D = « Le bois des arbres de la forêt ». Le dictionnaire $Dico$ nous donne alors 3 termes, après suppression des mots vides :

$$termes(D, Dico) = [\text{« bois »}, \text{« arbres »}, \text{« forêt»}]$$

La Figure 18 représente respectivement les premiers paliers des graphes des termes « bois », « arbres », « forêt ». Par construction ascendante, nous obtenons le graphe de la Figure 19.

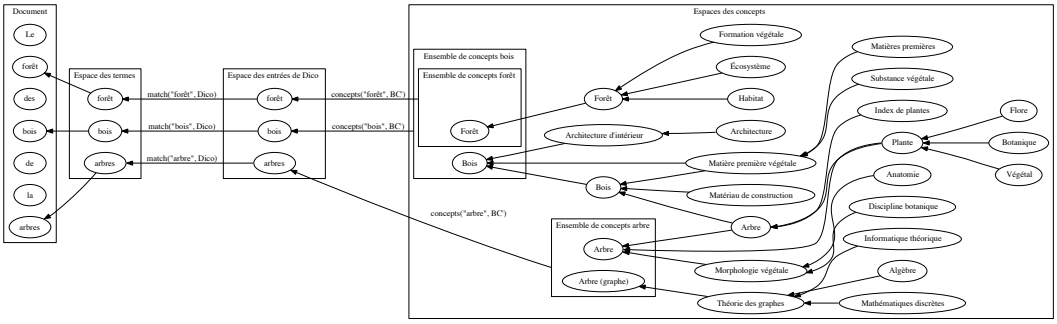


Figure 19 : Construction du graphe du texte « Le bois des arbres de la forêt » par méthode ascendante.

Ainsi pour chaque document, un graphe le représentant peut être construit. Ces graphes sont le socle de notre système d'aide à l'indexation et à la recherche d'information.

Un graphe de document est un graphe reliant les termes du document aux entrées d'un dictionnaire, les entrées aux concepts de la base de connaissances *BC'* et les concepts entre eux. Deux opérations sont définies sur les graphes de document : l'activation/désactivation des concepts et des termes, et l'extraction de sous-graphe.

1.3.2 Activation et désactivation

Désactiver un terme, un concept ou un lien entre un terme et un concept rend cet objet inaccessible. C'est une suppression réversible, dans la mesure où ces objets peuvent être réactivés. Elle consiste à désactiver définitivement les concepts et les termes qui a priori ne sont pas pertinents pour la recherche de descripteurs et à désactiver temporairement les termes et les concepts ambigus avant la phase de désambiguïsation.

L'activation/désactivation est valide sous certaines conditions :

1. l'activation ou la désactivation dans un graphe de document ne doit ni supprimer ni ajouter de racines au graphe. Ainsi, le graphe ne peut pas être vide ;
2. un graphe de document est nécessairement connexe et l'activation/désactivation doit préserver cette connexité ;
3. l'activation/désactivation doit préserver le fait que tous les concepts aient un chemin vers au moins un terme ;

4. l'activation/désactivation doit préserver le fait que tous les termes sont liés à au moins un concept.

Par exemple, dans le graphe de document de la Figure 20, nous souhaitons désactiver le « concept 12 ». Dans ce cas, le « concept 13 » devient une nouvelle racine du graphe. La règle 1 nous force à le désactiver, ce qui engendre, pour la même raison le « concept 4 ». Dans le même graphe, si nous désactivons le « concept 8 », les concepts « 1 », « 2 » et « 5 » deviennent des racines du graphe que nous désactivons à leur tour. Du coup, le « terme 2 » n'est lié à aucun concept, ce qui pose problème pour la règle 4. Le « terme 2 » est alors désactivé.

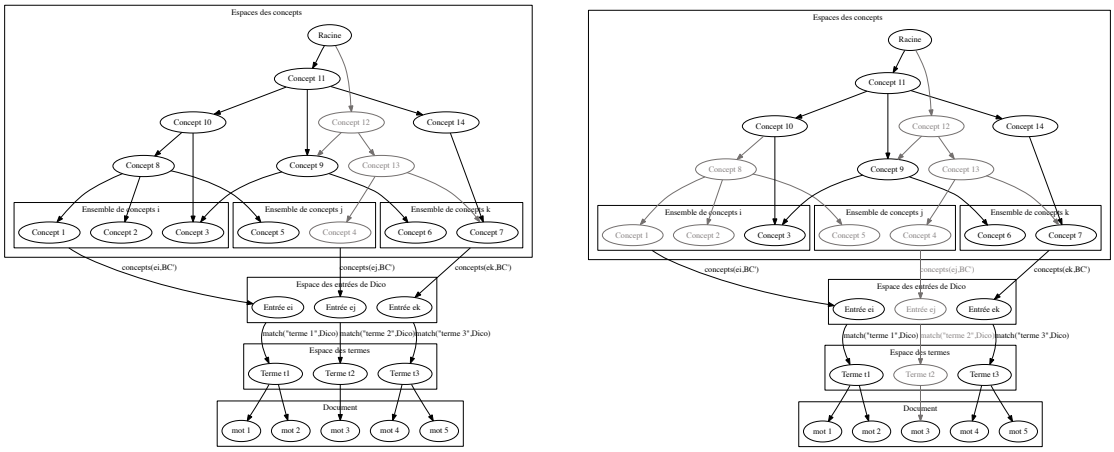


Figure 20 : désactivation des concepts 12 et 8

Cette opération d'activation/désactivation est indispensable pour l'élimination de concepts jugés a priori inutiles, mais également dans la phase de désambiguïsation présentée dans le Chapitre 2

1.3.3 Extraction de Sous-graphe

L'opération d'extraction de sous-graphe permet d'extraire le graphe d'une partie du document (une section, un chapitre, un document privé d'une section, etc.). Cette opération sert notamment pour le calcul de similarité pour la détection de rupture et de retour sémantique mais aussi pour la désambiguïsation.

Dans notre contexte, un sous-graphe de document est un graphe représentant une sélection de termes du document, construit par l'opération :

- sous-graphe(termes $\{t_1,\dots,t_n\}$, graphe de document GD) : sous-graphe composé des n termes.

Par exemple, la Figure 21 présente le sous-graphe du graphe de la Figure 20, composé des termes t2 et t3.

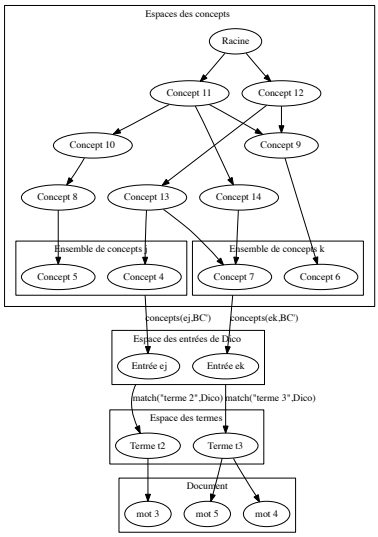


Figure 21 : Sous-graphe formé des termes t1 et t2

Cette opération permet également au système de recherche documentaire de ne retourner qu'une section du document susceptible d'intéresser l'utilisateur, et pas son intégralité.

1.4 Exemple de graphe de document créé à partir de Wikipédia

En utilisant, la base de connaissances hiérarchiques *BC'* (une partie du réseau de catégories Wikipédia), nous souhaitons construire le graphe de document. Les termes sont extraits du document grâce au dictionnaire *Dico*. L'algorithme de génération ascendante par palier est utilisé afin d'obtenir un graphe de document.

Un premier problème qui se pose est l'obtention d'une seule racine de ce graphe appelée « accueil » (page web d'accueil de Wikipédia). Cependant, lors de la construction du graphe, il est possible que certaines catégories ou articles soient placés dans d'autres catégories qui n'existent pas. Il faut simplement ne pas prendre en compte ces catégories pour n'obtenir qu'une seule racine.

Un deuxième problème est l'existence de cycles dans le réseau des catégories de Wikipédia qui complique le calcul des descripteurs. Nous proposons d'exclure les nœuds créant un cycle lors de la construction du graphe. Lorsqu'un nœud déjà dans le graphe est le nœud source d'une relation à ajouter, nous devons vérifier que celle-ci ne génère pas de cycle. Si c'est le cas, nous n'ajoutons pas la relation. Avec cette approche, nous supprimons le nœud le plus éloigné de l'ensemble des concepts de départ, puisque la construction s'effectue par palier ascendant.

Nous rappelons que le graphe construit à partir de Wikipédia répond aux critères suivants :

1. Le graphe n'a qu'une racine « Accueil » ;

2. Chaque terme est relié à au moins un concept ;
3. Chaque concept est relié à au moins un terme ;
4. Le graphe est connexe.

Ce type de graphe est qualifié de Graphe Orienté Acyclique (GOA) (en anglais, *Direct Acyclic Graph* (DAG)). A titre d'exemple, la Figure 22 montre le graphe du document « Albert Einstein a introduit la théorie de la relativité restreinte ». Ce graphe comporte 2 termes à l'origine du graphe (« Albert Einstein » et « théorie de la relativité restreinte »), 691 concepts, 1528 relations entre concepts. Ce GOA semble inexploitable pour le choix de concepts permettant une indexation. Cependant, le chapitre suivant décrit la méthode que nous proposons qui est fondée sur la notion de graphe de document.

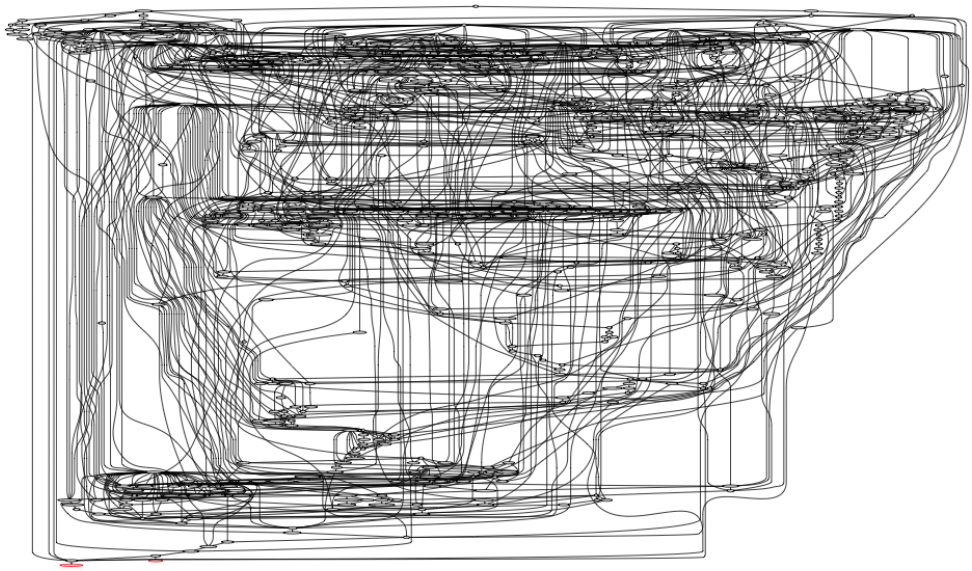


Figure 22 : graphe du texte « Albert Einstein a introduit la théorie de la relativité restreinte » (en rouge, les concepts « Albert Einstein » et « théorie de la relativité restreinte »)

1.5 Conclusion

Dans ce chapitre, nous avons présenté la construction d'un graphe de terme, d'un graphe de document, les premières opérations associées et un exemple utilisant nos approches sur le corpus de Wikipédia pour l'extraction des termes des documents et le réseau de catégories de Wikipédia pour la génération des graphes.

Ces graphes sont utilisés pour l'extraction des descripteurs du document. Il faut maintenant lister les conditions pour qu'un concept d'un graphe de document soit potentiellement un descripteur de ce document. A partir de ces concepts, que nous qualifions de forts ou d'importants, les mots-clés dans le texte et les thématiques (ou domaines) seront extraits.

Chapitre 2. Extraction des descripteurs d'un document à partir d'un graphe

2.1 Introduction

Ce chapitre a pour objectif d'étudier les graphes de document et notamment leur topologie afin de pouvoir mentionner les concepts pertinents pouvant être des descripteurs du document. Ces concepts fournissent des informations pour l'extraction des mots-clés dans le texte et des thématiques du document. Ils sont aussi exploités pour une première approche de la désambiguïsation des termes polysémiques d'un document.

Nous débutons ce chapitre par la mise en place du vocabulaire employé (section 2.2) ainsi que les prétraitements à appliquer aux graphes de document (section 2.3) avant de les analyser pour l'extraction des descripteurs. Ensuite, nous construisons un graphe de document, pour observer sa topologie et la position des concepts qui nous semblent importants et utilisables comme descripteurs du document (section 2.4). Nous présentons ensuite les trois propriétés que partagent les concepts qui nous serviront de descripteurs du document (section 2.5) : la fréquence terminologique, la généricité conceptuelle et la diversité conceptuelle. A partir de ces trois propriétés nous allons en déduire une formule permettant de coter les concepts importants du graphe (section 2.6). La cotation des concepts du graphe va ensuite nous permettre d'accomplir trois activités : L'extraction des mots-clés (section 2.7), l'extraction des thématiques (section 2.8) et la désambiguïsation des termes polysémiques (section 2.9). Finalement, nous allons évaluer ces approches en utilisant le corpus et le réseau de catégories de Wikipédia comme base de connaissances pour la cotation des graphes. C'est également les articles de Wikipédia que nous utilisons comme corpus pour tester nos approches.

2.2 Vocabulaire

Nous présentons le vocabulaire employé pour faire référence au graphe et au document. La Figure 23 nous sert d'exemple pour l'expliquer.

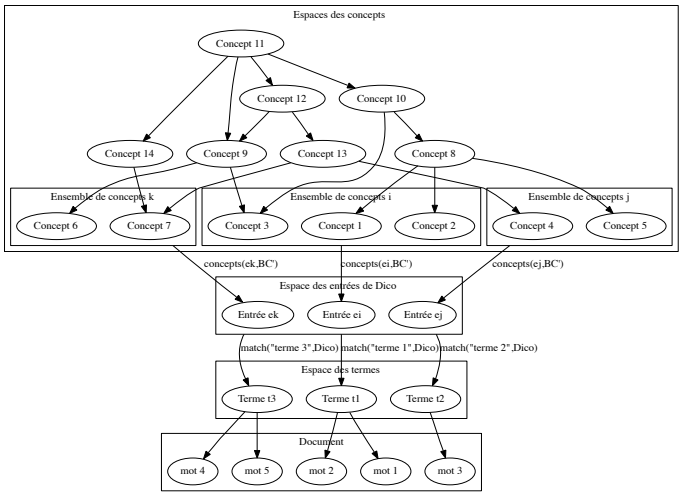


Figure 23 : Graphe de document

- ***Terme du graphe :*** c'est une représentation dans le graphe d'un terme du document. Dans la figure, les termes sont « Terme t1 », « Terme t2 », « Terme t3 ». Si dans le texte le même terme apparaît plusieurs fois, alors celui-ci sera présent plusieurs fois dans le graphe. Chaque terme du graphe est rattaché à un ou plusieurs concepts grâce au dictionnaire de correspondance ;
- ***Concept du graphe :*** c'est un nœud du graphe issu de la base de connaissances hiérarchique. Il représente une certaine idée ou notion d'une entité. Dans la Figure 23, les concepts sont « Concept 1 », … , « Concept 14 » ;
- ***Concept d'un terme :*** ce sont les concepts directement reliés à ce terme. Dans la figure, « Concept 6 » « Concept 7 » sont les concepts du terme « Terme t3 » ;
- ***Terme ambigu du graphe :*** un terme est ambigu si le nombre de concepts du terme est supérieur à 1. Un même terme peut faire référence à plusieurs concepts et dénoter plusieurs idées différentes. Dans la Figure 23, les trois termes du graphe sont ambigus ;
- ***Concept majoritaire d'un terme :*** lorsqu'un terme est ambigu, un de ses concepts peut avoir une fréquence d'utilisation (ou probabilité) supérieure à celle des autres concepts. Par exemple, le terme « France » dans Wikipédia a 19 sens différents mais le sens majoritaire est le pays. Cette information sur les sens majoritaires des termes est donnée par un acteur extérieur. En général, il s'agit d'une méthode statistique exploitant un corpus de termes annoté par des concepts, comme par exemple le corpus Wikipédia (nous rappelons que les liens internes de Wikipédia sont annotés par l'article qu'il renvoie). Pour un terme donné, le concept ayant la plus forte probabilité supérieure à x (fixé) est majoritaire pour ce terme ;
- ***Graphe idéal :*** un graphe est dit idéal lorsque chaque terme n'est relié qu'à un seul concept et que ce concept est le bon (après une phase de désambiguïsation).

- ***Racine conceptuelle :*** c'est un nœud du graphe qui n'a pas d'ancêtre. Dans la figure, le graphe n'a qu'une racine « Concept 11 » ;
- ***Feuille conceptuelle :*** c 'est un concept directement attaché à des termes. L'ensemble des feuilles conceptuelles est formé des premiers concepts générés lors de la construction ascendante par palier. Dans la figure, les feuilles sont « Concept 1 », … , « Concept 7 ». Toutes les feuilles conceptuelles ne sont pas des feuilles au sens de la théorie des graphes. Il est fréquent qu'une feuille conceptuelle ait des descendants comme le montre la Figure 24 ;

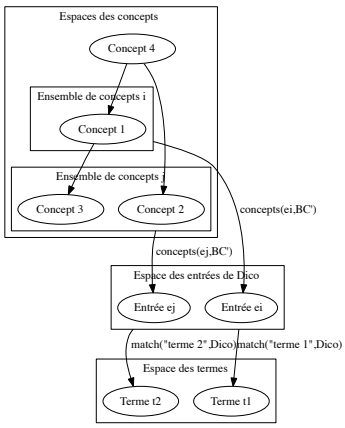


Figure 24 : la feuille conceptuelle « Concept 1 » à pour fils « Concept 3 »

- ***Descripteur :*** c'est une entité textuelle permettant de décrire un document. L'ensemble des descripteurs doit décrire le document de manière synthétique et exhaustive. Dans notre contexte, tous les descripteurs candidats se trouvent dans le graphe, c'est-à-dire que ce sont soit des concepts, soit des termes du graphe. Dans ce chapitre, nous proposons une méthode pour extraire ces descripteurs pour un document donné;
- ***Concept vide :*** c'est un concept du graphe dont nous savons, a priori, qu'il est inutile pour l'extraction des descripteurs ;
- ***Concept fort (important) du graphe :*** c'est un concept du graphe qui a toutes les conditions nécessaires pour être un descripteur potentiel du graphe et donc du document ;
- ***Mot-clé :*** une ambiguïté existe lorsque nous parlons de mot-clé. En effet, en général, un mot-clé est un descripteur du document qui se trouve ou non dans le document. Dans notre contexte, un mot-clé est un terme important du document, c'est-à-dire un descripteur dans le texte ;
- ***Thématique :*** c'est un concept du graphe qui est en général de plus haut niveau que les concepts forts. Il est plus générique mais aussi plus vague que les concepts forts ou les mots-clés. Par exemple, si « mécanique des fluides » est un concept fort, alors « physique » ou « science » peuvent être des thématiques.

2.3 Prétraitements sur un graphe de document

Avant l'étude d'un graphe, celui-ci doit subir un prétraitement :

1. Construction du graphe à partir du texte ;
2. Désactivation définitive des concepts vides ;
3. a - Soit désactivation des termes ambigus ;
   b - Soit désactivation des termes ambigus selon certaines conditions.

L'étape 2 permet de supprimer les concepts vides. Ce sont les concepts de Wikipédia liés à la catégorie « espace non-encyclopédique » (mentionnée au chapitre Wikipédia et la recherche d'information) qui informe sur l'article et non sur son contenu (meta-catégorie d'articles).

L'étape 3 consiste à n'avoir que des termes reliés à un seul concept. L'étape 3a) consiste à rendre les termes du graphe non ambigus. L'avantage de cette approche est bien sûr la non ambiguïté des termes du graphe. Son désavantage découle de la taille du graphe très limité en nombre de termes. En effet, le nombre de termes non ambigus est faible. Il se peut alors que le graphe non ambigu soit tellement réduit que l'extraction des concepts forts échoue, de par la trop grande différence du graphe avec celui d'origine.

Pour palier ce problème, dans l'étape 3b) nous ne conservons que les termes n'ayant qu'un seul concept ou le concept majoritaire de chaque terme. Lorsque nous prenons le risque d'utiliser le sens majoritaire, il est plus probable d'avoir un succès qu'un échec (avec x assez élevé) car le graphe obtenu est plus proche de celui d'origine (et probablement du idéal) que le graphe non ambigu. (Krovetz & Croft, 1992) préconise un seuil de x=80%.

Ainsi, tous les termes du graphe ne sont reliés qu'à un seul concept. Une fois l'étape de désactivation terminée, nous analysons le graphe et extrayons les concepts forts. A partir de ces concepts forts, nous désambiguïsons les termes du graphe. Cela signifie que nous pouvons revenir sur la décision du choix du concept pour certains termes et donc choisir un concept minoritaire.

2.4 Etude des graphes

Etudions le texte extrait de l'article de Wikipédia sur la turbulence[26] (phénomène physique). Le texte est le suivant :

La turbulence désigne l'état d'un fluide, liquide ou gaz, dans lequel la vitesse présente en tout point un caractère tourbillonnaire : tourbillons dont la taille, la localisation et l'orientation varient constamment. Les écoulements turbulents se caractérisent donc par une apparence très désordonnée, un comportement difficilement prévisible et l'existence de nombreuses échelles spatiales et

[26] http://fr.wikipedia.org/wiki/Turbulence

temporelles. De tels écoulements apparaissent lorsque la source d'énergie cinétique, qui met le fluide en mouvement, est relativement intense devant les forces de viscosité que le fluide oppose pour se déplacer. À l'inverse, on appelle laminaire le caractère d'un écoulement régulier. La découverte et l'étude des turbulences est très ancienne, elle a été par exemple faite par Léonard de Vinci.

Dans un premier temps, nous avons soumis ce texte à une ingénieure spécialiste en « énergétique et propulsion » et nous lui avons demandé : « Quelle est la liste de mots permettant de décrire exhaustivement et synthétiquement ce texte ? ». Sa liste est la suivante : turbulence, mécanique des fluides, écoulement, tourbillons/tourbillonnement, thermodynamique.

Dans un second temps, nous souhaitons extraire automatiquement des descripteurs pour ce texte. Pour cela, nous générons et prétraitons le graphe du document pour le texte ci-dessus (x=0.6 pour les concepts majoritaires, fixé empiriquement). Le résultat est présenté par la Figure 25. En rouge, les termes du document directement liés à leur concept. Ce graphe est constitué de 624 nœuds et 1160 relations.

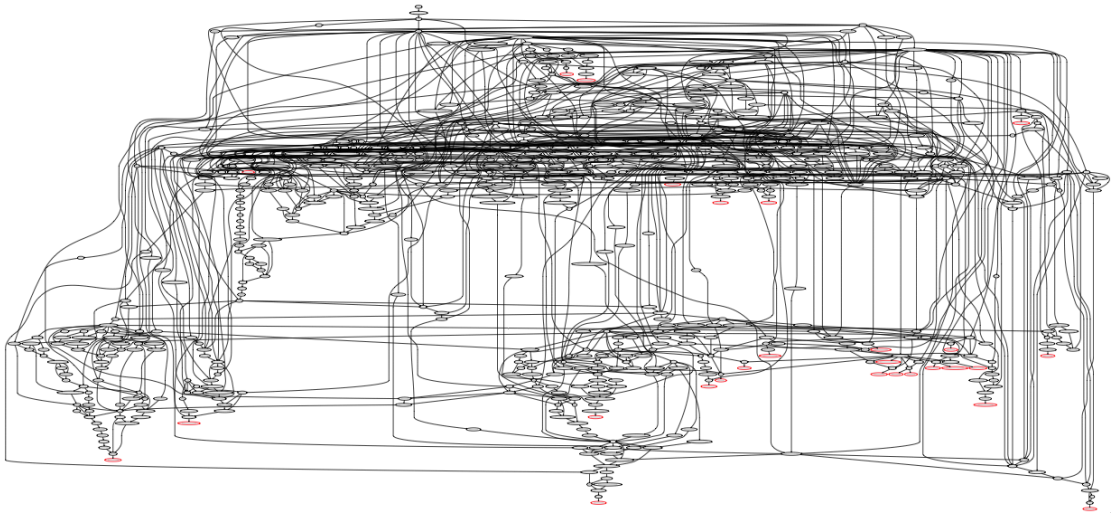


Figure 25 : Graphe du document traitant de turbulence. En rouge, les termes du document

En localisant dans le graphe les concepts liés aux termes de la liste de notre experte, considérés comme des concepts forts, nous observons une zone du graphe formée par 3 des 5 descripteurs (mécanique des fluides, turbulence, tourbillons). Dans la Figure 26, le terme « turbulence (1;3;13) » signifie que le terme est composé d'un mot qui commence à l'indice 3 et se termine à l'indice 13 dans le texte (nombre de caractères). Les concepts forts sont proches (en nombre d'arêtes à parcourir) de certains termes du texte et de feuilles conceptuelles (ou sont eux-mêmes des feuilles conceptuelles). Et c'est grâce à ces termes du texte que ces concepts forts sont présents dans le graphe. Les concepts forts sont donc générés dans le graphe grâce à plusieurs termes. Par exemple, le concept « Mécanique_des_fluides » a été généré par neuf termes (dont ceux visibles comme « fluide en mouvement », « viscosité », « turbulence », « écoulement turbulent » et deux fois le terme « fluide »),

« Turbulence » par deux termes (« turbulence » et « écoulements turbulents ») et « Tourbillon_(physique) » par un seul terme (tourbillons).

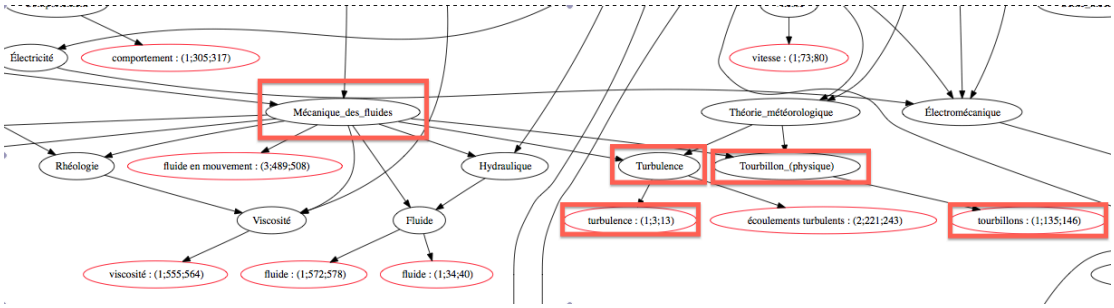


Figure 26 : Zoom sur la zone « turbulence », « mécanique des Fluides » et « tourbillons »

Nous supposons donc que les concepts importants ont été générés par un grand nombre de termes et sont proches des feuilles conceptuelles dans le graphe.

Lorsqu'un concept est ajouté au graphe lors de sa construction, celui-ci entraine aussi l'ajout de tous ses ancêtres. Du coup, il n'est pas nécessaire d'analyser tous les ancêtres pour l'analyse et la cotation des concepts, car nous souhaitons des concepts proches des feuilles conceptuelles,.

Pour un concept donné, nous ne pouvons pas savoir quels nœuds (concepts ou termes) sont responsables de sa présence dans le graphe. Cela signifie que lors de la cotation des concepts, nous devons analyser leurs descendants. Dans l'exemple, le concept « mécanique des fluides » a l'avantage d'avoir plusieurs concepts descendants directs (« turbulence », « fluide », « hydraulique », « viscosité », etc.).

Nous pouvons donc émettre les hypothèses suivantes :

1. Un concept fort est généré par plusieurs termes ;
2. Un concept fort est proche des feuilles conceptuelles et des termes ;
3. Un concept fort a plusieurs concepts fils.

2.5 Propriétés d'un concept dans un graphe

Nous identifions les propriétés d'un concept fort afin de proposer une formule de cotation des concepts permettant d'extraire les concepts forts.

2.5.1 Occurrence et fréquence d'un concept

L'occurrence terminologique d'un concept est le nombre de termes responsables de la présence de ce concept dans le graphe. C'est aussi le nombre de graphes de terme possédant ce concept. Par exemple, dans la Figure 26, le concept « turbulence » a une occurrence terminologique égale à 2 (2

termes turbulence et écoulements turbulents). A priori, plus un concept est proche de la racine, plus ce concept a un score élevé. A contrario, un concept proche des feuilles conceptuelles a un score faible. Nous notons l'occurrence terminologique avec la fonction :

$$occ(concept, graphe)$$

Pour le graphe de la Figure 25 :

- $occ($« Tourbillon_(physique) »$, graphe) = 1$
- $occ($« Turbulence »$, graphe) = 2$
- $occ($« Mecanique_des_fluides »$, graphe) = 9$
- $occ\big($« Mécanique »$, graphe\big) = 11$
- $occ\big($« Ingénierie »$, graphe\big) = 22$
- $occ\big($« Sciences »$, graphe\big) = 27$
- $occ\big($« Accueil »$, graphe\big) = 27$

Nous introduisons également la notion de fréquence terminologique ou occurrence normalisée :

$$freq(concept, graphe) = \frac{occ(concept, graphe)}{Max_i(occ\big(concept_{i \in concepts}, graphe\big))}$$

Par exemple :

- $freq($« Tourbillon_(physique) »$, graphe) = 1/27$
- $freq($« Mecanique_des_fluides »$, graphe) = 9/27 = 1/3$
- $freq\big($« Accueil »$, graphe\big) = 27/27 = 1$

Ces fonctions ne suffisent pas pour déterminer les concepts forts du graphe car les concepts de haut niveau auront un score trop fort (notamment la racine du graphe).

2.5.2 Profondeur et généricité d'un concept

La notion de profondeur d'un nœud d'un graphe n'est définie dans la théorie des graphes que pour un arbre. C'est la longueur du chemin unique allant de la racine au nœud. Dans un GOA, il existe plusieurs chemins menant des racines aux concepts et ces chemins sont de longueur différente. Par exemple, dans Wikipédia, entre « Léonard_de_Vinci » et « artiste » il existe 4 chemins de longueur allant de 3 à 5 (en nombre d'arcs) ; entre « humain » et « accueil », les longueurs des chemins vont de 5 à 30.

Dans ce cas de figure, nous pourrions considérer que le chemin le plus court est le meilleur. Or cette approche n'est pas intéressante car aucune information n'est donnée sur le poids des relations

entre deux catégories. Nous choisissons donc de ne pas considérer les notions de distance, de profondeur et de hauteur issue de la théorie des graphes.

Nous cherchons alors à mesurer la proximité d'un concept à la racine du graphe, ce que nous appelons la généricité d'un concept. La généricité est aussi liée au nombre de descendants d'un concept :

$$gen(concept, graphe)$$

Dans le graphe de la Figure 25 :

- $gen($« Tourbillon_(physique) »$, graphe) = 1$
- $gen($« Turbulence »$, graphe) = 1$
- $gen($« Mecanique_des_fluides »$, graphe) = 9$
- $gen\big($«Mécanique »$, graphe\big) = 23$
- $gen\big($« Ingénierie »$, graphe\big) = 310$
- $gen\big($« Sciences »$, graphe\big) = 524$
- $gen\big($« *Accueil* »$, graphe\big) = 624$

Comme pour l'occurrence terminologique, nous définissons la généricité normalisée :

$$genNorm(concept, graphe) = \frac{gen(concept, graphe)}{Max_i(gen\big(concept_{i \in concepts}, graphe\big))}$$

- $genNorm($« Tourbillon_(physique) »$, graphe) = 1/624$
- $genNorm($« Mecanique_des_fluides »$, graphe) = 9/624$
- $genNorm\big($« Accueil »$, graphe\big) = 624/624 = 1$

2.5.3 Corrélation entre occurrence et généricité

Les concepts de haut niveaux ont un score d'occurrence et de généricité élevé. Ces deux mesures sont corrélées. Nous mettons en évidence cette corrélation dans le graphique de la Figure 27. Nous avons analysé un ensemble de graphes de document et nous avons, pour une occurrence donnée, calculer la moyenne des généricités. De même, cette corrélation est linéaire avec un coefficient proche de 0.98.

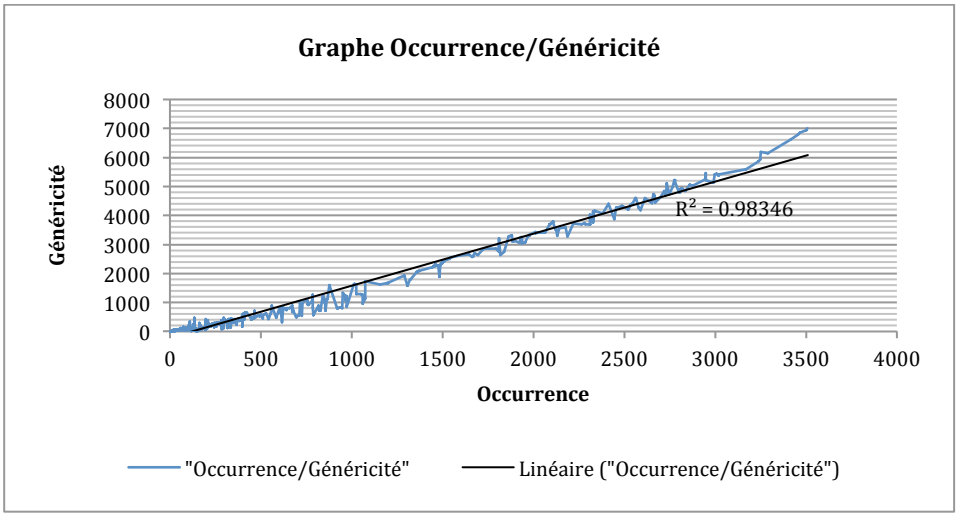


Figure 27 : Corrélation entre occurrence et généricité

Nous interprétons l'occurrence d'un concept comme l'expansion horizontale de ce concept dans le graphe et la généricité comme l'expansion verticale. La Figure 28 suggère une interprétation géométrique de ces mesures. Un concept et ses descendants forment un rectangle de hauteur représentant la généricité du concept et de largeur représentant l'occurrence.

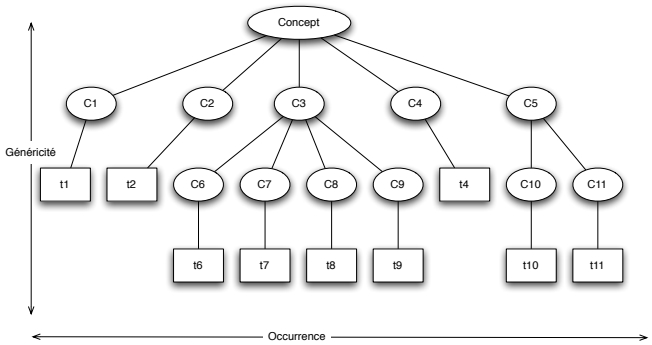


Figure 28 : Interprétation géométrique de la généricité et de l'occurrence

Un concept fort possède à la fois une forte occurrence et une faible généricité. Nous recherchons les concepts ayant un rapport occurrence/généricité le plus haut possible (faible hauteur et grande largeur).

La probabilité de présence d'un concept dans un graphe est corrélée à sa généricité et à son occurrence. Par exemple la généricité normalisée et l'occurrence normalisée de « accueil » sont toutes deux égales à 1.

Nous listons les cas possibles dans le Tableau 8.

	Occurrence Forte	Occurrence Faible
Généricité Forte	• Forte probabilité de présence • Concept trop générique donc faible	• Faible probabilité de présence • Nœud de trop haut niveau donc concept faible
Généricité Faible	• Forte probabilité de présence • Concept potentiellement fort	• Faible probabilité de présence • Concept Faible

Tableau 8 : Force d'un concept en fonction de sa généricité et de son occurrence

Lorsque l'occurrence d'un concept est élevée, nous avons un échantillon de texte suffisant pour permettre le calcul de la probabilité de présence de ce concept dans le graphe en fonction de sa généricité. A occurrence fixe, plus un concept est générique, plus il est probable qu'il soit présent dans le graphe. D'où le rapport suivant[27] :

$$P(Concept \in Graphe) \propto \frac{gen(Concept, Graphe)}{occ(Concept, Graphe)}$$

D'après (Resnik, 1995, 1999), la *quantité d'information* (en anglais : *Information Content – IC*) est fonction de la probabilité de présence d'un concept dans un corpus :

$$IC(Concept) = -log\ (P(Concept \in Corpus))$$

Par analogie, nous avons la formule suivante :

$$IC(Concept) = -log\left(\frac{gen(Concept, Graphe)}{occ(Concept, Graphe)}\right) = log\left(\frac{occ(Concept, Graphe)}{gen(Concept, Graphe)}\right)$$

$$ICNorm(Concept) = -log\left(\frac{genNorm(Concept, Graphe)}{freq(Concept, Graphe)}\right) = log\left(\frac{freq(Concept, Graphe)}{genNorm(Concept, Graphe)}\right)$$

Le Tableau 9 donne la quantité d'information normalisée de chaque concept du graphe de la Figure 25 (page 112):

[27] le symbole ∝ signifie « proportionnel à »

Concept	occ	gen	ICNorm
Turbulence	2	1	1.664850821
Fluide	2	1	1.664850821
Vitesse	1	1	1.363820826
Gaz	1	1	1.363820826
Faîtage	1	1	1.363820826
Temps_(grammaire)	1	1	1.363820826
MET	1	1	1.363820826
Comportement	1	1	1.363820826
Viscosité	1	1	1.363820826
Source_d'énergie	1	1	1.363820826
Désignation_de_Bayer	1	1	1.363820826
Léonard_de_Vinci	1	1	1.363820826
Exemple_(mathématiques)	1	1	1.363820826
Appelle	1	1	1.363820826
Laminaria	1	1	1.363820826
Hydraulique	2	2	1.363820826
Polytope_régulier	1	1	1.363820826
Échelles	1	1	1.363820826
Tourbillon_(physique)	1	1	1.363820826
Volant_d'inertie	1	1	1.363820826
Espace_(cosmologie)	1	1	1.363820826
Liquide	1	1	1.363820826
Opposé_(mathématiques)	1	1	1.363820826
État	1	1	1.363820826
Liste_des_présentateurs_des_Oscars	1	1	1.363820826
Mécanique_des_fluides	9	9	1.363820826
Théorie_météorologique	3	3	1.363820826
Mécanique_des_milieux_continus	9	10	1.318063335
Mécanique_classique	9	11	1.27667065
Météorologie	3	4	1.238882089
Phase	2	3	1.187729566
Théorie_scientifique	10	15	1.187729566
Énoncé_scientifique	10	16	1.159700843
Thermodynamique	2	4	1.06279083
Vocabulaire_des_mathématiques	1	2	1.06279083
.....	.....	.....	.....
Classification_chimique	3	60	0.06279083
Techniques_et_sciences_appliquées	25	507	0.056752875
Histoire	6	123	0.052066964
Sciences	27	560	0.046996563
Spécialité_paraclinique	1	21	0.041601531
Alimentation_humaine	1	21	0.041601531
Spécialité_médicale	1	21	0.041601531
Système_digestif	1	21	0.041601531
Société	27	582	0.030261605
Anatomie_animale	1	22	0.021398145
Système_digestif_et_excréteur	1	22	0.021398145
Anatomie	1	23	0.00209299
Physiologie_animale_et_humaine	1	23	0.00209299
Discipline_zoologique	1	23	0.00209299
Alimentation	1	23	0.00209299
Article	27	622	0.001394205
Espace_encyclopédique	27	623	0.000696543
Accueil	27	624	0

Tableau 9 : Score des concepts en fonction de leur quantité d'information

Nous adaptons cette mesure pour obtenir un score positif :

$$IC(Concept) = log\left(\frac{occ(Concept, Graphe)}{gen(Concept, Graphe)} + 1\right)$$

Nous remarquons deux choses concernant cette métrique :

- elle défavorise les concepts de haut niveau (« accueil » a une quantité d'information nulle) ;
- elle favorise les concepts issus des termes qui se répètent dans le texte et défavorise les concepts intermédiaires intéressants (comme « mécanique_des_fluides », par exemple).

2.5.4 Diversité conceptuelle d'un concept

Nous avons déjà mentionné que les concepts importants du graphe possèdent plusieurs descendants directs. Mais du fait que les graphes de document sont orientés acycliques, il peut y avoir plusieurs cas de figure, illustrés par la Figure 29.

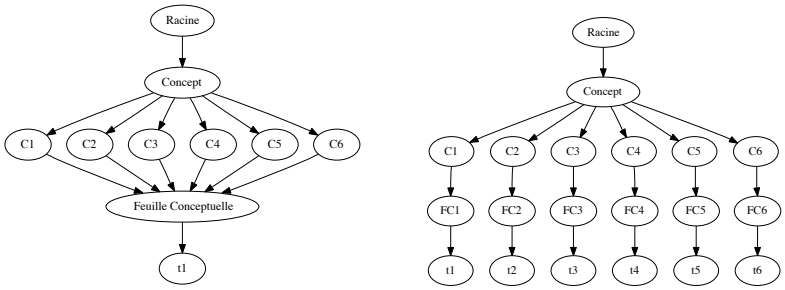


Figure 29 : Illustration de deux concepts ayant le même nombre de fils.

Nous constatons qu'un concept peut avoir plusieurs fils directs correspondant à une seule ou plusieurs feuilles conceptuelles. Dans le premier graphe, un seul terme a contribué à la génération du concept, alors que dans le second graphe, ce sont 6 termes. Ce dernier cas est intéressant car le concept a été généré par des termes différents. Nous appelons diversité conceptuelle d'un concept le nombre de feuilles conceptuelles responsables de l'existence des fils de ce concept.

Dans la Figure 29, la feuille conceptuelle « FC1 » est la seule responsable de la présence de « C1 » et de la relation entre « C1 » et « Concept ». Il en est de même pour toutes les feuilles conceptuelles du graphe. La diversité de « Concept » est égale à 6. En revanche, pour le concept « Racine », aucune feuille conceptuelle n'est l'unique responsable de sa présence dans le graphe. Sa diversité vaut 0.

Le calcul de la diversité utilise une notion étendue du *plus petit ancêtre commun* (en anglais, *least common subsumer* ou *ancester* LCS ou LCA).

Pour le calcul de la diversité d'un concept, nous avons besoin de connaître la contribution de chaque feuille conceptuelle dans le graphe. Il s'agit de comparer le graphe induit par une feuille conceptuelle et le reste du graphe. La Figure 30 montre un exemple de graphe composé de 3 feuilles conceptuelles ainsi que sa décomposition en un graphe composé d'une feuille conceptuelle {« FC3 »} et le reste du graphe {« FC1 », « FC2 »}.

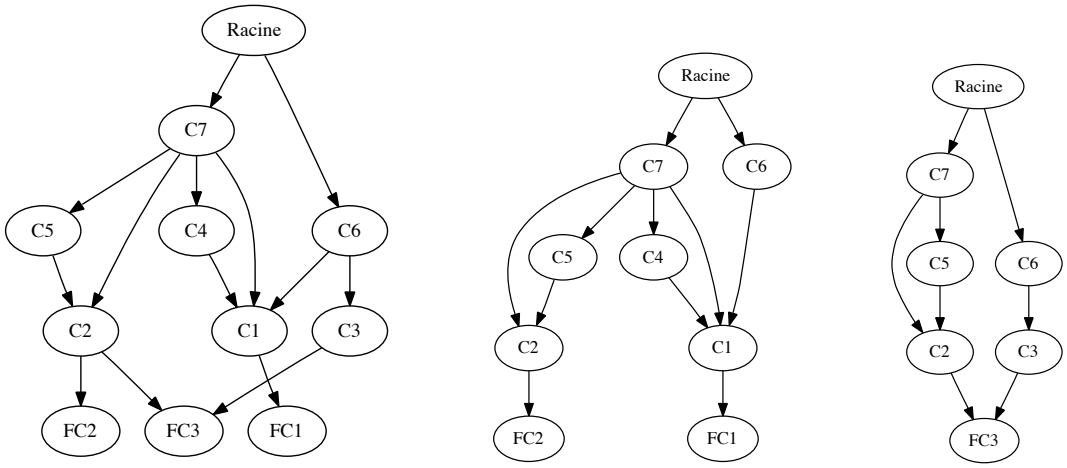


Figure 30 : Décomposition d'un graphe {« FC1 », « FC2 », « FC3 »} en deux sous-graphes { « FC1 », « FC2 »} et { « FC3 »}

Le concept « FC3 » contribue fortement à la création d'un fils direct de C2 (« FC3 » lui-même) et il est aussi responsable de la relation entre « C6 » et « C3 ». « FC3 » contribue donc à diversifier les descendants des concepts « C2 » et « C6 ».

Nous introduisons une nouvelle opération appelée Plus Petit Ancêtre Commun Généralisé (en anglais : *Generalized Least Common Ancestor* - GLCA), définie de la manière suivante avec C1≠C2 :

$GLCA(C1, C2) =$ Nœuds subsumant C1 et C2 possédant des concepts fils non présents dans C1

Nous appelons cette métrique GLCA car $LCA(C1, C2) \subset GLCA(C1, C2)$.

Par exemple,

$$LCA(\{\text{«}FC1\text{»}, \text{«}FC2\text{»}\}, \{\text{«}FC3\text{»}\}) = \{\text{«C2»}\}$$
$$GLCA(\{\text{«}FC1\text{»}, \text{«}FC2\text{»}\}, \{\text{«}FC3\text{»}\}) = \{\text{«C2»}, \text{«C6»}\}$$

Nous pouvons expliquer le GLCA de deux concepts ou de deux ensembles de concepts par la dissociation physique d'un graphe en deux graphes induits. Supposons que les concepts du graphe {« FC1 », « FC2 », « FC3 »} soient fixés dans l'espace et que les relations soient élastiques. Par exemple, si nous tirons sur le concept « FC3 », les élastiques craqueront aux concepts « C2 » et « C6 », qui sont les GLCA de ({« FC1 », « FC2 »},{ « FC3 »}), comme le suggère la Figure 31.

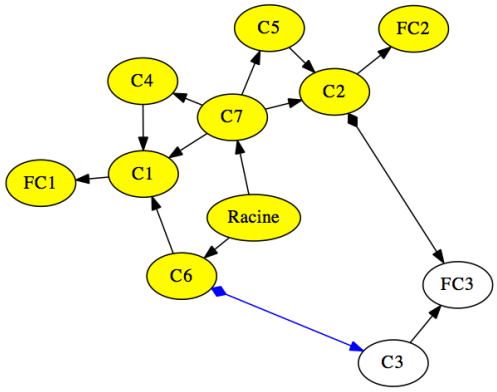


Figure 31 : Analogie physique pour l'extraction du GLCA de ({« FC1 », « FC2 »}, { « FC3 »})

2.5.5 *Calcul de la diversité conceptuelle*

Ainsi, pour la Figure 30, nous avons :

$$GLCA(\{\text{«}FC2\text{»}, \text{«}FC3\text{»}\}, \{\text{«}FC1\text{»}\}) = \{\text{«C6»}, \text{«C7»}\}$$
$$GLCA(\{\text{«}FC1\text{»}, \text{«}FC3\text{»}\}, \{\text{«}FC2\text{»}\}) = \{\text{«C2»}\}$$
$$GLCA(\{\text{«}FC1\text{»}, \text{«}FC2\text{»}\}, \{\text{«}FC3\text{»}\}) = \{\text{«C2»}, \text{«C6»}\}$$

Nous définissons la diversité conceptuelle d'un concept C comme suit :

$$div(C) = \#(\text{GLCA}_{FC_i}(\overline{\text{FC}_\iota}, \text{FC}_i) \supset \text{C})$$

La diversité conceptuelle d'un concept consiste à compter le nombre de fois où le concept est le GLCA des feuilles conceptuelles (FC_i) avec le reste du graphe ($\overline{\text{FC}_\iota}$). Dans l'exemple précédent, les concepts « C2 » et « C6 » ont une diversité égale à 2, « C7 » une diversité égale à 1 et les autres concepts une diversité nulle.

Nous remarquons que la diversité est inférieure ou égale au nombre de fils d'un concept. Cette mesure respecte donc bien l'hypothèse qu'un concept est important s'il a plusieurs fils issus de feuilles conceptuelles différentes.

Pour l'article de Wikipédia sur la turbulence, nous obtenons les scores de diversité suivants :

Concept	div
Mécanique_des_fluides	6
Techniques_et_sciences_appliquées	4
Linguistique	4
Histoire_thématique	4
Société	3
Culture	3
Ingénierie	3
Muséologie	3
Art	3
Discipline_de_la_biologie	2
Histoire	2
Sciences_humaines_et_sociales	2
Physiologie	2
Domaine_interdisciplinaire	2
Philosophie_politique	2
Science_auxiliaire_de_l'histoire	2
État	2
Voyage	2
Vivant	2
Continent	2
Anthropologie	2
Humain	2
Divertissement	2
Courant_philosophique	2
Aménagement_du_territoire	2
Analyse_artistique	2
Vie_quotidienne	2
Mathématiques	2
Anthropologie_sociale_et_culturelle	2
Branche_des_mathématiques	2
Énergie	2
Grandeur_physique	2
Mécanique	2
Phase	2
Théorie_météorologique	2
Classification	1
Sciences	1
Sociologie	1
Branche_de_la_sociologie	1
Fonctionnement_de_l'entreprise	1
Management	1
Terminologie_technique	1
Lexique	1
Lexicologie	1
Humanités	1
Géographie	1
Europe	1
Pensée	1
Vente	1
Technologie	1
Maritime	1
Champ_connexe_à_la_psychologie	1
Marketing	1
Outil_du_management	1
Océanographie	1
Philosophie	1
Politique	1
Branche_du_droit	1
Branche_de_la_philosophie	1
Thème_étudié_en_psychologie	1
Relation_humaine	1
Communication	1
Association_ou_organisme_lié_à_l'économie	1
Science_de_l'information	1
Géographie_de_l'Europe	1
Informatique	1
Finance	1
Association_ou_organisme_politique	1
Composé_chimique	1
Santé	1
Droit_public	1
Spiritualité	1
Philosophie_des_sciences	1
Philosophie_de_la_connaissance	1
Droit_administratif_général	1
Commerce	1
Pays	1
Membre_de_l'OMC	1
France	1
Culture_populaire	1
Astrophysique	1
État_membre_de_l'ONU	1
Objet_céleste	1
Transport	1
Service	1
Astronomie	1
Loisir	1
Physique_appliquée_et_interdisciplinaire	1
Science_politique	1
Territoire	1
Géographie_politique	1
Politique_de_la_France	1
Chimie	1
Secteur_d'activité	1
Secteur_industriel	1
Œuvre_créative	1
Littérature	1
Physique_classique	1
Thermodynamique	1
Théorie_scientifique	1
Mécanique_des_milieux_continus	1
Turbulence	0
Fluide	0
Vitesse	0
Gaz	0
Faîtage	0
Temps_(grammaire)	0

Tableau 10 : Diversité conceptuelle des concepts issus de l'article Wikipédia sur la turbulence

Nous constatons que de nombreux mauvais concepts ont une forte diversité (« histoire thématique » ou « linguistique »). Or ces nœuds ont une quantité d'information très faible. La cotation des concepts doit alors prendre en compte la diversité et la quantité d'information.

2.6 Cotation des concepts

Un concept est important si sa quantité d'information et sa diversité sont élevées. Ainsi, La cotation conceptuelle d'un concept est calculée en fonction de sa quantité d'information et de sa diversité conceptuelle :

$$Score(Concept) = IC(Concept) \times (div(Concept) + 1)$$

Cette formule donne un score $\in]0; +\infty[$ élevé lorsque le concept a une diversité et une quantité d'information élevées. Le « + 1 » ajouté à la diversité conceptuelle permet de classer les concepts qui ont une diversité nulle. Nous introduisons également le score normalisé d'un concept :

$$ScoreNorm(Concept) = \frac{Score(Concept)}{Max_{C_i}(Score(C_i))}$$

Pour l'article de Wikipédia sur la « turbulence », nous obtenons ainsi les dix meilleurs concepts en fonction de leur score normalisé :

Concept	occ	gen	div	ScoreNorm
Mécanique_des_fluides	9	9	6	1
Théorie_météorologique	3	3	2	0.428571429
Phase	2	3	2	0.315842397
Mécanique_des_milieux_continus	9	10	1	0.264571262
Mécanique	11	23	2	0.241671808
Turbulence	2	1	0	0.226423214
Fluide	2	1	0	0.226423214
Théorie_scientifique	10	15	1	0.210561598
Grandeur_physique	6	18	2	0.177873214
Thermodynamique	2	4	1	0.167132143

Tableau 11 : Cotation conceptuelle des concepts du graphe sur la "turbulence"

2.7 Extraction des mots-clés du texte

Pour extraire les mots-clés du texte, nous choisissons les termes associés aux N meilleurs concepts forts. Pour chaque terme, on somme les scores de leurs concepts :

$$ScoreMotClé(Terme, N) = Occ(Terme) \times \sum_{(C \in TopN) \cap (Terme \in C)} Score(C)$$

Par exemple, le terme « viscosité » apparaît une fois dans le document et est lié aux concepts forts « mécanique_des_fluides » « mécanique_des_milieux_continus », « mécanique » et « grandeur_physique » :

$$Score(\text{« } Viscosité \text{ »}, 10) = 1 \times (1 + 0.26 + 0.24 + 0.18) = 1.68$$

Voici les scores des termes correspondant aux cinq meilleurs concepts :

1. fluide
2. forces
3. gaz
4. liquide
5. viscosité
6. tourbillons
7. turbulence
8. écoulements turbulents
9. fluide en mouvement
10. vitesse
11. énergie cinétique
12. existence

Cette approche permet de choisir et de noter les termes du document en fonction des meilleurs concepts forts. Elle réalise une indexation conceptuelle mais aussi contextuelle : les concepts forts fixent le contexte du document et contribue à l'extraction des mots-clés contextualisés. Nous ne prenons en compte que les N premiers concepts car au-dessus d'un certain seuil, les concepts ne sont plus pertinents pour l'indexation. Le nombre N est fixé en fonction de la taille du texte (nombre de mots).

Nous proposons une seconde approche permettant d'affecter un score à tous les termes du document. Pour chaque terme, nous choisissons les N meilleurs concepts (pour garder les concepts pertinents) liés à ce terme et nous sommons les scores de ces concepts :

$$ScoreMotClé2(Terme, N) = Occ(Terme) \times \sum_{C \in \text{Les N meilleurs concepts contenant Terme}} Score(C)$$

L'intérêt de cette méthode est d'affecter un score à l'ensemble des termes du graphe, contrairement à l'autre métrique (ScoreMotClé) qui attribue un score nul à des concepts n'appartenant pas aux N meilleurs concepts.

2.8 Extraction des thématiques du document

De façon analogue à l'extraction des mots-clés, nous utilisons les scores et les rangs des concepts pour extraire les thématiques. La différence majeure est que, pour les mots-clés, nous connaissons la liste de termes candidats (les termes du document). Pour les thématiques, il faut aussi connaître la liste des concepts candidats. Ce travail se fait en général manuellement. Mais Wikipédia fournit la notion de Portail qui est une page agrégeant des articles traitant d'une même thématique. Actuellement, Wikipédia comporte environ 1000 thématiques. Nous utilisons cette liste de thématiques comme liste des concepts candidats. Ainsi, le score de la thématique se calcule de la même manière que les mots-clés :

$$ScoreThématique(T, N) = \sum_{(C \in Top\ N)} \begin{cases} 1 & si\ T = ArgMin_{T' \in \{T_i\}}\big(gen(T')\big) \\ 0 & sinon \end{cases}$$

Pour chacun des N meilleurs concepts, nous extrayons l'ensemble des thématiques $\{T_i\}$ ancêtres de C. Parmi ces thématiques $\{T_i\}$, nous incrémentons de 1 le score de la thématique qui a la généricité la plus faible (la plus précise).

Par exemple, pour N=5, les thématiques suivantes sont sélectionnées :

1 Chimie=2 (« mécanique_des_fluides » et « mécanique_des_milieux_continus » ont comme thématique ancêtre la moins générique « Chimie ») ;
2 Physique=1 (« mécanique ») ;
3 Météorologie=1 (« théorie_météorologique ») ;
4 Énergie=1 (« phase »).

Selon notre calcul, « Chimie » est la meilleure thématique. Par contre, d'après l'ingénieur, parmi ces thématiques sélectionnées, « Chimie » est la plus mauvaise, les trois autres sont correctes de façon équivalente. Cela signifie que dans la hiérarchie des catégories Wikipédia « mécanique_des_milieux_continus » est relié à Chimie, ce qui est une aberration d'après notre ingénieur. Quant à l'appartenance de « mécanique_des_fluides » à la catégorie « Chimie » celle-ci est plus acceptable. Pour résoudre ce problème, nous pouvons suggérer les deux ou trois concepts les plus génériques parmi les thématiques pour chacun des N meilleurs concepts.

2.9 Première approche pour la désambiguïsation

Dans cette section, nous proposons une approche pour la désambiguïsation exploitant la cotation conceptuelle. Dans le chapitre suivant, nous proposons une deuxième approche exploitant la similarité entre graphes de document et de concepts (voir section 3.4).

Pour l'instant, dans notre approche, chaque terme n'est rattaché qu'à un seul concept. Ce concept est soit le seul possible pour le terme, soit le meilleur avec une probabilité égale à X. Pour la désambiguïsation, nous réactivons l'ensemble des termes et leurs concepts liés. Nous proposons la formule suivante :

$$Désambiguisation(Terme, N) = ArgMax_{(C' rattaché\ à\ Terme) \cap (C'\ fils\ direct\ de\ C) \cap (C \in Top\ N)} \big(Score(C)\big)$$

Pour un terme donné, nous choisissons son concept de rattachement comme le concept fils direct d'un des N concepts forts pris dans l'ordre décroissant des scores. Il existe alors trois possibilités :

1. il n'y a qu'un seul concept fils direct du concept le plus fort. Nous désactivons ainsi les autres concepts du terme ;
2. il y a plusieurs concepts fils directs du concept le plus fort. Nous gardons celui qui a le sens majoritaire ;

3. aucun concept n'est fils direct des N meilleurs concepts. Nous gardons le concept qui a le sens majoritaire (comme précédemment).

Dans l'article sur la turbulence, le terme « laminaire » a plusieurs sens : laminaire dénote soit une algue, soit la caractérisation d'un écoulement non turbulent. Le sens majoritaire (à 60%) de ce terme est l'algue. Avant désambiguïsation, ce terme était rattaché dans le graphe au concept représentant l'algue. Cette anomalie n'a pas affecté la cotation des concepts. Lors de la phase de désambiguïsation, nous avons réactivé tous les concepts rattachés au terme « laminaire » : « laminaria » et « écoulement_laminaire ». Or, « écoulement_laminaire » est le fils direct de « mécanique_des_fluides » ; nous conservons donc ce concept et désactivons « laminaria ».

Après avoir désambiguïsé le graphe, il est très souhaitable de recommencer la cotation des concepts ainsi que celle des mots-clés et des thématiques. En effet, le graphe a été modifié et est surement plus proche du graphe idéal. Ainsi, dans l'exemple précédent le terme « laminaire » est classé en deuxième place du classement des mots-clés.

2.10 Evaluation de l'extraction des descripteurs

Afin d'évaluer nos méthodes, nous mettons en place un protocole d'évaluation de l'efficacité de l'extraction des concepts forts, des mots-clés dans le texte, des thématiques et de la désambiguïsation.

2.10.1 Protocole d'évaluation

Comme le suggèrent (Mihalcea & Csomai, 2007) pour les évaluations de Wikify!, nous avons choisi 100 articles de Wikipédia dits de qualité. Pour rappel, les articles de qualité de Wikipédia ont été promus pour la pertinence de leur contenu mais aussi pour la mise en forme (proportion correcte de liens internes, qualité syntaxique et orthographique etc.). Il existe à l'heure actuelle environ 800 articles de qualité dans Wikipédia en français. (Mihalcea & Csomai, 2007) proposent la mesure du keyphraseness qui permet de transformer un document textuel quelconque en une page ressemblant à un article Wikipédia, c'est-à-dire présentant des liens vers d'autres articles Wikipédia.

Dans un premier temps, nous allons extraire, pour chacun des 100 articles Wikipédia, le wiki code (la source du code MediaWiki). A partir de cette source, nous allons extraire les données suivantes :

- Le texte plat de l'article (sans lien, ni mise en page) ;
- Les liens internes ;
- Les catégories auxquelles l'article appartient ;
- Les thématiques auxquelles l'article appartient.

Nous allons utiliser en entrée de notre modèle les 100 textes plats ainsi que le réseau de catégorie Wikipédia. Pour obtenir le sens majoritaire des concepts, nous avons analysé les liens du

corpus de Wikipédia afin de déterminer pour chaque terme du document, l'ensemble des articles auxquels il peut faire référence ainsi que les proportions. Par exemple, le terme « France » redirige vers l'article « France » (au sens du pays) 129210 fois sur 132752 (97%).

Nous désactivons définitivement des graphes les catégories suivantes :

- Espace_non_encyclopédique
- Code
- Homonymie
- Catégorie_éponyme
- Code_AITA_des_aéroports
- Chronologie
- Concept_philosophique

Ces catégories et leurs descendants posent des problèmes car ils génèrent du bruit. Par exemple, « Concept_philosophique » est une catégorie contenant la quasi-totalité des concepts abstraits et ambigus (les termes associés sont souvent des verbes du type « Faire », « Mettre », « Posséder »).

Nous éliminons également les catégories contenant les termes suivants pour la même raison :

- « : » (le caractère deux points)
- Siècle
- Wikipédia:
- Sigle
- Homonymie
- _par_
- Article_avec_
- Wikimedia
- N'importe quel nombre

Avec les modifications apportées, les graphes représentés semblent issus d'une base de connaissances généralistes.

Nous appliquons ensuite les opérations et algorithmes afin d'extraire dans un premiers temps les concepts forts.

2.10.2 Evaluation du score des concepts forts

Pour l'évaluation des concepts forts, nous allons comparer, pour chacun des 100 textes, la liste des plus forts concepts avec la liste des catégories extraites. Il est évident que les concepts forts et les catégories sont deux choses bien distinctes. En revanche, comparer les listes de concepts forts et de catégories n'est pas dépourvu de sens. En effet, un concept est plus général qu'un simple terme et

possède la notion d'appartenance (la subsomption). Il en est de même pour les catégories de Wikipédia. Nous allons comparer notre approche avec la celle de la densité conceptuelle (Agirre & Rigau, 1996) :

$$CD(C,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendant(C)}$$

Où C est un concept et m le nombre de termes reliés à C. Nous rappelons que cette mesure n'est pas une mesure de cotation à l'origine, mais un début d'approche pour la désambiguïsation. Il est donc attendu que cette approche soit décevante. Pour évaluer les deux approches, nous allons tracer les courbes de rappel/précision obtenues. Le résultat est présenté dans la Figure 32.

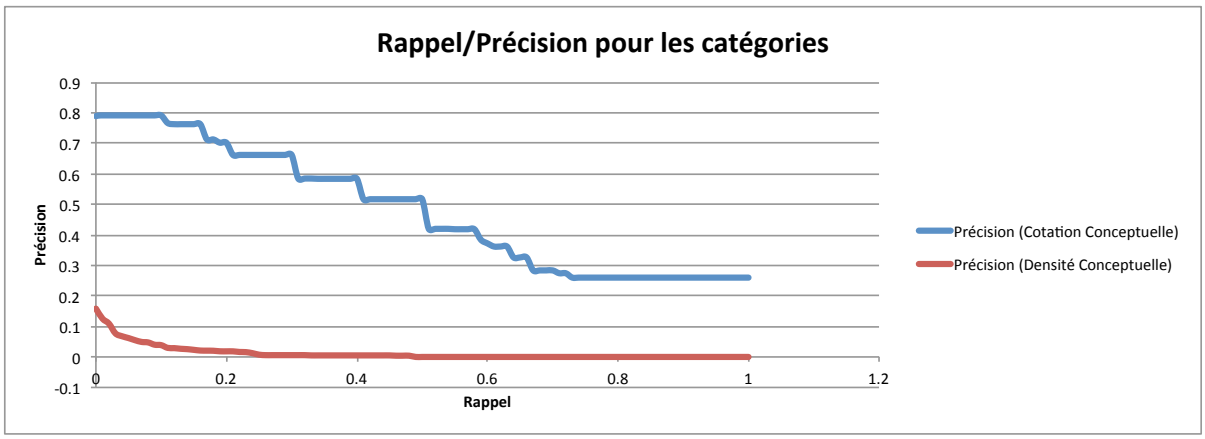


Figure 32 : Rappel/Précision pour la cotation conceptuelle

Cette figure montre le comportement de notre approche et celle de la densité conceptuelle. Nous observons que la densité conceptuelle n'est absolument pas adaptée à la comparaison avec les catégories des articles analysés. En revanche, nous estimons subjectivement que les concepts denses sont forts intéressants pour l'indexation.

Notre approche, en revanche, présente de bonnes correspondances avec les catégories des articles. Nous avons une précision de 52% pour un rappel à 50%. Cela signifie que pour un article appartenant à 100 catégories, il faudrait fournir une liste de 96 concepts pour avoir 50 catégories correctes. Nous sommes satisfaits de ce résultat car il démontre que les concepts forts sont assez généraux pour être souvent considérés comme des descripteurs mais aussi comme des catégories de classement de documents pédagogiques pour une bibliothèque numérique.

2.10.3 Evaluation de l'extraction des mots-clés

Nous allons évaluer notre système d'extraction de mots-clés en comparant la liste des liens internes et nos mots-clés contextualisés. Nous prenons les hypothèses de (Mihalcea & Csomai, 2007)

stipulant que les mots-clés d'un document étaient souvent des liens internes vers d'autres articles. Nous comparons, donc nos résultats avec ceux du Keyphraseness et de TF-IDF. Le résultat est présenté dans la Figure 33.

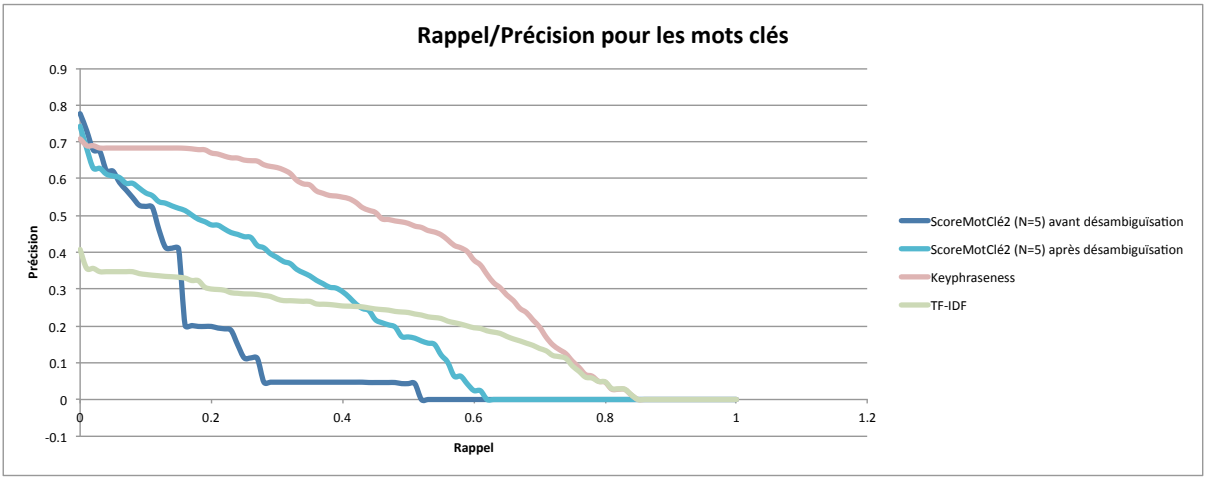


Figure 33 : Rappel/Précision pour l'extraction des mots-clés

Nous ne montrons que les résultats du second score proposées avec N= 5 pour l'extraction des mots-clés car les meilleurs résultats sont obtenus en utilisant cette métrique. Nous pouvons constater que notre métrique avant et après désambiguïsation est largement dépassée par la mesure du Keyphraseness en toutes circonstances mais également dépassé par TF-IDF à haut taux de rappel. A faible rappel, nos méthodes sont assez précises mais toujours en dessous du Keyphraseness. Nous remarquons également que nos algorithmes tendent vers 0 plus vite que les autres. Cela est dû à la désactivation des termes ambigus d'environ 40% avant la désambiguïsation et de 30% après désambiguïsation.

Le comportement de notre algorithme face à ce protocole d'évaluation est attendu. En effet, le Keyphraseness aura tendance à proposer les mots étant fréquemment des liens contrairement au TF-IDF. Un exemple frappant est le cas du terme « France » : pour le Keyphraseness, peu importe le contexte, le terme « France » est souvent un lien et est donc souvent sélectionné. A contrario, le score IDF de « France » est très faible car fréquent dans de nombreux documents et est donc rarement sélectionné. Notre approche a l'avantage de le sélectionner lorsque le contexte s'y prête.

(Mihalcea & Csomai, 2007) présente une autre approche pour l'évaluation de l'extraction des mots-clés. Dans les articles de qualité de Wikipédia seul 6% des termes du document sont des liens internes Wikipédia. Nous procédons donc à une évaluation à 6% de termes extraits sur les articles (Tableau 12).

	Rappel	Précision	F-Mesure
Score - N=5	18.12%	**50.41%**	26.66%
Score - N=10	22.35%	44.96%	29.85%
Score - N=15	26.07%	39.54%	31.42%
Score - N=20	27.35%	38.61%	32.02%
Score - N=25	29.21%	36.22%	32.34%
Score2 N=5	32.40%	34.05%	33.21%
Score - N=5 (après désambiguisation)	17.69%	31.86%	22.75%
Score - N=10 (après désambiguisation)	24.34%	30.72%	27.17%
Score - N=15 (après désambiguisation)	25.64%	29.25%	27.33%
Score - N=20 (après désambiguisation)	27.06%	30.03%	28.47%
Score - N=25 (après désambiguisation)	28.63%	30.83%	29.69%
Score2 - N=5 (après désambiguisation)	28.50%	29.80%	29.14%
Keyphraseness	**40.62%**	43.96%	**42.22%**
TF-IDF	25.99%	26.06%	26.02%

Tableau 12 : Tableau de rappel/précisons à 6% de termes sélectionnés

Nous remarquons encore une fois, l'efficacité du Keyphraseness par rapport à TF-IDF et à notre approche. En revanche nous remarquons le haut score de précision que nous obtenons avec notre approche avant désambiguïsation et en utilisant les 5 concepts des 5 premiers rangs. Cela est dû au fait que moins de 6% de termes sont extraits (nous rappelons que cette cotation ne prend en compte que les termes ancêtres des 5 premiers concepts). Cela confirme bien que la cotation conceptuelle est efficace pour positionner les documents.

Fort de ce constat, nous désirons mettre en œuvre une nouvelle mesure permettant de contextualiser le Keyphraseness. En effet, le Keyphraseness n'est pas sensible au contexte du document et nous introduisons donc la mesure du Keyphraseness contextualisé. Cette mesure permet de pondérer les scores Keyphraseness des termes en fonction de leur généalogie avec les concepts importants du document. La formule du Keyphraseness contextualisé (ou KeyphrasenessC) est :

$$KeyphrasenessC(terme, N) = \begin{cases} Keyphraseness(terme) \times \left(1 + \dfrac{N - Rang(terme)}{N}\right) & si\ N > Rang(terme) \\ Keyphraseness(terme) & sinon \end{cases}$$

Où Rang(terme) est le rang du premier concept contenant le terme (nous rappelons que le rang est établi en fonction du score de cotation conceptuelle) et N le nombre de meilleurs concepts à prendre en compte. Par exemple, si un terme appartient au premier concept, son score de Keyphraseness est

quasiment doublé. Avec le protocole précédent, nous obtenons le score de rappel et de précision du Tableau 13 :

	Rappel	Précision	F-Mesure
Keyphraseness	40.62%	43.96%	42.22%
KeyphrasenessC	**46.74%**	**50.89%**	**48.72%**

Tableau 13 : Comparaison du rappel et de la précision du Keyphraseness et du Keyphraseness contextualisé

Nous remarquons que la contextualisation du Keyphraseness permet d'améliorer sensiblement le score de rappel et de précison du Keyphraseness (+6.12% et +6.93% respectivement). Cela s'explique simplement par le fait que même les termes à faible score de Keyphraseness sont susceptibles d'être des liens d'un article de Wikipédia, à condition que ces termes soient utilisés dans un certain contexte. Par exemple, le terme « chaud » est très souvent utilisé, mais dans le contexte de la thermodynamique ce terme est souvent un lien interne de Wikipédia qui mène vers l'article sur la « Chaleur_latente» ou vers celui traitant de l'«Équation_de_la_chaleur».

Un dernier point intéressant est de remarquer que notre approche après désambiguïsation est plus mauvaise qu'avant la désambiguïsation. Cela peut-être dû à 2 phénomènes :

1. Soit la désambiguïsation échoue ;
2. Soit la désambiguïsation n'échoue pas mais les termes pouvant être ambigus (c'est-à-dire ayant plusieurs sens) ont une probabilité plus faible d'être des liens internes dans Wikipédia.

Nous répondrons à cette interrogation lors de l'évaluation de la désambiguïsation (Section 2.10.5.)

2.10.4 Evaluation de l'extraction des thématiques

Le protocole d'évaluation pour les thématiques est rigoureusement le même que pour les concepts forts. Nous allons mettre en correspondance la liste des portails avec les concepts extraits grâce à notre score de cotation thématique. Le résultat est représenté en Figure 34.

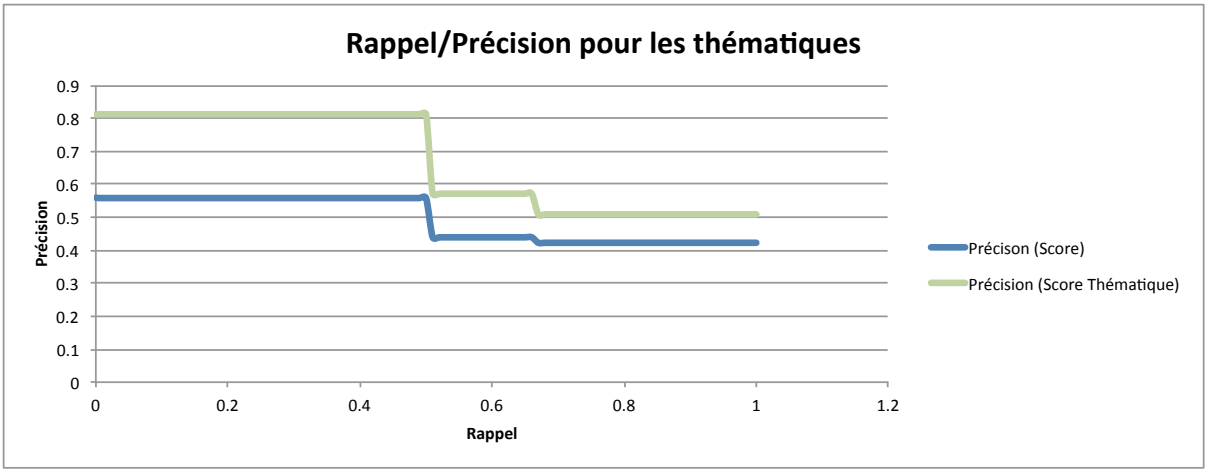


Figure 34 : Rappel/Précision pour l'extraction des thématiques

Nous remarquons que la mesure pour la cotation des thématiques est assez satisfaisante. En effet, nous l'appliquons sur notre corpus test et remarquons que la métrique de cotations thématiques est plus performante que la cotation conceptuelle. En effet, chaque thématique est un concept du graphe et a donc un score conceptuel ainsi qu'un score de thématique. Nous remarquons également les trois paliers analogues caractérisant le fait que chaque article du corpus test appartient au maximum à trois thématiques.

2.10.5 Evaluation de la désambiguïsation

Pour la désambiguïsation, nous allons mettre en place un nouveau protocole. Nous allons extraire pour chacun des 100 documents les liens internes ambigus, c'est-à-dire ceux dont le texte est différent du titre de l'article vers lequel il pointe (par exemple, [Écoulement laminaire|laminaire]). Nous allons ensuite calculer le taux de bonnes réponses fournies par le taux de réponses attendues. Par exemple, si nous avons 10 liens et que le système a désambigüisé correctement 5 d'entre eux, cela signifie que le taux de réussite est 0.5. Nous allons comparer notre approche avec les mesures de Leacock-Chodorow (Leacock & Chodorow, 1998) et le sens majoritaire.

Sens Majoritaire	68.39%
Score N=30	71.89%
Leacock-Chodorow	48.99%

Tableau 14 : Evaluation de la désambiguisation

Nous constatons que notre approche permet de faiblement améliorer le score de désambiguïsation par rapport au sens majoritaire (+3.5%). Cela nous permet de croire que l'extraction des concepts importants du graphe fonctionne en général. En effet, nous rappelons que la désambiguïsation se fait grâce à la cotation conceptuelle. Cela répond à notre interrogation concernant la baisse de performance de l'extraction des mots-clés après désambiguïsation. Il s'avère que les termes ambigus n'ayant pas de sens majoritaire sont rarement des liens internes d'articles Wikipédia.

2.11 Conclusion

La cotation conceptuelle des graphes de document permet d'effectuer 5 tâches importantes de l'indexation :

1. de fixer le contexte du document ;
2. d'extraire les concepts pouvant servir de descripteurs ;
3. d'extraire les mots-clés du texte pouvant, eux-aussi, servir de descripteurs ;
4. d'extraire les thématiques du document pouvant également servir de descripteurs ;
5. désambiguïser des termes et les promouvoir au rang de descripteurs ;

Les batteries de tests menés lors de l'évaluation donnent des résultats mitigés. Pour autant, en observant de manière subjective les descripteurs proposés, nous pensons que les résultats de l'extraction des descripteurs sont exploitables pour la mise en place d'un système d'aide à l'indexation. Nous démontrons cela en présentant le prototype à des bibliothécaires de UNIT. Le résultat de cette validation est présenté en Chapitre 4.

Le chapitre suivant est dans la continuité de la cotation conceptuelle. Nous nous demandons s'il est envisageable d'exploiter la cotation conceptuelle afin de développer un SRI permettant de retrouver des documents en utilisant une mesure de similarité sur les graphes. Le document est représenté par un graphe mais la requête sera également représentée par un graphe. Il suffira donc de trouver une mesure de similarité sur les graphes permettant de décider de la pertinence d'un document pour une requête donnée. Cette mesure de similarité pourrait aussi servir à faire de la similarité inter-documents (similarité entre 2 documents) et intra-document (similarité entre 2 partie d'un même document ou similarité entre un document et une partie du document). La similarité inter-document pourrait permettre la mise en place d'un système de recommandation documentaire et la similarité intra-document une plateforme d'analyse de la structure thématique de document pour la détection des ruptures et des retours sémantiques et thématiques.

Chapitre 3. Similarité entre graphes de document

3.1 Introduction

Dans ce chapitre, nous proposons d'exploiter la cotation conceptuelle pour développer un SRI permettant de retrouver des documents, en utilisant une mesure de similarité sur les graphes. Le document est représenté par un graphe, ainsi que la requête. Une mesure de similarité sur les graphes permet de donner la pertinence d'un document pour une requête donnée. Cette mesure permet aussi de faire des comparaisons inter-document et intra-document (sections de documents). La similarité inter-document permet la mise en place d'un système de recommandation documentaire et la similarité intra-document apporte une plate-forme d'analyse de la structure thématique de document pour la détection de ruptures et de retours sémantiques et thématiques.

Dans une première section (3.2), nous répertorions l'ensemble des modes de représentation et des métriques de similarité associées à ces représentations qui peuvent s'appliquer à notre contexte. La section 3.3 propose une évaluation des métriques recensées. Aussi, en section 3.4, nous proposons une nouvelle approche utilisant une métrique de similarité pour la désambiguïsation. Finalement, nous validerons l'apport des métriques de similarité pour trois activités différentes : la recommandation documentaire, la recherche documentaire et l'analyse de la structure thématique d'un document (section 3.5).

3.2 Représentation des documents et métriques de similarité associées

Il existe plusieurs approches pour mesurer la similarité entre deux documents ou une requête et un document. Ces approches ont été décrites dans le Chapitre 3 (Indexation et Recherche d'information sémantique et conceptuelle). Dans cette section, nous nous intéressons plus précisément au modèle vectoriel car il permet de comparer des objets en fonction des éléments qui les composent ainsi que leurs relations (pour le modèle généralisé).

Si document (ou une requête) peut être représenté par un sous-ensemble d'éléments pondérés ou non, le modèle vectoriel classique fournit un mode de représentation et des métriques de similarité associées.

Dans notre contexte, les éléments peuvent être les descripteurs des documents. Par exemple, si nous utilisons les descripteurs « mécanique des fluides », « turbulence » et « tourbillon », on aura un vecteur $\vec{D}$ à 3 composantes $w_{i,D}$ représentant le poids du descripteur *i* dans le document D :

$$\begin{matrix} Mécanique\ des\ fluides \\ Turbulence \\ Tourbillon \end{matrix} \qquad \vec{D} = \begin{pmatrix} w_{1,D} \\ w_{2,D} \\ w_{3,D} \end{pmatrix}$$

En général, la mesure du cosinus est utilisée :

$$Cos\left(\vec{D_i}, \vec{D_j}\right) = \frac{\vec{D_i}.\vec{D_j}}{\left\|\vec{D_i}\right\|\left\|\vec{D_j}\right\|}$$

Pour définir la base de notre l'espace vectoriel, ainsi que les poids des éléments pour un document donné, nous pouvons :

- utiliser l'ensemble des termes présents dans le corpus des documents avec les pondérations suivantes :
  - $w_{i,D} = tf - idf(t_i, D)$;
  - $w_{i,D} = keyprhaseness(t_i)$;
  - $w_{i,D} = ScoreMotClé2(t_i, 5)$ avant désambiguïsation.
- utiliser l'ensemble des concepts présents dans les graphes des documents du corpus avec les pondérations suivantes :
  - $w_{i,D} = ScoreNorm(C_i)$;
  - $w_{i,D} = ScoreNorm(C_i)$ si $C_i \in$ Top N des concepts, 0 sinon ;
  - $w_{i,D} = 1$ si $C_i \in$ Top N des concepts, 0 sinon ;
  - $w_{i,D} = ScoreNorm(C_i)$ après désambiguïsation ;
  - $w_{i,D} = ScoreNorm(C_i)$ si $C_i \in$ Top N des concepts après désambiguïsation, 0 sinon ;
  - $w_{i,D} = 1$ si $C_i \in$ Top N des concepts après désambiguïsation, 0 sinon.

Par exemple, si un document est indexé par les concepts « mécanique des fluides » et « tourbillon » et un autre avec « turbulence », alors le cosinus est nul, car les relations entre concepts ne sont pas représentées dans le modèle vectoriel classique. Or, le modèle vectoriel généralisé le permet. Il consiste simplement, pour deux vecteurs orthogonaux, à avoir un produit scalaire non nul. Algébriquement, cela signifie qu'il y a une dépendance entre les deux vecteurs.

(Ganesan et al., 2003) introduit la mesure de cosinus généralisée pour la similarité de deux collections d'objets hiérarchiques. Il propose le remplacement du produit scalaire de deux vecteurs par un calcul de similarité entre deux objets de la hiérarchie (cf..: Etat de l'art - 3.3.2.2).

Dans notre contexte, les objets sont les concepts des graphes de document. Soit $\overrightarrow{C_i}$ et $\overrightarrow{C_j}$ les vecteurs représentant les concept C_i et C_j, le produit scalaire est défini comme suit :

$$\overrightarrow{C_i} \cdot \overrightarrow{C_j} = \frac{2 \times profondeur\left(LCA(C_i, C_j)\right)}{profondeur(C_i) + profondeur(C_j)}$$

Deux concepts ont un produit scalaire élevé, si leur plus petit ancêtre commun est proche des deux concepts. Comme la notion de profondeur n'a pas de sens dans les DAG, nous la remplaçons par l'inverse de la généricité (en ajoutant +1 afin d'éviter une généricité nulle) :

$$\overrightarrow{C_i} \cdot \overrightarrow{C_j} = \frac{2 \times \frac{1}{gen\left(LCA(C_i, C_j)\right) + 1}}{\frac{1}{gen(C_i) + 1} + \frac{1}{gen(C_j) + 1}}$$

La généricité utilisée ici est celle d'un concept dans la totalité de la base hiérarchique. Son calcul est le même que celui de la diversité d'un concept dans un graphe.

3.3 Evaluation du calcul de similarité

Pour évaluer les métriques de similarité, nous avons sélectionné des articles de Wikipédia (70 au total : 5 articles par catégorie, 3 catégories par thématique et 4 thématiques) dans les thématiques et catégories suivantes :

- mathématiques :
  - algèbre linéaire ;
  - arithmétiques ;
  - analyse.
- informatique :
  - base de données ;
  - algorithmique ;
  - langage de programmation.
- physique :
  - cosmologie ;
  - mécanique des fluides ;
  - acoustique.
- art et littérature :
  - roman du XIXième siècle ;
  - conte philosophique ;
  - théâtre de l'absurde.

Notre objectif est de vérifier que notre mesure de similarité permet de dire qu'un document d'une catégorie est proche des autres articles de la même catégorie. Nous utilisons un tableau de similarité comme le montre Figure 35:

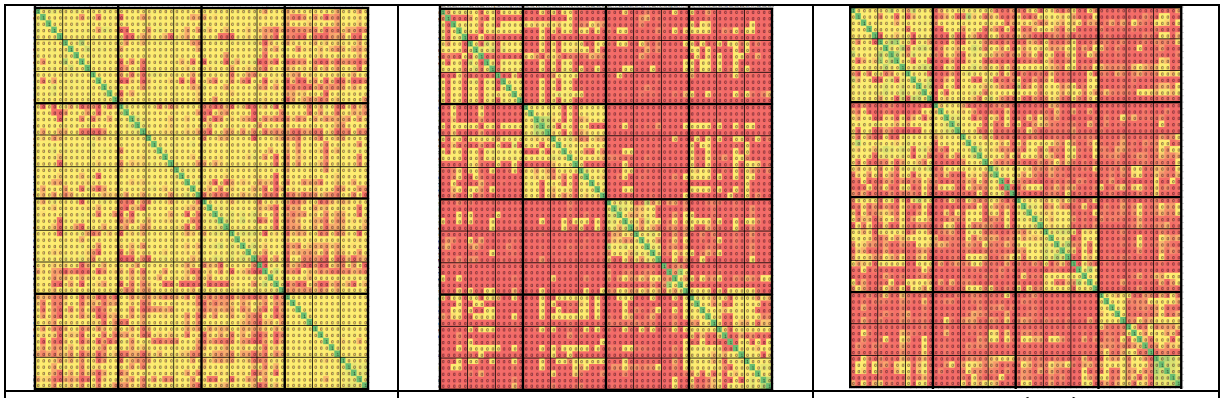
	Algèbre Linéaire	Arithmétique	Analyse	Base de données	Algorithmique	Langage de programmation	Cosmologie	Mécanique de fluides	Acoustique	Roman	Conte	Théatre
Algèbre Linéaire												
Arithmétique												
Analyse												
Base de données												
Algorithmique												
Langage de programmation												
Cosmologie												
Mécanique de fluides												
Acoustique												
Roman												
Conte												
Théatre												

Figure 35 : Tableau présentant l'organisation de l'évaluation (chaque case représente la comparaison de 5 articles d'une catégorie avec 5 articles d'une autre catégorie ; 720 calculs de similarité)

Les cellules proches d'une similarité égale à 1 sont coloriées en vert et les cellules proches d'une similarité nulle en rouge. Le score d'évaluation donné correspond au nombre d'articles en commun entre les 5 meilleurs articles retournés et les 5 articles attendus, c'est-à-dire ceux d'une même catégorie.

3.3.1 Evaluation du modèle vectoriel classique

Nous évaluons d'abord les métriques fondées sur l'espace composé des termes. Les scores TF-IDF, Keyphraseness et ScoreMotClé2 sont utilisés pour pondérer les termes.



TF-IDF : Score = 59,3%	Keyphraseness: Score = 63,3%	$ScoreMotClé2(t_i, 5)$: Score = 64,3%

Figure 36 : Tableau colorié des similarités inter-document utilisant les termes comme base de l'espace vectoriel.

La Figure 36 montre que le score TF-IDF ne discrimine pas assez les documents les uns des autres. En revanche, le score de Keyphraseness et notre cotation permettent de discriminer de manière plus flagrante les documents en fonction des thématiques et des catégories.

La Figure 37 présente les similarités entre documents utilisant la base de concepts avec notre cotation.

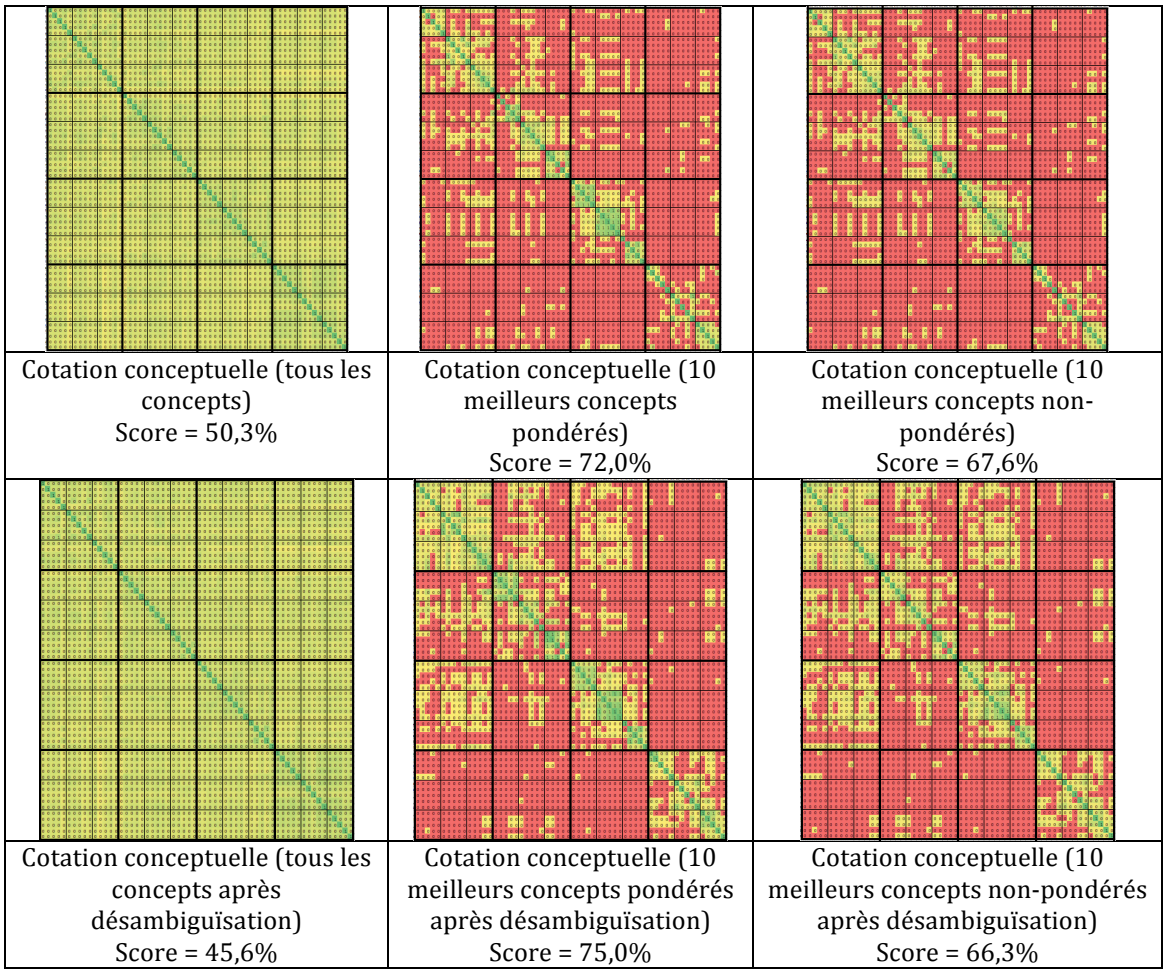


Figure 37 : Tableau colorié des similarités inter-document utilisant les concepts comme base de l'espace vectoriel.

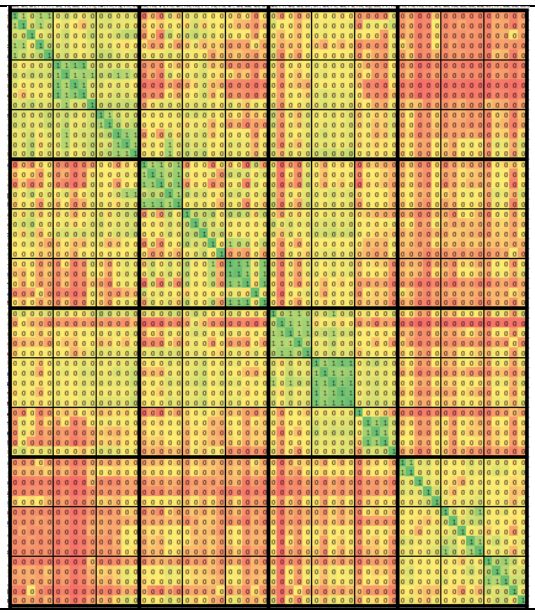
L'utilisation de l'espace des concepts pour la similarité entre documents s'avère plus efficace que l'espace des termes. En revanche, il ne faut utiliser qu'un nombre limité de concepts (les N plus importants). Le meilleur résultat (à 75%) est celui de la similarité entre documents utilisant l'espace des concepts et prenant les 10 meilleurs concepts pondérés après désambiguïsation.

Nous observons aussi que les trois thématiques scientifiques (mathématiques, informatique et physique) sont plus proches entre elles que celle artistique (art et littérature). Les articles traitant des mathématiques sont proches entre eux, mais aussi assez proches des documents traitant de

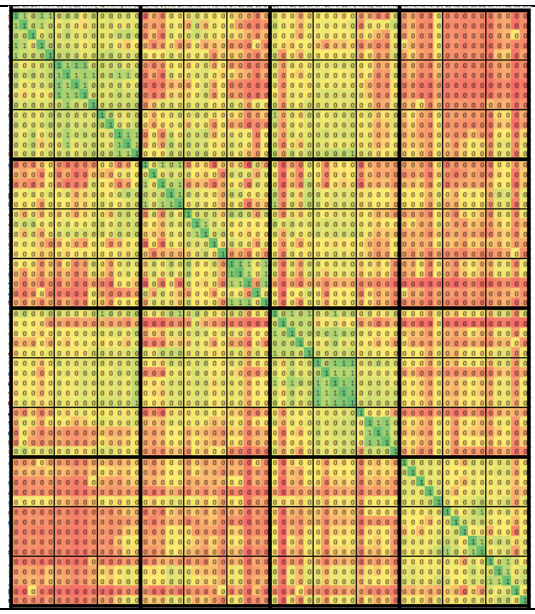
l'informatique et de la physique. En revanche, la similarité entre documents traitant de la physique et de l'informatique est assez faible.

3.3.2 Evaluation du modèle vectoriel généralisé

Dans cette section, nous évaluons l'apport de la prise en compte de la hiérarchie des concepts dans la mesure du cosinus. Le principal résultat attendu est qu'aucun document n'a une similarité nulle avec les autres documents car tous les concepts de la hiérarchie sont liés les uns aux autres (Figure 38).



Cotation conceptuelle 10 meilleurs concepts pondérés après désambiguïsation
Score = 76,3%



Cotation conceptuelle 10 meilleurs concepts non-pondérés après désambiguïsation
Score = 69,3%

Figure 38 : Tableau colorié des similarités inter-document utilisant les concepts comme base d'espace vectoriel et utilisant la hiérarchie des concepts.

Ainsi en utilisant la généalogie dans le calcul de similarité, les scores sont améliorés de respectivement 1% et 3% par rapport à la version classique.

Ceci démontre que nous pouvons discriminer les documents en fonction des concepts extraits des graphes de documents. Il est aussi intéressant de prendre en compte les relations entre concepts pour obtenir un score de similarité documentaire plus performant. Même si visuellement les matrices de similarité de la Figure 38 semblent moins bien discriminer les documents que les matrices de la Figure 37, en réalité ces similarités renforcent d'avantage les documents traitant des mêmes catégories et thématiques. Malgré l'aspect « orangeâtre » des matrices de la Figure 38, les documents a priori proches sont encore plus « verts » que dans les approches précédentes.

La similarité documentaire peut être utilisée pour un système de recommandation : un utilisateur qui s'intéresse à un document peut obtenir des autres documents traitant des mêmes

thématiques. En revanche, il est important de préciser que cette similarité ne permet de répondre qu'à la question « de quel sujet traite ce document ? ». Avec cette approche, il n'est pas envisageable de trouver un document en fonction de son type (« Cours », « Travaux dirigés », « Tutoriel », etc.), du niveau requis pour le comprendre (« Bac +1 », « Bac +5 », etc.) ou d'autres attributs.

3.4 Seconde approche pour la désambiguïsation des termes exploitant la similarité

Dans cette section, nous proposons une approche alternative pour la désambiguïsation. Un terme du document peut être lié à plusieurs concepts du graphe. Pour le calcul de désambiguïsation, l'idée est de calculer la similarité entre le graphe généré par un concept et le graphe du document privé de ce concept.

Pour chaque terme, nous calculons la similarité des graphes de chacun des concepts liés à ce terme et le reste du graphe de document. Nous évaluons la désambiguïsation avec la mesure du modèle vectoriel généralisé pondéré avant désambiguïsation (avec les 10 meilleurs concepts) en utilisant le même protocole d'évaluation que celui décrit dans la section 2.10.1 (Tableau 15).

Sens Majoritaire	68.39%
Score par cotation N=30	71.89%
Score par similarité (N=10)	61.09%
Leacock-Chodorow	48.99%

Tableau 15 : Score de mise en correspondance des termes avec les concepts adéquats

Le score de désambiguïsation par similarité est très inférieur à ceux par cotation et par sens majoritaire. Pour rappel, la désambiguïsation par cotation conceptuelle utilise le sens majoritaire dans le cas où le système ne peut pas décider. Or, la désambiguïsation par similarité n'utilise jamais le sens majoritaire, ce qui explique ce résultat. En revanche, cette mesure est nettement supérieure à celle proposée par (Leacock & Chodorow, 1998).

3.5 Validation des mesures de similarité

Nous avons mis en évidence dans les sections 3.2 et 3.3 plusieurs approches pour le calcul de similarité entre graphes. Nous ne conservons pour la validation que les trois approches les plus efficaces :

1. Modèle vectoriel généralisé : Cotation conceptuelle 10 meilleurs concepts pondérés après désambiguïsation (Score = 76,3%)
2. Modèle vectoriel classique : Cotation conceptuelle 10 meilleurs concepts pondérés après désambiguïsation (Score = 75,0%)

3. Modèle vectoriel classique : Cotation conceptuelle 10 meilleurs concepts pondérés avant désambiguïsation (Score = 72,0%)

Ces trois approches permettent d'avoir de très bonnes performances pour la similarité. En revanche, pour passer d'une méthode sans désambiguïsation à la méthode après désambiguïsation, c'est-à-dire de la méthode 3 à la méthode 2 dans la liste ci-dessus, le temps de calcul est quasiment doublé. En effet, la désambiguïsation consiste quasiment à reconstruire un nouveau graphe dont les concepts importants sont calculés une seconde fois. Cela signifie que pour avoir une amélioration de 3%, le temps de calcul est doublé.

La complexité algorithmique de l'approche la plus efficace est à l'heure actuelle beaucoup trop élevée pour un passage à l'échelle. Pour avoir un ordre d'idée sur le temps de traitement informatique, un modèle vectoriel classique ne va calculer la similarité d'un document qu'avec les documents contenant les mêmes coordonnées (c'est-à-dire les coordonnées non nulles du vecteur de document), la similarité avec les autres documents étant nulle. Le modèle vectoriel classique profite, en effet, de la modélisation algébrique des matrices creuses. En revanche l'approche généraliste va, pour chaque coordonnée d'un document, calculer la similarité avec chaque coordonnée de chaque document de la base. Il est donc inconcevable d'utiliser cette mesure à l'heure actuelle.

Nous proposons donc de valider les approches de la similarité en utilisant uniquement la troisième meilleure mesure. Celle-ci ne dénature pas les approches que nous avons développées tout au long de ce rapport, dans la mesure où les coordonnées du vecteur sont bien des concepts issus de la base de connaissances après analyse du document et cotation des concepts du graphe du document.

Pour cette validation, à l'instar du chapitre précédent (cf. Chapitre 2), nous jugeons nos similarités sur les articles de Wikipédia.

3.5.1 Validation des mesures entre requêtes et documents pour un système de recherche d'information

L'un de nos objectifs initiaux pour la mise en place d'une métrique de similarité était de concevoir un moteur de recherche conceptuelle. Nous traitons la requête comme un document, c'est-à-dire que nous construisons un graphe à partir des données textuelles de celle-ci. Ensuite, nous calculons la similarité entre les graphes des documents et le graphe de la requête.

Une des différences fondamentales entre une requête et un document est sa taille en terme de nombre de mots. Cette différence induit une difficulté : si l'utilisateur entre la requête « Java », à quel concept doit-on relier ce terme ? Ici, il n'y a aucune possibilité pour désambiguïser automatiquement le terme (ce constat a déjà été établi par (Voorhees, 1993) pour les requêtes courtes). Plutôt que de prendre le sens majoritaire du terme « Java », notre système ajoute une étape : la désambiguïsation manuelle de la requête comme le montre la Figure 39. Nous proposons à l'utilisateur de choisir le bon sens du concept « Java ».

Figure 39 : désambiguïsation manuelle de la requête "Java" dans un système de recherche d'information conceptuelle.

Lorsque nous sélectionnons «Java_(langage)» et que nous recherchons les documents similaires à cette requête sur la base des 65 documents de la section 3.3, le SRI nous renvoie 4 documents de la catégorie « Langage de programmation » et un document sur les « Bases de données ». Cette observation nous encourage donc à penser que l'approche conceptuelle est adaptée à la recherche documentaire.

3.5.2 Validation des mesures entre parties du document pour un outil d'analyse de la structure thématique

Il s'agit ici de localiser les ruptures thématiques dans un document, c'est-à-dire les changements de sujets traités (toutes les parties du document ne traitent pas des mêmes thèmes). Nous proposons donc de découper le graphe du document en N sous-graphes représentant chacun une partie du document. Actuellement, nous découpons le graphe en fonction du nombre de termes qui le compose (par exemple, pour un graphe de 100 termes, nous pouvons le scinder en 4 sous-graphes de 25 termes chacun). Ces termes sont connexes dans le texte et représentent une section non sécable du document original.

Pour ce faire,

- nous exploitons le graphe construit à l'étape précédente et nous le divisons en N sous-graphes ;
- nous utilisons la métrique de similarité (utilisant les 10 concepts les plus forts pondérés avant désambiguïsation). Cette métrique est la troisième meilleure (72% de bonne similarité) et est surtout la plus rapide (les deux meilleures sont beaucoup trop longues à calculer).

Nous analysons la similarité du graphe initial (document entier) avec chacun des N sous-graphes et nous comparons les ruptures détectées par le système avec celles détectées manuellement.

Par exemple, l'analyse de l'article « Principe de relativité » présente une rupture : une analogie entre la théorie de la relativité et le monde marin est expliquée par Galilée[28]. La similarité est calculée

[28] Dialogue sur les deux grands systèmes du Monde

entre le graphe du document et chacun des graphes des 20 parties du document. Le résultat est illustré par le graphique Figure 40.

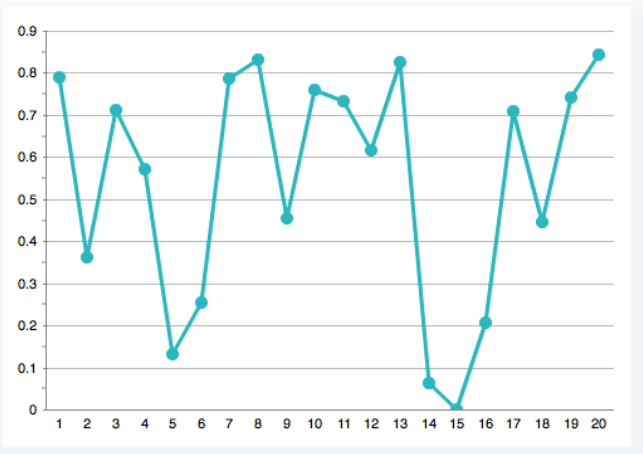


Figure 40 : Evolution thématique de l'article Wikipédia "Principe de relativité"

Le graphique montre deux grandes ruptures. La première concerne la cinquième partie du document qui contient des formules mathématiques que le système n'est pas capable d'interpréter. La seconde rupture concerne la quinzième partie du document qui traite du discours de Galilée.

Nous pensons que les mesures de similarité entre graphes sont aussi adaptées au calcul de rupture conceptuelle ou thématique d'un document.

3.6 Conclusion

Ce chapitre a permis de mettre en place et d'évaluer des métriques de similarité entre graphes de document et donc de similarité entre documents. Nous avons montré que les concepts des graphes sont très efficaces non seulement pour l'indexation des documents, mais aussi la recherche documentaire. Les concepts et leurs relations dans une base de connaissances peuvent aussi être utiles à la mise en place d'un système d'aide à l'indexation, à la recherche d'information et à analyse de la structure thématique (application directe des mesures de similarité entre graphes de document élaborées dans ce chapitre).

Chapitre 4. Prototype pour UNIT

4.1 Introduction

Nous avons construit un modèle permettant de représenter un document en utilisant les termes qui la composent et leurs relations avec des concepts issus d'une base de connaissances externe. Nous avons proposé des méthodes permettant d'extraire de ce modèle les descripteurs du document (les thématiques et les mots-clés du texte) à partir des concepts forts. Nous avons aussi développé deux approches pour la désambiguïsation ainsi qu'un calcul de la similarité inter-document et intra-document.

Nous décrivons, dans ce chapitre, le prototype d'un système d'aide à l'indexation et la recherche de documents qui

- fournit à partir du contenu textuel d'un document, les concepts forts, les thématiques et les mots-clés du document ;
- à partir d'une métrique de similarité entre graphes, localise les ruptures conceptuelles et thématiques dans le document ;
- à partir d'un corpus de documents indexés conceptuellement avec nos méthodes, restitue à l'utilisateur le résultat d'une recherche documentaire.

Dans une première section, nous présentons l'architecture du prototype avec ses caractéristiques techniques. Pour chacune des fonctionnalités du prototype, nous choisissons les méthodes et les métriques appropriées et expliquons nos choix. Enfin, nous expliquons le cheminement à effectuer pour accomplir les tâches d'indexation, d'analyse de ruptures thématiques et la recherche documentaire.

4.2 Architecture du prototype

Les utilisateurs de notre prototype sont les documentalistes d'une bibliothèque numérique pour les outils d'indexation et les usagers pour la recherche d'information. L'outil que nous proposons est très simple d'utilisation et permet de réaliser les trois tâches citées ci-dessus.

Pour construire la base de connaissances, nous partons de l'ensemble du corpus Wikipédia. Ce corpus[29] est au format XML faisant environ 1.5 Go. Pour chaque page de Wikipédia, le titre, le type (article, catégorie, etc.) ainsi que le WikiCode (pour obtenir les catégories) sont présents, ce qui nous permet de générer le réseau de catégories Wikipédia[30]. Nous analysons alors le corpus de Wikipédia pour extraire 3 bases de données comme le montre la Figure 41:

- Une base de graphe NEO4J[31] : NEO4J est une base de données « noSQL » qui permet de stocker une base sous forme de graphe. L'avantage de cette base est lié à ses performances pour la recherche et son parcours du graphe. Celui-ci est composé des pages de Wikipédia (articles, catégories, portails) et des relations entre pages et catégories. Lorsqu'une page appartient à une catégorie, une relation relie le nœud représentant la page au nœud représentant la catégorie. De même, nous construisons des relations entre termes et pages. Ces relations sont construites en analysant les liens internes de type (Titre_de_la_page|termes) ;
- Les pages en WikiCode sont stockées dans une base de données MySQL 5.1[32]. Cela permet d'accéder au contenu d'une page Wikipédia en local ;
- Lucene 3.1[33] pour les statistiques sur le corpus : avec l'outil Lucene-Search[34], nous pouvons stocker un ensemble important de statistiques tels que la fréquence des termes, la fréquence des liens, etc. Nous utilisons cette base pour le calcul de TF-IDF ou Keyphraseness, mais surtout pour le calcul des sens majoritaires d'un terme.

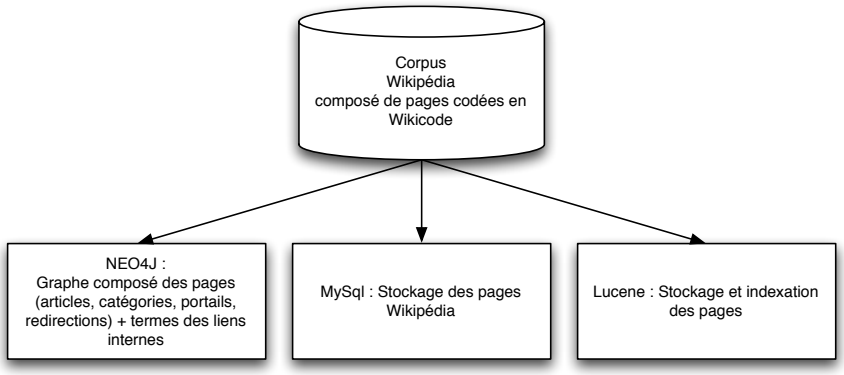


Figure 41 : Utilisation des trois bases de données : une base de graphe, une base de textes et une base d'indexation pour les statistiques

La Figure 42 montre le flot de traitements pour la construction et la cotation des graphes. Il y a une mise en correspondance des termes d'un document et les termes du graphe NEO4J. NEO4J permet la construction du graphe.

[29] http://download.wikimedia.org/frwiki/20110201/frwiki-20110201-pages-articles.xml.bz2
[30] Ce graphe existe également au format SQL : http://download.wikimedia.org/frwiki/20110201/frwiki-20110201-categorylinks.sql.gz
[31] http://neo4j.org/
[32] http://www.mysql.com/
[33] http://lucene.apache.org/java/docs/index.html
[34] http://www.mediawiki.org/wiki/Extension:Lucene-search

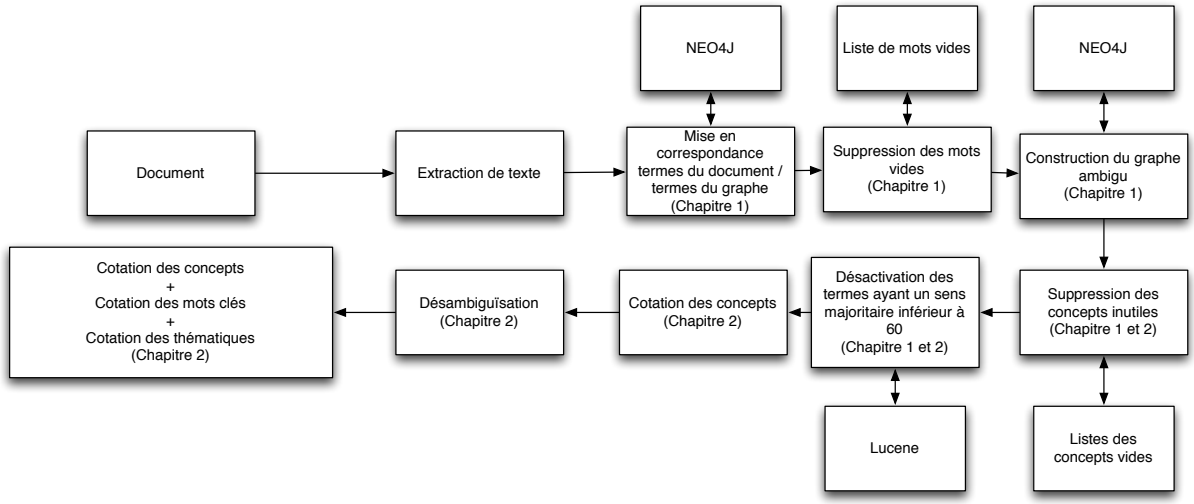


Figure 42 : Flot de traitements pour la construction et la cotation des graphes

Tous les modules présentés dans ce flot de traitements sont codés en JAVA et les échanges entre modules ce font via le framework Spring[35]. Ce flot prend en entrée un document textuel et fournit en sortie un graphe de document ainsi que trois listes de cotations (concepts forts, mots-clés, thématiques). Les données de sortie sont stockées dans deux bases :

- une base MySql qui mémorise le graphe et les cotations via le framework Hibernate[36] ;
- une base Lucene qui ne mémorise que les listes de cotations (sans le graphe). Cette base est utilisée pour la recherche d'information ne prenant pas en compte la hiérarchie des concepts.

Nous avons également conçu un module écrit en JAVA de calcul de similarité entre deux graphes. Enfin, pour la création du prototype, nous avons utilisé les technologies JSF[37] et PrimeFaces[38] permettant de manipuler nos modules via une page web.

4.3 Description des fonctionnalités du prototype

Dans cette section, nous détaillons l'utilisation des trois tâches proposées par notre prototype : l'indexation, la détection des ruptures sémantiques et la recherche d'information.

4.3.1 Extraction des descripteurs et indexation d'un document

Pour l'indexation, le documentaliste entre le texte brut à analyser. Ensuite, le traitement se fait en plusieurs étapes :

[35] http://www.springsource.org/
[36] http://www.hibernate.org/
[37] http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html
[38] http://www.primefaces.org/

- construction du graphe ambigu ne gardant que les termes liés à un concept dont la probabilité est supérieure à 60% ;
- suppression des concepts vides (voir Chapitre 2) ;
- calcul des concepts forts utilisant notre cotation ;
- extraction des thématiques à partir des 5 meilleurs concepts forts ;
- par contre, pas de désambiguïsation du graphe car cette étape est longue et n'augmente pas significativement les performances du système ;
- pas d'extraction de mots-clés : ceux-ci sont surlignés lorsque le documentaliste choisit manuellement les concepts qui l'intéressent. Les mots surlignés sont les termes générateurs des concepts forts.

Pour expliquer le fonctionnement de notre outil d'aide à l'indexation, nous prenons l'exemple d'un texte extrait d'un PDF du site UNIT comme le montre la Figure 43. Le cours en question traite du protocole HTPP[39].

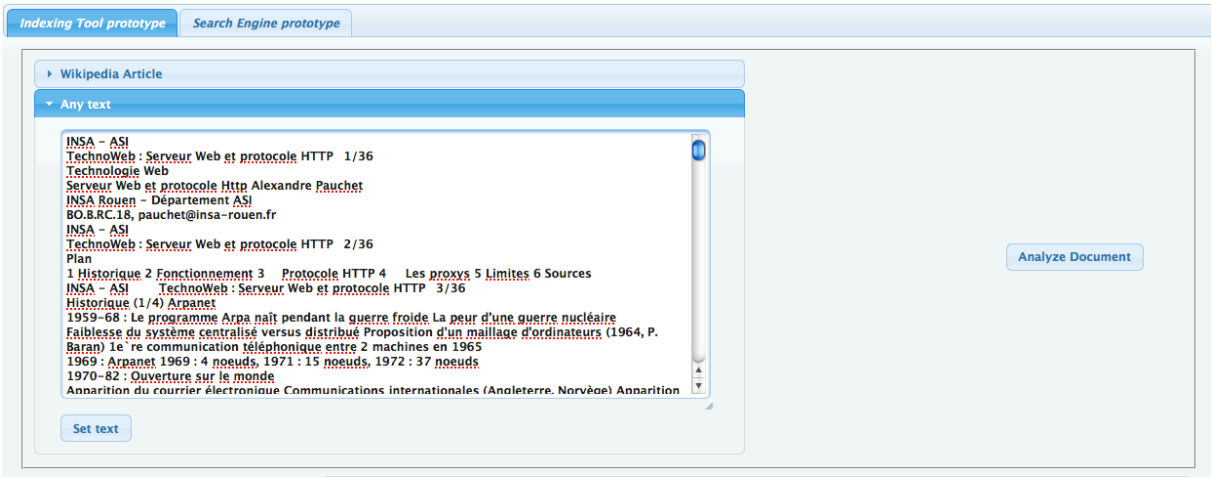


Figure 43 : Aide à l'indexation d'un cours UNIT sur le protocole HTTP

Après analyse du document, les 10 meilleurs concepts forts ainsi que les thématiques (calculées grâce aux 5 meilleurs concepts forts) sont proposés. La Figure 44 montre le résultat de cette extraction pour le texte en exemple. Si le documentaliste estime qu'une thématique proposée est bonne, il la sélectionne. Ceci modifie alors la liste des concepts forts. Par exemple, la Figure 45 montre que la sélection de la thématique « Télécommunication » fait apparaître de nouveaux concepts qui sont les meilleurs concepts liés à la thématique « Télécommunication ». Le documentaliste peut sélectionner plusieurs thématiques parmi celles proposées. Les concepts apparaissant sont alors l'union des meilleurs concepts liés aux thématiques sélectionnées.

[39] http://www.unit.eu/ori-oai-search/notice/view/unit-ori-wf-1-4265

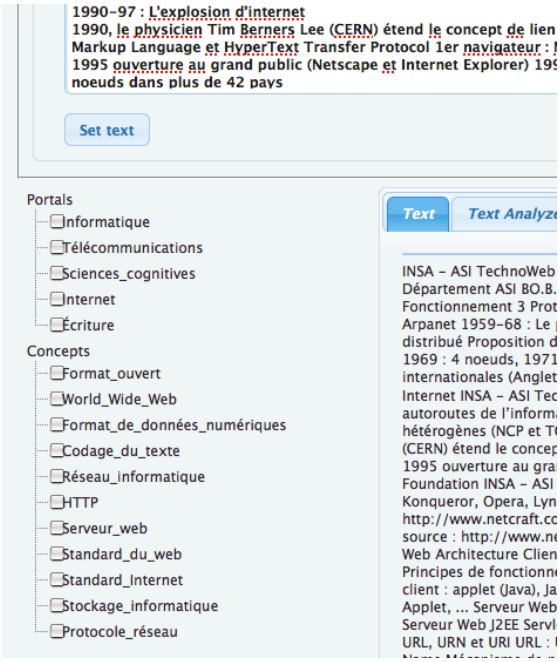


Figure 44 : thématiques et concepts forts extraits d'un cours UNIT sur le protocole HTTP

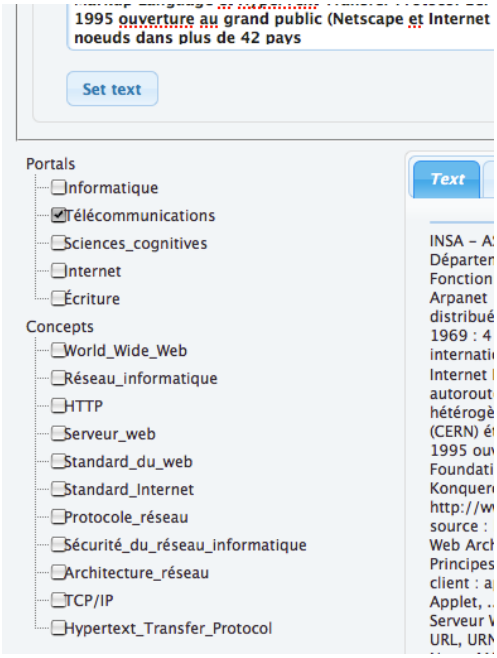


Figure 45 : nouveaux concepts après sélection de la thématique « Télécommunication »

Une fois les thématiques choisies, le documentaliste peut sélectionner les concepts importants qui font apparaître dans le texte les termes liés à ces concepts. La Figure 46 illustre la sélection de la thématique « Télécommunication » et du concept « HTTP ». Les termes relatifs à ce concept sont surlignés en jaune dans le texte. Les concepts forts appartiennent à la liste proposée selon la thématique choisie. La sélection de plusieurs concepts fera apparaître en jaune l'union des termes liés à ces concepts.

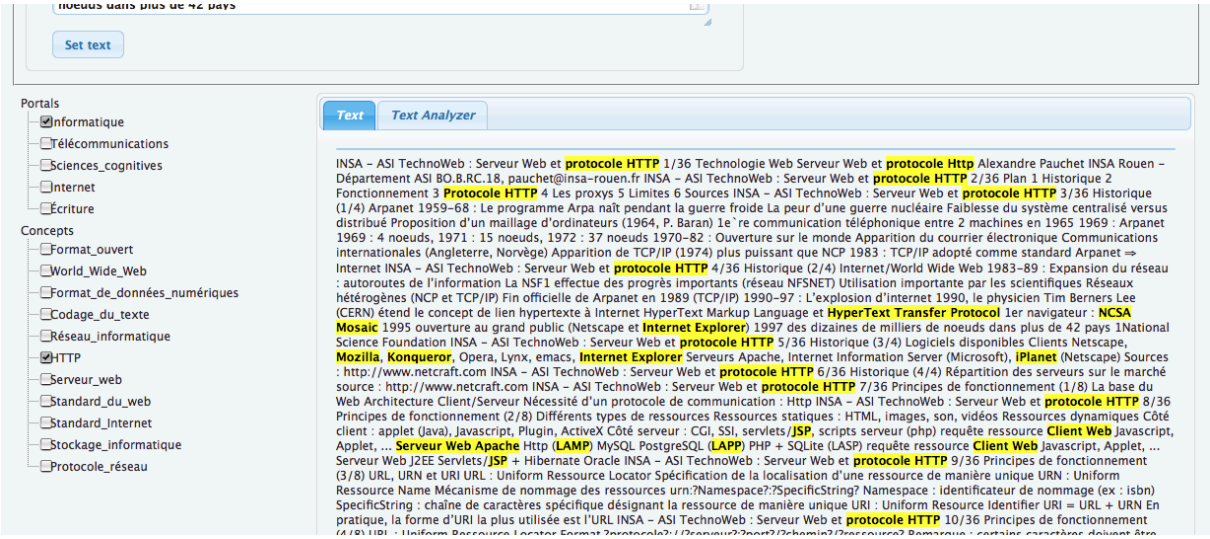


Figure 46 : Sélection de la thématique « Télécommunication » et du concept « HTTP ». Les termes traitant de HTTP sont surlignés en jaune.

En terme de complexité, le temps d'analyse est approprié à une analyse document par document, mais inapplicable à grande échelle pour le moment. En effet, la construction du graphe est assez rapide, mais la acyclisation du graphe est complexe : lors de la phase de construction, 80% du calcul est dédié à l'acyclisation. Celle-ci est-elle nécessaire à l'extraction des descripteurs ? Par exemple, pour un texte de 14000 mots, le graphe est construit en 1 minute et 20 secondes. Ensuite, l'analyse statistique sur les termes pour extraire les concepts avec leur sens majoritaire prend 2 minutes. Finalement, la cotation elle-même est assez rapide et prend 20 secondes. L'ensemble des algorithmes a un comportement quasi-linéaire par rapport au nombre de mots dans le document. En réalité, le calcul de la complexité est difficile à réaliser dû à l'indéterministe des algorithmes et à la taille des graphes ardue à connaître.

4.3.2 Analyse des ruptures et des retours thématiques

Comme expliqué dans le chapitre précédent, nous analysons la similarité du graphe initial (document entier) avec chacun des N sous-graphes.

Par exemple, nous découpons l'article UNIT qui traite du langage PHP[40] en 20 sous-graphes. La Figure 47, montre la similarité entre le graphe du document et chacun des 20 sous-graphes extrait. Nous remarquons quatre chutes importantes de la courbe de similarité : sous-graphes numéro 3, 10, 12 et 20. Cela s'explique très bien car ces parties du document sont des exemples de code écrit en PHP. En réalité dans le document, il existe beaucoup plus que quatre parties comportant du code mais ces quatre parties sont composées quasi-intégralement de code.

Cet exemple montre la force et la faiblesse de notre prototype pour la détection des ruptures thématiques. Nous sommes désormais convaincus de la pertinence des mesures de similarité pour la

[40] http://www.unit.eu/ori-oai-search/notice/view/unit-ori-wf-1-4291

détection des ruptures. En revanche, nous devons, à l'avenir, trouver un moyen de diviser dynamiquement le graphe non plus en nombre de parties désiré, mais en fonction des chutes et des hausses brutales de similarité. Par contre, nous pensons que cette pratique sera extrêmement difficile à mettre en œuvre (nous en rediscutons dans le chapitre Perspectives).

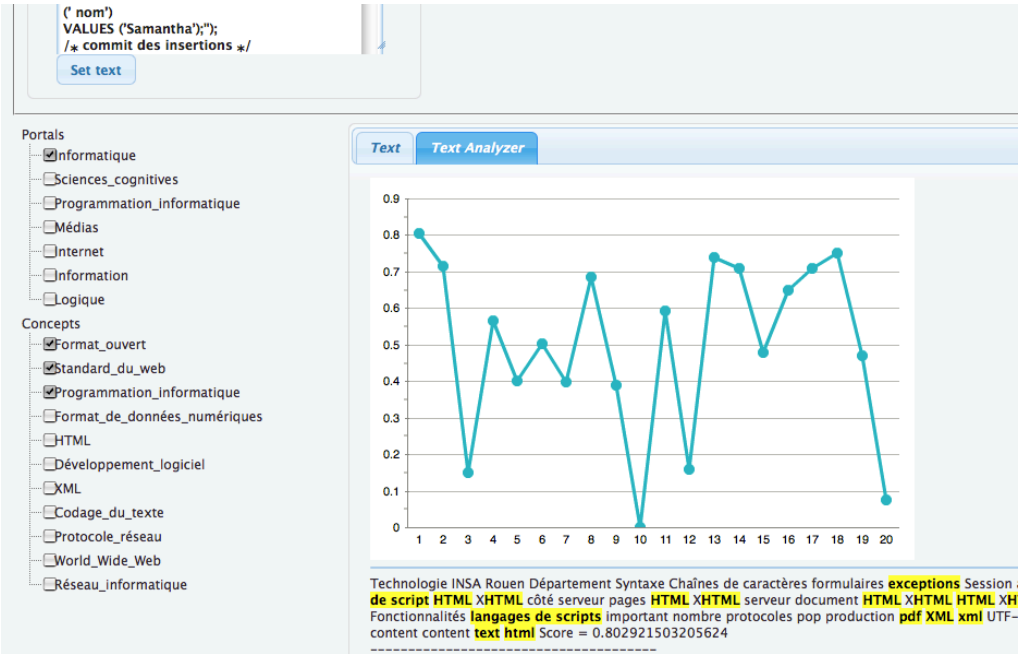


Figure 47 : Analyse des ruptures thématiques pour un cours sur le langage PHP

L'analyse des ruptures thématiques par similarité des concepts des graphes semble être efficace et confirme donc la pertinence des mesures de similarité proposées.

4.3.3 Recherche conceptuelle de documents

Pour la recherche de documents, nous utilisons les fiches LOM des documents UNIT qui présentent, entre autres, trois champs intéressants : le titre, le résumé et les mots-clés. Nous créons les graphes de documents en utilisant la concaténation de ces trois champs. Ce procédé biaise la cotation conceptuelle, mais dans le bon sens : le texte composé permet d'extraire des concepts avec une forte précision car ces trois champs fournissent déjà une indication sur le contexte.

Pour la recherche de documents,

- la requête est traitée comme un document : elle est transformée en graphe et les concepts sont cotés grâce à la cotation conceptuelle ;
- la similarité entre la requête et le document est calculée en utilisant les 10 meilleurs concepts avant désambiguïsation (pour les mêmes raisons pour les ruptures thématiques).

Notre prototype permet à l'utilisateur de désambiguïser lui-même sa requête. Par exemple, comme montre les Figure 48 et Figure 49, pour la requête « produit scalaire vecteur », l'internaute peut choisir les concepts correspondant aux termes « produit scalaire » et « vecteur ».

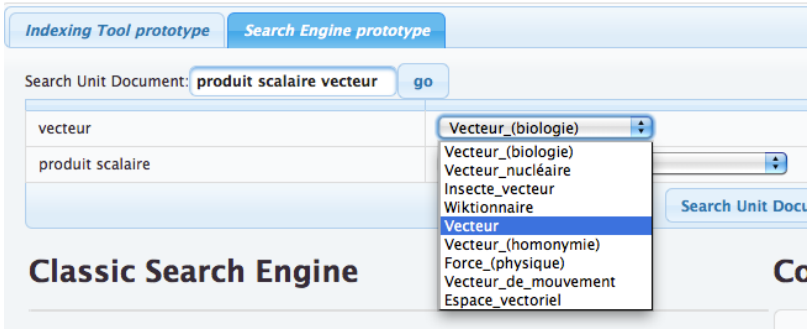


Figure 48 : Désambiguïsation manuelle du terme « vecteur »

Figure 49 : Désambiguïsation manuelle du terme « produit scalaire »

Dans un premier exemple, illustré par la Figure 50, nous exécutons la requête « clique ». Nous désambiguïsons le terme « clique » en choisissant celui traitant de la « théorie des graphes ». A gauche de l'image sont présentés les résultats obtenus en utilisant le mot simple « clique » (recherche dans les mots-clés des fiches LOM). Nous pouvons voir que les deux fiches contiennent le mot-clé « clique ». Dans la partie droite de l'image, nous avons les résultats de la recherche conceptuelle. Deux documents traitent de la théorie des graphes. Pour chaque document de la recherche conceptuelle, nous fournissons cinq informations :

- Le titre (issu de la fiche LOM) ;
- Le résumé (issue de la fiche LOM) ;
- Les mots clés (issus de la fiche LOM) ;
- Les 10 meilleurs concepts issus de la concaténation du titre, du résumé et des mots-clés.

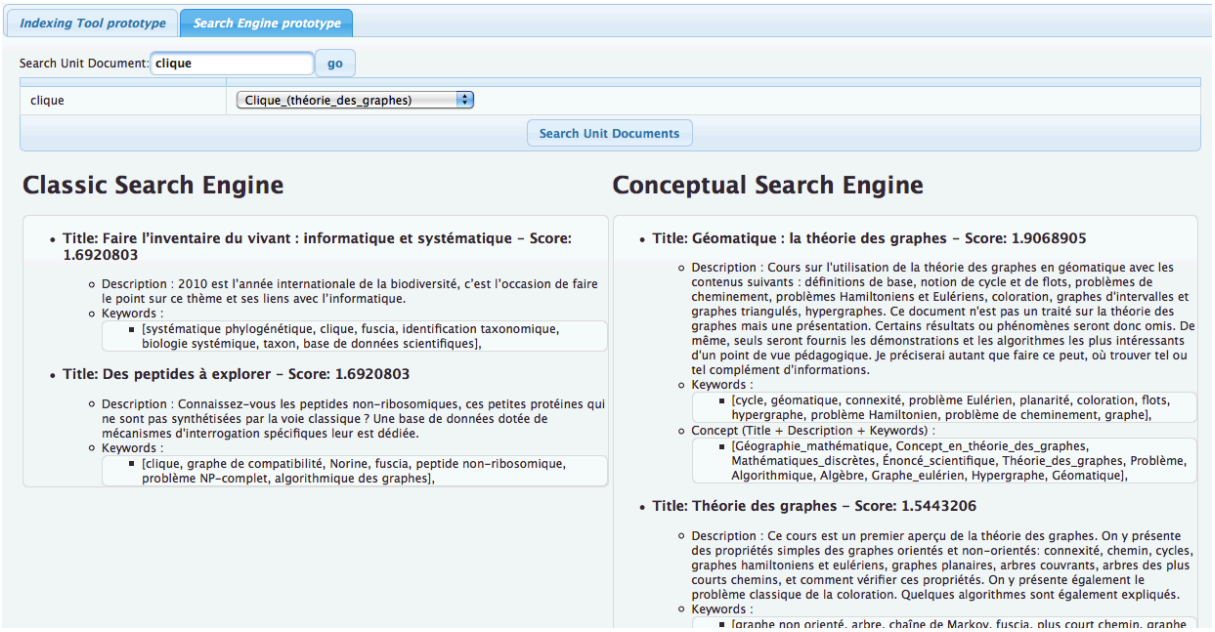


Figure 50 : Recherche des documents correspondant à la requête « clique ».

Dans un deuxième exemple, nous exécutons la requête « mal de tête ». Aucun document ne traite de ce sujet, mais la recherche fournit tout de même des résultats par similarité conceptuelle. La Figure 51 montre que les premiers documents retournés traite du cerveau. Notons que mal de tête n'est rattaché qu'à un seul concept « céphalée ». Cela signifie que le système de recherche d'information conceptuelle est capable de fournir un résultat proche des concepts de la requête, même en l'absence de ces concepts dans les descripteurs du document.

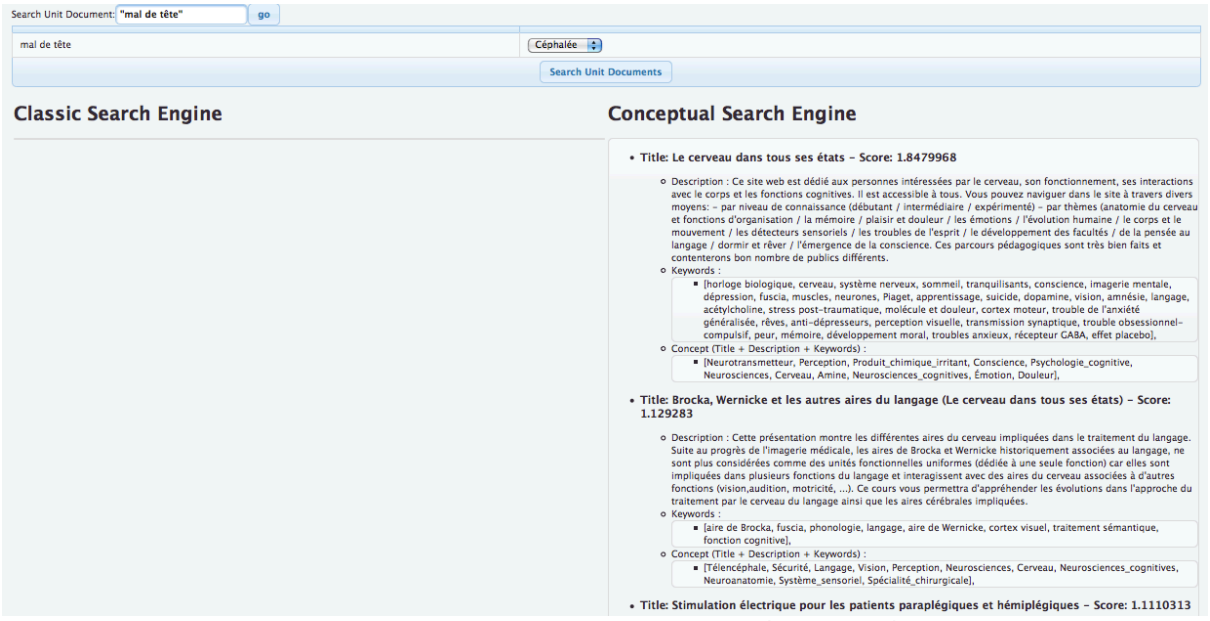


Figure 51 : Recherche des documents similaires à « Mal de tête »

Dans le dernier exemple, pour la requête « clefs étrangères », l'unique concept rattaché à cette requête est « clefs étrangères ». La Figure 52 montre que les deux premiers documents trouvés traitent de base de données géographique. Par contre, le troisième document retourné traite de système d'exploitation, qui ne correspond pas à la requête demandée.

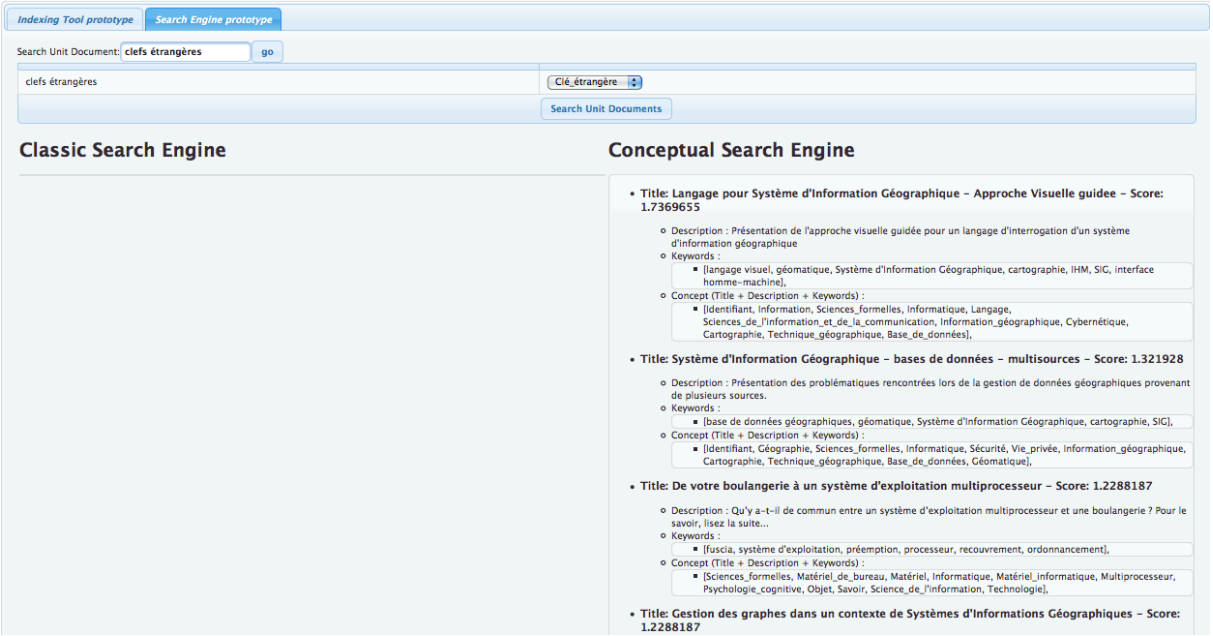


Figure 52 : Recherche des documents similaire à « Clefs étrangères »

Cet exemple permet de mettre en avant les faiblesses de notre approche. En effet, les documents et les requêtes courtes, en terme de nombres mots, génèrent de petits graphes avec peu de concepts. L'extraction est ainsi moins précise et peut contenir du bruit. A l'instar des métriques TF-IDF, il serait intéressant de normaliser le score de cotation des concepts pour la recherche d'information en fonction de la taille des graphes générés.

4.4 Avis d'un enseignant et d'un documentaliste

Dans un premier temps, nous nous sommes entretenus avec l'enseignant auteur du cours UNIT traitant du langage PHP et du protocole HTTP. Nous lui avons posé la question suivante : « Pensez-vous que notre outil peut vous aider à décrire le document ? ». Après manipulation de l'outil, l'enseignant nous a listé rapidement les manques du prototype d'après lui :

1. Les thématiques et les concepts doivent être visiblement hiérarchisés (par exemple, « Informatique » doit être au dessus de « Langage de Programmation » et d'« Internet ») ;
2. Les mots-clefs affichés en jaune donnent une bonne idée du contexte du document, mais il en manque. Les termes manquants sont polysémiques, sans sens majoritaire. Nous avons décidé

de créer un prototype sans désambiguïsation pour des raisons de temps de calcul, mais ce choix s'avère ne pas être le bon.

3. Aussi, certains termes en jaune ne devraient pas l'être : ce sont des termes des concepts sélectionnés. Si un terme traite du concept « Langage de Programmation », mais qu'il est donné en exemple (le terme « Java » dans un document traitant de « PHP »), ce dernier ne doit pas apparaître en jaune. Cela reflète encore un mauvais choix de notre part : nous n'utilisons pas la mesure de score des mots-clés dans le prototype. Nous pensions à tord qu'utiliser tous les termes reliés à concept suffirait.

Dans un second temps, nous avons rencontré le responsable de la bibliothèque de l'INSA de Rouen. Nous lui avons soumis la question suivante « Pensez-vous que ce prototype peut aider un documentaliste à indexer un document ? ». Après manipulation de l'outil, le documentaliste a dressé une liste de remarques :

1. L'outil est simple à utiliser et l'utilisateur prend vite ses marques.
2. Les thématiques proposées sont trop générales et souvent fausses. Cette largesse est volontaire de notre part car nous avons préféré choisir trop de thématiques que pas assez (en d'autres termes avoir une liste à fort taux de rappel en sacrifiant la précision).
3. La liste des concepts, après sélection des bonnes thématiques, a toujours satisfait le documentaliste pour les documents qu'il a choisis. La sélection des concepts faisant apparaître les termes en jaune est une bonne idée, mais l'outil gagnerait en lisibilité en utilisant plusieurs couleurs en fonction des concepts.
4. L'outil est très intéressant car il a l'avantage de proposer des descripteurs auxquels le documentaliste n'a pas pensé ou même ne connait pas. Lorsque le documentaliste ignorait le sens d'un concept, il cliquait dessus pour marquer en jaune les termes y faisant référence. Ce processus lui a permis de comprendre le sens des concepts qu'il ne connaissait pas.
5. Les bons descripteurs pour les documents sont les concepts. Les thématiques ou mots-clés du texte sont respectivement trop générales ou trop spécifiques.
6. Enfin, le documentaliste envisage d'utiliser ce prototype pour le référencement des thèses INSA. L'outil pourrait ainsi l'aider à trouver des descripteurs.

4.5 Conclusion

Le prototype mis en place a permis de valider trois tâches :

- l'aide à la recherche de descripteurs des documents ;
- l'analyse de la structure thématique et des ruptures ;
- la recherche d'information conceptuelle.

Le prototype semble faire ses preuves pour les tâches d'extraction des descripteurs et l'analyse thématique. Le moteur de recherche conceptuelle, quant à lui, nécessite encore des améliorations. Il permet à l'utilisateur d'être non ambigu car celui-ci peut attribuer un sens aux termes de sa requête. Les résultats de la recherche conceptuelle présentent des faiblesses dues à la petite taille des requêtes et de l'index des documents. Deux possibilités existent pour résoudre ce problème. La première est la normalisation de la métrique de cotation conceptuelle en fonction de la taille du graphe. La seconde est l'utilisation des métriques de similarité utilisant la notion de hiérarchie des concepts. Malheureusement, cette métrique ne peut pas être utilisée à l'heure actuelle pour un système de recherche d'information car trop couteuse en temps.

Ce prototype est le début d'une application réelle. Le temps de calcul est raisonnable et tous les aspects théoriques et techniques développés dans les chapitres précédents sont transparents pour l'utilisateur.

Conclusions et perspectives

Bilan

L'objectif du projet AiCoTICE, et donc de cette thèse, était de proposer et concevoir un outil d'aide à l'indexation et à la recherche de documents pédagogiques de l'entrepôt UNIT.

Les documents UNIT sont hétérogènes tant sur la forme (format et présentation) que sur le domaine traité. Pour faciliter la tâche des documentalistes, nous avons conçu un outil permettant à ces derniers d'extraire, à partir d'un document, un ensemble de descripteurs définis sur trois niveaux : les thématiques, les concepts forts et les mots-clés du texte. Les thématiques sont très générales et représentent le plus souvent le domaine principal du document (mathématiques, physique, chimie, droit, etc.). Les concepts forts sont à un niveau intermédiaire et représentent des sous-disciplines ou des catégories auxquelles le document peut se référer. Enfin, les mots-clés du texte ont la particularité de représenter succinctement les notions importantes du document.

L'extraction des descripteurs fait appel à une structure complexe transparente pour l'utilisateur : le graphe de concepts. Le système prend en entrée un document et le transforme en graphe grâce à une base de connaissances hiérarchique à partir des termes du document. Chaque terme est lié à des concepts de la base, eux-mêmes liés à d'autres concepts générant ainsi un graphe du document. Nous avons fait l'hypothèse que pour trouver les thématiques et les mots-clés d'un document, il est intéressant de trouver les concepts forts de ce document. Un concept est fort s'il est à la fois ni trop générique, ni trop spécifique et fédère un nombre important de termes du document. Nous avons proposé une heuristique de cotation des concepts, appelée cotation conceptuelle, utilisant trois propriétés du graphe : l'occurrence terminologique, la généricité conceptuelle et la diversité conceptuelle. La liste des meilleurs concepts est calculée et permet de déduire les thématiques du document (concepts parents de concepts forts du graphe) et les mots-clés (termes fils des concepts forts du graphe). Aussi, à partir de ces concepts forts, nous avons proposé de désambiguïser les termes polysémiques du document.

Les résultats de l'indexation, évalués de manière objective et subjective, sont encourageants et permettent de faire évoluer notre outil vers un système de recherche documentaire.

Les descripteurs des documents sont normalisés car ils sont issus de la terminologie de la base de connaissances hiérarchique. Or, les requêtes des utilisateurs ne le sont pas (par exemple, utilisation de synonymes en contexte ou fléchissement des termes de la terminologie). Nous avons appliqué à une requête utilisateur le même processus de création de graphe et d'extraction des descripteurs qu'aux documents indexés. Nous avons calculé la similarité entre les deux graphes représentant respectivement le document et la requête. Plusieurs métriques de similarité ont été proposées prenant en compte la construction du graphe en amont. Parmi ces métriques, deux ont retenu notre attention :

celle considérant uniquement les concepts les plus forts des graphes et celle considérant les concepts les plus forts ainsi que leur hiérarchie. Actuellement, seule la première méthode est applicable pour la recherche d'information ; la seconde, bien que plus performante, est très complexe en terme de temps de calcul.

Les applications de la similarité entre graphes de document ne se limitent pas à la simple recherche de documents ; plusieurs autres sont réalisées exploitant la similarité :

- inter-document pour concevoir un système de recommandation,
- intra-document (entre les parties d'un même document) pour un outil d'aide à l'analyse de la structure thématique du document (détection de ruptures et retours thématiques),
- pour une deuxième approche de la désambiguïsation des termes polysémiques.

Nous avons développé un prototype permettant de regrouper toutes ces fonctionnalités présentées. Il permet d'analyser un document et de suggérer à un documentaliste une liste de descripteurs, une analyse de la structure thématique d'un document ainsi que la recherche de documents. Le prototype montre la faisabilité de nos approches en un temps de calcul acceptable, ainsi que la qualité très satisfaisante des résultats retournés.

Nous pensons que l'attrait majeur de nos approches est l'exploitation de Wikipédia pour réaliser un système d'aide à l'indexation et à la recherche documentaire. En effet, en produisant les articles de Wikipédia, les internautes ont fait émerger involontairement une véritable structure de connaissances hiérarchiques. Nous avons exploité deux parties :

- le réseau de catégories pour la construction des graphes de concepts ;
- le corpus des articles pour l'analyse statistique des liens internes de Wikipédia (apportant un dictionnaire de termes et des relations termes/pages transformées en relations termes/concepts).

Des aberrations existent dans Wikipédia : le réseau de catégories contient des cycles inutiles involontairement générés. En ajoutant des catégories, l'utilisateur n'a pas une vision globale de la base et localement aucune anomalie n'est visible. Les cycles construits échappent alors au contrôle des contributeurs. Cependant, les décideurs de Wikipedia pourraient décider de supprimer ces cycles dans un futur proche.

De plus, la base de Wikipédia est en perpétuelle mutation et amélioration. L'apparition de nouveaux articles et catégories améliorera l'indexation et la recherche des documents par notre système.

Un autre attrait de nos travaux est l'utilisation conjointe de notre score de cotation conceptuelle avec Wikipédia. Cette heuristique nous fournit des scores extrêmement encourageants dans un

contexte d'aide à l'indexation et de recherche d'information. La cotation conceptuelle semble être un atout pour la recherche du contexte du document, des thématiques et des termes importants du texte. Le contexte induit par les concepts forts a notamment permis d'améliorer la mesure du Keyphraseness (métrique de recherche des liens internes de Wikipédia) en y ajoutant ce contexte.

Plusieurs améliorations peuvent être apportées à nos travaux :

1. Le premier constat est le temps important passé à la construction du graphe. Il est possible d'améliorer les performances de cette construction en optimisant les différentes opérations impliquées dans la construction (notamment l'étape d'acyclisation du graphe et l'étape de recherche des sens majoritaire des termes).
2. Le second constat concerne le nombre de descripteurs à extraire. En effet, nos méthodes utilisent un nombre de concepts forts fixé à l'avance (en général, 5 ou 10 meilleurs concepts en fonction des algorithmes utilisés). Nous pensons que ce nombre dépend de deux facteurs : la taille du graphe (en général proportionnelle à la taille du document en nombre de mots) et la diversité des thèmes abordés. Si un document est long et traite de plusieurs thématiques, il faut alors suggérer un nombre plus important de descripteurs que pour un document court ou une requête d'un utilisateur. Nos approches fonctionnent convenablement pour des documents techniques assez longs ; en revanche, pour les requêtes utilisateurs seuls quelques concepts (2 à 5 concepts) sont intéressants. Ainsi, le fait d'avoir fixé à l'avance le nombre de concepts génère du bruit (mauvaise précision).
3. La troisième amélioration possible concerne le mode de stockage et d'interrogation des graphes. En effet, si nous souhaitons utiliser la similarité exploitant la hiérarchie des concepts pour la recherche d'information, il faut trouver une manière de stocker les graphes permettant un calcul rapide de la similarité. Ceci est fait en utilisant les structures de matrices creuses pour les SRI fondés sur le modèle vectoriel ou des approches de type K plus proches voisins.

Perspectives

Nous avons volontairement occulté la phase de prétraitement utilisant des méthodes du TAL pour l'extraction des termes. Nous souhaitions nous concentrer sur l'aspect conceptuel de nos approches. Il est évident qu'un prétraitement TAL est utile, voire indispensable, pour une amélioration drastique des performances de notre système. Par exemple, nous pourrions filtrer les termes en fonction de leur nature et leur fonction grammaticale. Aussi, les termes peuvent être transformés en forme canonique, simplifiant ainsi la recherche des termes dans un dictionnaire. Nous pouvons, comme suggéré dans le chapitre 1 de la partie Contribution, utiliser l'outil FFFOR développé par (Mycek, 2011) pour détecter et traiter les formes figées et semi-figées dans un document.

Aussi, nous pensons que l'application des métriques de similarité pour la recherche de ruptures conceptuelles et thématiques est une idée à développer. Nous avons proposé une approche qui permet de découper le graphe d'un document manuellement et de calculer la similarité entre les différents sous-graphes extraits. Il serait bien plus intéressant de proposer un système de découpage automatique d'un graphe en détectant les points précis de ruptures et de retours thématiques. Nous estimons, à l'heure actuelle, cette tâche très complexe. En effet, il faut trouver le plus grand sous-graphe ayant une similarité faible avec les plus grands autres sous-graphes. Cela induit de très nombreux calculs de similarité et de construction de sous-graphes.

D'autre part, nous avons choisi de ne pas utiliser des méthodes d'apprentissages. Nous souhaitions un système utilisant uniquement un document et une base de connaissances généralistes plutôt que plusieurs documents manuellement indexés et des méthodes de classifications. L'avantage de nos approches est de fournir un outil « clé en main » sans analyse de corpus préalable et indépendant des domaines traités par ce corpus. En revanche, une fois le système d'indexation conceptuelle mis en place, nous pensons qu'introduire un système d'apprentissage exploitant les précédentes indexations serait intéressant. En effet, nos approches offrent des résultats indépendants du contexte d'utilisation et représentatifs du document mais pas de l'index souhaité par le documentaliste. Si celui-ci a choisi certains descripteurs pour indexer un document, il souhaiterait choisir les mêmes pour un document similaire. Pour ce faire, avant de fournir les descripteurs aux documentalistes, il faut calculer la similarité entre le graphe du nouveau document et les autre graphes des documents déjà indexés. S'il existe des documents similaires, alors les descripteurs proposés pourraient être étendus aux descripteurs déjà sélectionnés pour ces documents. Cette amélioration peut aussi être appliquée au système de recherche d'information, pour laquelle nous pourrions reformuler la requête d'un utilisateur en fonction des descripteurs des documents similaires.

Une toute autre perspective est d'exploiter l'aspect multilingue de Wikipédia. En effet, environ 280 sites Wikipédia différents existent selon la langue. Nous avons donc à notre disposition environ 280 bases de connaissances (une par langue). Nous pourrions donc appliquer nos approches sur des documents rédigés dans une de ces langues (en connaissant la langue du document).

En allant plus loin pour les approches multilingues d'indexation et de recherche d'information, les liens externes mettant en correspondance les articles avec leurs équivalents dans les autres langues sont encore plus intéressant. En effet, nous pouvons imaginer un système de recherche d'information où la requête est rédigée dans une langue et le résultat permet de retourner tous les documents répondant à cette requête dans d'autres langues. Nous pouvons ainsi aider les documentalistes à :

- indexer un document écrit en langue étrangère avec des descripteurs de sa propre langue (en l'occurrence le français) ;
- indexer un document en français avec des descripteurs écrits dans d'autres langues de manière automatique.

Pour nos travaux, nous avons traité des textes en langue française pour indexer un graphe de concepts. Ces concepts ont certes des labels en français, mais en théorie ces concepts ne sont que des entités de représentation et sont donc indépendants de la langue du document. L'idée est de ne plus utiliser une seule base de données Wikipédia (dans une langue donnée), mais de construire une nouvelle base unifiée et étendue des différents Wikipédia. Cette tâche est ardue car les structures des réseaux de catégories des Wikipédia sont différentes. Il faut alors concevoir une méthode d'alignement des réseaux de catégories Wikipédia.

Finalement, nous avons conçu, validé et évalué le triplet Wikipédia en français, cotation conceptuelle et documents à vocation techniques et pédagogiques. Nous pensons que le cœur de nos travaux réside dans l'utilisation de Wikipédia avec notre mesure de cotation conceptuelle. La recherche de cette mesure a été guidée et donc biaisée par l'observation des graphes de documents induits par la structure de Wikipédia en français, à l'instar des mesures de similarité sémantique guidées par WordNet. Nous devons donc valider et évaluer nos méthodes avec d'autres types de bases de connaissances hiérarchiques et pour d'autres types de documents. Un exemple possible est l'utilisation de la terminologie CISMeF[41], pour l'indexation des documents médicaux.

[41] http://www.chu-rouen.fr/cismef/

Bibliographie

Agirre, E., & Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the Web. *In Proceedings of the 18th International Conference on Computational Linguistics* (pp. 11-19).

Agirre, E., & Rigau, G. (1996). Word sense disambiguation using conceptual density. *In Proceedings of the 16th conference on Computational linguistics* (pp. 16–22).

Alemzadeh, M., & Karray, F. (2010). An Efficient Method for Tagging a Query with Category Labels Using Wikipedia towards Enhancing Search Engine Results. *In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 192-195).

Allen, J. (1995). *Natural language understanding* (2nd ed.). Benjamin/Cummings Pub. Co.

Aouicha, B. (2009). *Une approche algébrique pour la recherche d'information structurée*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *In Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference* (pp. 722–735).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley Longman Publishing Co., Inc.

Ballerini, J. P., Büchel, M., Knaus, D., Mateev, B., Mittendorf, M., Schäuble, P., Sheridan, P., et al. (1996). SPIDER retrieval system at TREC 5. *In proceedings of TREC-5* (pp. 217-228).

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (Vol. 18, pp. 805–810).

Baziz, M. (2005). Indexation conceptuelle guidée par ontologie pour la recherche d'information. *Thèse de doctorat, Université Paul Sabatier, Toulouse, France*.

Baziz, M., Boughanem, M., Aussenac-Gilles, N., & Chrisment, C. (2005). Semantic cores for representing documents in IR. *In Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1011–1017).

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of documentation*, *42*(2), 84–113.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, *35*(12), 29–38.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 154-165.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, *28*(3), 289–299.

Bookstein, A., & Swanson, D. (1974). Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, *25*(5), 312-316.

Boulogne, A., & Institut national des techniques de la documentation (Paris, France). (2004). *Vocabulaire de la documentation* (3rd ed.). ADBS éditions.

Bourda, Y. (2001). Objets pédagogiques, vous avez dit objets pédagogiques? *Cahiers GUTenberg*, *39-40*, 71-79.

Chaignaud, N., Delavigne, V., Holzem, M., Kotowicz, J. P., & Loisel, A. (2010). Étude cognitive des processus de construction d'une requête dans un système de gestion de connaissances médicales. *Technique et science informatiques*, *29*(8-9), 991–1021.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on knowledge and data engineering*, *19*, 370–383.

Clauson, K. A., Polen, H. H., Boulos, M. N. K., & Dzenowagis, J. H. (2008). Scope, completeness, and accuracy of drug information in Wikipedia. *The Annals of pharmacotherapy*, *42*(12), 1814-1821.

Cleveland, D. B., & Cleveland, A. D. (2000). *Introduction to Indexing and Abstracting:* (3rd ed.). Libraries Unlimited.

Cleverdon, C. W., Mills, J., & Keen, M. (1966). *Aslib Cranfield research project - Factors determining the performance of indexing systems : Test results*.

Coursey, K., & Mihalcea, R. (2009). Topic identification using Wikipedia graph centrality. *In*

Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 117–120).

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using encyclopedic knowledge for automatic topic identification. *In Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 210–218).

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. *In Proceedings of the 2007 Empirical Methods in Natural Language Processing* (pp. 708–716).

Daelemans, W., Van Den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, *34*(1), 11–41.

Dagan, I., Lee, L., & Pereira, F. (1997). Similarity-based methods for word sense disambiguation. *In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 56–63).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Van Dijk, T. A. (1985). Structures of news in the press. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, 69–93.

Van Dijk, T. A. (1997). The study of discourse. *Discourse as structure and process*, *1*, 1–34.

Dillon, A. (1991). Readers' models of text structures: the case of academic articles. *International Journal of Man-Machine Studies*, *35*(6), 913–925.

Dumais, S. T. (1995). *Latent semantic indexing (LSI): TREC-3 report* (pp. 219–230).

Ellis, D. (1999). *From language to communication* (2nd ed.). Lawrence Erlbaum Associates.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, *59*(10), 1662–1674.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Fogarolli, A. (2009). Word sense disambiguation based on wikipedia link structure. *In proceedings of the 2009 IEEE International Conference on Semantic Computing* (pp. 77–82).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence* (pp. 6–12).

Ganesan, P., Garcia-Molina, H., & Widom, J. (2003). Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)*, *21*(1), 64–93.

Gibbs, R. W. (1987). Mutual knowledge and the psychology of conversational inference. *Journal of Pragmatics*, *11*(5), 561 - 588.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. *In proceedings of the 1998 COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

Grandbastien, M., Huynh-Kim-Bang, B., & Monceaux, A. (2009). Knowledge framework supporting semantic search of learning resources. *Metadata and Semantics*, 259–268.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199–220. doi:10.1006/knac.1993.1008

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. *Towards very large knowledge bases: knowledge building and knowledge sharing*, *1*(9), 9.

Hafer, M. A., & Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, *10*(11-12), 371–385.

Han, E. H., & Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. *Principles of Data Mining and Knowledge Discovery*, 116–123.

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, *26*(5), 280–289.

Harter, Stephen P. (1986). *Online Information Retrieval: Concepts, Principles and Techniques*.

Academic Press.
Hatcher, E., & Gospodnetic, O. (2004). *Lucene in action*. Manning Publications.
Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 784–796.
Hearst, M. A., & Plaunt, C. (1993). Subtopic structuring for full-length document access. *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 59–68).
Hoch, R. (1994). Using IR techniques for text classification in document analysis. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 31–40).
Hutchins, W. J. (1985). Information retrieval and text analysis. *Discourse and communication: new approaches to the analysis of mass media discourse and communication* (pp. 106–125).
Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham Publishing.
Janik, M., & Kochut, K. J. (2008). Wikipedia in action: Ontological knowledge in text categorization. *In proceedings of the 2008 IEEE International Conference on Semantic Computing* (pp. 268–275).
Karov, Y., & Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational Linguistics*, *24*(1), 41–59.
Kieras, D. (1982). *Thematic processes in the comprehension of technical prose : final report*. Department of Psychology - University of Arizona.
Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. *In Proceedings of the 27th international conference on Human factors in computing systems* (pp. 1509–1512).
Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. *in Proceedings of the 5th annual international conference on Computing and combinatorics* (pp. 1–17).
Koskenniemi, K. (1983). Two-level model for morphological analysis. *Proceedings of the 8th international joint conference on artificial intelligence* (pp. 683–685).
Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, *10*(2), 115–141.
De La Passardière, B., & Grandbastien, M. (2003). Présentation de LOM v1. 0, standard IEEE. *Revue Sciences et techniques éducatives Hors série*, 211-218.
Lancaster, F. W., & Fayen, E. G. (1973). *Information retrieval on-line*. Melville.
Lancaster, F. W. (2003). *Indexing & Abstracting in Theory & Practice* (3rd ed.). University of Illinois Press.
Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, *49*(2), 265–283.
Lee, Y. K., & Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *In Proceedings of the 2002 Empirical Methods in Natural Language Processing* (pp. 41–48).
Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation* (pp. 24–26).
Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. *In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 246–254).
Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *In Proceeding of the 5th International Symposium on Online Journalism*.
Lin, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning* (pp. 296–304).
Loisel, A. (2008). *Modélisation du dialogue Homme-Machine pour la recherche d'informations: approche questions-réponses*. Thèse de doctorat, INSA de Rouen, France.
Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309–317.

Manning, C. D., Raghavan, Prabhakar, & Schütze, H. (2008). *Introduction to Information Retrieval* (1st ed.). Cambridge University Press.

Medelyan, O., Witten, I. H., & Milne, D. (2008). Topic indexing with Wikipedia. *In Proceedings of the AAAI WikiAI workshop*.

Medin, D. L. (1989). Concepts and conceptual structure. *American psychologist*, *44*(12), 1469-1481.

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. *In Proceedings of the 2007 North American Chapter of the Association for Computational Linguistics* (Vol. 2007).

Mihalcea, R., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. *In Proceedings of the 2007 Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (CIKM)* (pp. 233–242).

Mihalcea, R., & Moldovan, D. (2000). Semantic indexing using WordNet senses. *In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 35–45).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Milne, David. (2010). *Applying Wikipedia to Interactive Information Retrieval* (Thesis). Thèse de doctorat, University of Waikato, Nouvelle Zélande.

Moens, M.-F. (2000). *Automatic indexing and abstracting of document texts*. Springer.

Mooney, R. J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing* (pp. 82–91).

Müller, C., & Gurevych, I. (2009). Using wikipedia and wiktionary in domain-specific information retrieval. *In Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access* (pp. 219–226).

Mycek, P. (2011). *Développement de modules de traitement linguistique*. Mémoire de Master, Université du Havre.

Nastase, V., & Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. *In Proceedings of the 23rd national conference on Artificial intelligence* (pp. 1219–1224).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 1–69.

Ng, H. T. (1997). Getting serious about word sense disambiguation. *In Proceedings of the 1997 ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How* (pp. 1–7).

Ogawa, Y., Morita, T., & Kobayashi, K. (1991). A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy sets and systems*, *39*(2), 163–179.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web* (Technical Report). Stanford InfoLab.

Paice, C. D. (1991). The rhetorical structure of expository text. *In Proceedings of Informatics 11 Conference* (pp. 1–25).

Passardière, B. D. L., & Jarraud, P. (2005). LOM et l'indexation de ressources scientifiques Vers de bonnes pratiques pour l'Université en Ligne. *In proceedings of the 2005 Environnements informatiques pour l'apprentissage humain*, 57-68.

Pedersen, T. (2006). Unsupervised corpus-based methods for WSD. *Word Sense Disambiguation*, 133–166.

Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0*. Walter de Gruyter.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275–281).

Ponzetto, S. P., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. *In Proceedings of the 2007 national conference on artificial intelligence* (Vol. 22, p. 1440).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. *In Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 259–268).

Quillian, M. R. (1968). Semantic Memory. In M. Minsky (Ed.), *Semantic Information Processing* (pp. 227-270). MIT Press.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man and cybernetics*, *19*(1), 17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, *11*(95), 130.

Rijsbergen, C. J. van. (1979). *Information Retrieval*. Butterworth.

Rivest, R. L. (1987). Learning decision lists. *Machine learning*, *2*(3), 229–246.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for information Science*, *27*(3), 129–146.

Rocchio, J. (1971). Relevance Feedback in Information Retrieval. *The SMART Retrieval System* (pp. 313–323).

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, *17*(2), 69–76.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005a). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. *Advances in Web Intelligence*, 380–386.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005b). Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. *Natural Language Processing and Information Systems*, 67–79.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2006). From Wikipedia to semantic relationships: a semiautomated annotation approach. *In proceedings of the Third European Semantic Web Conference (ESWC 2006)*.

Salton, G. (1971). *The SMART Retrieval System; Experiments in Automatic Document Processing*. Prentice-Hall, Inc.

Salton, G., Buckley, C., & Smith, M. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management*, *26*(1), 73–92.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, *26*(11), 1022–1036.

Salton, Gerard. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

Salton, Gerard, & Buckley, Christopher. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Salton, Gerard, & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill computer science series. McGraw-Hill.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 142–151).

Schamber, L. (1996). What is a document? Rethinking the concept in uneasy times. *Journal of the American Society for Information Science*, *47*(9), 669–671.

Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems*, *7*(2), 195–207.

Scinto, L. (1983). Functional Connectivity and the Communicative Structure of Text. *Micro and macro connexity of texts*.

Smeaton, A. F. (1992). Progress in the application of natural language processing to information retrieval tasks. *The computer journal*, *35*(3), 268-278.

Sowa, J. F. (1983). Conceptual structures: information processing in mind and machine. *Addison Weshley*.

Sowa, John F. (1976). Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, *20*(4), 336–357.

Spärck Jones, K. (1973). Index term weighting. *Information Storage and Retrieval*, *9*(11), 619-633.

Spärck Jones, Karen. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

Spärck Jones, Karen, & Galliers, J. R. (1995). *Evaluating natural language processing systems: an analysis and review*. Springer.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *In Proceedings of the 20th international joint conference on Artifical intelligence* (pp. 1606-1611).

Styltsvig, H. B. (2006). *Ontology-based information retrieval*. Thèse de doctorat, Roskilde University, Denmark.

Suchanek, F. M., Ifrim, G., & Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 712–717).

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. *In Proceedings of the 16th international conference on World Wide Web* (pp. 697–706).

Syed, Z., Finin, T., & Joshi, A. (2008). Wikipedia as an ontology for describing documents. *In Proceedings of the Second International Conference on Weblogs and Social Media* (pp. 136-144).

Turtle, H., & Croft, W. B. (1989). Inference networks for document retrieval. *In Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 1–24).

Vocabulaire de la documentation. (1987). . Association française de normalisation.

Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. *in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 171–180).

Wilkinson, R., & Hingston, P. (1991). Using the cosine measure in a neural network for document retrieval. *In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 202–210).

Witten, I. H., & Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *In Proceedings of the 2008 AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA* (pp. 25–30).

Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. *In Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 18–25).

Woods, W. A. (1997). *Conceptual indexing: A better way to organize knowledge*. Technical Report of Sun Microsystems.

Wu, F., & Weld, D. S. (2008). Automatically refining the Wikipedia infobox ontology. *In Proceeding of the 17th international conference on World Wide Web* (pp. 635–644).

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133–138).

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 88–95).

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Zirn, C., Nastase, V., & Strube, M. (2008). Distinguishing between instances and classes in the wikipedia taxonomy. *In Proceedings of the 5th European semantic web conference on The semantic web: research and applications* (pp. 376–387).