ViewpointS : vers une émergence de connaissances collectives par élicitation de point de vue

Guillaume Surroca

Collège Doctoral Languedoc-Roussillon

THÈSE
Pour obtenir le grade de
Docteur

Délivrée par **l'Université de Montpellier**

Préparée au sein de l'école doctorale I2S
Et de l'unité de recherche SMILE

Spécialité : **Informatique**

Présentée par **Guillaume Surroca**

ViewointS : vers une émergence de connaissances collectives par élicitation de points de vue

Soutenue le 30/06/2017 devant le jury composé de

Dr. Stefan Trausan-Matu, Professeur des universités, Université Polytechnique de Bucarest	rapporteur
Dr. Isabelle Mirbel, Maître de Conférences, HDR, Université Côte d'Azur, Inria, CNRS	rapporteur
Dr. Claude Frasson, Professeur, Université de Montréal	examinateur
Dr. Violaine Prince, Professeur des Universités, Université de Montpellier, LIRMM	examinateur
Dr. Marie-Hélène Abel, Professeur des Universités, Université de Technologie de Compiègne	examinateur
Dr. Stefano A. Cerri, Professeur des Universités émérite, Université de Montpellier, LIRMM	directeur
Dr. Philippe Lemoisson, Chercheur, CIRAD	co-encadrant
Dr. Clément Jonquet, Maître de Conférences, Université de Montpellier, LIRMM	co-encadrant
Dr. Jean-Pierre Muller, Cadre Scientifique, HDR, CIRAD	invité

Remerciements

Ce fût une sacrée expérience d'ingénierie que de ramener la plus grande partie de nos idées dans le réel. Et cela fût très formateur pour moi-même en tant que doctorant et encadrant de stage mais aussi pour tous les stagiaires qui ont travaillé au projet. Une expérience de recherche de doctorant c'est surtout une entreprise personnelle, une expérience d'entreprenariat dans la recherche, c'est se rendre petit à petit autonome sur la façon d'entreprendre et de disséminer le projet dans lequel on est impliqué. J'ai rejoint le projet en tant que stagiaire pour apporter une première implémentation et je me suis impliqué pour laisser en « legs » à la fin de la thèse au-delà de ce manuscrit toute l'ingénierie derrière le projet. Je suis donc en charge maintenant de léger tous les chantiers de viewpoints et nos outils aux prochains étudiants curieux.

Je dois donc beaucoup à beaucoup de monde :

A mes parents pour commencer. Si j'en arrive là aujourd'hui à présenter devant un jury constitué de mes pairs chercheurs ce travail c'est bien que j'ai eu les parents les plus investis dans mon éducation. Je n'ai jamais été dans le besoin. J'ai eu peur. Ils étaient là. J'ai pu mener sans problème mon cheminement d'autodidacte grâce à eux. J'espère faire preuve un jour d'autant d'abnégation.

A mes amis. A tous mes camarades de promotion avec qui j'ai un temps cheminé. J'ai retenu quelque chose de chacun d'entre vous. Ne serait-ce que le souvenir des années à la fac'. Quand je serai vieux, je pense que je relaterai toujours cette période de ma vie.

A mes stagiaires. Chers stagiaires. Vous avez, malgré vous, pris part à une expérimentation pédagogique d'un doctorant enthousiaste. J'espère avoir été au niveau en tant que pédagogue/coach parceque vous l'avez été en tant que stagiaires. Les briques que vous avez construites ou participé à construire vont durer !

A mes collègues. Tout ceci n'arrive qu'à la suite d'une longue collaboration (5 ans !) avec Philippe Lemoisson. Ce partenariat a été pendant ces années et reste précieux. Merci Philippe de m'avoir embarqué dans cette aventure ! Et il m'a fallu apprendre beaucoup. Pour cela j'ai eu la chance d'avoir trois directeurs de thèse (Stefano A. Cerri, Clément Jonquet et Philippe Lemoisson) très engagés dans leur tâche. A l'heure qu'il est je dois admettre que je ne pourrai jamais totalement savoir ce que représentent tous les efforts qu'ils ont consacré à ma formation. Je ne peux que vous remercier pour ce que je suis devenu et ce dont je ne suis pas encore conscient. J'espère faire preuve un jour d'autant d'abnégation.

A mes anciens professeurs. Ils m'ont donné bien plus que ce qu'il ne fallait pour en arriver ou je suis. Je me rappelle avec nostalgie le lycée technique et l'implication des professeurs. Grâce à eux il rôde dans ma tête ce rêve de revenir un jour dans le lycée technique comme enseignant. Boucler la boucle en prenant mes anciens professeurs comme modèle ?

A la France. Aux écoles, aux bibliothèques, aux universités, lycées, au service public à qui je dois avec humilité reconnaître le rôle extrêmement positif des institutions de la République dédiées à l'épanouissement collectif dans ma vie.

Préface

La collaboration autour de ViewpointS commença pendant le stage recherche en seconde année de Master. C'est là que nous avons créé avec Philippe Lemoisson la première implémentation de ViewpointS. Une première version de l'API ViewpointS fût produite. Celle-ci permit de produire une preuve de concept qui fût publiée [1] ainsi qu'un prototype de moteur de recherche. Cela donnait à l'approche ViewpointS une crédibilité suffisante pour créer le sujet de thèse dont le présent document est le produit.

Contexte historique du projet

Le projet ViewpointS était le fruit de la réflexion et des échanges de Philippe Lemoisson avec Stefano A. Cerri à la suite de la thèse de Philippe [2]. Il initia alors la mise au point de ce formalisme de représentation des connaissances. La contribution scientifique de cette thèse consiste à développer ViewpointS comme formalisme, à l'évaluer en le comparant à d'autres approches d'ingénierie des connaissances et à l'opérationnaliser dans plusieurs cas d'étude afin de favoriser la dissémination scientifique du projet.

Financements

Cette thèse a été financée par un projet ANR Jeune Chercheur[1] dirigé par Clément Jonquet dans le domaine de la représentation des connaissances. Le projet SIFR[2] a pour objectif d'améliorer l'indexation et les capacités d'exploitation des documents biomédicaux français. L'objectif est d'améliorer l'indexation des documents grâce à un service d'annotation sémantique. Ce service d'annotation indexe les documents biomédicaux en se basant des ontologies.

Après avoir cherché à améliorer l'indexation le projet SIFR investigue l'une des principales problématiques en bio-informatique : la valorisation des données existantes. Le domaine biomédical est un domaine très prolifique en termes de production de résultats. L'un des challenges de la bio-informatique est de faire face à cette surabondance de productions scientifiques et de proposer des services de fouille de données apte à en retirer de nouvelles découvertes.

ViewpointS – en tant que formalisme de représentation des connaissances – se positionne d'abord comme réceptacle des connaissances créées par le service d'annotation de SIFR. Mais ViewpointS est conçu comme un formalisme intégrateur. Nous souhaitons pouvoir créer une plus-value sur les données biomédicales en intégrant les connaissances créées par les services SIFR ainsi que celles provenant d'autant de sources que possible.

[1] http://www.agence-nationale-recherche.fr/suivi-bilan/recherches-exploratoires-et-emergentes/jcjc-generalite-et-contacts/
[2] http://www.lirmm.fr/sifr/

L'originalité de ViewpointS est de proposer une approche complémentaire par rapport au raisonnement qui opère sur les ontologies. L'approche ViewpointS ne cherche pas à raisonner sur la connaissance – comme le font la plupart des approches se basant sur les ontologies – mais la représente sous forme de graphe et cherche à y appliquer des algorithmes de graphes qui se basent uniquement sur la topologie. Ce choix des méthodes topologiques par rapport à celles qui se basent sur la logique est argumenté dans l'État de l'art. Nous représentons donc des connaissances biomédicales telles que celles qui intéressent le projet SIFR dans ViewpointS et essayons d'appliquer cette originalité pour proposer des services comme la recherche d'information, la suggestion d'alignement d'ontologies ou la découverte de connaissances.

Collaborations

Autour du projet ont gravité plusieurs étudiants qui dans divers stages ont contribué à la mise au point technique de certains aspects du projet. Nous mentionnerons plus en détails l'écosystème ViewpointS dans l'annexe « La société ViewpointS ».

Abstract

Nowadays, the Web is formed by two types of content which are linked: structured data of the so-called Semantic Web and users' contributions of the Social Web. The ViewpointS approach was designed as an integrative formalism capable of mixing these two types of content while preserving the subjectivity of the interactions of the Social Web. ViewpointS is a subjective knowledge represention approach. Knowledge is represented by means of viewpoints which are micro-expressions of individual semantics tying the relation between two Web resources. The approach also provides a second level of subjectivity. Indeed, the *viewpoints* can be interpreted differently according to the user through the *perspective* mechanism. In addition to a top-down approach where collective semantics of a group is established by consensus, collective semantics of a ViewpointS community is emerging from the exchange and confrontation of *viewpoints* and evolve fluidly. In our framework, resources from the Web are tied by *viewpoints* in a *Knowledge Graph*. From the *Knowledge Graph* containing *viewpoints* and Web *resources* a Knowledge Map consisting of "*synapses*" and resources is created as a result of the interpretation and aggregation of viewpoints. The evolution of the ViewpointS *synapses* may be considered analog to the ones in the brain in the very simple sense that each *viewpoint* contributes to the establishment, strengthening or weakening of a *synapse* that connects two *resources*. The exchange of *viewpoints* is the selection process ruling the synapses evolution like the selectionist process within the brain.

We investigate in this study the potential impact of our subjective representation of knowledge in various fields: information search, recommendation, multilingual ontology alignment and methods for calculating semantic distances.

Keywords

Knowledge representation, Subjectivity, Semantic distances, Semantic Web, Social Web

Résumé

Le Web d'aujourd'hui est formé, entre autres, de deux types de contenus que sont les données structurées et liées du Web sémantique et les contributions subjectives des utilisateurs du Web social. L'approche ViewpointS a été conçue comme un formalisme creuset apte à intégrer ces deux types de contenus, en préservant la subjectivité des interactions du Web Social. ViewpointS est une approche de représentation subjective des connaissances. Les connaissances sont représentées sous forme de points de vue – des *viewpoints* – qui sont des éléments de base d'une sémantique individuelle déclarant la proximité de deux ressources. L'approche propose aussi un second degré de subjectivité. En effet, *viewpoints* peuvent être interprétés différemment selon l'utilisateur grâce au mécanisme de *perspective*. Il y a une subjectivité dans la connaissance capturée ainsi que dans la manière de l'exploiter. En complément aux approches top-down où la sémantique collective d'un groupe est établie par consensus, la sémantique collective d'une communauté ViewpointS émerge de façon « bottom-up » de l'échange et la confrontation des *viewpoints* et évolue de manière fluide au fur et à mesure de leur émission. Les *ressources* du Web sont représentées et liées par les *viewpoints* dans le *Graphe de Connaissances*. A l'utilisation, les *viewpoints* entre deux *ressources* sont agrégés pour créer une « *synapse* ». A partir du *Graphe de Connaissances* contenant les *viewpoints* et les *ressources* du Web une *Carte de Connaissances* composée de *synapses* et de *ressources* est créée qui est le fruit de l'interprétation et de l'agrégation des *viewpoints*. Chaque viewpoint contribue à la création, au renforcement ou à l'affaiblissement d'une *synapse* qui relie deux ressources. L'échange de *viewpoints* est le processus de sélection qui permet l'évolution des synapses d'une manière analogue à celles qui évoluent dans le cerveau au fil d'un processus sélectionniste neuronal.

Nous investiguons dans cette étude l'impact que peut avoir la représentation subjective des connaissances dans divers scénarii de construction collective des connaissances. Les domaines traités sur les bénéfices de la subjectivité des connaissances représentées sont la recherche d'information, la recommandation, l'alignement multilingue d'ontologies et les méthodes de calcul de distance sémantique.

Mots-clés

Représentation des connaissances, Subjectivité, Distances sémantiques, Web Sémantique, Web Social

Sommaire

Chapitre 1. Introduction

ViewpointS : vers une émergence de connaissances collectives par élicitation de points de vue

Ce premier chapitre introduit le contexte général de la thèse ainsi que les concepts et paradigmes qui nous ont inspirés. La contribution de cette thèse consiste à proposer une approche de représentation des connaissances apte à intégrer les diverses connaissances présentes sur le Web et à préserver la subjectivité de ces connaissances.

Nous présentons donc en premier lieu un historique résumé du Web depuis l'invention du protocole HTTP (HyperText Transfert Protocol[3]). Nous divisons cette histoire en trois grands âges allant de l'éclosion du Web à l'émergence du Web Sémantique en passant par l'apparition des réseaux sociaux et du Web Social. Nous nous concentrons par la suite sur la dernière évolution du Web couramment appelée Web 3.0 ou Web Sémantique.

Nous commençons ensuite une discussion sur les fondements de cette dernière phase de l'évolution du Web en matière de co-construction de connaissances partagées entre les utilisateurs humains et les machines. Les concepts autour de la découverte de connaissance – notamment celui de Sérendipité – seront abordés. Nous axons ce développement à propos de la connaissance sur la capacité pour une communauté d'agents humains ou artificiels de participer à l'émergence de nouvelles connaissances, mais aussi de les découvrir. Cette discussion amènera le lecteur à une présentation brève des notions de connaissance subjective et de connaissance interprétée. Nous dégageons la problématique de cette thèse basée sur les pistes de recherche évoquées dans cette introduction. Le chapitre se conclut sur une présentation du projet ViewpointS dont certains aspects dépassent le cadre de cette thèse.

1.1 Brève histoire du Web

Ce qu'on appelle le Web est la contraction du terme complet World Wide Web – "toile à échelle mondiale" – qui était à l'origine un réseau de ressources hypertextes permettant de consulter à distance des pages reliées entre elles par des liens HyperText. Nous allons relater synthétiquement dans la section suivante son histoire.

1.1.1 L'âge de Bronze

Il n'y aurait pas de cyberculture sans contre-culture : c'est la thèse défendue dans l'ouvrage de Fred Turner [3]. On y comprend comment, dans les années 1960, l'informatique passe des militaires aux hippies des communautés. La place de la cybernétique grandit selon les analyses de MacLuhan [4] qui est l'un des fondateurs des études contemporaines sur les médias d'aujourd'hui, le glissement

[3] https://www.w3.org/Protocols/

vers l'entreprenariat et la nouvelle économie. On retrouvera plus tard dans les communautés d'Internet l'identité relativement égalitariste du mouvement hippie des 70's. On retrouve dès 1968 dans l'œuvre de Brand Stewart – le Whole Earth Catalog[4] – un imaginaire de réseau d'articles interconnectés entre eux qui se mettait à jour régulièrement par des ajouts. On peut trouver dans cet ouvrage un caractère très encyclopédique qu'on retrouve dans Wikipédia de nos jours.

Le premier à matérialiser cette interconnexion d'articles fut un ingénieur (Tim Berners Lee) qui proposait pour le CERN[5] un protocole – HTTP (HyperText Transfert Protocole)[6] – ainsi qu'un langage de description pour page Web (HTML[7]) permettant de les relier entre eux et de naviguer dans un réseau qu'on nomma simplement plus tard le Web. On retrouve dans ce réseau l'horizontalité que promouvait la contre-culture des années 70 et la connexion de tous à tous. Jusqu'en 1993, le Web est essentiellement développé sous l'impulsion des ingénieurs en charge de la création d'HTTP.

Les choses changent avec l'apparition de NCSA[8] Mosaic[9]. Ce nouveau navigateur Web jette les bases de l'interface graphique des navigateurs modernes en intégrant les images au texte et cause un accroissement exponentiel de la popularité du web. Tous les brevets autour du Word Wide Web (WWW) furent ensuite versées dans le domaine public ce qui permit l'implication d'un plus grand nombre dans son développement. Ce réseau était matérialisé dans une architecture numérique distribuée : l'Internet, qui se bases sur TCP/IP[10], un protocole de transfert de données de bas niveau [5].

Mais la taille de ce réseau était proportionnelle au nombre de ses contributeurs, c'est-à-dire très peu avant 1995. En 1993, il existait 123 sites Web[11]. L'apparition de navigateurs comme Netscape en 1994 popularisa l'usage du Web qui augmenta le nombre de sites à 2788. Le million de sites fût atteint en 1997. Des moteurs de recherche tell que Yahoo! permettait alors de venir à bout de la masse d'information en indexant les articles du WWW. Un an plus tard le brevet de PageRank – l'algorithme de recherche proposé par Google – est déposé. Alors que la navigation sur le Web était originalement une navigation de lien en lien ces nouveaux moteurs de recherche offrent une approche globale du Web, recensent l'ensemble des pages grâce à des algorithmes d'indexation qui cheminent aléatoirement entre elles et qui les cataloguent pour la recherche.

Mais afin de s'accroître en dépassant le seul cadre des ingénieurs et chercheurs maîtrisant les langages de publication Web (HTML) et de passer au prochain stade de la diffusion de cette technologie la création de contenu devait à la fois être rendue accessible aux non-techniciens et proposer des usages motivants pour le plus grand nombre. Nous passons d'un « Web archive » centré sur les documents et les connections entre eux à un « Web Social » centré sur l'utilisateur et ses relations sociales.

[4] https://fr.wikipedia.org/wiki/Whole_Earth_Catalog
[5] Organisation européenne de recherche sur le nucléaire
[6] https://www.w3.org/History/1989/proposal.html
[7] https://fr.wikipedia.org/wiki/Hypertext_Markup_Language
[8] National Center for Supercomputing Applications
[9] https://fr.wikipedia.org/wiki/NCSA_Mosaic
[10] https://fr.wikipedia.org/wiki/Suite_des_protocoles_Internet
[11] http://www.internetlivestats.com/total-number-of-websites/

1.1.2 L'âge de Fer

La prochaine phase de l'évolution du Web fût nommée en 2004 : le Web 2.0. Il sera aussi appelé Web Social vu la nature très sociale de ce changement. Cette nouvelle évolution allait inclure une nouvelle masse d'utilisateurs, producteurs et consommateurs, en plus des ingénieurs et académiques qui avaient débuté le projet ainsi que de nouveaux types de contenus (vidéo, musique). Nous en donnons la définition suivante :

> ***Définition – Web Social, Web 2.0***
>
> *L'expression* ***« Web 2.0 »*** *désigne l'ensemble des techniques, des fonctionnalités et des usages qui ont suivi la forme originelle du* World Wide Web*, caractérisée par plus de simplicité et d'interactivité. Les internautes peuvent d'une part contribuer à l'échange d'informations et interagir de façon simple, à la fois au niveau du contenu et de la structure des pages, et d'autre part entre eux, créant notamment le Web social. L'internaute devient, grâce aux outils mis à sa disposition, une personne active sur la toile.*

C'est alors l'explosion des services d'échange instantané ou asynchrone qui donna l'impulsion initiale à ce changement. Dans les nouveaux réseaux sociaux, les blogs et fora, on pouvait créer des groupes de discussion, créer des évènements ou partager des contenus ne nécessitant aucune connaissance technique. La démocratisation de la création de contenus sur le Web fut facilitée par l'évolution des technologies permettant l'amélioration de l'ergonomie des interfaces telles qu'Ajax[12]. Toutes ces fonctionnalités très banales de nos jours visant à améliorer le quotidien du plus grand nombre permirent la popularisation de l'usage du Web. La plateforme MSN avec son client Windows Live Messenger[13], en 1995, était un précurseur dans le marché sur ce chemin de l'appropriation du Web par les masses. Ceci dit, le boom de l'utilisation de ces services du Web Social est également dû à la démocratisation de l'accès au matériel informatique et à une connexion Internet. D'ailleurs, MSN est le lointain descendant d'un système d'enseignement à distance nommé PLATO développée dans les années 1960. PLATO était une plateforme d'apprentissage à distance qui offrait plusieurs moyens (chats, fora, visualisations etc). La Figure 1 montre une visualisation offerte par PLATO pour l'apprentissage de la chimie.

[12] https://fr.wikipedia.org/wiki/Ajax_(informatique)
[13] https://en.wikipedia.org/wiki/Windows_Live_Messenger

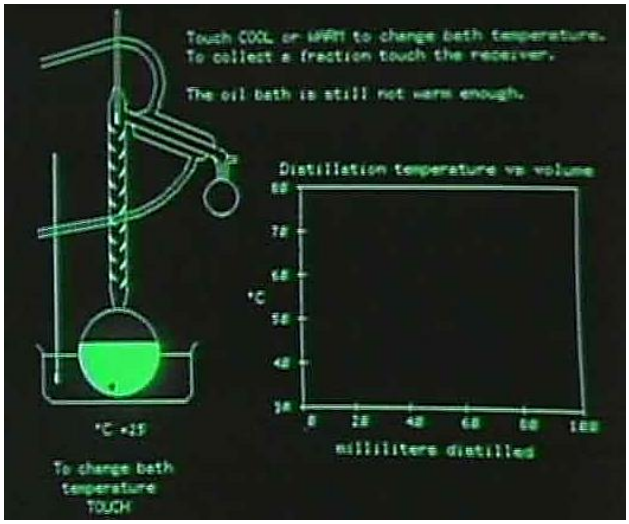


Figure 1 Illustration du système PLATO (documentaire BBS: The Documentary de Jason Scott[14]).

Cette nouvelle dimension sociale du Web étendait les modes de navigation. Il était possible de naviguer de page en page, il était possible de faire une recherche sur l'ensemble du Web. L'interconnexion sociale représentée dans les réseaux sociaux permet maintenant aussi une navigation dans l'environnement social. Et les résultats de cette navigation sont le produit d'une indexation classique et du réseau social (partages, feedbacks). Il s'agit d'une nouvelle façon de filtrer l'information sur le Web. Le boom des usages du Web et des services que nous mentionnons s'incarne aussi dans la façon dont les connaissances se créent par l'effort collectif de masse. De plus en plus de sources démontrent le nombre croissant de références à Wikipédia dans la presse académique la complétude de Wikipédia. Il existe par exemple des études sur des sujets spécifiques comme les médicaments [6] ou celle de Temin Kim Park [7] sur la visibilité des publications de Wikipédia sur Google Scholar. Ainsi, la transition principale du Web des documents aux Web Social est celle d'une archive documentaire à une communauté qui construit sur ses interactions.

La navigation et la découverte des connaissances qui s'opéraient soit dans une approche la plus locale avec la navigation de site en site soit dans une approche globale de recherche sur tout le Web s'enrichi d'un nouveau mode de navigation. En effet, le Web Social comble l'espace entre ces deux « extrêmes » en permettant à l'utilisateur de découvrir de la connaissance dans son voisinage social (un ami ayant partagé un lien sur le mur de son réseau social favoris) et de naviguer sur le Web de voisinage social en voisinage social.

Mais l'évolution des services qui ont permis cette popularisation de l'usage du Web encouragea le développement d'une nouvelle couche du Web. Les services devaient disposer d'une représentation au moins partielle de la connaissance humaine afin de raisonner sur cette connaissance pour s'améliorer. Une grande entreprise d'élicitation des connaissances devait permettre cette évolution. Cela mena à l'ajout d'une nouvelle couche dans le Web : le Web sémantique. Autour de l'utilisateur – central dans le paradigme du Web Social – les connaissances se structurent aussi autour des Concepts – brique élémentaires de sens – qui permettent aux machines de raisonner dessus.

[14] https://en.wikipedia.org/wiki/BBS:_The_Documentary (5min8s)

1.1.3 L'âge d'or : Vers une interconnexion maximale dans le Web

Les développements en Intelligence Artificielle sur la représentation des connaissances permettaient la création de services se basant sur la connaissance en opérant des raisonnements.

> ***Définition – Web Sémantique***
>
> *Le **Web sémantique** est une extension du Web standardisée par le World Wide Web Consortium. Cette nouvelle couche du Web enrichit la toile qui relie les pages et d'un nouveau type de relation de sens – des relations sémantiques – capables d'être exploitées par les nouvelles applications de l'Intelligence Artificielle dans la traitement des connaissances.*

La connaissance était alors représentée dans une ontologie. Une ontologie est une structuration de la connaissance humaine mettant en relation des concepts avec des relations de sens, des relations sémantiques. T. Gruber définit les ontologies comme « la spécification d'une conceptualisation » [8].

En prenant du recul sur l'évolution du Web on la voit comme une quête permanente d'une plus grande interconnexion entre ses éléments (pages Web, médias, concepts etc.). Au début, les pages Web – *des supports de connaissance* – étaient liés par les liens HTTP. Ensuite, à la fois les utilisateurs et les nouveaux algorithmes de fouille et d'indexation – tous deux *fournisseurs de connaissance* – devinrent des éléments centraux par la démocratisation du Web. Le maillage social tissé par ces fournisseurs de connaissance venant des blogs ou fora constitua une nouvelle couche d'interconnexion dans le Web ainsi que la propagation d'un nouveau type de ressource : le *descripteur de connaissance*.

> ***Définition – Ontologie***
>
> *En philosophie, l'**ontologie** (de* onto-*, tiré du grec* ὤν, ὄντος *« étant », participe présent du verbe* εἰμί *« être ») est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe. L'ontologie représente l'ensemble des connaissances d'un domaine en reliant ses concepts par des relations sémantiques telles que « partie de » entre le concept Roue et Voiture ou « spécialisation de » entre Humain et Mammifère.*

Les pages Web pouvaient alors être liées entre elles par des relations sociales (Ex. : Si un même individu « like » deux pages sur Facebook, il y a alors une relation indirecte entre ces pages). L'arrivée du Web Sémantique allait poser la dernière pierre à l'édifice en connectant les ressources du Web d'un réseau de concepts et de relations sémantiques. La Figure 2 résume cette évolution dans les termes que nous emploierons dorénavant dans notre argumentation : supports, fournisseurs et descripteurs de connaissance.

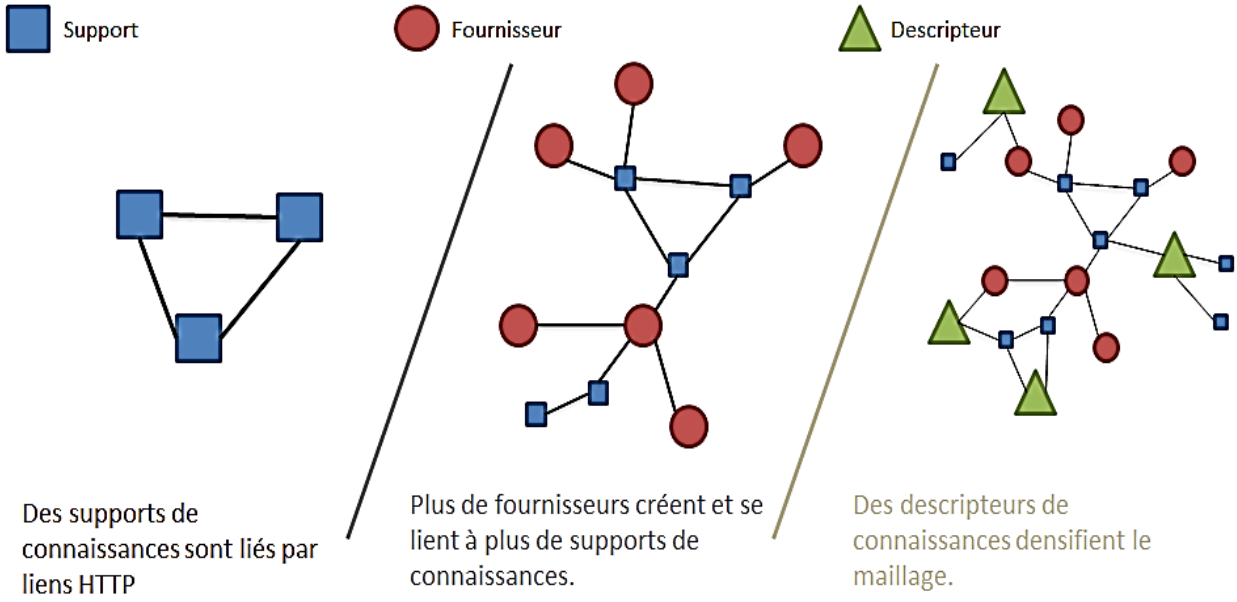


Figure 2 Illustration de l'évolution du Web

1.2 Découverte de la connaissance

Comment découvrir de nouvelles connaissances, de nouveaux liens ? Sur quels types de raisonnement peut-on se baser pour découvrir de nouvelles connaissances à partir de la connaissance existante ? Ces savoirs pratiques se caractérisent notamment par la combinaison de l'expérience et de l'information et permettent d'appréhender la singularité des situations

Ce sont des pratiques abductives, au sens où l'on adopte des hypothèses plausibles susceptibles d'être vérifiées ultérieurement. Charles S. Peirce, philosophe et logicien américain, voyait l'abduction comme « un aperçu créatif – *« creative insight »* – pour résoudre un problème surprenant, une expérience qui déçoit une anticipation, ou un évènement qui entame une habitude ». Alors que l'induction va du cas expérimental à la règle, l'abduction va de la règle au cas. C'est la recherche des causes, ou d'une hypothèse explicative. Nous pratiquons l'abduction dans la vie courante, lorsque nous recherchons les causes d'un phénomène ou d'un fait surprenant.

> ***Histoire scientifique – Quelques découvertes surprenantes***
>
> *Le four micro-ondes fut le produit d'un incident impliquant le magnétron d'un radar et une barre chocolatée. Les magnétrons devinrent plus petits et on utilisa leur capacité chauffante.*

Charles S. Peirce a introduit la notion d'abduction en épistémologie en reprenant les trois types de raisonnement proposés par Aristote : « étant donné un fait B et la connaissance que A implique B, A est une abduction ou une explication de B ». Dans le processus de construction du savoir, l'abduction guide l'induction, elle est un moment préalable de l'induction. Mais seule l'abduction est créative et apporte de nouvelles connaissances, bien qu'elle soit imprévisible et incertaine, et en cela très proche de la Sérendipité.

Le conte très ancien « Voyages et Aventures des trois Princes de Serendip » dont s'est inspiré l'écrivain anglais Horace Walpole pour forger le mot serendipity (la faculté de « découvrir, par hasard et sagacité, des choses qu'on ne cherche pas ») illustre en effet un processus épistémologique très proche de l'abduction.

> ***Histoire scientifique – Le conte de la Sérendipité***
>
> *Les trois princes de Serendip, voyageant pour s'instruire, rencontrent en chemin un chamelier qui leur demande s'ils n'auraient pas vu, « par hasard », un de ses chameaux égaré. Les princes le lui décrivent sans hésiter : « N'est-il pas borgne ? Ne lui manque-t-il pas une dent ? Ne serait-il pas boiteux ?» Le conducteur ayant acquiescé, c'est donc bien son chameau qu'ils ont trouvé et ont laissé loin derrière eux. Par la suite, le chamelier ayant cherché en vain son animal et pensant avoir été volé, les trois frères sont arrêtés et jugés. C'est alors qu'ils démontrent comment des indices observés sur le sol leur ont permis de reconstruire (par abduction) l'aspect d'un animal qu'ils n'avaient jamais vu.*

Le lecteur pourra trouver en section 2.2 une discussion sur le rôle que la Sérendipité joue dans la découverte des connaissances et quelles leçons on peut en tirer dans les services comme les moteurs de recherche.

1.3 Partage de connaissances

La connaissance que le lecteur tire de ce manuscrit est le résultat d'une double subjectivisation : une première subjectivité – la mienne comme auteur quand je donne mon point de vue – et la vôtre (cher lecteur) quand vous appliquez votre perspective. La Figure 3 illustre ce processus communicationnel. La connaissance (subjective) est d'abord encodée par le destinateur puis décodée par le destinataire selon sa grille de lecture (subjective).

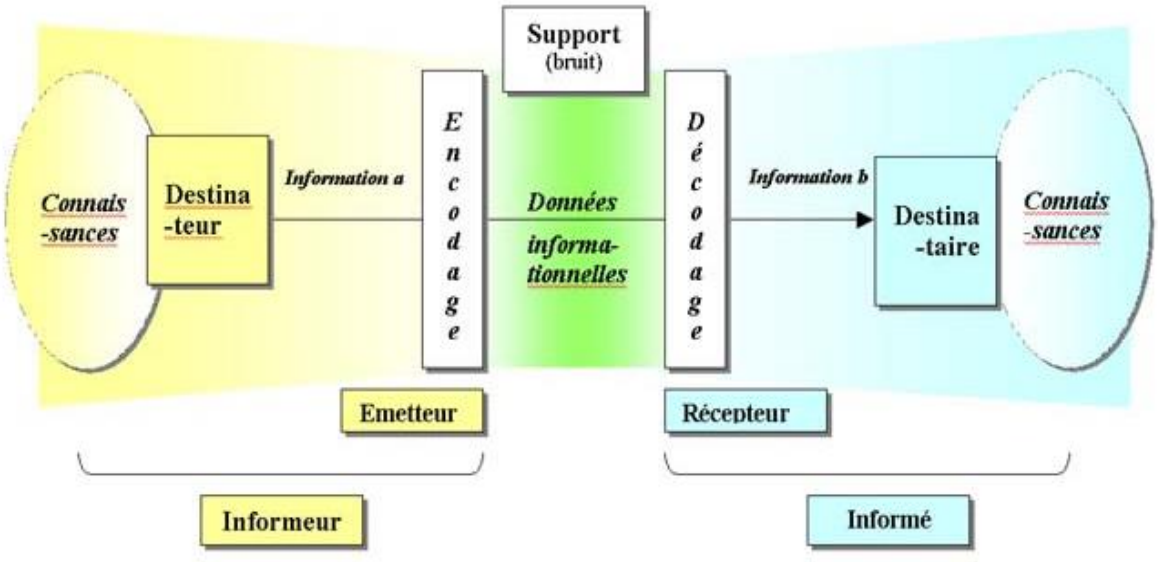


Figure 3 Processus de subjectivisation de la connaissance [9]

Pour donner un exemple de cette chaîne prenons l'exemple de la subjectivisation de l'information sur les sondages. Un sondage en lui-même est une donnée brute – une compilation de statistiques démographiques – mais un journaliste qui publie en appuyant son article sur ce sondage apportera un premier degré de subjectivité dans la présentation, la mise en forme des « données informationnelles ». L'article est ensuite lu en fonction de la sensibilité du lecteur (sa subjectivité). La connaissance que chaque lecteur obtient est le fruit de cette double subjectivisation : celle du lecteur et celle du journaliste. Dans cette chaîne de l'information aussi bien le journaliste que le lecteur adopte des points de vue. Nous intégrons dans la conceptualisation de ViewpointS cette double subjectivité : celle de celui/celle qui émet la connaissance (via le viewpoint) et celle de celui qui la reçoit (via la perspective).

Cette pensée se place dans l'héritage du modèle STROBE [10] proposé par Clément Jonquet et Stefano A. Cerri. On retrouve dans STROBE la subjectivité dans l'interprétation des connaissances. Il s'agit d'un modèle incluant l'agent dans différents environnements conversationnels uniques permettant permettant un traitement contextuel des connaissances. « Ces Agents sont capables d'interpréter des messages dans des environnements donnés incluant un interpréteur qui apprend de la conversation et donc qui représente l'évolution de sa connaissance au niveau méta. »

1.4 Problème abordé dans la thèse

Notre réflexion centrale dans le cadre de cette étude porte sur la captation, la représentation et l'exploitation de la subjectivité des connaissances. Notre objectif est double : (i) prendre en compte la subjectivité des connaissances à l'intérieur des briques de base de l'information afin d'en préserver toute la richesse et (ii) permettre à l'utilisateur d'interpréter et d'agréger les briques de base selon sa propre subjectivité pour exploiter à son profit les connaissances stockées. Il y a donc une double subjectivité : celle de l'émetteur de l'information et celle de l'utilisateur qui interprète ces informations.

Notre état de l'art est motivé par le questionnement suivant : Dans l'état actuel des formalismes de représentation des connaissances quel est le degré de subjectivité de l'information qui est conservé ?

A partir de notre constat, nous mettrons au point un formalisme de représentation des connaissances apte (i) à retenir au mieux cette subjectivité et (ii) offrir de nouvelles fonctionnalités ou en améliorer certaines existant déjà en exploitant la subjectivité. La mise au point de ce formalisme ainsi que les diverses expérimentations qui suivent dans cette thèse permettront d'aborder la question suivante : Quel est l'impact de la préservation de la subjectivité de l'information et de son exploitation, aussi subjective, en terme de recherche d'information, de capacité de recommandation et d'apprentissage collectif ?

1.5 Plan

Nous entamons cette dissertation par une étude de l'état de l'art dans laquelle nous inspecterons (i) les avancées dans les formalismes de représentation des connaissances, (ii) les mesures sémantiques exploitant ces formalismes, (iii) une réflexion sur la notion de point de vue et de subjectivité ainsi que (vi) un développement sur la découverte des connaissances. C'était la base qui a nourri notre réflexion dans la conception du formalisme qui se construit sur (i) les forces et faiblesses des approches de représentation des connaissances existantes, (ii) des méthodes pour exploiter ce nouveau formalisme afin de donner de nouveaux services aux utilisateur favorisant la découverte de connaissances. Le mécanisme de perspective que nous proposons se base sur (iii) la préservation de

la subjectivité de la connaissance dans tout son parcours de la donnée brute à la connaissance interprétée. Nous nous positionnons aussi par rapport aux deux trajectoires de l'élicitation des connaissances : la construction consensuelle de l'intelligence collective ou son émergence venant de la confrontation des interactions entre agents. Ce tour d'horizon que nous proposons dans le Chapitre 2 nous permet de commencer à spécifier le formalisme que nous présentons dans le Chapitre 3. Dans ce chapitre central nous introduisons tout le vocabulaire ViewpointS – les objets, méthodes – ainsi que l'implémentation de cette approche dans une API (Interface de Programmation Applicative) permettant de le rendre opérationnel. Nous y décrivons les structures de stockage pour les connaissances dans l'approche Viewpoints ainsi que les choix architecturaux. Nous expliquons le rôle de la double clé de voute de l'architecture de Viewpoints que sont les viewpoints et les perspectives. L'ensemble des méthodes qui exploitent la topologie de connaissances subjective que nous proposons y sont aussi détaillées.

Le Chapitre 4 décrit les expérimentations qui ont été menées et les réflexions qui en sont le produit et qui nous ont permis d'évaluer l'approche. Nous nous basons sur le vocabulaire développé dans le chapitre précédent pour proposer au lecteur plusieurs mises en situation de l'approche ViewpointS. Nous abordons plusieurs jeux de données aux structures différentes afin de démontrer (i) la capacité d'apprentissage du graphe de connaissances, (ii) l'efficacité des méthodes exploitant ce graphe (voisinage et distance sémantique) et (iii) l'opérationnalité de l'approche grâces aux prototypes développés.

Pour finir, le Chapitre 5 résume toute la contribution scientifique et d'ingénierie de cette thèse par rapport au projet ViewpointS. Nous examinerons ce que cette thèse laisse comme outils mais aussi comme opportunités, pistes de recherches, à la prochaine génération d'étudiants, de chercheurs ou d'ingénieurs qui contribuera au projet. Nous faisons état du prototype de moteur de recherche ViewpointsWebApp ainsi que l'API qui permet l'implémentation du modèle ViewpointS. Nous synthétisons les résultats précédents dans le Chapitre 6 afin de conclure cette thèse.

Chapitre 2. État de l'art

Nous nous posions la question dans le chapitre précédent : Dans l'état actuel des formalismes de représentation des connaissances quel degré de subjectivité de l'information peut être conservé ? Comment cette subjectivité peut être exploitée? Ce chapitre – en cherchant à répondre à cette question par un état de l'art pluridisciplinaire – introduit les travaux qui seront présentés dans la suite de cette thèse en présentant le contexte et le positionnement scientifique de l'approche ViewpointS. Nous commençons par présenter les divers champs de recherches qui ont inspiré nos travaux. Nous abordons ensuite chacun d'entre eux dans une partie qui lui est dédiée. Chacune de ces parties propose une description détaillée des approches qui font l'histoire des domaines de recherche impliqués et conclue en listant les verrous scientifiques et les pistes de recherche sur lesquelles nous espérons contribuer. La thématique de nos travaux nous a d'abord intéressés à la découverte des approches de représentation des connaissances de la littérature. C'est pourquoi nous proposons en premier lieu un état de l'art sur le domaine de l'ingénierie des connaissances et le rôle qu'a pris le domaine de l'intelligence artificielle dans la création d'approches qui permettent à la machine d'exploiter la connaissance numériquement élicitée et représentée. L'exploration de notre état de l'art amènera ensuite le lecteur à la découverte des méthodes inventées pour exploiter cette connaissance représentée. Nous verrons qu'elle joue un rôle essentiel dans le processus cognitif humain. Nous discutons ensuite de la notion de points de vue qui est, pour nous, essentielle afin de représenter la subjectivité des échanges et des connaissances du Web. Pour terminer, nous allons examiner comment cette subjectivité est exploitée dans la diversité des systèmes de recommandation.

2.1 Défi de l'élicitation des connaissances

Définition – Intelligence Artificielle

L'Intelligence Artificielle (IA) est un domaine en informatique qui traite de la résolution de problèmes complexes nécessitant de fortes capacités cognitives (l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique, la planification) pour substituer dans certains contextes l'homme par la machine.

Le Web n'a jamais contenu autant de données depuis les débuts de l'âge de Fer. Le problème posé par cette surabondance est de transformer ce « Big Data » en « Big Knowledge ». Pour cela, l'IA a, dans son histoire, donné de nouveaux outils aptes à acquérir et tirer parti de connaissances. Ce domaine rassemble des scientifiques de divers horizons dans le but de permettre à la machine de percevoir notre environnement, d'y apprendre des connaissances, de raisonner et de planifier à partir de ces connaissances, de les communiquer et de les mettre à jour.

Une des raisons qui a motivé le développement de cette discipline est la limitation que représentait le stockage de connaissances représentées « en extension ». Tout devait être élicité. Par exemple : Jean est mon père. Isabelle et Henri sont les parents de Jean et Isabelle et Henri sont les grands-pères de Guillaume. C'est ainsi qu'on peut représenter en extension une partie de mes liens de parenté proches. Mais si avec l'outil logique je pouvais économiser l'espace de la connaissance de «Pierre et Henri sont les grands-pères de Guillaume » en le déduisant à partir de trois informations ? Il y avait donc aussi un challenge du stockage des connaissances en partie derrière les débuts du domaine de l'IA. Nous nous intéressons en particulier – parmi les divers champs de recherche de l'IA visant à synthétiser l'intelligence des créatures du vivant – à la représentation des connaissances. Une succession de contributeurs dans le domaine de l'IA se sont posés la question suivante : Comment représenter et rendre intelligible la connaissance humaine pour un ordinateur[11] ? Comment synthétiser la connaissance humaine en s'inspirant des procédés neuropsychologiques de notre cerveau [12] ? L'ingénierie des connaissances (IC) est le domaine qui se pose ces questions sur la représentation des connaissances et cherche à recourir à d'importantes sources de connaissances afin de répondre à des problèmes complexes tels que le diagnostic médical.

Dans cette section, nous évoquons en premier l'histoire de l'ingénierie des connaissances. Nous abordons ensuite quelques exemples d'application qui tirent bénéfice de bases de connaissances produites dans le domaine de la représentation des connaissances. Dans la section 2.1.2 nous présentons quelques exemples de représentations des connaissances pour le raisonnement et leurs limitations. Nous montrons ensuite quelques exemples d'une représentation des connaissances intégrant l'agent puis la notion de «point de vue ».

2.1.1 Histoire de l'ingénierie des connaissances

Un part importante du travail de la communauté IC a été premièrement de représenter les connaissances humaines dans un grand nombre de domaines pour qu'elles puissent être traitées par les algorithmes. Mais qui élicite cette connaissance et pourquoi ?

Les systèmes à base de connaissance trouvent application dans divers domaines. La représentation des connaissances permet premièrement une meilleure indexation des données [13] car elle prend en compte les relation sémantiques entre documents grâce aux ontologies. En effet l'annotation, c'est-à-dire le processus de rattacher des concepts d'ontologie dans les documents ou autres ressources permet de rattacher les ressources du Web au Web Sémantique. Par exemple, le projet SIFR[15] qui soutient financièrement cette thèse a comme sujet central l'indexation sémantique des ressources biomédicales afin d'améliorer leur recherche. L'annotateur français de SIFR[14] utilise des méthodes statistiques de Traitement Automatique du Langage Naturel (TALN) afin de repérer les mots-clés les plus représentatifs de la ressource que nous souhaitons indexer puis fait correspondre ces mots-clés avec des concepts dans des ontologies biomédicales françaises telles que MeSH français[16] produit par l'Inserm qui est une traduction de MeSH anglais.

Si l'indexation sémantique permet d'améliorer la recherche d'information ce n'est pas le seul apport du Web Sémantique à la Recherche d'Informations (RI). Il permet aussi de compléter des requêtes incomplètes et/ou aussi de retrier les résultats de manière plus efficace [15], [16].

[15] http://www.lirmm.fr/sifr/
[16] http://MeSH.inserm.fr/MeSH/

De la même manière que l'IC ouvre des voies d'amélioration en RI les systèmes de recommandation peuvent tirer les mêmes bénéfices comme ces exemples de systèmes de recommandation pour l'apprentissage[17], [18].

Les communautés IC ont envisagé diverses approches complémentaires pour organiser collectivement cette connaissance. Des représentations très structurées sont constituées par consensus par des cercles d'experts, acteurs de la construction du Web sémantique (ex., les ontologies[19] ou les données liées[20]). Une ontologie[21],[22] est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. C'est le cas par exemple pour certaines ontologies dans le domaine biomédical nécessitant de l'expertise comme la GeneOntology[23]. Il s'agit d'un projet de bio-informatique inscrit dans la démarche plus large d'Open Biomédical Ontologies[17] visant à représenter nos connaissances génétiques actuelles. Le domaine biomédical est particulièrement riche d'ontologies telles qu'UMLS (Unified Medical Language System), MeSH (Medical Subjects Headings) ou MedLinePlus[18]. D'autres domaines nécessitant une expertise comme l'agronomie ont représentés les connaissances qu'ils ont récoltées sur les phénotypes des plantes dans la CropOntology[24].

Toutefois l'élicitation de la connaissance humaine pour permettre à l'IA de la traiter est une très grande entreprise qui nécessite la participation du plus grand nombre. Le Web Social (ou Web 2.0) a ouvert l'accès de masse à la création de contenus sur le Web. Il s'agit d'une opportunité clairement identifiée par Donan ou Quinn dans leurs réflexions sur les perspectives du crowdsourcing aussi désignée comme « Distributive Human Computation » [25], [26].

> ***Définition – Crowd sourcing***
>
> *Le crowdsourcing – ou myriadisation en français – fait appel au grand nombre en demandant des micro-contributions. L'étude de la myriadisation est un domaine émergent de la représentation des connaissances.*

La participation peut être soit active dans le cas où les contributeurs participent activement à l'élicitation de leurs connaissances, soit passive. Un exemple de contribution passive est une recherche sur un moteur de recherche. Google peut donc observer « une relation étroite entre le nombre de personnes faisant une recherche sur le mot et le sujet « grippe » et le nombre de personnes ayant des symptômes grippaux ». Le nombre de recherches sur le mot grippe augmente en période d'épidémie et de la part des gens grippés[27], ce qui peut intéresser les épidémiologistes ou l'OMS.

Toutefois, James Surowiecki revient dans son livre sur la sagesse des foules[28] sur deux conditions préalable au bon déroulement de la myriadisation se basant sur la contribution active : (i) l'indépendance des contributeurs et (ii) la diversité de la foule. Qui plus est les éléments de la foule doivent posséder un savoir minimum et pertinent par rapport à l'objectif de cette myriadisation.

[17] http://www.obofoundry.org/

[18] L'ensemble des ces ressources sont disponible sur le portail de la Librairie National de Médecine américaine https://www.nlm.nih.gov.

Les plateformes du Web Social appartenant à ce que nous appelions précédemment « l'âge de fer » du Web ont donné le moyen à des communautés d'éliciter par micro-contributions leurs connaissances. Certaines comme JeuDeMots[29] utilisent le jeu comme ressort de motivation pour la construction d'une connaissance partagée sur la langue française. La catégorie la plus représentée de ces plateformes du Web 2.0 contient les systèmes de « Social Bookmarking » – ou systèmes de partage de signets – qui permettent aux internautes d'associer des tags (descripteurs de connaissances) à divers types de ressources. Un tag (ou étiquette, marqueur, libellé) est un mot-clé ou terme associé ou assigné à de l'information (par exemple une image, un article, ou un clip vidéo), qui décrit une caractéristique de l'objet et permet un regroupement facile des informations contenant les mêmes mots-clés. Last.fm[19] donne les moyens à ses utilisateurs de structurer la connaissance musicale en catégorisant les musiques par tags. Flickr[20] quant à lui nous permet de catégoriser nos photos de vacances ou autres images de la même manière. Ce mode de catégorisation est communément appelé « folksonomie».

> ***Définition – Folksonomie***
>
> *Le mot est un néologisme produit de la combinaison des mots « folk » (le peuple) et « taxonomy » (taxonomie). Une taxonomie est un arbre hiérarchique de concepts.*

Contrairement à l'élicitation de connaissances par consensus d'experts les folksonomies et les autres modes de construction collaborative de connaissances qui viennent du Web Social produisent la connaissance par émergence. Comme le disaient Aberer et al. dans leurs travaux sur l'émergence de sémantique collective[30] le sens émerge des interactions entre utilisateurs d'une communauté. Par exemple, si nous taggons plusieurs documents, films et autres ressources du Web par le tag « IA » alors nous contribuons avec d'autres utilisateurs à donner progressivement du sens au concept d'IA. L'utilisation du langage dans divers contextes donne donc du sens aux mots. L'approche par émergence donne une sémantique collective évoluant de manière très fluide au gré des usages du langage et la démocratisation de l'utilisation des services du Web 2.0 permet une évolution rapide de cette connaissance par l'abondance des données récoltées. L'évolution de ce Web de connaissance tient de la plasticité du cerveau. Bernstein évoquait d'ailleurs le Web comme un cerveau global ; le « Global Brain »[31]. Nous aborderons cette inspiration neurologique et « l'émergence de sentiers » dans la connaissance dans une autre section. La construction de folksonomies par myriadisation a donc plusieurs avantages par rapport à l'établissement par consensus d'experts d'ontologies : car c'est un moyen (i) peu coûteux, (ii) sur lequel tout le monde peut contribuer et voir les contributions des autres et (iii) à l'évolution rapide et fluide.

[19] http://www.last.fm/
[20] https://www.flickr.com/

Définition – Pragmatique (linguistique)

*La **pragmatique** est un champ de recherche de la linguistique qui étudie les façons dont le contexte contribue à la signification des termes.*

Toutefois le vocabulaire des folksonomies est librement choisi par les utilisateurs et cela pose plusieurs problèmes. Premièrement il existe des problèmes de sémantique que les folksonomies ne proposent pas de désambiguïser. L'absence de groupe standard de tags (ex. : singulier ou pluriel, casse), de relations sémantiques entre tags ou le mauvais étiquetage dû aux fautes d'orthographe provoque une multiplication de tags identiques en signification ou de tags utilisés dans des contextes différents (ex : glacier) donc polysémiques.

Ces deux approches de construction du Web Sémantique correspondent à deux natures du Web qu'on pourrait caractériser comme : le Web Computationnellement Sémantique et le Web Cognitivement Sémantique[32]. D'une part, le web computationnellement sémantique correspond à un web en grande partie traité par des agents logiciels nécessitant une connaissance représentée de la manière la plus formelle possible (cf. ontologies, thesaurus etc.). D'autre part, le web cognitivement sémantique vise avant tout à semi-automatiser certaines tâches pour accroître l'intelligibilité du Web pour des utilisateurs humains qui sont dans une démarche d'exploration et de création de contenus (cf. systèmes de social bookmarking). Cette seconde approche de l'enrichissement du Web Sémantique se base donc avant tout sur l'interaction humain-machine et in fine humain-humain via l'asynchronicité des supports offerts par le Web 2.0. Nous résumons les bénéfices apportés par cette approche dans le Tableau 1.

Le Web computationnellement sémantique, par sa représentation des connaissances totalement non-ambigüe, permet de produire des inférences logiques valides et de rechercher l'information de façon précise. Le Web cognitivement sémantique ne permet généralement pas de faire ce genre d'inférence logiquement valide de manière automatique mais se base sur le dialogue de la machine avec l'humain qui est garant de la validité cognitive des informations qu'il décide de retenir. Le fait de faciliter l'élicitation des connaissances avec cette représentation plus simple et intuitive des connaissances a permis à la démarche de l'élicitation de couvrir un nombre plus vaste de domaines qu'avec le web computationnellement sémantique. Une modélisation moins formelle, ou plutôt moins « sémantiquement opérationnelle » pour un ordinateur mais plus légère, lisible et malléable du point de vue d'un utilisateur, pourra apparaître comme plus adaptée à des documents évolutifs mobilisant des points de vue différents. Elle semble plus conforme à la vision des fondateurs du Web, sans doute en partie utopique, selon laquelle la « notion de communauté ouverte de spécialistes qui fournissent chacun des parcelles de connaissance doit se substituer à la figure de l'expert de référence unique. Dans ce contexte, il est important que des points de vue multiples, parfois contradictoires, puissent s'exprimer et être représentés et que les concepts et les relations qui les relient soient fréquemment remis en cause. A contrario, dans l'optique du web computationnellement sémantique, une attention particulière sera d'abord portée à l'identification des experts, puis à l'obtention d'un consensus dans la définition d'un concept afin d'obtenir une représentation logiquement valide et univoque. EnCOrE[33] était un projet de construction consensuelle d'ontologie dans le domaine de la chimie organique dont les difficultés d'établissement du consensus peuvent être un frein à l'établissement d'une sémantique collective. Ces deux approches ont donc leurs avantages et inconvénients. Voilà pourquoi la communauté IC aujourd'hui s'intéresse au rapprochement

entre le Web Social et le Web Sémantique[34]–[37] afin de tirer le meilleur parti de ces deux modes d'élicitation de la connaissance.

Tableau 1 Bénéfices et désavantages et web computationnellement sémantique et cognitivement sémantique.

	Web computationnellement sémantique	**Web cognitivement sémantique**
Coût initial de modélisation	Elevé	Faible
Coût de mise à jour	Elevé	Faible
Intuitivité de la représentation	Faible	Elevée
Réponses aux requêtes	Faible quantité de documents mais précision élevée	Grande quantité de documents mais faible précision
Possibilité de découverte fortuite	Faible car due à des questions ciblées	Elevée car le contenu est plus évolutif et les réponses sont parfois indirectement liées

L'approche ViewpointS aspire à bénéficier des avantages du Web Social pour co-construire rapidement et de manière fluide la connaissance grâce à la démocratisation et la massification des moyens de création de connaissances. Elle vise aussi à pallier les défauts que nous mentionnions à propos des folksonomies en intégrant à ces connaissances du Web cognitivement sémantique celles du Web computationnellement sémantique. Du moins c'est ce que nous tenterons de montrer au fil des expérimentations décrites au chapitre 4. Parmi les connaissances que nous intégrons dans ViewpointS certaines sont très subjectives venant du Web Social, d'autres – provenant d'ontologies – sont objectives. A chaque fois il fallait donc choisir entre (i) objectiver les connaissances du Web social et perdre la provenance et subjectivité de cette connaissances ou (ii) subjectiviser les connaissances des ontologies en leur attribuant une provenance (ex. : Si MeSH contient l'information « Le mélanome est une forme de cancer » alors nous allons l'intégrer aux connaissances subjectives du Web Social en créant un agent logiciel MeSH qui exprimera l'information subjective « MeSH dit que la mélanome est une proche de cancer »). Nous avons pris le parti de subjectiviser toute la connaissance. Nous argumentons ce choix dans ce chapitre et les chapitres suivants.

2.1.2 Représentation du Web Computationnellement Sémantique

2.1.2.1 RDF/OWL

Un format de représentation des connaissances sur le Web semble s'imposer dans les pratiques de la communauté Web Sémantique. En effet, Resource Description Framework (RDF[21])[38] est une représentations des connaissances sous forme de graphe reliant entre elles les ressources du Web. Chacune de ces ressources est identifiée par un URI unique. Ainsi les pages Web sont identifiées par des URL alors que les concepts d'ontologies sont identifiés par des URN car ce sont des concepts intangibles. Il s'agit d'un langage à base de triplet. Chaque triplet contient trois informations : la ressource à décrire (ex. : un livre), le prédicat ou la propriété décrite de cette ressource (ex. : la propriété auteur d'un livre) et l'objet qui correspond à la valeur de cette propriété (ex. :l'auteur du livre en question). Il s'agit pour le W3C (Word Wide Web Consortium) de la syntaxe standard de représentation des connaissances du Web sémantique.

> ***Définition – URI, URL***
>
> *Un URI (Universal Resource Identifier[22]) permet d'identifier une ressource du Web. Les URI sont la technologie de base du World Wide Web car tous les hyperliens du Web sont exprimés sous forme d'URI. Un URI peut être soit un URN (Universal Resource Name) qu'on utilise pour désigner une ressource sans avoir à la localiser ou une URL (Universal Resource Locator) qui donne un accès physique à la ressource.*

Plusieurs variantes de cette syntaxe furent proposées. L'une des syntaxes de ce langage est RDF/XML[23]. D'autres syntaxes de RDF sont apparues ensuite, cherchant à rendre la lecture plus compréhensible ou à compresser l'information en utilisant moins de caractères ; c'est le cas par exemple de Notation3 (ou N3) ou Turtle[24]. Le W3C propose en 2008 une syntaxe RDF – RDFa – permettant d'intégrer des connaissances en RDF à l'intérieur d'une page HTML.

Nous montrons ci-dessous un exemple de connaissances représentées en RDF en format RDF-XML, N3 et RDFa. Si on devait exprimer la connaissance de l'exemple suivant en langage naturel ce serait : « On parle du livre 1984. Georges Orwell l'a écrit. Il est paru en 1949 ». On note que N3 est une syntaxe beaucoup plus compacte que RDF-XML. On voit dans l'exemple que pour décrire le livre « 1984 » nous faisons appel à des attributs comme dc:title ou dc:creator. En effet pour décrire des ressources nous utilisons des vocabulaires tels que DublinCore[25]. Ces vocabulaires sont décrits en OWL ou une de ses alternatives plus légères : RDFs. Il existe une très grande variété de vocabulaires pour décrire les ressources du Web[26].

---

[21] *https://www.w3.org/**RDF***
[22] https://www.w3.org/wiki/URI
[23] https://www.w3.org/TR/rdf-syntax-grammar/
[24] https://www.w3.org/TR/turtle/
[25] http://dublincore.org/
[26] Le site http://lov.okfn.org/dataset/lov/ recense l'ensemble de ces vocabulaires.

Les bonnes pratiques du Web Sémantique et notamment du Web des Données Liées[20] comprennent entre autre l'utilisation des vocabulaires recommandés par le W3C afin d'éviter que les ressources soient décrites à l'aide de vocabulaires différents mais ayant le même but. On peut y trouver des vocabulaires populaires comme Friend Of A Friend (FOAF) permettant de décrire les relations sociales entre individus.

Exemple RDFa

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/"
    about="http://www.example.com/books/1984">
    <span property="dc:title">1984</span>
    <span property="dc:creator">Georges Orwell</span>
    <span property="dc:date">01-01-1984</span></div>
```

Exemple RDF-XML

```
<rdf:Description rdf:about="http://www.example.com/books/1984"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
    <dc:title>1984</dc:title>
    <dc:creator>Georges Orwell</dc:creator>
    <dc:date>01-01-1984</dc:creator>
</rdf:Description>
```

Exemple N3

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.

<http://www.example.com/books/1984>
    dc:title "1984";
    dc:creator "Georges Orwell";
    dc:date "01-01-1984".
```

Web Ontology Language (OWL) est un langage de représentation des connaissances construit sur le modèle de données de RDF. Il fournit les moyens pour définir des ontologies web structurées. Le langage OWL est basé sur les recherches effectuées dans le domaine de la logique de description. Il peut être vu en quelque sorte comme un standard informatique qui met en œuvre certaines logiques de description, et permet à des outils qui comprennent OWL de travailler avec ces données, de vérifier que les données sont cohérentes, de déduire des connaissances nouvelles ou d'extraire certaines informations de cette base de données. Dans l'exemple ci-dessus nous créons un vocabulaire sur les plantes que nous allons ensuite utiliser dans la description des Magnolias. Nous décrivons un vocabulaire sur les plantes en définissant les types d'objets (plant type et sa spécialisation flowers) ainsi qu'un attribut « family » pour décrire nos plantes. Nous décrivons ensuite le Magnolia comme une fleur (i.e. de type flowers) de la famille des Magnoliacées.

C'est un formalisme de représentation des connaissances qui s'est popularisé par sa manière de s'intégrer aux pages Web grâce à RDFa, ainsi que dans beaucoup de secteurs de l'industrie qui veulent représenter les connaissances – leur « logique métier – afin de faciliter la transmission des con-

naissances. Le succès de cette approches se mesure au nombre de technologies qui ont ensuite été créées pour traiter RDF : Apache Jena[27] (que nous utilisons pour l'export/import de données rdf) ou Rdf4J[28] pour Java, SPARQL[29] le langage de requête homologue de SQL pour RDF, EasyRDF en PHP, etc. L'adoption par une majorité de développeurs et l'effervescence des nouvelles librairies traitant RDF en font un des grands standards dans la représentation des connaissances.

Exemple OWL

```
<rdf:RDF
   <!— Définition de l'attribut famille -->
   <owl:DatatypeProperty rdf:about="http://www.example.com/plants#family"/>
    <!— Définition de la classe Plante -->
    <owl:Class rdf:about="http://www.example.com/plants#planttype">
         <rdfs:label>The plant type</rdfs:label>
          <rdfs:comment>The class of all plant types.</rdfs:comment>
    </owl:Class>

    <!—Définition de la sous-classe Fleur -->
    <owl:Class rdf:about="http://www.example.com/plants#flowers">
          <rdfs:subClassOf rdf:resource="http://www.example.com
/plants#planttype"/>
          <rdfs:label>Flowering plants</rdfs:label>
          <rdfs:comment>Flowering plants, also known as angio-
sperms.</rdfs:comment>
    </owl:Class>
```

Instanciation du type Fleur

```
   <rdf:Description rdf:about="http://www.example.com/plants#magnolia">
       <rdf:type rdf:resource="http://www.example.com/plants#flowers"/>
       <plants:family>Magnoliaceae</plants:family>
   </rdf:Description></rdf:RDF>
```

2.1.2.2 Topic Map

Les cartes topiques[39] (en anglais Topic Maps) sont un standard ISO de représentation des connaissances, dont le but est de catégoriser des documents autour de descripteurs de connaissances (les Topics). Ces topics sont ensuite reliés entre eux dans un réseau sémantique. Il s'agit du produit d'un travail débuté en 1993 par le GCA Research Institute[30] visant à définir des vues multiples et concurrentes d'un ensemble d'information. Les principaux concepts derrière cette approche sont le Topic, le Sujet, la Ressource et l'Occurrence. D'autres concepts secondaires comme le Rôle enrichissent le modèle. La Figure 4 illustre ces concepts. Le concept central des Topic Maps est le Topic. Il repré-

[27] https://jena.apache.org
[28] http://rdf4j.org/
[29] https://www.w3.org/TR/rdf-sparql-query/
[30] http://www2.gca.org/knowledgetechnologies/2001/proceedings/ahmed/

sente un sujet unique et clairement identifié dans le contexte et est une instance d'au moins une classe. Un Topic est décrit par son (ses) nom(s), occurrences et rôle(s) dans les associations. Le sujet est ce que le Topic essaie de représenter formellement. L'identification du sujet est problématique. Une Topic Map aux sujets ambigus est sinon inutilisable -- du moins source de confusions. Les sujets ressources adressables (par exemple documents sur le Web, bases de données, etc.) sont identifiés de manière non ambiguë et optimale par leurs URI. L'URI de cette ressource permet l'accès à une définition écrite, sonore ou visuelle dudit sujet. Les ressources contiennent les informations sur les sujets des Topics. Elles peuvent être des bases de données, des documents en lignes, des pages Web, etc. Une ressource concernant le sujet d'un Topic définit une occurrence de ce Topic. Les ressources peuvent être classées par type en utilisant par exemple leurs métadonnées. Une occurrence est un lien vers une ressource sur le sujet du Topic. Les occurrences sont classifiables par type : document texte, image, statistiques etc. les occurrences sont valides dans un contexte.

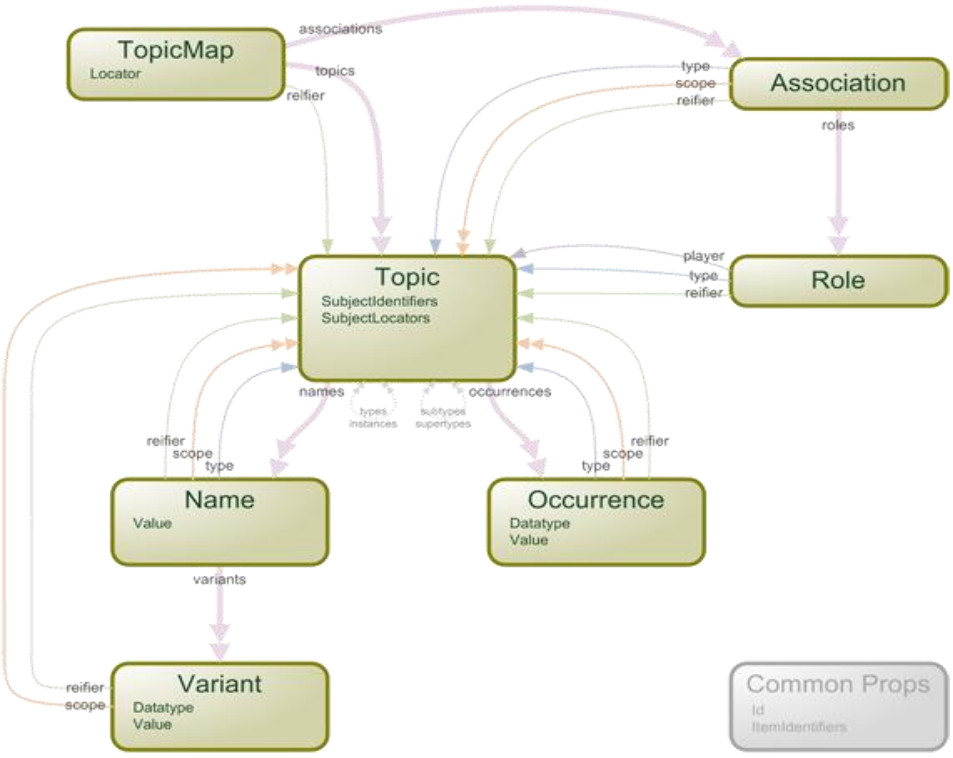


Figure 4 Illustration Topic Map

Un modèle Topic Map s'exprime dans une syntaxe dérivée d'XML qu'on appelle XTM[31]. L'extrait de code précédent illustre par l'exemple l'approche Topic Map. L'exemple précédent décrit Hamlet – la pièce de théâtre écrite par William Shakespeare. Le Topic Hamlet est relié à la ressource textuelle correspondant à l'œuvre. Topic Map propose ensuite son analogue à SQL - TMQL[32]d – permettant de requêter.

[31] La spécification XTM est définie par le consortium TopicMaps.org
[32] http://www.isotopicmaps.org/tmql/

Exemple Topic Map

```
<topic id="hamlet">
    <instanceOf><topicRef xlink:href="#play"/></instanceOf>
    <baseName>
        <baseNameString>Hamlet, Prince of Denmark</baseNameString>
    </baseName>
    <occurrence>
        <instanceOf>
            <topicRef xlink:href="#plain-text-format"/>
        </instanceOf>
        <resourceRef
xlink:href="ftp://www.gutenberg.org/pub/gutenberg/etext97/1ws2610.txt"/>
    </occurrence>
</topic>
```

Toutefois, même si Topic Map a su développer une interopérabilité avec RDF/OWL/SPARQL il n'a, à notre avis, pas su proposer suffisamment de différences par rapport à RDF pour se trouver une niche d'usage. Il s'agit d'un formalisme qui est selon nous abandonné.

2.1.2.3 Logiques descriptives

Les Logiques Descriptives (LD)[40] sont une famille de formalismes de représentation des connaissances basés sur la logique. Les LD ont vocation à permettre à des services de raisonnement de faire de l'aide à la décision.

Le développement des LD fut fortement influencé par les travaux sur la logique des prédicats, les schémas[41] et les réseaux sémantiques. Des correspondances existent entre les LD et ces formalismes. La modélisation des connaissances d'un domaine avec les LD se réalise en deux niveaux. Le premier, le niveau terminologique ou TBox, décrit les connaissances générales d'un domaine alors que le second, le niveau factuel ou ABox, représente une configuration précise. Une TBox comprend la définition des concepts et des rôles, alors qu'une ABox décrit les individus en les nommant et en spécifiant en termes de concepts et de rôles, des assertions qui portent sur ces individus nommés. Plusieurs ABox peuvent être associés à une même TBox ; chacune représente une configuration constituée d'individus, et utilise les concepts et rôles de la TBox pour l'exprimer. Le Tableau 2 exemplifie la construction d'une base de connaissance en LD.

Ce tableau contient des entités atomiques qui peuvent être soit des entités (ex. : Humain, Femelle), soit des rôles (ex. : relationParentEnfant) soit des entités composées. Les concepts et rôles atomiques peuvent être combinés au moyen de constructeurs pour former respectivement des concepts et des rôles composés. Par exemple, le concept composé Mâle ⊓ Femelle résulte de l'application du constructeur ⊓ aux concepts atomiques Mâle et Femelle. Le concept Mâle ⊓ Femelle s'interprète comme l'ensemble des individus qui appartiennent aux concepts Mâle et Femelle.

Les différentes LD se distinguent par les constructeurs qu'elles proposent. Plus les LD sont expressives, plus les chances sont grandes que les problèmes d'inférence soient non décidables ou de complexité très élevée. Par contre, les LD trop peu expressives démontrent une inaptitude à représenter des domaines complexes. Ainsi toute une famille d'extensions a été créée pour enrichir la séman-

tique des LD. La LD minimale à partir de laquelle toutes les autres ont été dérivées est notée $\mathcal{AL}$ et fût créée en 1991 par Schmidt-Schaub et Smolka[42]. Il existe trois façons proéminentes d'étendre $\mathcal{AL}$: (i) ajouter des constructeurs de concepts, (ii) ajouter des constructeurs de rôles et (iii) énoncer des contraintes sur l'interprétation des rôles.

Toutefois l'ajout de couches d'expressivité aux LD augmente sensiblement la complexité du raisonnement. Le Tableau 3 présente un aperçu non-exhaustif de la complexité du raisonnement en fonction de l'expressivité des LD[43]. On observe donc que la complexité du raisonnement dans les LD les plus expressives est exponentielle ou linéaire-exponentielle. Il s'agit d'un verrou majeur pour la représentation des connaissances du Web et en particulier de la représentation de la richesse des interactions du Web. Beaucoup de LD se basent sur des structures d'arbre binaire qui s'inspirent des fragments à deux variables de la logique du premier ordre.

Tableau 2 Exemple d'une base de connaissance en LD.

TBox	*ABox*
$Femelle \sqsubseteq \top \sqcap \neg Mâle$	$Humain(Anne)$
$Mâle \sqsubseteq \top \sqcap \neg Femelle$	$Femelle(Anne)$
$Animal \equiv Mâle \sqcup Femelle$	$Femme(Sophie)$
$Humain \sqsubseteq Animal$	$Humain(Robert)$
$Femme \equiv Humain \sqcap Femelle$	$\neg Femelle(Robert)$
$Homme \equiv Humain \sqcap \neg Femelle$	$Homme(David)$
$Mère \equiv Femme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Sophie, Anne)$
$Père \equiv Homme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Robert, David)$
$MèreSansFille \equiv Mère \sqcap$ $\forall relationParentEnfant.\neg Femme$	
$relationParentEnfant \sqsubseteq \top_R$	

Il existe deux fonctionnalités que les LD ne partagent pas avec les autres formalismes de représentations des connaissances : (i) deux conceptions avec deux noms différents peuvent par inférence être désignées comme équivalentes (cf. Unique name assumption) et (ii) les LD se basent sur l'hypothèse du monde fermée. L'hypothèse du monde fermée (HMF) est le présupposé qu'un fait est considéré comme *faux* si, en un temps fini, on échoue à montrer qu'il est vrai, ce qui revient à dire que tout ce qui est vrai doit être connu (inclus dans la base de connaissances) ou démontrable en temps fini, il n'y a pas de monde extérieur qui pourrait contenir des éléments de preuve inconnus du programme.

Tableau 3 Résumé de la complexité du raisonnement par rapport à l'expressivité croissante des LD.

Complexité	Logiques de description
P	$\mathcal{AL}$, $\mathcal{ALN}$
NP	$\mathcal{ALE}$
PSpace	$\mathcal{ALC}$, $\mathcal{ALEN}$
ExpTime	$\mathcal{SHIQ}$, $\mathcal{SHOQ}$
NExpTime	...

Les logiques descriptives sont à notre avis une approche qui n'est pas envisageable pour des questions de complexité et de présupposé épistémologique (HMF) d'appliquer à la surabondance de connaissances qui émerge du nouveau Web socio-sémantique composés souvent d'éléments contradictoire. Nous abordons dans la section suivante des approches qui tentent de dépasser la représenta-

tion computationnellement sémantique des connaissances et voir des approches qui essayent d'intégrer le web cognitivement sémantique.

2.1.3 Représentation des connaissances par point de vue

Si un formalisme a pour but d'intégrer les connaissances du web computationnellement sémantique et cognitivement sémantique, il devrait à notre avis représenter l'agent qui est le fournisseur de connaissances et préserver ainsi la provenance de la connaissance. Nous réduisons donc dans les sections suivante notre champ d'investigations aux formalismes agent-centriques en IC. Dans la même optique d'intégration – et en particulier l'intégration des connaissances venant des interactions du Web Social – nous nous intéressons spécialement aux formalismes représentant la subjectivité de la connaissance. Les approches que nous présentons visent principalement à permettre la structuration collaborative des tags ou concepts.

En 2005 HyperTopic était proposé comme une extension de Topic Map prenant en compte la notion de point de vue[44]. CartoDD[45] se base sur ce formalisme pour cataloguer collaborativement du contenu. Dans cette approche la notion de point de vue correspond à la notion de perspective sur un champ de connaissance. Ces points de vues correspondent aux différents perspectives offertes par diverses expertises (ex. : point de vue médical, point de vue biologique, point de vue éthique etc.).

La Figure 5 illustre le rapport entre le point de vue d'HyperTopic et les ressources du Web. Ce diagramme suffit d'ores et déjà à soulever une limitation d'HyperTopic : un Topic ne peut appartenir qu'à un seul point de vue.

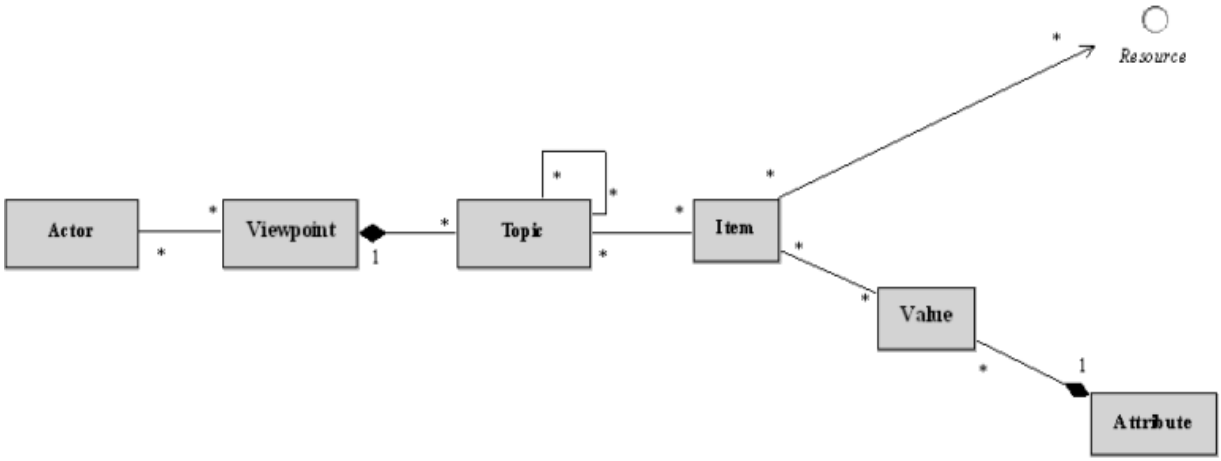


Figure 5 Formalisme HyperTopic résumé tirée de

A contrario de cette vision du point de vue comme « vue intégrée sur un ensemble », Limpens proposait une autre vision du point de vue[46]. En effet, le point de vue est alors considéré comme une micro-expression d'une sémantique individuelle. Chacun de leurs points de vue correspond alors à une information subjective qui devrait être interprété de différentes façons en une connaissance interprétée. Toutefois le processus d'interprétation personnalisée n'a pas été implémenté dans cette approche.

La Figure 6 illustre ce processus de structuration collaborative de tags. Dans l'exemple, au fil de sa recherche sur le terme pollution, plusieurs suggestions de relations entre le tag recherché et d'autres tags semblant être en relation sont faites à l'utilisateur qui fait ensuite le tri. Cela correspond aux procédés du Web cognitivement sémantique. L'utilisateur peut donc lier le tag « pollution » au tag

« pollution organique » en créant un point de vue rapprochant ces deux concepts via un quadruplet : (utilisateur, pollution, pollution organique, spécialisation).

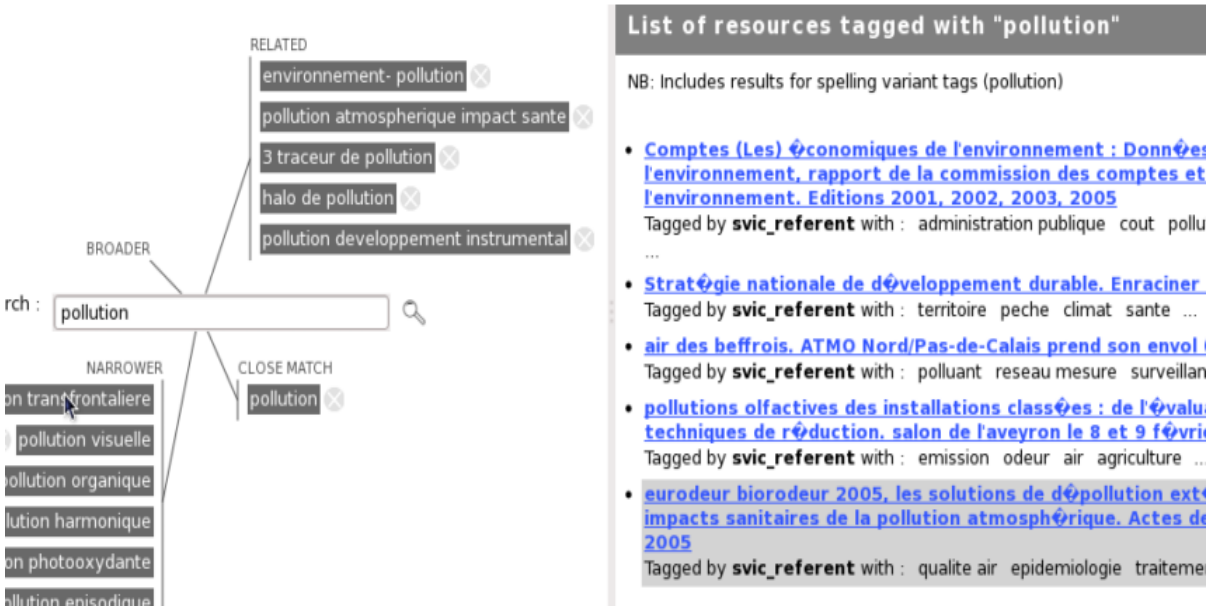


Figure 6 Processus de structuration collaborative de tags au fil des recherches des utilisateurs tirée de [44]

Toutefois les points de vue exprimés peuvent entrer en conflit. Ces conflits sont résolus dans la plupart des cas de manière automatique mais dans certains cas la résolution du conflit doit faire appel à une contribution supplémentaire. Un « super-utilisateur » joue donc l'arbitre dans le conflit entre points de vue. La résolution des conflits de points de vue n'est donc pas transparente car la logique du « super-utilisateur » n'est pas connue et peut varier et n'est donc pas entièrement automatique.

La création du formalisme ViewpointS souhaite répondre au besoin d'un formalisme de représentation des connaissances apte à représenter des connaissances subjectives pouvant être interprétées de multiples façons et confrontant les points de vue de de manière entièrement transparente et automatique grâce à des algorithmes se basant sur la topologie des connaissances. Par rapport à l'approche présentée par Limpens nous souhaitons donc une double subjectivité dans notre formalisme : (i) celle des briques de base des sémantiques individuelles des utilisateurs créateur de connaissances et (ii) celle de l'interprétation libre de ces points de vue par l'utilisateur exploitant ces connaissances.

Nous allons voir dans la section suivante un des bénéfices de l'approche topologique dans ViewpointS pour la découverte de connaissances surprenante : la Sérendipité.

2.2 Découverte des connaissances, la surprise de la Sérendipité

La Sérendipité est définie dans le dictionnaire Oxford comme « la faculté de faire d'inattendues et heureuses découvertes par accident »[33]. Le mot lui-même a été utilisé pour la première fois par Horace Walpole dans une lettre du 28 janvier 1754 à son ami Horace Mann, diplomate du roi George II à

[33] Définition : http://www.oxforddictionaries.com/fr/definition/anglais_americain/serendipity

Florence, évoquant une énigme en matière d'armoiries vénitiennes qu'il venait de résoudre accidentellement. Ce mot était dérivé d'un ancien comte persan : « Les trois princes de Serendip »[47]. Merton et Barber écrivaient à propos du phénomène de Sérendipité qu'il « concerne l'expérience assez générale de l'observation d'une donnée non-anticipée, anormale et stratégique qui devient l'occasion du développement d'une nouvelle théorie, ou l'extension d'une théorie existante. » [48]. Toutefois la définition de Sérendipité évolua et Solly en 1880 la décrivait comme une « sagacité accidentelle » ou le fait de « chercher quelque chose pour trouver autre chose ». Plus récemment, Perriault[49] disait « L'effet Serendip (…) consiste à trouver par hasard et avec agilité une chose que l'on ne cherche pas. On est alors conduit à pratiquer l'inférence abductive, à construire un cadre théorique qui englobe grâce à un bricolage approprié les informations jusqu'alors disparates ». L'abduction est avec la déduction – procédé par lequel on va de la cause aux effets - l'une des trois formes de raisonnements en science. D'ailleurs, selon le sémioticien et philosophe américain Charles Sanders Peirce, fondateur du courant pragmatiste en sémiologie, l'abduction est une troisième forme de raisonnement, qui complète la déduction et l'induction. Selon lui, l'abduction est le seul mode de raisonnement par lequel on peut aboutir à des connaissances nouvelles. Cependant, selon Pek van Andel la méthode hypothético-déductive et la méthode anomalie-abductive ne s'excluent pas[34]. En effet, il décrit la science comme une démarche se trouvant dans une systématique alternance entre 1/ tester une hypothèse par une démarche déductive et 2/ expliquer une anomalie par méthode abductive. Ces définitions font appel à des mots comme « accident » ou « chance », ainsi nous percevons que la notion de hasard est importante dans le phénomène de Sérendipité. Toutefois, dans « Three principles of Serendip : Insight, Chance, and Discovery »[50] les auteurs rejetaient le principe d'une Sérendipité uniquement liée au « divin jet de dés ». En effet, comme le disait Pasteur « la chance favorise les esprits préparés » et – pour ne citer que l'un des exemples historiques – elle a en l'occurrence favorisé celui d'Alexander Fleming qui s'il n'avait pas été expert n'aurait pas reconnu la pénicilline comme résultat accidentel de son travail qui consistait à l'origine à faire des cultures de staphylocoques dans le but d'étudier l'effet antibactérien du lysozyme. Ses boîtes de Pétri furent contaminées accidentellement et il se rendit compte qu'autour des champignons qui avait contaminé ses boîtes les staphylocoques ne poussent plus. La préparation, l'entraînement et la connaissance ne garantissent pas la découverte par Sérendipité mais elles la rendent plus probables.

Cette Sérendipité existe d'autant plus sur le Web au vu de l'immense quantité de données qu'il contient et des chances que l'on a de s'y perdre. Et vu que les gens surfent sur internet pour le plaisir mais aussi pour la découverte et l'apprentissage on peut parler d'apprentissage « sérendipiteux » sur internet. Dans ce contexte, Bowles écrivait que la recherche de connaissance par l'apprentissage sérendipiteux peut arriver par chance ou comme sous-produit d'une tâche principale. Par exemple, un utilisateur fait une recherche initiale qui le mène, au fur et à mesure de l'exploration des résultats, sur une trajectoire tangente qui in fine s'avère plus productive que sa première requête. Dans de tels cas Bowles nous dit que l'apprentissage sérendipiteux a lieu. D'ailleurs selon Allen Tough presque 80% de l'apprentissage est informel et non planifié professionnellement [51]. Qui plus est, Marchionini disait de la navigation sérendipiteuse qu'il s'agissait « d'une loterie intellectuelle … peu de probabilité mais gros gain potentiel » [52]. Nous gagnons aussi « de nouveaux points de vue (façons de voir) ou associations pour notre problème en parcourant des sources alternatives utilisant des outils,

[34] Serendipité ou l'art de faire des trouvailles : http://www.automatesintelligents.com/echanges/2005/fev/serendipite.html

des techniques et des structures de données différentes ». De nos jours les moteurs de recherche et exploration du Web qui impliquent du hasard comme StrumbleUpon[35], Banana Slug[36] ou des systèmes de bookmarking social comme del.icio.us[37] gagnent de plus en plus en popularité. Toutefois, la Sérendipité est un état d'esprit à cultiver pour faire des trouvailles. C'est le rôle que peut jouer la Sérendipité comme mode opératoire de l'acquisition de connaissances de notre communauté scientifique. Le groupe 3M est un exemple en la matière en amenant les scientifiques à consacrer 15% de leur temps à des axes de recherche en dehors de ceux définis par la R&D.

Si nous nous référons à la littérature nous ne parlons pas de Sérendipité mais de Sérendipités. Swinners & Briet dans [53] mettent en évidence 4 types de sérendipités :

1. Fait de trouver (découvrir, inventer) par hasard, par chance ou par accident, autre chose et, parfois tout autre chose, et, même, parfois, le contraire de ce que l'on cherchait (et de trouver en l'état) ; et de se rendre compte de son intérêt et de son importance. Ex : découvertes du Téflon et du Kevlar.
2. Fait de trouver (découvrir, inventer) quelque chose que l'on cherchait (objet, solution, etc.) mais, à la suite d'un accident plus ou moins malheureux ou d'une erreur, par un moyen imprévu ; et de s'en rendre compte. Ex: découverte de l'imprimante à jet d'encre.
3. Fait de découvrir par hasard, par accident, par chance ou par malchance, une application imprévue à quelque chose, une autre application que celle à laquelle on pensait ; et de s'en rendre compte. Ex : la super-glue et le four micro-onde dont le magnétron était à l'origine censé être le cœur d'un radar.
4. Faculté de trouver par accident, hasard ou chance l'idée d'une innovation. Ex : passer d'un flacon de verre tombé d'une table, et qui ne se casse pas, au pare-brise d'automobile.

On retrouve ces types de Sérendipité en naviguant sur le Web avec un but bien défini, partiellement défini ou en « errant » sans but précis (cf. cyber-glandouille). L'ethnographe Mark de Rond dans The Structure Of Serendipity [54] structure ces quatre types de Sérendipité de la façon suivante dans le Tableau 4.

L'approche ViewpointS et la recherche d'information qu'elle propose permet ces types de Sérendipité. En effet le moteur de recherche basé sur l'approche renvoie pour une requête les résultats qui y sont soit directement soit indirectement liés. Cela permet des résultats qui vont au-delà de ce qu'un moteur de recherche classique basé sur des mots-clés aurait pu fournir. C'est donc en prenant le risque de renvoyer des informations indirectement liées aux résultats souhaités que l'on ouvre la possibilité de situations de découverte, d'apprentissage fortuit. Nous verrons que ce qu'on pourrait appeler les circonstances dans le concours de circonstances précédent si on se place dans l'approche ViewpointS sont en fait les recherches d'autres utilisateurs qui ont été soit valorisées soient critiquées par le système de feedback qui est au cœur de notre approche. En effet le système de feedback sous forme de points de vue permet le filtrage communautaire de ces ressources indirectement liées quand elles ne sont pas pertinentes par rapport au besoin d'information exprimé.

[35] https://www.stumbleupon.com/
[36] http://bananaslug.com/
[37] https://delicious.com/

Tableau 4 Structuration des 4 types de Sérendipité

	Concours de circonstance	Hasard pur
On découvre ce que l'on cherche (pseudo-Sérendipité)	On découvre grâce à un concours de circonstances favorables ce que l'on cherchait : le PCR (Réaction en chaîne par polymérase).	On découvre par hasard ce que l'on cherchait : la structure de l'ADN
On découvre autre chose que ce que l'on cherchait (pure Serendipité)	On découvre par hasard ce que l'on ne cherchait pas : l'aspirine	On découvre autre chose que ce que l'on cherchait grâce à un concours de circonstances favorables : la pénicilline.

Nous constatons aussi que sur la majorité des moteurs de recherche aucun mécanisme de feedback n'est donné pour que l'utilisateur puisse donner un retour sur la qualité des résultats que lui renvoie le moteur de recherche contribuant ainsi à l'amélioration du service pour lui-même et tous les autres utilisateurs. C'est une fonctionnalité que nous remarquons de nos jours principalement sur les systèmes de recommandation. Cette absence de feedback a d'ailleurs créé pour Google Scholar un effet « junk science »[38].

Cependant, une des caractéristiques clé de la Sérendipité est sa fugacité, il est quasiment impossible de retrouver le chemin qui a conduit à l'information sérendipiteuse. Il faut l'enregistrer immédiatement et l'indexer en clair systématiquement donc proposer également un mécanisme de feedback pour l'information non renvoyée initialement par le moteur de recherche, non-recherché initialement par l'utilisateur mais qui participent à l'enrichissement de sa recherche et de celles de autres.

Nous souhaitons adopter dans ViewpointS cette idée itérative d'une interaction entre un service qui se base sur la topologie des connaissances et non pas le raisonnement, qui renvoie « Beaucoup de Résultats mais avec une faible précision » (cf. Tableau 1) et un utilisateur qui valide, rejette et navigue activement dans l'espace des connaissances. Les feedbacks émis par les utilisateurs aident ensuite aux recherches pour les utilisateurs suivants. Ces feedbacks sont autant de mis de pains qui permettront aux utilisateurs de demain de bénéficier de ces cheminements donc certains ont mené à la découverte fortuite.

2.3 Positionnement de l'approche ViewpointS

L'approche ViewpointS se positionne donc dans la succession des approches de représentation des connaissances telles quel celle développée par F. Limpens et Gandon intégrant l'agent comme élé-

[38] http://scholarlyoa.com/2014/11/04/google-scholar-is-filled-with-junk-science/#more-4371

ment central ainsi que la notion de point de vue. Qui plus est, nous donnons le moyen d'une interprétation personnalisée des points de vue. L'utilisation d'une approche topologique basée sur l'évaluation de distances sémantiques permet de rendre à la fois transparente et automatique la confrontation des points de vue et elle permet également la découverte fortuite de connaissances. En effet, nous assumons de diminuer la précision des recherche en renvoyant à l'utilisateur des résultats indirectement liés à sa requête. Cela a pour nous l'avantage principal de permettre les découvertes de nouvelles associations au fur et à mesure des recherches et à l'utilisateur d'enrichir le moteur de recherche ViewpointS par l'émission de feedbacks.

2.4 Méthodes topologiques d'exploitation des connaissances

2.4.1 Etat de l'art des mesures de similarité sémantique

La majorité des systèmes à base de connaissances utilisent l'inférence à partir de la connaissance exacte d'un domaine c'est-à-dire qu'à partir de faits établis sont déduits d'autres faits. La déduction est une approche qui est tout à fait adaptée à la recherche d'information quand l'objectif est clairement déterminé. Or, comme c'est souvent le cas en recherche d'information, l'objectif n'est pas clairement défini – du moins au début – et se définit au fur et à mesure de la recherche. Ainsi, l'approche déductive ne comprend tout simplement pas les questions telles que « Quel article est similaire à celui que je viens d'écrire ? » car cette question ne peut être traitée par des opérateurs logiques qui se basent sur des variables booléennes. Mais il y a dans les représentations des connaissances les éléments qui permettent, par exemple, de dire que ce document est plus proche du concept « Algorithmes de graphe » que du concept « Raisonnement logique ». Le raisonnement approximatif typique du Web cognitivement sémantique utilise des méthodes capables de comparer les entités.

Ces méthodes de comparaison sémantique sont utiles pour classifier des ressources, pour obtenir le « voisinage sémantique » d'une entité composé de toutes les entités relativement similaires à une recherche d'information. Ce mécanisme de similarité sémantique est d'ailleurs au cœur du raisonnement cognitif humain. La capacité d'apprécier des distances entre concepts abstraits comme s'ils faisaient partie d'un espace topologique de la connaissance avant de servir en informatique était étudiée par la communauté des psychologues[55]. En effet, l'apprentissage humain se base sur un renforcement synaptique et ce renforcement se produit au fur et à mesure que nous vivons des situations similaires. La Similarité sémantique joue donc un rôle très important dans l'apprentissage humain, la reconnaissance de formes ou la planification[56]. Etant donné que nous nous inspirons des procédés mentaux humains dans la création de systèmes à base de connaissance il est alors évident que cette mesure de similarité sémantique joue un rôle très important dans la création d'intelligences artificielles.

> ***Définition – Distance/Similarité sémantique***
>
> *La similarité sémantique est une notion définie entre deux concepts soit au sein d'une même hiérarchie conceptuelle, soit - dans le cas d'alignement d'ontologies - entre deux concepts appartenant respectivement à deux hiérarchies conceptuelles distinctes. La similarité sémantique indique que ces deux concepts possèdent un grand nombre d'éléments en communs.*

Pour commencer nous discutons dans cette section de différents exemples d'application de mesure de similarité sémantique. Nous discutons ensuite des catégories de mesures de similarité ou de distance sémantique de la littérature que Sébastien Harispe et al. catégorisait dans son travail [57] sur l'unification des mesures de similarité/distance sémantique.

Parmi les défauts que nous évoquions à propos des folksonomies, la similarité sémantique propose d'en corriger un : la multiplication de tags faisant références au même concept (ex : comme la création de tag est libre on retrouve plusieurs variantes ou même avec une faute d'orthographe). Mais, la topologie des folksonomies permet le calcul de similarité sémantique entre tags et certains tags qui sont jugés très similaires peuvent être jugés comme équivalents. Il est alors possible de créer la connexion forte entre ces deux expressions d'un même concept et de corriger le problème de la polysémie[58] ou de les fusionner.

La mesure de similarité sémantique trouve aussi application en recherche d'information quand il s'agit de trouver le voisinage sémantique d'un terme recherché comme le montrent ces cas d'étude sur WordNet[59] ou MeSH[60].

En classification on utilise également la similarité sémantique afin de partitionner l'espace des connaissances, c'est-à-dire de regrouper les entités en paquets d'objets fortement liés entre eux. On appelle ce procédé la clusterisation. Le lecteur pourra trouver dans les travaux suivants [61], [62] un état de l'art des méthodes de clustering.

La similarité sémantique sert aussi par exemple dans le cas concret de l'extension des utilisations des médicaments. Cette pratique considère que certains médicaments peuvent être exploités dans d'autres cas que ceux qui sont originalement prévus. Pour une maladie spécifique traitée par un médicament précis nous cherchons les maladies similaires en nous basant sur les bases de connaissance biomédicales afin de trouver d'autres usage imprévus du médicament[63].

La similarité sémantique est une notion définie entre deux concepts soit au sein d'une même hiérarchie conceptuelle, soit - dans le cas d'alignement d'ontologies - entre deux concepts appartenant respectivement à deux hiérarchies conceptuelles distinctes. Ainsi, il y a similarité sémantique entre deux concepts si : (i) d'un point de vue intensionnel, les deux concepts partagent une grande proportion de leurs propriétés descriptives et fonctionnelles, (ii) d'un point de vue expressionnel, les deux concepts partagent une grande proportion des termes qui les dénotent et (iii) d'un point de vue extensionnel, les deux concepts partagent une grande proportion de leurs instances.

Certaines mesures calculent la similarité sémantique soit entre deux concepts (pairwise), d'autres entre deux groupes de concepts (groupwise). On parle aussi de distance sémantique mais il ne s'agit pas uniquement de construire une fonction inverse de celle qui calcule la similarité car la distance doit valider plusieurs propriétés mathématiques. En effet pour être considérée comme métrique une mesure de distance doit respecter trois propriétés : la séparation, l'inégalité triangulaire et la symétrie.

Pour un ensemble d'entités E, la distance sémantique est l'application :

$$d : E \times E \rightarrow \mathbb{R}$$

Vérifiant les propriétés suivantes :

Nom	Propriété
Séparation	$\forall (a,b) \in E^2,\ d(a,b) = 0 \Leftrightarrow a = b$
Inégalité triangulaire	$\forall (a,b,c) \in E^3,\ d(a,c) \leq d(a,b) + d(b,c)$
Symétrie	$\forall (a,b) \in E^2,\ d(a,b) = d(b,a)$

Nous partitionnons les mesures de similarité en 2 groupes : (i) les mesures qui se basent sur la topologie c'est-à-dire sur les relations entre concepts et (ii) les mesures qui en plus de considérer la topologie se base sur le contenu informationnel des concepts.

2.4.1.1 Mesures basées sur la topologie

Rada proposait en 1989 une mesure de distance sémantique à laquelle il donna son nom[64]. Il s'agit d'une mesure qui calcule dans une taxonomie le plus court chemin entre deux concepts et considère le nombre d'arcs composant ce plus court chemin comme la distance sémantique entre ces deux concepts. Par exemple, d'après la Figure 7, les concepts « Vegetable » et « Berry » sont à une distance de 3 car 3 arc séparent ces deux concepts. Nous considérons aussi que « Berry » est plus proche de « Fruit » que de « vegetable ». Dans cet exemple, le plus court chemin entre « Berry » et « Vegetable » passe par « Food » qui est ce qu'on appelle le Least Common Ancestor (LCA), c'est-à-dire le terme le plus générique qui est spécialisé directement ou indirectement par « Berry » et « vegetable ».

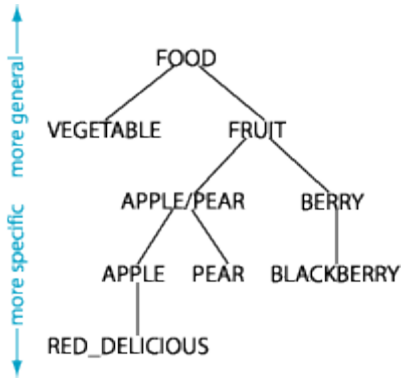


Figure 7 Exemple de taxonomie pour l'application de la mesure Rada.

Leacock & Chodorov[65] proposèrent d'adapter la distance Rada pour qu'elle tienne compte de la profondeur de la taxonomie. En effet, Rada propose le même résultat pour la distance entre « Fruit » et « Vegetable » que pour la distance entre « Apple » et « Pear ». Si nous prenons nEdges la distance Rada entre deux concepts et D la profondeur de la taxonomie (4 sur l'exemple) alors la similarité Leacock & Chodorov entre deux concepts c1 et c2 se défini comme :

$$sim_{LeacockChodorov}(c_1, c_2) = -\log(\frac{nEdges}{2 \times D})$$

Wu & Palmer proposaient ensuite un score de similarité normalisé dont la formule est la suivante et prend en compte la profondeur des deux concepts à comparer et celle du LCA[66] :

$$sim_{WuPalmer}(c_1, c_2) = \frac{2 \times depth(LCA)}{depth}$$

2.4.1.2 Mesures basées sur les attributs de nœuds et la topologie

Toutefois nous ne disposons pas systématiquement de taxonomies pour calculer une similarité ou une distance sémantique. C'est pourquoi plusieurs mesures proposent de baser en plus leurs calculs sur leurs contenus informationnels. Une première approche – Vector Space Model[67] – calculant la similarité entre ressources propose de représenter ces documents par des vecteurs descriptifs.

Dans une base de connaissance où les ressources sont en totalité décrites par n descripteurs nous allons décrire chacune des ressources par des vecteurs à n dimensions. Prenons l'exemple du calcul de similarité entre 2 films chacun étant décrit par un vecteur de genres de la façon suivante.

$$V_{film}(Star\ Wars) = \begin{matrix} Science\ Fiction \\ HeroicFantasy \\ Romance \\ \dots \\ War \end{matrix} \begin{vmatrix} 1 \\ 1 \\ 0 \\ \dots \\ 1 \end{vmatrix}$$

$$V_{film}(Lord\ Of\ The\ Rings) = \begin{matrix} Science\ Fiction \\ HeroicFantasy \\ Romance \\ \dots \\ War \end{matrix} \begin{vmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 1 \end{vmatrix}$$

On voit alors que ces deux films partagent plusieurs descripteurs en commun. Nous calculons la distance entre les deux films comme le produit scalaire des deux vecteurs :

$$d_{VSM}(Star\ Wars, Lord\ Of\ The\ Rings) = V_{film}(Lord\ Of\ The\ Rings) \cdot V_{film}(Star\ Wars)$$

Toutefois cette méthode a le défaut de fermer l'ensemble des descripteurs à un ensemble défini à l'avance. Pour cette raison, d'autres mesures ont été proposées basée sur la théorie de l'information de Shannon[68] et se basant sur l'information partagée par les entités à comparer. L'approche se base sur la notion de Contenu Informationnel. Le niveau de spécificité des concepts d'une taxonomie est variable. Cette notion de spécificité d'un concept se base sur sa profondeur. On considère en effet que plus un terme est spécifique plus il contient l'Information. Nous pouvons définir une fonction de spécificité $\theta: C \rightarrow \mathbb{R}$ de la façon suivante :

$$\theta(c) = depth(c)$$

D'après la définition ci-dessus, si deux concepts ont un nombre égal de descendants/ancêtres ils auront alors la même spécificité.

D'autres mesures de spécificité furent proposées par la suite pour faire face à ce problème. Une stratégie alternative se base sur la Théorie de l'Information. Le contenu informationnel est une expression de θ et mesure la spécificité d'un concept comme la quantité d'information qu'un concept transmet. Resnik proposa une mesure de calcul du contenu informationnel sur laquelle il a basé une mesure de similarité sémantique:

$$p(c) = \frac{|I(c)|}{|I|}.$$

Ou I(c) est le nombre d'occurrences de c dans un corpus ou instances dans une ontologie. La valeur du contenu informationnel est donc calculée de la façon suivante :

$$IC_{Resnik} = -\log(\mathrm{p(c)})$$

En se basant sur cette mesure de contenu informationnel il est possible de trouver « l'ancêtre commun le plus informationnel » (MICA) qui est le LCA qui maximiseIC_{Resnik}. La similarité selon Resnik s'exprime donc selon la formule suivante :

$$sim_{Resnik}(c_1, c_2) = IC_{Resnik}(MICA(c_1, c_2))$$

Cette mesure fut par la suite adaptée par Lin[69], Jiang & Conrath[70] ainsi que d'autres auteurs. Voici, entre autres, comment sont calculées les mesures de Lin et de Jiang & Conrath :

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC_{Resnik}(MICA(c_1, c_2))}{IC_{Resnik}(c_1) + IC_{Resnik}(c_1)}$$

$$sim_{JiangConrath}(c_1, c_2) = IC_{Resnik}(c_1) + IC_{Resnik}(c_1) - 2 \times IC_{Resnik}(MICA(c_1, c_2))$$

Les mesures de similarité sémantique que nous présentions dans les deux dernières sections évaluent la similarité entre deux concepts d'ontologie et beaucoup dépendent d'une certaine structuration de la connaissance représentée. Mais si l'agent prend de plus en plus sa place dans les représentations des connaissances, nous pensons qu'il devrait être possible d'évaluer des similarités avec ce nouveau type de ressource représentée. Ainsi, nous identifions dans la section suivante les perspectives de développement d'une similarité/distance sémantique qui s'applique à tout couple de ressources (agents, documents, descripteurs).

2.4.2 Verrous technologiques et perspectives

Nous avons identifié, dans ce tour d'horizon des méthodes de calcul de similarité ou de distance sémantiques, plusieurs verrous technologiques. Tout d'abord, ces méthodes ont besoin d'une structure taxonomique pour fonctionner. De plus, elles ne s'appliquent qu'au calcul de distance/similarité sémantique entre concepts de taxonomie. Enfin, dans le paradigme du Web cognitivement sémantique – ou nous voyons l'agent comme un élément central – nous pensons qu'une distance sémantique devrait être capable de calculer des distances sémantiques entre des couples de ressources de types différents : agents (humains ou artificiels), ressources (documents, films, musique) et descripteurs de connaissances (tags de folksonomie, concepts d'ontologie). Par exemple, il serait intéressant qu'une mesure puisse évaluer la distance sémantique entre un auteur et le sujet (descripteur) sur lequel il écrit.

Une distance sémantique exploitant toute la topologie des connaissances devrait permettre d'évaluer des distances entre …

- Un utilisateur et des documents pour lui proposer des recommandations.
- Des utilisateurs pour exploiter le réseau de supports, connaissances et descripteurs de connaissances afin de suggérer des collaborations.

- Des descripteurs et des utilisateurs pour donner par exemple la capacité à l'étudiant de trouver pour une discipline les auteurs les plus proches.
- Des descripteurs. Il s'agirait alors d'utiliser la connaissance générée par les interactions pour proposer de nouvelles relations entre ontologies.

2.5 Le Point de Vue, brique de base de sémantique individuelle

Afin de clore cet état de l'art et pour faire suite au besoin de subjectivité de la connaissance représentée dans les approches de la section 2.1 nous proposons d'approfondir la notion de point de vue, qui joue notamment un rôle important en linguistique et en pragmatique.

Nous positionnons ViewpointS comme une approche pragmatique de la sémantique, c'est-à-dire que le sens des mots s'établit au fur et à mesure de leurs utilisations dans divers contextes ou de l'expression de divers points de vue sur ces mots.

Les travaux de CS Pierce[71] et la notion d'interprétant sont essentiels pour baser notre réflexion sur le point de vue. En effet pour Pierce toute pensée de base sur des Signes qui sont des triades : un Signe matériel dénote un objet de pensée grâce à un interprétant. Un signe, s'il est perçu dans divers contextes, prends différentes significations : des sémioses. C'est un triplé signe-contexte-signification. Les points de vue dynamiques sont des interprétants dynamiques de signes exprimés par des individus évoluant au fur et à mesure du changement de l'environnement ou de l'échange entre personnes. Le point de vue institué par rapport au point de vue dynamique peut être considéré comme un point d'arrêt institué par décision collective.

On peut faire correspondre ces deux types de points de vue aux deux processus de co-construction de la sémantique collective : la convergence et l'émergence. Le point de vue dynamique est alors l'interaction de base, la micro-expression d'une sémantique, qui permet l'émergence d'une sémantique collective. Le point de vue institué est construit par un consensus en fonction des besoins d'un groupe. Le point de vue dynamique est donc pour nous à la base d'une co-construction fluide de la connaissance alors que les points de vue institués correspondent à des jalons dans le temps marquant des consensus.

Dans l'approche ViewpointS nous considérons tous les points de vue comme points de vue dynamiques car nous pensons que seules les expressions de sémantique (dans un contexte très précis à un temps données) sont aptes à représenter la dynamique de la connaissance sur le Web. Qui plus est si on considère un groupe dans son ensemble alors le point de vue institué peut être compris comme l'expression de la sémiose de cet individu-groupe au moment et dans le contexte du consensus. Nous prenons donc le parti de représenter la connaissance à partir d'un ensemble de points de vue dynamiques qui sont des expressions de sémioses et de mettre au même niveau le point de vue institué résultat d'un consensus et le point de vue dynamique.

Dans notre approche, donner la priorité au résultat du consensus ou à l'émergence reste le choix de l'utilisateur. C'est à l'utilisateur d'interpréter les points de vue ; ceci nous amènera dans le chapitre suivant à définir la notion de Perspective. Dans l'approche ViewpointS, il est possible de donner plus ou moins de poids à un point de vue en fonction de qui l'a émis. Les institutions qui seront considérées comme crédibles par les utilisateurs dans un domaine particulier comme le biomédical verront leurs points de vue mis en valeur. Mais c'est au final au choix de chacun de faire confiance à chaque source de donnée que ce soit des données intentionnelles établies par consensus, le résultat du trai-

tement d'un texte par un algorithme de TALN ou les interactions sociales. Nous présentons dans le chapitre suivant ce double mécanisme de subjectivité.

2.6 La subjectivité dans les systèmes de recommandation

Un système de recommandation a pour but de fournir un utilisateur en ressources en fonction des préférences qu'il a exprimées. En plus de ce que fait un moteur de recherche traditionnel, il filtre les résultats et rajoute certaines suggestions qui ne sont pas forcément directement liées à sa recherche. Le filtrage des résultats permet de faire face au problème inhérent aux moteurs de recherche qu'est la surabondance de résultats. Ainsi les moteurs de recherche comme Google se sont adapté en intégrant un filtrage des résultats en fonction des recherches précédentes. Cette analyse des traces est l'une des bases du filtrage d'informations sur lequel s'appuient les systèmes de recommandation [72]. Ainsi, on différencie les systèmes de recommandation dans laquelle la participation de l'utilisateur est non volontaire aux systèmes de recherche d'information dans lesquels la demande de l'utilisateur d'être guidé et orienté vers les choix approprié est explicite [73]. On trouve parmi les systèmes de recommandation trois grandes catégories qui se basent différemment sur la subjectivité de leurs utilisateurs : (i) la recommandation par filtrage de contenu, (ii) la recommandation par filtrage collaboratif et (iii) la recommandation hybride qui combine les deux.

Les systèmes de recommandation basés sur le filtrage par contenu se basent sur les évaluations qu'un utilisateur a portées sur un ensemble de documents. Le système lui propose alors un ensemble de documents proches des documents qu'il avait appréciés. On peut citer comme exemple le système INFOSCOPE qui faisait de la recommandation à des utilisateurs de groupes de discutions en fonction des requêtes qu'ils soumettent [74]. On peut résumer l'approche de recommandation par contenu aux 4 étapes suivantes : (i) constitution d'un profil utilisateur, (ii) construction de classes d'items à prédire, (iii) recommandation d'items aux utilisateurs en utilisant une mesure de proximité et (iv) enrichissement du profil utilisateur. Ce type de système de recommandation a pour qualité que seul la connaissance de l'utilisateur est requise et plus l'utilisateur l'utilisera plus la pertinence des items proposés sera bonne. Mais il a aussi comme défaut d'avoir un démarrage difficile pour l'utilisateur qui n'a construit aucun profil. Qui plus est, le filtrage par contenu a pour défaut la redondance thématique. Quelqu'un qui ne s'intéresse qu'aux items liés au cinéma ne se verra jamais proposer du contenu sur un sujet totalement différent comme la politique.

Le filtrage collaboratif corrige ces deux problèmes du « démarrage à froid » et de la redondance thématique. Il ne se base non plus sur une proximité potentiel « nouvel item – profil utilisateur » mais cherche à rapprocher l'utilisateur d'autres utilisateurs ayant des goûts à priori similaires. Les items suggérés sont alors les items appréciés par les utilisateurs de même goût. Pour ce faire, le profil de l'utilisateur doit être connu de tous. Ce profil utilisateur est conçu soit de manière passive, soit de manière active en impliquant l'utilisateur par une interrogation par exemple sur la qualité de tel ou tel jeu ou film. Nichols s'est intéressé à cette problématique en défendant la construction implicite et passive du profil de l'utilisateur [75]. On peut citer comme exemple de ce style de filtrage plusieurs approches présentées dans [76] sur les applications du filtrage collaboratifs comme le système MovieLens pour la recommandation de films ou le système FlyCasting recommandant des radios en ligne [77]. Le filtrage collaboratif manque toutefois de la notion de thème. Par exemple, un utilisateur peut ne s'intéresser qu'aux romans de science-fiction mais – étant jugé proche d'un autre utilisateur – se faire proposer des romans romantiques.

Une combinaison des deux précédents modes de filtrage fût proposée et définie dans [78] afin de tirer le meilleur parti des deux approches. Cette approche hybride est la plus présente dans la littérature car considérée comme la plus efficace [79]. Claypool proposait, par exemple, d'hybrider les deux méthodes en combinant linéairement les deux mesures liées aux deux approches [80]. Par ailleurs, d'autres études comme [81] s'intéressent à cette hybridation et résolvent le problème du démarrage à froid. La tendance actuelle des systèmes de recommandation est plutôt axée sur des méthodes hybrides, multicritères et se fondant sur des notions comme les émotions et les opinions.

On voit donc que la subjectivité de l'utilisateur, et l'intersubjectivité d'une communauté dans le cas du filtrage collaboratif joue un rôle essentiel dans les systèmes de recommandation. Il en ressort que ce que nous faisons avec ViewpointS s'apparente à de la recommandation par filtrage collaboratif. En effet, l'ensemble des viewpoints exprimés par un utilisateur/agent peut être considéré comme son profil ; une perspective centrée sur un agent est donc filtre correspondant à ce profil.

EN RESUME

Les approches actuelles de représentation des connaissances ne proposent aucun modèle centré utilisateur qui fasse émerger la connaissance à partir des interactions.

Quand la subjectivité est représentée dans ces formalismes sous forme de points de vue leur confrontation ne se fait pas systématiquement de manière automatique et transparente mais peut demander l'intervention d'arbitre.

Il est important, pour une approche qui essaye de favoriser la Sérendipité, d'enregistrer les traces de ceux qui explorent le graphe de connaissance en y creusant des chemins qui seront empruntés par les futurs utilisateurs. Quand la découverte fortuite se produit il faut permettre la sauvegarde du chemin qui y a mené.

Les distances sémantiques portent sur une classe particulière de ressources de Web – les descripteurs de connaissance – et nécessite une structuration taxonomique des connaissances. Il n'y a aucune distance sémantique calculant de manière générique des distances descripteur-agent, document-agent, document, agent-agent, etc.

Chapitre 3. L'approche ViewpointS

Ce chapitre introduit les concepts et méthodes qui fondent l'approche ViewpointS.[39] Nous commençons par expliquer la structure du graphe de connaissances et le mécanisme de Perspective. Nous abordons ensuite les méthodes exploitant le graphe de connaissances qui permet à ces méthodes d'interpréter sa subjectivité.

3.1 Introduction

Notre but avec ViewpointS est de fournir une approche de représentation des connaissances pouvant s'appliquer à n'importe quel scénario de co-construction et d'exploitation de connaissances partagées. Les scénarios de co-construction des connaissances impliquent souvent des connaissances pluridisciplinaires. Le but de notre approche est donc de permettre l'intégration de données de champs différents afin de permettre la découverte qui est le fruit de cette intégration de jeux de connaissances divers par leurs structures et leurs domaines. Dans notre représentation des connaissances nous gardons le plus petit dénominateur commun à des connaissances qui peuvent appartenir à plusieurs domaines. Ce dénominateur commun est la relation de similarité sémantique perçue par un individu entre deux concepts.

Nous avons choisi comme relation de base de ViewpointS la relation sémantique la plus basique (la similarité sémantique) car il s'agit d'un type de relation sémantique qu'on est sûr de pouvoir retrouver sous plusieurs formes dans tous les jeux de connaissances auxquels nous pourrions nous intéresser. Par ailleurs, nous souhaitons aussi préserver la valeur des contributions des utilisateurs du Web Social. Il est donc nécessaire de garder le plus longtemps possible toute la subjectivité de ces expressions de sémantiques individuelles. Nous choisissons donc la relation de proximité sémantique subjective, c'est-à-dire liée à une interprétation spécifique, comme unique relation pouvant lier deux ressources du Web.

Il y a dans ViewpointS un double niveau de subjectivité. En effet, alors que la subjectivité de la création de connaissances est préservée il y a aussi dans ViewpointS la subjectivité de l'interprétation des connaissances. L'utilisateur du graphe de connaissances peut décider d'interpréter comme il le souhaite la connaissance en fonction par exemple de qui l'a émise, de quand elle a été créée ou des types de relations qui lient les concepts. Nous introduirons dans la section suivante les notions de « point de vue » et de « ressource » qui ensemble forment le graphe de connaissances de ViewpointS. Nous aborderons le passage du graphe de connaissances subjectif à la carte de connaissances interprétées via le mécanisme de Perspective.

[39] ViewpointS est le nom propre que nous avons donné à notre approche, qui se base sur une représentation de connaissances à partir de points de vue ou viewpoints. Le 'S' final est important.

3.2 Formalisme

Dans ViewpointS, les Ressources du Web sont associées par des micro-expressions sémantiques qu'on appelle les viewpoints. Comme dans ViewpointS toute la connaissance est subjective les Ressources ne peuvent être liées que par les Viewpoints. L'ensemble des ressources et viewpoints constitue un graphe biparti que nous appelons graphe de connaissances.

3.2.1 Graphe de connaissances

Nous représentons le *graphe de connaissances* (KnowledgeGraph – KG) comme un graphe biparti $G = (R \cup V, E)$ ou R et V sont les deux ensembles de nœuds du graphe biparti. R est l'ensemble des Ressources, V est l'ensemble des viewpoints et E l'ensemble des arêtes connectant les Ressources aux Viewpoints.

3.2.1.1 Ressources

L'ensemble R est constitué par des types de ressources qui correspondent aux types de ressources qu'on trouve sur le Web ou aux ressources du « monde physique » que nous proposons de représenter (ex : un livre non-numérisé). Nous considérons les divers types d'agents – soit qu'ils représentent directement les utilisateurs du Web, soit qu'ils soient des agents logiciels contenant par exemple des algorithmes de fouille de données ou des jeux de données (Artificial Agent) ou bien soit qu'ils représentent des organisations (Legal Person) – comme un élément clé dans notre approche de représentation des connaissances. Ces agents sont les fournisseurs de connaissances. Le contenu du Web est représenté par les supports de connaissance.

Les pages Web, les images les vidéos – tant qu'elles sont identifiables et localisables par une URL sont considérées comme documents numériques (Numeric Document). Par opposition à ceux-ci, les documents physiques n'ayant pas d'existence sur le Web sont des Physical Document.

Les connaissances dans le graphe de connaissances sont structurées par un ensemble de descripteurs de connaissances qui sont soit des concepts d'ontologies ou des tags de folksonomies. Toutes ces ressources sont liées entre elles dans l'espace des connaissances par les Viewpoints. Ce sont des micro-contributions subjectives portant sur la proximité ou la distance entre deux ressources. La Figure 8 illustre l'ensemble des types de ressources qu'on lie avec les Viewpoints.

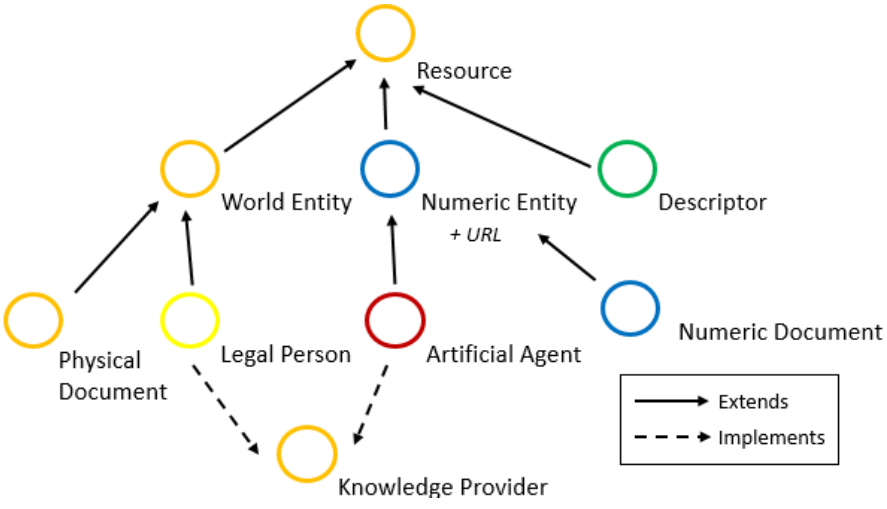


Figure 8 Structuration de types de ressources dans ViewpointS

3.2.1.2 Points de Vue

Le Viewpoint – illustré dans la Figure 9 – est la seconde classe de nœud du graphe de connaissance. Il s'agit d'un quadruplet de type (émetteur, {r_1, r_2}, {Θ,+/-}, τ) avec :

- Un agent émetteur du point de vue
- Un couple de ressources {r_1, r_2} liées par une relation de distance sémantique
- Un couple {Θ,+/-} type et polarité du Viewpoint (positive ou négative soit pour rapprocher ou séparer deux ressources)
- Une date d'émission τ

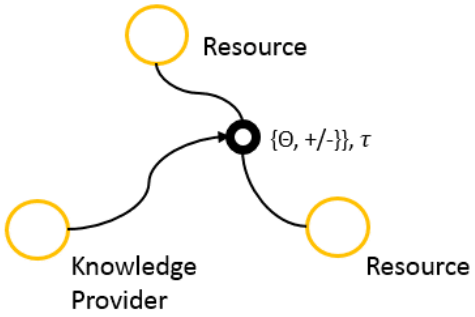


Figure 9 Représentation du viewpoint.

La polarité des viewpoints – positive ou négative – permettra au viewpoint de rapprocher ou de d'éloigner les deux objets qu'il connecte. Le fait de donner un type de viewpoint permet à l'approche ViewpointS d'intégrer la diversité des relations sémantiques qui lient les ressources du Web sémantique telle que l'hypéronymie ou la métonymie. Par exemple, parmi les cas d'étude sur lesquels nous nous sommes penchés, celui sur les données bibliographiques de la plateforme HAL LIRMM[40] nous générons trois types de Viewpoints : (i) les Viewpoints rapprochant les auteurs de leurs articles (de type Authorship), (ii) les Viewpoints rapprochant les articles des mots-clés déclarés par les auteurs (de type hasKeyword) ou (iii) ceux qui sont générés par l'annotation (générés par un service d'annotation de texte rapprochant un document textuel de concepts d'ontologies).

Les Viewpoints ne seront évalués qu'à l'utilisation quand un utilisateur interrogera le graphe de connaissances. Chaque utilisateur explore le graphe de connaissances sous l'angle d'une Perspective qu'il choisit librement. Il est donc impossible pour un utilisateur de créer des viewpoints et de prévoir l'impact de cette modification sur les futures recherches d'autres utilisateurs étant donné qu'il ne connaît pas leur perspectives. Il devient par exemple impossible d'émettre sciemment beaucoup de points de vue pour faire évoluer le graphe de connaissance dans un sens particulier. PageRank fut un temps abusé par la science de la redirection afin de promouvoir certaines pages car les résultats de l'opération de base étaient prévisibles [82]; ce biais potentiel est pallié dans notre approche.

[40] http://hal-lirmm.ccsd.cnrs.fr/

3.2.2 Perspectives et Knowledge Maps (KMs)

La Perspective (illustrée dans la Figure 10) permet de passer d'un graphe de connaissances subjectives non-interprété (KG) à une *Carte de connaissance* (KM). Chaque Viewpoint est évalué et participe à la construction de liens entre ressources qu'on appelle les Synapses. La Synapse représente un lien de proximité sémantique entre deux ressources. La KM est alors un simple graphe non dirigé $G = (R, S)$ ou S est l'ensemble des Synapses constituées par l'interprétation et l'agrégation des Viewpoints.

Chaque algorithme appliqué est un algorithme « subjectif », c'est-à-dire qu'il doit être lié à une Perspective. L'acte d'interprétation des Viewpoints procède en deux étapes. Premièrement, chaque Viewpoint est évalué. Ils peuvent être évalués selon plusieurs paramètres comme l'agent émetteur si on souhaite donner de la valeur aux Viewpoints d'un collègue ou d'un type ϑ ainsi que le temps τ.

La KM est créée à la volée au fur et à mesure de l'exploration par nos algorithmes. La polarité sert par exemple à affaiblir une relation : si l'utilisateur qui parcourt le graphe de connaissances par recherches successives obtient à un moment un résultat improbable, il corrige les connaissances collectives en exprimant au sein du KG que la ressource qu'il cherchait et ce résultat n'ont rien de « similaire ». Lorsque de nouvelles interactions sont enregistrées (Viewpoints) certaines synapses futures (créées à la prochaine exploitation du graphe) seront renforcées, d'autres s'atténueront voir disparaitront. Le KG évolue ainsi de manière très fluide et plastique à la manière des neurones du cerveau, en respectant la métaphore du darwinisme neuronal énoncé par G. M. Edelman dans son travail sur le système évolutif qu'est le cerveau [83]. L'hypothèse centrale d'Edelman est que la cartographie neuronale hypercomplexe du cerveau se construit par un processus sélectif. Les connexions les plus utilisées vont donc se renforcer et les autres disparaître, façonnant ainsi des réseaux de neurones uniques à chaque individu. Les circuits sélectionnés forment ce qu'Edelman appelle des cartes neuronales.

La fonction d'interprétation créant ces synapses à partir des Viewpoint est une application de l'ensemble des viewpoints V vers l'ensemble des réels :

$$^{u}map : V \rightarrow \mathbb{R}$$

Ensuite pour chaque couple $\{r_1, r_2\}$ connecté par un ou plusieurs Viewpoints les évaluations des Viewpoints sont agrégées. U est l'utilisateur fournissant sa Perspective à la fonction pour l'évaluation des Viewpoints. La fonction d'agrégation est donc l'application :

$$^{u}reduce : \mathbb{R}^n \rightarrow \mathbb{R}$$

Avec n le nombre de Viewpoints reliant(r_1, r_2). La façon la plus commune d'agréger les Viewpoints est de sommer leurs évaluations de la manière suivante :

$$^{u}reduce(V) = \sum_{i=1}^{n} {}^{u}map(v_i)$$

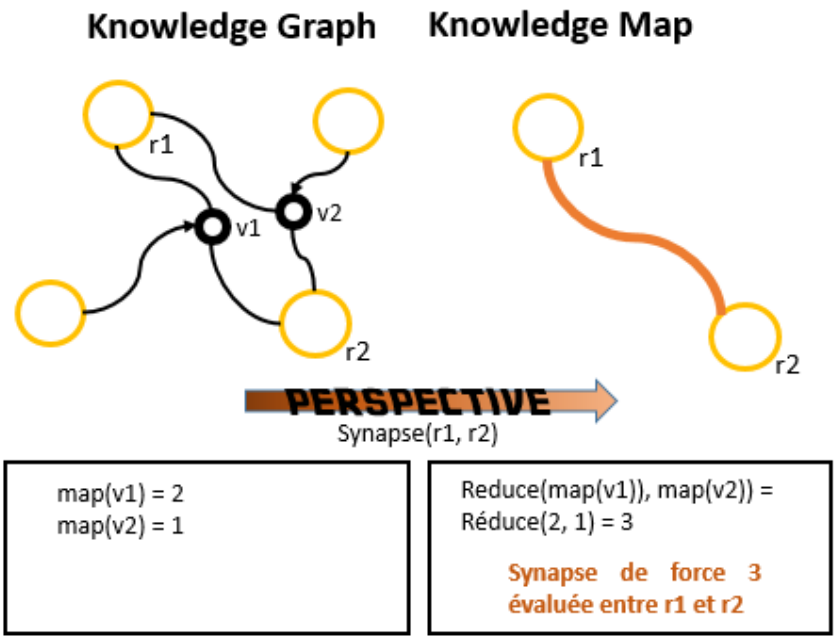


Figure 10 Mécanisme de Perspective.

Si le résultat de cette agrégation de Viewpoints est négatif alors nous ne créons aucune synapse.

Prenons un exemple de fonction d'interprétation des Viewpoints appartenant à la perspective d'un utilisateur: une base de données bibliographique est indexée sous forme de Viewpoints. Un ensemble de Viewpoints de types différents est créé : (i) les Viewpoints « authorship » rapprochant l'auteur à son document, (ii) les Viewpoints fruit de la catégorisation par annotation sémantique de texte rapprochant les documents aux descripteurs. Quand nous parcourons un KG muni d'une Perspective, nous pouvons premièrement interpréter les viewpoints en fonction de leur auteur en construisant une fonction map munie de la liste associative suivante associant chaque auteur à un coefficient de « réputation » :

$$reputations = Liste < A, \mathbb{R} >.$$

Nous empruntons les termes map et reduce qui appartiennent à la terminologie de la programmation orientée flux car le fonctionnement de la fonction d'interprétation et d'agrégation est similaire à celui des fonctions map et reduce de ce domaine. En effet, lors de l'exécution des méthodes sur KG, la KM est évaluée au fur et à mesure et seules les Synapses nécessaires au calcul sont construites.

L'ensemble des Viewpoints constitutifs de ces synapses peut être donc considéré comme un flux de Viewpoints sur lequel on « map » à chaque Viewpoint le résultat de la fonction d'interprétation et les résultats sont ensuite accumulés pour créer la Synapse. La fonction map peut s'écrire de la façon suivante :

$$\begin{cases} polarité\ positive: {}^{u}map(v) = reputations(émetteur(v)) \\ polarité\ négative: -{}^{u}map(v) = reputations(émetteur(v)) \\ polarité\ nulle: 0 \end{cases}$$

L'utilisateur averti peut également faire peser dans son utilisation plus le résultat d'algorithmes de catégorisation par fouille de texte ou les interconnexions du réseau d'auteur. Pour cela il est nécessaire de joindre une liste associant chaque type de Viewpoint à son « poids ».

$$poidsTypes = Liste < T, \mathbb{R} >$$

Avec T l'ensemble des types de Viewpoints. Nous pouvons alors panacher la précédente fonction map avec celle-ci :

$$\begin{cases} polarité\ positive\colon\ {}^{u}map(v) = reputations(émetteur(v)) \times poidsType(type(v)) \\ polarité\ négative\colon\ -\ {}^{u}map(v) = reputations(émetteur(v)) \times poidsType(type(v)) \\ polarité\ nulle\colon\ 0 \end{cases}$$

La Perspective et ses deux composants – map et reduce – forment donc un mécanisme extensible capable d'être adapté facilement. Le paramétrage des listes associatives que nous avons vues précédemment permet de spécialiser les algorithmes rattachés à la perspective au cas d'étude. Ces listes associatives correspondent en quelque sorte au génome de l'algorithme. Le nombre de paramètres dans la Perspective peut donc très vite augmenter. Nous évoquerons en annexe (Annexe 3) la capacité de transformer de problème NP-complet de recherche d'une perspective optimale en se basant sur l'algorithmique génétique.

3.3 Méthodes de gestion et d'exploitation du KG

Dans cette section, nous abordons les méthodes que nous avons développées afin d'exploiter le graphe de connaissances et la subjectivité qu'il contient. L'exploitation du graphe de connaissance nécessitait pour commencer de deux types de méthodes : (i) le calcul de distance sémantique entre deux ressources et (ii) le calcul du voisinage sémantique avec un rayon donné autour d'une ressource. Le résultat de la seconde méthode renvoie pour une ressource donnée l'ensemble des ressources proches sémantiquement. Il s'agit d'un résultat similaire à une recherche d'information.

3.3.1 Création de ressources et viewpoints

Le graphe de connaissances évolue sans cesse par l'ajout de nouvelles ressources et Viewpoints. Cet ajout peut être fait manuellement. Mais l'infrastructure logicielle donne également au développeur le moyen d'intégrer, grâce à des algorithmes de moissonnage adaptés aux différents formats standards du Web (json, rdf, xml), des masses de données pour les transformer en ressources et en Viewpoints. Plusieurs exemples allant de connaissances cinématographiques basées sur une base de données relationnelle à l'indexation de base de connaissances biomédicales en rdf sont fournis dans le Chapitre 4.

Dans les expérimentations que nous avons conduites jusqu'à présent, ViewpointS est purement cumulatif. C'est-à-dire que nous n'avons mis en œuvre que l'ajout de nouveaux viewpoints et de nouvelles ressources. Cela dit nous avons envisagé de permettre la suppression d'une ressource avec la suppression des viewpoints qui s'y rattachent. Ou l'anonymisation, comme forme de suppression de viewpoints. Un utilisateur se désinscrivant ne souhaitant plus savoir qu'une connaissance est gardée sur lui pourrait choisir d'anonymiser cette connaissance afin de la laisser tout de même en héritage aux utilisateurs du service.

3.3.2 Méthodes exploitant le graphe de connaissances

Les méthodes que nous présentons correspondent aux critères que nous nous sommes fixés d'après notre lecture de l'état de l'art. Premièrement ces méthodes sont applicables sur n'importe quel type de ressources. Il est par exemple possible de calculer le voisinage autour d'un agent ou une distance

sémantique entre deux agents. Ensuite, ces méthodes – très simples dans leur conception – ne nécessitent aucune structure spécifique des connaissances (ex. : structure taxonomique).

3.3.3 Calcul de voisinage sémantique

Le calcul de voisinage sémantique est relativement similaire à une recherche d'information. Toutefois il diffère fondamentalement par rapport à une recherche basée sur des métadonnées. La pleine exploitation de la transitivité des relations dans le graphe de connaissance prend en effet le risque de renvoyer des résultats indirectement liés. Mais dans l'approche du Web cognitivement sémantique il est selon nous tout à fait approprié d'adopter une approche basé sur une le voisinage topologique et se baser sur l'interaction (les feedbacks) avec l'utilisateur pour valider les résultats. Nous présentons trois méthodes.

La première (SPN – Shortest Path Neighbourhood) est une adaptation de l'algorithme de Dijkstra[41] et se base sur l'ensemble des plus courts chemins partant d'une ressource centrale pour déterminer le voisinage sémantique de cette ressource.

La seconde (MPN – Multiple Paths Neighbourhood) se base elle sur l'arbre de tous les chemins partant de la ressource centrale. Il s'agit d'une fonction de voisinage beaucoup plus lente mais complète car elle prend en compte pour déterminer la distance sémantique entre deux ressources l'ensemble des chemins de longueur bornée entre elles.

La troisième (MFN – Multiple Flows Neighbourhood) est une adaptation de SPN pour prendre en compte les chemins parallèles entre ressources. Il s'agit d'un compromis entre la complétude de MPN et la vitesse d'exécution de SPN. Si on définit la fonction générique de distance sémantique ${}^{u}\Psi(r_1, r_2)$ alors le voisinage s'écrit ${}^{u,m}Neighbourhood(r_0) = \{\forall r_i \in R \mid {}^{u}\Psi(r_0, r_i) \leq \mathrm{m}\}$.

Il existe aussi une quatrième méthode que nous avions développée afin de pallier au manque de performance de MPN et qui se basait sur un échantillonnage de chemins choisis aléatoirement dans l'arbre des chemins sur lequel se base MPN. Nous l'avons appelé Weighted Random Walk Neighbourhood (WRWN).

3.3.3.1 SPN

L'algorithme de Dijkstra a été conçu pour calculer le plus court chemin entre deux nœuds d'un graphe. Pour cela il calcule l'ensemble des plus courts chemins partant du premier nœud. Le plus court chemin entre les nœuds fait partie de cette arborescence de chemins. SPN prend cette première étape et l'adapte : nous calculons l'ensemble des plus courts chemins de longueur bornée à partir de la ressource centrale. Le paramètre principal de SPN est le diamètre du voisinage (m). Toutes les ressources présentes dans cette arborescence de plus courts chemins font partie du voisinage sémantique.

Le fait de borner la longueur des chemins explorés économise beaucoup de calculs. En effet, sans borner la longueur des plus courts chemins à explorer la complexité dans le pire des cas est $O(|V|^2 + |R|^2)$. Nous n'atteignons toutefois ce pire des cas que si nous fixons $m = diamètre(KG)$. Le diamètre étant la distance maximale entre deux nœuds d'un graphe fixer m à cette valeur revient à

[41] fr.wikipedia.org/wiki/Algorithme_de_Dijkstra

supprimer la borne. Toutefois le fait de fixer m à une valeur basse rend l'algorithme indépendant en termes de complexité de la taille du graphe en nombre de nœuds.

Pour trouver de manière dynamique la bonne valeur de m il est possible de l'exprimer comme une fraction du diamètre de KG. Ainsi, SPN devient:

$$^{u,p}SPN(r_0) = \{\forall r_i \in R \mid {}^{u}\Psi(r_0, r_i) \leq \mathrm{p} \times \text{diamètre(KG)}\}$$

Algorithme 1 Shortest Path Neighbourhood

Input: Resource r, float m
Result: $^{m}SPN - Neighbourhood(r)$
initialization;
todo = r;
while *todo* $\neq \emptyset$ **do**
  Resource $o_i = Todo.premier()$
  forall the *Resource* $o_k \in DirectNeighbourhood(o_i)$ **do**
    $newDistance = o_i.distance + 1/synapse(o_i, o_k)$;
    if $o_k.distance = \infty \wedge newDistance < o_k.distance$ **then**
      $Todo \cup \{o_k\}$;
      $ok.distance = newDistance$;
      if $o_k.distance \leq m$ **then**
        $Result \cup \{o_k\}$ **end**
      end
    end
  end
  return *Result*

3.3.3.2 MPN

Alors que SPN se base uniquement sur le plus court chemin pour rapprocher sémantiquement deux ressources MPN est une autre méthode qui exploite tous les chemins non-cycliques de longueur bornée entre ces ressources. Alors que SPN construit un arbre de tous les plus courts chemins à partir d'un nœud de départ MPN construit l'arbre de tous les chemins non-cycliques de longueur bornée. Ensuite, à partir de l'arbre de parcours nous construisons les distances de tous les nœuds présents dans cet arbre en fonction de tous les chemins qui y mènent.

Comme précédemment nous trouvons la borne en fonction en faisant un ratio du diamètre du graphe. Pour calculer le Diamètre en nous basant sur une distance au sens SPN nous nous basons sur l'ensemble des chemins non-cycliques entre deux ressources. Chacun de ces chemins de longueur d est alors simplifié sous la forme de sa synapse équivalente de force 1/d. Les deux ressources sont alors connectées par un ensemble de synapses que nous sommons pour obtenir une (« super-synapse »). La distance correspond donc à l'inverse de cette valeur. Formellement, pour deux ressources r_1, r_2 connectées par un ensemble de n chemins de longueurs $d_1, d_2, d_3 \ldots d_n$ nous calculons ainsi la distance entre r_1 *et* r_2 :

$$^{u}d_{MPN}(r_1, r_2) = \frac{1}{\frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3} + \cdots + \frac{1}{d_n}}$$

Notre méthode de calcul de distance MPN est donc très analogue à ces deux approches de calcul de flot maximum. On exprime ce calcul en d'autres termes en le présentant comme l'inverse du flot maximum entre deux ressources que nous appelons la source et le puits car nous ne cherchons pas un flot – équivalent à une proximité – mais une distance.

Algorithme 2 Multiple Paths Neighbourhood

Input: Resource r, float m
Result: $^{m}MPN - Neighbourhood(r)$
traversalTree : Tree;
result : Map(Resource, Float);
initialization;
$traversalTree = buildTraversalTree(r, m)$;
$result = buildDistances(traversalTree)$;
return *result*

Input: Resource r, float m
Result: TraversalTree(r)
traversalTree : Tree;
todo : List(Resource);
$traversalTree.root = TreeNode(r)$;
$todo \cup traversalTree.root$;
while $todo \neq \emptyset$ **do**
    TreeNode $current = todo.first$;;
    TreeNode $next = nextPossible(current, m)$;
    if $next \neq NIL$ **then**
        todo.removeFirst();
    else
        todo.addFirst(next);
    end
end
return *traversalTree*

Toutefois, comme cette fois-ci nous ne prenons pas en compte les plus courts chemins mais l'ensemble des chemins non-cycliques de longueur bornée, ainsi, plus cette borne sera grande plus nous prendrons en compte de chemins et meilleure sera la construction d'une distance basée sur les subjectivités des viewpoints.

3.3.3.3 WRWN

Weighted Random Walk neighbourhood (WRWN) est une méthode de calcul de voisinage sémantique qui reprend le principe de MPN. Toutefois au lieu d'explorer l'arbre de tous les chemins de longueur bornés partant du nœud central il n'explore qu'un échantillon aléatoire de ces chemins. Il s'agit donc d'une version échantillonnée et approximée de MPN. Comme précédemment, nous nous basons sur la construction d'un arbre de parcours. Ce choix aléatoire-pondéré est pondéré par le

poids des synapses menant aux ressources du voisinage direct. Pour ce faire nous construisons une distribution statistique du pourcentage de chance d'aller vers les voisins directs de la ressource donnée. Nous nous basons sur le poids des synapses pour déterminer ces probabilités. La Figure 11 illustre cette distribution statistique et comment le choix est fait. La construction de l'arbre de parcours une fois que nous avons construit le nombre souhaité de chemins. Dans l'exemple de la Figure 11, la synapse partant vers la ressource 3 étant beaucoup plus forte, le chemin passant par cette ressource aura beaucoup plus de chances d'être parcouru (60%). Le taux d'échantillonnage – c'est-à-dire le pourcentage des chemins explorés parmi l'ensemble des chemins de longueurs bornés partant du nœud central – permet d'ajuster la précision de l'algorithme. Si l'algorithme se base sur un taux d'échantillonnage faible alors il sera beaucoup plus rapide mais l'évaluation des distances entre deux ressources se basera sur moins de chemins et représentera moins de viewpoints dans la construction subjective de la distance.

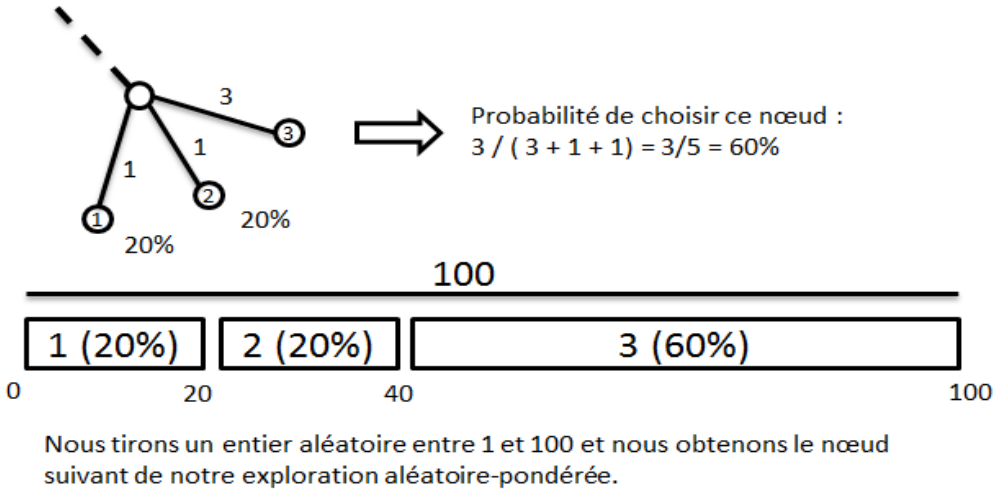


Figure 11 Fonctionnement du choix aléatoire pondéré.

Algorithme 3 WeightedRandom Walk

```
Input: Resource r, float m
Result: TraversalTree(r)
traversalTree : Tree;
todo : List(Resource);
traversalTree.root = TreeNode(r);
todo ∪ traversalTree.root;
while todo ≠ ∅ do
    TreeNode current = todo.first;;
    TreeNode next = nextRandomPossible(current, m);
    if next ≠ NIL then
        todo.removeFirst();
    else
        todo.addFirst(next);
    end
end
return traversalTree
```

3.3.3.4 MFN

L'algorithme MFN calcule le voisinage sémantique d'une ressource donnée en déterminant les éléments de ce voisinage de la même manière que SPN. Toutefois, toutes les synapses qui constitueront les plus courts chemins constitutifs du voisinage de SPN, seront ensuite post-traitées. Pour chaque synapse connectant directement deux ressources nous prendrons en compte dans le post-traitement tous les chemins indirects de longueur bornée connectant ces deux ressources. Cela permet de bénéficier de la vitesse de traitement de l'approche SPN tout en la corrigeant pour qu'elle comptabilise dans la construction subjective de la réponse à la fois les viewpoints contribuant au plus court chemin mais aussi les viewpoints contribuant à des chemins indirects.

Nous obtenons donc un voisinage de ressources dont les distances sont post-traitées. Il est une dérivation de SPN et se base au début sur le même fonctionnement inspiré de l'algorithme de Dijkstra. Si nous souhaitons calculer le voisinage autour d'une ressource r d'un diamètre donné (m) autour d'une ressource centrale nous explorons tous les plus courts chemins partant de cette ressource. L'ensemble des ressources sur les chemins que nous venons de calculer font partie du voisinage. La distance de chacune des ressources du voisinage avec la ressource centrale est la longueur du plus court chemin entre cette ressource voisine et la ressource r. L'idée est de réévaluer ces distances en prenant en compte non plus le seul plus court chemin mais l'ensemble des chemins connectant ces ressources. Cela nous permet de bénéficier de la rapidité de SPN pour la détermination du voisinage tout en utilisant l'approche basée sur de multiples chemins pour retrier les éléments de ce voisinage.

La Figure 12 donne un exemple de fonctionnement de MFN. Sur le graphe de départ (a), nous souhaitons obtenir le voisinage de la ressource 0, ainsi, nous calculons l'ensemble des plus courts chemins de longueur bornée partant de 0. Pour chaque chemin partant de 0 dans cet arbre nous recalculons chacune des synapses en prenant en comptes d'autres chemins (en traits fins sur (b)).

Par exemple, le chemin entre 0 et 7 est mis à jour deux fois. Premièrement (b), lorsque r0 et r4 sont connectés directement et deuxièmement, par les chemins r0-r3-r7-r4 et r0-r8-r4. La synapse r0 est donc renforcée en accord avec la formule détaillée dans la fiche MPN. La synapse r4-r7 est également remise à jour suivant le même principe (c).

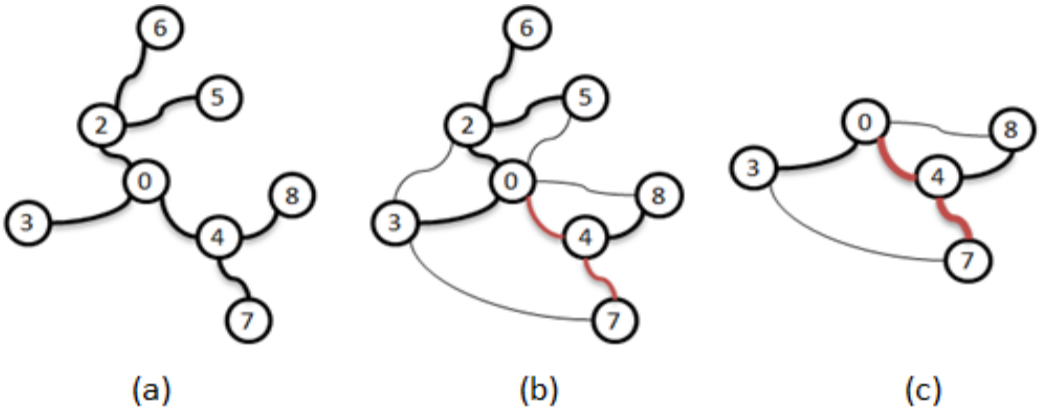


Figure 12 Exemple sur le fonctionnement de MFN

Algorithme 4 Multiple Flows Neighbourhood

Input: Resource r, float m, float p
Result: MFN(r, m, p)
traversalTree : Tree;
$shortestPathsTree = buildShortestPathsTree(r, m)$;
forall the *Path path* $\in$ *shortestPathsTree starting from r* **do**
  forall the *Synapse* $s(r_1, r_2) \in path$ **do**
    computeFlow(s, p);
  end
end
$result = buildDistances(shortestPathsTree)$;
return *result*

Input: Synapse s(r_1, r_2), float p
Result: computeFlow(s, p)
$pathMaxLength = \frac{1}{p \times s.strength}$;
$traversalTree = buildTraversalTree(r, pathMaxLength)$;
return getDistance(r_1, r_2);

3.3.3.5 Exemple comparatif

Montrons maintenant un exemple. La Figure 13 montre une Knowledge Map de test sur lequel nous allons essayer trois des méthodes de voisinage que nous présentons : SPN, MPN et MFN. WRWN étant une version approximée de MPN nous l'ignorerons, son résultat avec 100% de précision étant équivalent à MPN. Nous recherchons le voisinage sémantique de r_0 (Tableau 6) trié par distance décroissante pour chaque méthode avec les paramètres précisés dans la Tableau 5. Pour la simplicité de la comparaison nous disons que chacune des synapses à une force de 1.

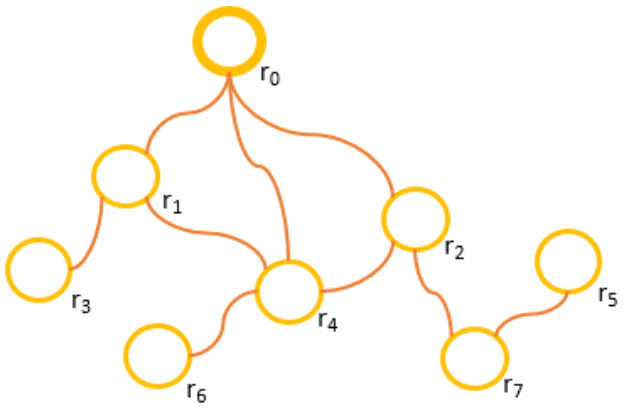


Figure 13 Exemple de Knowledge Map

Tableau 5 Paramètres des méthodes de voisinage sémantique

Paramètre	Valeur
m	3
p (MFN)	3

Tableau 6 Voisinage de r_0

SPN	MPN	MFN
r_1 (1)	r_4 (0.5)	r_4 (0.5)
r_4 (1)	r_1 (0.54)	r_1 (0.54)
r_2 (1)	r_2 (0.54)	r_2 (0.54)
r_3 (2)	r_6 (0.85)	r_6 (0.85)
r_6 (2)	r_3 (1.2)	r_3 (1.2)
r_7 (2)	r_7 (1.2)	r_7 (1.2)
r_5 (3)	r_5 (3)	r_5 (3)

On observe dans l'ordre des résultats par rapport à la topologie du KG de la Figure 13 que les ressources que MFN et MPN renvoie en premier sont les ressources les plus connexes par lesquelles passent le plus de chemins possibles. Des ressources qui étaient avec SPN à la même distance sont maintenant différenciées en fonction du nombre de chemins qui y mènent. Le tri se fait donc plus « finement » dans les méthodes se basant sur de multiples chemins. Dans MPN, m prend un rôle plus grand que la définition du seul rayon de voisinage. En effet, plus m est grand plus MPN prendra en compte de chemins plus les résultats s'affinent. Il en est de même pour la valeur de p pour MFN qui a un fonctionnement différent que pour MPN. En effet, pour la réévaluation d'une synapse plaçant deux ressources à distance donnée d l'une de l'autre nous nous permettons de chercher l'ensemble des chemins de longueur p x d connectant ces deux ressources. La borne de longueur maximum pour les chemins parallèles à explorer qu'on obtient à partir de p s'adapte donc à la synapse contrairement à MPN pour laquelle la borne est fixe. Nous observons ce gain de performance de MFN par rapport à MPN dans l'Annexe 2. Cela fait de MFN une version optimisée de MPN.

3.3.4 Calcul de distance sémantique

Chacune des méthodes de calcul de voisinage sémantique – SPN, MPN ou MFN – peut être dérivée en une méthode de calcul de distance sémantique (respectivement : SPD, MPD et MFD). Pour calculer la distance entre deux ressources $r_1 \, et \, r_2$ nous calculons le voisinage ${}^{u,perimeter(KG)}Neighbourhood(r_1)$. Si $r_1 \, et \, r_2$ sont connectés dans KG alors r_2 fait partie de ce voisinage, sa distance sémantique a donc été évaluée et nous la renvoyons.

Par exemple, pour calculer la Shortest Path Distance (d_{SPN}) entre deux ressources r_1 et r_2 à partir de la méthode de calcul de voisinage SPN nous calculons le voisinage de r_1 avec un rayon m maximum. Si r_1 et r_2 sont connexes alors r_2 fera partie de ce voisinage et sa distance à r_1 sera la longueur du plus court chemin entre ces ressources.

3.3.5 Métriques sur la structuration des connaissances

Lors de l'expérimentation des princes de Serendip (0), nous avons été confrontés au besoin de décrire, qualifier, mesurer la structuration des connaissances avec des méthodes de graphes. Les premières tentatives ont produit des mesures spécifiques telles que celles présentées en section 0. Au fil des expérimentations, nous avons abouti à une mesure unique et générique appelée « LocalHomogeneity » et présentée dans [84] (acceptée et en attente d'être publiée) .

L'idée générale peut être illustrée dans un exemple. Dans l'application de ViewpointS à la recommandation de films (4.4) nous observons qu'au voisinage d'un film gravitent des films de même genre. Au sein de la Knowledge Map, les films sont donc « homogènes » en genre. D'après Klir [85] on peut interpréter cette homogénéité locale en disant que la topologie du jeu de données sur lequel on se base contient l'information sur le genre des films.

Plus formellement, si on considère un graphe de connaissances KG contenant R ressources et D descripteurs, nous mesurons si des éléments appartenant au voisinage sémantique de R ont des éléments similaires de D dans leurs voisinages respectifs. L'homogénéité locale est la probabilité de trouver des éléments similaires, au sens défini dans des Vector Space Model (VSM), à ceux de D dans les voisinages des éléments proches de R. Nous définissons la fonction « local homogeneity » de la façon suivante :

Prenons une perspective d'un utilisateur u, m une borne pour le calcul du voisinage et R et D deux ensembles de ressources, un vecteur descriptive servant au VSM qu'on appellera Desc(r_i) tel que Desc(r_i) est le nombre d'occurrences de d_j dans le ${}^{u,m}Neighbourhood(r_i)$.

^{U,m,D}localHomogeneity(R)est la valeur moyenne de toutes les distances $d_{VSM}(r_i, r_j)$ calculées sur toutes les paires (r_i, r_j) tel que $d_{VSM}(r_i, r_j) \leq m$.

3.3.6 Renforcement et affaiblissement des synapses et influence sur les voisinages

Dans l'approche ViewpointS la distance sémantique est une propriété qui émerge des interactions, c'est-à-dire de la création de viewpoints. La Knowledge Map, constituée des synapses construites par le mécanisme de Perspective, est une carte où des chemins se tracent, se renforcent ou s'estompent au fur et à mesure des interactions. A la manière du cerveau comme système évolutionniste tel que le décrit G. Edelman dans [83], [86] un processus de sélection fait évoluer avec fluidité et plasticité la population des synapses. Le processus de sélection dans ViewpointS est l'expression des sémantiques individuelles à l'aide de viewpoints.

Le processus de feedback dans ViewpointS renforce ou atténue les synapses et des chemins dans le graphe de connaissances. Prenons un exemple (Figure 14) : Un utilisateur obtient une ressource r_2 pour une recherche du voisinage sémantique de r_1.

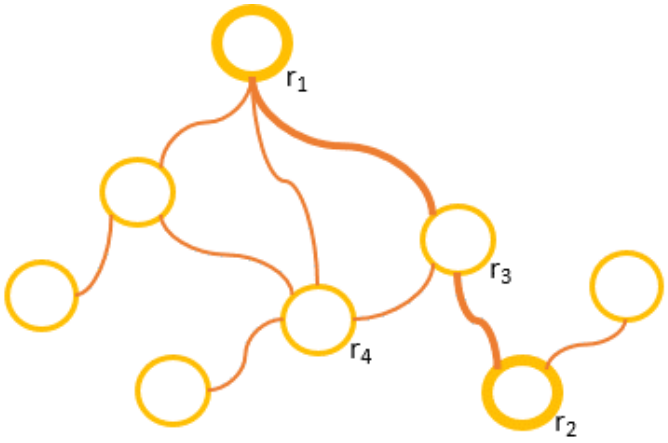


Figure 14 Exemple d'un voisinage sémantique

Il peut soit « liker »/ « disliker » r_2. Un viewpoint le connectant à r_2 est alors créé avec une polarité positive ou négative. Il peut aussi juger de la pertinence de r_2 par rapport à r_1. Toutes les synapses s_i à s_n du plus court chemin de r à r_0 sont alors renforcées ou atténuées par des viewpoints de type FragmentedFeedback. Le poids du type FragmentedFeedback est distribué en part égale sur chaque synapse du chemin avec un viewpoint qui sera interprété de manière spécifique. Chaque viewpoint de feedback fragmenté ajouté pour modifier les synapses du chemin sera interprété de la manière suivante :

$$polarité\ positive:\ {}^{u}map(v_{frag}) = \frac{poidsType(type(v))}{n}$$

$$polarité\ négative:\ -\ {}^{u}map(v_{frag}) = \frac{poidsType(type(v))}{n}$$

Voyons maintenant l'impact du renforcement des synapses du chemin r_1-r_2 sur les trois distances sémantiques que nous proposons. $d_{SPN}(r_1, r_2)$ change entièrement puisque les synapses du plus court chemin sur lequel elle base son calcul sont renforcées ou atténuées. Dans le cas des deux autres mesures l'impact est plus atténué. En effet, le fait que d_{MPN} et d_{MFN} se basent aussi sur des chemins indirects telle que r_1-r_4-r_3. Etant donné que le feedback n'as rien changé des synapses de ce chemin une partie ce que qui fait le résultat de d_{MPN} et d_{MFN} est inchangé. Pourd_{MFN}, le plus court chemin est certes utilisé dans la construction initiale de la distance mais le chemin r_1-r_4-r_3 sert ensuite à réévaluer la synapse r_1-r_4. d_{MPN} considère à la base de son fonctionnement l'exploration de chemins parralèles.

EN RESUME

Nous proposons dans le formalisme ViewpointS :

- Un graphe de connaissances apte à intégrer des connaissances de domaines variés et structures très diverses et à représenter leur subjectivité. Pour cela nous réduisons la richesse des relations sémantique à un minimum n'incluant que la relation sémantique la plus universelle : celle de la similarité sémantique.

- Une perspective propre à chaque utilisateur permet d'interpréter de façon personnalisée les subjectivités représentées dans le graphe de connaissances.

- Des méthodes de calcul de distance sémantique permettent d'obtenir une distance sémantique entre n'importe quel couple de ressources sans prérequis sur la structuration des connaissances.

- Des méthodes de calcul de voisinage sémantique basées sur la distance sémantique qui permettent la recherche d'information.

- Un mécanisme de feedback qui, après une recherche, renforce ou affaibli les chemins qui mènent de la ressource recherchée à la ressource renvoyée. Cela permet de creuser des chemins dans le graphe de connaissances qui auront plus de chances d'être empruntés par d'autres utilisateurs.

Chapitre 4. Expérimentations

Article référence:

P. Lemoisson, G. Surroca, and S. A. Cerri, "Viewpoints: An Alternative Approach toward Business Intelligence," in eChallenges e-2013 Conference, 2013, p. 8.

Ce chapitre fait état de toutes les expérimentations et évaluations que nous avons menées. Il commence par une preuve de concept sur la capacité d'apprentissage du graphe de connaissances ViewpointS [1]. Une première application de ViewpointS s'axa sur la recherche d'information. Des données bibliographiques (publications, auteurs, journaux, conférences) permettaient la constitution du premier exemple de graphe de connaissances ViewpointS à partir duquel nous construisions notre premier prototype de moteur de recherche [87]. Cette première preuve de concept est ensuite améliorée dans le cadre une expérimentation sur l'impact des stratégies de navigation des utilisateurs sur la construction du Web [88][89]. En particulier, nous souhaitons capturer le phénomène de Sérendipité (i.e., de l'apprentissage fortuit) à l'aide d'un formalisme de représentation des connaissances subjectives où un ensemble de points de vue forment un graphe de connaissances interprétable de façon personnalisée. Nous établissons une preuve de concept sur la capacité d'apprentissage collectif que permet ce formalisme appelé Viewpoints en construisant une simulation de la diffusion de connaissances telle qu'elle peut exister sur le Web grâce à la coexistence des données liées et des contributions des utilisateurs.

A l'aide d'un modèle comportemental paramétré pour représenter diverses stratégies de navigation Web, nous cherchons à optimiser la diffusion de systèmes de préférences. Nos résultats nous permettent d'identifier les stratégies les plus adéquates pour l'apprentissage fortuit et d'approcher la notion de Sérendipité. Nous évaluons ensuite dans la section 1.2 l'approche Viewpoints sous l'angle des systèmes de recommandation. Cette évaluation a pour objectif de confronter le formalisme à un premier cas d'étude sur les films et de comparer entre elles les méthodes d'exploitation du graphe des connaissances. Une fois les forces et faiblesses de nos méthodes identifiées nous comparons dans la section 1.3 une partie d'entre elles – les mesures de distance sémantique – à plusieurs autres mesures de distance sémantique de l'état de l'art [90]. Pour cela, nous utilisons un jeu de données qui a précédemment servi à d'autres évaluations de distances sémantiques : WordNet.

La section 1.5 apporte l'évaluation d'un des potentiels usages des distances sémantiques : l'alignement d'ontologies. Nous utilisons la distance sémantique comme heuristique pour déterminer si deux concepts appartenant à deux ontologies distinctes peuvent être alignés. Ce chapitre se conclu sur la présentation de la dernière version du prototype de moteur de recherche ViewpointS que nous appelons VWA (Viewpoints Web Application) et qui est détaillé dans le Chapitre 5.

4.1 Preuve de concept sur la capacité de d'apprentissage du graphe de connaissances

4.1.1 Objectifs

Ayant conçu une première version du formalisme nous souhaitions tester la capacité d'apprentissage du graphe de connaissance. Pour cela nous simulons l'évolution d'un Web fictif qui est parcouru par des utilisateurs. Les utilisateurs enrichissent par leurs feedbacks le graphe de connaissance. Chaque feedback est un viewpoint. Et chacun de ces viewpoints de feedback améliore les prochaines recherches sur le graphe de connaissance. Par rapport à un résultat r, l'utilisateur simulé u peut :

- « Liker » un résultat de recherche et émettre un viewpoint (u, {u, r}, {Like,+}, τ),
- Déclarer le résultat comme pertinent par rapport à sa recherche initiale s en créant un viewpoint (u, {r, s}, {Similar,+}, τ).

Alors que nous développions cette première expérimentation nous ne disposions que d'une seule méthode de calcul de voisinage sémantique : SPN.

4.1.2 Graphe de connaissance

Un graphe de connaissances est initialisé contenant un ensemble fixe constitué de documents (D), agents (A) et de topics(T) $O=A \cup D \cup T$. Un ensemble de viewpoints V_0 est créé entre des ressources aléatoires de O.

4.1.3 Déroulement de l'expérimentation

Nous suivons ensuite la séquence suivante :

- Un agent a_i est choisi au hasard dans A,
- L'agent a_i fait une recherche autour d'une ressource aléatoire o_q dans O,
- SPN calcule le voisinage sémantique de o_q que nous appelons R. Nous considérons ce voisinage comme l'ensemble des résultats pertinents pour cette recherche,
- Nous appelons ${}^{known}R$ l'ensemble des résultats connus dans R, c'est-à-dire les ressources sur lesquelles a_i n'ont émis aucun viewpoint. Prenons ensuite ${}^{new}R = R - {}^{known}R$ l'ensemble des ressources nouvelles,
- Avec une probabilité β a_i émet un viewpoint (a_i, {o_q, o_k }, {Like,+}, τ) sinon émet un viewpoint (a_i, { o_q, o_k }, {Like,-}, τ) β est ce que nous appelons la « perméabilité » de l'agent à une nouvelle information,
- Nous calculons ensuite la satisfaction qui est la proportion de ressources dans ${}^{new}R$ sur lesquelles a_i a émis un viewpoint positif,
- Nous calculons la pertinence qui est la proportion de ressources de R sur lesquelles a_i a émis un viewpoint positif.

La simulation se découpe en plusieurs Go. Chaque Go est composé de plusieurs Runs exécutant la séquence précédente. Ainsi nous faisons évoluer ce Web fictif sur plusieurs étapes chacune apportant plusieurs enrichissements au graphe de connaissance par feedback.

4.1.4 Résultats

La Tableau 7 résume l'ensemble des paramètres que nous avons utilisés pour produire les résultats de cette section.

Tableau 7 Paramètres de l'expérimentation

Card(A) = 20	Nombre d'agents
Card(D)=10	Nombre de documents
Card(T)=20	Nombre de topics
Card(V_0)=10	Nombre initial de viewpoints
m=1	Paramètre de rayon de voisinage pour SPN
X=10	Nombre de Runs
Y=50	Nombre de Gos
β1= 0%	Perméabilité pour la simulation 1
β2= 10%	Perméabilité pour la simulation 2
β3= 30%	Perméabilité pour la simulation 3

Les graphes de la Figure 15 nous fournissent les résultats suivants :

- La satisfaction moyenne des agents augmente avec la perméabilité. Mais cela était attendu.
- La courbe de pertinence est beaucoup plus intéressante. Quand la perméabilité est nulle, la pertinence croît de 0 à 0.5 d'une façon linéaire. Cela peut être expliqué par le fait que les ressources trouves qui sont inconnues deviennent de plus en plus distantes. Il y a donc dans les voisinages sémantiques de moins en moins de résultats. Cela mène à l'augmentation du ratio de réponses pertinentes.
- Quand les agents acceptant d'apprendre de la communauté, la pertinence moyenne augmente proportionnellement à la perméabilité. Avec β=30%, la pertinence moyenne est plus élevée qu'avec β=10%. Une perméabilité faible exclue des résultats qui ne seront jamais reconnus comme pertinents.

Le résultat principal à montrer dans cette expérimentation est la capacité d'apprentissage du graphe de connaissance. Au fur et à mesure que les viewpoints sont ajoutés les distances évalués sur le graphe de connaissances évoluent car les synapses sur lesquelles elles sont basées se renforcent, s'atténuent, se créent ou disparaissent. Le graphe de connaissances s'adapte et fait bénéficier à tous les agents de chaque feedback.

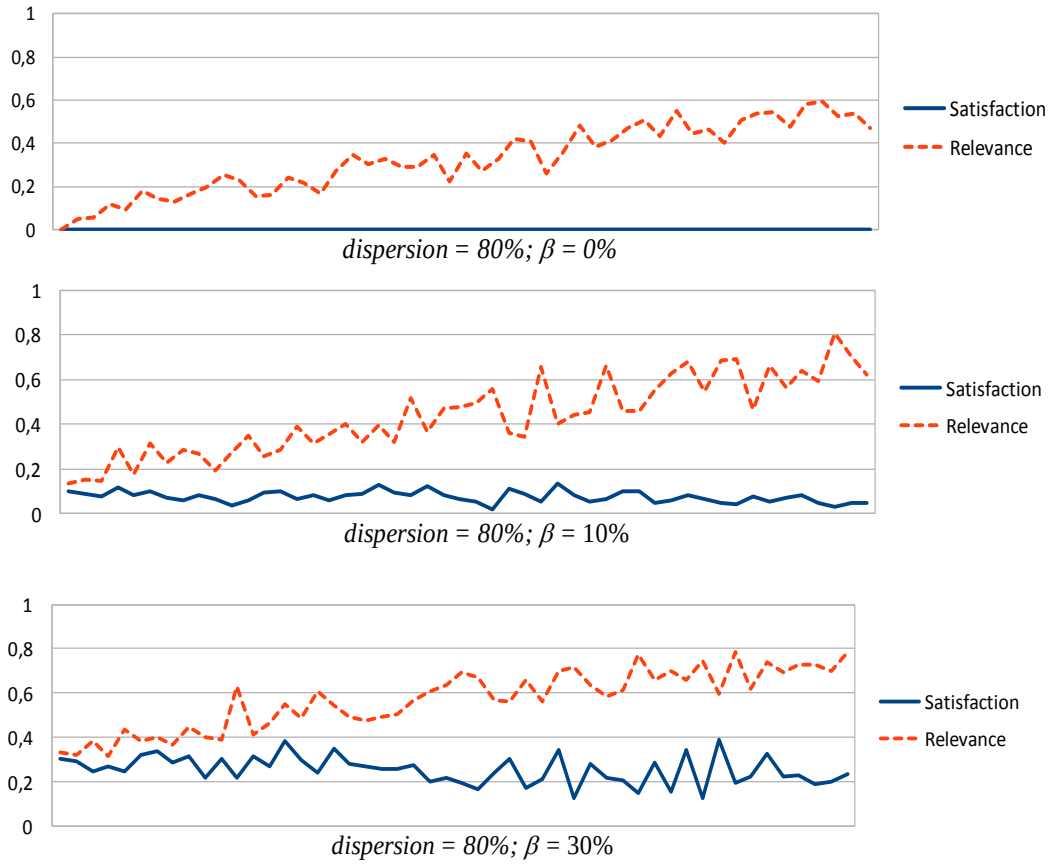


Figure 15 Courbes de satisfaction et pertinence

EN RESUME

La premier algorithme de calcul de voisinage sémantique fonctionne dans une simulation de recherches & interactions et s'améliore sur la base de l'apprentissage du graphe de connaissances par feedback. Au fur et à mesure de l'enrichissement du graphe de connaissances par les feedbacks on observe l'amélioration de la satisfaction moyenne des utilisateurs simulés.

4.2 Recherche de connaissances dans une base de publications scientifiques

Article référence:

G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, "Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique," IC - 25èmes Journées francophones d'Ingénierie des Connaissances. Clermont-Ferrand, France, pp. 175–186, 2014.

4.2.1 Objectifs

Afin de favoriser la dissémination et d'améliorer la présentation de notre projet nous décisions pour commencer d'exemplifier notre approche en construisant un prototype. Il s'agissait d'un prototype de moteur de recherche et son but était de nous laisser explorer les usages possibles de ViewpointS en termes de Recherche d'Information (RI).

4.2.2 Graphe de connaissance

Pour construire une application et tester cette approche sur des données réelles, nous avons choisi les ressources bibliographiques venant de la plateforme HAL-LIRMM[42] et HAL-LIRMM [43]comme corpus de données à indexer avec ViewpointS. C'est une base de données de toutes les publications du LIRMM et du CIRAD. Notre choix s'est porté vers cette ressource car :

- Chaque document est accompagné de métadonnées telles que les auteurs et les mots-clés choisis par ces auteurs pour décrire leurs publications. Cela en fait un jeu de donnéess approprié pour illustrer le potentiel du formalisme.
- Nous avons accès à ce jeu de donnéess de taille raisonnable.
- Ces données concernent nos collègues et notre laboratoire ce qui est donc pertinent pour évaluer/calibrer l'approche ViewpointS et motiver nos collègues à fournir leur évaluation et leurs viewpoints lors du feedback.

Dans un souci de comparaison et d'alignement avec le moteur de recherche fourni par HAL-LIRMM et HAL-CIRAD, nous avons sélectionné les métadonnées les plus simples pour l'initialisation du graphe de connaissance : pour un NumericDocument d, pour chaque auteur a (HumanAgent) et pour chaque mot-clé t (Descriptor) nous créons un viewpoint (a, {d, t}, {Match,+}, 0). Nous ne nous basons pas sur la dimension temporelle dans cette expérimentation. Cette procédure est répétée sur chaque NumericDocument. Dans notre application, basée sur les données de septembre 2013, 1663 HumanAgents, 5219 NumericDocuments et 5846 Descriptors nous donnent 42860 viewpoints.

4.2.3 Fonctionnalités

Le moteur de recherche permet à partir de n'importe quelle ressource du KG d'obtenir les documents, agents ou topics pertinents (c'est-à-dire proches sémantiquement). Il a été conçu comme un outil d'exploration d'état de l'art. Commencer un état de l'art c'est souvent commencer avec quelques noms d'auteurs, quelques mots-clés et un article ou deux. A partir de cela le moteur de re-

[42] http://hal-lirmm.ccsd.cnrs.fr
[43] http://hal.cirad.fr/

cherche doit permettre par exemple pour un mot-clé d'obtenir tous les mots-clés liés et d'apprendre le vocabulaire, la liste des auteurs les plus représentatifs de ce domaine ou des articles qui sont liés. Il permet aussi d'identifier des communautés d'agents partageant des sujets de recherche. Il fait cela en calculant le voisinage sémantique autour de la ressource souhaitée (Fournisseur, descripteur ou support de connaissance). Faire une recherche sur un descripteur comme Sérendipité revient à calculer son voisinage ${}^{u,m}\mathrm{Neighbourhood}(r_{\text{Sérendipité}})$ avec une Perspective u. Ce moteur de recherche ne se basait alors que sur une variante de l'algorithme de calcul de voisinage sémantique : SPN. Cela serait impossibles si (i) nous ne représentions pas l'agent de manière intrinsèque dans notre formalisme et (ii) nous nous basions sur une indexation de document par mot-clé. La notion plus floue de proximité sémantique permet d'obtenir des voisinages sémantiques d'éléments liés mais le plus souvent indirectement liés.

Prendre le parti d'une sémantique floue uniquement à base de relation de distance c'est aussi prendre le risque de renvoyer des résultats incorrects. C'est là que l'interaction humaine – dans le paradigme du Web Cognitivement Sémantique – est essentielle. Les feedbacks permettent de corriger les erreurs mais peuvent être surtout créateurs de nouvelles associations originales qui auraient été éclipsées dans une recherche par mot-clé. On peut réagir sur ce prototype en « récompensant » ou « punissant » les réponses. Cela a le même effet que dans un cerveau de renforcer ou atténuer les synapses.

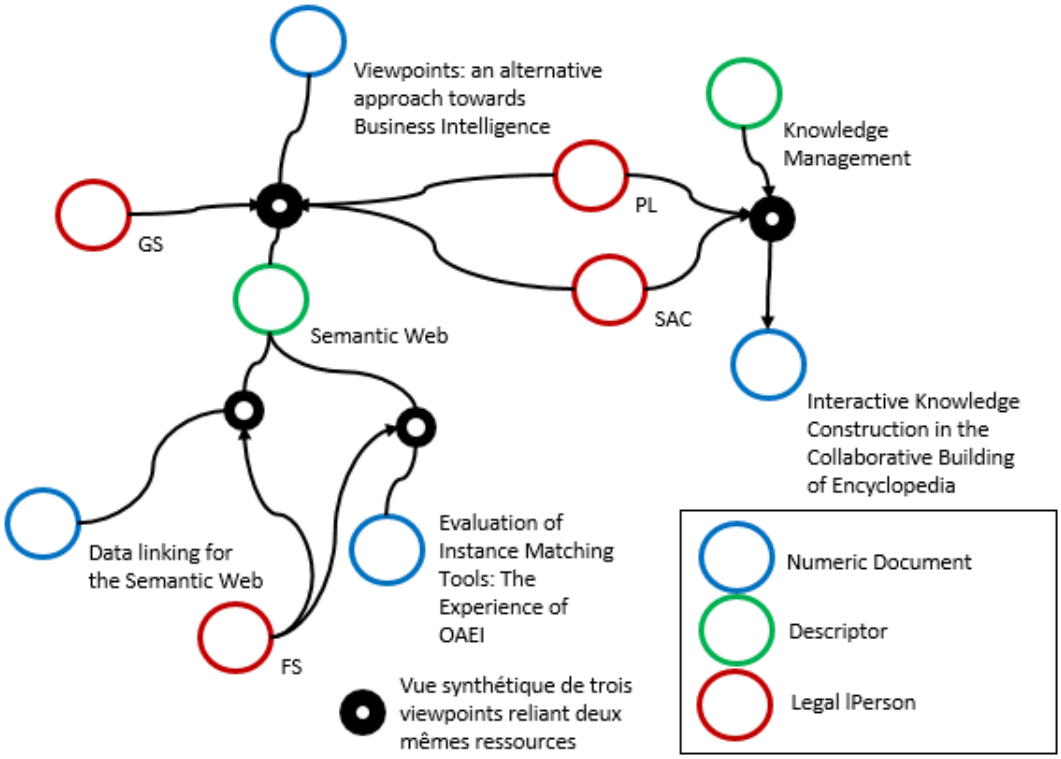


Figure 16 Sous graphe d'objets au voisinage de 'Semantic Web' extraits à partir de KG.

La Figure 16 illustre un sous-ensemble des objets au voisinage du Topic 'Semantic Web'.[44] Une recherche de 'Semantic Web' sur HAL-LIRMM ne retourne que les objets qui sont directement liés à

[44] Pour garder les schémas lisibles nous affichons dans le graphe de connaissances seulement les nœuds servant à l'illustration.

cette requête, c.-à-d. les articles ayant 'Semantic Web' dans leurs mots-clés ainsi que les auteurs de ces articles. Cependant, dans notre prototype, grâce à l'utilisation de la distance sémantique, la requête 'Semantic Web' renvoie en une seule recherche le ^{u,m}Neighbourhood de ce topic, c.-à-d. toutes les ressources pour lesquelles il existe un chemin de longueur inférieure ou égale à 'm' vers ce Topic; par exemple 'Data Linking' ou l'article 'Viewpoints : an alternative approach towards business intelligence'.

4.2.4 Exemple d'utilisation

Cette section montre un exemple d'utilisation du prototype et l'impact des feedbacks sur le graphe de connaissances.

Étape 1 : Guillaume Surroca (GS) exécute une recherche sur le Topic 'Knowledge Management'.[45] L'interface présente alors les résultats dans trois onglets ('Documents', 'Agents', 'Topics') comme illustré dans la Figure 17. L'agent François Scharffe (FS) y figure à une distance de 0,58 de l'objet recherché. Le prototype donne à l'utilisateur l'explication des résultats qu'il propose : en effet, l'utilisateur peut visualiser pour chaque résultat un des plus courts chemins reliant l'objet de la requête et le résultat dans le graphe de connaissances (bouton 'Path'). De plus, l'utilisateur peut valider chaque résultat (boutons vert) et émettre ainsi en guise de feedback de nouveaux viewpoints qui viendront nourrir le graphe pour les prochaines requêtes comme illustré dans la suite du scénario.

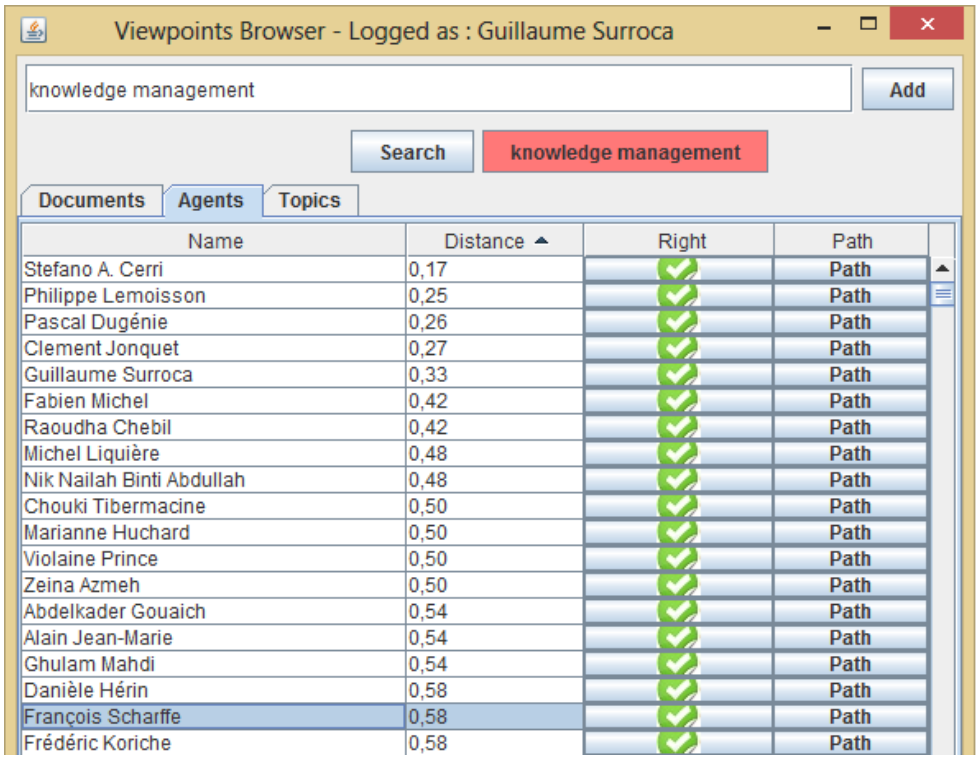


Figure 17 Illustration d'une recherche sur 'Knowledge Management' dans l'interface. Le nom de l'utilisateur connecté apparait et permettra d'identifier l'émetteur des viewpoints lors du feedback.

[45] Dans le prototype, un utilisateur saisit une chaine de caractères qui permet d'identifier l'objet de la requête (document, agent ou topic) par auto-complétion c'est-à-dire en se limitant explicitement aux objets de connaissance déjà présents dans le graphe.

Étape 2 : Ensuite, Clément Jonquet (CJ) fait une recherche sur le Topic 'Semantic Web', et obtient comme résultat l'article 'Interactive Knowledge Construction in the Collaborative Building of an Encyclopedia' (IKC) ; son feedback consiste à approuver le résultat en émettant un nouveau viewpoint reliant ce document au Topic 'Semantic Web'. La Figure 18 illustre le graphe de connaissances après la contribution de CJ. Ce nouveau viewpoint, de poids α=3 contribue à une synapse (IKC, Semantic Web) plus forte que les synapses précédentes (SAC, Semantic Web) ou (PL, Semantic Web) ; ainsi il existe un nouveau plus court chemin et la distance diminue.

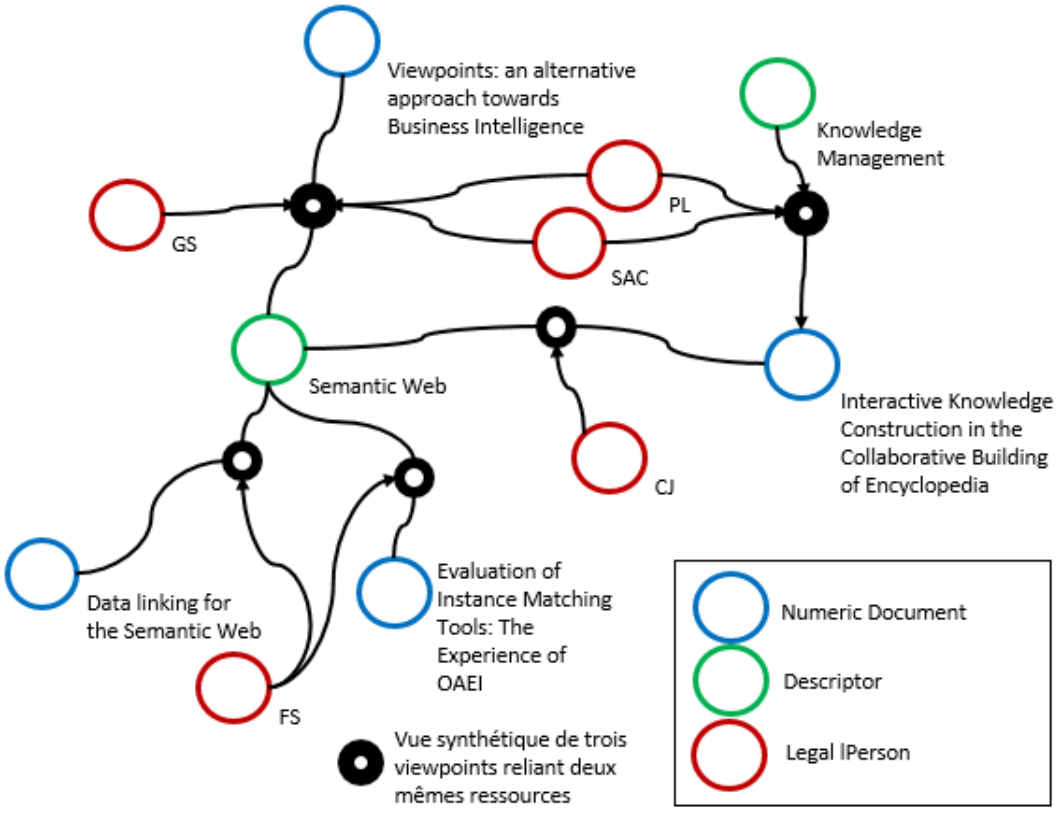


Figure 18 Impact du feedback de CJ sur le graphe de connaissances.

Étape 3 : Finalement, GS refait une recherche (Figure 19) sur 'Knowledge Management', et cette fois-ci l'agent FS apparaîtra plus haut dans la liste des résultats (ordonnés par distances) étant donné que ces deux objets se sont rapprochés (0.50 < 0.58).

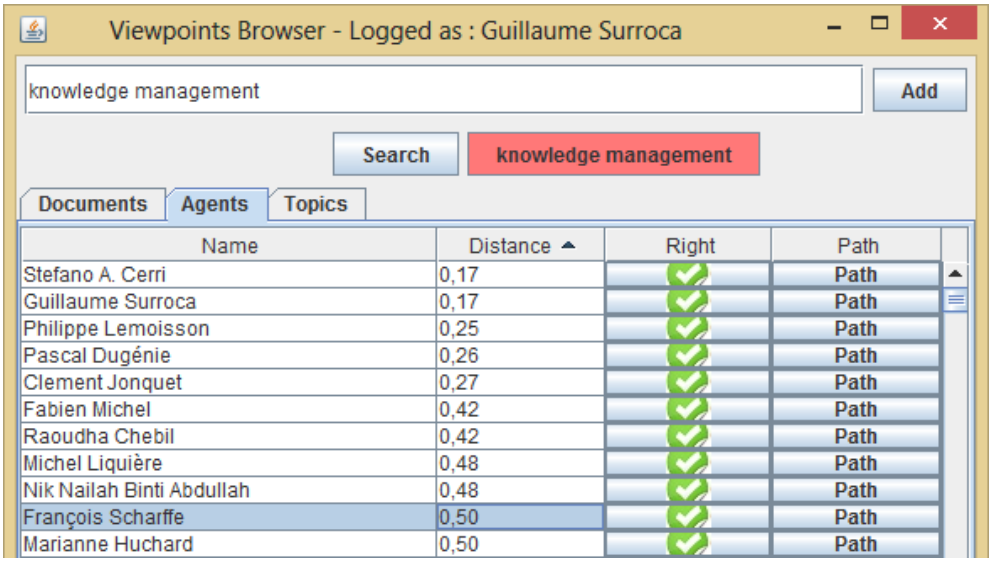


Figure 19 Impact du feedback de CJ sur la recherche.

ViewpointS un potentiel accru pour la recherche d'information :

- Pour un agent donné, le prototype retourne les agents au voisinage permettant d'identifier d'autres contributeurs ou collaborateurs potentiels. Il permet en outre d'identifier les documents proches de cet agent (sans se limiter aux publications dont il est auteur) et ses topics d'intérêts explicites (mot clés de ses publications) ou implicites (mots clés d'autres publications dont il est proche).
- Pour un Topic donné, le prototype permet non seulement d'identifier les documents pertinents (comme n'importe quel moteur de recherche par mot clé) mais permet également d'identifier les experts pour ce Topic et les topics proches dans le graphe de connaissance (illustration de la terminologie spécifique à une base de connaissances).
- Pour un document donné, un utilisateur peut trouver d'autres documents similaires (en plus des topics et agents proches).

4.2.5 Discussions

Transformer des métadonnées en viewpoints suppose un ensemble de choix de modélisation. Ainsi, le choix exposé section 4.2 : « dans un document d, pour chaque auteur a et pour chaque mot-clé t nous créons un viewpoint (a, {d, t}, {Match,+}, 0) est celui d'un modèle simple et immédiat. En outre, étant donné que les topics sont pour le moment des mots-clés librement choisis par les auteurs lorsqu'ils ont enregistré leurs publications sur HAL-LIRMM le graphe de connaissance connaît les problèmes classiques des folksonomies (ambiguïté, polysémie, multilinguisme, etc.). Par exemple, certains topics peuvent être ambigus et créer de faux chemins ; ils peuvent également représenter le même concept. Une stratégie de remédiation serait l'intégration d'un agent artificiel exploitant une ontologie suffisamment riche pour éliminer les ambiguïtés et s'appuyant sur celle-ci pour exprimer sous forme de viewpoints des proximités sémantiques précises. Il est à noter que ce problème lié aux folksonomies n'existe pas avec des sources telles que PubMed (publications biomédicales) respectant une terminologie standardisée (MeSH).En traitant le corpus des publications des chercheurs du

LIRMM, nous avons montré l'opérationnalité de l'approche VIEWPOINTS dans un contexte de recherche d'information, et nous avons apporté des éléments de réponse aux questions énoncées dans l'introduction :

- Le graphe de connaissances réifie en toute transparence la sémantique collective de la communauté. Il contient toutes les explications concernant l'émergence de cette sémantique à partir des contributions. Le processus de réponse aux requêtes est donc lui aussi transparent, permettant à l'utilisateur de trouver des documents ou des données dignes de confiance sur ses sujets d'intérêt.
- La géographie de la connaissance ainsi produite permet aisément de trouver les bonnes personnes pour échanger, argumenter et capitaliser sur un sujet particulier.
- Au fil des interactions, la structure du graphe élicite une proximité sémantique entre certains topics, en reflétant les points de vue des membres de la communauté. Ceci est un premier pas vers des processus d'agrégation susceptibles de faire émerger de nouvelles connaissances.

Ce premier prototype nous a permis d'expérimenter le calcul de la distance sémantique sur des données réelles, mais surtout de valider l'aptitude du formalisme à supporter un modèle traitant la recherche d'information scientifique.

EN RESUME

L'intégration de jeux de données bibliographiques du CIRAD, du LIRMM et l'intégration des deux montrent une première la capacité d'intégration des données de ViewpointS.

Un prototype de moteur de recherche permet de faire des recherches sur des auteurs, des descripteurs ou des documents et d'avoir l'ensemble des ressources sémantiquement les plus proche.

Le mécanisme de feedback permet d'enrichir la base de connaissance et le mécanisme de feedback peut être paramétré.

4.3 Simulation des stratégies de navigation web en regard de l'apprentissage par Sérendipité

Articles référence:

G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, "Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité," in IC 2015, 2015, p. 12.

G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, "Preference Dissemination by Sharing Viewpoints : Simulating Serendipity," KEOD: Knowledge Engineering and Ontology Development, vol. 7th Intert, no. 2. pp. 402–409, 12-Nov-2015.

Nous souhaitons simuler le comportement d'utilisateurs qui parcourent sur un Web fictif. Pour ce faire, l'utilisateur se sert d'un moteur de recherche qui renvoie toutes les ressources du web ou d'une portion du web qui lui paraissent proches de l'objectif de recherche d'information qui a été exprimé. A partir de ces résultats l'utilisateur part ensuite dans une série de cheminement de proche en proche, de lien en lien qui le mène parfois – souvent même – au-delà de l'ensemble de ressource qui lui ont été données pour sa recherche par le moteur de recherche. Ce cheminement au-delà des résultats de recherche permet de « trouver (découvrir, inventer) par hasard, par chance ou par accident, autre chose et, parfois tout autre chose, et, même, parfois, le contraire de ce que l'on cherchait (et de trouver en l'état) ; et de se rendre compte de son intérêt et de son importance. »[91]. Hors, s'il existe parfois un mécanisme de feedback sur les résultats d'un moteur de recherche rien ne permet à notre connaissance de valoriser les cheminements qui se font à partir de ces résultats. En effet, le problème qui se pose alors est la gestion de cette Sérendipité[92].

Dans le domaine de la recherche documentaire, la Sérendipité est encouragée. Cependant, une des caractéristiques clé de la Sérendipité est sa fugacité, il est quasiment impossible de retrouver le chemin qui a conduit à l'information sérendipiteuse. Il faut l'enregistrer immédiatement et l'indexer en clair systématiquement donc proposer également un mécanisme de feedback pour l'information non renvoyée initialement par le moteur de recherche, non-recherché initialement par l'utilisateur mais qui participent à l'enrichissement de sa recherche et de celles de autres. Nous verrons plus précisément que les goûts exprimés par les internautes pendant leur navigation constituent des traces de la Sérendipité.

4.3.1 Objectifs

Nous nous posons les questions suivantes :

- Quelles sont les stratégies de navigation sur le Web qui permettent la diffusion optimale des systèmes de préférences des utilisateurs ?
- Comment positionner les conditions propres à l'apprentissage fortuit, c'est-à-dire à la Sérendipité, dans l'étude des systèmes de préférences ?

Nous parlerons de système de préférences d'un agent pour identifier l'ensemble des goûts ou attirances qu'il exprime sous forme de relations de proximité ou de distance entre ressources du Web. Dans notre première contribution nous avons démontré la capacité d'apprentissage d'une base de connaissances construite à partir d'une première ébauche de notre formalisme. Toutefois, cette preuve de concept ne se basait que sur une modélisation très pauvre du comportement des agents qui naviguaient au hasard au sein de la base de connaissances pour y contribuer ; nous nous intéressions alors seulement à la satisfaction des utilisateurs sans prendre en compte leurs systèmes de préférences. Dans une autre contribution, nous avons montré comment Viewpoints permet la recherche et la découverte de connaissances grâce à un prototype de recherche de publications scientifique. Dans la modélisation du comportement des agents que nous proposons aujourd'hui, nous incluons un paramètre d'« ouverture à la Sérendipité » qui est la propension d'un agent à s'orienter vers des ressources hors de son système de préférences pour guider sa recherche ; cela nous permet d'évaluer la diffusion des systèmes de préférences selon si agent est plutôt ouvert d'esprit ou plutôt

focalisé sur ce qu'il connait et préfère. A partir de cette modélisation, nous construisons une simulation dans laquelle nous créons des règles de comportement individuel (niveau microscopique) et observons l'effet sur l'apprentissage collectif et la diffusion des systèmes de préférences (niveau macroscopique). Cette simulation donne les grandes lignes de l'effet de l'utilisation de Viewpoints pour encapsuler des données du Web sémantique et social.

4.3.2 Graphe de connaissance

Nous souhaitons simuler l'évolution d'une base de connaissances telle que le Web à partir de règles de comportement individuelles qui décrivent les navigations d'agents sur le Web et la diffusion de leurs systèmes de préférences respectifs. Nous commençons par expliquer comment nous représentons les systèmes de préférences dans un graphe de connaissances Viewpoints, puis nous proposons un modèle du comportement simulant différentes stratégies de navigation paramétrables. Il est possible dans viewpoints de créer de nouveaux types de viewpoints, nous allons créer dans ce scénario deux types de viewpoints – de type Like et Knows (un nouveau type créé pour les besoins de l'expérimentation) – pour représenter les relations des Princes aux.

4.3.2.1 Représentation des systèmes de préférences

Chaque ressource de KG est caractérisée par une forme, une taille et une couleur qui serviront à les rapprocher. Les informations de forme et de taille seront déjà présentes dans le graphe de départ de la simulation ; ces informations simulent les données du Web sémantique. Les informations de couleur des ressources seront ajoutées au fur et à mesure de la simulation par 3 agents (rouge, vert, bleu), les princes de Serendip, qui connaissent et aiment respectivement une couleur distincte ; ces informations simulent les contributions du Web social. Le système de préférences d'un prince est traduit par les viewpoints qu'il émet pour se rapprocher des ressources de sa couleur ou rapprocher entre elles des ressources de même couleur (la sienne). La diffusion d'un système de préférences est donc équivalente à la diffusion de l'information de couleur dans le graphe. Ainsi, l'apprentissage de la couleur par le graphe représente l'émergence d'une intelligence collective de la communauté. Nous considérons ici deux types de viewpoints : (i) rapprochant deux ressources de même couleur (Knows) (ii) rapprochant un prince d'une couleur à une ressource de la même couleur (Likes). Par exemple, si le prince rouge fait une recherche sur une ressource r qui est rouge et obtient parmi les résultats une ressource r' qui est aussi rouge alors il créera les deux types de viewpoints : (prince rouge, {prince rouge, r}, {Likes,+}, τ) et (prince rouge, {r, r'}, {Knows,+}, τ). Dans la section suivante nous présenterons les stratégies de navigation dans KG qui permettent à un prince de diffuser la connaissance de sa couleur.

4.3.2.2 Modèle comportemental des Princes

L'automate à état (Figure 20) décrit le comportement des princes quand ils naviguent dans KG, et diffusent au fur et à mesure de leurs feedbacks (émissions de viewpoints) leurs systèmes de préférences. Plus généralement, cet automate nous permet de décrire le comportement d'un utilisateur explorant le contenu d'une base de connaissances telle que le Web. Nous capturons ainsi des comportements tels que : la requête sur moteur de recherche, l'exploration des résultats, l'exploration des liens inclus dans ces résultats et le retour éventuel au moteur de recherche avec une autre requête, etc. Les probabilités qui conditionnent les transitions dans cet automate dépendent de trois paramètres :

- β, qui est la probabilité de revenir en arrière pendant la navigation, c.-à-d. soit de revenir à la recherche d'origine (état de départ) ou à la dernière recherche effectuée (état précèdent).
- μ, qui est le choix parmi les outils de navigation disponibles : soit l'utilisation du moteur de recherche opérant globalement sur le graphe soit l'exploration locale des résultats de proche en proche en suivant les liens qui les connectent.
- σ, qui est la capacité à diriger sa navigation vers des ressources qui n'appartiennent pas forcément à son propre système de préférences : l'ouverture à la Sérendipité.

Dans notre simulation le comportement d'un prince de Serendip correspond à un paramétrage spécifique de β, μ et σ ; nous parlerons de *stratégie de navigation*. Ces stratégies simulent des stratégies de navigation sur le Web (ou autre base de connaissances). La simulation se divise en cycles qui correspondent à des explorations successives de KG. Au début d'un cycle, un prince commence par une interaction avec KG qui simule l'utilisation d'un moteur de recherche : une ressource de KG est sélectionnée aléatoirement et nous utilisons la fonction de voisinage indirect pour obtenir une liste de résultats (autres ressources) triés. A partir des résultats proposés, le prince poursuit (β faible) ou abandonne cette recherche et en fait une nouvelle (β fort). S'il poursuit, il va évaluer (relativement à la couleur correspondant à son système de préférences) ces résultats un par un et opter pour le premier non-visité en fonction du paramètre σ. Si le prince est ouvert à la Sérendipité (σ fort), alors il ne se dirigera pas systématiquement vers une ressource de même couleur que lui, sinon (σ faible[46]) il privilégiera sa couleur. Ayant choisi une ressource, le prince passera à la prochaine étape de son cheminement, en fonction de μ, soit en faisant une recherche sur cette ressource (μ fort) soit en explorant localement autour de cette ressource (μ faible). La première interaction simule le fait d'ouvrir une page Web après avoir cliqué sur une des URL proposées par le moteur de recherche ; l'interaction suivante simule soit une nouvelle recherche sur par exemple le titre de la page, soit le clic sur un lien inclus dans celle-ci.

Dans la simulation, un prince dispose d'un budget d'interactions qui diminue à chaque interaction (recherche ou exploration). Ce budget représente la quantité d'effort qu'il est prêt à faire dans sa navigation. Si au moment du retour en arrière il n'y a plus d'étapes précédentes, s'il n'y a plus de ressources non-visitées ou si son budget d'interaction a été dépensé alors le cycle s'achève.

[46] S'il avait choisi le moteur de recherche Qwant.com il aurait donc commencé par le premier résultat qui lui semble correspondre le plus à ses goûts (σ faible).

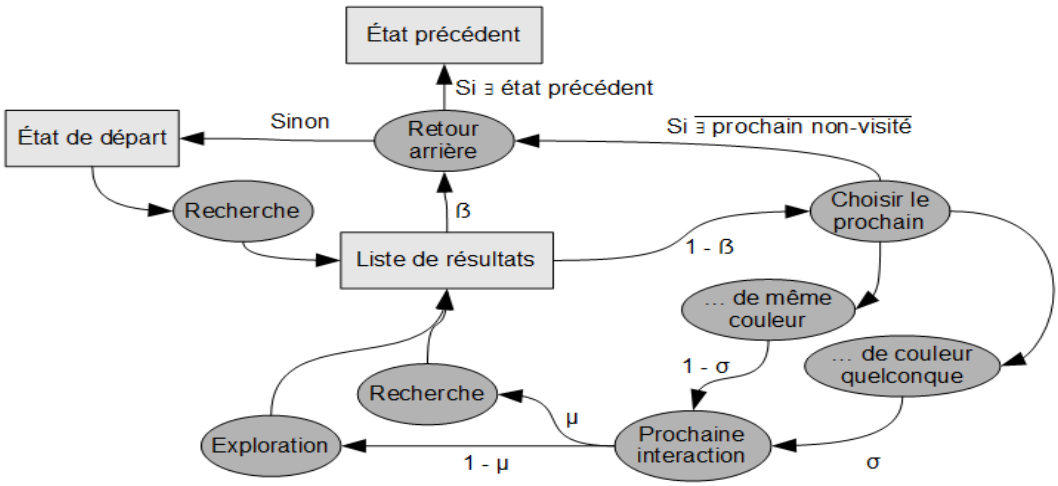


Figure 20 Automate de comportement des princes : stratégies de navigation.

Nous représentons dans la Figure 21 les trois paramètres relatifs aux stratégies de navigation dans un espace à trois dimensions. Ces stratégies mises en place dans la simulation des princes de Serendip simulent des stratégies de navigation sur le Web. En terme de parcours de graphe plus β est élevé plus on se rapproche d'un parcours en LARGEUR et plus β est faible plus il s'agit d'un parcours en PROFONDEUR.

Dans une démarche de recherche d'information le parcours en LARGEUR reviendrait à évaluer de manière superficielle l'ensemble des résultats pour avoir une idée d'ensemble de tous les résultats et l'approche par PROFONDEUR reviendrait plutôt à se concentrer sur ce qui paraîtrait être la meilleure réponse et la creuser plus en profondeur. μ conditionne le style de navigation. Quand μ est élevé on utilise majoritairement des moteurs de RECHERCHE renvoyant des résultats triés et indirectement liés tandis que quand μ est faible on explore de proche en proche en récupérant des résultats non-triés et directement liés (EXPLORATION). Par exemple, la navigation entre vidéos suggérées sur YouTube est un bon cas de figure d'une exploration de proche en proche tandis que l'utilisation répétitive de Google dans une recherche est plutôt un exemple de parcours en LARGEUR. Nous représentons l'ouverture à la Sérendipité (σ) comme une troisième dimension. Quand σ est grand c'est que l'utilisateur est dans une démarche d'OUVERTURE et qu'il est disposé à cheminer aussi bien parmi des ressources qui correspondent à ses préférences que d'autres ressources qui n'y correspondent pas mais qui pourrait l'amener à la découverte fortuite. Dans le cas opposé (FERMETURE), l'utilisateur parcourt le Web entièrement guidé par ses préférences.

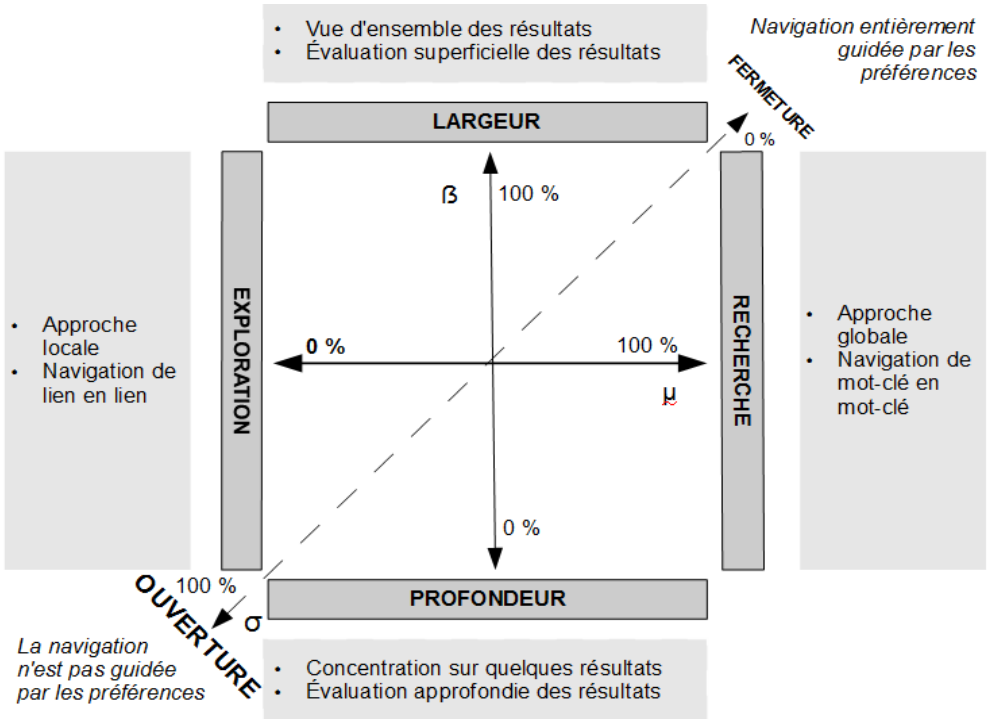


Figure 21 Différentes stratégies de navigation en fonction des paramètres β, μ et σ.

4.3.3 Déroulement de l'expérimentation

A l'initialisation, un KG de taille déterminée est généré. Les ressources de ce KG sont des ressources caractérisées par une taille (petit, moyen, grand), une forme (carré, cercle, triangle) en plus de posséder une couleur (rouge, vert, bleu). Pour chaque combinaison possible de tailles, formes et couleurs N ressources sont créées. Il y a donc initialement 27N ressources dans KG. Deux agents artificiels que nous appellerons péons sont ajoutés à KG. L'un d'entre eux partage son appréciation de la notion de forme au graphe de connaissances en reliant tous les couples de ressources de même forme par des viewpoints de type `svp:initial`. L'autre péon fera de même pour la caractéristique de taille. Ainsi, après le passage des péons, KG ne connaît pas la couleur car les ressources ne sont liées que par les deux caractéristiques de taille et de forme. Pour finir la phase d'initialisation trois autres agents sont ajoutés à KG : les princes. Chacun est caractérisé par une couleur unique et a la capacité d'apprécier cette couleur et de partager cette appréciation en émettant de nouveaux viewpoints de type `svp:like` et `svp:knows` dans le graphe de connaissances. Il y a donc une connaissance implicite que les princes sont seuls aptes à partager, par émission de viewpoints de feedback. Les paramètres de la simulation sont résumés dans le

Tableau 8. Le prince suit le modèle comportemental que nous avons précédemment défini et diffuse ses préférences (la connaissance de sa couleur) en émettant des viewpoints de type svp:like et svp:knows. Le poids associé à chaque type de viewpoint est indiqué. La fonction d'agrégation des viewpoints pour le calcul de la valeur des synapses est la somme. A la fin de chaque cycle les mesures suivantes sont effectuées. Elles nous permettent d'évaluer la diffusion de la connaissance des couleurs dans KG :

- `M1 Couleur X` : Il s'agit du ratio : distance[47] moyenne entre ressources quelconques / distance moyenne entre ressources de couleur X.

[47] La mesure de distance employée est une distance aux propriétés métriques (symétrie, séparation et inégalité triangulaire) basée sur le calcul du plus court chemin de Dijsktra (cf. [19]).

- `M2 Couleur X` : Il s'agit de la probabilité d'obtenir au voisinage d'une ressource de couleur spécifique des ressources de la même couleur.

Tableau 8 Résumé des paramètres de la simulation et de leurs valeurs fixes.

Catégorie	Paramètre	Valeur (si fixée)	
Paramètres d'échelle	Facteur d'échelle (N)	3	
	Nombre de cycles	100	
	Nombre d'interactions par cycle	50	
Paramètres de perspective	Poids associé aux viewpoints de type `vps:initial`	1	
	… de type `Knows`	2	
	… de type `Likes`	1	
Paramètres de stratégie de navigation	ß		
	µ		
	σ		
Répartition de l'activité	Prince rouge	33%	80%
	Prince vert	33%	10%
	Prince bleu	33%	10%
Paramètres d'algorithme	Borne de distance pour le calcul de voisinage sémantique (m)	2	

Étant donné le nombre important de paramètres, nous ne présenterons des résultats obtenus (courbes) que pour certaines simulations, que nous avons jugées les plus significatives pour l'étude des stratégies de navigation. Cependant, nous expliquerons les effets de tel ou tel paramètres dans la section discussion. La Figure 22 illustre une évolution du graphe de connaissances. Elle exemplifie six viewpoints émis à l'initialisation (en haut). La flèche de la perspective montre ce que donne ce graphe de connaissance une fois interprété sous forme de KM.

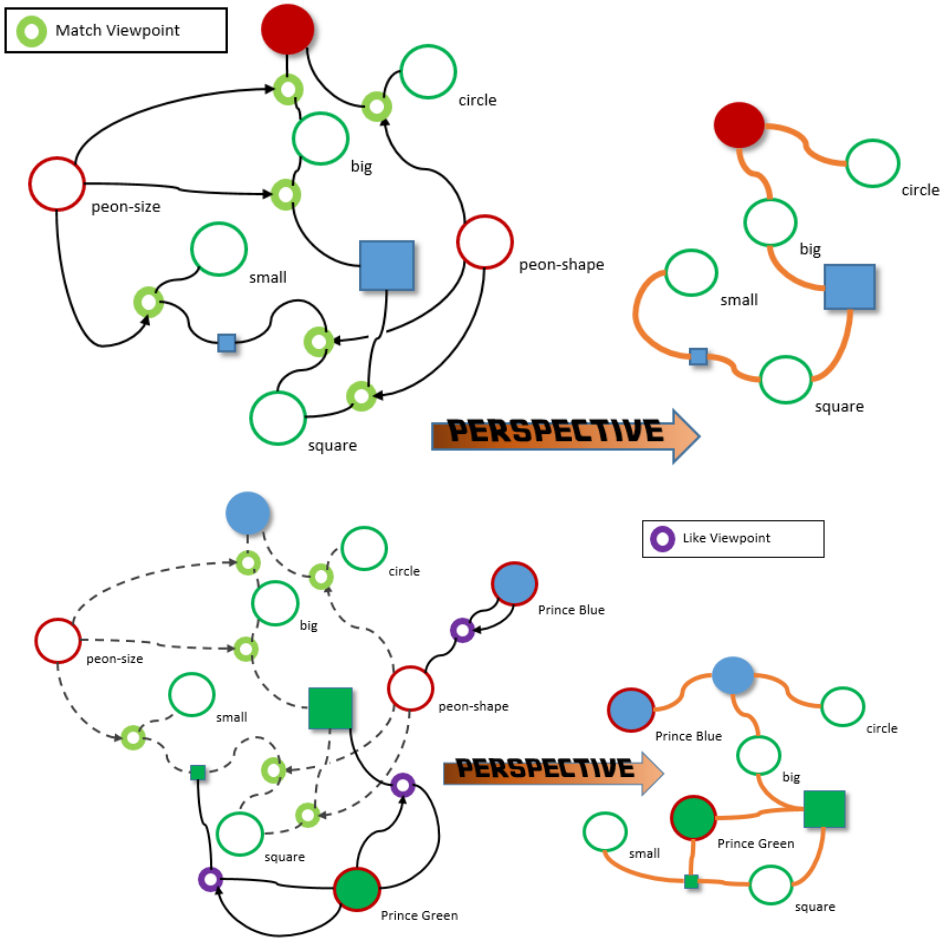


Figure 22 Illustration de l'évolution du graphe de connaissances dans la simulation des princes de Serendip.

4.3.4 Hypothèses

Au fur et à mesure que les princes contribuent à KG ils partagent leurs appréciations des couleurs avec les autres utilisateurs grâce au mécanisme de feedback. Nous souhaitons observer comment, après leurs contributions, KG aura « appris » au niveau global la notion de couleur qui n'était pas dans les connaissances originalement représentées par les viewpoints.

Chaque système de préférences individuel d'un prince devient ainsi, grâce aux viewpoints, une part de la connaissance collective représentée dans KG où il cohabite avec les systèmes de préférences des autres princes. Nous souhaitons expérimenter différentes stratégies de navigation et démontrer que les systèmes de préférences diffusés de façon concurrente ne se neutralisent pas. Nous souhaitons également mesurer l'effet de la Serendipité. Ainsi, nous nous attendons à ce que la mesure M1 augmente, c'est-à-dire à ce que la distance moyenne entre ressources de même couleur décroisse plus vite que la distance moyenne entre ressource quelconques. En effet, les princes rapprochent les ressources de même couleur d'eux-mêmes et les unes des autres, sans jamais rapprocher deux ressources de couleurs différentes. Pour les mêmes raisons, la mesure M2 devrait augmenter aussi car elle reflète la probabilité de trouver une ressource de même couleur dans le m-voisinage d'une ressource.

4.3.5 Résultats

Dans un premier temps, nous faisons varier les stratégies de navigation en conservant la symétrie dans le comportement des trois princes et dans leur répartition de l'activité. Nous observons comment KG « apprend » la couleur rouge en mettant l'accent sur le paramètre σ (ouverture à la Sérendipité). Dans un second temps, nous nous restreignons à deux stratégies de navigation contrastées et jouons sur des répartitions d'activité différentes pour les trois princes ; nous comparons alors les apprentissages respectifs des trois couleurs par KG.

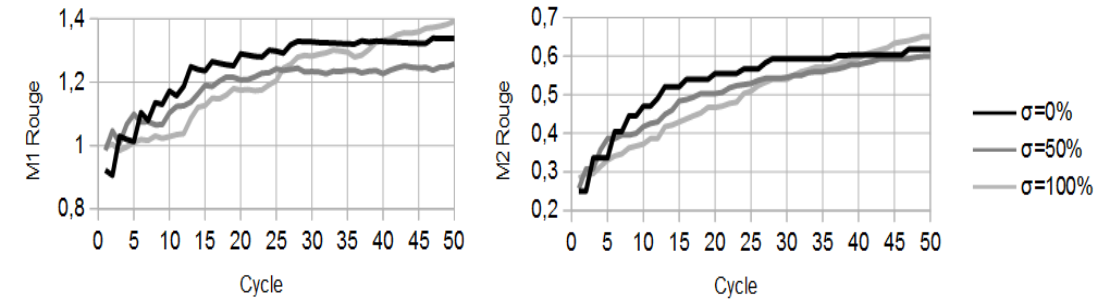
4.3.5.1 Impact de l'ouverture à la Sérendipité

Nous commençons par évaluer l'impact de σ sur la diffusion de la couleur rouge grâce aux mesures `M1 Rouge` et `M2 Rouge`. Nous remarquons d'après la Figure 23 que dans le cas d'une utilisation majoritaire du moteur de recherche M1 et M2 croissent plus vite quand l'ouverture est faible mais qu'inversement quand l'ouverture est élevée elles atteignent des valeurs finales plus élevées. La recherche renvoie des résultats indirectement liés et permet de créer des viewpoints originaux. L'ouverture à la Sérendipité permet au final une diffusion plus grande de la connaissance des couleurs grâce à l'exploration de plus de ressources qui n'auraient pas été rencontrées avec des stratégies fermées. Par exemple, le prince rouge peut rencontrer une autre ressource rouge cachée derrière une ressource verte s'il ose explorer la ressource verte.

Par contraste, nous observons que dans une approche d'exploration locale où seuls sont renvoyés des résultats directement liés l'ouverture à la Sérendipité n'apporte rien ni en terme de croissance des valeurs M1 et M2, ni en terme de valeur finale obtenue. L'idée, avec une telle stratégie, est d'explorer localement et en profondeur les résultats, ainsi le fait de passer par des résultats moins intéressants en chemin a plutôt tendance à freiner la diffusion des systèmes de préférences.

L'effet de μ (outil de navigation) est donc très important sur la Serendipité. Nous nous rendons toutefois compte de la relative homogénéité de notre graphe par rapport à la structure du Web. Nous pensons que la Sérendipité peut apporter en condition réelle un saut qualitatif plus substantiel que celui que nous mesurons sur ce graphe 'jouet'. Dans cette simulation les trois princes sont également actifs (33%) et ß=10%[48]

[48] Nous avons pu étudier au fur et à mesure de nos simulations que la variation du paramètre ß ne change pas les résultats que nous présentons ci-après. Ainsi, nous le fixons dans toutes les simulations présentées à 10% donnant ainsi priorité aux stratégies en profondeur.



– Pour µ=70% et ß=10% (Recherche Profondeur plus ou moins Ouverte).

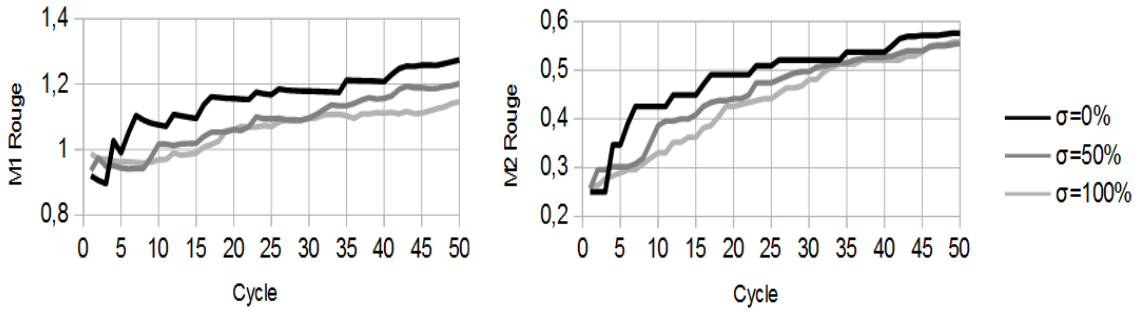


Figure 23 Pour µ=30% et ß=10% (Exploration Profondeur plus ou moins Ouverte).

4.3.5.1 Impact de la répartition de l'activité entre les princes

Dans cette partie nous analysons l'impact de la répartition de l'activité des princes sur la diffusion des couleurs. Pour cela nous observons comparativement `M1 Rouge`, `M1 Vert` et `M1 Bleu` qui évaluent chacune la diffusion d'une couleur dans le graphe. Dans cette simulation, nous faisons varier les probabilités associées aux degrés d'activité respectifs des princes et considérons successivement deux configurations contrastées pour les stratégies de navigation : Recherche Largeur Fermée (µ=80%, ß=40%, σ =10%) et Exploration Profondeur Ouverte (µ=20%, ß=10%, σ =70%). Nous comparons les résultats (Figure 24 et Figure 25) obtenus pour ces stratégies avec une réparation homogène de l'activité des princes et avec une répartition non-homogène.

Dans les deux cas, nous remarquons que la diffusion de chaque couleur se fait même si la concurrence ralentit cette diffusion. Lorsque les princes sont en concurrence, l'apprentissage d'une couleur se fait bien au détriment d'une autre (lorsque M1 augmente pour un cycle donnée, les autres diminuent) et le prince le plus actif diffuse plus efficacement sa couleur.

Cependant, la somme des `M1 Rouge`, `M1 Vert` et `M1 Bleu` finales a une valeur plus élevée quand toutes les connaissances sur les couleurs peuvent être diffusées (Figure 24, la somme des valeurs finales vaut respectivement 3.9 et 3.6) que quand une couleur domine dans la diffusion des couleurs (Figure 25, la somme des valeurs finales vaut respectivement 3.6 et 3.5). D'après ces résultats, on peut déduire que les contributions des utilisateurs du Web social ne neutralisent pas celles des autres mais peut les occulter en passant au premier plan.

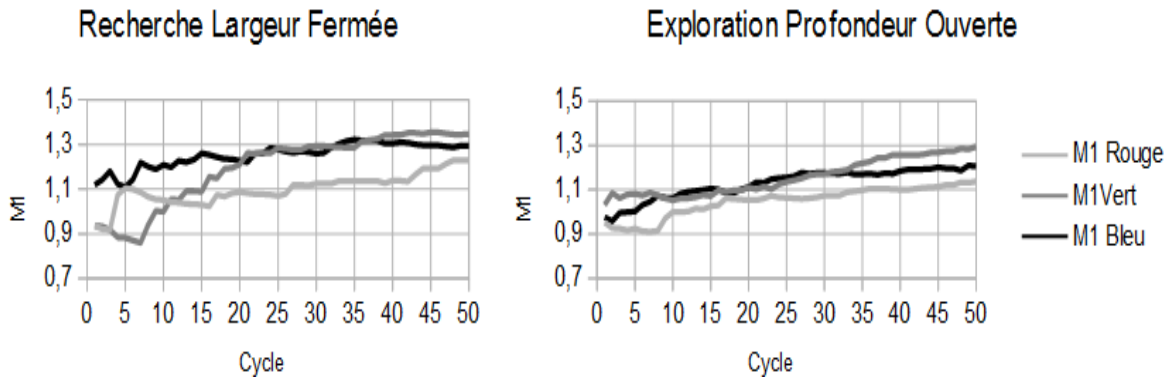


Figure 24 Tous les princes contribuent autant (33%)

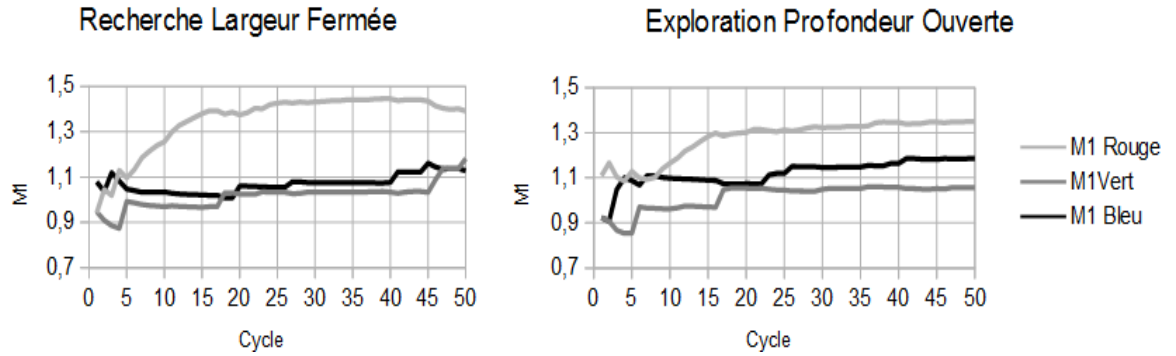


Figure 25 Le prince rouge est plus actif (80%) que les autres (10%).

4.3.6 Discussions

Nous nous attendons à ce que le rapport distance moyenne entre ressources de même couleur / distance moyenne entre ressource quelconques baisse. En effet, comme les princes rapprochent tous les objets de même couleur d'eux-mêmes, alors ils rapprochent indirectement chaque objet de même couleur. Donc normalement la distance moyenne entre ressources de même couleur décroit plus vite que la distance moyenne entre objet quelconques. Par la même causalité, la proportion moyenne de ressources de même couleur que r au voisinage de r devrait augmenter car si la distance moyenne entre ressources de même couleur diminue plus vite que la distance moyenne entre ressources quelconques donc toutes les ressources de même couleur se rapprochent alors il y aura incidemment plus d'objet de couleur c au voisinage d'un objet de couleur c. Pour finir, comme le mécanisme de voyage de proche en proche est basé sur une mécanique d'aléatoire pondéré sur la force des synapses alors on peut penser que le cheminement qui part de r a de forte chances de rester dans le voisinage de r donc ce cheminement a de forte chance de comporter une couleur de plus en plus dominante qui est celle de r.

A travers la simulation des princes de Serendip nous présentons un essai de modélisation de la Sérendipité sur le Web. Nous sommes toutefois conscients que ce modèle du comportement des utilisateurs du Web ne rend pas complètement compte de la réalité et de la diversité des méthodes d'explorations du Web. Malgré cela, nous espérons avoir démontré la capacité d'apprentissage du graphe de connaissances de Viewpoints. Les résultats de la simulation nous permettent d'évaluer l'apport de l'ouverture à la Sérendipité dans diverses stratégies de navigation ainsi que son impact sur la diffusion des systèmes de préférences ; nous avons donc consolidé la preuve de concept de Viewpoints en le confrontant à une modélisation d'usage plus réaliste que lors de nos dernières simulations.

EN RESUME

Cette expérimentation est une simulation de stratégies de navigation Web en regard de l'apprentissage par Sérendipité. Au cours de leurs parcours du Web les internautes diffusent leurs systèmes de préférences. La simulation d'un écosystème d'agents consommateurs et producteurs de viewpoints dans une simulation sur la diffusion de préférences démontre quelle est la stratégie de navigation du Web la plus appropriée pour découvrir le plus du Web à court et long terme.

L'ouverture à des ressources n'appartenant pas à son propre système de préférences. Nous démontrons qu'au court terme une navigation uniquement guidé par les préférences d'un utilisateur procure beaucoup de nouveaux résultats, toutefois cela occulte certaines ressources potentielles et au long terme la quantité maximale de découvertes positives – c'est-à-dire de son propre système de préférence – est atteinte par l'internaute simulé qui s'ouvre le plus à d'autres systèmes de préférences.

4.4 Recommandation de films

4.4.1 Objectifs

Après avoir expérimenté ViewpointS sur deux graphes de connaissances-jouet nous entreprenons d'appliquer ViewpointS à un nouveau cas d'étude concret. Nous avons choisi une base de données de films afin de créer un graphe de connaissances sur les films. Ce graphe de connaissances nous permet de proposer un service de recommandation de films. Nous allons, dans cette section, évaluer le potentiel des méthodes de calcul de voisinage sémantique.

4.4.2 Graphe de connaissance

Nous choisissons un jeu de donnéess cinématographiques contenant à la fois des métadonnées « objectives » sur les films (genres, année de sortie) ainsi que sur un ensemble d'utilisateurs (sexe, âge, métier) et des données « subjectives » qui sont des notes/ratings créées par ces utilisateurs. La connaissance sur les films est d'une part une connaissance explicite sur les films grâce aux métadonnées de films mais aussi d'une connaissance implicite liées aux goûts et appréciations qui attend d'être élicitée. Ce sont des données rendues anonymes venant d'un système de recommandation de films : MovieLens[49].

Ce jeu de donnéess inclus plusieurs types de relations et en jouant sur le poids que nous donnons à chaque type nous pouvons voir quel est le bénéfice apporté par la subjectivité des points de vue (notes) sur la sémantique collective qui émerge de ce jeu de donnéess. La Figure 26 montre la structure globale de notre jeu de donnéess interprété sous forme de ressources et de viewpoints. Les ressources « principales » de ce jeu de donnéess sont les Films (considérés comme des Numeric Resources) et les ressources de description (considérés comme Descriptors) comme les genres, année de sortie et note pour les films et les métiers, âges et sexes. D'une part des viewpoints sont créés entre les films et leurs descripteurs (année de sortie, genre, note). D'autre part les utilisateurs qui ont donné les notes sont liés à des d'autres descripteurs (sexe, âge, métier). Ces deux composantes de la connaissances des films sont ce qu'on appelle la connaissance explicite. Mais les utilisateurs créent aussi des viewpoints qui les rapprochent des films et cela rapproche ces deux groupe d'information l'un axé sur l'utilisateur l'autre autour du film.

Ainsi, grâce aux notes des utilisateurs une connexion se fait entre deux jeux de données : les métadonnées de films et celles sur les utilisateurs. Nous allons voir en quoi ces données d'utilisateurs peuvent grâce à la transitivité rapprocher différemment les films que si nous ne considérions que les genres. Nous pouvons voir aussi à quels points les goûts exprimés sur les films rapprochent les utilisateurs entre eux autrement que par leurs données d'âge, de sexe ou de profession.

Voici quelques chiffres sur le jeu de données MovieLens :

- 18 genres
- 1682 films
- 943 utilisateurs
- 100k notes données

[49] http://grouplens.org/datasets/movielens

- 68 ans de films représentés

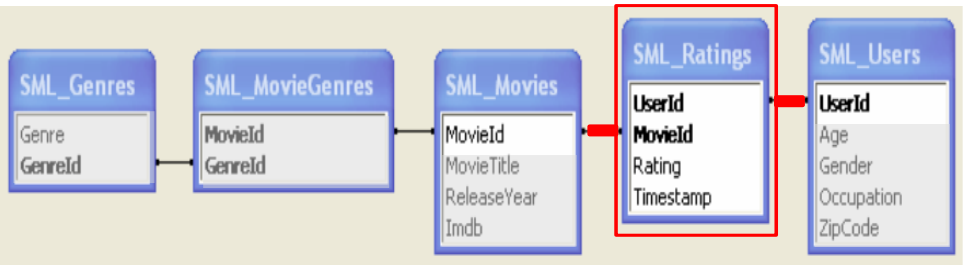


Figure 26 Schéma relationnel du jeu de données MovieLens[50]

Une fois indexées ce dataset nous donne un graphe de connaissances de 207471 viewpoints reliant 2752 ressources (films, genres, années de sortie, utilisateurs, sexe, catégories d'âge, jobs). La Figure 27 illustre le graphe de connaissance résultant de l'indexation des données MovieLens. Ces viewpoints sont soit des viewpoints :

- Film-genre : 2893
- Film-année : 1691
- Film-rating : 100k
- Utilisateur-Film : 100k
- Utilisateur-Catégorie d'âge : 943
- Utilisateur-Job : 943
- Utilisateur-Sexe : 943

Nous avons commencé par construire un outil pour explorer ce jeu de donnéess et observer ses spécificités. Il s'agit d'une légère amélioration par rapport au prototype que nous avons conçu pour la recherche bibliographique. Il permet de chercher au voisinage de n'importe quelle ressource et donner un retour positif ou négatif aux résultats comme le montre la Figure 28. Il offre aussi la capacité de voir la KM locale autour d'une ressource spécifique (Figure 29). La recherche dans cet outil emprunte la même démarche que le précédent prototype de la section 4.2. L'utilisateur peut cette fois choisir entre SPN et MPN pour le calcul de voisinage sémantique.

[50] https://grouplens.org/datasets/movielens/

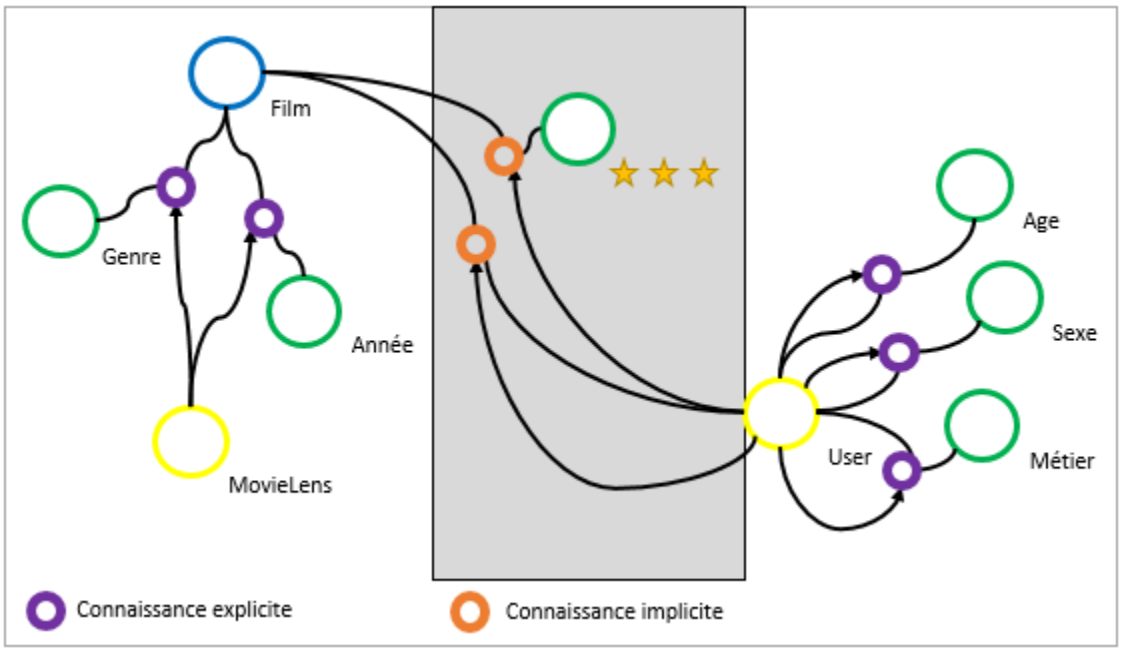


Figure 27 Illustration du graphe de connaissance MovieLens

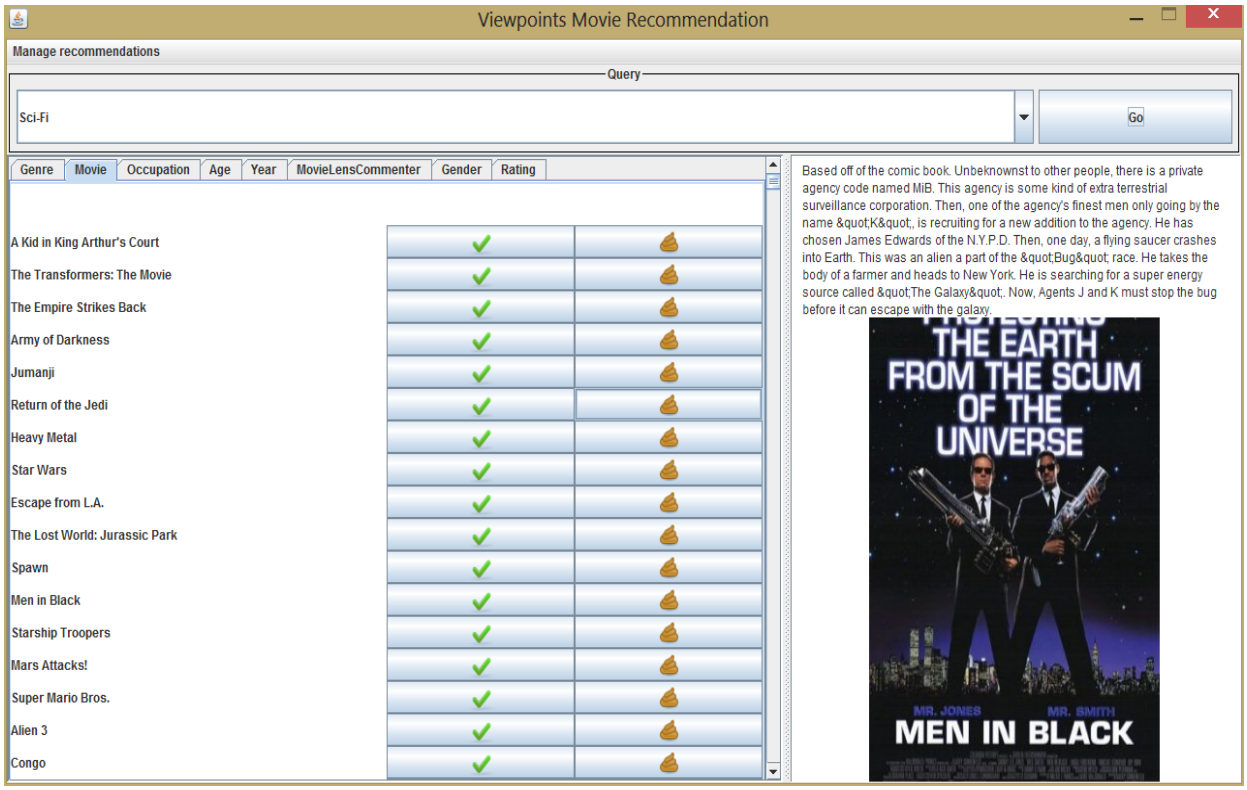


Figure 28 Prototype Viewpoints Movie Recommender

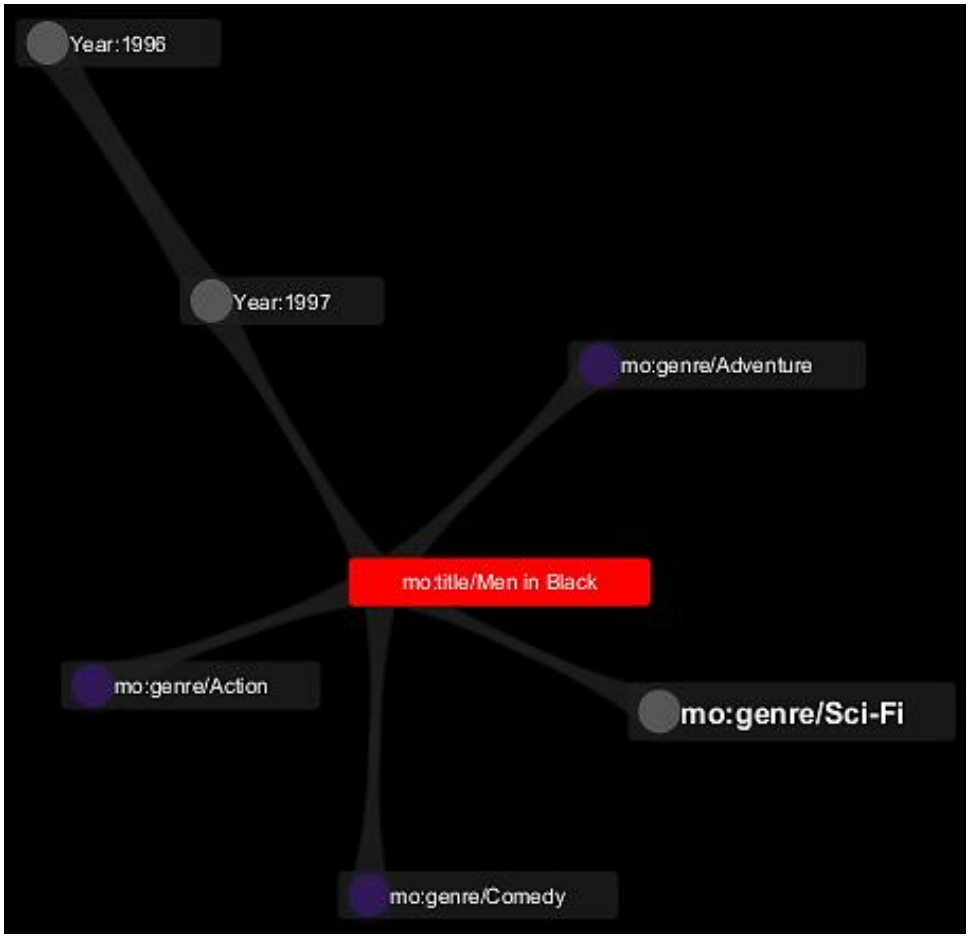


Figure 29 KM locale autour du film Men In Black

4.4.3 Déroulement de l'expérimentation

Grâce à l'API Viewpoints nous avons construit un prototype de système de recommandation de films se basant sur les données de MovieLens et utilisant SPN et MPN pour explorer le graphe de connaissance obtenu. Nous jugeons la qualité de chacune des méthodes d'exploration du graphe dans une situation de recommandation de films en la comparant à un système de recommandation référence : TheMovieDatabase[51] (TMDB). Comme nous ne connaissons pas les méthodes, ni les données employées par TMDB nous ne pouvons pas réellement nous comparer directement à TMDB. En effet, les données ne sont pas exactement les mêmes (films, notes, etc.) et nous ne connaissons pas les critères de TMDB pour recommander des films.

Toutefois, TMDB nous permet de comparer nos méthodes entre elles en supposant que qu'il représente une référence (gold standard). Ainsi, nous pouvons mesurer le rappel et la précision en comparaison à TMDB. Dans la suite, nous expérimentons deux scénarios de tests, décrits ci-après, pour lesquels la perspective utilisée ne prend en compte que les relations entre films et genres. Nous expérimenterons aussi d'autres types de perspectives prenant en compte une plus grande variété de types de viewpoints. Etant donné que nous nous comparons à TMDB qui contient certains films non compris dans notre jeu de donnéess (MovieLens) nous nous baserons sur les films qui appartiennent aux deux datasets.

[51] https://www.themoviedb.org/documentation/api

4.4.3.1 Scénario de test 1 : Recommandation de films par genre

Nous commençons par demander à TMDB des recommandations de films pour un genre donné. Il s'agit de notre gold standard. Ensuite nous comparons les voisinages que renvoie chacune de nos méthodes avec le gold standard mais sur notre jeux de données MovieLens. Nous répétons l'opération pour l'ensemble des 19 genres.

4.4.3.2 Scénario de test 2 : Recommandation de films similaires

Nous demandons à TMDB de suggérer pour un film donné un ensemble de films similaires. Il s'agit de notre gold standard. . Ensuite nous comparons les voisinages que renvoie chacune de nos méthodes avec le gold standard. Nous répétons l'opération sur un ensemble de films aléatoirement choisis. Nous prendrons un échantillon de 50 films pour tous nos tests.

4.4.4 Résultats

4.4.4.1 Scénario de test 1 : Recommandation de films par genre

Le Tableau 9 résume les résultats obtenus pour les trois méthodes SPN, MPN, WRWN. Pour WRWN nous nous fixons un nombre de chemins aléatoires à explorer à 50 (attribut n du pseudocode).

Tableau 9 Mesures pour 50 réponses renvoyées par SPN, MPN et WRWN et un gold standard de 100 films.

Méthode	Rappel	Précision	F-mesure
MPN	0,2	0,69	0,31
SPN	0,18	0,57	0,26
WRWN	0,2	0,64	0,22

La meilleure méthode pour le cas genre-film est MPN car celle-ci offre les meilleures mesures (rappel, précision et f-mesure). En effet, c'est la meilleure pour évaluer la distance entre deux films qui partagent plusieurs genres en commun. Juste derrière se trouve WRWN qui est une version échantillonnée (approximée) de MPN. Remarquons aussi que SPN qui se base sur le plus court chemin est une méthode qui, dans nos autres résultats, est très sensible a une répartition très inégale des poids sur les types de viewpoints. Si un type de viewpoint venait à avoir un poids beaucoup plus élevé le plus court chemin change radicalement ainsi que le résultat avec l'ajout de viewpoints de ce type. L'approche basée sur de multiples chemins (MPN) est beaucoup moins influencée par l'émission de viewpoints de poids fort.

Nous comparons ensuite (Tableau 10) la méthode qui semble renvoyer les meilleurs résultats (MPN) avec une méthode classique de la recherche d'information : PageRank[52]. PageRank construit en parcourant un graphe aléatoirement une distribution statistique qui donne pour chaque nœud la probabilité de l'atteindre en venant de n'importe nœud d'origine. Nous personnalisons PageRank en calculant pour chaque nœud la probabilité d'être parcouru en venant d'un nœud de d'origine. Le voisinage topologique au sens de PageRank est l'ensemble des nœuds ayant une probabilité d'être atteint en partant du nœud donné supérieure ou égale à p. p permet de doser le diamètre de voisinage sou-

[52] https://fr.wikipedia.org/wiki/PageRank

haité par le voisinage PageRank. Nous prendrons en compte les 50 premiers résultats du voisinage PageRank. PageRank ne faisant pas de distinction entre hyperliens nous dirons que toute synapse à un poids de 1.

Tableau 10 Comparatif MPN vs. PageRank. Nous utilisons les mêmes paramètres de mesures.

Méthode	Rappel	Précision	F-mesure
MPN	0,2	0,39	0,31
PageRank	0,12	0,46	0,19

PageRank considère au même niveau tous les hyperliens reliant les documents. Le comportement du voisinage PageRank est similaire à celui de WRWN en cela qu'il se base sur une succession de cheminements aléatoires. Ces deux méthodes se basent toutes deux sur de multiples chemins entre deux ressources. Toutefois WRWN se base sur un choix aléatoire pondéré par la force des synapses.

Nous allons ensuite expérimenter l'impact de deux autres perspectives. Dans la première (P1) nous considèrerons en plus des viewpoints film-genre, les viewpoints film-année et film-note. Ensuite nous considèrerons l'ensemble des types de viewpoints (dont film-utilisateur, utilisateur-sexe, utilisateur-métier, utilisateur-age) dans la perspective P2.

Poids des types dans P2 :

- genre-film : 6
- année-film : 3
- note film : 0,1
- Poids des types dans P3 :
- genre-film : 6
- année-film : 3
- note-film : 0,1
- utilisateur-sexe : 1
- utilisateur-age : 3
- utilisateur-métier : 3

Tableau 11 Résultats obtenus pour une perspective prenant en compte la totalité des métadonnées sur les films (P1).

Méthode	Rappel	Précision	F-mesure
MPN	0,07	0,2	0,1
SPN	0,1	0,31	0,15
WRWN	0,05	0,18	0,08

Tableau 12 Résultats obtenus pour une perspective prenant en compte tous les types de relations offerts par MovieLens (P2).

Méthode	Rappel	Précision	F-mesure
MPN	0,02	0,01	0.01
SPN	0,08	0,16	0.1
WRWN	0,02	0,01	0.01

On remarque que, dans le scénario de recommandation de film par genre, le fait de prendre en compte d'autres données d'autres natures comme celles des utilisateurs mènent par des chemins trop indirects à des résultats inexacts. Le fait que les films de même année soient reliés indirectement pollue en quelque sorte le rapprochement des films de même genre. Il en est de même pour les relations entre films et notes ainsi que pour les nombreux chemins beaucoup plus indirects passant par les utilisateurs et leurs descripteurs. Toutefois comme nous avons déterminé tout de même un poids dominant pour les viewpoints film-genre SPN a moins été affecté que MPN. Ces résultats de la transitivité de nos méthodes sont peut être intéressants dans une approche exploratoire du jeu de données car ils mènent potentiellement à la découverte mais dans un cas d'utilisation orienté recommandation nos approches pouvant se ramener à des méthodes plus classique de l'IR sont meilleures.

4.4.4.2 Scénario de test 2 : Recommandation de films similaires

Le Tableau 13 présente les résultats obtenus dans le scénario Film → Films pour deux des trois mesures que nous proposons. WRWN est une version approximée de MPN et qui dans le meilleur des cas renvoient le même résultat que MPN. Nous ne comparerons dès lors plus WRW aux autres méthodes. Nous utiliserons dans les tests de ce scénario 10 films choisis aléatoirement. Pour perspective nous utiliserons celle qui, dans l'expérimentation précédente nous a offert les meilleurs résultats : P1.

Tableau 13 Comparatif MPN vs. SPN

Méthode	Rappel	Précision	F-mesure
MPN	0,4	0,19	0,26
SPN	0,29	0,14	0,18

D'après ces résultats MPN est meilleure. Cela s'explique par le fait que les films sont reliés par les genres avec la perspective que nous utilisons (P1) et que MPN permet de rapprocher les films entre eux en fonction du nombre de genres qu'ils ont en commun. SPN au contraire ne saura pas différencier les distances entre deux films s'ils sont rapprochés par un ou trois genres par exemple. Nous comparons ensuite MPN à VSM (Tableau 14).

Tableau 14 Comparatif MPN vs.VSM

Méthode	Rappel	Précision	F-mesure
MPN	0,38	0,18	0,24
VSM	0,41	0,2	0,27

Pour ce test nous avons pris une valeur de diamètre de voisinage grande pour MPN (m = 3). Cela a pour défaut de rapprocher deux films par un ensemble de chemins beaucoup trop indirects. Voilà pourquoi VSM est meilleure. Le fait de rapprocher indirectement les films constitue une prise de risque qui pénalise la méthode MPN. Nous allons voir dans le prochain test si ce handicap s'amoindri quand on diminue la valeur de m (nous prendrons m = 1,5).

Tableau 15 Comparatif MPN vs. VSM avec m petit (1,5)

Méthode	Rappel	Précision	F-mesure
MPN	0,21	0,1	0,13
VSM	0,23	0,2	0,14

Quand nous diminuons le degré de transitivité (m) les résultats de MPN se rapprochent de ceux de VSM (Tableau 14). En effet les chemins indirects reliant deux films deviennent donc beaucoup moins indirects. A partir d'une certaine valeur de m les chemins entre deux films parcourus par MPN deviennent donc des chemins directs. Nous pouvons donc en limitant la transitivité reproduire avec MPN les résultats de VSM. Nous pouvons donc reproduire les résultats d'une méthode comme VSM qui n'est pas générique grâce à MPN qui est une méthode générique en diminuant la transitivité. En effet, VSM est une méthode qui doit être spécifiée pour chaque modèle car nous devons donner pour chacun de ces modèles les représentations algébriques (i.e. les vecteurs de descripteurs) de chaque ressource qui le constitue. Ici nous avons du spécifier notre façon de représenter sous forme vectorielle les films pour pouvoir calculer les distances film-film. Qui plus est, la distance VSM ne s'applique que sur deux objets de classe homogène alors que la distance que nous utilisons dans MPN et SPN peuvent s'appliquer entre deux ressources de classe différentes.

4.4.5 Discussions

Comme nous l'avons vu dans nos résultats sur la recommandation, pouvoir s'adapter sans coût pour imiter le fonctionnement de méthodes éprouvées en RI permet de garantir les meilleurs résultats dans les scénarios RI et recommandation. Toutefois la transitivité de nos méthodes peut se montrer utile dans une approche exploratoire sur les données sans spécialement d'objectif fixe visant à la découverte via des relations indirectes. Le fait de renvoyer des résultats indirectement voir très indirectement liés est un risque d'un point de vue RI car cela diminue notre précision. Mais si nous prenons ce risque, nous augmentons aussi notre rappel et les résultats indirectement liés donnent l'opportunité à l'utilisateur – par feedback – de créer de nouvelles connections que nous n'aurions pu envisager en prenant une faible transitivité. Dans le cas d'étude sur les films ce sont les méthodes qui se basent sur de multiples chemins qui montrent les meilleurs résultats. Il faut garder toutefois à l'esprit que ce que nous observons ici se base sur une comparaison des comparaisons de nos méthodes avec les fonctionnalités de recommandation de l'API TheMovieDataBase.

Nous retenons de cette période d'évaluation que pour répondre à des scénarios classiques de recommandation ou recherche d'information il valait mieux configurer nos méthodes afin qu'elles reproduisent le fonctionnement les méthodes classiques répondant à ces scénarios. Nous avons récemment publié un article [84] réutilisant ce jeu de données cinématographique dans le cadre d'une

expérimentation sur l'évolution de la structuration de connaissances cinématographique au fur et à mesure des interactions des utilisateurs.

Dans cette expérimentation, nous nous comparons avec d'autres approches à un gold standard dont la donnée nous est connue mais dont la méthode nous est inconnue. En effet, TheMovieDatabase ne communique pas sur son algorithme de recommandation. Les méthodes que nous comparons à ce gold standard – dont les nôtres – sont donc loin de celles qui sont utilisées dans le gold standard. Dans cette expérimentation, ce qui nous intéresse ce n'est pas le niveau absolu des rappels et précisions par rapport à ce gold standard mais leurs niveaux relatifs entre les approches que nous testons. Nous en tirons principalement une leçon sur le paramétrage le plus adapté pour nos méthodes afin de nous rapprocher au mieux d'un service de recommandation de films comme TMDB.

EN RESUME

Dans ce KG, nous abordons un sujet actuel touchant de nombreux internautes : la recommandation de films. Le prototype est au même niveau que le précédent mais propose en plus quelques options d'ergonomie. L'affichage de synopsis et affiches et visualisation de la Carte de Connaissance autour d'une ressource permet d'inspecter de plus près le jeu de donnéess.

Pour répondre à des scénarios classiques de recommandation ou recherche d'information il vaut mieux configurer nos méthodes afin qu'elles reproduisent le fonctionnement les méthodes classiques répondant à ces scénarios.

Si nous limitons la transitivité de nos méthodes nous nous rapprochons des résultats d'approches plus classiques comme VSM. Toutefois, même si elle réduit la précision, cette transitivité des méthodes ViewpointS – en particulier dans un scénario sur la recommandation de films – fait remonter des résultats plus indirectement liés mais pouvant mener à la découverte fortuite.

4.5 Benchmark des distances sémantiques de ViewpointS

Article référence:

G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, "Subjective and generic distance in ViewpointS," in Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS '16, 2016, pp. 1–6.

4.5.1 Objectifs

Le calcul de distance sémantique est l'élément central de ViewpointS puisque la seule relation sémantique que nous utilisons est celle de proximité ou distance dans une approche topologique de la

connaissance. Elles nous permettent de calculer des voisinages sémantiques. Elles sont aussi utiles dans des algorithmes de graphe tels que la détection de communauté. L'étude de l'état de l'art nous a inspiré la création de trois méthodes de calcul de distances basées sur un graphe de connaissance ViewpointS. Nous cherchons par l'étude suivante à comparer dans un même scénario ces méthodes et celles que nous mentionnions dans l'état de l'art.

4.5.2 Graphe de connaissances

Pour cela, nous nous basons sur un jeu de données différent pour offrir un benchmark complémentaire. Nous examinerons ici la qualité de nos méthodes dans le calcul de distances sémantiques. Pour ce faire nous disposons d'un gold standard de 354 distances entre paires de mots[53] (354D). Ces distances ont été obtenues par une étude en psychologie dans laquelle on demandait à un groupe de participants de juger la similarité entre mots. Nous allons indexer le jeu de données WordNet 2.0[54] afin de disposer d'un graphe de connaissances contenant les mots de notre gold standard. Grâce à nos méthodes nous allons calculer les 354 distances et les comparer à 1/ celles du gold standard et 2/ aux distances sémantiques qu'obtiennent des mesures reconnues implémentée dans une librairie spécialisée dans le calcul de distance sémantique[55]. La Figure 30 illustre le jeu de données WordNet. Des mots (Words) possèdent plusieurs sens (des WordSenses). « Dog » pourrait avoir plusieurs sens (des WordSenses). Mais « ws_dog » est le seul sens du mot « dog ». Des sens de mots peuvent être regroupés en groupes de synonymes (SynSets). C'est le cas pour le SynSet « s_(dog, domestic dog, Canis familiaris) ». Les SynSets peuvent être connectés par plusieurs relations sémantiques comme l'hypéronymie (isA).

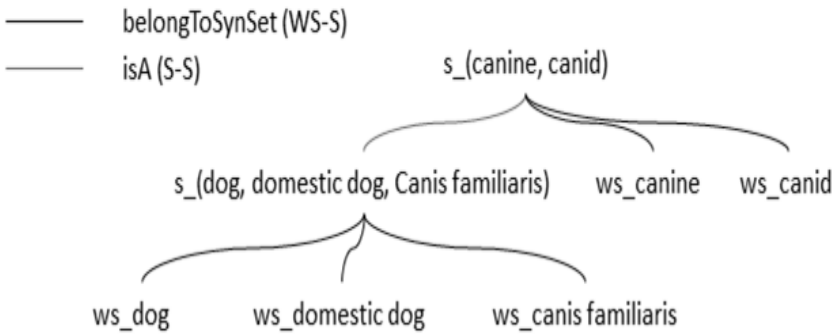


Figure 30 un exemple dans WordNet

Voici la liste des relations dans le jeu de données WordNet que nous avons choisi :

- Word-WordSense
- WordSense-WordSense: relation de type « see also ».
- WordSense-SynSet : appartenance d'un WordSense à un SynSet
- SynSet-SynSet :
    - Similarité sémantique (Similar)
    - Hyperonymie (isA)
    - Méronymie

[53] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/
[54] https://datahub.io/dataset/w3c-wordnet
[55] http://www.semantic-measures-library.org/sml/

4.5.3 Déroulement de l'expérimentation

Nous donnons un ensemble de Perspectives différentes qui ont chacune un sens propre dans la façon de voir les données qu'elles proposent. Le Tableau 16 résume les perspectives que nous employons dans ce benchmark sur la distance sémantique. Comme les méthodes que nous utilisons sont des méthodes topologiques celles-ci se basent sur l'exploration de chemins dans le graphe de connaissances. Les différents paramétrages de Perspective correspondent donc à des ordres de priorité dans les chemins à explorer. Ainsi, s'il existe plusieurs chemins possibles entre deux mots nous choisirons de privilégier tel ou tel chemin. Privilégie-t-on les chemins se basant sur les relations entre SynSets ou les chemins passant par les relations entre WordSenses ?

Nous allons voir dans les résultats l'impact de ces différents paramétrages de Perspective. Le Tableau 17 résume les ordres de priorité de chemins pour chaque Perspective du Tableau 16. Par exemple, avec la Perspective P4 nous favorisons en premier lieux la relation de Similarité entre SynSets donc nous emprunterons avant tout les chemins W – WS – S –Similar– S – WS – W.

Tableau 16 Perspectives pour le benchmark sur les distances sémantiques

	W – WS	WS – S	Méronymie (m)	Hypéronymie (h)	Similarité (s)	SeeAlso (sA)
P1	1	1	1	5	1	5
P2	5	1	4	5	7	1
P3	5	1	5	4	7	1
P4	5	1	4	7	5	1
P5	1	1	5	4	7	7
P6	1	1	4	7	5	7

Tableau 17 Ordre de priorité en chemin pour chaque Perspective.

P1	P2	P3	P4	P5	P6
See Also	Hypéronymie	Hyperonymie	Similarité	See Also	See Also
Similarité	Similarité	Méronymie	Hyperonymie	Hyperonymie	Similarité
	Méronymie	Similarité	Méronymie	Méronymie	Hyperonymie
				Similarité	Similarité

Nous utilisons les implémentations des mesures de distance ou de similarité sémantique de SML[56]. Il s'agit d'une libraire java unifiant un grand nombre de mesures de distance sémantiques de l'état de l'art. Nous utiliserons les mesures de Wu & Palmer et de Lin. Les résultats seront exprimés en termes de ratio de précision par rapport au gold standard. Pour une distance à évaluer d_{test} et une distance d'évaluation d_{gold} (appartenant au gold standard) la précision est le pourcentage :

[56] http://www.semantic-measures-library.org/sml/

$$precision = 100 - \frac{abs(d_{test} - d_{gold}) \times 100}{d_{gold}}$$

4.5.4 Résultats

Nous comparons premièrement les distances obtenues par nos deux méthodes utilisant le plus court chemin et l'ensemble des chemins entre deux concepts aux 354 distances du gold standard dans La Figure 31. Les résultats sont exprimés en pourcentage de précision par rapport au gold standard.

On peut noter tout d'abord que l'effet du changement de Perspective affecte beaucoup plus SPD que MPD. En effet, le plus court chemin peut changer radicalement d'une perspective à l'autre donc la SP-distance. On remarque ensuite que la meilleure Perspective est P4, c'est-à-dire celle qui donne la priorité aux relations entre SynSets 1/ de similarité, 2/ hyponymie et 3/ méronymie. Cela paraît logique que ce soit en mettant le focus sur la relation de similarité entre collections de synonymes que nous obtenons les meilleurs résultats. La meilleure combinaison méthode-perspective est MPD-P4.

Comme pour le benchmark précédent, on remarque que MPD tire vraiment avantage de la diversité des relations sémantiques tandis que SPD ne se base que sur le plus court chemin et n'exploite que peu du potentiel du graphe de connaissance WordNet. Nous comparons ensuite dans les graphiques de la Figure 32 nos mesures (SPD et MPD) aux méthodes de calcul de distance sémantique.

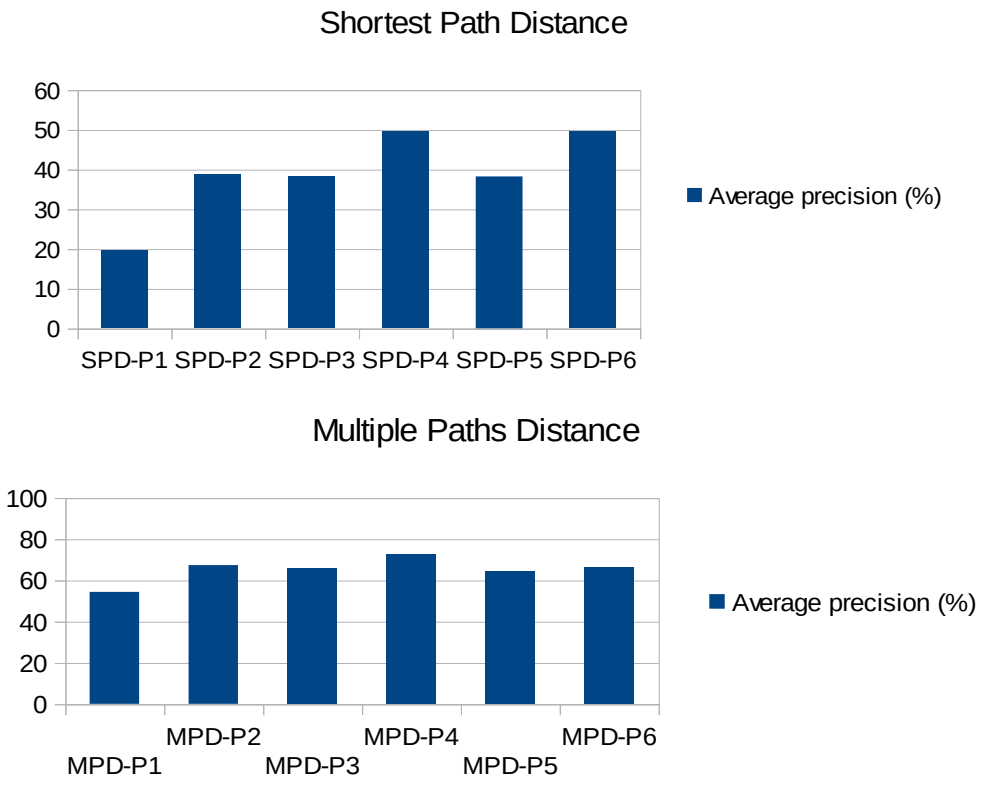


Figure 31 Comparaison de nos méthodes utilisant le panel de perspectives de test avec le gold standard des 354 distances.

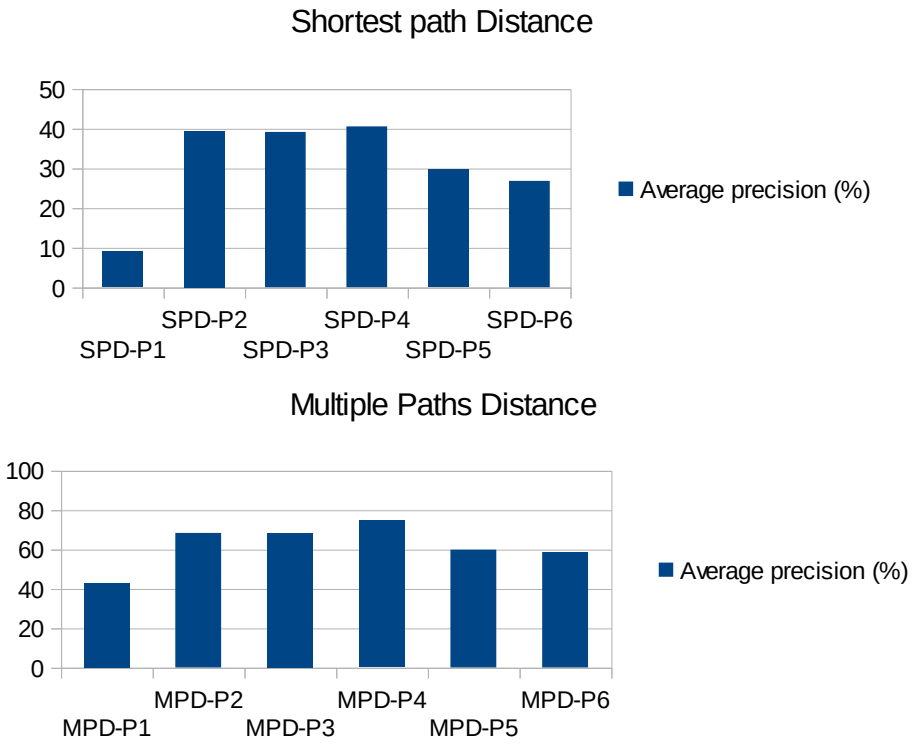


Figure 32 Comparaison de nos méthodes utilisant le panel de perspectives de test avec la distance de Lin.

Lin étant une méthode basé sur le contenu informationnel, nous ne sommes pas surpris de trouver de meilleurs résultats grâce à la méthode se basant sur des chemins multiples pour calculer la distance sémantique. Nous comparons ensuite les résultats de SPD et MPD avec les résultats de la méthode Wu & Palmer dans la Figure 33.

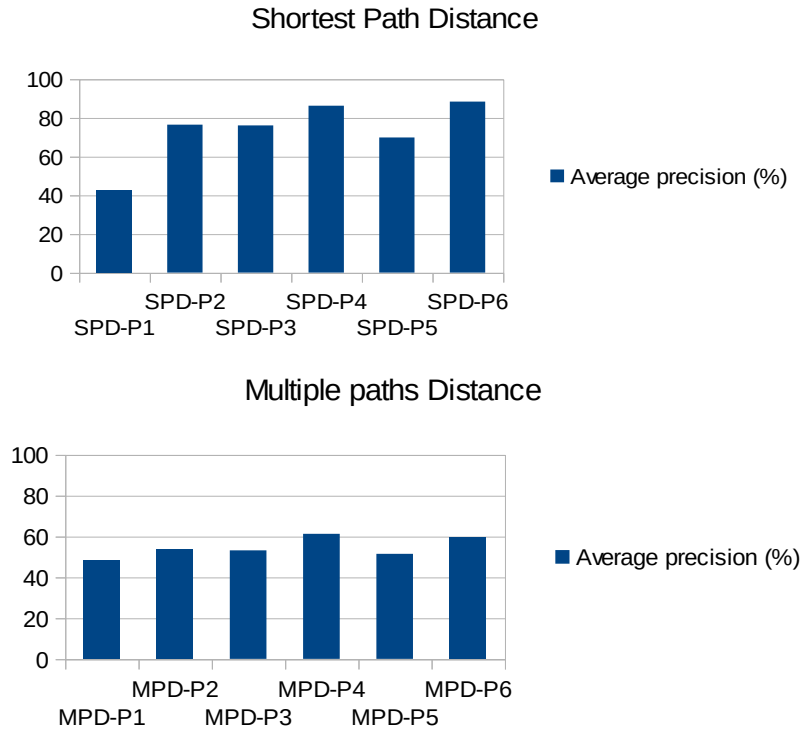


Figure 33 Comparaison de nos méthodes utilisant le panel de perspectives de test avec la distance de Lin.

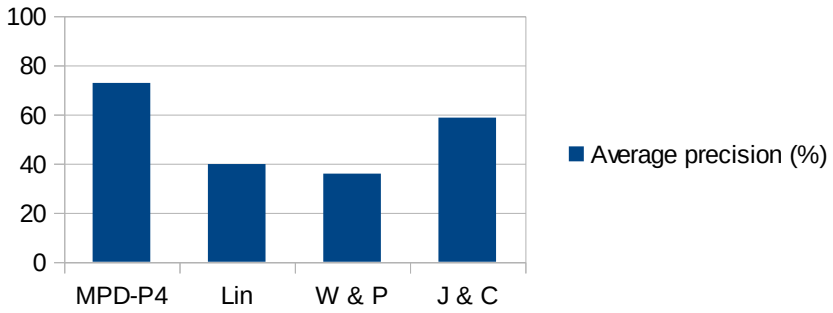


Figure 34 Récapitulatif des résultats par rapport au gold standard 354D.

4.5.5 Discussions

Pour conclure cette évaluation de nos méthodes nous pouvons grâce à un à une diminution de la transitivité de nos méthodes nous ramener au fonctionnement de méthodes classiques comme Vector Space Model. Dans nos méthodes génériques nous n'avons pas à spécifier, comme pour VSM, les vecteurs de descripteur pour chaque type de ressources. Et nous pouvons calculer une distance grâce à la méthode MultiPath ou ShortestPath entre deux ressources de classes différentes ce qui est impossible dans le Vector Space Model. Nous montrons aussi que nous pouvons nous ramener à des résultats en termes de calcul de distance sémantique assez proche de ceux qu'obtiennent les méthodes comme celles de Lin et Wu & Palmer. Toutefois ces méthodes ont été conçues pour être utilisées sur une base de connaissance possédant une structure taxonomique alors que nos méthodes sur le graphe de connaissance ViewpointS peuvent s'appliquer sans cette restriction de structure.

EN RESUME

Nous avons évalué l'efficacité en termes de précision des méthodes de distance sémantiques de ViewpointS (SPD, MPD) en les comparant sur un même jeu de donnéess (WordNet) et sur la base d'un même gold standard (woldsim 354) à des méthodes connues.

En paramétrant correctement la perspective, nous parvenons à obtenir des résultats similaires voir meilleurs que les méthodes empruntés à la littérature.

Alors que les méthodes comme Vector Space Model doivent être spécifiées pour le jeu de donnéess, notre approche topologique et générique ne demande aucune spécification préalable en particulier en termes de structuration des connaissances.

4.6 Évaluation de la suggestion de traductions dans ViewpointS

4.6.1 Objectifs

Le but de cette partie est de nous appuyer sur un nouveau scénario concret d'exploitation de l'approche ViewpointS qui s'inscrit dans le contexte du projet SIFR. Le projet SIFR vise à indexer des données biomédicales (ex. : rapports de recherche, tests cliniques, données génomiques) grâce aux connaissances structurées des ontologies biomédicales. L'indexation de ressources biomédicale à l'aide d'annotations permet la recherche sémantique de ces ressources : quelles sont les publications sur les cancers apparentés au cancer du sein ? Quels sont les auteurs qui publient sur ce sujet ?

L'indexation sémantique enrichie les méthodes d'indexation classique (généralement associant des mots-clés aux ressources) à l'aide de toutes les relations sémantiques définies dans les ontologies utilisées (hypéronymie, proximité sémantique, métonymie etc.). Cette indexation fournit une meilleure recherche sur des termes biomédicaux comme « cancer ».

> ***Définition – Annotation***
>
> *L'annotation est une liaison entre un support de connaissance et un descripteur de connaissance du Web Sémantique. Il s'agit par exemple de repérer les mots-clés dans un texte par l'analyse statistique et de les faire correspondre avec des concepts d'ontologie portant le même nom. Cela relie le document au Web Sémantique.*

Par exemple, la liste d'article résultats, bénéficiera de divers types de relations appartenant à la fois au réseau d'auteurs (relations de publication) et à la classification par des descripteurs (mots-clés). Cette connaissance structurée est le fruit du travail d'une communauté d'experts et rassemble plusieurs ontologies et alignements d'ontologie. Des méthodes de Traitement Automatique du Langage Naturel (TALN) telles que celles construites par Juan Antonio Lossio-Ventura [93] extraient les mots-clés des documents par traitement statistique (ex., fréquence d'apparition des mots-clés). Si les mots-clés trouvés sont des concepts présents dans nos ontologies de référence alors l'annotation est faite entre la ressource annotée et le concept d'ontologie La spécificité du projet SIFR est de s'axer sur une annotation de données textuelle en français et d'apporter une valeur ajoutée multilingue aux projets NCBO BioPortal développé par l'Université de Stanford[57].

La recherche devient multilingue et même si elle ne s'axe que sur une seule langue – si nous ne demandons que des documents français par exemple – elle peut bénéficier des connaissances formalisées dans des ontologies utilisant principalement d'autres langages ou multilingues mais alignées les unes les autres. Il existe d'ores et déjà un annotateur français, le SIFR Annotator[58], permettant de lier des ressources textuelles écrites en français aux concepts des ontologies ou terminologies médicales françaises. La Figure 35 illustre le processus d'annotation mis en place dans le projet SIFR.

[57] http://bioportal.bioontology.org/
[58] http://bioportal.lirmm.fr/annotator

4.6.2 Graphe de connaissance

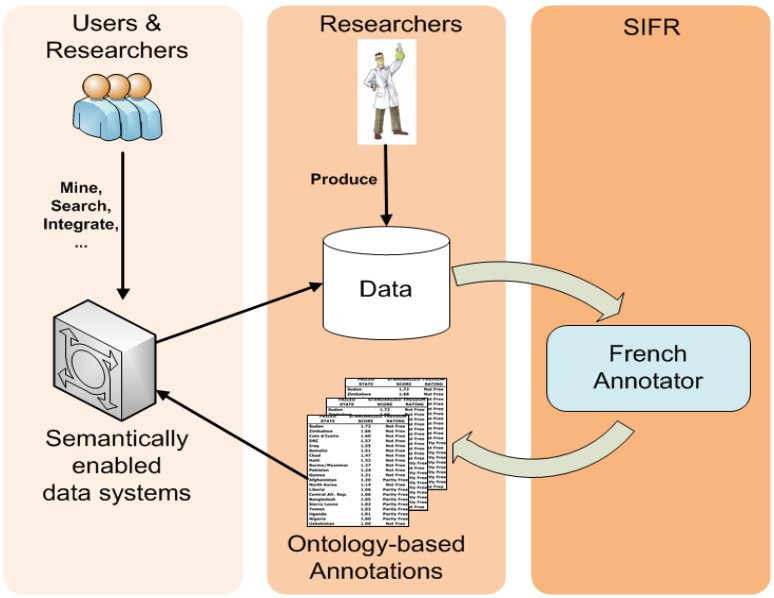


Figure 35 Illustration du processus d'annotation (origine : projet SIFR).

ViewpointS permet d'interpréter sous forme de subjectivités (de viewpoints) des connaissances venant 1/ des jeux de données eux-mêmes (ontologies, listes d'articles, etc.) et 2/ des annotations que produisent les annotateurs (français, anglais, automatique ou pas). Les données du web sémantique servent à organiser les données documentaires car ces documents sont liés à des descripteurs (concepts d'ontologie) qui sont structurés entre eux. Une fois cette connaissance représentée sous forme de graphe nous pouvons offrir plusieurs procédures d'exploitation des connaissances : 1/ une exploration du jeu de données grâce à un prototype de moteur de recherche, 2/ un service de suggestion d'alignements de façon à ce que les données documentaires permettent l'enrichissement des ontologies.

Nous allons nourrir un graphe de connaissance à partir d'un extrait de données bibliographiques venant de PubMed. Nous allons ensuite annoter les articles PubMed à partir de deux ontologies qui sont la traduction l'une de l'autre et intégrer ces annotations dans le graphe de connaissance sous forme de viewpoints émis par chacun des annotateurs. Chacun de ces annotateurs relie donc à la version de MeSH correspondant à sa langue les articles de PubMed. Si les articles sont doublement annotés (anglais et français) ils nous servirons de passerelle entre les versions françaises et anglaises de MeSH.

Nous partons donc de l'hypothèse que l'activité des auteurs pourrait permettre de rapprocher les concepts de deux langues différentes qui sont aussi de potentielles traductions. Par exemple le fait que Jacques Ferber aie écrit un article français et un article anglais, l'un annoté par « systèmes multi-agents » et l'autre par « multi-agent systems » rapproche ces deux concepts qui sont alors suggérés comme traduction.

Nous choisissons d'indexer les données suivantes présentes sur le SIFR BioPortal. Une liste d'article est doublement annotée par les concepts de deux ontologies de langues différentes (française et anglaise). Ces articles (NumericDocuments) sont connectés à leurs auteurs (Authors, spécialisation de HumanAgent rajoutant des informations comme l'adresse courriel).

L'activité scientifique des auteurs que nous représentons ouvre la voie pour le rapprochement des connaissances représentées en plusieurs langues car beaucoup d'auteurs ont publié dans plus d'un langage. Cela créé un pont entre les connaissances collectives de langues différentes. En outre, étant donné que de tels alignements (traduction) ont d'ores et déjà été produits pour plusieurs ontologies comme MeSH, nous pourrons ignorer ces informations dans notre démarche d'indexation et les utiliser pour l'évaluation de nos résultats.

4.6.2.1 Données PubMed

Ces données sont contenues dans un fichier XML recensant de 2000 articles PubMed et leurs auteurs. Dans un KG spécifique, nous créons trois types de ressources : les annotations, les articles et les auteurs. La Figure 36 illustre la connaissance subjective extraite du jeu de données PubMed.

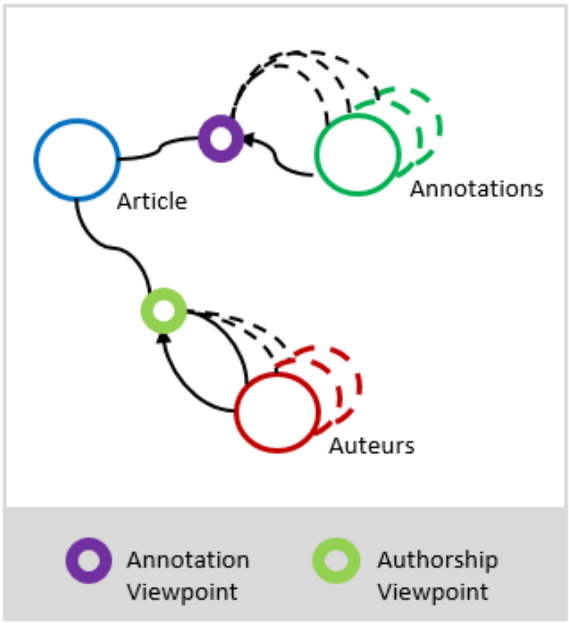


Figure 36 Illustration des viewpoints et ressources extraites du jeu de données PubMed.

4.6.2.2 Annotations françaises

Grâce au web service du SIFR Annotator[59] nous annotons tous les articles en se basant sur les titres français en se basant sur des requêtes http. Pour chaque annotation sur un article nous créons donc la ressource correspondante à l'annotation (si elle n'existe pas déjà dans le graphe de connaissance) et la relions à l'article par un viewpoints de type « Annotation Viewpoint » qui est émis par un agent artificiel représentant l'annotateur. Nous créons un type de viewpoints spécifique – type Annotation-Viewpoint – qui dérive du type Match afin de représenter les annotations.

[59] http://services.bioportal.lirmm.fr/annotator

4.6.2.3 Annotations anglaises

De la même manière nous enrichissons notre graphe des annotations anglaises sur les titres anglais grâce au web service du NCBO Annotator[60].

4.6.2.4 Données MESH

Nous nous servons en données sur les deux BioPortals respectifs pour avoir les ontologies françaises et anglaises au format rdf/xml. L'ontologie établie des proximités sémantiques (de diverses natures) entre des concepts biomédicaux.

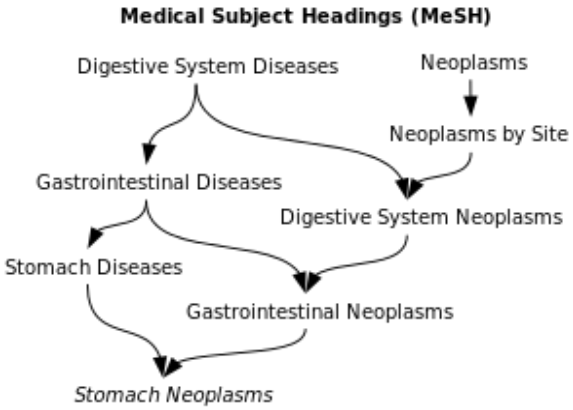


Figure 37 Un exemple que l'on peut trouver dans MESH FR.

MESH permet ainsi de relier les unes aux autres les concepts reliées aux ressources par annotation. Ces concepts correspondent à des domaines (microbiologie, traumatologie etc.), des organes, des maladies. La Figure 37 illustre en partie ces connaissances. Nous créons des viewpoints de type Match à partir de ces relations.

4.6.2.5 Challenge du passage à l'échelle

La taille importante des jeux de données pose le problème du passage à l'échelle. Le seul jeu de donnéess PubMed annoté par les annotateurs francophones et anglophones produit 23k ressources et 273361 viewpoints. En prenant en plus en compte la totalité des concepts des ontologies et les annotations étendues nous dépassons les 2M de viewpoints. Alors que SPN (ShortestPathNeighbourhood) est indépendant de la taille du graphe dans sa complexité due à la simplicité du calcul du plus court chemin, MPN ne permet pas en l'état (sans optimisations techniques) d'obtenir une réponse en un temps de calcul suffisant pour le systématiser dans un benchmark. Pour répondre à cette difficulté nous allons agir premièrement sur la taille des données en réduisant la quantité d'annotations et le nombre d'articles PubMed. Mais nous utiliserons également une nouvelle méthode de calcul de voisinage sémantique qui représente le meilleurs compromis temps de calcul/qualité. MultiFlowsNeighbourhood est la fonction que nous avons choisie. Pour plus de détails sur la complexité et les temps de calculs constaté le lecteur peut trouver en annexe un benchmark (Annexe 2) de passage à l'échelle des méthodes de ViewpointS.

[60] https://data.bioontology.org/annotator

La Figure 38 représente l'ensemble des données indexées dans notre cas d'étude. Le bas de la figure représente trois types de chemins représentatifs pouvant rapprocher deux concepts de langues différentes (ici mélanome et melanoma). Premièrement, mélanome et melanoma sont rapprochés directement par un document qui a été annoté à la fois par l'annotateur anglais et français (1). Pour cela le document devrait avoir (idéalement) du contenu français et anglais. Deuxièmement, si mélanome et melanoma ne sont pas rapprochés directement par un document doublement annoté alors il se peut qu'un auteur aie écrit un article en français et en anglais parlant de mélanome/melanoma (2). Mélanome et melanoma sont alors rapprochés par l'auteur a. On peut imaginer des chemins encore plus indirects comme (3) passant par les relations ontologiques entre concepts t. Mais il est possible d'imaginer des chemins beaucoup plus indirects passant à la fois le réseau d'auteur comme dans (2) et par les relations ontologiques (3).

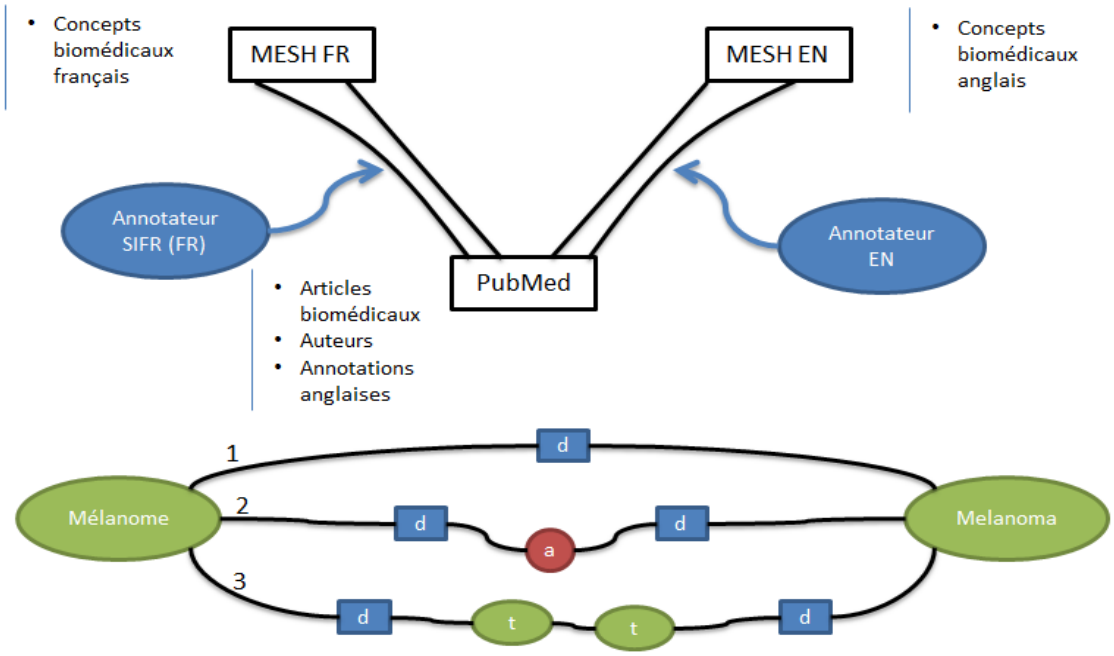


Figure 38 Illustration des données du cas d'étude.

4.6.3 Déroulement de l'expérimentation

Nous allons évaluer le multilinguisme sur la base d'un ensemble de concepts choisis aléatoirement dans MESH FR. Pour chaque concept C_i de cet échantillon E nous obtenons sont voisinage $^{m,p}MFN(C_i)$. Nous notons d'abord pour chaque concept si la traduction est – selon l'alignement/gold standard – dans $^{m,p}MFN(C_i)$. Le score de traduction @n est le nombre moyen de traduction trouvées dans les n premiers voisins. L'ensemble des résultats peut toutefois contenir des termes qui ne sont pas des traductions directes mais incluant la traduction comme « malignant melanoma » pour le terme mélanome. Nous ne pourrons détecter ces positifs que s'ils sont inclus dans l'alignement de référence. Nous comptabiliserons comme bonnes uniquement les traductions présentes dans le gold standard.

Voici deux exemples de voisinages obtenus :

MFN(melanoma, m = 2.0, p = 1.5)	
Insuffisance	
Colon	
Cancer colorectal métastasique	
Mélatonine	
Tumeurs	
Tumeur myofibroblastique	
Mélanome achromique	Traduction contenue dans le terme
Tumeurs ovariennes	
Kyste épidermoïde	
Gastrotectomie	

MFN(menopause, m = 2.0, p = 1.5)	
Fibromes utérins	
Occlusion	
Anesthésique	
Ménopause	Traduction directe
Léiomyome	
Cancer de la prostate métastasique	
Mort cellulaire	
Léiomyome	
Statines	
Gliosarcome	

On peut voir dans les deux voisinages précédents deux exemples de traductions directe ou contenu. Alors que les traductions directes comme « Ménopause » dans le second exemple sont présentes dans notre alignement-référence il est nécessaire de faire quelques traitements pour détecter les traductions « indirectes ». Nous vérifions si parmi les voisins la ou les traductions correctes sont contenues dans les labels des voisins.

4.6.4 Résultats

Premièrement nous donneront un poids maximal au lien d'article à auteur dans la partie « Les auteurs alignent ». Ensuite nous testons dans « Les annotations » une perspective donnant de l'importance aux relations d'annotation à article. Pour finir, nous montrons les résultats de la perspective qui nous a semblé la meilleure à suggérer des traductions. Dans chacune des parties suivantes, nous expérimentons plusieurs facteurs de m (suivant la définition précédente). Sachant que nous nous basons sur 2000 articles PubMed nous limiterons la taille du graphe résultant en ne prenant en compte que 10 annotations maximum par article (5 par langue). Nous allons jouer sur le facteur de parallélisme de MFN et observer l'effet de son augmentation. Plus nous l'augmenterons plus le post-traitement basé sur de multiples chemins de MFN sera complet et prendra de chemins parallèles.

4.6.4.1 Les auteurs alignent

Les résultats du Tableau 18 nous montrent qu'une forte transitivité des chemins entre concepts passant par les réseaux d'auteurs bénéfice beaucoup à l'exploitation de multiples chemins. Toutefois on note qu'à partir d'un seuil du facteur de parallélisme l'exploitation de chemins beaucoup plus indirects (on utilise des chemins 5 fois plus longs pour post-traiter une synapse pré-calculée par SPN) pollue nos résultats. En effet, si un auteur publie sur deux sujets qui sont très différents il n'y a que l'un (ou peu) d'entre eux qui nous puisse contribuer à notre recherche. On ne peut pas passer à côté d'exploiter la totalité de la production de cet auteur et ses contributions sur le second sujet (celui qui ne nous intéresse pas) pollue nos résultats de recherche en plus d'alourdir de beaucoup les calculs de MFN.

Tableau 18 Perspective auteur-centrée.

Type de viewpoint	Match	Annotation	Authorship
Poids	1	1	10

Fonction de voisinage	$^{3,1}MFN(C_i)$	$^{3,2}MFN(C_i)$	$^{3,5}MFN(C_i)$
$score_{traduction}$	0.54	0.58	0.27

4.6.4.2 Les annotations alignent

Notre hypothèse – basée sur les observations précédentes – est que l'exploitation de chemins très indirects remontant par exemple plusieurs relations d'hypéronymie risque de polluer nos résultats de recherche.

Tableau 19 Perspective annotation-centrée.

Type de viewpoint	Match	Annotation	Authorship
Poids	2	5	1

Fonction de voisinage	$^{3,1}MFN(C_i)$	$^{3,2}MFN(C_i)$	$^{3,4}MFN(C_i)$
$score_{traduction}$	0.61	0.52	0.16

Dans KG le sous-graphe constitué des viewpoints et ressources venant de MeSH FR/EN est beaucoup plus dense que le réseau d'auteurs provenant des articles PubMed. L'augmentation du facteur de parallélisme fait directement chuter le score de traduction. Nous pensons qu'avec un meilleur calibrage de perspective nous obtiendrons des résultats de MFN tirant parti de son parallélisme améliorant le score de traduction.

4.6.4.3 Essai de perspective optimisée

On peut d'ores et déjà faire quelques remarques concernant la perspective qui nous est apparue comme la meilleure pour notre problème. On s'aperçoit que le système de double annotateur a un impact positif sur notre capacité à suggérer des traductions. Toutefois en réduisant le poids des relations venant de MeSH on pourra augmenter le facteur de parallélisme sans faire chuter le score de traduction. Après plusieurs tests, la perspective du Tableau 20 a obtenu les meilleurs scores de précision.

Tableau 20 Perspective jugée optimale.

Type de viewpoint	Match	Annotation	Authorship
Poids	2	4	6

Fonction de voisinage	$^{3,1}MFN(C_i)$	$^{3,2}MFN(C_i)$	$^{3,5}MFN(C_i)$
$score_{traduction}$	0.65	0.73	0.12

On remarque premièrement que quand on configure la perspective pour exploiter les données venant du réseau d'auteurs de PubMed et celles qui viennent des MeSHs l'effet négatif d'un grand parallélisme de MFN s'accentue. Toutefois, nous obtenons de meilleurs en élevant un peu le facteur de parallélisme.

4.6.5 Discussions

Ce sont toutefois des résultats produits par une indexation spécifique et réduite de MeSH que nous exploitons en utilisant une perspective construite arbitrairement. Nous pensons – d'après les résultats obtenus – pouvoir proposer un service qui suggère 10 traductions pour un terme biomédical et dans 73% des cas arrivent à trouver une traduction parmi ses 10 essais. Ce service tire parti des derniers développements de nos algorithmes de voisinage afin de tirer parti de multiples chemins dans l'exploration de KG tout en ayant un certain passage à l'échelle. Alors que le travail d'alignement pour des ontologies majeures comme MeSH se base principalement sur un travail d'expertise, le service de suggestion pourrait réduire la masse de travail pour l'agent humain dans la création de ces alignements. Le rôle de la perspective et de la subjectivité dans ce service de suggestion d'alignement multilingue est de donner de plus amples options de customisation de MFN et de trouver le bon équilibre des forces dans l'interprétation des viewpoints. Si nous souhaitons que chacune des sources de connaissances (PubMed, MeSH, annotateurs) contribuent au mieux nous devons disposer d'un moyen de réguler l'importance des très gros jeux de données et de promouvoir les liens à l'intérieur de jeux de données plus petits. C'est ce que nous avons fait en allouant au viewpoints d'annotation un poids inférieur au viewpoints de type Authorship.

Au final, ce qui nous intéressait principalement dans cette expérimentation c'est de voir le potentiel multilingue de l'approche ViewpointS. C'est pourquoi nous proposons des résultats surtout en matière de suggestion de traduction plutôt qu'en termes d'alignement d'ontologie.

EN RESUME

Nous mettons en œuvre ViewpointS au service d'un système de suggestion d'alignement multilingue entre une ontologie biomédicale française et une autre anglaise.

Nous nous basons sur le réseau d'auteurs qui publie dans les termes de ces deux ontologies pour permettre à l'activité scientifique des auteurs de rapprocher sémantiquement les descripteurs de ces ontologies. Pour un terme biomédical français nous suggérons 10 propositions de traduction uniquement en nous basant sur la distance sémantique et un bon paramétrage de la perspective.

Nos résultats montrent que dans 73% des cas, sur un graphe de connaissance de 22k ressources et plus de 200k viewpoints, une traduction sur les 10 suggérées est une traduction correcte.

Chapitre 5. ViewpointS Web Application

Cette thèse contient une dimension d'ingénierie essentielle. C'est pourquoi à chaque étape dans la maturation du projet nous avons cherché à produire des prototypes démontrant les capacités de l'approche ViewpointS sur des jeux de données réels et à déterminer à quels besoins pouvait correspondre notre approche. Le lecteur peut accéder à l'adresse http//viewpoints.cirad.fr/home un site présentant le projet, nos publications et les diverses applications de la ViewpointS' Web App (VWA). Nous permettons au lecteur de parcourir – grâce à VWA – la plupart des jeux de données sur lesquels nous nous sommes basés pour les expérimentations du Chapitre 4. Ainsi il lui sera possible d'expérimenter par lui-même les fonctionnalités de ViewpointS dans un scénario impliquant des données réelles. Nous présentons trois versions de VWA. L'une d'entre elle se base sur un mix de deux jeux de données bibliographiques du CIRAD et du LIRMM présentée dans la section 4.2. La seconde se construit sur un graphe de connaissances créé à partir de données cinématographiques que nous explorons dans le benchmark sur la recommandation de film de la section 4.4. Pour finir, un graphe de connaissances alimentées de jeux de données biomédicales – présentés dans la section 4.6 – permet la version BioMed de VWA. Ce Chapitre présente VWA ainsi que l'API qui permet l'implémentation du modèle ViewpointS.

5.1 Objectifs

Les précédents prototypes ont permis de tester les fonctionnalités offertes par l'approche ViewpointS en termes de Recherche d'Information dans le cadre fermé de notre équipe. Nous souhaitons permettre à quiconque d'essayer la fonctionnalité de Recherche d'Information dans ViewpointS. Ce prototype démonstrateur public a vocation à faciliter la dissémination du projet.

Notre objectif avec VWA est de fournir une interface générique qui permet d'explorer et d'interagir avec n'importe quel graphe de connaissance ViewpointS quelle que soit la nature et la structure des connaissances qu'il contient. Nous pensons VWA comme un moteur de recherche ayant des fonctionnalités complémentaires à la recherche par mot-clé. Ici nous cherchons à offrir un service de recherche pouvant répondre à une demande précise ainsi qu'une possibilité d'exploration pour l'utilisateur qui n'a pas d'objectif entièrement clair en tête. Cependant, comme nous l'évoquons dans le Tableau 1 de la section 2.1.1, notre approche qui se base sur la sémantique la plus pauvre ne comprenant que la relation sémantique la plus universelle – la similarité sémantique – se base surtout sur l'interaction avec l'utilisateur. L'approche topologique renvoie plus de résultats qui sont souvent in-

directement lié à la requête initiale et prend de fait le risque de renvoyer des résultats distants. La création de viewpoints par feedback permet de valoriser/dévaloriser les chemins dans le graphe de connaissance qui ont mené de la requête à la réponse. Cela permet en cas de découverte fortuite de garder la trace de la Sérendipité.

5.2 Spécifications

Ce prototype démonstrateur doit remplir toutes les fonctionnalités des versions précédentes : (i) faire une recherche au voisinage d'une ressource, (ii) paramétrer la Perspective et (iii) enrichir le graphe de connaissance soit par feedback soit par l'ajout manuel de ressources et de viewpoints soit par l'indexation de jeux de données. La Figure 39 est le diagramme d'utilisation correspondant à cet usage de base de VWA. Mais étant donné que le prototype illustre l'approche ViewpointS nous avons ajoutés plusieurs fonctionnalités qui permettent à l'utilisateur d'inspecter le fonctionnement interne de ViewpointS. Pour ce faire, nous avons donné la possibilité de manipuler le graphe de connaissance « à la main » en ajoutant des viewpoints créés en dehors du mécanisme de feedback. L'utilisateur peut donc créer un petit graphe-jouet en créant manuellement les ressources et viewpoints qu'il veut et y appliquer les méthodes de ViewpointS. Toutefois, du graphe de connaissance aux résultats de son exploitation, l'utilisateur ne voit pas l'étape intermédiaire qu'est la construction de la Knowledge Map. C'est pourquoi nous avons décidé d'ajouter deux fonctionnalités :

- La Carte de Connaissances locale permet la représentation de la partie de la Carte des Connaissances au voisinage d'une ressource. Cela permet de visualiser le voisinage d'une autre manière que par la liste de résultats en représentant les synapses construites par le mécanisme de Perspective. Le lecteur pourra expérimenter lui-même cette fonctionnalité sur les trois prototypes ouvrant le panneau « Knowledge Graph Vizualisation » et en lançant une recherche (bouton « Search »).
- La fonctionnalité « Shortest Path » permet de visualiser le plus court chemin entre deux ressources sélectionnées dans la Carte de Connaissances. Par exemple, après une recherche, afficher les plus courts chemins entre la ressource recherchée et l'une des ressources du voisinage revient à afficher une « justification » du résultat. En effet, on peut voir alors le cheminement dans la Knowledge Map qui a mené de la ressource recherchée à la ressource-résultat.

La Figure 40 montre un aperçu de l'application. Dans cet exemple, Viewpoints Web App est un moteur de recherche bibliographique du même style que celui que nous présentions dans la section 4.2. Il est basé sur l'indexation des données bibliographiques HAL LIRMM[61] et HAL CIRAD[62]. Nous cherchons le voisinage sémantique de Guillaume Surroca. On y trouve ses articles, ses collaborateurs, les sujets sur lesquels il a écrit et il est possible pour chacun de ces résultats de :

- d'aimer ou non un résultat pour le repositionner par rapport à soit en le distanciant ou en le rapprochant ;
- de juger les résultats pertinents ou pas en renforçant ou atténuant grâce au feedback fragmenté le chemin « synaptique » qui a mené au résultat ;

[61] https://hal-lirmm.ccsd.cnrs.fr/
[62] http://hal.cirad.fr/

- d’avoir une meilleure appréciation du résultat en voyant la justification d’un résultat. Il est ainsi possible de visionner le chemin entre deux ressources connexes ;
- d’observer graphiquement la Knowledge Map autour de la ressource recherchée comme autre visualisation du voisinage sémantique ;
- de paramétrer la perspective et les méthodes de calcul de voisinage sémantique.

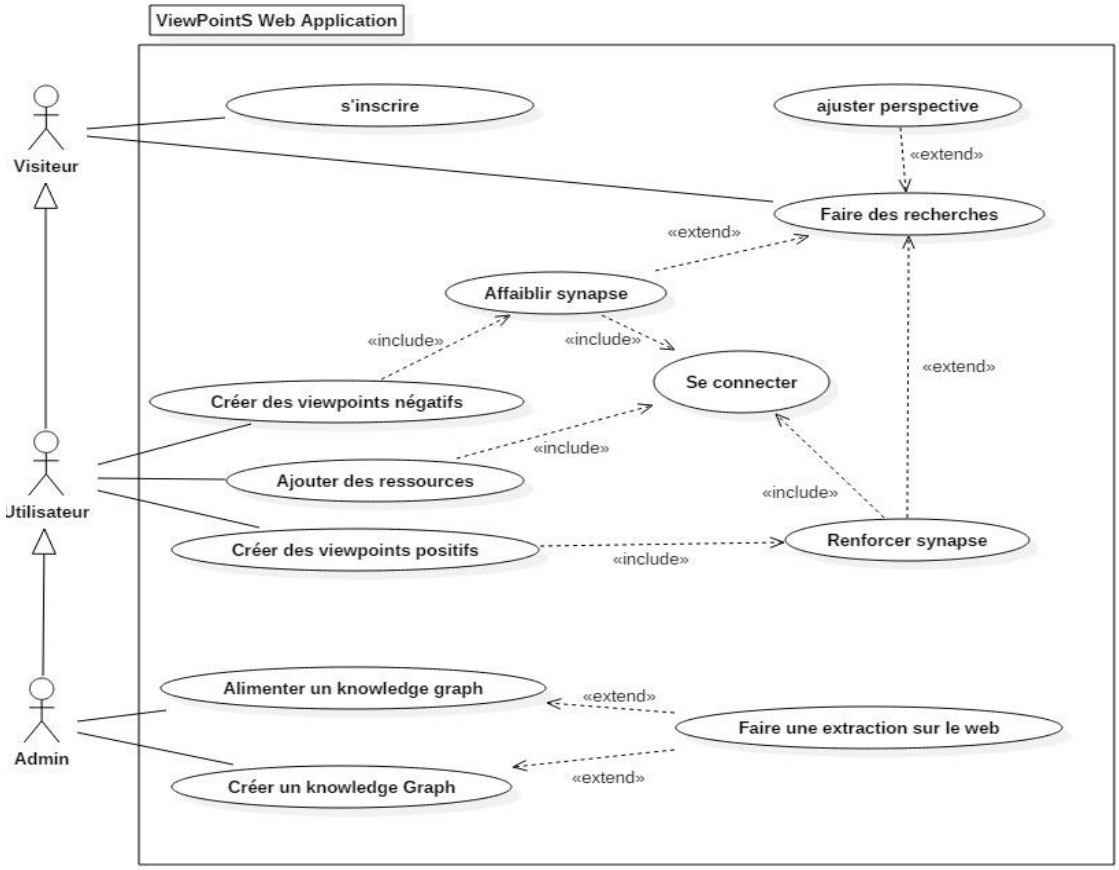


Figure 39 Diagramme de cas d’utilisation du moteur de recherche VWA

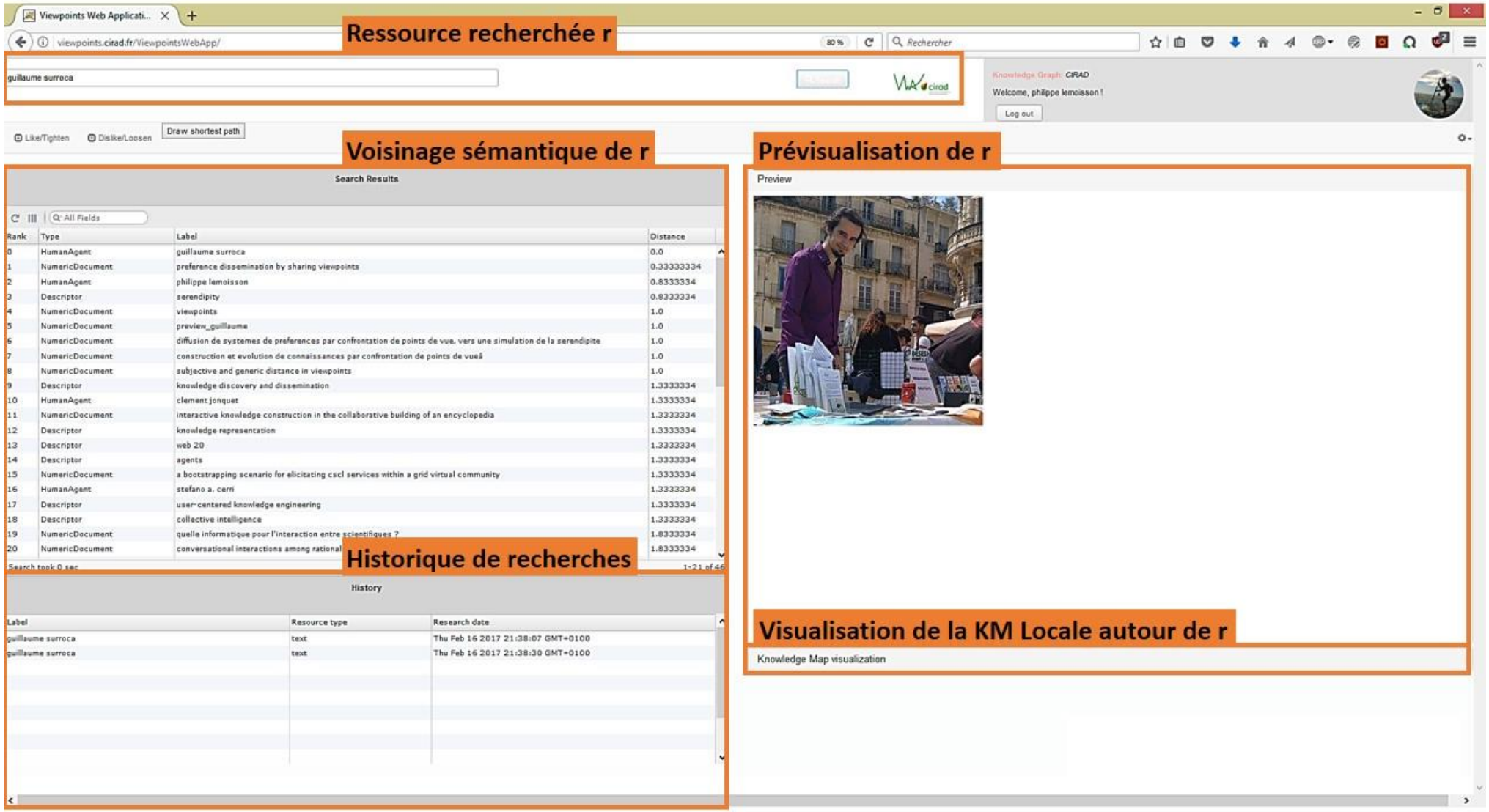


Figure 40 Illustration d'ensemble de VWA

5.3 Présentation de VWA

La Figure précédente montre l'accueil de VWA. Après s'être identifié, l'utilisateur peut interagir avec le graphe de connaissances. On y trouve en haut à gauche une barre qui permet de trouver une ressource par auto-complétion et de chercher son voisinage sémantique. Une liste de résultats que l'utilisateur peut trier et filtrer à sa guise apparaît ensuite dans le tableau central. En bas du tableau des résultats se trouve un historique des dernières recherches de l'utilisateur. La partie droite de l'interface est dédié à une visualisation complémentaire composée d'une fenêtre de prévisualisation ainsi qu'à une visualisation graphique de la carte de connaissances. Il suffit de double cliquer sur une des ressources de la liste de résultats pour afficher soit sa prévisualisation dans la fenêtre de « Preview » soit la carte de connaissances locale autour de cette ressource dans le panneau « Knowledge Map Visualization ». Dans l'exemple de la figure précédente nous nous sommes connectés en tant que Philippe Lemoisson et avons cherché Guillaume Surroca. Il est possible de reproduire la manipulation sur http://viewpoints.cirad.fr/ViewpointsWebApp-CIRIRMM.

En ouvrant le panneau « Knowledge Map » on peut observer la carte de connaissances locale autour de l'agent Guillaume Surroca. La Figure 41 montre la Knowledge Map autour de Guillaume Surroca (GS). Un document d – « Diffusion de préférences par confrontation de points de vue : vers une simulation de la Sérendipité »[88] – relie GS à Clément Jonquet (CJ). La Figure 42 montre un plus court chemin affiché dans VWA. Il s'agit du plus court chemin entre l'agent Guillaume Surroca et le descripteur « semantic similarity » par l'intermédiaire de l'article « Subjective and generic distance in ViewpointS : an experiment on WordNet » [94]. La Figure 46 montre un exemple de recherche sur Clément Jonquet menée avec les trois méthodes de calcul de voisinage sémantique que VWA propose. En sélectionnant deux ressources dans les résultats ou l'historique l'utilisateur peut dessiner le plus court chemin liant ces deux ressources.

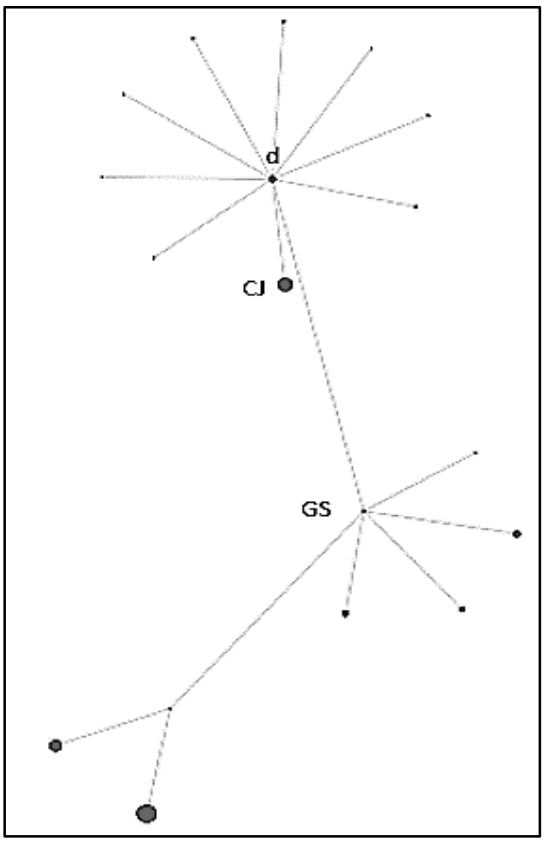


Figure 41 Illustration de la fonctionnalité Knowledge Map locale

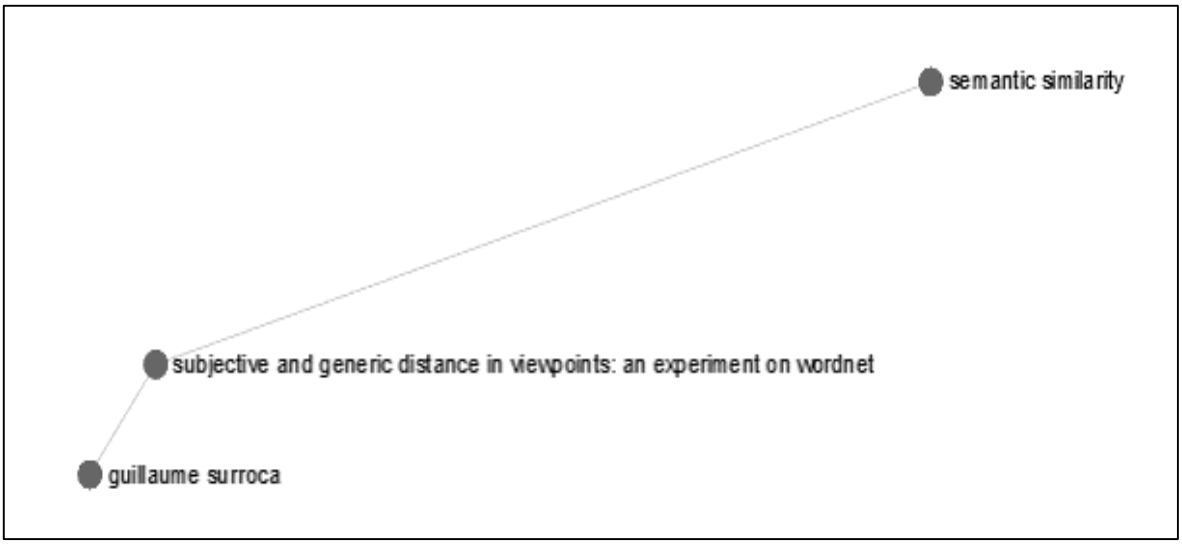


Figure 42 Illustration d'un plus court cheminaffiché dans VWA

Le choix de la Perspective influent à la fois sur les résultats de recherche mais aussi les visualisations de la carte de connaissances. En effet la création de la Knowledge Map dépend d'une interprétation spécifique des viewpoints. Le plus court chemin peut donc changer radicalement selon les types de connexion qu'on privilégie dans le paramétrage de la Perspective. Ce paramétrage de VWA se fait dans un panneau de configuration accessible à partir du menu de VWA représenté par un petit engrenage (Figure 43 et Figure 44) qui permet d'associer à chaque type de viewpoint un poids et de donner le rayon du voisinage et d'autres paramètres en fonction de la méthode. L'utilisateur y configure les méthodes de calcul de voisinage sémantique en donnant le rayon de voisinage (« Neighbourhood radius ») et, pour certaines méthodes comme MFN, d'autres paramètres comme le « parallelization factor ». Dans le jeu de donnéess de cet exemple, 5 types de viewpoints existent. On donne à chacun d'eux une importance relative.

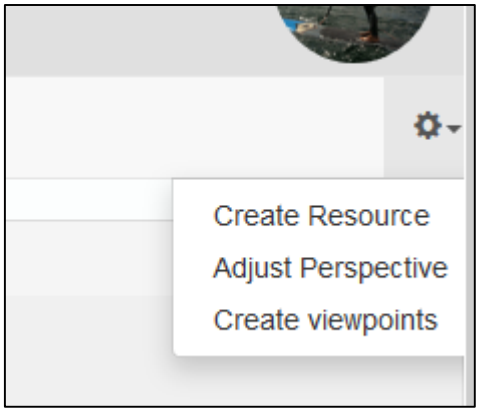


Figure 43 Menu de VWA

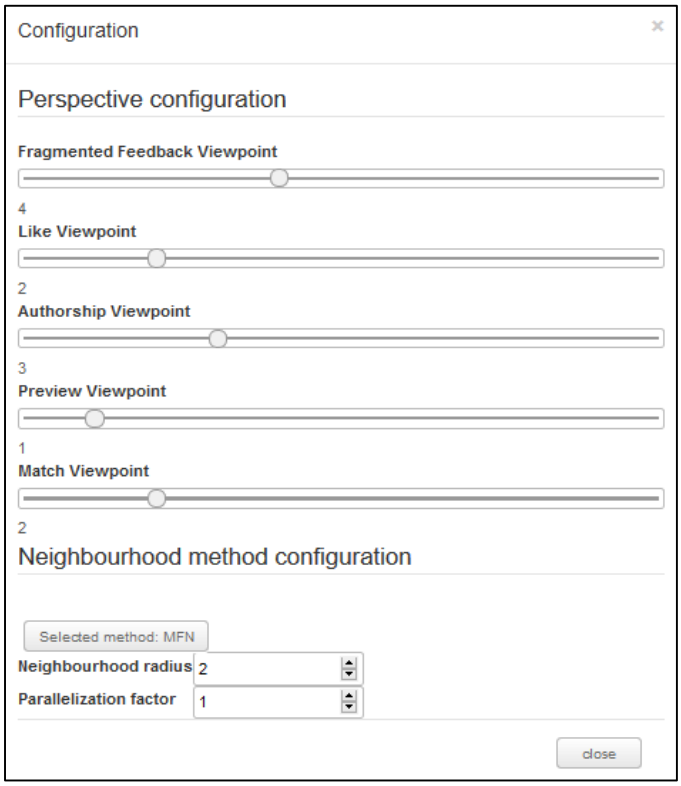


Figure 44 Configuration de VWA

Comme il est possible de l'observer, le menu de VWA peut servir également à saisir « manuellement » des ressources et des viewpoints. Il suffit de cliquer sur « Create Resource » ou « Create viewpoints et des menus s'ouvrent pour demander l'information nécessaire (Figure 45).

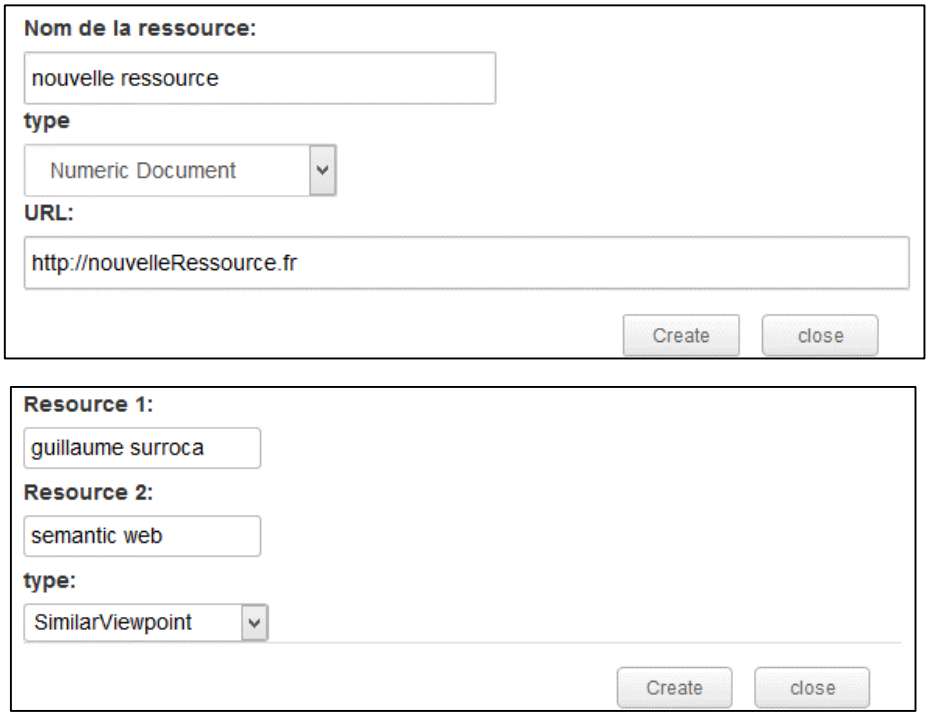


Figure 45 Création de ressources et de viewpoints

Créer une ressource nécessite un nom et une url s'il s'agit d'une Numeric Document comme le montre le haut de la figure. Le dialogue de création (en bas de la figure) demande les noms des ressources à connecter par un viewpoint et le type de ce viewpoint. Il est à noter que certains types de viewpoints comme le viewpoint de type Similar sont restreints à certains types de ressources connectables. La création « manuelle » de viewpoints est la façon la moins immédiate pour enrichir le graphe de connaissances. En effet avec deux boutons « Like/Tighten » et « Dislike/Loosen » nous offrons 4 possibilités de feedbacks différents. Si une seule ressource est sélectionnée dans les résultats de recherche alors « Like/Tighten » est un Like vous rapprochant de la ressource sélectionnée et « Dislike/Loosen » est un bouton Dislike au fonctionnement inverse. Si deux ressources sont sélectionnées, il est possible de les rapprocher (« Tighten ») en renforçant le plus court chemin entre ces deux ressources ou de les éloigner l'un de l'autre (« Loosen ») en affaiblissant les synapses du plus court chemin.

Observons par exemple sur la Figure 46 l'impact du choix de la méthode de voisinage sémantique sur les résultats. On remarque que les résultats obtenus par SPN et MFN sont similaires. Ils partagent en effet en partie une approche similaire puisqu'ils utilisent tous deux une propagation des plus courts chemins à partir d'un nœud central. MPN se base sur l'ensemble de tous les chemins de longueur bornée. Comme beaucoup de chemin sont pris pour évaluer la distance entre deux ressources – chaque chemin diminuant cette distance – les distances à la ressource centrale des ressources du voisinage est nettement inférieure. D'autant plus que, se basant sur de multiples chemins, MPN fait remonter le réseau des auteurs qui ont eu de multiples collaborations avec Clément et qui augmente les chemins parallèles passant par des articles entre l'agent Clément Jonquet et ses collègues.

SPN

Ra...	Type	Label	Distance
0	HumanAgent	clement jonquet	0.0
1	NumericDocument	biomedical terminology extraction	0.5
2	NumericDocument	biotex	0.5
3	NumericDocument	combining c-value and keyword extraction methods for biomedical terms extraction	0.5
4	NumericDocument	reusing the ncbo bioportal technology for agronomy to build agroportal	0.5
5	NumericDocument	les agents comme des interpreteurs scheme	1.0

MPN

Ra...	Type	Label	Distance
0	HumanAgent	sandra bringay	0.039525...
1	HumanAgent	mathieu roche	0.052631...
2	HumanAgent	maguelonne teisseire	0.052631...
3	HumanAgent	pascal poncelet	0.053763...
4	NumericDocument	discovering novelty in gene data	0.0625
5	HumanAgent	juan antonio lossio-ventura	0.08
6	NumericDocument	towards an on-line analysis of tweets processing	0.1
7	NumericDocument	discovering novelty in sequential patterns	0.1
8	NumericDocument	biotex	0.125

MFN

Ra...	Type	Label	Distance
0	NumericDocument	biomedical terminology extraction	0.5
1	NumericDocument	biotex	0.5
2	NumericDocument	combining c-value and keyword extraction methods for biomedical terms extraction	0.5
3	NumericDocument	reusing the ncbo bioportal technology for agronomy to build agroportal	0.5
4	NumericDocument	les agents comme des interpreteurs scheme	1.0
5	NumericDocument	actes de la 9eme conference des technologies de l'information et de la communication pour l...	1.0

Figure 46 Illustration de l'impact de la méthode de voisinage sur les résultats

5.4 Architecture

Faire passer le prototype à un nouveau stade ouvert à l'usage et évaluation sur le Web a demandé un passage en pré-production et l'utilisation de technologies très utilisées dans l'industrie. Cela nécessite aussi l'utilisation d'une architecture logicielle adaptée : l'architecture 3-tiers.

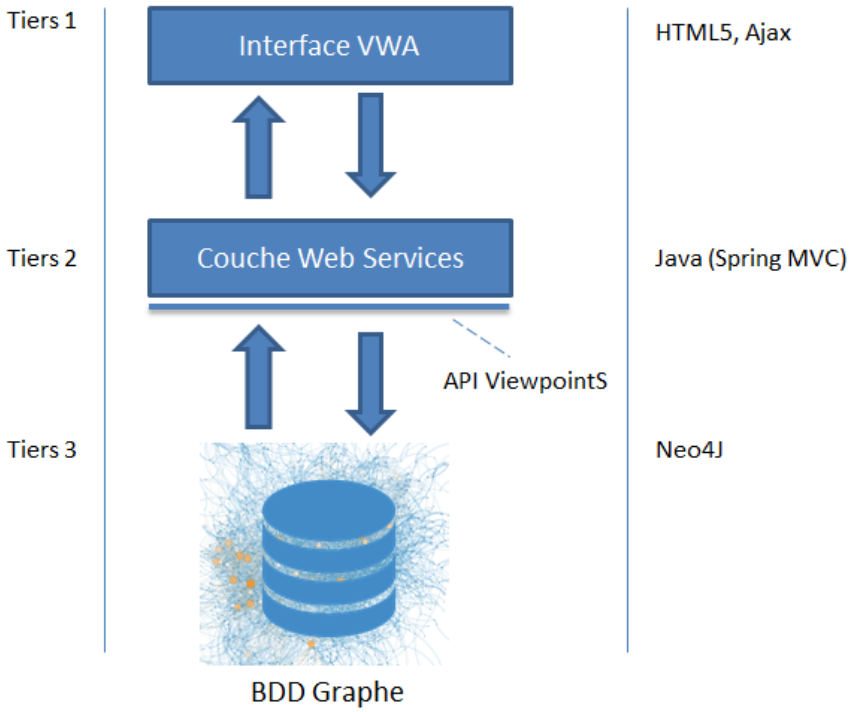


Figure 47 Illustration de la pile 3-tiers.

Comme l'illustre la Figure 47, cette architecture est l'empilement de trois étages :

- L'interface VWA est une interface Web dynamique classique construite en HTML5/Ajax[63]. Généralement il s'agit de la couche de présentation des données et du dialogue avec l'utilisateur. Quand une interaction se produit à partir de l'interface (ex. : la recherche d'un voisinage sémantique) une requête est envoyée à la couche de services Web.
- La couche Web Services comprend toute la logique de ViewpointS. C'est cette couche qui contient par exemple les méthodes de calcul de voisinage sémantique. Elle rend disponible à une ou plusieurs interfaces clientes via des requêtes HTTP toutes les méthodes de l'API ViewpointS qui était jusqu'alors uniquement disponible en local. L'offre de technologies accomplissant ce rôles est très diversifiée et utilise des langages très diversifiés respectant les contraintes REST[64] (ex. : Node.js[65] en Javascript, Struts[66] en Java ou Ruby on Rails[67] en Ruby).

[63] Ajax, pour Asyncronous Javascript And Xml, est une technologie qui a participé à l'essort du Web 2.0 en rendant les pages dynamiques car cette technologie permet de dépasser le chargement synchrone des pages et permet de charger et de modifier uniquement certaines parties d'une page. Telles VWA, les applications Ajax ne sont plus constituées que d'une page Web entièrement dynamique qui n'est chargé qu'une fois.
[64] https://fr.wikipedia.org/wiki/Representational_state_transfer
[65] https://nodejs.org/en/
[66] https://struts.apache.org/
[67] http://rubyonrails.org/

Nous avons choisi Spring MVC[68] car nous préférons limiter au maximum le nombre de langages que nous utilisons et il existe beaucoup de ressource documentant les usages de cette technologie. Ces technologies sont la clef de voute de la construction d'une application en architecture 3-tiers. Une fois l'application lancée sur un serveur d'application Apache Tomcat[69], le ou les graphe(s) de connaissances y sont chargés en mémoire vive. Pour que les modifications de ces graphes de connaissances persistent même après l'arrêt du serveur d'application des transactions vers une base de données graphe transmettent chacune de ces modifications.

- Le Tiers 3. C'est ici qu'une base de données embarquée Neo4J[70] stocke les modifications qui ont été apportées aux graphes de connaissances. Neo4J se charge de gérer un système de fichier qui contient les graphes de connaissances.

5.5 API ViewpointS

L'API ViewpointS – implémentation de l'approche ViewpointS dans une API Java – est contenue dans plusieurs librairies Java. Nous allons expliquer en détails comment a été structurée cette API et comment y accéder. Nous expliquons aussi comment ces structures ont été choisies pour optimiser les opérations les plus employées sur le graphe de connaissances ou occuper une taille optimale en mémoire par rapport à la nature des graphes qu'on a rencontrés lors de nos expérimentations (Chapitre 4).

5.5.1 Architecture

5.5.1.1 Vue d'ensemble de l'API

L'API est constituée d'un noyau contenant toutes les structures de données et méthodes essentielles à la construction/modification d'un graphe de connaissances ainsi que les méthodes les plus basiques l'exploitant telles que le calcul de voisinage ou de distance sémantique. Autour de ce noyau gravitent des modules qui l'enrichissent de nouvelles fonctionnalités.

5.5.1.2 Noyau ViewpointS

Dans le noyau il y a d'abord les structures de données contenant les ressources et viewpoints d'un graphe de connaissances que nous abordons en premier. La deuxième partie du noyau est dédié au mécanisme qui permet de passer du graphe de connaissances à la carte de connaissances. Ce mécanisme de perspective est essentiel pour les méthodes d'exploitation du graphe de connaissances dont nous expliquons en dernière partie la structure.

5.5.1.2.1 Graphe de connaissances

Le graphe de connaissances est un graphe biparti, c'est-à-dire un graphe composé de deux classes de nœuds tels que le représente le diagramme UML de la Figure 48. Le Knowledge Graph est la classe centrale contenant les viewpoints et ressources. Il contient plusieurs nœuds (Node) qui sont soit des ressources soit des viewpoints. Le Knowledge Provider est la classe généralisant tous les types

[68] http://spring.io/
[69] https://tomcat.apache.org/
[70] http://neo4j.com

d'agents producteurs ou consommateurs de viewpoints. Il est lui-même une ressource capable d'être connectée au graphe de connaissances aussi en tant qu'objet de recherche.

Il existe plusieurs façons de représenter un graphe dans une structure de données : les listes d'adjacence, les matrices d'adjacence ou les matrices d'incidence. Chacune de ces structures est optimisée pour une opération en particulier. Par exemple : récupérer la liste des voisins d'un nœud ou la liste des arrêtes entre deux nœuds. Les algorithmes que nous avons conçus explorent le graphe de voisin en voisin. Ainsi l'opération que nous utilisons le plus est celle qui obtient la liste des voisins d'un nœud.

Nous avons donc opté pour les listes d'adjacence car elles permettent de faire cette opération en temps constant. Les listes d'adjacences forment en outre une structure de données particulièrement adaptée aux graphes creux. En effet l'occupation mémoire d'une telle structure de donnée est directement proportionnelle au nombre d'arrêtes/arcs alors que la taille des matrices est proportionnelle au carré de ce nombre. On remarque que les graphes du Web, les « graphes de terrain » sont très souvent des graphes creux.

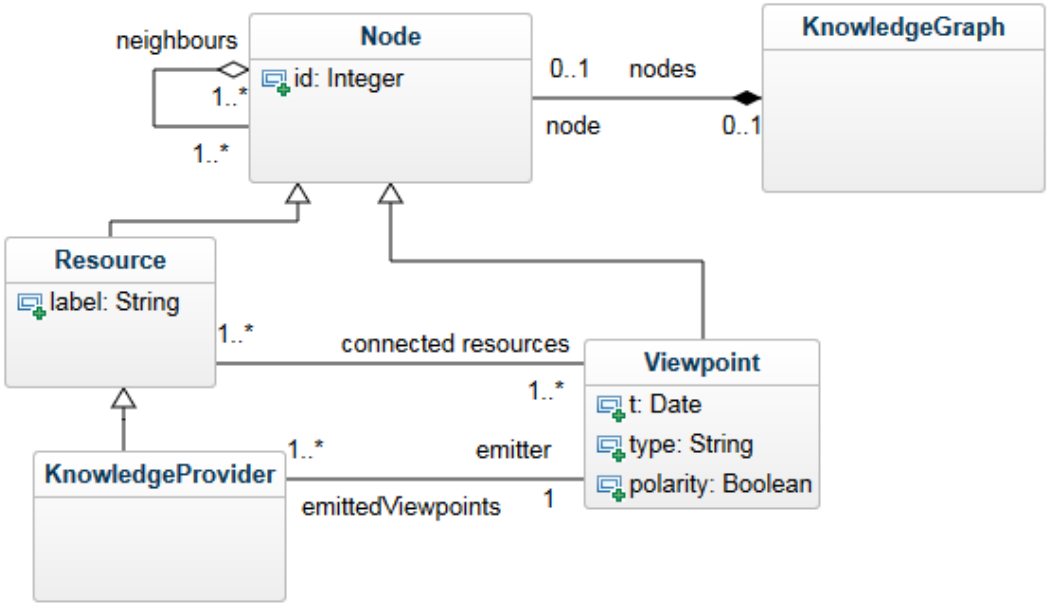


Figure 48 Diagramme UML du graphe de connaissances.

Les classes Ressource et Viewpoint sont étendues pour donner l'ensemble de classes de viewpoints et ressources présentées dans ce chapitre. Ces types sont extensibles à volonté pour adapter un graphe de connaissances à n'importe quel domaine de connaissances.

5.5.1.2.2 Mécanisme de Perspective

Chaque méthode exploitant le graphe de connaissances doit interpréter la subjectivité des viewpoints. C'est pourquoi ces méthodes nécessitent chacune un observateur, c'est-à-dire un agent qui possède une Perspective (Figure 49) que la méthode utilisera pour interpréter les viewpoints et les agréger sous forme de synapses.

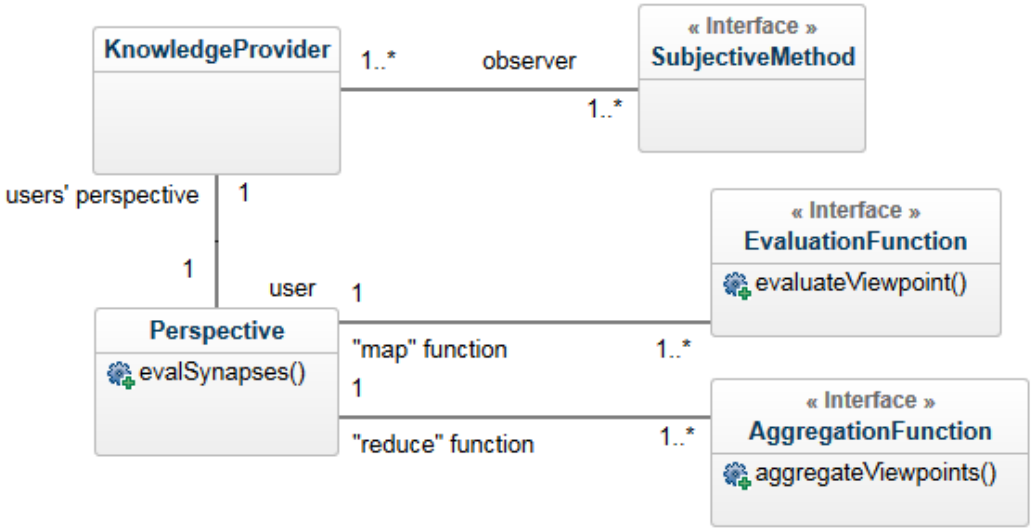


Figure 49 Diagramme UML du mécanisme de Perspective

La Perspective contient deux interfaces EvaluationFonction et AggregationFunction que l'utilisateur doit étendre si besoin. Deux implémentations par défaut de ces classes - DefaultEvaluationFunction et DefaultAggregationunction – sont les implémentations les plus simples de ces interfaces. Là encore, l'optique est de permettre une extensibilité maximale. Ainsi, l'utilisateur peut construire lui-même sa propre fonction d'évaluation qui évalue par exemple les viewpoints en fonction de son type et de son agent émetteur.

5.5.1.2.3 Méthodes d'exploitation

Toutes les méthodes exploitant le graphe de connaissances sont une implémentation de SubjectiveMethod. Nous les avons catégorisées comme le montre la Figure 50. Là encore toute la place est laissée à de futurs ajouts de méthodes de calcul de voisinage ou de distance sémantique. On voit que chaque méthode de voisinage (SPN, MPN, MFN) donnent lieu chacune à une méthode de distance sémantique (SPD, MDP, MFD).

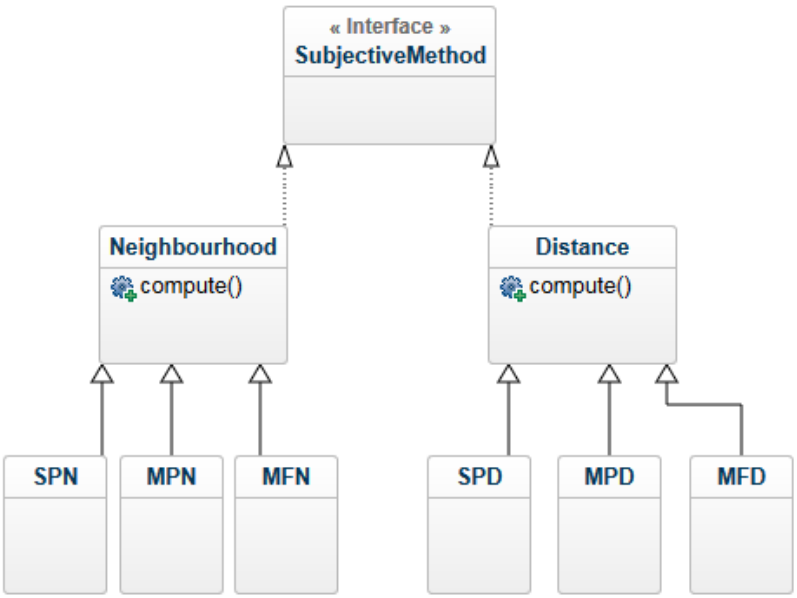


Figure 50 Diagramme UML de la hiérarchie des méthodes subjectives d'exploitation

5.5.2 Module d'import/export/indexation

Ce module gère tout ce qui est échange de données :

- La persistance du graphe de connaissances grâce à Neo4J,
- L'indexation de données dans un graphe de connaissances. Les formats de données gérés sont : XML, RDF, CSV, Json ;
- L'export d'un graphe de connaissance en format ViewpointS XML.

Le module contient des classes à étendre permettant à l'utilisateur d'écrire ses propres méthodes d'indexation de données ViewpointS qui enrichissent un graphe de connaissances. Le diagramme UML de la Figure 51 montre cet ensemble de classes à étendre. Par exemple dans le cas d'étude se basant sur les données bibliographiques de la plateforme HAL nous avons étendu XMLImportModule afin d'indexer son contenu.

La connaissance du graphe de connaissances est rendue persistante grâce à une base de données NoSQL embarquée Neo4J. Chaque modification sur le graphe de connaissances est répercutée dans un système de fichier qui est géré en partie grâce à Lucene[71]. Cela permet de profiter également des fonctionnalités de l'indexation d'un moteur de recherche classique permettant par exemple de retrouver rapidement les ressources ayant un nom donné. Neo4j est une technologie de stockage de graphe qui offre un ensemble de fonctionnalités comme un langage de requête spécialisé – Cypher[72] – semblable à SQL. Cela permet de requêter un graphe de connaissances d'une manière similaire à une base de données relationnelles. En effet, la structure des requêtes de type « SELECT * WHERE » nous rappelle celle des requêtes SQL.

Comme le montre l'expérimentation sur l'alignement d'ontologies (4.6), nous sommes en mesure d'intégrer dans un graphe de connaissances en partie le contenu des bases de connaissances du Web Sémantique. De cet import du Web Sémantique nous ne gardons que le minimum nécessaire et suffisant pour exploiter avec les méthodes que nous avons développées et de la façon la plus optimale possible ce que nous prenons. La valeur ajoutée des processus que nous proposons comme la création de nouvelles connexions par exemple par feedback au fur et à mesure de l'exploration par l'utilisateur ou par le mécanisme de suggestion de traduction peut ensuite être exportée et rendu utilisable dans le Web Sémantique. Pour cela nous proposons d'exporter le graphe de connaissances au format RDF/XML. Un viewpoint (a, {r_1, r_2}, {isAuthor,-}, 2008-09-01) est exporté de la sorte :

Exemple d'export d'un viewpoint dans le format RDF/XML

```
<rdf:Description rdf:about="http://www.viewpoints.org/viewpoints/5823">
   <foaf:made rdf:resource="http://www.viewpoints.org/resources/a" />
   <poder:connection rdf:resource=" http://www.viewpoints.org/resources/r1" />
   <poder:connection rdf:resource=" http://www.viewpoints.org/resources/r2" />
   <dc:date>2008-09-01</dc:date>
   <marl:hasPolarity rdf:resource="http://purl.org/marl/ns#Negative" />
   <rdf:type rdf:resource="http://lsdis.cs.uga.edu/projects/semdis/opus#author" />
</rdf:Description>
```

[71] https://lucene.apache.org/core/
[72] https://neo4j.com/developer/cypher-query-language/

Nous réutilisons au maximum les vocabulaires du Web Sémantique tels que FOAF, poder, opus ou Marl. Afin de rendre à l'export le graphe de connaissances en RDF le plus connecté avec le Web Sémantique existant. Qui plus est nous essayons tant que faire se peut de réutiliser d'associer les types de viewpoints à des types de relations du Web Sémantique. Par exemple nous associons au type « isAuthor » la relation « author » du vocabulaire Opus.

Un graphe de connaissance peut également être exporté au format ViewpointS XML permettant l'échange de graphes de connaissances.

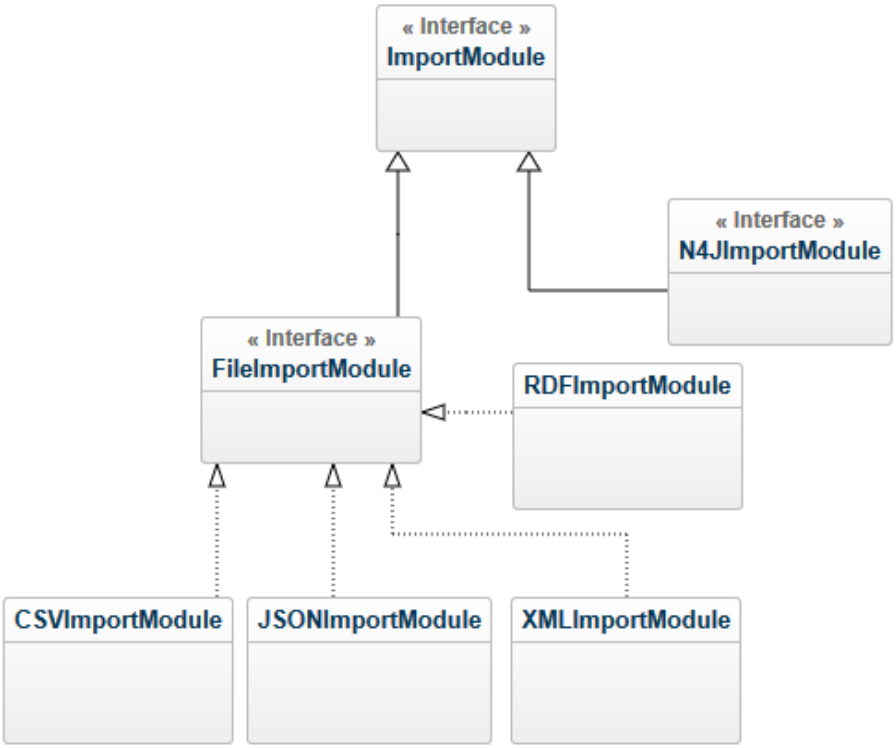


Figure 51 Diagramme UML du module d'indexation/import/export

5.5.3 Accessibilité

5.5.3.1 Serveur de versionnement

ViewpointS étant l'un des projets appartenant au projet SIFR il est hébergé sur GitHub dans le groupe SIFR : https://github.com/sifrproject. L'API est un projet Maven[73] qui peut être importé et enrichi dans n'importe quel IDE : Netbeans[74], IntelliJ[75] ou Eclipse[76].

Maven gère automatiquement les nombreuses sous-dépendances du projet allant des dépendances essentielles comme Spring MVC, en passant par diverses librairies gérant les différents formats de données d'échange (xml, json, csv ou rdf) jusqu'aux librairies offrant les structures de données performantes au cœur du graphe de connaissance. Il s'agit pour nous d'une structure ouverte à l'évolution ; les modules peuvent être considérés comme des plugins du noyau. Tous les choix archi-

[73] https://fr.wikipedia.org/wiki/Apache_Maven
[74] https://netbeans.org
[75] https://www.jetbrains.com/idea/
[76] https://eclipse.org/

tecturaux dont nous parlons dans les sections suivantes sont motivés par la volonté d'une API très flexible qui puisse être étendue facilement selon les besoin du futur utilisateur. La Figure 52 montre les deux dépendances typiques d'une application Viewpoints telle que celle que nous présentons au Chapitre 5. Toutefois la plupart des expérimentations (Chapitre 4) que nous avons faites ne nécessitait que le noyau.

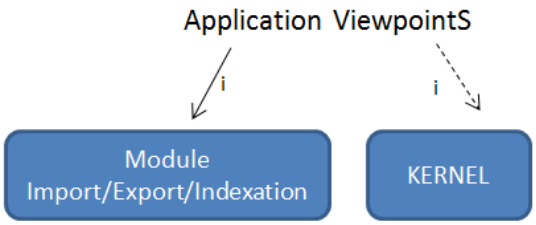


Figure 52 Schéma de dépendance d'une application ViewpointS classique.

5.5.3.2 Documentation

Une documentation technique JavaDoc est disponible à l'URL suivante : http://viewpoints.cirad.fr/api-documentation.

5.6 Cas d'utilisation

Grâce à VWA nous souhaitons donner l'occasion au lecteur d'examiner chacun des jeux de données à partir desquels nous avons construit les expérimentations du Chapitre 4. Ainsi l'utilisateur pourra choisir d'explorer 3 graphes de connaissances :

- Le graphe de connaissances HAL CIRAD-LIRMM qui contient les données bibliographiques des deux communautés de chercheurs. Il sera possible dans cet exemple d'examiner le lien interdisciplinaire qui relie ces deux laboratoires d'agronomie et d'informatique.
- Le graphe de connaissances Movielens. L'utilisateur pourra en faire les mêmes usages qu'avec le système de recommandation de films que nous avons présenté en 4.4.2.
- Le graphe de connaissance Biomédical constitué de l'intégration de plusieurs jeux de données biomédicaux. Ce cas d'utilisation se base sur un graphe de connaissances de très grande taille (de l'ordre du million de viewpoints) et constitue un défi de passage à l'échelle pour VWA.

Le lecteur peut ainsi explorer chacun de ces jeux de données, voir par lui-même leurs différentes topologies et reproduire certaines des expérimentations du Chapitre 4.

5.7 Pistes d'amélioration

La visualisation graphique est, nous pensons, la meilleure façon d'appréhender l'approche ViewpointS. C'est pourquoi l'une des principales pistes d'amélioration du prototype VWA est d'ordre ergonomique : nous souhaitons étendre ce genre de visualisation à un maximum des usages possibles de VWA. D'ailleurs, afficher un graphe contenant plusieurs milliers de nœuds de façon lisible est déjà un challenge en soit. Mais nous envisageons la future navigation dans VWA dans un graphe dans lequel l'utilisateur se déplace de proche en proche et découvre progressivement les voisinages sémantiques successifs. La navigation sur la Toile qui se faisait tantôt sous forme de listes de résul-

tats de recherche pourrait ainsi se faire sur la base de la visualisation d'une Toile sémantique en perpétuelle évolution au grès de la création des points de vue.

La perspective principale de VWA reste une évaluation par l'usage. Il est envisagé d'employer ViewpointS dans des scénarios de co-constructions de connaissances territoriales liés aux activités du CIRAD. Il est donc envisagé d'employé VWA comme moteur de recherche sur un graphe de connaissance représentant les acteurs d'un territoire et de représenter les connaissances de ces acteurs. Pour cela nous pensons intégrer les éléments de la modélisation DPSIR[77] sous forme de types de ressources ou viewpoints spéciaux. Il est aussi envisagé d'expérimenter VWA comme moteur de recherche bibliographique au sein de communautés de test comme l'UMR TETIS[78].

[77] Drivers Pressure, States, Impact, Resources. Il s'agit d'un modèle de représentation de la dynamique territoriale https://en.wikipedia.org/wiki/DPSIR

[78] https://www.tetis.teledetection.fr/index.php/fr/

Chapitre 6. Conclusion

En Introduction nous nous posions la question suivante : Quel est l'impact de la préservation de la subjectivité de l'information et de son exploitation aussi subjective en termes de recherche d'information, de capacité de recommandation et d'apprentissage collectif ? Au fil des expérimentations nous avons pu juger de l'apport de la double subjectivité de l'approche ViewpointS. Nous présentons dans la section qui suit un résumé de nos résultats permettant d'apporter des éléments de réponse par rapport à cette problématique. Nous résumons pour commencer toutes les pistes que nous avons entre-ouvertes pendant nos recherches mais que nous n'avons pas eu le temps d'exploiter ainsi que les perspectives qui se dessinent maintenant que nous disposons d'un prototype exploitable sur le Web. Nous proposons ensuite une prise de recul sur nos résultats expérimentaux.

6.1 Résultats obtenus

6.1.1 Subjectivité de la Perspective

La mesure de distance sémantique est centrale dans l'approche ViewpointS. Nous avons vu la généricité des méthodes de calcul de distances sémantiques qui s'appliquent sur un graphe de connaissance dans l'approche ViewpointS. Toutefois il est pratique de pouvoir rendre une mesure générique plus spécifique pour tenir en compte des particularismes des différents jeux de données du Web. C'est le rôle qu'à tenu la Perspective. En changeant la façon dont on regarde le graphe de connaissances – autrement dit, la façon dont on interprète les viewpoints – nous avons indirectement paramétré les méthodes viewpoints pour les adapter au cas d'utilisation. Ce mécanisme de Perspective, qui est le second degré de subjectivité de ViewpointS, a tenu le même rôle dans l'évaluation sur la recommandation de films. En choisissant une méthode appropriée pour le calcul de voisinage sémantique et en l'accompagnant d'une perspective que nous avons calibrée en fonction du jeu de données nous nous sommes rapprochés des résultats d'une plateforme de recommandation de film qui nous servait d'étalon de référence.

Nous avons aussi étudié comparativement quels étaient les résultats d'approches de recherche d'information classiques comme PageRank ou Vector Space Model. La Perspective a permis d'avoir des résultats similaires à ces méthodes avec les méthodes génériques de ViewpointS. Le mécanisme de Perspective ne peut laisser présager à celui qui enrichi le graphe avec des viewpoints le résultat que cela aura sur l'exploitation des autres du graphe de connaissances. Chacun regarde le graphe avec une Perspective qui lui est propre. Il est donc impossible de manipuler intentionnellement le graphe de connaissance pour détourner l'exploitation qu'en font les autres. Cela peut éviter les effets de toute intervention d'agents "de mauvaise foi" comme c'est le cas des "malware".

6.1.2 Subjectivité des viewpoints

L'une des finalités de ViewpointS est de jouer le rôle de formalisme de connaissance creuset apte à intégrer la plupart des types de connaissances du Web. Lors de notre exploration de divers jeu de donnéess du Web nous avons trouvé une multitude de relations sémantique pouvant connecter les

ressources du Web. Mais elles avaient toutes en commun de rapprocher ou distancier sémantiquement ces ressources. C'est le constat qui a mené à la création du viewpoint – une unité d'expression d'une sémantique personnelle – sur la proximité ou la distance de deux ressources du Web. Ainsi nous avons pu intégrer des jeux de données très divers.

Toutefois nous payions le prix de cette intégration par la perte d'une partie de la sémantique originale contenue dans ces jeux de données. Mais étant donné que l'objectif de ViewpointS n'est d'intégrer les connaissances du Web telles quelles et dans leur entièreté, mais d'intégrer jusque ce qu'il nous en faut pour pouvoir proposer de nouvelles associations; alors cette pauvreté sémantique peut suffire aux opérations de ViewpointS. Nous avons aussi pris le parti du tout subjectif.

De la même manière les interactions du Web Social représentent des subjectivités. Nous avons donc décidé d'interpréter – et c'est la prise de position forte de cette thèse – toute connaissances sur le Web comme subjectives. Il existe toutefois une vérité, certains diront, échappant à la subjectivité. Mais elle est le fruit d'un consensus, donc de l'acceptation par une majorité d'un point de vue qui était parfois originalement marginal. La vérité dans un graphe de connaissances ViewpointS évolue de manière fluide au fil de l'expression des viewpoints. En rendant la connaissance dans ViewpointS purement subjective nous avons pu intégrer les connaissances venant de jeux de données structurés sans perdre leur provenance et les interactions du Web Social sans perdre la subjectivité de ces échanges.

6.2 Pistes pour le passage à l'échelle et l'optimisation de perspectives

Lors de nos expérimentations sur des jeux de données réels nous nous sommes confronté au problème du passage à l'échelle des méthodes d'exploitation du graphe de connaissances. Hors nous avons laissé dans le choix de notre architecture du mécanisme de Perspective – principal composante du calcul dans ViewpointS – une porte ouverte à la parallélisation du traitement. En effet la Perspective est dotée de deux composantes d'interprétation des viewpoints et d'agrégation des résultats d'interprétation directement inspirés du modèle map-reduce qu'on trouve dans la programmation orientée flux. La parallélisation de nos algorithmes – jusqu'alors non-parallèle– permettrait un passage à l'échelle et le traitement de jeux de données plus conséquents. Mais cette direction vers le paradigme de la programmation par flux n'a pas pour seul but d'optimiser le temps de calcul de nos algorithmes. En effet, à l'heure actuelle nos méthodes s'exécutent sur un graphe de connaissances entièrement contenu dans la mémoire vive.

Même si nous avons fait le maximum pour diminuer au maximum l'occupation mémoire du graphe de connaissances nous allons être confrontés tôt ou tard au problème suivant : les graphes que nous utilisons ne peuvent plus être contenus entièrement dans la mémoire vive. Le challenge sera alors de ne charger en mémoire vive uniquement la partie du graphe de connaissance qui peut ou pourrait éventuellement être utilisée. Nous pourrons alors exploiter des graphes de taille bien supérieure à ceux que nous traitions lors de cette thèse. Pour ce faire, nous devrions aussi adapter les méthodes d'exploitation du graphe de connaissances pour qu'elles puissent fonctionner uniquement avec des données partielles. C'est là tout l'intérêt de l'approche map-reduce et de la programmation orientée flux. Pour résumer, un passage à l'échelle de l'approche ViewpointS correspondrait pour nous à un glissement vers le paradigme des flux.

Dans les expérimentations du Chapitre 4 nous paramétrions les Perspectives nous-mêmes selon notre connaissance des jeux de données. Cela dit nous sommes persuadés que bon nombre des

Perspectives que nous avons choisies ne sont pas optimales pour l'usage que nous souhaitions en faire. C'est pourquoi il fût question de trouver une méthode permettant de trouver la Perspective optimale pour répondre à un problème. Nous abordons en Annexes l'embryon d'une expérimentation sur laquelle nous avions obtenus quelques premiers résultats à ce sujet (Annexe 3). L'intérêt pour l'algorithmique génétique nous paraît cohérent dans ce travail qui trouve quelques inspirations biomimétiques principalement en rapport au cerveau et aux mécanismes de l'évolution.

Malgré tout, la principale problématique de passage à l'échelle dans nos perspectives est celle du passage à l'échelle des usages. C'est pourquoi nous envisageons de confronter VWA à un public restreint. Nous pensons pouvoir proposer VWA comme moteur de recherche/système de recommandation bibliographique à l'échelle d'une umr du CIRAD. D'autres perspectives liées aux activités de gestion du territoire du CIRAD ouvrent le domaine de la représentation des dynamiques territoriales à ViewpointS. L'unité de recherche TETIS, appartenant au CIRAD, pourrait constituer une première implantation de l'approche ViewpointS.

6.3 Viewpoint final de l'auteur

En prenant du recul, je pense que nous aurions peut-être plus du nous fondre technologiquement en développant ViewpointS dans les technologies du Web Sémantique comme le stockage triple store rdf. Nous avons choisi une approche plus expérimentale, moins enracinée dans les usages du Web Sémantique, afin avant tout d'optimiser le stockage des connaissances pour l'usage qu'on en faisait. Cela nous a permis d'accéder plus vite à l'expérimentation.

Les premiers bénéfices qui sont apparus à propos de la préservation de la provenance de la connaissance est le renforcement de la place de l'agent dans nos cartes de connaissances. L'agent catégorise, il établit des similarités ou se revendique comme auteurs de documents. Il est à la fois l'objet de plus de relations sémantiques le rapprochant à toutes sortes de ressources mais prend aussi une place nouvelle, en tant qu'émetteur de viewpoint, une place dans ce que nous avons appelé la Perspective des autres agents.

Je pense aussi que ce qui peut faire la force d'un modèle c'est surtout sa simplicité. A défaut de chercher la complétude dans les relations sémantiques et concepts que nous représentons, nous avons essayé de trouver une manière simple et générique de traiter des jeux de données très hétérogènes en les rapprochant par la relation sémantique la plus basique : la similarité. Le formalisme devait être également adaptable à n'importe quelle structuration ou nature de connaissances. Nous sommes loin de penser que les types de base de ressources et de viewpoints que nous avons inclus dans ViewpointS et son API suffisent à représenter les connaissances du Web. C'est pourquoi l'architecture ViewpointS a été conçue pour être facilement spécifiable.

Pour finir, nous espérons avoir contribué à l'état de l'art de la représentation des connaissances, au moins des démarches d'intégration des diverses types de connaissances – objectives et subjectives – avec les concepts de proximité dans les viewpoints et la double subjectivité décrite ci-dessus.

Annexe 1 Guide de départ rapide

Dans cette section nous donnons un guide de départ rapide pour le développeur ViewpointS. Ce guide amènera le lecteur de la création d'un graphe de connaissance à son exploitation. Nous commençons par instancier un graphe de connaissances.

```
KnowledgeGraph KG = new KnowledgeGraph() ;
```

Créons ensuite plusieurs ressources que nous ajoutons à KG.

```
LegalPerson guillaume = new LegalPerson('Guillaume Surroca') ;
NumericDocument these = new NumericDocument('Thèse ViewpointS');
Descriptor desc1 = new Descriptor('Knowledge Representation');
Descriptor desc2 = new Descriptor('Semantic Distances');
KG.addRessource(guillaume);
KG.addRessource(these);
KG.addRessource(desc1);
KG.addRessource(desc2);
```

Nous avons créé plusieurs ressources correspondant à l'agent qui me représente, à cette thèse et à deux des mots-clés qui la décrivent. Il est temps de créer les viewpoints qui lieront ces ressources entre elles.

```
KG.addViewpoint(new AuthorShipViewpoint(guillaume, guillaume, these,
   ViewpointPolarity.POSITIVE)) ;
KG.addViewpoint(new MatchViewpoint(guillaume, these, desc1,
   ViewpointPolarity.POSITIVE)) ;
KG.addViewpoint(new MatchViewpoint(guillaume, these, desc2,
   ViewpointPolarity.POSITIVE)) ;
```

Nous avons ajouté des viewpoints de type différent. Il est désormais possible de paramétrer la Perspective de Guillaume en donnant par exemple un poids plus fois aux viewpoints de type Authorship Viewpoint. Par défaut le poids d'un type de viewpoint est de 1.

```
Guillaume.getPerspective().setTypeWeight(AuthorShipViewpoint.class, 5) ;
```

Nous allons maintenant instancier une méthode de calcul de voisinage sémantique afin de calculer le voisinage de guillaume.

```
MultipleFlowsNeighbourhood mfn = new MultipleFlowsNeighbourhood() ;
mfn.setObserver(guillaume) ;
TreeMap<Ressource, Float> neighbourhood;
neighbourhood = mfn.computeAndSort(guillaume, 5);
System.out.println(neighbourhood);
```

Nous avons instancié MFN et donné l'agent guillaume comme son observateur. MFN utilisera donc la perspective de guillaume que nous avons tantôt paramétrée. Nous calculons le voisinage autour de guillaume composé des ressources à distance inférieure ou égale à 5. Le résultat de computeAndSort est une liste associative <Ressource, distance> que nous trions par distance croissante. Nous affichons ensuite le voisinage composé de guillaume, de la thèse ainsi que les deux descripteurs.

Annexe 2 Benchmark de passage à l'échelle

Dans cette Annexes nous proposons au lecteur un aperçu des performances des principales méthodes exploitant le graphe de connaissances. Nous évaluons ainsi la capacité de passage à l'échelle de l'approche ViewpointS. Nous testons ces méthodes sur plusieurs graphes de connaissances générés aléatoirement de tailles et densités différentes.

1. Méthode

Les graphes de connaissances générés sont de densités différentes. Par densité nous entendons le rapport nombre de viewpoints / nombre de ressources. Plus ce ratio est élevé plus les ressources du graphe sont interconnectées. Nous allons tester sur deux scénarios de densité forte et faible des montées en charges de SPN, MPN et MFN. Les paramètres de ce benchmark sont les suivants (Tableau 21) :

Tableau 21 Paramètres généraux du benchmark de passage à l'échelle.

Paramètre	Description
nRessources	Nombre de ressources
Densité	Rapport Nombre de viewpoints / nombre de ressources
nTests	Nombre de tests de vitesse pour chaque méthode

Nous allons exécuter pour un graphe de connaissances de nRessources nTest tests de vitesse pour chaque méthode. Nous ferons évoluer nRessources afin d'observer le passage à l'échelle. Chaque méthode possède des paramètres propres que nous allons fixer ainsi :

Tableau 22 Paramètres spécifiques aux algorithmes du benchmark de passage à l'échelle.

Méthode	Paramètre	Valeur
SPN	m	2
MPN	m	2
MFN	m	2
	p	1.5

Nous faisons évoluer la taille du graphe de connaissances d'un graphe de 1000 à 500k ressources. La densité faible correspond à un rapport nombre de viewpoints / nombre de ressources de 0.5 alors que la densité forte correspond à un ratio de 5 (pour chaque ressource créées 5 viewpoints seront créés). Pour chaque test nous choisissons aléatoirement nTest ressources sur lesquelles appliquer SPN, MPN et MFN qui seront paramétrés tels que le montre le Tableau 22.

2. Résultats

Comme le montre la Figure 53, a plupart des algorithmes que nous avons conçus et qui se basent sur une propagation dans le graphe de connaissances sont – d'après ce que montrent les figures suivantes - assez peu sensibles à la taille du graphe de connaissance. En effet la limitation de l'exploration du graphe de nos algorithmes pour obtenir un voisinage topologique permet à ces algorithmes de ne pas dépendre de la taille du graphe d'exécution pour leur temps d'exécution.

Cela dit, les résultats montrent aussi que les méthodes d'exploration de graphe sont beaucoup plus impactées par la densité (Figure 54). Une recherche sur un « nœud-hub » - un nœud de fort degré et entouré d'un maillage dense – sera donc beaucoup plus longue. L'ordre que nous obtenons parmi nos méthodes correspond à ce que nous nous attendions.

SPN fût développée comme une méthode rapide, capable de calculer un voisinage de diamètre maximal afin d'obtenir la distance entre deux ressources quelconques du graphe. C'est la méthode la plus rapide. MPN fût créée comme la méthode la plus complète afin de prendre en compte tous les chemins connectant deux ressources pour évaluer la distance qui les sépare. Il s'agissait d'une méthode complète mais beaucoup trop lente pour pouvoir systématiser son utilisation dans nos expérimentations en particulier sur des jeux de données de taille importante comme les jeux de données biomédicaux. La création de MFN correspondait à un besoin de compromis entre SPN et MPN.

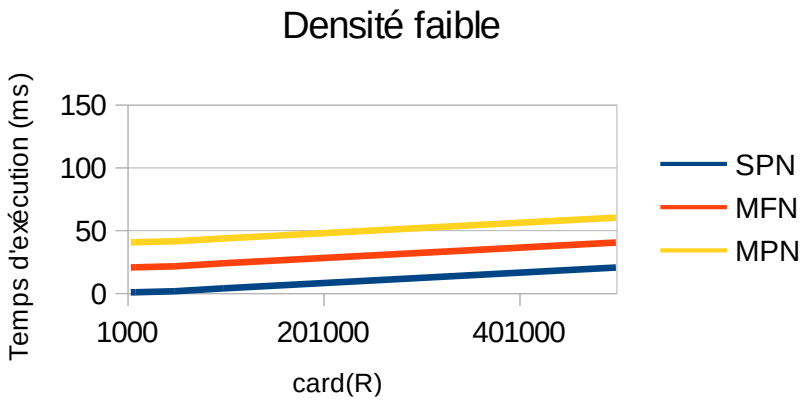


Figure 53 Temps d'exécutions moyens des méthodes ViewpointS pour un KG de densité faible.

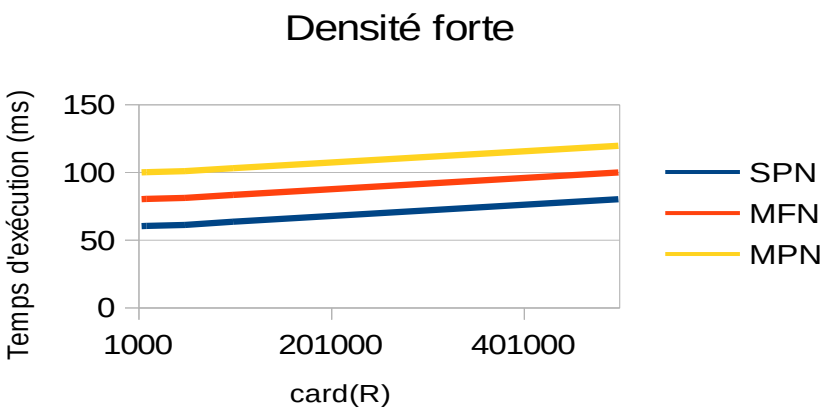


Figure 54 Temps d'exécutions moyens des méthodes ViewpointS pour un KG de densité faible.

3. Discussions

Nous en somme toutefois resté à l'étape de prototypage sur ces algorithmes. Ceux-ci ont été ont été optimisé algorithmiquement. Toutefois aucune amélioration technique n'a été appliquée à ces méthodes. Ce sont des algorithmes monothread qui ont le potentiel d'être parallélisés à l'avenir en utilisant le pattern Fork/Join[79] apparu avec Java7 ou d'autres API permettant de faire du parallélisme soit sur CPU soit sur GPU (GPGPU[80]).

[79] "Le problème que Fork/Join se propose de résoudre concerne le traitement de larges quantités de données, numériques ou non. Il arrive souvent que ces quantités peuvent être traitées paquet par paquet, chaque paquet pouvant être pris en compte de façon indépendante des autres. Le traitement de chaque paquet fournit un résultat partiel. Ces résultats sont ensuite fusionnés, d'une façon ou d'une autre, dans le résultat global du traitement. On constate que l'utilisation de ce framework permet de gagner bien plus qu'un facteur 4 sur l'exécution dans un unique thread : on gagne quasiment un facteur 6, à condition de ne pas fixer des paramètres trop délirants." Ref: http://blog.paumard.org/2011/07/05/java-7-fork-join/

[80] https://fr.wikipedia.org/wiki/General-purpose_processing_on_graphics_processing_units

Annexe 3 Ouverture sur l'optimisation de Perspective

1. Problématique soulevée

Dans les expérimentations du Chapitre 4 nous paramétrions les Perspectives nous-mêmes selon notre connaissance des jeux de données. Cela dit nous sommes persuadés que bon nombre des Perspectives que nous avons choisies ne sont pas optimales pour l'usage que nous souhaitions en faire. C'est pourquoi il fût question de trouver une méthode permettant de trouver la Perspective optimale pour répondre à un problème.

2. Introduction sur les algorithmes génétiques

Parmi tous les types d'algorithmes existants, certains ont la particularité de s'inspirer de l'évolution des espèces dans leur cadre naturel. Ce sont les algorithmes génétiques. Les espèces s'adaptent à leur cadre de vie qui peut évoluer, les individus de chaque espèce se reproduisent, créant ainsi de nouveaux individus, certains subissent des modifications de leur ADN, certains disparaissent. Un algorithme génétique va reproduire ce modèle d'évolution dans le but de trouver des solutions pour un problème donné. Il sera fait usage dans ce cours de termes empruntés au monde des biologistes et des généticiens et ceci afin de mieux représenter chacun des concepts abordés :

- Dans notre cas, une population sera un ensemble d'individus.
- Un individu sera une solution à un problème donné.
- Un gène sera une partie d'une solution, donc d'un individu.
- Une génération est une itération de notre algorithme.

Un algorithme génétique va faire évoluer une population dans le but d'en améliorer les individus. Et c'est donc, à chaque génération, un ensemble d'individus qui sera mis en avant et non un individu particulier. Nous obtiendrons donc un ensemble de solutions pour un problème et pas une solution unique. Les solutions trouvées seront généralement différentes, mais seront d'une qualité équivalente. Le déroulement d'un algorithme génétique peut être découpé en cinq parties :

1. La création de la population initiale
2. L'évaluation des individus
3. La sélection des individus
4. La création de nouveaux individus par croisement et mutations

5. Réitération du processus jusqu'à 2

Le patrimoine génétique d'une population est renouvelé à chaque génération par deux opérations : le croisement et la mutation. Le croisement prend

3. Fonctionnement de l'optimisation de Perspective

Un paramétrage de Perspective consiste à donner un poids à chaque type de viewpoints dans le graphe de connaissance. Dans le paradigme de la programmation génétique une Perspective peut alors être vue comme un individu dont le génome est cette table des poids associés aux types de viewpoints. Si nous reprenons les étapes du processus évolutionniste nous avons besoin d'une méthode pour générer la population initiale, d'une fonction d'évaluation des individus, une stratégie de sélection qui décide lesquels survivent jusqu'à la prochaine génération ainsi que d'une stratégie de croisement des génomes et de mutations pour créer la population d'une nouvelle génération.

a. Génération de la population initiale

Nous représentons chaque individu p par son génome grâce à une matrice associant chaque type de viewpoint du graphe de connaissance à traiter $t_1 \dots t_n$ à un poids entre 0 et 10.

$$genome(p) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ \dots \\ t_n \end{matrix} \begin{vmatrix} 0 \\ 9 \\ 4 \\ \dots \\ 7 \end{vmatrix}$$

Un individu de la population initiale est généré en remplissant cette matrice d'entiers naturels compris entre 0 et 10. Nous générons une population initiale de n_individus individus. Le croisement entre deux individus p_1 et p_2 se note p_1 x p_2. Une moitié de gênes des deux parents p_1 et p_2 se retrouvent aléatoirement dans le génome du nouvel individu. Par exemple :

$$genome(\mathrm{p1}) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} \begin{vmatrix} 0 \\ 9 \\ 4 \\ 2 \\ 7 \end{vmatrix}, genome(p2) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} \begin{vmatrix} 6 \\ 2 \\ 10 \\ 4 \\ 5 \end{vmatrix}, genome(p1 \times p2) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} \begin{vmatrix} 6 \\ 9 \\ 4 \\ 4 \\ 5 \end{vmatrix}$$

L'opération de mutation consiste à changer aléatoirement – toujours entre 0 et 10 – un des gènes d'un individu. Par exemple :

$$genome(\mathrm{p1}) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} \begin{vmatrix} 0 \\ 9 \\ 4 \\ 2 \\ 7 \end{vmatrix}, mutation(genome(p1)) = \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{matrix} \begin{vmatrix} 0 \\ 2 \\ 4 \\ 2 \\ 7 \end{vmatrix}$$

La mutation est une opération essentielle car le patrimoine génétique a tendance, au fil des générations, à converger vers un ensemble de gênes optimaux et les mutations permettent de renouveler ce patrimoine génétique pour tenter de dépasser l'optimum local atteint par le croisement.

b. Evaluation des individus

A chaque génération la fonction d'évaluation donne un score à chaque individu qui déterminera si il survivra jusqu'à la prochaine génération et donc si il propagera une partie de son génome par croi-

sement. Si nous souhaitions par exemple trouver des perspectives optimales pour le calcul de distances sémantiques sur WordNet alors nous utiliserions le score de précision que nous présentons dans le benchmark de distance sémantique du Chapitre 4. La population est ensuite triée par score décroissant.

c. Sélection

Il existe plusieurs stratégies de sélection : l'élitisme, la sélection par rang, la sélection par tournoi et la roulette. Chacune de ces méthodes a ses avantages.

Élitisme

Cette méthode de sélection permet de mettre en avant les meilleurs individus de la population. Ce sont donc les individus les plus prometteurs qui vont participer à l'amélioration de notre population. Cette méthode a l'avantage de permettre une convergence (plus) rapide des solutions, mais au détriment de la diversité des individus. On prend en effet le risque d'écarter des individus de piètre qualité, mais qui aurait pu apporter de quoi créer de très bonnes solutions dans les générations suivantes.

Sélection par tournoi

Le principe de la sélection par tournoi augmente les chances pour les individus de piètre qualité de participer à l'amélioration de la population. Ce principe est très rapide à implémenter. Un tournoi consiste en une rencontre entre plusieurs individus pris au hasard dans la population. Le vainqueur du tournoi est l'individu de meilleure qualité. On peut choisir de ne conserver que le vainqueur comme on peut choisir de conserver les 2 meilleurs individus ou les 3 meilleurs. On peut aussi faire participer un même individu à plusieurs tournois. La méthode par tournoi laisse beaucoup de liberté sur la façon d'organiser le tournoi.

Roulette

La sélection des individus par le système de la roulette s'inspire des roues de loterie. A chacun des individus de la population est associé un secteur d'une roue. L'angle du secteur étant proportionnel à la qualité de l'individu qu'il représente. On tourne la roue et on obtient un individu (**Erreur ! Source u renvoi introuvable.**). Les tirages des individus sont ainsi pondérés par leur qualité. Les meilleurs individus ont plus de chance d'être croisés et de participer à l'amélioration de notre population. Mais cette méthode laisse également une chance aux individus les moins bien noter de diffuser une partie de leur génome. En effet, certains des gênes de ces individus auraient pu s'exprimer très positivement dans d'autres génomes.

Sélection par rang

La sélection par rang est une variante du système de roulette. Il s'agit également d'implémenter une roulette, mais cette fois ci les secteurs de la roue ne sont plus proportionnels à la qualité des individus, mais à leur rang dans la population triée en fonction de la qualité des individus.

D'une manière plus parlante, il faut trier la population en fonction de la qualité des individus puis leur attribuer à chacun un rang. Les individus de moins bonne qualité obtiennent un rang faible (à partir de 1). Et ainsi en itérant sur chaque individu on finit par attribuer le rang N au meilleur individu (où N est la taille de la population). La suite de la méthode consiste uniquement en l'implémentation

d'une roulette basée sur les rangs des individus. L'angle de chaque secteur de la roue sera proportionnel au rang de l'individu qu'il représente.

d. Création d'une nouvelle population

Après la sélection, il faut régénérer la population en utilisant son patrimoine génétique. Le choix peut être fait soit de maintenir la population stable en créant autant d'individus de la génération précédentes qui n'ont pas passé la sélection soit en augmentant la taille de la population à chaque génération. Les individus survivants sont regroupés par paires aléatoires afin de créer un nouvel individu par croisement. Après le croisement chaque nouvel individu a un pourcentage de chance mutation_ratio de muter.

4. Utilisations

Hélas nous n'avons pas pu mener cette expérimentation suffisamment pour obtenir des résultats publiables. Nous avons expérimenté à petite échelle cette méthode d'optimisation de perspective dans le contexte du benchmark des distances sémantiques. C'est l'un des usages que nous pensons intéressant de cette fonction d'optimisation. Si on utilise une première méthode m_1 afin de calculer un grand nombre de distances que nous considérons ces distances comme un golden standard et donc la précision d'une des méthodes de ViewpointS m_v comme une mesure à optimiser. L'optimisation de la perspective pour la méthode m_v sert donc à réduire l'écart entre les résultats de m_v et de m_1 et à faire en quelque sorte « mimer » m_1 par m_v. On peut aussi envisager un usage sur le temps ou la mesure à optimiser est par exemple le nombre d'interactions de l'utilisateur. L'optimisation fait donc évoluer au fil du temps fait donc évoluer la Perspective de cet utilisateur afin qu'il interagisse au plus avec le graphe de connaissances.

Annexe 4 La société ViewpointS

Alors que nous terminions à trois sur le stage de recherche des débuts avec Philippe Lemoisson Stefano A. Cerri et moi-même, Clément Jonquet nous rejoignit dans le « noyau dur » de la petite équipe de R&D que nous étions. Nous avons collaborés avec plusieurs stagiaires et chercheurs qui ont nourri notre réflexion.

1. Encadrements de stage

L'équipe a encadré au total 5 stages. Parmi les 5 stagiaires donc nous disposions 3 ont été affecté au développement de VWA (Awa dia, Harish Sankar et Guilhaume Moraud).Ils permirent chacun de progresser jusqu'à la publication de VWA. Nous avons affecté un autre stagiaire (Luc Méric) au développement d'un algorithme de clusterisation pour ViewpointS. Le dernier (Alexandre Lerbet) a passé 2 semaines à développer un Module IO permettant d'intégrer des données biomédicales dans un KG spécial SIFR.

2. Chercheurs associés

En tant que co-encadrant de mon stage de M2 recherche Jacques Ferber a aussi contribuée à la façon de penser notre approche mais aussi surtout à la façon de la présenter. Matthieu Roche, alors en contact avec Philippe nous a rejoint et alimenté notre réflexion sur le traitement du langage naturel. François Bousquet finit par rejoindre le Comité de Suivi de thèse pour apporter un regard extérieur sur notre travail.

3. Publications de la thèse

P. Lemoisson, G. Surroca, et S. A. Cerri, « Viewpoints: An Alternative Approach toward Business Intelligence », in *eChallenges e-2013 Conference*, Dublin, 2013, p. 8. URL: http://bit.ly/2o8Lv1A

G. Surroca, P. Lemoisson, C. Jonquet, et S. A. Cerri, « Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique », *IC - 25èmes Journées francophones d'Ingénierie des Connaissances*. Clermont-Ferrand, France, p. 175-186, 13-mai-2014. URL: http://bit.ly/2nKowZq

G. Surroca, P. Lemoisson, C. Jonquet, et S. A. Cerri, « Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité », in *IC 2015*, Rennes, France, 2015, p. 12. URL: http://bit.ly/2oeTRkU

G. Surroca, P. Lemoisson, C. Jonquet, et S. A. Cerri, « Preference Dissemination by Sharing Viewpoints : Simulating Serendipity », *KEOD: Knowledge Engineering and Ontology Development*, vol. 7, n° 2., Lisbonne, Portugal 2015, p. 402-409. URL: http://bit.ly/2nfovsN

G. Surroca, P. Lemoisson, C. Jonquet, et S. A. Cerri, « Subjective and generic distance in ViewpointS », in *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS '16*, Nîmes, France, 2016, p. 1-6. URL: http://bit.ly/2nyP93f

P. Lemoisson, G. Surroca, C. Jonquet, et S. A. Cerri, « ViewpointS: When Social Ranking Meets the Semantic Web », *FLAIRS 17 Conference, California, USA*, 2017. URL: http://bit.ly/2owg5yh

Liste des figures

Liste des tables

Liste des algorithmes

Bibliographie

[1] P. Lemoisson, G. Surroca, and S. A. Cerri, “Viewpoints: An Alternative Approach toward Business Intelligence,” in *eChallenges e-2013 Conference*, 2013, p. 8.

[2] P. Lemoisson, “Construction collaborative de théorie : vers une machine abstraite conversationnelle,” *http://www.theses.fr*, 2006.

[3] F. Turner, L. Vannini, H. Le Crosnier, and D. Cardon, *Aux sources de l’utopie numérique de la contre-culture à la cyberculture, Stewart Brand, un homme d’influence*. C&F editions, 2012.

[4] M. McLuhan, *The social impact of cybernetics*. University of Notre Dame Press, 1966.

[5] B. A. Forouzan and B. A., *TCP/IP protocol suite*. McGraw-Hill, 2003.

[6] K. A. Clauson, H. H. Polen, M. N. K. Boulos, and J. H. Dzenowagis, “Scope, completeness, and accuracy of drug information in Wikipedia.,” *Ann. Pharmacother.*, vol. 42, no. 12, pp. 1814–21, Dec. 2008.

[7] T. K. Park, “The visibility of Wikipedia in scholarly publications,” *First Monday*, vol. 16, no. 8, Jul. 2011.

[8] T. R. Gruber, “A translation approach to portable ontologies,” *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.

[9] M. du Plessis, “The role of knowledge management in innovation,” *J. Knowl. Manag.*, vol. 11, no. 4, pp. 20–29, Jul. 2007.

[10] C. Jonquet and S. A. Cerri, “Les Agents comme des interpréteurs Scheme : Spécification dynamique par la communication,” in *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, 2004, vol. 2, pp. 779–788.

[11] R. Davis, H. Shrobe, and P. Szolovits, “What Is a Knowledge Representation?,” *AI Magazine*, vol. 14, no. 1. p. 17, 15-Mar-1993.

[12] J. T. Abbott, J. L. Austerweil, and T. L. Griffiths, “Human memory search as a random walk in a semantic network,” in *NIPS*, 2012, pp. 3050–3058.

[13] A. Hammache and R. Ahmed-Ouamer, “Système d’Inférence pour une Indexation de Documents Basée sur une Ontologie de Domaine.,” *INFORSID*, pp. 895–910, 2006.

[14] C. Jonquet, “A few contributions of the SIFR (Semantic Indexing of French biomedical Resources) project and how we reuse NCBO technology,” *Septième Atelier Rech. d’Information Semant. RISE*, p. 4, 2015.

[15] A. (Paul) Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, “Semantic concept-based query expansion and re-ranking for multimedia retrieval,” in *Proceedings of the 15th international conference on*

Multimedia - MULTIMEDIA '07, 2007, p. 991.

[16] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Inf. Process. Manag.*, vol. 43, no. 4, pp. 866–886, 2007.

[17] Z. Yu, Y. Nakamura, S. Jang, S. Kajita, and K. Mase, "Ontology-based semantic recommendation for context-aware e-learning," in *International Conference on Ubiquitous Intelligence and Computing*, 2007, pp. 898–907.

[18] B. Vesin, M. Ivanović, A. Klašnja-Milićević, and Z. Budimac, "Protus 2.0: Ontology-based semantic recommendation in programming tutoring system," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 12229–12246, Nov. 2012.

[19] S. Karapiperis and D. Apostolou, "Consensus building in collaborative ontology engineering processes," *J. Univers. Knowl. Manag.*, vol. 1, no. 3, pp. 199–216, 2006.

[20] C. Bizer, T. Health, and T. Berners-Lee, "Linked Data - The Story So Far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, vol. 5, no. 3, 2009, pp. 1–22.

[21] T. R. Gruber and G. R. Olsen, "An ontology for engineering mathematics," in *4th International Conference on Principles of Knowledge Representation and Reasoning, KR'04*, 1994, vol. 94, pp. 258–269.

[22] N. Guarino, "Formal ontology, conceptual analysis and knowledge representation," *Int. J. Hum. Comput. Stud.*, vol. 43, no. 5–6, pp. 625–640, Nov. 1995.

[23] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nat. Genet.*, vol. 25, no. 1, pp. 25–9, May 2000.

[24] R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G. F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, and G. McLaren, "Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature," *AoB Plants*, vol. 2010, 2010.

[25] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Commun. ACM*, vol. 54, no. 4, p. 86, Apr. 2011.

[26] A. J. Quinn and B. B. Bederson, "Human computation," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 2011, p. 1403.

[27] M. Casey, "Citizen mapping and charting: How crowdsourcing is helping to revolutionize mapping & charting," *Hydro 2009 Conf. Norfolk, Va*, 2009.

[28] J. Surowiecki, *The Wisdom of Crowds*, Anchor. 2005.

[29] M. Lafourcade, "Making people play for Lexical Acquisition with the JeuxDeMots prototype," in *SNLP'07: 7th International Symposium on Natural Language Processing*, 2007, p. 7.

[30] K. Aberer, P. Cudr, T. Catarci, M. Hacid, A. Illarramendi, M. Mecella, E. Mena, E. J. Neuhold, O. De, T. Risse, and M. Scannapieco, "Emergent Semantics Principles and Issues," in *Database Systems for Advanced Applications*, vol. 2, D. Lee, YoonJoon and Li, Jianzhong and Whang, Kyu-Young and Lee, Ed. Springer Berlin Heidelber, 2004, pp. 25–38.

[31] A. Bernstein, “The global brain semantic web - Interleaving human-machine knowledge and computation,” in *ISWC2012 Workshop on What will the Semantic Web Look Like 10 Years From Now?*, 2012, pp. 1–6.

[32] J. Caussanel, J.-P. Cahier, M. Zacklad, and J. Charlet, “Les Topic Maps sont-ils un bon candidat pour l’ingénierie du Web Sémantique ?,” in *13èmes journées francophones d’Ingénierie des Connaissances, IC’02*, 2002, p. 12.

[33] P. Lemoisson, E. Untersteller, S. A. Cerri, M. A. S. N. Nunes, A. Krief, and F. Paraguacu, “Interactive Construction of EnCOrE (Learning by Building and Using an Encyclopedia),” in *GLS’04: 1st Workshop on GRID Learning Services at ITS’04*, 2004, pp. 78–93.

[34] J. G. Breslin, A. Passant, and D. Vrandečić, “Social semantic Web,” in *Handbook of Semantic Web Technologies*, S. B. Heidelberg, Ed. 2011, pp. 467–506.

[35] T. Gruber, “Collective knowledge systems: Where the Social Web meets the Semantic Web,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 6, no. 1, pp. 4–13, Feb. 2008.

[36] A. Mikroyannidis, “Toward a social semantic web,” *Computer (Long. Beach. Calif).*, vol. 40, no. 11, 2007.

[37] P. Mika, “Ontologies are us: A unified model of social networks and semantics,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 5, no. 1, pp. 5–15, Mar. 2007.

[38] W. W. Web Consortium, “Resource Description Framework (RDF) : concepts and abstract syntax.” World Wide Web Consortium, 10-Feb-2004.

[39] R. Kannan, “Topic Map: An Ontology Framework for Information Retrieval,” Mar. 2010.

[40] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, 2003.

[41] M. Minsky, “A Framework for Representing Knowledge,” *Psychol. Comput. Vis.*, vol. 73, pp. 211–217, Jun. 1974.

[42] M. Schmidt-Schauß and G. Smolka, “Attributive concept descriptions with complements,” *Artif. Intell.*, vol. 48, no. 1, pp. 1–26, Feb. 1991.

[43] F. M. Donini, “Complexity of reasoning,” *Descr. Log. Handb.*, pp. 96–136, Jan. 2003.

[44] M. Zacklad, J. Cahier, A. Benel, Lh. Zaher, C. Lejeune, and C. Zhou, “Hypertopic : une métasémiotique et un protocole pour le Web socio-sémantique,” *Actes des eme journees Francoph. dingenierie des connaissances*, pp. 217–228, 2007.

[45] J.-P. Cahier and L. Zaher, “Cartodd.org: un Web2.0 basé sur Hypertopic pour les initiatives de développement durable,” *Atelier spécial “ IC 2.0 ” Assoc. à la Conférence " Ingénierie des Connaissances*, 2008.

[46] F. Limpens and F. Gandon, “Un cycle de vie complet pour l ’ enrichissement sémantique des folksonomies,” in *Extraction Gestion de Connaissance EGC 2011*, 2011, pp. 389–400.

[47] R. K. Merton and E. Barber, *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*, vol. 2006. 2006.

[48] E. Hodges, *The Three Princes of Serendip*, Atheneum. 1964.

[49] J. Perriault, “Effet diligence, effet serendip et autres défis pour les sciences de l’information,” in *Pratiques collectives distribuées sur Internet*, 2000.

[50] G. A. Fine and J. G. Deegan, “Three principles of Serendip: insight, chance, and discovery in qualitative research,” *Int. J. Qual. Stud. Educ.*, vol. 9, no. 4, pp. 434–447, Oct. 1996.

[51] M. Bowles, “Relearning to E-learn: Strategies for Electronic Learning and Knowledge,” *Educ. Technol. Soc.*, vol. 7, no. 4, pp. 212–220, 2004.

[52] G. Marchionini, *Information Seeking in Electronic Environments*, Cambridge. Cambridge university press, 1997.

[53] J. Swiners and J. Briet, “L’intelligence créative : Au-delà du brainstorming, innover en équipe,” 2004. [Online]. Available: http://www.amazon.fr/Lintelligence-créative-Au-delà-brainstorming-innover/dp/2840013851. [Accessed: 20-Nov-2014].

[54] M. de Rond, *The structure of serendipity*. Judge Business School, University of Cambridge, 2005.

[55] E. L. Rissland, “AI and Similarity,” *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 39–49, May 2006.

[56] D. Gentner and A. B. Markman, “Structure mapping in analogy and similarity.,” *Am. Psychol.*, vol. 52, no. 1, p. 45, 1997.

[57] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, “A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain.,” *J. Biomed. Inform.*, vol. 48, pp. 38–53, Apr. 2014.

[58] B. Elayeb, I. Bounhas, O. Ben Khiroun, F. Evrard, and N. Bellamine Ben Saoud, “A comparative study between possibilistic and probabilistic approaches for monolingual word sense disambiguation,” *Knowl. Inf. Syst.*, vol. 44, no. 1, pp. 91–126, May 2014.

[59] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis, and E. E. Milios, “Semantic similarity methods in wordNet and their application to information retrieval on the web,” in *Proceedings of the seventh ACM international workshop on Web information and data management - WIDM ’05*, 2005, p. 10.

[60] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, “Information Retrieval by Semantic Similarity,” *Int. J. Semant. Web Inf. Syst.*, vol. 2, no. 3, pp. 55–73, 2006.

[61] S. E. Schaeffer, “Graph clustering,” *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[62] U. Brandes, D. Delling, and M. Gaertler, “On finding graph clusterings with maximum modularity,” in *International Workshop on Graph-Theoretic Concepts in Computer Science*, 2007, pp. 121–132.

[63] L. Eronen and H. Toivonen, “Biomine: predicting links between biological entities using network models of heterogeneous databases.,” *BMC Bioinformatics*, vol. 13, no. 1, p. 119, Jan. 2012.

[64] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and application of a metric on semantic nets,” *IEEE Trans. Syst. Man. Cybern.*, vol. 19, no. 1, pp. 17–30, 1989.

[65] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification,” vol. 49, pp. 265--283, Jan. 1998.

[66] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* -, 1994, pp. 133–138.

[67] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[68] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, p. 3, Jan. 2001.

[69] D. Lin, “An Information-Theoretic Definition of Similarity,” *ICML*, pp. 296–304, Jul. 1998.

[70] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” p. 15, Sep. 1997.

[71] C. Pierce, “Écrits sur le signe,” *Paris, Fr. Ed. du Seuil*, 1978.

[72] N. J. Belkin and W. B. Croft, “Information filtering and information retrieval: two sides of the same coin?,” *Commun. ACM*, vol. 35, no. 12, pp. 29–38, Dec. 1992.

[73] D. Maltz and K. Ehrlich, “Pointing the way,” in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’95*, 1995, pp. 202–209.

[74] G. Fischer and C. Stevens, “Information access in complex, poorly structured information spaces,” in *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI ’91*, 1991, pp. 63–70.

[75] D. M. Nichols and D. M. Nichols, “Implicit Rating and Filtering,” *Proc. FIFTH DELOS Work. Filter. Collab. Filter.*, pp. 31--36, 1997.

[76] D. B. Terry and D. B., “A tour through Tapestry,” in *Proceedings of the conference on Organizational computing systems - COCS ’93*, 1993, pp. 21–30.

[77] D. B. Hauver and J. C. French, “Flycasting: using collaborative filtering to generate a playlist for online radio,” in *Proceedings First International Conference on WEB Delivering of Music. WEDELMUSIC 2001*, pp. 123–130.

[78] R. Burke, “Hybrid Recommender Systems: Survey and Experiments,” *User Model. User-adapt. Interact.*, vol. 12, no. 4, pp. 331–370, 2002.

[79] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, “Methods and metrics for cold-start recommendations,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’02*, 2002, p. 253.

[80] T. Miranda, T. Miranda, M. Claypool, M. Claypool, A. Gokhale, A. Gokhale, T. Mir, P. Murnikov, P. Murnikov, D. Netes, D. Netes, M. Sartin, and M. Sartin, “Combining Content-Based and Collaborative Filters in an Online Newspaper,” *Proc. ACM SIGIR Work. Recomm. Syst.*, 1999.

[81] Q. Li and B. M. Kim, “An approach for combining content-based and collaborative filters,” in *Proceedings of the sixth international workshop on Information retrieval with Asian languages -*, 2003, vol. 11, pp. 17–24.

[82] J. Bar-Ilan, “Manipulating search engine algorithms: the case of Google,” *J. Information, Commun. Ethics Soc.*, vol. 5, no. 2/3, pp. 155–166, Oct. 2007.

[83] G. Edelman, *Neural Darwinism: The theory of neuronal group selection.* Basic Books, 1987.

[84] P. Lemoisson, G. Surroca, C. Jonquet, and S. A. Cerri, “ViewpointS: When Social Ranking Meets the Semantic Web,” *FLAIRS 17 Conf. Proc.*, 2017.

[85] G. J. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, vol. 2005. John Wiley & Sons, 2005.

[86] G. M. Edelman and G. N. Reeke, “Selective networks capable of representative transformations, limited generalizations, and associative memory.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, no. 6, pp. 2091–5, Mar. 1982.

[87] G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, “Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d’information scientifique,” *IC - 25èmes Journées francophones d’Ingénierie des Connaissances*. Clermont-Ferrand, France, pp. 175–186, 13-May-2014.

[88] G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, “Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité,” in *IC 2015*, 2015, p. 12.

[89] G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, “Preference Dissemination by Sharing Viewpoints : Simulating Serendipity,” *KEOD: Knowledge Engineering and Ontology Development*, vol. 7th Intert, no. 2. pp. 402–409, 12-Nov-2015.

[90] G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, “Subjective and generic distance in ViewpointS,” in *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS ’16*, 2016, pp. 1–6.

[91] S. Catellin, *Sérendipité: du conte au concept*, Seuil. 2014.

[92] J. Corneli, A. Pease, and S. Colton, “Modelling serendipity in a computational context,” *arXiv Prepr. arXiv1411.0440*, 2014.

[93] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, “BIOTEX: a system for biomedical terminology extraction, ranking, and validation,” *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*. CEUR-WS.org, pp. 157–160, 2014.

[94] G. Surroca, P. Lemoisson, C. Jonquet, and S. A. Cerri, “Subjective and generic distance in ViewpointS,” in *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS ’16*, 2016, pp. 1–6.