Université Libanaise
Faculté des Sciences

No d'ordre :

Université Paul Sabatier
Toulouse III- IRIT

Master 2 Recherche
Informatique et Télécommunication

Année universitaire 2007/2008

Une plateforme pour l'étude et la visualisation de liens lexicaux, sémantiques et structuraux dans les Réseaux Petits Mondes Hiérarchiques (RPMH).

Réalisé par : Antoun Yaacoub

Sous la direction de Dr. Ali Awada

Composition du Jury :

Dr. Ali Awada	Directeur
Dr. Bilal Chebaro	Examinateur
Dr. Kabalan Barbar	Examinateur

Soutenu le 17 Décembre 2008

Consigne

Dans de ce rapport, nous avons été obligé, en vue de le restreindre à un nombre raisonnable de pages, d'omettre

- Les notions liées au formalisme de graphe (Graphe non orienté - Graphe orienté - Graphe réflexif - Graphe symétrique - Voisin ou contact – Chemin – Distance - Diamètre - Graphe augmenté - Chemin de longueur m - Graphe fortement connexe - Matrice d'adjacence d'un graphe - Matrice Markovienne d'un graphe G réflexif)[1].
- L'approche mathématique relative aux Réseaux Petits Mondes Hiérarchiques (RPMH)
- L'approche mathématique de l'analyse en composante principale.
- La notion d'Ontologie et sa confrontation avec
  - le vocabulaire contrôlé
  - la taxonomie
  - le thésaurus
  - le réseau sémantique

  et sa relation avec
  - la philosophie
  - l'intelligence artificielle
  - le web sémantique
- L'approche informatique de la lemmatisation utilisée.

Toutes les parties omises, une grande partie de la bibliographie et l'ensemble des ressources électroniques sont disponibles sur le site conçu à cet effet: http://antoun.yaacoub.org/m2r

[1] Pour plus de détails sur les graphes: http://www.math.fau.edu/locke/graphthe.htm

Remerciements

Selon une enquête menée en 2005 par la société libanaise M.E.S. (Middle East Studies[1]) auprès d'un échantillon de 2023 personnes de la population éduquée au Liban comportant des professeurs, docteurs et élèves universitaires de divers spécialisations, les résultats ont montré que le nombre d'achat de livres et de gens qui lisent sont en baisse permanente. Ce qui m'a poussé à rédiger ce remerciement d'une manière totalement différente à ce qui est admis, relève du fait que 80.23% de la population libanaise intérogée prête en rédigeant ou en lisant une très grande importance au remerciement, voire 19.77% ne lisent que le remerciement!!, ce qui selon l'étude pousse les rédacteurs à bien soigner leurs remerciements et à y inclure toute personne potentiellement acheteuse leurs œuvres pour augmenter les ventes (selon l'étude aussi, la plupart des libanais achète un livre si leur nom y apparaît).

Une enquête similaire[2], mais cette fois ci menée auprès du peuple anglais révèle des résultats majoritairement différents. Les résultats montrent que les chercheurs et les universitaires ont tendance à se préoccuper de la bibliographie parce que ça leur permet d'estimer et d'évaluer l'importance des articles (un article qui est très cité a souvent plus de chance d'être lu) et ainsi d'orienter leurs choix de lecture vers des articles de grande informativité. Quant aux français, aucune enquête (selon ma recherche) ne s'est intéressée de fournir des réponses à ces questions mais il me semble que les français ne vont pas trop s'éloigner par leurs opinions de ceux de leurs voisins anglais et si vous ne me croyez pas, vous n'avez qu'à demander Dr. Ali Awada ; il a plein d'anecdotes à ce sujet à vous raconter.

Etant libanais, je veux bien remercier toutes les personnes que je connais !!!!, mais je tiens avant tout à remercier les rapporteurs de mon mémoire. A ne pas oublier tout au long de mon cursus universitaire, les personnes qui m'ont procuré une base solide en informatique et m'ont incité à y approfondir mes connaissances.

Je tiens également à exprimer ma gratitude envers toutes les personnes qui ont contribué scientifiquement ou humainement à la réalisation de ce travail de recherche, spécialement mes amis Abir, Khodor et Rabih. Un très grand merci à mes amis Soha et Tannous pour leur soutien et la relecture minutieuse de ce rapport. Je pense aussi à mes parents, mon frère et ma sœur, qui m'ont apporté un soutien gigantesque, non seulement au niveau sentimental, mais également par leurs encouragements, dont j'avais besoin pour mener à sa meilleure fin ce travail.

Je remercie de même le corps administratif du master recherche à l'UL pour leur soutien et le corps professoral à l'UPS pour les informations qu'ils m'ont transférées (via le net :P) .

Le plus grand merci s'adresse à Dr. Ali Awada, le directeur de ma mémoire. Comment le remercier de sa confiance en moi, en acceptant de tenter la chance dans ce périlleux chemin de cette mémoire ? (Il faut lui demander s'il va de même rempiler pour une 3ème fois ?!?!) Sa philosophie et sa psychologie m'ont été d'un grand secours. De tout mon cœur, un grand merci pour lui. Il m'a aidé à surmonter les multiples difficultés rencontrées. Merci pour son soutien, son inspiration constante et sa confiance, sa lecture attentive et pour toutes ses remarques pertinentes et constructives sur le manuscrit. Ses remarques et ses critiques m'ont permis de mieux visionner mon travail et il n'a guère cessé de m'encourager à améliorer d'avantage la qualité de mes raisonnements. Merci finalement de m'avoir aidé à progresser dans mon travail.

---

[1] Je tiens à remercier le directeur du M.E.S., Général Tannous Mouawad pour la consultation complète et gratuite de l'enquête.

[2] http://www.literacytrust.org.uk et http://www.literacytrust.org.uk/Database/Mori.html [30 Novembre 2008]

Résumé

Cette recherche présente une méthode pour l'étude de différents types de relations dans divers univers tels que les dictionnaires, les pages Web et les textes. Dans un premier temps, il s'agit de répertorier les différents types de relations existant entre des objets des univers précédemment mentionnés. La deuxième étape consiste à mettre en place une méthode permettant de quantifier la relation à étudier. Les entités constitutives de ces univers (les pages web, les articles associés aux entrées d'un dictionnaire, …) forment un graph RPMH (réseau de petits mondes hiérarchiques) dont les sommets sont les entités et les arcs traduisent un lien (hypertextuel, définitionnel,…) direct entre deux sommets. Ce qui revient donc à présenter une méthode pour l'étude de la structure des grands graphes de type petits mondes hiérarchiques. Notre approche se base sur l'utilisation des matrices markoviennes et sur le fait de multiplier une matrice représentant un graphe k fois pour quantifier la relation entre tous les nœuds et pour calculer des informations sur l'ensemble du graphe. Pour illustrer cette approche, des exemples et résultats concrets sont présentés sur des graphes web et d'origine linguistique.

Mots-clés: *Graphe, Réseau petit monde hiérarchique, structure de lien, polysémie, synonymie, géométrisation du sens, visualisation.*

Abstract

This research provides a method for studying different types of relationships in various worlds such as dictionaries, Web pages and texts. As a first step, we'll identify different types of relationships between objects of the universes previously mentioned. The second step consists in developing a method to quantify the relationship to study. The constituents of the universe (web pages, items associated with entries in a dictionary, …) form a graph HSWN (hierarchical small worlds network) whose nodes are entities and edges reflect a direct link (hypertext, definitional …) between two nodes. This leads to introduce a method for studying the structure of large graphs of hierarchical small worlds type. Our approach is based on the use of Markovian matrices and the fact that we multiply the matrix k times to quantify the relationship between all the nodes and for calculating information on the whole graph. To illustrate this approach, examples and results are displayed on web and linguistic graphs.

Keywords: *Graph, Hierarchical small world network, link structure, polysemy, synonymy, visualization*

Table des matières

Chapitre 0
Introduction

Les recherches en sémantique lexicale s'appuient de plus en plus sur des ressources électroniques de grande taille (dictionnaires informatisés, corpus, ontologies) à partir desquelles on peut obtenir diverses relations sémantiques entre unités lexicales. Ces relations sont naturellement modélisées par des graphes. Bien qu'ils décrivent des phénomènes lexicaux très différents, ces graphes ont en commun des caractéristiques bien particulières. On dit qu'ils sont de type petit monde (RPMH – réseau petit monde hiérarchique). Ainsi en va-t-il des réseaux des interactions protéiques, du graphe du web, du graphe des appels téléphoniques, du graphe des co-auteurs scientifiques, des réseaux lexicaux, … Ces graphes ont une topologie bien particulière, dans laquelle la relation entre structure locale et structure globale n'a rien à voir avec celle des graphes (aléatoires ou réguliers) classiquement étudiés. Ceci explique l'intérêt considérable que les RPMH ont suscité dans les communautés scientifiques concernées. En effet, on peut penser que ces caractéristiques reflètent les propriétés des systèmes dont ces grands graphes de terrain rendent compte, et donc que l'étude de leurs structures permettra une meilleure compréhension des phénomènes dont ils sont issus.

Nous voulons mener une étude théorique mathématique et informatique de la structure de ces graphes pour le lexique et pour le web. Il s'agit de les géométriser afin de faire apparaître l'organisation (du lexique/des hyperliens), qui est implicitement encodée dans leur structure.

La problématique de la construction de réseaux sémantiques pour modéliser les liaisons que constituent les individus entre les différents items d'un univers donné est une problématique ancienne. Ross Quillian avait dès 1966 jeté les bases de ce domaine en proposant un réseau sémantique destiné à rendre compte de la part objective du sens des mots [93]. De récents outils issus du domaine de la visualisation d'information permettent de présenter à l'écran et de manipuler de très gros graphes, comportant plusieurs milliers de nœuds. L'affichage de tels graphes repose sur des algorithmes de placement des points pour respecter au mieux des règles "esthétiques" de construction. Des métriques combinatoires mesurant l'aspect structurel du graphe permettent de constituer des sous-graphes pour en simplifier l'affichage.

Nous nous sommes donc penchés sur la construction de vastes réseaux sémantiques, sur la mesure de la structure de ces graphes et sur leur visualisation.

Dans ce mémoire, nous commencerons tout d'abord, par répertorier l'ensemble des relations entre les univers de dictionnaires (mots), textes et de pages web. Ensuite dans le chapitre 2, nous présenterons ce qu'est le petit monde et les RPMH. Nous évoquons au chapitre 3, quelques approches pour géométriser les graphes lexicaux. Nous passerons en revue des principaux outils de visualisation pour la recherche et la classification des documents et de visualisation de la structure du World Wide Web au chapitre 4. Ces quatre chapitres dressent l'état de l'art de notre mémoire. Le chapitre 5 est le cœur de notre contribution. Il concerne l'application de notre approche sur les RPMH. Nous détaillerons notre contribution, et présenterons le module implémenté ainsi que les outils utilisés. Différents résultats correspondant à différentes exemples sont ensuite rapportés et commentés, pour enfin terminer par une conclusion-bilan sur notre approche.

Chapitre 1

Types de relations dans l'univers des dictionnaires, des textes et des pages Web. Etat de l'art

Nous allons nous pencher au cours de ce premier chapitre sur les différents types de relations dans les univers cités. On commencera par expliquer ces différentes relations, en mettant en relief celles qui pourront nous fournir des réponses à notre problématique. On ne prétend pas avoir répertorié la totalité des univers et des relations existantes, mais on a juste passé en revue par ceux et celles qui conviennent le plus à notre domaine de recherche.

1.1 Type de relation existant dans l'univers des dictionnaires (types de relation entre les mots)

La **sémantique** [1] est une branche de la linguistique qui parmi ses plusieurs objets d'étude analyse les rapports de sens entre les mots. On relève plusieurs types de relations [2][3][4] :

1. Selon une définition (généralement attribuée à Leibniz), deux expressions sont **synonymes** si la substitution de l'une par l'autre ne change pas la valeur de vérité de la phrase dans laquelle la substitution est faite. Par cette définition, et si on postule vraiment l'existence, les vrais synonymes sont ainsi rares. Une version affaiblie de cette définition permettrait de définir la synonymie relativement à un contexte: deux expressions sont synonymes dans un contexte linguistique *C* si la substitution de l'un par l'autre en *C* ne modifie pas sa valeur de vérité. Par exemple, la substitution de "*plank*" par "*board*" va rarement modifier les valeurs de vérité dans le contexte de menuiserie, mais il existe d'autres contextes de "*board*" où cette substitution serait totalement inappropriée. Il est à noter que la définition de la synonymie en termes de substituabilité nécessitera le partitionnement du lexique en noms, verbes, adjectifs et adverbes. Ceci pour dire que si les synonymes doivent être interchangeables, alors les mots dans ces différentes catégories syntaxiques ne peuvent pas être synonymes, car ils ne sont pas interchangeables. Les noms expriment des concepts nominaux, les verbes expriment des concepts verbaux et les modificateurs fournissent des moyens pour qualifier ces concepts. Un argument pourrait être aussi présenté en faveur d'une partition encore: certains mots dans la même catégorie syntaxique (en particulier les verbes) expriment des concepts très similaires, mais ils ne peuvent pas être échangés qu'après avoir rendu la phrase non-grammaticale. La définition de la synonymie en termes de valeurs de vérité semble rendre la synonymie comme une relation discrète: deux mots sont soit synonymes ou ils ne le sont pas. Mais, comme certains linguistes ont argumenté, la synonymie sera meilleure perçue comme une extrémité d'un continuum où la similitude de sens pourrait être évaluée et notée. Il est facile de supposer que la relation est symétrique: si *x* est semblable à *y*, alors *y* est également similaire à *x*. La synonymie est alors un rapport de proximité sémantique entre des mots. La proximité sémantique indique que les mots ont des significations très semblables dans un contexte donné et que deux synonymes sont obligatoirement de même nature.
2. Une des relations les plus familières aussi est l'**antonymie**, qui se révèle être difficile à définir. L'antonyme d'un mot *x* est parfois *non-x*, mais pas toujours. Par exemple, *riche* et *pauvre* sont des antonymes, mais dire que quelqu'un n'est pas riche ne signifie pas qu'il doit être pauvre. En fait, beaucoup de personnes se considèrent eux-mêmes ni riches ni pauvres. L'antonymie, qui semble être une simple relation symétrique, est en fait assez complexe ; même en anglais, il y a de difficultés à reconnaître les antonymes.

L'antonymie est une relation lexicale entre les formes des mots, mais pas une relation sémantique entre les sens des mots. Par exemple, les significations "*rise, ascend*" et "*fall, descend*" peuvent être conceptuellement opposées, mais elles ne sont pas des antonymes; [*rise/fall*] sont des antonymes et le sont aussi [*ascend/descend*]. Ce qui nécessite la distinction entre les relations sémantiques entre les formes des mots et les relations sémantiques entre les sens des mots. De même, si deux mots ont, dans un contexte donné, un sens opposé alors ces deux antonymes sont obligatoirement de même nature.

3. Contrairement à la synonymie et l'antonymie, qui sont des relations lexicales entre les formes des mots, l'**hyponymie** / l'**hyperonymie** est une relation sémantique entre les sens des mots: par exemple, "*maple*" est un hyponyme de "*tree*" et "*tree*" est un hyponyme de "*plant*". Une grande attention a été consacrée à cette relation (autrement appelée subordination / superordination, sous-ensemble / sur-ensemble, ou la relation *IS-A*). Un concept représenté par *x* est dit hyponyme du concept représenté par *y* si on peut dire que « une *x* est une (sorte de) *y* ». L'hyponymie est transitive et asymétrique ; elle génère une structure sémantique hiérarchique, dans laquelle un hyponyme est en dessous de son supérieur (*superordinate*). Ces représentations hiérarchiques sont largement utilisées dans la construction de systèmes de recherches d'information, et sont appelées systèmes d'héritage: un hyponyme hérite toutes les caractéristiques du concept générique et a au moins une caractéristique qui le distingue de son supérieur et de tout autre hyponyme supérieur. Par exemple, "*maple*" hérite des caractéristiques de son supérieur, "*tree*", mais se distingue des autres arbres par la dureté de son bois, de la forme de ses feuilles, l'usage de sa sève pour le sirop, etc.
4. Synonymie, antonymie, et hyponymie sont des relations familières. Elles s'appliquent dans tout le lexique et les gens n'ont pas besoin d'une formation spéciale en linguistique afin de les apprécier. Une autre relation sémantique qui bénéficie de ses avantages, est la relation partie-ensemble (ou *Has-a*), connue sous le nom **méronymie / holonymie**. Un concept représenté par *x* est un méronyme d'un concept représenté par *y* si on peut dire que « un *y* a un *x* (en partie) ou un *x* est une partie de *y* ». La relation de méronymie est transitive et asymétrique, et peut être utilisée pour construire une hiérarchie en partie (avec certaines réserves, car un méronyme peut avoir de nombreux holonymes). Il est supposé que le concept d'une partie de son ensemble peut être une partie d'un concept d'ensemble.
5. Dans la suite, on présente, une série de relations lexicales entre les mots mais il n'existe pas à notre jour un dictionnaire publié que ce soit informatique ou pas permettant de repérer ces différentes relations au sein d'un texte donné. L'**éponymie** est le fait de « donner son nom à » quelque chose. Ce qui donne son nom est un éponyme. Par une utilisation abusive, « éponyme » est souvent utilisé comme signifiant "de même nom", sans que soit considéré lequel des deux termes a donné son nom à l'autre (ce qui est donc une incorrection sémantique). Ex : Harry Potter est éponyme du livre, puisque c'est le héros du roman.

   La **métonymie** consiste à employer un mot ou une expression dans un sens différent de son sens propre où ce changement de sens s'opère entre des termes qui évoquent un même ensemble logique structuré en parties. La métonymie vise une relation privilégiée entre les parties : la cause pour l'effet, ou le contenant pour le contenu, l'artiste pour l'œuvre, la nourriture typique pour le peuple qui la mange, la localisation pour l'institution qui y est installée…

   La **synecdoque** est un cas particulier de métonymie où une relation d'inclusion (matérielle ou conceptuelle) lie le terme cité et le terme évoqué.

   La **catachrèse** est l'un des procédés par lesquels le lexique d'une langue s'enrichit. Elle donne un sens nouveau à un mot ou à une expression qui existe déjà. Par exemple, poubelle est catachrèse du nom du préfet de la Seine Eugène-René Poubelle qui imposa en 1884 l'usage de ce récipient.

   Une **antonomase** utilise un nom propre comme nom commun, ou inversement.

   Une **métaphore** est une pratique du langage qui consiste à utiliser dans un contexte B un terme antérieurement usité dans un contexte A plus ancien ou différent.

La **pantonymie** consiste à désigner un terme par un autre, beaucoup plus générique, dans l'ordre de l'hyperonymie. Ces termes passe-partout tels que « truc », « machin », « chose », « bidule », « toutim » qui renvoient à des personnes, des objets ou des notions, sont des pantonymes.
La **paronymie** est une relation lexicale qui porte entre deux mots dont les sens sont différents mais dont l'écriture et/ou la prononciation sont fort proches. En somme, il s'agit d'une homonymie approximative.
Un **rétronyme** est un mot nouveau ou une expression nouvelle créé pour désigner un vieil objet ou concept dont le nom original est devenu utilisé pour quelque chose d'autre, ou qui n'est plus unique. La création d'un rétronyme est généralement la conséquence d'une avancée technologique.
L'**homonymie** est la relation entre des mots d'une langue qui ont la même forme orale ou écrite mais des sens différents.
Tout en ayant des sens différents, les homonymes peuvent :

- s'écrire de la même manière et se prononcer différemment (**Homographie**)
- se prononcer de la même manière et s'écrire différemment (**Homophonie**)
- se prononcer et s'écrire de la même manière (**Homonymie parfaite**)

En linguistique, on décrit l'homonymie comme la relation entre plusieurs formes linguistiques ayant le même signifiant, graphique ou phonique, et des signifiés entièrement différents. Cette acception généralise la notion habituelle d'homonymie à des formes qui ne sont pas des mots, par exemple des locutions.

6. La **polysémie** [92] est la qualité d'un mot ou d'une expression qui a deux voire plusieurs sens différents (on le qualifie de polysémique). Il ne faut pas confondre polysémie et homonymie. Deux mots homonymes ont la même forme (phonique ou graphique) mais sont des mots totalement différents. Ils ont deux entrées distinctes dans le dictionnaire. La polysémie analyse bien les différents sens d'un même mot. Ils sont présentés dans la même entrée du dictionnaire. C'est le cas d'une très grande majorité des mots courants du dictionnaire. Il arrive même qu'un mot désigne ainsi à la fois une chose et son contraire. Ainsi en est-il des mots français :
   - hôte, désignant selon le contexte celui qui reçoit ou celui qui est reçu ;
   - amateur, désignant selon le contexte une personne avertie ou ignorante ;
   - plus : il y en a plus (il y en a davantage) ou il n'y en a plus (il n'en reste pas).

   L'évolution du langage (due au fait qu'il faut bien décrire soit un monde qui évolue, soit un monde dont au moins notre connaissance évolue) conduit à utiliser parfois un mot dans un nouveau sens, le plus souvent par extension de sens. On parlera par exemple d'une feuille de papier ou du pied d'un arbre, par analogie avec une feuille d'arbre ou avec le pied d'un animal.
   Les chaînes sémantiques permettent souvent, en jouant sur la polysémie, de passer de synonyme en synonyme d'un mot à son contraire.
   Exemple : Léger = inconséquent = ... = gauche = lourd
   On peut passer de même de « vie » à « mort », d'« homme » à « femme », etc., le plus souvent par des chaînes ne comportant pas plus de dix mots. L'astuce réside dans le fait que si A est synonyme de B dans un certain contexte, et B synonyme de C dans un autre contexte, cela n'implique nullement que A soit synonyme de C dans quelque contexte que ce soit : la relation de synonymie n'est pas transitive.

1.2 Type de relation et distance intertextuelle

Le texte a une réalité matérielle qui se prête à l'analyse. Les éléments comptables peuvent être soumis aux instruments de mesure, des plus menus (les atomes des lettres) aux plus volumineux (les grosses molécules des structures syntaxiques, des schémas narratifs ou topoi, des constellations lexicales ou thématiques).

Le calcul de distance entre les textes, développé à des fins de classification ou d'attribution d'auteur, est la démarche par laquelle on tente d'évaluer la plus ou moins grande

ressemblance entre divers textes en prenant appui sur des éléments susceptibles d'être mesurés ou dénombrés, qui permettront d'une part de quantifier et ordonner ces degrés de ressemblance, et d'autre part de reproduire le calcul autant de fois qu'on le souhaite sur différents types de textes en éliminant tout impact de la subjectivité.

Ce sujet est connu sous le nom de « **connexion lexicale** ». Celui-ci est défini comme « l'intersection du vocabulaire de deux textes » [5][6][7]. La connexion est donc le complémentaire de la **distance** [8].

L'indice de la distance intertextuelle[1] [9] fournit un bon outil pour la mesure des ressemblances et des dissemblances entre textes. Ses propriétés facilitent la recherche des meilleures partitions possibles, au sein de vastes bases de données textuelles, grâce à des techniques comme la classification hiérarchique ou l'analyse arborée [10].

La statistique commence à proposer des outils qui seront d'une grande utilité pour l'analyse des grandes bases de textes enregistrées sur support électronique. Mais ce nouvel intérêt soulève inévitablement le problème des normes d'enregistrement des textes et celui des conventions de mesure car il ne sert à rien de mesurer finement un phénomène dont la saisie n'aurait pas été assurée, au préalable, avec un minimum de rigueur.

1.3 Structure des liens et relation dans le World Wide Web (WWW)

La structure du réseau web peut être une riche source d'information concernant le contenu de cet environnement, sachant que nous avons des moyens efficaces pour la comprendre [29].

1.3.1 Analyse de la structure de lien.

Analyser la structure *hyperlien* entre les pages webs permet de fournir un moyen efficace pour résoudre le problème du partitionnement (Clustering) des pages webs [11][28]. Ce dernier traite la question de disséquer une population hétérogène en des sous-populations qui sont en quelque sorte plus cohésive. Dans le cadre de la WWW, ceci peut impliquer la distinction des pages liées à différents significations ou sens des termes d'une requête formulée dans un moteur de recherche.

1.3.2 Construction d'un sous-graphe réduit de la WWW

On peut voir toute collection V de pages hyperliens comme un graphe orienté G = (V, E): les nœuds correspondent à des pages, et une arête orientée $(p; q) \in E$ indique la présence d'un lien de P à Q [80].

1.3.3 Partitionnement des structures de lien

La bibliométrie [12] est l'étude des documents écrits et de leur structure de référence. La recherche en bibliométrie est depuis longtemps préoccupée par le recours à des citations pour produire des estimations quantitatives de l'importance et de l'«impact» des articles scientifiques et de revues.

[1] Nous avons omis l'état de l'art concernant les diverses distances intertextuelles proposées dans la littérature pour des raisons citées ultérieurement. On peut les retrouver sur le site http://antoun.yaacoub.org/m2r

Le partitionnement basé sur les liens dans le contexte de la bibliométrie, hypertexte, et le Web a essentiellement porté sur le problème de la décomposition d'une collection de nœuds explicitement représentée en des sous-ensembles "cohésif - solidaire". À ce titre, il a été surtout appliqué pour des ensembles d'objets de taille modérée - Par exemple, une collection réduite de revues scientifiques, ou de l'ensemble des pages sur un seul site www.

À un niveau très élevé, le partitionnement exige une fonction de similarité sous-jacente entre les objets, et une méthode pour produire des *clusters* à partir de cette fonction de similarité.

Deux fonctions de similarité émergent des études de la bibliométrie et sont *bibliographic coupling* (en raison de Kessler [13]) et *co-citation* (en raison de Small [14]). Pour une paire de documents p et q, la première quantité est égale au nombre de documents cités par p et q ensemble, et la seconde quantité est le nombre de documents qui citent les deux p et q. ***Co-citation*** a été utilisée comme une mesure de la similitude des pages web par Larson [15] et par Pitkow et Pirolli [16]. Weiss et al. [17] définissent les mesures de similarité à base de lien pour les pages dans un environnement hypertexte qui généralise *co-citation* et *bibliographic coupling* pour permettre des chaînes arbitrairement longues de liens.

Plusieurs méthodes ont été proposées dans ce contexte pour produire des *clusters* à partir d'un ensemble de nœuds annoté avec ces informations de similitude. Small et Griffith [18] utilisent une recherche par largeur pour calculer les composantes connexes du graphe non orienté dans lequel deux nœuds sont reliés par une arête si et seulement si ils ont une valeur de co-citation positive. Pitkow et Pirolli [16] appliquent cet algorithme pour étudier les relations basées sur les liens dans une collection de pages web. On peut également utiliser l'analyse en composantes principales [19][20] et les techniques de réduction de la dimension. Dans ce cadre, on commence avec une matrice M contenant les informations de similitude entre les paires des nœuds, et on calcule ensuite une représentation (basée sur cette matrice) de chaque nœud *i* comme un vecteur *{vi}* de plus grande dimension. On utilise les premiers vecteurs propres non-principaux de la matrice de similitude M pour définir un sous-espace de dimension plus petite dans lequel les vecteurs *{vi}* peuvent être projetés. Une variété de techniques à base de visualisations géométriques peut être employée pour identifier les *clusters* denses dans cette faible dimension. Les théorèmes standards de l'algèbre linéaire fournissent un sens précis dans lequel la projection sur les *k* premiers vecteurs propres produit une distorsion minimale sur toutes les projections de k-dimensions des données. Small, McCain, et d'autres ont appliqué cette technique aux revues. L'application des techniques de la réduction de la dimension aux *clusters* des pages www basées sur la co-citation a été employée par Larson et par Pitkow et Pirolli.

Le partitionnement des documents ou des pages hypertextes peut bien sûr compter sur des combinaisons d'informations textuelles et à base de liens. Des combinaisons de ces mesures ont été étudiées par Shaw [21][22] dans le cadre de la bibliométrie. Pirolli, Pitkow, et Rao [23] ont utilisé une combinaison de la topologie des liens et de la similitude intertextuelle pour regrouper et classer les pages sur le www.

Aussi, le domaine de *spectral graph partitioning* a été lancé par les travaux de Donath et Hoffman et Fiedler[2]. Les méthodes de partitionnement du graphe spectral relient les partitions peu connectées d'un graphe G non-orienté aux valeurs et vecteurs propres de sa matrice d'adjacence A. Chaque vecteur propre de A a une seule coordonnée pour chaque nœud de G, et donc peut être considérée comme une assignation de poids aux nœuds de G. Chaque vecteur propre non-principal a à la fois des coordonnées positives et négatives. Une heuristique fondamentale qui ressort de l'étude spectrale de ces méthodes est que les nœuds correspondant à la plus grande coordonnée positive d'un vecteur propre donné ont tendance à être très peu connectés aux nœuds correspondants à la plus grande coordonnée négative du même vecteur propre.

[2] voir le livre de Chung [24] pour une vue d'ensemble

Dans une autre direction, *centroid scaling* est une méthode de partitionnement conçue pour représenter deux types d'objets dans un espace commun. Prenons, par exemple, un ensemble de personnes qui ont fourni des réponses aux questions d'un sondage - on souhaite représenter à la fois le peuple et les réponses possibles dans un espace commun, de manière que chaque personne est "proche" des réponses qu'elle a choisies, et que chaque réponse est "proche" de la population qui l'a choisie. *Centroid scaling* se base sur une méthode fondée sur les vecteurs propres pour atteindre cet objectif. Les méthodes à *Centroid scaling* ne sont généralement pas concernées seulement par l'interprétation de la plus grande coordonnée dans les représentations qu'ils produisent, mais l'objectif est de déduire une notion de similitude depuis un ensemble d'objets à travers des moyens géométriques. Dans le cadre de la recherche d'information, la méthode *Latent Semantic Indexing* de Deerwester et al. [25] qui consiste à appliquer une approche *centroid scaling* à un modèle d'espace-vectoriel de documents [26], a permis de représenter les termes et les documents dans un seul espace commun de faible dimension.

Enfin, une autre approche à ce problème est consignée dans un système appelé HyperClass [27], qui fait usage de modèles statistiques robustes telles que les chaines aléatoires de Markov (Markov random fields MRF's) couplée à une technique de relaxation d'étiquetage. En utilisant cette approche, on obtient une catégorisation améliorée de l'exactitude en exploitant les informations des liens au voisinage du document. L'usage de la MRF découle du fait que les pages qui sont les mêmes ou ont des sujets connexes ont tendance à être liés plus fréquemment que ceux dont les sujets sont non-connexes.

Chapitre 2

Réseau Petit Monde Hiérarchique (RPMH) / *Hierarchical Small World Network*. Etat de l'art

Le fait que les réseaux sociaux semblent être des petits mondes a été longtemps l'objet d'observations anecdotiques.

L'hypothèse que "n'importe quelle deux personnes dans le monde sont reliées par une petite chaîne de connaissances sociales" est connu sous le nom du phénomène ***petit monde*** *[84]. L'hypothèse a été testée pour la première fois par Stanley Milgram en 1967. Dans l'expérience qu'il a dirigée, Milgram a constaté qu'il y avait une moyenne de six connaissances entre n'importe quels paire de participants. Une rafale de travail a suivie cette découverte. Le phénomène a été capturé dans plusieurs domaines, notamment en filmographie, et en mathématiques [85].*

On appelle réseau d'interactions tout ensemble d'entités interagissant de façon individuelle. Les grands réseaux d'interactions recouvrent ainsi des réseaux aussi divers que le réseau des routeurs d'Internet, le réseau des contacts sociaux entre individus, ou le réseau des réactions chimiques entre protéines dans le métabolisme d'un être vivant. On parlera également de ***réseaux réels****. [31]*

*L'augmentation récente des capacités de traitement et de collecte d'un grand nombre de données statistiques sur ces réseaux a permis l'essor des études de ces objets. En particulier, on a observé expérimentalement que ces réseaux partageaient des propriétés macroscopiques communes. Une de ces propriétés est l'****effet petit monde****.*

Nous aborderons tout d'abord dans ce chapitre l'expérimentation de Milgram en mettant en relief sur ces résultats obtenus concernant l'effet petit monde. Ensuite, nous exposerons les propriétés identificatrices et les modèles proposés dans la littérature pour la navigation de ce type de réseaux. Nous finirons par mentionner quelques exemples de réseaux petits mondes dont le plus important est le réseau lexical.

2.1 L'expérimentation de Milgram (1967)

En 1967, Stanley Milgram a tenté de vérifier quantitativement le phénomène petit monde. Il a sélectionné au hasard des gens depuis Kansas et Nebraska, et les a initiés à démarrer une chaîne de transfert de lettres. Les cibles de ces lettres ont été à Cambridge, MA et Boston. Chaque initiateur a à envoyer une lettre à travers la poste à la personne cible. Mais le jeu avait des règles. Les initiateurs peuvent seulement envoyer le dossier aux gens qu'ils connaissaient directement. Chaque personne devait à son tour envoyer la lettre à une autre personne, ...

Son premier objectif était une personne à Cambridge, et les initiateurs sont situés à Wichita, Kansas. Milgram a constaté que les premières lettres ont été reçues en quatre jours et ont pris seulement deux intermédiaires de connaissances. Dans une deuxième étude, les initiateurs sont situés à Nebraska, et la personne cible de Sharon, MA, et travaille à Boston. Milgram a indiqué que «les chaînes ont variés de deux à dix intermédiaires de connaissances, avec une médiane à cinq". Toute personne semble atteindre la cible avec une moyenne de six sauts.

En termes de graphe, Milgram a essayé de trouver des plus courts chemins dans un réseau social donné de personnes aux États-Unis. Au lieu d'utiliser la «bonne» approche de la BFS (Breadth First Search - parcours en largeur), il a dû compter sur DFS (Deep First Search - parcours en profondeur), comme le facteur typique de branchement de plusieurs centaines de connaissances sociales qui ont abouti dans les deux expériences à beaucoup de travail pour tous les participants, et une chaîne de lettre illégale en utilisant le système postal des États-Unis . Alors, ce que Milgram a vraiment trouvé était juste une limite supérieure sur les plus courts chemins. Parmi les 296 chaînes, 217 chaînes seulement ont commencé, et 64 sont

achevés. Le nombre de nœuds intermédiaires a varié entre de 2-10, avec une médiane de 5 et une moyenne de 6.

Kleinfeld [30] a passé en revue l'expérience de Milgram et a constaté qu'il existe un certain nombre de lacunes dans son expérience. La technique d'échantillonnage de Milgram afin de déterminer les points de départ des chaînes a été biaisée vu que sa cible était un actionnaire, et un certain nombre des initiateurs ont été aussi des actionnaires, qui sont en principe plus susceptibles de connaître des actionnaires. Il a recruté des gens par le biais de la publicité pour trouver des gens qui sont bien « connectés », et l'expérience a été généralement conçue pour attirer des gens qui sont économiquement bien connus, et qui ont tendance à avoir des connexions de plus longue distance. En outre, sur les 2 études de Kansas et de Nebraska, seule celle de Nebraska, qui a eu beaucoup plus de succès, a été publiée.

2.2 Test de reconnaissance des RPMH

Nous en venons à présent aux trois principales propriétés qui ont été observées expérimentalement pour reconnaître un RPMH : ***le petit diamètre***, ***la distribution des degrés suivant une loi de puissance*** et ***la forte densité locale ou clustering***.

1. **Petit diamètre**. Le diamètre d'un réseau est formellement le plus long des plus courts chemins entre deux entités, ou nœuds, du réseau, via ses connexions. Dans l'étude des grands réseaux d'interactions, il a souvent fait référence à un petit diamètre pour décrire l'observation d'une petite distance moyenne, le diamètre étant généralement trop coûteux à obtenir (puisqu'il s'agit d'un pire cas). Dans de nombreux réseaux d'interactions, la distance moyenne observée est, de façon surprenante, de l'ordre du logarithme du nombre total de nœuds, alors que le nombre de connexions, ou liens, reste très inférieur au carré du nombre de nœuds. Newman donne l'exemple d'un réseau de co-auteurs d'articles de biologie de 1 520 251 nœuds et 11 803 064 liens, dans lequel la longueur moyenne d'un chemin est 4,9. Un autre exemple est l'étude du réseau Internet de 10 597 nœuds et 31 992 liens effectuée par Faloutsos et al. [32], où la longueur moyenne d'un chemin est de 3,3.

   Si les chemins sont très courts par rapport à la densité de liens des réseaux, la pertinence de cette propriété peut toutefois être remise en cause lorsque l'on remarque qu'un réseau aléatoire uniforme, où chaque paire de nœuds est reliée indépendamment avec une probabilité fixée, présente également un diamètre logarithmique en son nombre de nœuds lorsque le nombre d'arête est très faible (de l'ordre de n log n pour n nœuds).

2. **Distribution des degrés suivant une loi de puissance**. On dit qu'un réseau présente une distribution des degrés suivant une loi de puissance si le nombre de nœuds de degré k est proportionnel à $1/k^{\alpha}$, pour une constante $\alpha > 0$, sur un intervalle de plusieurs ordres de grandeur (par exemple entre $k = 10$ et $k = 10^6$). En 1999, Faloutsos et al. [32] ont observé que le réseau Internet présentait cette propriété. Par la suite, elle a également été observée dans des réseaux de pages web, et des réseaux de distribution d'électricité. Cette découverte a été cruciale pour les travaux sur la propagation des virus dans les réseaux réels. Avant cette découverte, le modèle usuel pour ce type d'étude était un réseau aléatoire uniforme, sur lequel on observe un effet de seuil pour la transmission d'un virus, c'est-à-dire qu'en dessous d'une fraction d'individus infectés, le virus cesse de se répandre. Mais une étude similaire menée sur un modèle présentant une distribution de degrés en loi de puissance a donné des résultats différents, en particulier l'effet de seuil disparaît. Une telle observation a donc remis en cause les mécanismes mis en place pour freiner la propagation des virus et les modèles utilisés jusqu'alors.

Récemment, les travaux d'Achlioptas et al. [33] ont toutefois mis en doute la pertinence de la loi de puissance comme propriété caractéristique, en mettant en évidence un risque de biais lié à la méthode de collecte des données.

Ils démontrent qu'en parcourant un réseau aléatoire uniforme selon un processus similaire à celui utilisé pour les parcours expérimentaux du réseau des routeurs Internet (le parcours par traceroute à partir d'un faible nombre de sources), on obtient des données statistiques présentant une distribution des degrés suivant une loi de puissance. Il est pourtant connu que le réseau aléatoire uniforme présente une distribution des degrés suivant une loi de Poisson. Cette étude souligne le risque de baser la construction de modèles sur une propriété qui pourrait n'être qu'un biais de la mesure expérimentale.

3. **Forte densité locale ou clustering**. On parle de forte densité locale d'un réseau lorsque les voisins d'un même nœud sont très reliés entre eux. Dans un réseau social, par exemple, cela signifie que les amis d'un même individu ont une grande probabilité d'être amis entre eux. Pour quantifier cette propriété, Watts et Strogatz [34] ont introduit en 1998 la notion formelle de coefficient de clustering. Il s'agit de la moyenne, sur tous les nœuds u, du ratio du nombre de voisins de u qui sont reliés entre eux sur le nombre total de liens qui pourraient potentiellement exister entre ces voisins. Ils observent que ce coefficient s'élève à 0,2 dans un réseau de collaborations entre acteurs de cinéma comprenant 449 913 nœuds et 25 516 582 liens, alors qu'en construisant un réseau aléatoire uniforme ayant le même nombre de nœuds et de liens, on obtient un coefficient de l'ordre de 10^{-4} seulement.

   Toutefois, cette propriété semble encore mal définie par le coefficient de clustering. Il est en effet possible de construire des réseaux ayant un fort coefficient de clustering alors que, majoritairement, les nœuds n'ont pas de voisins reliés entre eux. Un exemple est illustré sur la figure ci-dessous, où le coefficient de clustering vaut environ 1/2 alors que n/2 nœuds sont disposés en chaîne et n'ont aucun de leurs voisins reliés entre eux.

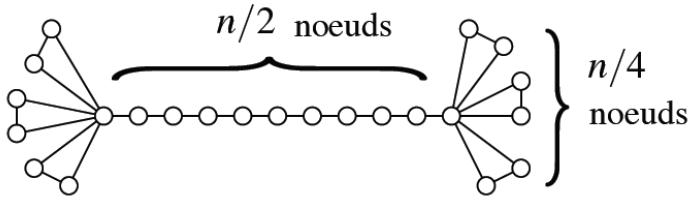


Figure 1: *Exemple montrant que le coefficient de clustering vaut environ 1/2 alors que n/2 nœuds sont disposés en chaîne et n'ont aucun de leurs voisins reliés entre eux.*

2.3 Modèles pour les réseaux d'interactions

Un réseau se représente de façon naturelle par un graphe. Nous utiliserons dans toute la suite l'un ou l'autre vocabulaire de façon indifférenciée : un nœud correspond à un sommet et un lien à une arête.

1. **Graphe aléatoire uniforme d'Erdös et Rényi**. Le premier graphe étudié comme un modèle possible pour les réseaux réels fut le graphe aléatoire G(n, p) d'Erdös et Rényi [35]. Il s'agit d'un graphe aléatoire uniforme sur n sommets où il existe une arête entre deux sommets avec une probabilité p constante. Cet objet mathématique a été très étudié[1]. Nous nous intéressons seulement ici à son rôle historique en tant que modèle

[1] Se référer à *B. B. Bollobás. Random Graphs. Academic Press, London, 1985.* pour une vue d'ensemble des résultats

pour les réseaux d'interactions, car il était considéré jusqu'à récemment comme le seul modèle, par défaut, pour les réseaux réels.

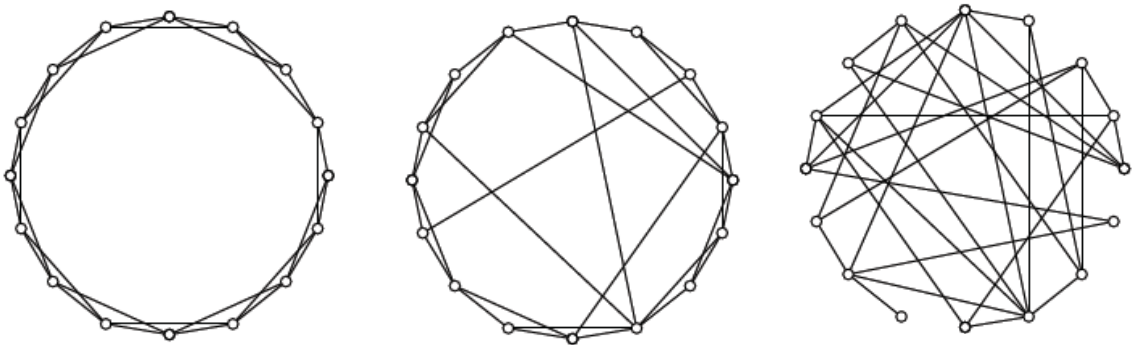
Or les récentes observations expérimentales (qui étaient impossibles ou très partielles auparavant) ont mis en lumière d'importantes différences. Ainsi, la distribution des degrés suit une loi de Poisson (exponentielle), alors qu'il s'agit d'une loi de puissance pour la grande majorité des réseaux réels. Il s'agit là d'une différence importante puisqu'elle caractérise l'hétérogénéité du réseau. Les nœuds jouant tous le même rôle dans un graphe d'Erdös-Rényi, le graphe ne possède pas de propriété discriminante, et les degrés sont naturellement répartis de façon égale autour de la moyenne. Pour la même raison, ce graphe présente un faible coefficient de clustering, puisque les voisins d'un nœud n'ont aucune raison d'être davantage reliés entre eux que deux nœuds pris au hasard. Enfin, même si ce graphe présente un diamètre polylogarithmique en le nombre de sommets, on peut montrer que tout algorithme décentralisé y calcule des chemins de longueur au moins polynomiale [36]; ce n'est donc pas un petit monde navigable.

Ces différences ont montré l'importance de trouver un modèle plus fidèle.

2. **Modèle d'Albert et Barabasi pour la distribution des degrés**. Suite à la découverte de distributions de degrés suivant une loi de puissance dans de nombreux réseaux réels, la construction de nouveaux modèles a été dirigée vers la reproduction de cette propriété. En 1999, Albert et Barabasi [37] ont popularisé un modèle dynamique permettant d'obtenir une distribution des degrés suivant une loi de puissance. Ce modèle consiste à construire un réseau nœud par nœud, en reliant chaque nouveau nœud préférentiellement aux sommets existants de plus hauts degrés. Il est connu sous le nom de l'attachement préférentiel. Des processus similaires avaient été introduits dès les années 20 par des mathématiciens, puis étudiés en sociologie, mais ce fut la première étude de ce processus en tant que modèle d'un phénomène physique. L'intérêt de ce modèle est sa construction dynamique, puisque dans de nombreux réseaux réels, des nœuds et des liens sont fréquemment ajoutés et enlevés au cours du temps (on peut penser au réseau des pages web par exemple).

   Nous présentons maintenant les deux principaux modèles qui ont été développés spécifiquement pour reproduire l'effet petit monde.

3. **Modèle de petit monde non navigable de Watts et Strogatz**. En introduisant la notion formelle de coefficient de clustering, Watts et Strogatz [34] ont proposé un modèle qui présente à la fois un petit diamètre et un fort coefficient de clustering.
   Une variante du modèle a également été développée et analysée par Newman et Watts. Le modèle est construit de la façon suivante : à partir d'un anneau régulier de n sommets et k arêtes par sommet, distribuées régulièrement par rapport à leur origine, on redirige indépendamment chaque extrémité d'arête avec une probabilité p constante, donnée en paramètre, vers un sommet de l'anneau choisi de manière aléatoire uniforme. La figure suivante illustre ce modèle pour $p = 0$, $p = 1$ et une valeur intermédiaire $0 < p < 1$ qui donne lieu à l'apparition des deux propriétés de petit diamètre et fort clustering simultanément. Le nombre k d'arêtes de départ n'influe pas sur le modèle.



p=0
Grand diamètre
et fort *clustering.*

0<p<1
Petit diamètre
et fort *clustering.*

p=1
Petit diamètre
et faible *clustering.*

Figure 2 : *Modèle de Watts et Strogatz*

En cherchant à reproduire à la fois le petit diamètre présent dans les graphes aléatoires uniformes et le fort coefficient de clustering présent dans les grilles régulières, Watts et Strogatz ont eu l'idée d'introduire une part d'aléatoire sur un réseau régulier et ils ont obtenu les deux propriétés simultanées pour un assez grand intervalle de valeurs du paramètre p. Toutefois, ce modèle ne présente pas la propriété dynamique de navigabilité. Précisément, Kleinberg [36] a montré que tout algorithme de routage décentralisé calcule, entre deux sommets de ce graphe, un chemin de longueur au moins polynomiale en n. Cela signifie que, même lorsque le paramètre p donne naissance à des chemins polylogarithmiques en n entre toute paire de sommets, il n'existe pas d'algorithme qui puisse les découvrir en n'ayant qu'une vue locale du graphe.

4. **Modèle de petit monde navigable de Kleinberg**. Le premier modèle de petit monde présentant la propriété dynamique de navigabilité a été introduit par Kleinberg en 2000 [36][82]. Il s'agit d'une grille régulière de dimension 2, augmentée d'un arc aléatoire par sommet u, dont la destination est v avec probabilité proportionnelle à $1/|u - v|^s$, pour chaque v, où $|u - v|$ est la distance l_1 entre les deux nœuds dans la grille régulière, et $s > 0$ un paramètre. La principale différence avec le modèle de Watts et Strogatz est la distribution non uniforme des liens aléatoires, elle est en effet fortement liée aux positions des nœuds sur la grille sous-jacente. Kleinberg montre que lorsque $s = 2$, un algorithme de routage glouton, très simple et décentralisé, calcule un chemin de longueur polylogarithmique en la taille du graphe ($O(\log^2 n)$ pour n nœuds) entre toute paire de sommets.

   La figure suivante illustre un exemple de réseau de Kleinberg de dimension 2 ayant un lien aléatoire sortant par nœud (en gras). Les liens aléatoires ne sont pas tous représentés pour la lisibilité.

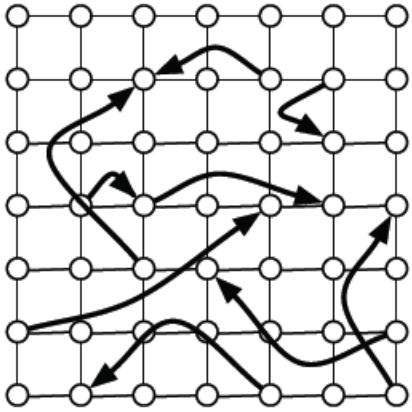


Figure 3 : *Exemple de réseau de Kleinberg* $\mathcal{K}_{1,n}^{2}$

Ce modèle met en valeur deux types d'informations contenues dans un réseau : une information globale, mais partielle, partagée par tous (représentée par la grille) ; et une information locale qui permet la navigation (représentée par les liens aléatoires). Un algorithme est alors dit décentralisé s'il calcule des chemins en connaissant les positions de tous les nœuds sur la grille, mais en n'accédant à la position de la destination d'un lien aléatoire qu'en visitant son origine. Kleinberg propose une analogie avec un réseau social, où la grille serait l'ensemble des positions géographiques des individus et les liens aléatoires les connaissances amicales d'un individu dont lui seul connaît les positions. Comme les liens supplémentaires sont distribués selon une loi reliée à la structure sous-jacente de l'information globale, ils stockent cette information dans un certain sens, puisque les décisions de routage, prises en fonction des liens aléatoires, sont corrélées à la position des nœuds dans la structure sous-jacente.

Un résultat récent de Clauset et Moore vient par ailleurs renforcer la pertinence du modèle de Kleinberg et de sa distribution. Clauset et Moore ont simulé un processus de reconnexion dynamique des liens dans l'anneau que l'on peut comparer, par exemple, au comportement d'un internaute qui créerait un nouveau marque-page vers une page web lorsque la recherche d'une page a pris un temps supérieur à un certain seuil. Ils ont alors observé expérimentalement que la distribution des liens convergeait vers une loi harmonique (la distribution des arcs aléatoires du modèle de Kleinberg pour l'anneau) tandis que l'espérance de longueur des chemins calculés par l'algorithme de routage glouton convergeait vers $O(\log^2 n)$, pour n nœuds.
Du point de vue informatique, le modèle de Kleinberg présente des caractéristiques algorithmiques intéressantes que ne présentaient pas les modèles précédents (modèles d'Erdös et Rényi, et de Watts et Strogatz). Ajoutons que si le modèle de Kleinberg reproduit bien la propriété de petit monde navigable, une critique fréquente est qu'il ne reproduit pas la distribution des degrés observée sur les réseaux réels, puisque chaque sommet a un degré sortant constant.

2.4 Quelques exemples des graphes-réels

1. **Graphe de Connaissances :** Parmi les grands graphes-réels, les plus intéressants sont ceux dont les nœuds représentent l'être humain. Une arête se trace entre A et B si A a une relation avec B (A connaît B, A travaille dans la même entreprise que B, A est un co-auteur de B ,...). Ces graphes sont devenus populaires dans les années 1990 avec le slogan «six degrés de séparation» qui exprime l'idée que le graphe des connaissances humaines a un diamètre moyen de 6, (c'est-à-dire la moyenne du nombre minimal d'arêtes à couvrir – d'une connaissance à une autre connaissance - afin de passer d'un nœud à un autre est de 6).

2. **Graphes des appels téléphoniques aux États-Unis** : Les nœuds sont les numéros de téléphone aux États-Unis et une arête est tracée de A à B chaque fois que A demande B. Un graphe pareil construit pendant une période de 20 jours a 290 millions de nœuds et 4 milliards d'arêtes. Abello, Pardalos et Resende ont étudié un tel graphe pour une période d'un jour (plus de 50 millions de nœuds actifs et 170 millions d'arêtes).

3. **Le graphe de collaboration entre acteurs** : Les nœuds sont les 225.000 acteurs énumérés par le syndicat américain du film, et une arête est tracée de A à B si A et B ont joué ensemble dans le même film. Watts et Strogatz [34] ont mesuré le diamètre moyen de ce graphe qui est de 3.65.

4. **Le graphe neural des *Caenorhabditis elegans*** : *Caenorhabditis elegans* est un petit ver composé de 959 cellules dont 300 forment son réseau neuronal. Watts et Strogatz ont également mesuré le diamètre moyen du graphe de la *C. elegans* depuis ces 300 neurones qui est de 2,65. Même si le graphe de la *C. elegans* n'est pas très grand, car il ne dispose que de 300 nœuds, aujourd'hui, nous pouvons retracer l'histoire de chacune de ses 959 cellules depuis la première cellule (l'arbre généalogique de chacune de ses cellules), ce qui rend la *C . Elegans* très intéressante pour mener des études plus poussées concernant l'évolution de son graphe.

5. **Le graphe du World Wide Web** : Les nœuds sont les 800 millions de pages disponibles sur Internet, et une arête est tracée de A à B si un lien hypertexte à la page B apparaît dans la page A [81]. Albert, Jeong et Barabási [38] ont montré que le diamètre moyen du graphe web est de 19 (c'est-à-dire, la moyenne du nombre minimal de liens à suivre pour aller d'une page à une autre est de 19) et que le WWW dispose d'une distribution selon une loi de puissance(2000) [87] ; mais c'est Lada Adamic [39][40] qui a vérifié la propriété de clustérisation du world wide Web tout en opérant sur plusieurs sites.

6. **Les graphes lexicaux :** Il existe plusieurs types de réseaux lexicaux, suivant la nature de la relation sémantique qui définit les arcs du graphe (les sommets représentant les unités lexicales d'une langue **–** de quelques dizaines de milliers à quelques centaines de milliers d'éléments, suivant la langue et la couverture du corpus utilisé).

   Plusieurs articles ont analysé les structures des différents graphes-réels mais peu sont les articles qui ont traité les graphes d'origine linguistique. (On mentionne l'exception notable du graphe de Wordnet analysé dans [41] et dans [83]).

   Les trois principaux types de relations utilisées sont les suivantes :
   - **Relations syntagmatiques**, ou plutôt de cooccurrence ; on construit une arête entre deux mots si on les trouve dans un grand corpus au voisinage l'un de l'autre (typiquement à une distance maximale de deux/trois mots ou plus).
   - **Relations paradigmatiques**, notamment de synonymie ; à partir de bases de données lexicales, comme le célèbre WordNet (Fellbaum 1999), on construit un graphe dans lequel deux sommets sont reliés par une arête si les mots correspondants entretiennent une relation synonymique [42] [http://www.crisco.unicaen.fr ou http://dico.isc.cnrs.fr/dico/fr/chercher]
   - **Relations de proximité sémantique** ; il s'agit de relations moins spécifiques qui peuvent prendre en compte à la fois l'axe paradigmatique et l'axe syntagmatique. Gaume [46][47] a construit un graphe du lexique du Français, en définissant les arêtes de la manière suivante : il construit une arête entre un verbe A et B si l'un est dans la définition de l'autre dans un dictionnaire général (construction du même type que celle appliquée par Veronis & Ide); comme une entrée de dictionnaire général comporte la plupart du temps des définitions, des exemples, des

synonymes, et même des antonymes, les arêtes sont alors étiquetées par le type de relation qu'elles représentent : on peut donc, selon les besoins, restreindre le graphe à certaines combinaisons de relations : syntagmatiques et/ou paradigmatiques et/ou même logico sémantiques.

Tous ces graphes sont à l'évidence de type RPMH. Outre leur intérêt propre dans l'étude du lexique, ils peuvent donc aussi nous permettre de mieux comprendre les grands graphes de terrain dans leur ensemble.

De manière générale, si les définitions d'un dictionnaire sont porteuses de sens, c'est au moins par le réseau qu'elles tissent entre les mots qui en sont les entrées. Exploiter ce réseau de type petit monde a permis d'en tirer parti de l'hypothèse que les zones de densité fortes en arcs (les agrégats) identifient des zones de sens proches. Gaume, Veronis, Ide,.. ont illustré leurs approches sur deux types de dictionnaires : un dictionnaire de langue, le Grand Robert et DicoSyn un dictionnaire de synonymes constitué de **sept dictionnaires** classiques (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert) dont ont été extraite les relations synonymiques[2].

Les dictionnaires sont représentés par des graphes dont les sommets et les arêtes peuvent être définis de multiples façons. L'une d'entre elles consiste à prendre pour sommets du graphe les entrées du dictionnaire et d'admettre l'existence d'un arc d'un sommet A vers un sommet B si et seulement si le mot B apparaît dans la définition du mot A. Les graphes ainsi obtenus sont des RPMH typiques [45].

[2] Ce premier travail de fusion, effectué à l'Institut National de la Langue Française (aujourd'hui ATILF: http://www.atilf.fr/) a produit une série de fichiers ; les données de ceux-ci ont été regroupées et homogénéisées au sein du laboratoire CRISCO http://elsap1.unicaen.fr/ par un important travail de correction (par adjonctions ou suppressions de liens synonymiques) sur le fichier final (Ploux & Victorri 1998).

Chapitre 3

Approches existantes de la géométrisation[1] des graphes lexicaux. Etat de l'art

Grâce au développement de nouvelles technologies informatiques, les recherches en traitement automatique des langues s'appuient de plus en plus sur des ressources lexicales à grande échelle (corpus, ontologies, dictionnaires électroniques …). Ces ressources permettent d'obtenir de façon automatique des informations sémantiques sur les mots et les relations qu'ils entretiennent entre eux. Ces relations peuvent être représentées naturellement par des réseaux lexicaux. Les sommets en sont les mots d'une langue [88][91].

Dans ce chapitre, nous passons en revue des différentes approches de géométrisation de ces graphes lexicaux en introduisant les notions de cliques, d'espaces sémantiques, de distances sémantiques, de composantes N-connexes et de composantes de sens.

3.1 Espaces sémantiques et notion de cliques

3.1.1 Utilisation des cliques

Le but de l'exploration présentée ici est d'avoir accès à la topologie sous-jacente du graphe étudié. Nous nous appuyons pour cela sur sa structure de graphe petit-monde à invariance d'échelle. Ravasz et Barabási [45] ont montré qu'un fort coefficient de clustering associé à une structure à invariance d'échelle détermine une combinaison originale de modularité et d'organisation hiérarchique. Un coefficient de clustering élevé traduit la présence de nombreux clusters, qui sont très interconnectés. Ces clusters s'associent entre eux pour former des groupes plus grands mais moins connectés, qui se combinent à nouveau pour former des clusters encore plus gros et encore moins connectés. Et ainsi de suite.

L'invariance d'échelle fait que le nombre de sommets très connectés par rapport aux autres sommets est constante à chaque niveau de clusterisation. Cela fait qu'aucun sommet ne peut être vu comme dominant les autres. L'unité structurelle est donc le cluster: à la base nous avons des clusters vraiment petits et très connectés, qui deviennent de plus en plus gros et de moins en moins interconnectés. De plus, les groupes de sommets se recoupent à tous les niveaux. L'unité d'étude de la topologie du graphe doit donc être un petit groupe de sommets très fortement connectés les uns aux autres.

Ploux et Victorri [42] ont utilisé la clique comme unité d'étude de leur graphe. Une clique est en effet un ensemble de sommets deux à deux connectés le plus grand possible. La figure 4 présente un extrait du graphe adjectival. Ce graphe présente ainsi 3 cliques : < *bas ;brutal ; mauvais ; méchant* > (On ne peut pas ajouter *faible* à cette clique car il n'est pas synonyme de *brutal*), < *bas ; faible ; mauvais* ; *méchant* > et < *fier ; intraitable ; méchant ; sauvage* >. La clique est un bon candidat de cluster de base dans la structure du graphe. Le taux de clusterisation élevé dans un graphe petit-monde assure la présence d'un grand nombre de cliques. L'utilisation des cliques comme unité minimale de groupements de sommets a par ailleurs fait ses preuves dans l'exploration de petits graphes de synonymie, constitués par un mot vedette et l'ensemble de ces synonymes. [42][48]

[1] Ce terme est emprunté depuis l'article de Venant [90]

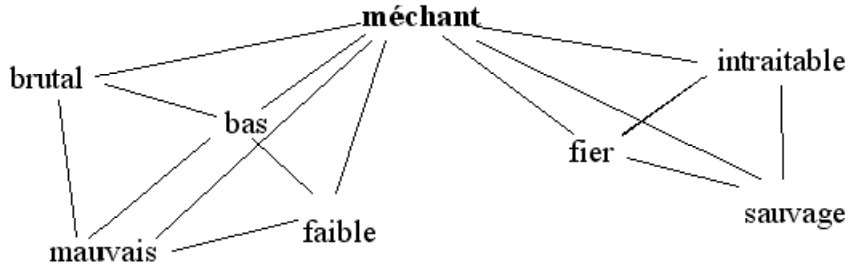


Figure 4. *Un extrait du graphe adjectival*

3.1.2 Espaces sémantiques

La structure hiérarchique que nous venons de décrire peut, la plupart du temps, être mise en relation avec un espace géométrique sous-jacent. Prenons l'exemple des réseaux sociaux, archétypes des graphes petit-monde à invariance d'échelle. Ces réseaux sont modélisés par des graphes dont les sommets sont des personnes. Les arêtes du graphe modélisent une relation sociale : être amis, connaître le prénom de l'autre personne, travailler ensemble… La structure d'un tel réseau est clairement reliée à un espace géographique sous-jacent, présentant une structure hiérarchique duale de celle du graphe. Les petits clusters de personnes très connectées correspondent à des petites zones géographiques où les gens se rencontrent régulièrement (bureaux, lieux d'habitation…). Ces lieux se regroupent ensuite en des zones géographiques plus étendues (société, quartiers, villages…) elles-mêmes très interconnectées (holdings, villes, communauté de communes…). L'hypothèse que nous faisons est que, à tout graphe de type petit-monde à invariance d'échelle, nous pouvons associer un espace conceptuel sous-jacent, de nature souvent plus abstraite qu'une simple carte géographique, dont la topologie peut révéler celle du graphe, plus difficile d'accès. Dans l'exploration du graphe adjectival, il s'agira donc de construire l'espace sémantique associé au graphe. Nous pouvons considérer en première approximation qu'une clique correspond à un emploi adjectival et que ce sont donc les cliques qui constitueront les points de l'espace sémantique [89].

3.1.3 Une métrique pour l'espace des cliques

On peut définir l'espace sémantique adjectival comme l'espace euclidien engendré par les adjectifs. Chaque clique y est représentée par un point dont les coordonnées sont calculées en fonction des synonymes qu'elle contient : soient $a_1, a_2, \dots, a_n$ les adjectifs, et $c_1, c_2, \dots, c_p$ les cliques, l'adjectif a_i correspond au $i^{\text{ème}}$ vecteur de base de cet espace, et la clique c_k à un point dont les coordonnées x_{ki} valent 0 ou 1 suivant que l'adjectif correspondant appartient ou non à la clique : $x_{ki} = 1$ si $a_i \epsilon\, c_k$ et $x_{ki} = 0$ si $a_i \notin c_k$.
La distance entre deux cliques c_k et c_l est alors donnée par la *métrique canonique* sur cet espace euclidien, définie de la façon suivante :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} (x_{ki} - x_{li})^2$$

Ploux et Victorri (98), dans leurs travaux sur la construction d'espaces sémantiques, montrent par l'analyse de quelques exemples que cette distance se révèle totalement inadéquate. Ils expliquent cela par le fait que cette distance donne le même « poids » à tous les synonymes, et qu'elle traite de la même manière toutes les cliques, quel que soit leur cardinal. « Or certains synonymes peuvent recouvrir une grande partie des emplois […], alors que d'autres sont plus « spécifiques », dans la mesure où ils ne s'appliquent qu'à un ensemble très restreint d'emplois. De plus, certaines cliques possèdent beaucoup plus d'éléments que d'autres. Ces différences doivent être prises en compte dans la définition de la distance, si l'on veut représenter correctement la proximité sémantique de deux cliques. » Ils proposent donc

d'utiliser une métrique bien connue en analyse de données *la métrique du* χ^2: deux cliques c_k et c_l étant données, la distance entre les deux est donnée par :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_{.i}} \left(\frac{x_{ki}}{x_{k.}} - \frac{x_{li}}{x_{l.}} \right)^2$$

$$avec\ x_{.i} = \sum_{j=1}^{p} x_{ji}\ , x_{k.} = \sum_{i=1}^{n} x_{ki}\ , x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji}$$

Cette métrique possède l'avantage d'une part de pondérer chaque synonyme en fonction du nombre de cliques dans lequel il intervient (plus un synonyme apparaît dans des cliques différentes, moins il est spécifique et moins son rôle dans la discrimination des sens de l'unité est important), et d'autre part de diviser les coordonnées de chaque clique par son nombre d'éléments : le point représentant la clique est d'autant plus proche de l'origine que la clique correspondante comporte plus de synonymes. La distance du χ^2 confère donc à l'ensemble des cliques une structure géométrique qui semble respecter la notion intuitive de proximité entre emplois.

Exemple: Ploux et Victorri [42] ont appliqué leur méthode à l'unité lexicale *maison*, et se sont limités aux deux premiers axes de l'analyse en composantes principale. Chaque point de la représentation obtenue représente une clique. Ils ont noté aussi que les cliques se distribuent le long d'une courbe, allant de sens de type «établissement commercial », jusqu'à des sens exprimant la notion de descendance, en passant par une valeur centrale de lieu d'habitation, représentée par une majorité de cliques.

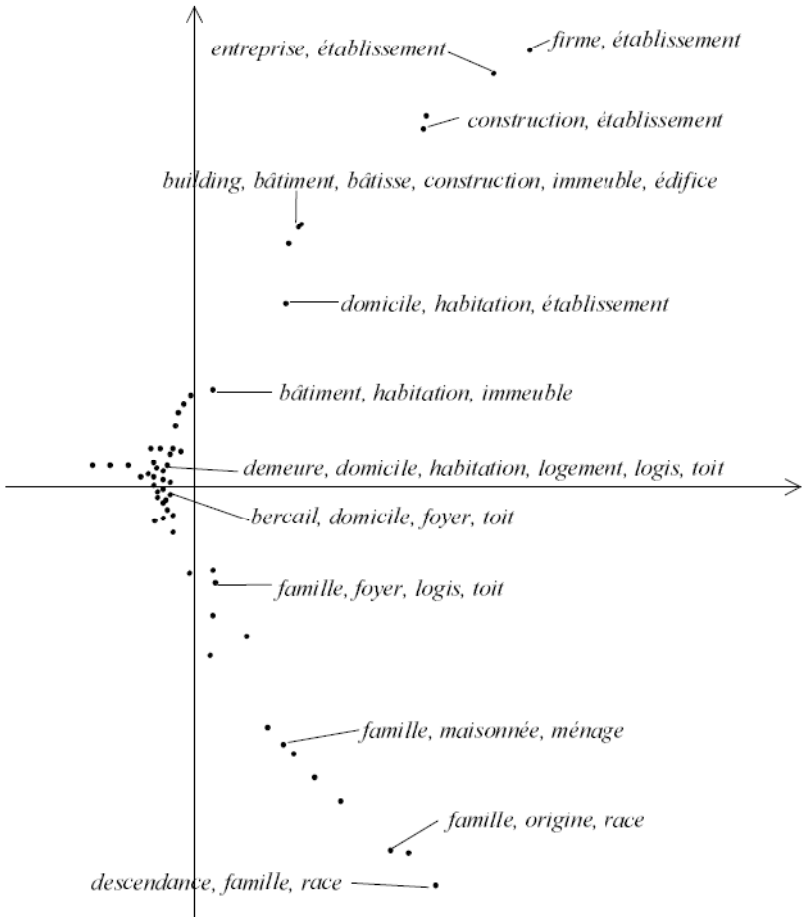


Figure 5 : *Représentation des cliques associées à* ***maison***

Ils ont relevé aussi plusieurs caractéristiques :

- Les 2 axes sont susceptibles de recevoir une interprétation sémantique.
- L'intérêt d'avoir une représentation qui épouse approximativement la forme d'une courbe. Cela signifie que si l'on ne cherche pas de finesse excessive, nous pouvons

parcourir l'ensemble des sens de maison à l'aide d'un seul paramètre. Autrement dit l'espace sémantique associé à maison est en première approximation un espace unidimensionnel.

3.2 Extraction des synonymes en utilisant les chaînes de Markov

Gaume [47][49][50] propose une méthode stochastique pour la métrologie et la cartographie des structures locales et globales des RPMH. La méthode *Prox* consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question. Des particules se baladent aléatoirement de sommets en sommets dans le graphe en empruntant ses arcs. Ce sont les dynamiques des trajectoires des particules qui cartographient les propriétés structurelles des graphes étudiés.
Pour illustrer leur approche, il a mis au point un logiciel[2] pour cartographier la forme de sens dans les RPMH.

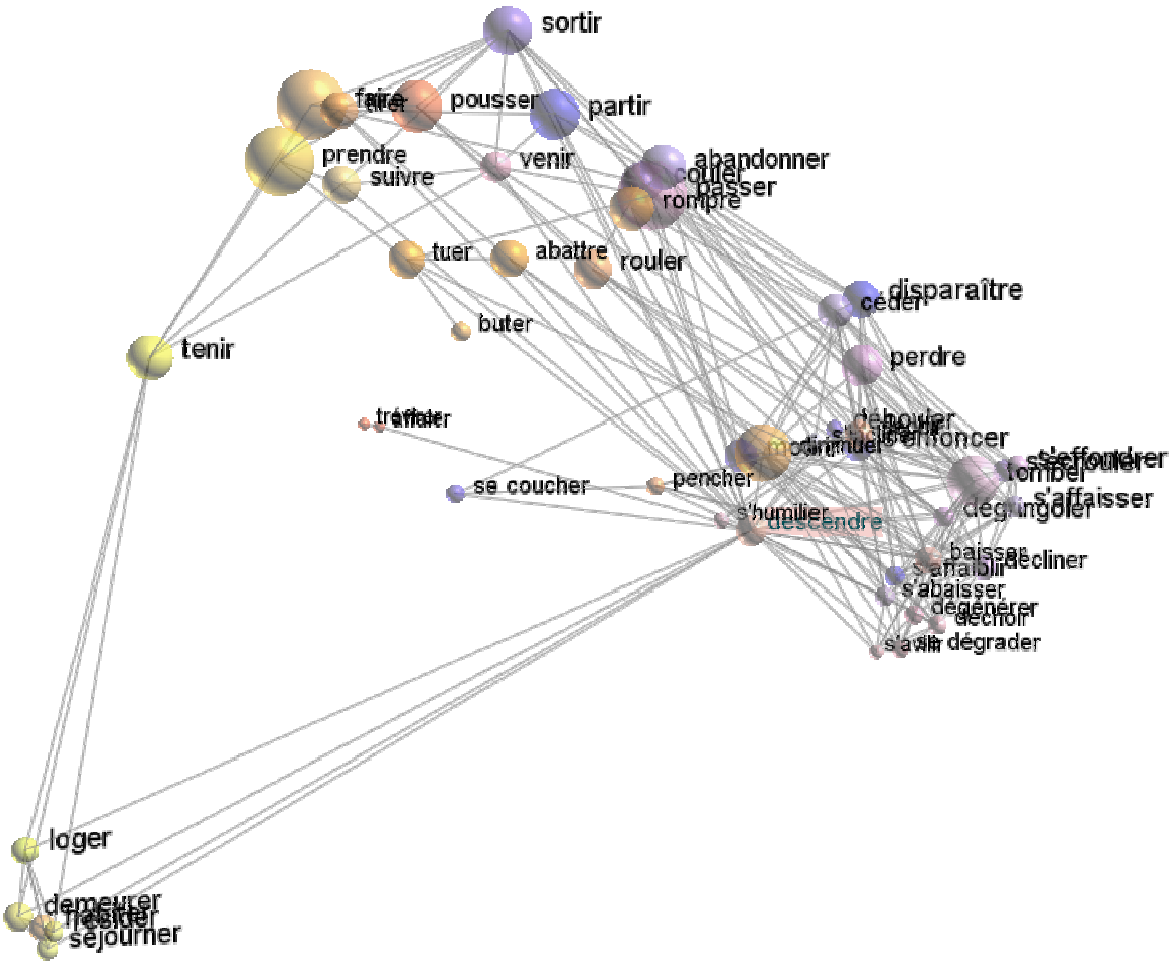


Figure 6 : *Représentation de l'espace sémantique par prox associées à* ***descendre***

3.3 Extraction des synonymes en utilisant une distance sémantique sur un dictionnaire

Muller, Hathout et Gaume [51] ont davantage travaillé sur les Chaînes de Markov et ont présenté une méthode pour exploiter la structure des graphes sémantiques et calculer une

[2] http://prox.irit.fr

distance entre les mots. Cette distance a été utilisée pour isoler les synonymes candidats pour un mot donné.

L'idée comme présentée à la section précédente est de voir le graphe comme une chaîne de Markov dont les états sont les nœuds du graphe et les transitions sont ses arêtes, évalués par des probabilités. Ils supposent que la distance moyenne parcourue par une particule entre deux nœuds est une indication de la distance sémantique entre ces nœuds.

Formellement, si G = (V,E) est un graphe réflexif (chaque nœud est connecté à lui-même) avec |V| = n, ils notent [G] la matrice d'adjacence n × n de G telle que [G]i,j est non nulle s'il existe une arête entre les nœuds i et j et 0 autrement.
La première étape consiste à transformer la matrice en une matrice Markovienne. Ils notent $[\hat{G}]$ la matrice Markovienne de G, telle que

$$[\hat{G}]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in V}([G]_{r,x})}$$

La somme de chaque ligne de G est différente de 0 car le graphe est réflexif.

Ils notent aussi $[\hat{G}]^i$ la matrice $[\hat{G}]$ multipliées i fois par elle-même. Ils définissent une fonction PROX(G, i, r, s) comme $[\hat{G}]^i_{r,s}$, c'est en fait la probabilité qu'une particule aléatoire qui vient de quitter le nœud r va se déposer sur le nœud s après i étape(s). Ils notent que le choix de i dépend du graphe et est déterminé empiriquement.

Leur expérimentation les a conduits à conclure que leur méthode est capable de récupérer un grand nombre de synonymes directs et indirects dans la définition des mots, et que leur méthode se base sur un choix arbitraire qu'il faut exploiter davantage.

3.4 Extraction de composantes N-connexes dans les graphes de dictionnaires de verbes

Awada et Chebaro [52] ont proposé une méthode pour résoudre le problème de la polysémie en définissant :

- La synonymétrie comme notion de mesure de la proximité de sens entre deux verbes d'une langue ainsi permettant de détecter et d'éliminer des utilisations métaphoriques de certains verbes. Ils postulent que la mesure de cette proximité de sens entre deux verbes donnés doit être déterminée en examinant uniquement le plus court chemin entre les deux sommets correspondants dans le graphe. Cependant l'existence d'un chemin, même très court, entre deux sommets ne traduit pas forcément le fait qu'ils aient le même sens. En effet, ils montrent par un exemple que la présence d'un verbe polysémique sur ce chemin suffit pour corrompre la relation de synonymie détectée entre les deux verbes. La synonymie entre différentes entrées d'un dictionnaire se traduit dans le graphe correspondant par des concentrations de relations (arcs) entre tous les verbes (sommets) ayant le même sens. Il en résulte que les verbes synonymes sont regroupés dans la même composante connexe et donc il devrait y avoir équivalence entre la notion de sens et celle de composante connexe dans le graphe. Cette hypothèse aurait été vraie en l'absence de verbes polysémiques du dictionnaire. La présence d'un verbe polysémique se traduit, dans le graphe, par l'appartenance de son sommet à toutes les composantes connexes correspondant à ses différentes acceptions. Un tel sommet joue ainsi le rôle de passerelle entre différentes composantes connexes en les unissant à l'intérieur d'une même composante connexe ce qui permet parfois d'interpréter 2 mots non-synonymes comme des synonymes. Ils ont séparé les composantes sur le graphe tout en introduisant la notion de N-connexité.
- Ils définissent la N-connexité comme un nouveau critère mathématique de regroupement des verbes synonymes : Un sous graphe forme une composante N-

connexe si et seulement si chaque sommet de ce sous graphe est en liaison directe (un arc) avec au moins N sommets du même sous graphe. Tout sommet du graphe vérifiant cette propriété appartient à la composante N-connexe.

Ils reformulent la notion de la synonymie en disant que deux verbes sont synonymes s'ils appartiennent à la même composante N-connexe du graphe de dictionnaire. Leurs expérimentations les ont amenés à noter qu'une valeur élevée de N permet d'isoler de vrais synonymes alors qu'une faible valeur de N permet l'introduction des verbes métaphoriques dans une même composante N-connexe. Cependant la valeur de N repose sur un choix empirique et dépend du verbe examiné.

3.5 Regroupement de synonymes en composantes de sens dans un dictionnaire

Awada [53] a de même proposé une autre méthode en vue de résoudre le problème de la polysémie mais cette fois ci il a réparti les synonymes d'un verbe en groupes appelés composantes de sens correspondant chacun à une acception du verbe. Etant donné que le dictionnaire est représenté par un graphe, le regroupement des verbes-synonymes sur un même circuit implique que ces verbes doivent appartenir à la même composante de sens. Voulant regrouper des synonymes d'un verbe donné en une acception tout en garantissant l'existence d'un nombre minimal de circuits les contenant, Awada a introduit cette notion de seuil d'acception qui joue le rôle d'un filtre qui empêche de regrouper certains synonymes d'un verbe donné dans une même composante de sens, et permet par opposition, d'en regrouper d'autres.

Ces expériences ont montré qu'une valeur faible de ce seuil fait entrer dans la même composante de sens des verbes qui ont peu ou pas assez de relations entre eux en tant que synonymes du verbe de départ car peu de circuits les réunissent. D'un autre côté, une valeur accrue de ce seuil a pour effet d'empêcher le regroupement de verbes pouvant correspondre à une même acception, voire d'éliminer carrément certains verbes qui seraient ainsi considérés comme des synonymes non acceptables du verbe de départ.

Cependant, il s'est confronté à une difficulté inhérente liée au choix du seuil d'acceptation. C'est pour cette raison qu'il a minimisé le rôle du seuil d'acceptation en le combinant à un autre facteur qui est la longueur du circuit. En effet, plus un circuit est long, plus il y a de chance d'y trouver des verbes, et donc des verbes polysémiques et par conséquent de mélanger différentes composantes de sens. Cependant, la prise en compte de circuits trop courts uniquement a pour effet de scinder une même composante de sens en plusieurs.

Il a finalement proposé sa nouvelle reformulation du principe de regroupement : "le regroupement de synonymes d'un verbe donné en une acception se fait quand il existe au moins un certain nombre de circuits de longueur inférieure ou égale à une longueur donnée".

Chapitre 4

Interfaces de visualisation. Etat de l'art

Les interfaces de visualisation de l'information étudie la représentation visuelle des grandes collections d'information non numérique, tels que les fichiers et les lignes de code dans les systèmes informatiques [55], et l'utilisation des techniques graphiques pour aider à comprendre et analyser des données [56]. Les interfaces de visualisation traitent des ensembles de données abstraits, tels que les textes non structurés ou des points dans un espace de grande dimension, qui n'ont pas une structure géométrique 2D ou 3D [54] [57].

Dans ce chapitre, nous passons en revue des différentes interfaces de visualisation pour la recherche et la classification des documents ainsi que celles conçues pour visualiser la structure du World Wide Web.

4.1 Interfaces de visualisation pour la recherche et la classification des documents

Diverses classifications des interfaces de visualisation ont été proposées comme dans [58] qui traite des visualisations pour la RI ou encore [59] et [60] qui effectuent une taxonomie des différentes techniques de visualisation générale d'informations. Nous ne présentons dans cette section que la partie de la classification proposée par Zamir [58] qui traite du domaine de notre étude.

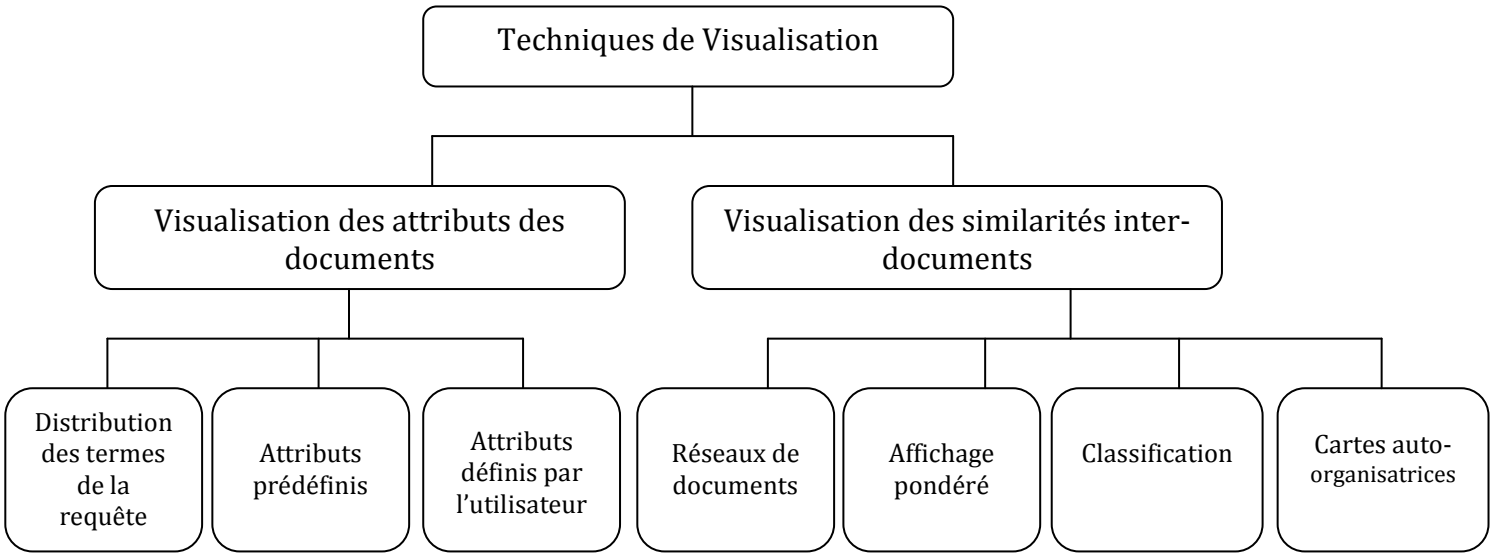


Figure 7. *Classification des techniques de visualisation [58]*

Cette classification met en évidence deux types de techniques selon leur but :

- les visualisations des attributs des documents (informations issues des documents). Il y a trois sous-catégories de techniques.
  - la distribution des termes de la requête. Elle permet de savoir comment chaque mot-clé utilisé dans la requête est réparti dans les documents,
  - les attributs prédéfinis. Elle permet de montrer la relation qu'a le document avec des attributs tels que la taille, l'auteur, etc.,
  - les attributs formulés par l'utilisateur. Elle permet de montrer la relation qu'a le document avec des critères choisis par l'utilisateur (requête par exemple…),
- les visualisations de similarité inter-documents. Il y a quatre sous-catégories.
  - les réseaux de documents. Les documents sont reliés entre eux selon leur similarité,

- « les affichages pondérés ». Les documents sont répartis visuellement selon des forces qui les repoussent ou les rapprochent des autres par rapport à leur similarité,
- les « classifications ». Ces visualisations représentent les documents sous forme de groupes de documents (par similarité de contenu, selon les liens hypertextes…),
- les cartes auto-organisatrices (ou SOM). Ces techniques permettent d'afficher sur une « carte » 2D les documents par rapport à leur similarité de contenu.

Cette classification ne permet toutefois pas une catégorisation globale des interfaces de visualisation. En effet, la catégorie de la visualisation des similarités inter-documents n'est pas relative aux éléments visualisés mais aux techniques employées. Pour pallier cela, Chevalier [61] a modifié cette classification en proposant une nouvelle qui fait abstraction délibérément des techniques de visualisation car leur nombre est très important et elles évoluent continuellement. Ainsi, la catégorie des visualisations des similarités inter-documents a été reconsidérée en s'appuyant sur la visualisation des documents les uns par rapport aux autres ou sur la visualisation des relations entre classes de documents. La branche concernant la visualisation a été conservée telle quelle (Figure suivante).

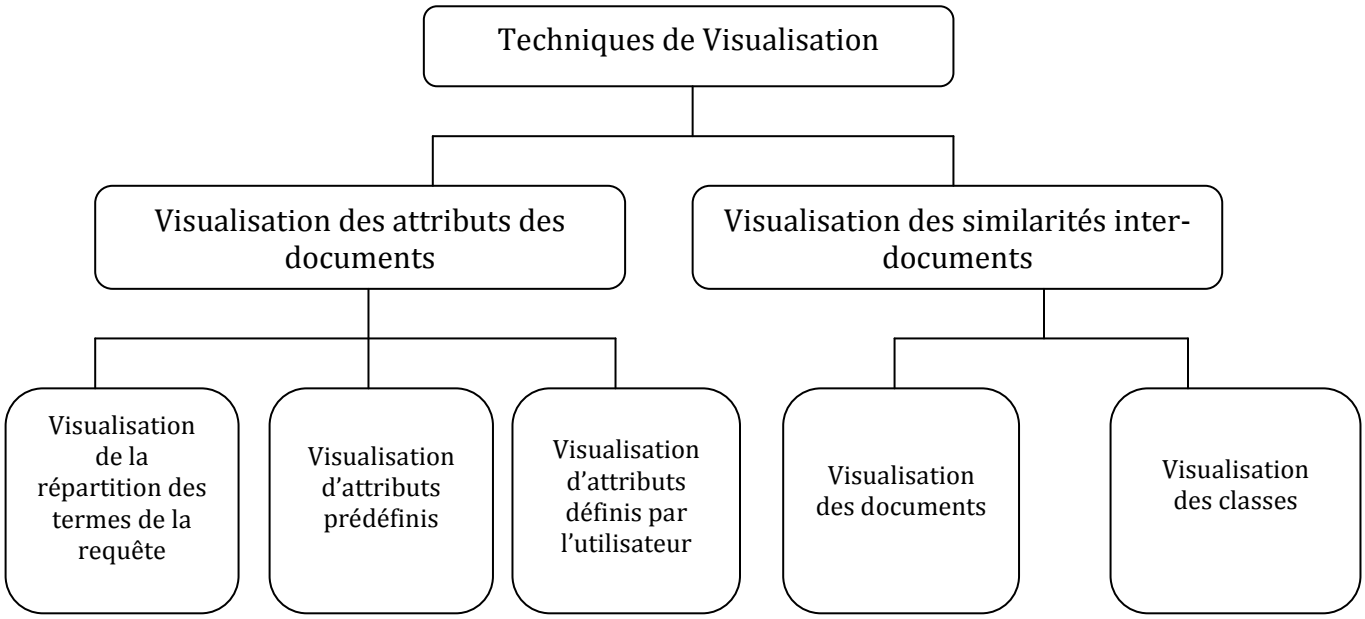


Figure 8. *Une nouvelle classification des techniques de visualisation proposée par Chevalier [61].*

Pour chacune de ces catégories, nous allons donner succinctement dans ce qui suit une description ainsi que des exemples d'interfaces de visualisation.

4.1.1 Visualisation des attributs des documents

4.1.1.a Répartition des termes de la requête

Cette technique de visualisation vise à présenter la répartition des différents mots-clés de la requête au sein des documents retrouvés. L'exemple le plus représentatif est *TileBars* [62] qui présente les résultats dans une liste de résultats mais accompagnée, pour chaque document, d'un bloc visuel représentant la répartition des termes de la requête. Le but recherché est de pouvoir suggérer simultanément à l'utilisateur :

- la longueur relative du document (taille du bloc),
- la fréquence des termes de la requête dans le document (par la nuance de gris),

- la distribution visuelle des termes de la requête dans le document, mais aussi dans les autres résultats.

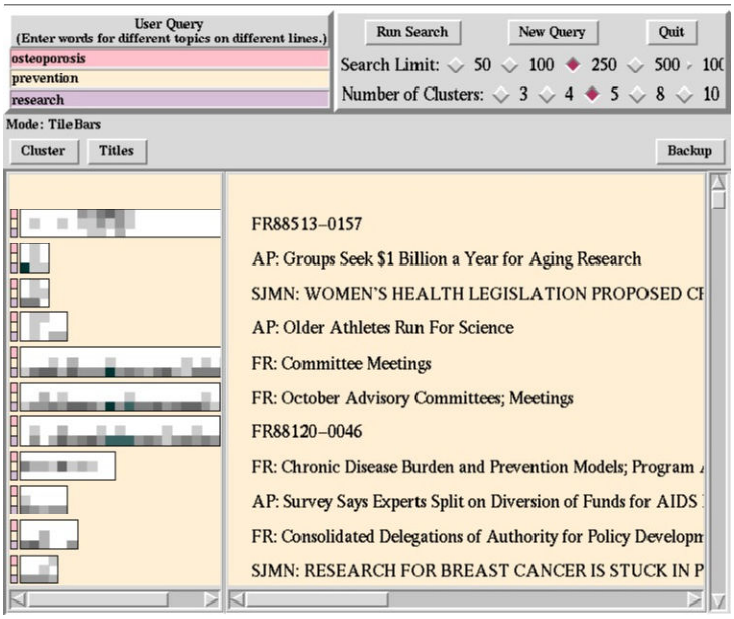


Figure 9. *TileBars*

Nous pouvons également citer comme exemple l'interface *Seesoft* [55].

4.1.1.b Attributs prédéfinis

Dans cette catégorie, les documents sont visualisés par rapport à des attributs prédéfinis comme des auteurs ou encore des thèmes prédéfinis par exemple. Ainsi *Cougar* [63] permet de visualiser un ensemble de documents par rapport aux thèmes qu'ils abordent. La technique utilisée repose sur les diagrammes de Venn (Figure 10). Chaque thème est représenté par un cercle et chaque document est situé dans les cercles correspondant aux thèmes qu'il aborde. Il n'y a au maximum que 3 thèmes sélectionnés dans cette interface car l'affichage proposé est en 2D. *InfoCrystal* [64] fait également partie de cette catégorie d'interfaces.

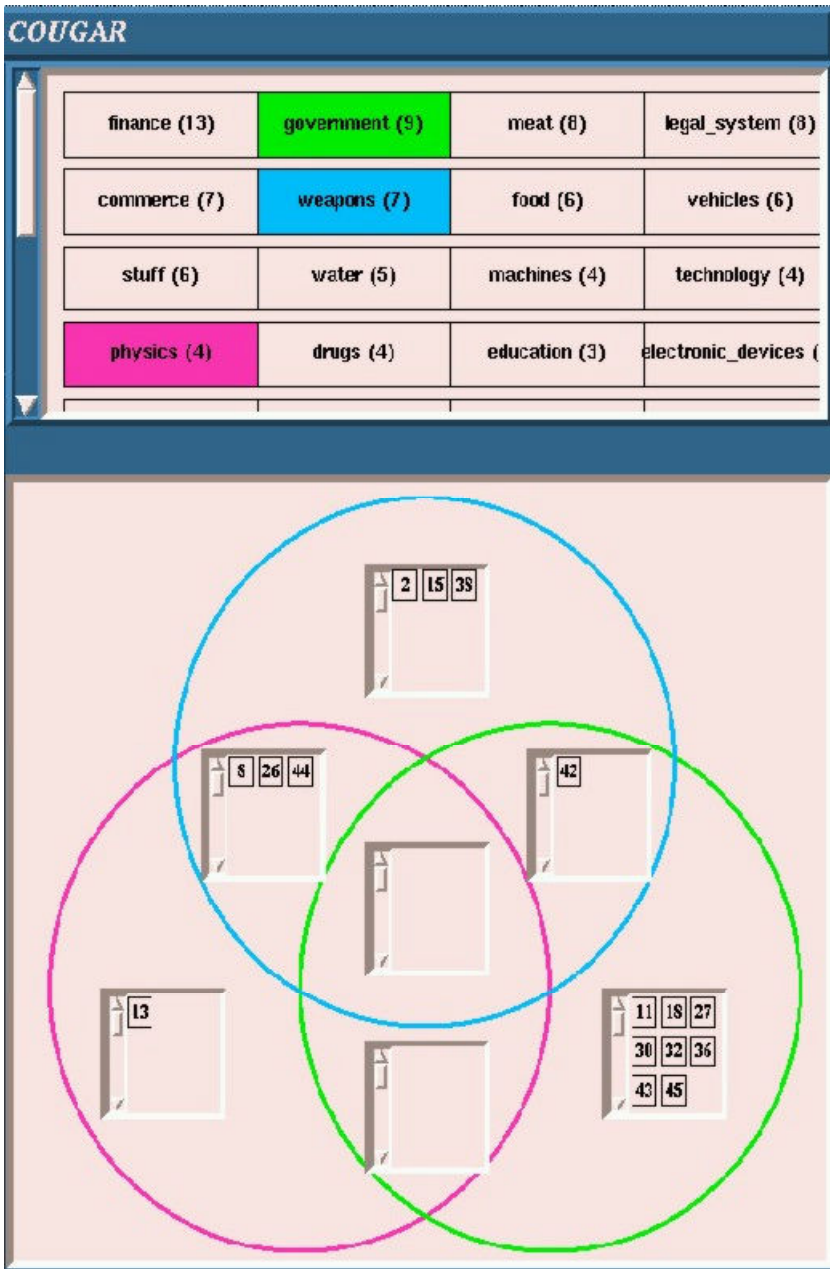


Figure 10. *Cougar*

4.1.1.c Attributs formulés par l'utilisateur

Dans cette catégorie, l'utilisateur peut choisir les attributs des documents suivant lesquels il souhaite visualiser les résultats. En général, ces attributs correspondent aux termes de la requête. Par exemple dans *Vibe* [65] et *VR-Vibe* pour sa version 3D (Figure 11), l'utilisateur peut visualiser les documents par rapport à des termes qui l'intéressent.

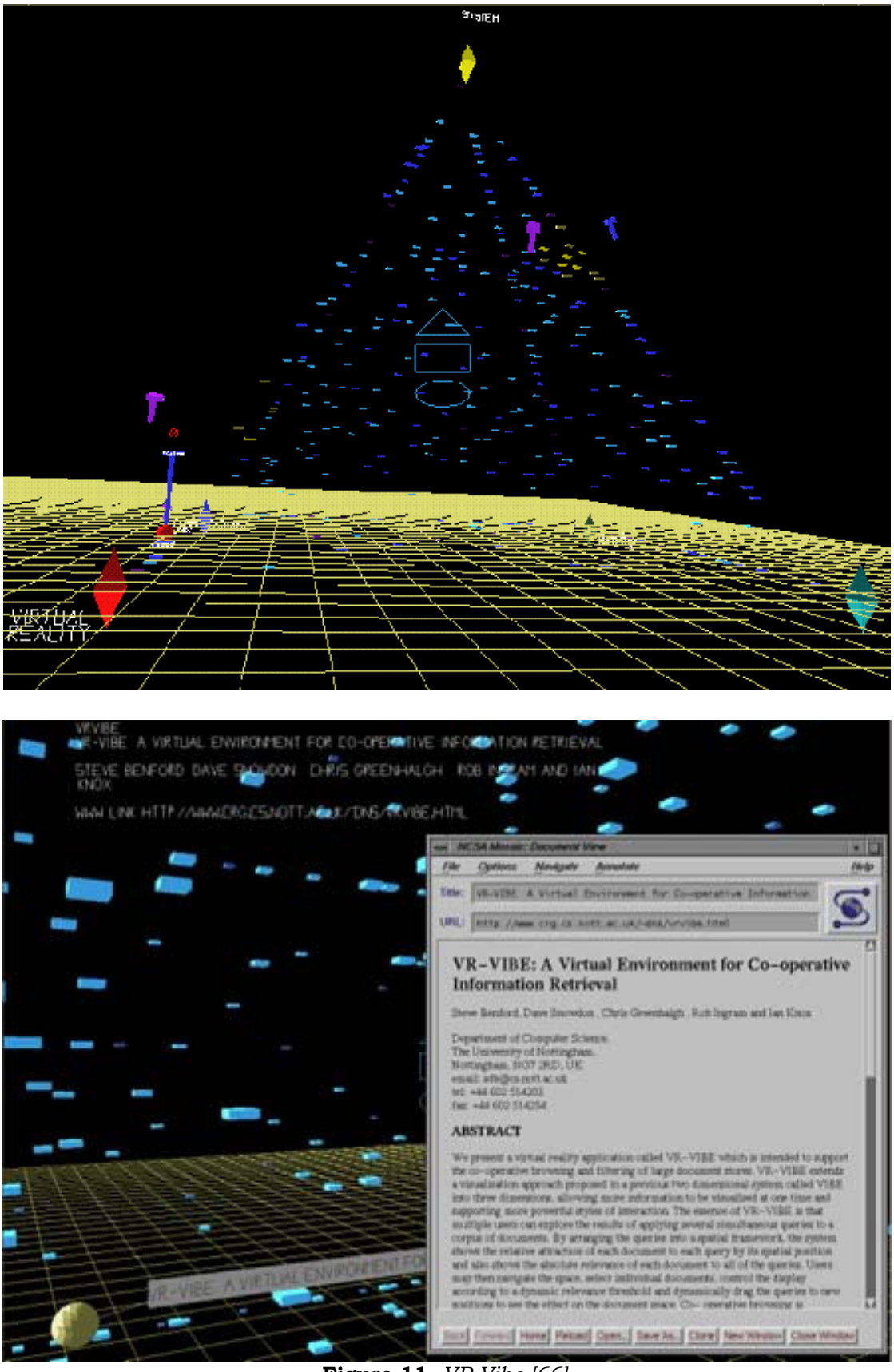


Figure 11. *VR-Vibe [66]*

Une autre approche intéressante est celle proposée par [67] au travers de l'interface *Three-Keywords Axes Display* (Figure 12). Cette interface permet de visualiser des termes ou combinaisons de termes sur des axes orthogonaux. Chaque document est représenté par un carré bleu dont la taille est proportionnelle à la taille du document. Dans chaque carré, l'importance d'un critère est représentée au moyen de lignes de couleurs plus ou moins longues proportionnellement à leur importance.

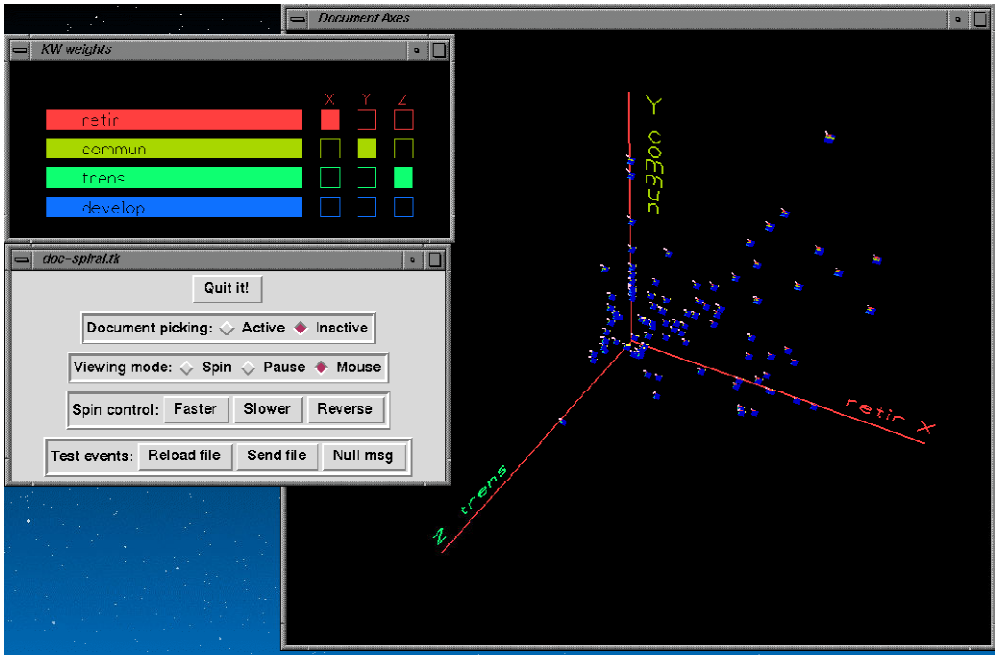


Figure 12. *Three-Keywords Axes Display [67]*

Par contre, dans *DocCube* [68] (Figure 13), la visualisation ne se base plus sur les termes de la requête mais sur des hiérarchies de concepts. Les documents sont visualisés au travers d'une interface en 3D relatives aux différentes entrées des branches des hiérarchies de concepts sélectionnées.

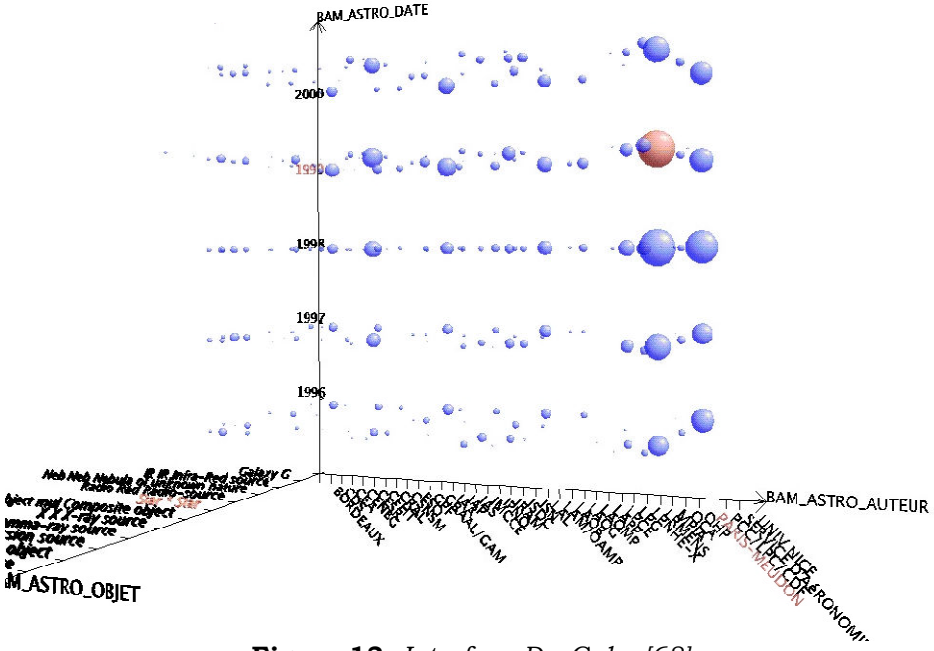


Figure 13. *Interface DocCube [68]*

4.1.2 Visualisation des relations inter-documents

Ces techniques de visualisation visent à mettre en évidence les relations entre les différents documents. Ces projets reposent essentiellement sur les notions de similarité et de classification.

4.1.2.a Relations document-document

La visualisation des relations entre documents permet d'apprécier les documents similaires pour un document donné.
L'approche la plus commune est la visualisation des documents au travers d'un réseau comme le propose l'outil de recherche *Kartoo*[1] dans lequel les liens entre les documents correspondent aux termes que les documents ont principalement en commun (Figure 14).

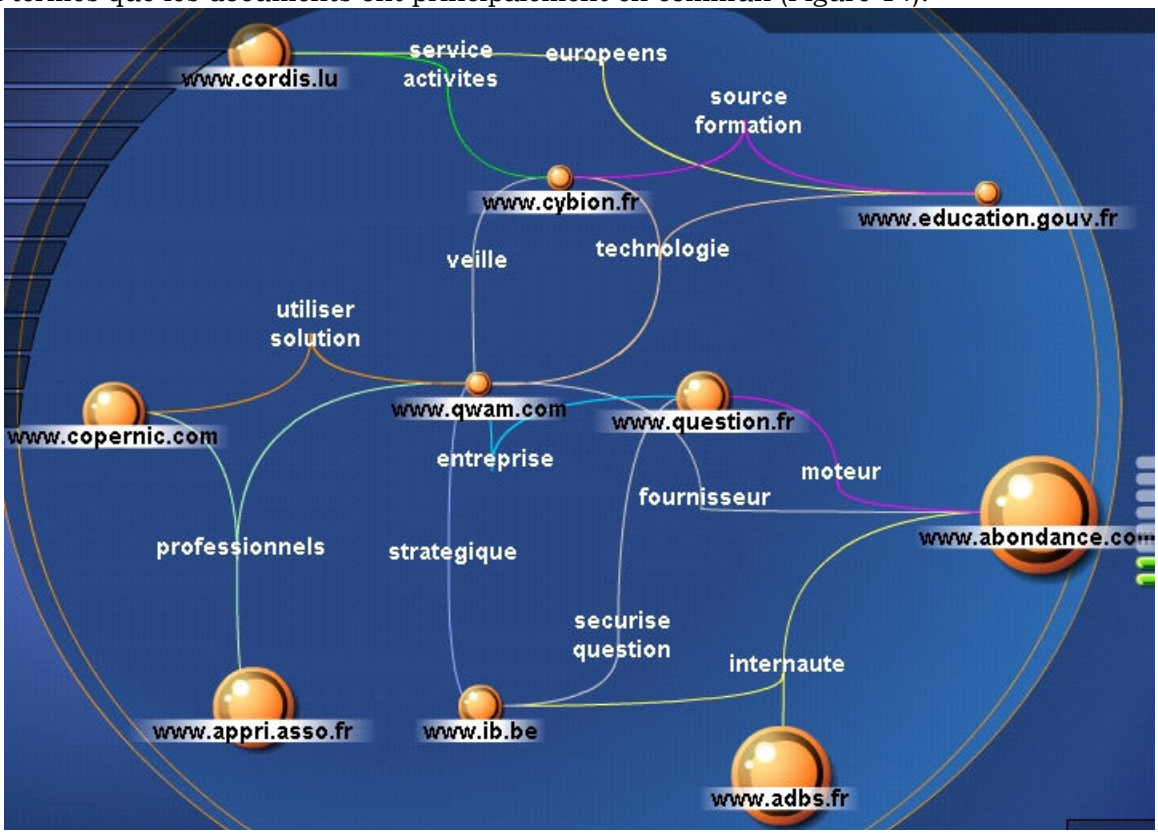


Figure 14. *Kartoo*

Une alternative à la représentation en réseau est l'utilisation d'un affichage pondéré des documents. Par exemple, l'utilisation d'un procédé de force d'attraction et répulsion permet de réorganiser les documents spatialement selon leur similarité. Grâce à ces affichages, l'utilisateur comprend à partir d'un document quels sont ceux qui sont proches de celui-ci. *Bead* [69] est un autre exemple d'interface basée sur ce type de représentation.

[1] http://www.kartoo.com

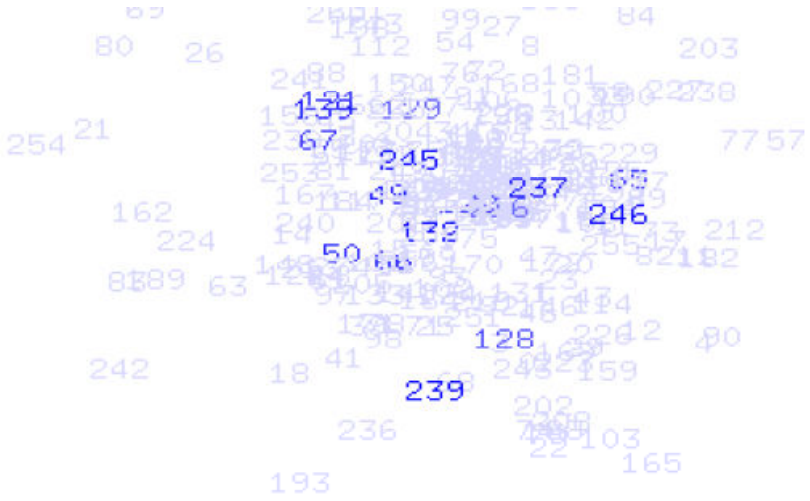


Figure 15. *Bead*

4.1.2.b Relations classes de documents – classes de documents

La classification des documents permet de regrouper les documents similaires permettant ainsi de diminuer le nombre d'éléments visualisés puisque l'on ne représente plus les documents de façon indépendante mais les classes de documents.

Différentes techniques peuvent être utilisées pour représenter les relations entre des classes de documents.

La représentation la plus simple des classes de documents sont les classes elles-mêmes. Par exemple, *Grouper* [70] (Figure 16) ou encore *Scatter/Gather* reposent sur une liste de résultats au sein de laquelle apparaissent des classes de documents (définies par un ensemble de groupes de mots caractérisant la classe de documents).

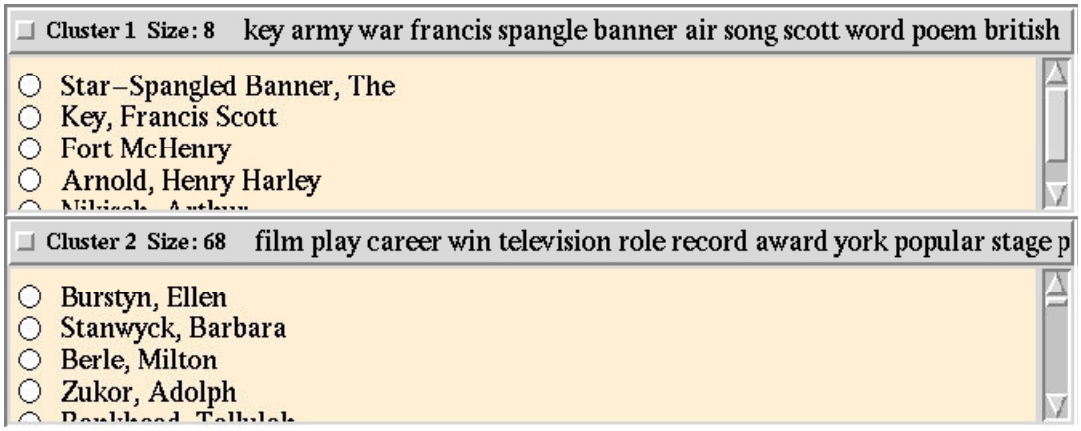


Figure 16. *Grouper*

Par ailleurs, la visualisation proposée par l'outil de recherche *Mapstan* (Figure 17) vise à présenter les classes de documents au moyen d'une métaphore d'un réseau urbain. Les quartiers (cercles) représentent des classes de documents similaires tandis que les « rues » représentent la similarité entre les quartiers. Ainsi, plus une rue est courte et épaisse, plus la similarité entre les quartiers sera importante.

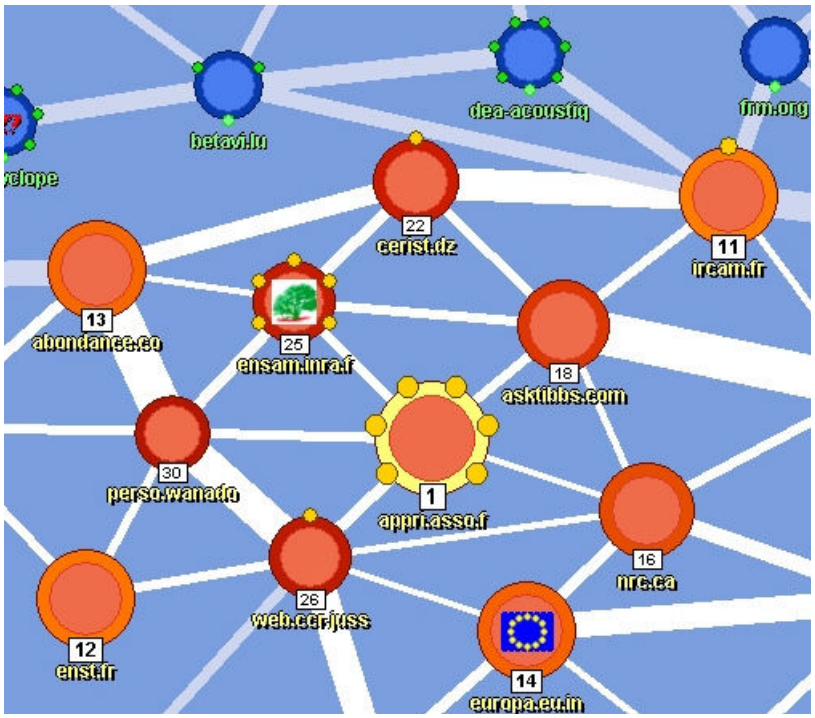


Figure 17. *MapStan*

Plutôt que de représenter les relations des classes sous la forme d'un réseau, dans les cartes auto-organisatrices [71], la carte est une grille où chaque case correspond à une classe de documents similaires. Les classes sont positionnées les unes par rapport aux autres suivant leur similarité respective. Lesteven propose une utilisation de ces cartes dans le domaine de l'astronomie tandis que *Websom* en propose une application sur le web.

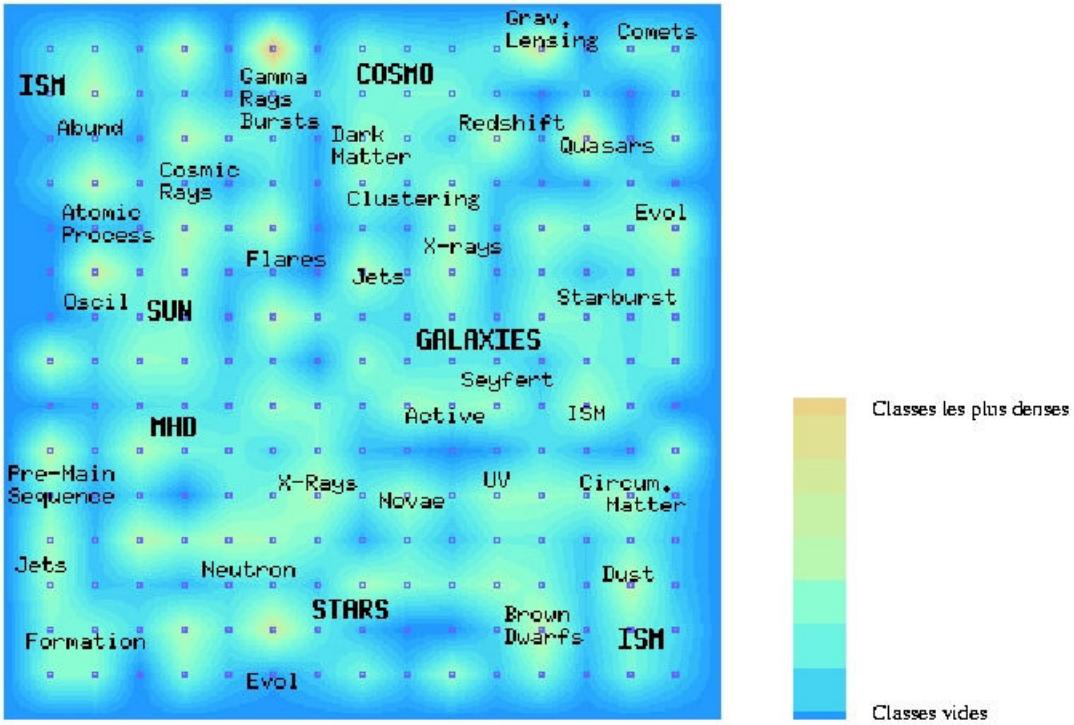


Figure 18. *Carte auto-organisatrice dans le domaine de l'astronomie*

Une autre approche de visualisation des relations entre classes de documents est celle de *Umap* (Figure 19). Elle repose sur les « Arbres de connaissance® » et propose une carte «

maritime » où chaque îlot correspond à un thème décrit par un ensemble de termes issus des documents retrouvés. Chaque élément de la carte représente un terme dont la position est calculée en fonction de sa corrélation avec les autres termes issus des documents. Cette visualisation permet de voir la cohérence du contenu des documents retrouvés au regard des termes qu'ils contiennent.

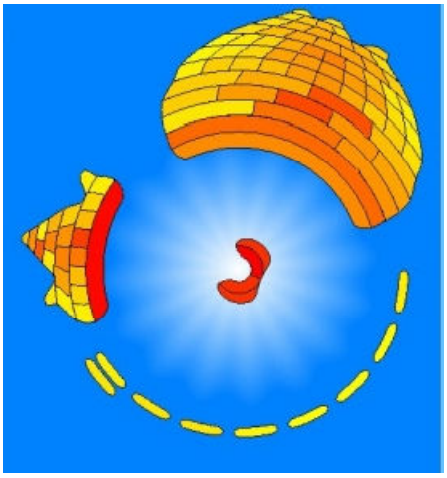


Figure 19. *Umap*

4.1.3 Comparaison des interfaces de visualisation

Dans cette section, nous proposons une évaluation comparative des interfaces présentées selon différents axes. En effet, une interface de visualisation utilise une certaine métaphore (représentation) des informations. Cette métaphore peut être construite à partir de plusieurs axes visuels qu'il est nécessaire de prendre en compte pour comparer les interfaces car ils conditionnent l'interprétation de la représentation graphique. Pour cette comparaison, nous avons fait le choix de présenter l'axe « espace de visualisation» et l'axe « couleur » car ils se retrouvent couramment dans la plupart des métaphores utilisées.

4.1.3.a L'espace de visualisation : Texte, 2D ou 3D ?

Au travers de la catégorisation des interfaces précédentes nous avons pu constater qu'il existe trois catégories de représentation des informations : L'affichage textuel, l'affichage en deux dimensions et l'affichage en trois dimensions. Ces représentations correspondent au type d'affichage utilisé et ce quel que soit le nombre de caractéristiques visualisées des informations. Le choix de la dimension utilisée n'est pas sans conséquence pour l'utilisation générale de l'interface.

Ce choix conditionne :

- la mise en œuvre du système,
- le nombre d'éléments visualisés,
- l'expérience requise pour que l'utilisateur puisse s'orienter dans l'espace des informations,
- l'attrait de l'utilisateur.

Intuitivement, la complexité de la mise en œuvre et le coût (en temps ou en ressources) d'une interface augmente avec le nombre de dimensions. Ainsi les interfaces textuelles sont plus simples à mettre en œuvre que le sont les interfaces en 2D. Du fait du nombre de calculs accrus, les interfaces 3D sont encore plus coûteuses que les interfaces 2D. Le coût (temps et ressources) peut être aujourd'hui amorti par les performances actuelles des microordinateurs personnels.

Par contre, le nombre d'éléments visualisés augmente avec le nombre de dimensions. Les interfaces textuelles permettent de visualiser un nombre restreint d'éléments comparativement

aux interfaces 2D et 3D. Les interfaces 2D quant à elles permettent de visualiser un nombre d'éléments inférieur aux interfaces 3D.

En ce qui concerne l'expérience requise, elle augmente avec le nombre de dimensions utilisées. En effet, le fait d'utiliser une interface textuelle ne nécessite quasiment aucun apprentissage car elle est naturellement interprétée tandis qu'une interface 3D peut nécessiter un apprentissage relativement important.

L'attrait visuel pour un utilisateur a également son importance dans l'utilisation d'une interface. Ainsi, Sebrechts [72] montre que l'utilisateur est plus sensible aux représentations graphiques qu'à une représentation textuelle et que pour une même tâche la représentation 3D permet d'obtenir globalement de meilleurs résultats. Pour résumer, nous pouvons dire que la visualisation en 2D est peut-être la mieux adaptée à tous utilisateur car elle permet d'obtenir un bon compromis entre le nombre d'éléments visualisés et l'apprentissage nécessaire. Cependant, les interfaces 3D, sous réserve d'un apprentissage minimum, peuvent donner d'aussi bons, voire de meilleurs résultats que les autres types de visualisations [72].

4.1.3.b La couleur

La couleur est un axe exploité dans la plupart des visualisations proposées dans la littérature. En effet, la couleur permet de faire une distinction visuelle entre les différents éléments. Keim explique que la coloration a un fort impact sur l'interprétation des résultats. Cugini souligne d'ailleurs que le balayage visuel des couleurs (qui est un processus automatique) demande moins de temps et d'efforts que le balayage visuel des termes. Par ailleurs, l'utilisation des dégradés de couleurs permet de réaliser visuellement une hiérarchisation des éléments [62].

Pour comparer l'utilisation des couleurs, nous avons défini deux catégories d'utilisation de la couleur. Elles peuvent être utilisées à des fins :

- ***Pragmatiques***. Les couleurs correspondent à des actions réalisées par l'utilisateur ou à des caractéristiques uniquement destinées à distinguer les éléments. Par exemple, le fait de faire passer un document de la couleur verte à la couleur bleue traduit l'action « document visité »,
- ***Sémantiques***. Les couleurs traduisent l'importance d'un critère ou d'une caractéristique des métaphores. Par exemple, un document de couleur verte traduit un document pertinent alors qu'un document de couleur rouge traduit un document non pertinent.

Le tableau suivant synthétise les principaux travaux présentés dans cette section.

Légende du Tableau (« N/d » signifie « non disponible »)	
Dimensions	***Couleurs***
T : dimension textuelle	***P*** : Pragmatiques
2 : 2D	***S*** : Sémantiques
3 : 3D	***D*** : Dégradés

Projet	*Dimension*				*Couleur*			
Nom	***T***	***2***	***3***	***Détails***	***P***	***S***	***D***	***Détails***
Cougar	X	X		Chaque catégorie sélectionnée est représentée par un cercle. Les documents sont positionnés selon leur appartenance aux différentes catégories.	X			La couleur des cercle correspond à la couleur de la catégorie sélectionnée.
DocCube			X	A l'intersection des différentes entrées des hiérarchies de concepts, une sphère est affichée dont la taille est proportionnelle au nombre de documents contenus dans toutes ses concepts.	X			Selon les actions réalisées par l'utilisateur, les éléments changent de couleur (pour la sélection notamment).
Grouper	X			Les documents sont rassemblés en classes.				N/d
InfoCrystal		X		Ne présente que le nombre d'éléments par rapport aux combinaisons des différents critères.				N/d
Kartoo		X		Les documents sont représentés par des cercles dont la taille correspond à la pertinence globale. Ils sont reliés par des traits correspondant à des termes qu'ils partagent.	X			Chaque trait reliant les documents possède une couleur qui correspond au terme partagé entre les différents documents.
MapStan		X		Les documents sont répartis en quartiers (cercles) selon leur similarité. Ils sont reliés par des rues dont la taille dépend de la similarité entre les quartiers.	X			Les quartiers retrouvés par l'outil de recherche sont présentés en rouge alors que les documents recommandés sont présentés en bleu.
Three-Keywords Axes Display			X	3 critères maximum sont visualisés de façon orthogonale. Plusieurs mots clés peuvent être assignés au même axe. Les documents sont visualisés par un carré positionné selon l'importance des différents critères. La taille de l'élément dépend de la taille du document.	X			Chaque document est représenté par un carré bleu. A l'intérieur de ce carré, chaque critère est représenté par une droite de la couleur spécifique au critère et dont la longueur dépend de l'importance du critère.
TileBars	X	X		Chaque *Tile* (petit carré) présente la longueur du document. Le document est découpé en sections où est visualisée l'importance de chaque	X	X	X	L'intensité de la couleur (dégradé du blanc au noir) correspond à l'importance du critère dans la section en cours. Une autre utilisation des couleurs sert à

				critère.				caractériser chaque critère.
Umap		X		Les îlots représentent des termes issus des documents et très corrélés.		X	X	La couleur dépend du nombre de documents dans lesquels apparaît le terme.
Vibe		X		Chaque critère est représenté par un cercle et tous les documents sont positionnés par rapport à ces critères. Plus ils sont proches d'un critère plus celui-ci est important dans les documents.				N/d
VR-Vibe			X	Chaque critère est présenté par un octaèdre. Les documents sont représentés par des cubes qui sont positionnés comme dans *Vibe* selon l'importance des différents critères. La taille des cubes correspond à la pertinence globale des documents.	X		X	Tous les documents ont une même couleur initiale (bleue). Par contre, l'intensité de celle-ci correspond à la pertinence globale du document. La teinte dépend des actions de l'utilisateur : ouverture d'un document, désignation d'un document pertinent.
WebSom		X		Les documents sont présentés sous forme d'une grille.		X	X	La couleur dépend du nombre de documents présents dans l'entrée de la grille.

Tableau 1. *Catégorisation des principales interfaces de visualisation*

Outre ces axes de visualisation, un problème important des interfaces pour la RI reste la multitude des tâches (recherche précise d'un document, survol d'un domaine…) et la diversité humaine. [73], par exemple, souligne l'intérêt de l'adaptation de l'interface aux utilisateurs. Ainsi, en plus des critères cognitifs « standards », Vernier [74] a introduit le critère « *d'affordance* » qui indique que l'utilisateur doit comprendre qu'il est nécessaire de changer de visualisation selon sa tâche et son niveau d'expérience. Ceci implique donc l'intégration au sein du même système de plusieurs visualisations possibles, que ce soit pour répondre au grand nombre de tâches ou pour s'adapter à l'utilisateur en lui proposant des outils exploitables et ce quelque soit son niveau d'expérience.

4.2 Interfaces de visualisation de la structure du World Wide Web

Les graphes orientés sont une cible attrayante pour la visualisation en raison de leur omniprésence dans les systèmes d'information. Un bon nombre de structures qui imprègnent l'informatique peut être représenté en tant que graphes (nœuds-liens).

Le Web est ainsi un problème intéressant pour la visualisation parce qu'il est associé à un graphe dont les nœuds sont fortement interconnectés. Généralement le concepteur d'un site Web a une notion claire de l'hiérarchie à l'intérieur de son site. Visualiser le Web est devenu un thème récurrent pour la visualisation de l'information. De nombreux chercheurs se sont efforcés d'améliorer le problème qui afflige les internautes à utiliser des navigateurs traditionnels disposant d'un historique sous forme d'une liste à une dimension. Les

Webmasters sont intéressés à voir à la fois la structure statique de leur site et le trafic dynamique à travers la structure du site.

4.2.1 Visualisation dans un espace 3D hyperbolique

Le problème classique avec la mise en page d'un arbre dans un espace euclidien est : que le nombre de nœuds croît de façon exponentielle, mais la circonférence du cercle ou l'air de la sphère croît seulement polynomiallement. Pour éviter les collisions, nous devons allouer moins de place aux nœuds qui se produisent plus en profondeur dans l'arbre. Lorsque qu'on zoome en arrière pour voir un aperçu de l'ensemble de l'arbre, les seuls nœuds qu'on peut voir en détail sont ceux qui entourent le nœud racine.

Dans l'espace hyperbolique [76], la circonférence et l'air augmentent de manière exponentielle au lieu d'une manière géométrique. Il y a assez de place pour allouer la même quantité d'espace pour chaque nœud, quelle que soit son profondeur dans l'arbre.

Bien que l'espace hyperbolique soit infini, nous pouvons le projeter dans un volume fini de l'espace euclidien. Lorsqu'on pose et déplace des arbres en utilisant des distances hyperboliques, nous pouvons voir les détails tout autour du nœud actuel tout en conservant une vue d'ensemble de la totalité de la structure.

Par exemple. L'outil de visualisation H3 [75] se base sur un outil de visualisation hyperbolique.

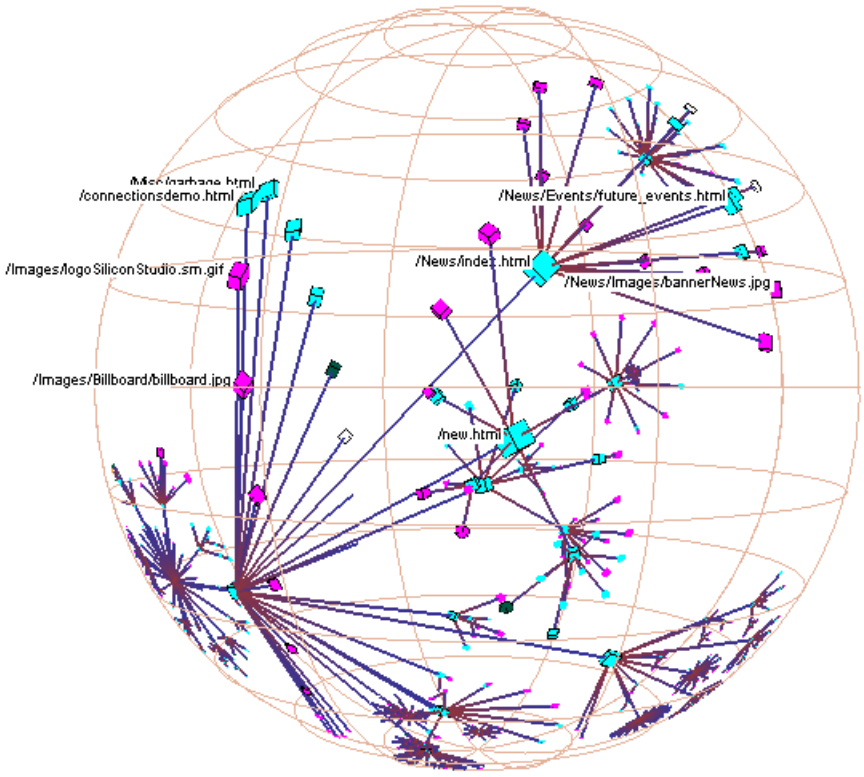


Figure 20. *Exemple d'affichage d'un site web dans H3 [77]*

4.2.2 Visualisation dans un espace 3D hyperbolique en utilisant le procédé fisheye (œil de poisson)

Lors de l'utilisation d'un espace hyperbolique 3D pour afficher les graphes dans le cadre d'une distorsion fisheye, et à tout moment, le montant de grossissement, et le niveau de détail visible, varient à travers la fenêtre d'affichage. Cela permet à l'utilisateur d'examiner les détails

d'une petite région du graphe, tout en ayant une vue d'ensemble de la totalité du graphe disponible en tant que cadre de référence.

Les graphes sont rendus dans une sphère qui contient la projection euclidienne de l'espace hyperbolique 3D. Les points à l'intérieur de la sphère sont amplifiés en fonction de leur distance radiale depuis le centre. Les objets près du centre sont amplifiés, tandis que ceux près de la frontière sont rétrécis. Le montant de grossissement diminue continuellement à un rythme accéléré à partir du centre vers la frontière, jusqu'à ce que les objets soient réduits à une taille de zéro. En apportant différentes parties d'un graphe à la région centrale amplifiée, l'utilisateur peut examiner chaque partie du graphe en détail.

Walrus est un outil pour visualiser interactivement de grands graphes orientés dans un espace tridimensionnel. Il est techniquement possible d'afficher des graphes contenant plusieurs million de nœuds ou plus, mais ça va diminuer son efficacité.

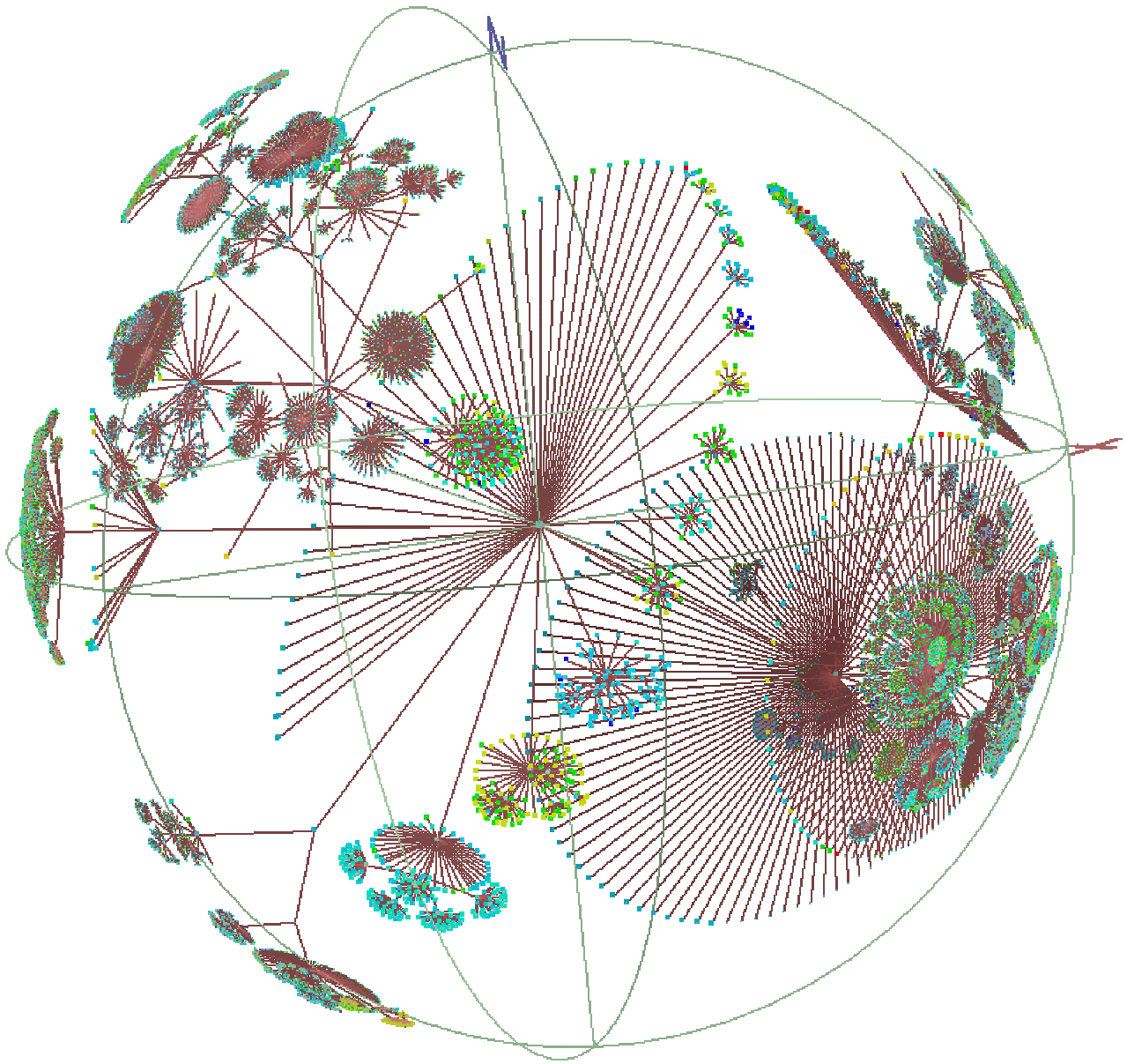


Figure 21. *Exemple d'affichage d'un site web dans Walrus*

4.2.3 Approche fractale pour la visualisation

Comme il est bien connu, le mot « fractale » a été proposé par B. Mandelbrot et représente une grande variété d'objets similaires.

Par exemple, l'arbre binaire (figure suivante) est une fractale. [78] Chaque longueur d'arête est r fois plus grande que celle de la précédente. Il existe une relation : D = - log r N entre le facteur de branchement, N, et le facteur d'échelle, R. La constante D s'appelle la dimension de fractale.

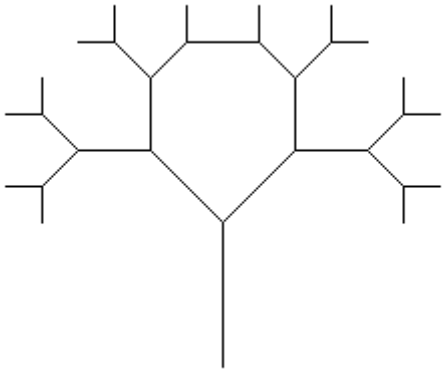


Figure 22. *Un arbre de fractale*

Bien que les arbres généraux ne soient pas similaires, si la relation : D = - log rx Nx existe entre le facteur de branchement Nx au nœud X et le facteur d'échelle rx, cet arbre est dit aussi une fractale.

En utilisant la caractéristique de similitude d'une fractale, il est possible de normaliser la vue à chaque niveau de l'arbre. Dans un arbre de fractale, une vue semblable est obtenue quand nous zoomons sur un sous-arbre. Quand une partie de sous-arbre est agrandie, une vue semblable apparaît aussi. La figure suivante représente ce concept. Toutefois même si l'arbre à afficher est énorme, la vue obtenue en se concentrant sur une partie de l'arbre est presque identique. Dans un système de visualisation pour les structures hiérarchiques, les nœuds ont des rôles plus importants que les arêtes. Ainsi, la taille des nœuds doit être réduite avec la longueur de l'arête en utilisant les mêmes facteurs d'échelle.

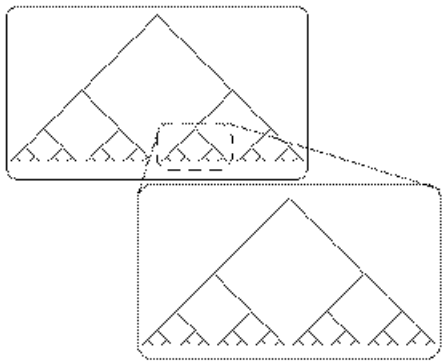


Figure 23. *Le concept de la disposition de fractale*

Bien que la visualisation en utilisant les fractales permet de garder le montant total de nœuds reconnaissables presque constant, elle n'est pas assez pour l'instant un système pratique de visualisation. Comme tous les nœuds affichés, qu'ils soient reconnaissables ou pas, réduisent le temps de réponse du système, ces nœuds méconnaissables comme ceux qui sont en dehors du champ de visionnement doivent être effacés. Comme système de visualisation, nous citons VOGUE [79].

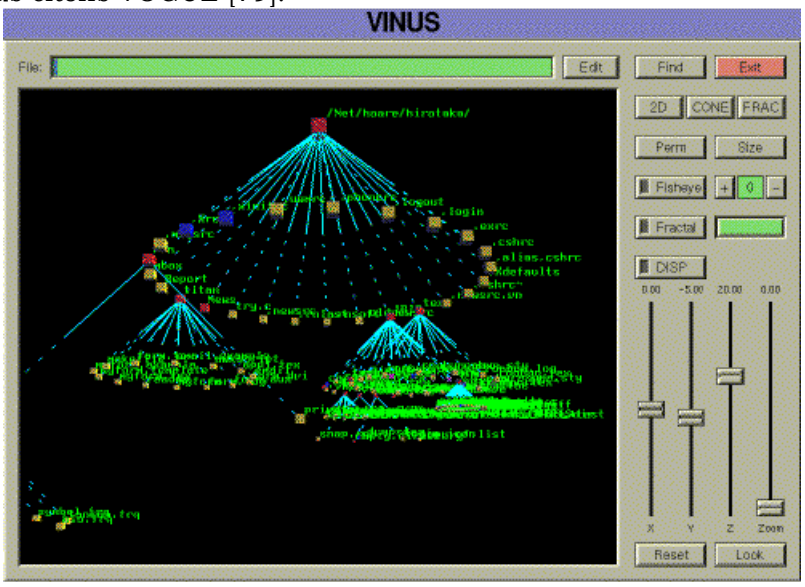


Figure 24. *VOGUE*

Chapitre 5
Mise en œuvre de la plateforme commune

Nous nous proposons dans ce chapitre de poser une solution pour chacun des univers étudiés. Nous expliciterons les méthodes de formation des réseaux sémantiques sous forme d'un graphe RPMH et de visualisation des relations retenues. Nous proposerons un exemple illustratif et deux études de cas l'une sur l'univers des dictionnaires et l'autre sur celui des pages web.

5.1 Hypothèses de recherche et solutions

Notre premier but dans cette recherche était d'étudier différents types de relations dans divers univers tels que les dictionnaires, les pages Webs et les textes. Les entités constitutives de ces univers (les pages web, les articles associés aux entrées d'un dictionnaire, …) forment un graph RPMH pour certaines relations. Ainsi, dans le cas de l'univers de dictionnaire, nous ne gardons que la relation de synonymie (les autres relations ne forment pas de RPMH) (le graphe de synonymie dans WordNet est un RPMH pour chacune des fonctions grammaticales : verbes - noms - adjectifs – adverbes), alors que dans l'univers des textes, aucune relation entre les textes ne vérifie la propriété RPMH. Quant à la structure de lien entre les pages web, cette propriété devient valide.

Ensuite nous nous sommes intéressés à clustériser dans les univers cités un ensemble d'items exhibant une relation vérifiant la propriété RPMH pour la visualiser sous forme d'un nuage de points.

Nous sommes donc parti à la recherche d'une méthode qui permet de visualiser localement et globalement un graphe RPMH (lexical ou de pages webs) tout en prenant en compte sa structure globale (les relations entre les différents éléments étudiés). La solution que nous avons proposée ici se base sur l'utilisation d'une matrice de transition (il existe un arc d'un sommet A vers un sommet B si et seulement si l'entité B possède un lien avec l'entité A). Parcourir x fois ce graphe revient à multiplier la matrice de transition x fois. La matrice obtenue en tant que matrice de coordonnées dans $\mathbb{R}^n$ contient une information calculée sur l'ensemble du graphe (la multiplication de la matrice par elle-même renvoie des informations sur la relation de tous les nœuds entre eux) qu'il serait possible de représenter dans $\mathbb{R}^2$ au moyen d'une Analyse en Composante Principale (ACP) en gardant les 2 premiers axes.

5.2 Schémas synoptiques

5.2.1 Solution proposée pour l'univers des dictionnaires

Le schéma général de ce module est représenté ci-dessous. De façon générale, ce module utilise l'annotateur Tree-Tagger pour récupérer les formes de base qui sont des mots dans l'univers des dictionnaires, pour ensuite faire appel à WordNet et enclencher le processus de construction du graphe RPMH induit des mots introduits après avoir choisi la relation sur laquelle va porter l'analyse grammaticale des mots (si une relation de synonymie existe entre 2 mots, alors une valeur de 1 sera attribuée, sinon une valeur de 0). Le graphe ainsi obtenu sera transformé en une matrice (de dimension $\mathbb{N}^p$) et cette dernière sera envoyée à MatLab pour être transformée en une matrice markovienne (passage à $\mathbb{R}^p$). Une série de multiplication (le nombre de multiplication sera introduit par l'utilisateur) sera appliquée à la matrice Markovienne obtenue permettant ainsi de regrouper le plus possible les différents mots rencontrés lors de la construction du graphe. Ensuite, nous appliquons une analyse en composante principale pour passer de $\mathbb{R}^p$ à $\mathbb{R}^2$, et ainsi visualiser la relation de synonymie sous forme d'un nuage de point en 2D.

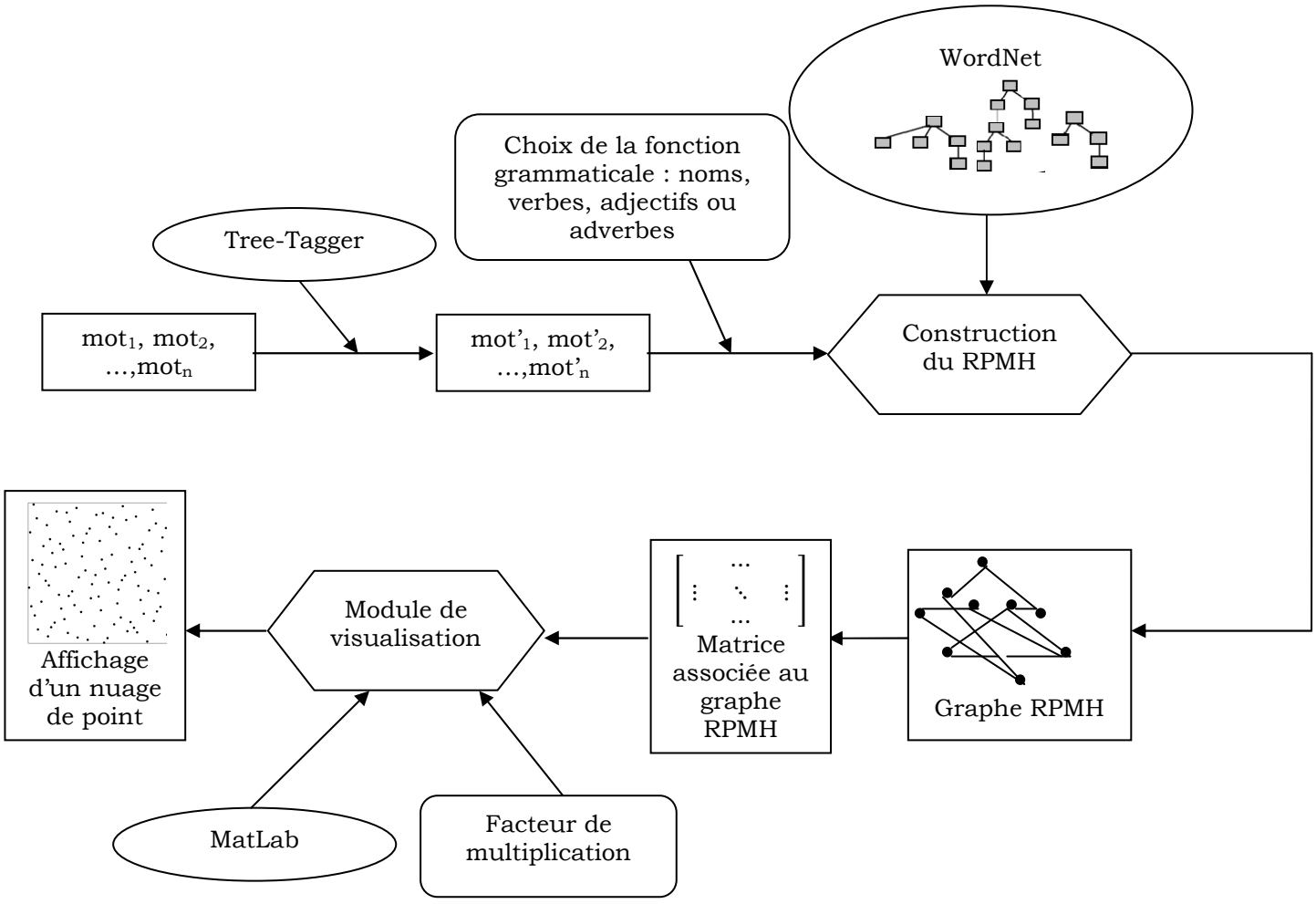


Figure 25. *Schéma de fonctionnement général du module proposé pour l'univers des dictionnaires*

5.2.2 Solution proposée pour l'univers des textes

De même le schéma général de ce module utilise l'annotateur Tree-Tagger pour récupérer les formes de bases de l'univers des textes qui sont un concept et un ensemble de textes constituant le corpus étudié, pour ensuite faire appel à WordNet et enclencher le processus de construction du graphe RPMH induit de tous les mots constituant aussi bien le concept que les mots constituants les différents textes du corpus après avoir choisi sur quoi va porter l'analyse grammaticale de cet ensemble de mots (si une relation de synonymie existe entre 2 mots, alors une valeur de 1 sera attribuée, sinon une valeur de 0). Le graphe ainsi obtenu sera transformé en une matrice (de dimension $\mathbb{N}^p$) et cette dernière sera envoyée à MatLab pour être transformée en une matrice markovienne (passage à $\mathbb{R}^p$). Ensuite, nous appliquons une analyse en composante principale pour passer de $\mathbb{R}^p$ à $\mathbb{R}^2$; nous calculons pour l'ensemble des mots constituant chaque texte donné du corpus son résumé pour permettre ainsi la visualisation du concept ainsi que l'ensemble des textes du corpus sous forme d'un nuage de point en 2D. L'une des méthodes pour avoir le résumé est de calculer le barycentre des points associés aux mots du texte.

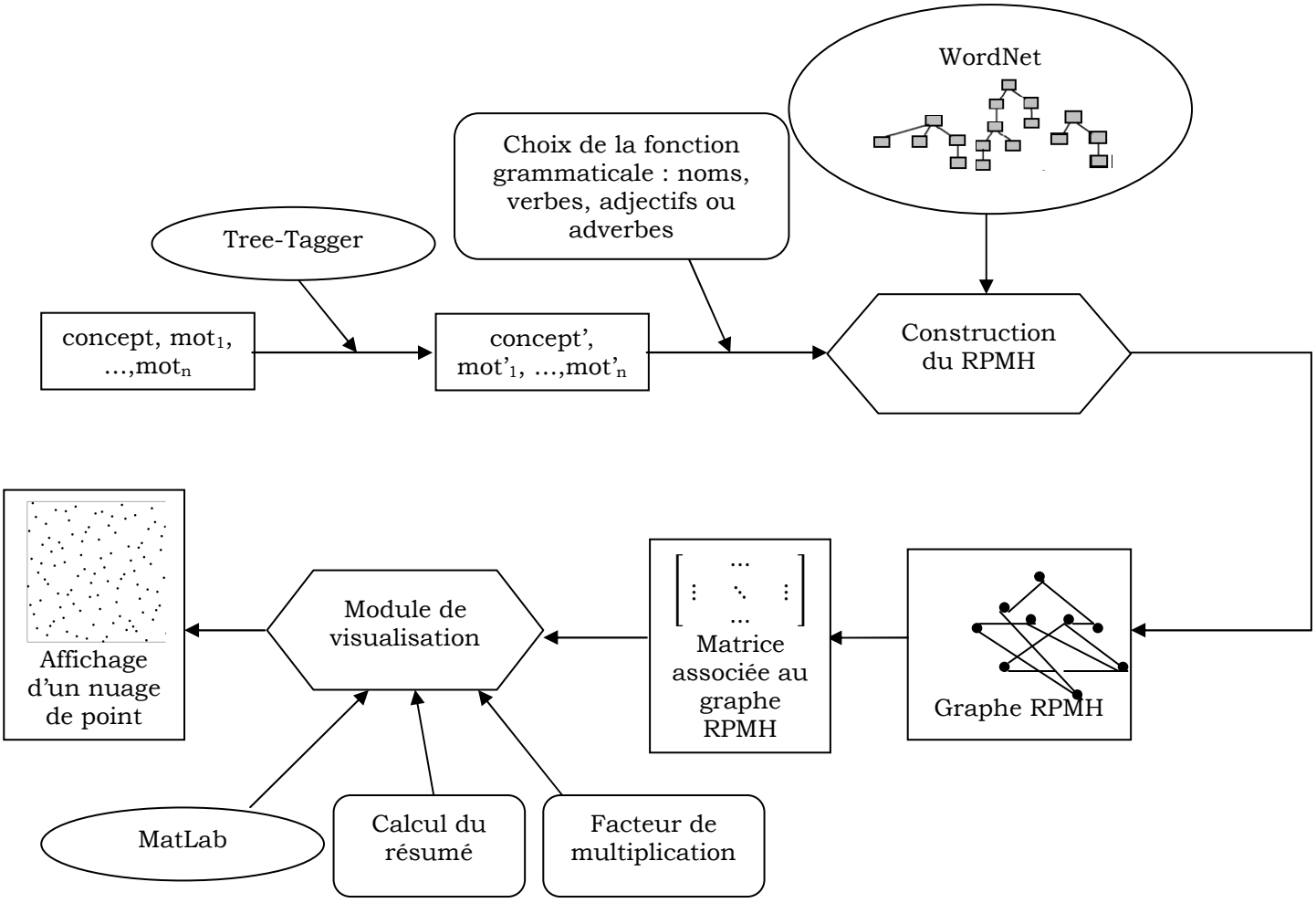


Figure 26. *Schéma de fonctionnement général du module pour l'univers des textes*

5.2.3 Solution proposée pour l'univers des pages web

Le schéma général de ce module est représenté ci-dessous. De façon générale, ce module utilise un crawler pour récupérer les différents hyperliens entre les pages webs d'un site donné pour ensuite faire enclencher le processus de construction du graphe RPMH (si un lien hyperlien existe entre 2 pages du site, alors une valeur de 1 sera attribuée, sinon une valeur de 0). Le graphe ainsi obtenu sera transformé en une matrice (de dimension $\mathbb{N}^p$) et cette dernière sera envoyée à MatLab pour être transformée en une matrice markovienne (passage à $\mathbb{R}^p$). Une série de multiplication (le nombre de multiplication sera introduit par l'utilisateur) serait effectuée à la matrice Markovienne obtenue permettant ainsi de regrouper le plus possible les différents pages rencontrés lors de la construction du graphe. Ensuite, nous appliquons une analyse en composante principale pour passer de $\mathbb{R}^p$ à $\mathbb{R}^2$; et ainsi visualiser la relation de l'analyse de liens sous forme d'un nuage de point en 2D.

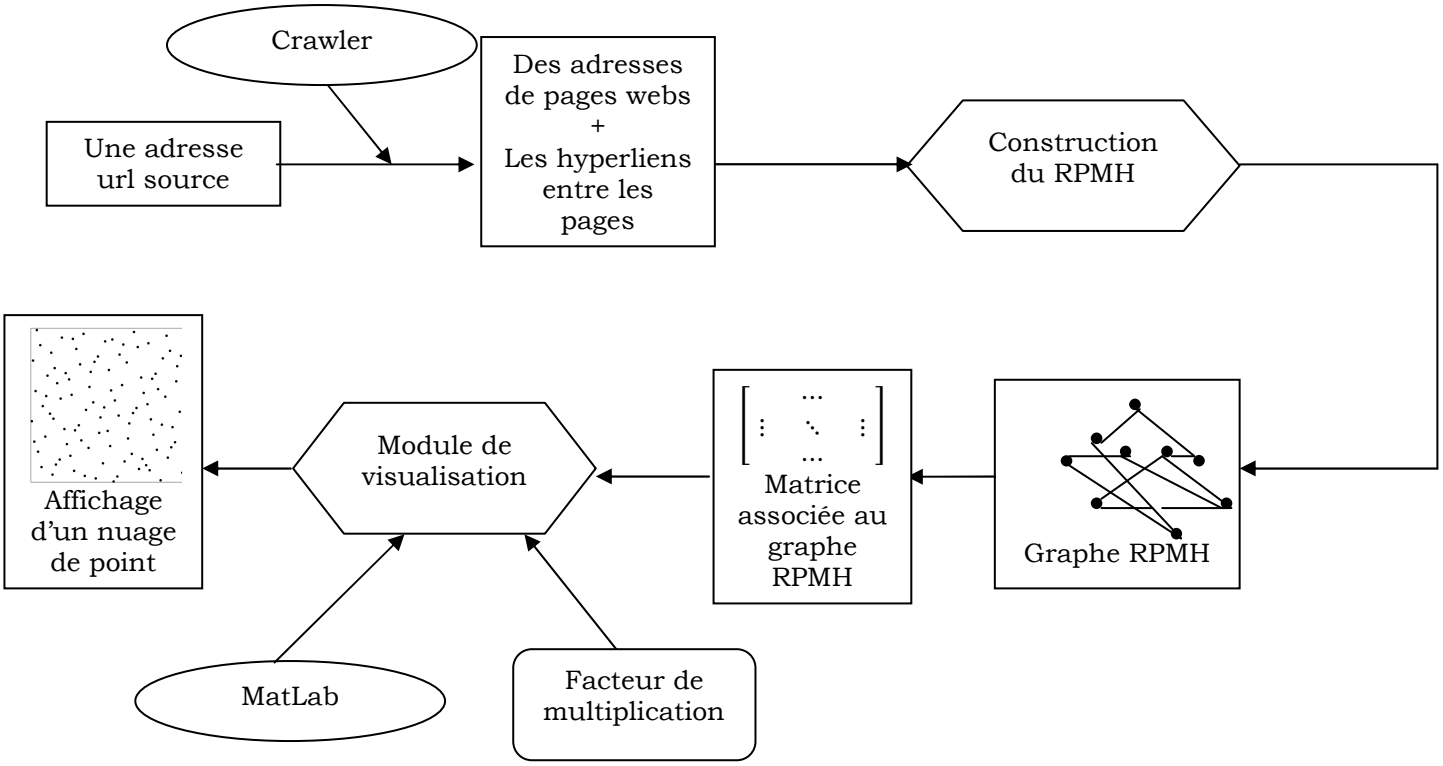


Figure 27. *Schéma de fonctionnement général du module pour l'univers des pages webs*

Nous présentons dans les paragraphes suivants les différents composants de ces schémas, WordNet, Tree Tagger puis MatLab et nous terminons en détaillant les modules de construction du réseau sémantique sous forme d'un graphe RPMH et de visualisation en passant par l'analyse en composante principale. Nous justifions leur apport et décrivons leur fonctionnement.

5.3 Outils

5.3.1 WordNet

WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise (des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour).

La composante atomique sur laquelle repose le système entier est le **synset** (*synonym set*), un groupe de mots interchangeables, dénotant un sens ou un usage particulier. Chaque synset dénote une acception différente du mot en question.

À l'instar d'un dictionnaire traditionnel, WordNet offre ainsi, pour chaque mot, une liste de synsets correspondant à toutes ses acceptions répertoriées. Mais les synsets ont également d'autres usages : ils peuvent représenter des concepts plus abstraits, de plus haut niveau que les mots et leurs sens, qu'on peut organiser sous forme d'ontologies. Une ontologie est un système de catégories permettant de classifier les éléments d'un univers. Les systèmes de catégorisation qui nous intéressent correspondent aux différentes relations sémantiques avec

lesquelles il est possible de regrouper de manière cohérente les composantes d'un univers linguistique (les mots, les sens et les concepts par exemple).

La relation sémantique servant de critère pour l'agrégation d'un groupe de concepts définira le type de l'ontologie. WordNet répertorie ainsi une grande variété de relations sémantiques permettant d'organiser le sens des mots (et donc par extension les mots eux-mêmes) en des systèmes de catégories qu'on peut consulter de manière cohérente et uniforme. On pourra ainsi interroger le système quant aux hyperonymes d'un mot particulier. On peut également interroger le système quant à la relation inverse de l'hypernymie, l'hyponymie. WordNet offre en fait une multitude d'autres ontologies, faisant usage de relations sémantiques plus spécialisées et restrictives. On peut ainsi interroger le système quant aux méronymes d'un mot ou d'un concept, les parties constitutives d'un objet (HAS-PART).

WordNet est un système d'une étonnante ampleur : la version la plus récente (2.1) répertorie plus de 200 000 mots de classes ouvertes (pour lesquelles l'ajout d'éléments lexicaux est possible) ainsi que plus de 115 000 synsets. À chaque nouvelle version, le lexique s'enrichit de nouveaux mots, et des relations sémantiques sont ajoutées, modifiées, ou encore rendues désuètes. Si on examine par exemple l'ontologie générée par la relation d'hyperonymie, il est notable qu'elle est la plus complète dans son embranchement nominal (le lexique de WordNet est séparé en quatre grandes super-catégories lexicales: les noms, les verbes, les adjectifs et les adverbes. Les noms sont ainsi classés en un système de catégories complet et précis comprenant plusieurs niveaux d'imbrication (on retrouve notamment certaines sections de cette ontologie où la profondeur dépasse 10 niveaux). On retrouve en revanche un système de classification beaucoup moins élaboré pour les verbes, qui sont organisés en un système hiérarchique beaucoup plus « plat » (moins de niveaux d'imbrication), où on passe très rapidement d'un concept spécialisé à un concept très général. À ce jour, il n'y a aucune catégorisation hiérarchique définie pour les embranchements des adjectifs et des adverbes. Ce déséquilibre potentiellement problématique se retrouve à l'intérieur même des super-catégories, où il est évidemment beaucoup plus apparent dans la branche nominale : certains mots sont ainsi liés à une grande chaîne de concepts finement graduée, tandis que d'autres sont très proche des concepts les plus généraux.

WordNet jouit d'une énorme et grandissante popularité au sein de la communauté scientifique, et joue également un rôle important dans plusieurs projets commerciaux. Sa richesse et sa précision en font un outil de choix, susceptible d'être mis à profit par une multitude de techniques et de théories diverses. Son utilisation fait en sorte de procurer aux algorithmes et applications une importante plate-forme de connaissances a priori du langage et du monde dans lequel il s'articule. Un exemple particulièrement représentatif et ingénieux de son utilisation est donné par les métriques heuristiques de "distance sémantique" entre les concepts d'une ontologie particulière, basées sur la distance à parcourir dans le graphe. Cette distance peut permettre de quantifier par exemple la similarité de deux concepts. Elle peut également servir à faire de la désambiguïsation.

5.3.2 Tree Tagger

Le TreeTagger[1] est un outil mis au point pour l'annotation grammaticale de données textuelles, par l'association à chacun des mots de l'information de la "partie du discours" et du lemme.

Il a été mis au point dans le cadre du projet TC[2] à l'Institut de linguistique informatique de l'Université de Stuttgart. Il est utilisé avec succès pour l'étiquetage dans les principales langues européennes dont l'anglais, le français, le bulgare et le russe, et est adaptable à d'autres langues dans la mesure où l'on dispose d'un lexique et d'un corpus d'entraînement étiqueté manuellement.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[2] http://www.ims.uni-stuttgart.de/projekte/tc/

TreeTagger utilise un arbre de décision binaire pour calculer la taille du contexte à utiliser pour estimer les probabilités de transition.

Le lexique implémenté dans TreeTagger contient la liste des possibilités d'étiquetage pour chaque mot. Il se divise en trois parties :

- un ***lexique pleine forme***, contenant les deux millions d'entrées du corpus Penn Treebank. Les mots dont la fréquence relative était inférieure à 1% ont été supprimés : ils étaient le plus souvent dus à des erreurs d'étiquetage;
- un ***lexique de suffixes***, organisé dans une structure arborescente. Chaque nœud de l'arbre, à l'exception du nœud racine, correspond à un caractère. Les feuilles de l'arbre contiennent des vecteurs de probabilité des tags;
- et une ***entrée par défaut***, obtenue en soustrayant aux fréquences du nœud racine les fréquences des tags dans les feuilles de l'arbre, et en normalisant les fréquences résultantes.

La recherche d'un mot dans le lexique démarre par une recherche dans le premier fichier (avec changement de la casse du mot si la recherche s'avère infructueuse avec la casse originelle); puis dans le second si le mot n'a pas été trouvé dans le premier.

L'intégration du Tree Tagger s'est avérée indispensable dans notre programme pour récupérer la catégorie grammaticale (verbe, nom, adverbe, adjectif) des mots fournis par l'utilisateur. Il choisira une de ces catégories pour ensuite laisser le programme interroger l'ontologie WordNet en vue de récupérer seulement les synonymes des mots ayant la même fonction grammaticale.

5.3.3 MatLab

MATLAB[3] est une abréviation de Matrix LABoratory. Écrit à l'origine, en Fortran, par C. Moler, MATLAB était destiné à faciliter l'accès au logiciel matriciel développé dans les projets LINPACK et EISPACK.

MATLAB est un environnement puissant, complet et facile à utiliser destiné au calcul scientifique. Il apporte aux ingénieurs, chercheurs et à tout scientifique un système interactif intégrant calcul numérique et visualisation. C'est un environnement performant, ouvert et programmable qui permet de remarquables gains de productivité et de créativité.

MATLAB est un environnement complet, ouvert et extensible pour le calcul et la visualisation. Il dispose de plusieurs centaines (voire milliers, selon les versions et les modules optionnels autour du noyau Matlab) de fonctions mathématiques, scientifiques et techniques. L'approche matricielle de MATLAB permet de traiter les données sans aucune limitation de taille et de réaliser des calculs numériques et symboliques de façon fiable et rapide.

MATLAB possède son propre langage, intuitif et naturel qui permet des gains de temps de CPU spectaculaires par rapport à des langages comme le C, le TurboPascal, le Fortran et le JAVA. Avec MATLAB, on peut faire des liaisons de façon dynamique, à des programmes C, Fortran, et JAVA échanger des données avec d'autres applications (via la DDE : MATLAB serveur ou client) ou utiliser MATLAB comme moteur d'analyse et de visualisation.

5.3.4 Web Crawler

Un robot d'indexation (ou littéralement araignée du Web ; en anglais web crawler ou web spider) est un logiciel qui explore automatiquement le Web. Il est généralement conçu pour

[3] http://www.mathworks.fr

collecter les ressources (pages web, images, vidéos, documents Word, PDF ou PostScript, ...), afin de les exploiter (un moteur de recherche les indexe).

Fonctionnant sur le même principe, certains robots (*spambots*) sont utilisés pour archiver les ressources ou collecter des adresses électroniques auxquelles envoyer des pourriels.

Pour retrouver de nouvelles ressources, un robot procède en suivant récursivement les hyperliens trouvés à partir d'une page pivot. Par la suite, il est avantageux de mémoriser l'URL de chaque ressource récupérée. Toutefois, de nombreuses ressources échappent à cette exploration récursive, car seuls des hyperliens créés à la demande, donc introuvables par un robot, permettent d'y accéder. Cet ensemble de ressources inexploré est parfois appelé web profond.

Un fichier d'exclusion (*robots.txt*) placé dans la racine d'un site web permet de donner aux robots une liste de ressources à ignorer. Cette convention permet de réduire la charge du serveur web et d'éviter des ressources sans intérêt. Par contre, certains robots ne se préoccupent pas de ce fichier.

Deux caractéristiques du Web compliquent le travail du robot d'indexation : le grand volume de données et la bande passante. Un très grand nombre de pages sont ajoutées, modifiées et supprimées chaque jour. Si la capacité de stockage d'information, comme la vitesse des processeurs, a augmenté rapidement, la bande passante n'a pas bénéficié de la même progression. Le problème est donc de traiter un volume toujours croissant d'information avec un débit limité. Le robot a donc besoin de donner des priorités à ses téléchargements.

5.4 Construction du réseau sémantique sous forme d'un graphe RPMH

5.4.1 Pour l'univers de dictionnaires et de textes

Ayant un ensemble de mot $\{m_1, m_2, \ldots, m_n\}$ et partant du mot donné m_1 (un des mots dans le cas dans l'univers de dictionnaires et le concept dans le cas de l'univers des textes), et ayant une fonction grammaticale précise (nom, verbe, adjectif ou adverbe) nous commençons par interroger l'ontologie WordNet pour renvoyer l'ensemble des mots synonymes ayant la même fonction grammaticale du mot m_1. Comme les mots renvoyés entretiennent une relation de synonymie symétrique entre eux, la relation sera assignée une valeur de 1 indiquant sa présence. Pour chaque mot synonyme retrouvé, nous réitèrons le même calcul, jusqu'à rencontrer l'ensemble des mots $\{m_2, \ldots, m_n\}$. Ceci est possible parce que chacun des graphes des différentes formes grammaticales dans l'ontologie WordNet forme une seule composante connexe. Une fois l'ensemble des mots rencontrés, une matrice symétrique $n \times n$ sera formée où la valeur de l'entrée $[i,j]$ correspondante à la ligne i (mot m_i) et au colonne j (mot m_j) représente la présence de relation de synonymie entre les mots i et j, autrement dit une valeur de 1 traduit la présence de cette relation et 0 son absence.

Il est à noter que nous avons opté pour un parcours en largeur des graphes en WordNet parce que dans le cas contraire, il y a un risque de diverger rapidement en empruntant des chemins polysémiques (les chemins polysémiques sont des chemins dans le graphe ayant un ou plusieurs mots polysémiques) et ainsi parcourir tout le graphe avant de retrouver les autres mots. La figure 28 schématise ce phénomène.

Supposons que l'utilisateur souhaite étudier la relation de synonymie entre les mots "*dark*", "*shadow*" et "*night*".

Voici le réseau sémantique issu de l'ontologie WordNet.

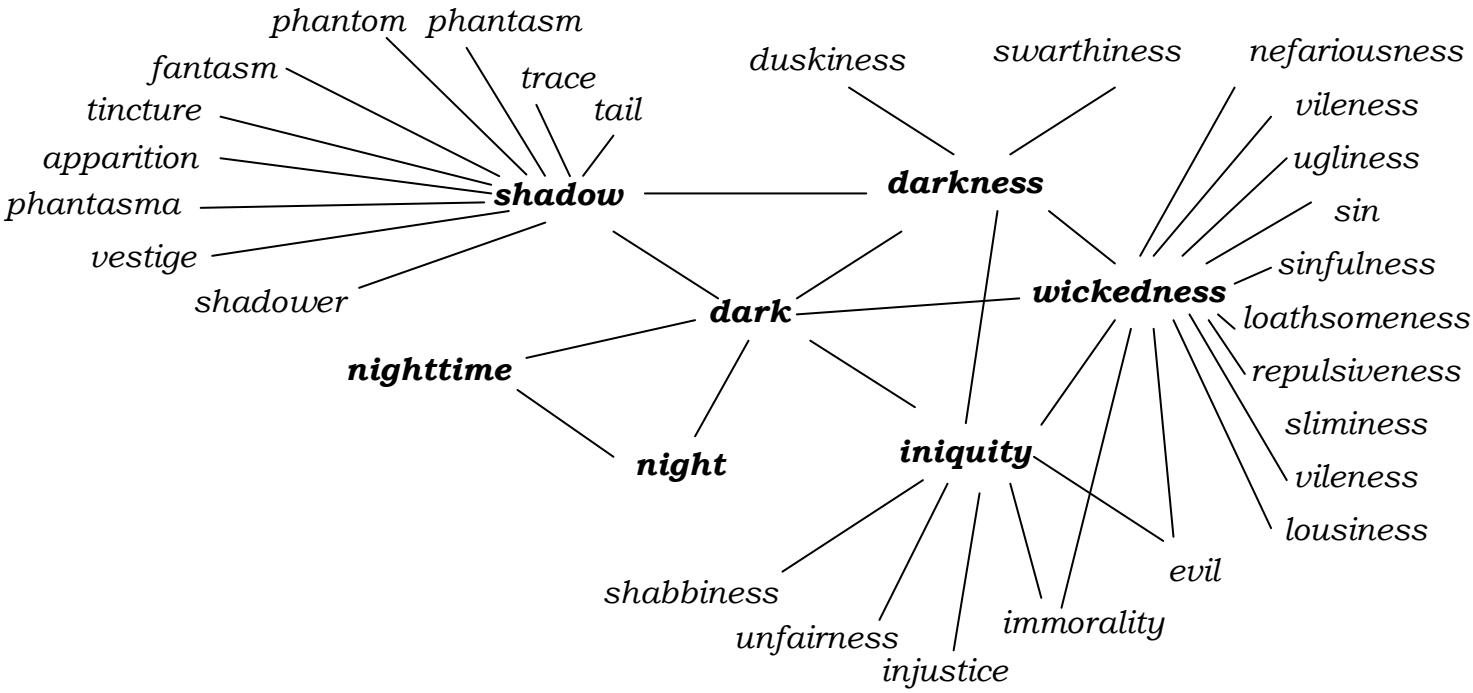


Figure 28. *Réseau sémantique du mot **dark**. Les mots dont les synonymes sont affichés sont en gras.*

On voit bien que les trois mots entretiennent une relation de synonymie directe. Si par exemple nous commençons par parcourir le graphe en profondeur en commençant par le premier synonyme *shadow*, nous allons nous retrouver avec d'autres mots aussi bien polysémiques alors qu'en parcourant en largeur, il y a plus de chance de rencontrer les mots recherchés vu que la relation étudiée est la synonymie. Or, plus nous s'éloignons du mot initial, plus la probabilité de retrouver des mots qui sont synonymes baisse ; ce qui contrait à la recherche de ces mots dans l'entourage du mot initial d'où la justification de l'utilisation d'un parcours en largeur du graphe.

5.4.2 Pour l'univers des pages web

Dans le cas de l'univers des pages web, l'utilisateur fournit au crawler l'adresse d'une page web pivot d'un site donné et c'est au crawler de parcourir cette page à la recherche des hyperliens. Chaque page pointée sera de nouveau analysée de la même manière. La présence d'un hyperlien entre 2 pages p et q se traduit par une valeur de 1 dans l'entrée $[p,q]$ de la matrice correspondante.

5.5 Visualisation des relations

Les méthodes multifactorielles d'analyse des données permettent d'obtenir des représentations graphiques qui constituent le meilleur résumé possible de l'information contenue dans un grand tableau ou matrice de données. Pour cela, il faut consentir à une perte d'information afin de gagner en lisibilité. En fonction des phénomènes que l'on veut étudier et de la nature du tableau de données dont on dispose, on appliquera telle ou telle méthode multifactorielle. En effet, il n'existe pas une méthode factorielle d'analyse des données, mais un ensemble de méthodes, reposant toutes sur les mêmes théories mathématiques. Ainsi, on trouvera les principales méthodes suivantes :

- **ACP** : Analyse en Composantes Principales, pour les tableaux de variables quantitatives.
- **AFTD** : Analyse Factorielle d'un Tableau de Distances, pour les tableaux de distances.
- **AFC** : Analyse Factorielle des Correspondances, pour les tableaux de contingence.
- **ACM** : Analyse des Correspondances Multiples, pour les tableaux de variables qualitatives.
- **STATIS** : Structuration des Tableaux A Trois Indices de la Statistique, **AFM** : Analyse Factorielle Multiple, **DACP** : Double Analyse en Composante Principale, sont quelques méthodes basées sur les précédentes et adaptées à l'étude de phénomènes temporels ou de répétition.
- la liste n'est pas exhaustive.

Dans la plupart des situations, on dispose de plusieurs observations sur chaque individu constituant la population d'étude. On a donc à prendre en compte *p* variables par individu, *p* étant strictement supérieur à 1. L'étude séparée de chacune de ces variables donne quelques informations mais reste toutefois insuffisante car elle laisse de côté les liaisons entre elles, ce qui est pourtant souvent ce que l'on veut étudier.

C'est le rôle de la statistique multifactorielle : analyser les données dans leur ensemble, en prenant en compte de toutes les variables.

L'Analyse en Composantes Principales est alors une bonne méthode pour étudier les données multidimensionnelles, lorsque toutes les variables observées sont de type numérique, de préférence dans les mêmes unités, et que l'on veut voir s'il y a des liens entre ces variables.

Dans la littérature, on trouve deux approches différentes de l'ACP[4] :

- Elle peut être présentée comme la recherche d'un ensemble réduit de variables non-corrélées, combinaisons linéaires des variables initiales résumant avec précision les données (approche anglo-saxonne).
- Une autre interprétation repose sur la représentation des données initiales à l'aide de nuage de points dans un espace géométrique. L'objectif est alors de trouver des sous-espaces (droite, plan,...) qui représentent au mieux le nuage initial.

Nombreux sont les logiciels qui sont disponibles sur le marché et qui permettent d'effectuer une ACP, on cite par exemple le logiciel SPAD et l'environnement informatique du métalangage MatLab.

Une fois l'analyse réalisée, on obtient les coordonnées des différentes variables dans un espace de dimension faible en 2D.

5.6 Applications

5.6.1 Application théorique à un petit graphe

Soit un graphe sémantique comportant 13 nœuds et un ensemble d'arêtes traduisant l'existence d'une relation de synonymie entre les nœuds. La figure suivante schématise l'ensemble des nœuds et des arêtes.

[4] Pour la formulation mathématique de l'ACP et les règles d'interprétation des résultats, consultez le site http://antoun.yaacoub.org/m2r

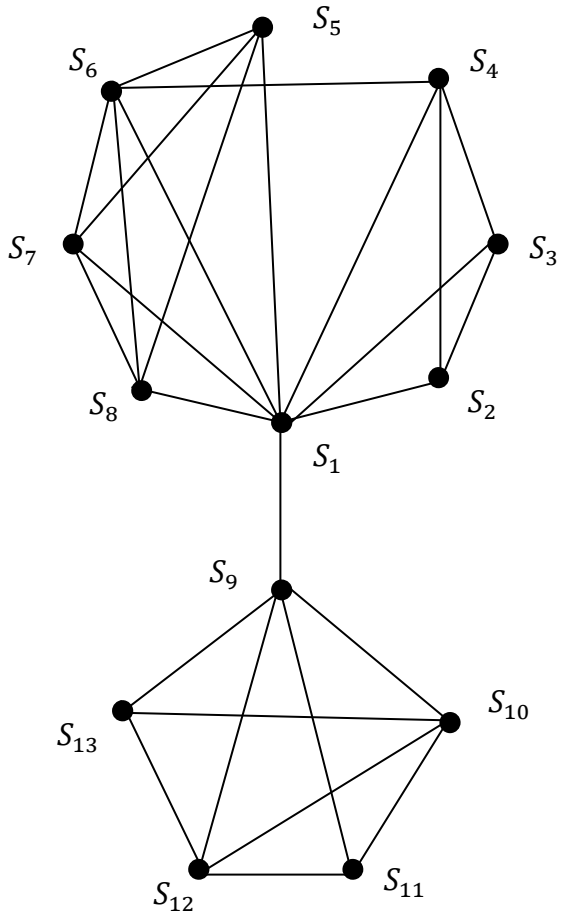


Figure 29. *Exemple théorique de graphe comportant 13 nœuds*

Il est clair que la composante S_1 ne se regroupera jamais avec la composante S_9 car il n'y a pas de cycle possible entre elles.

Soit D la matrice d'adjacence de ce dictionnaire de 13 mots.

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Soit DD la matrice markovienne de D

$$DD = \begin{bmatrix} 0 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 \\ 1/5 & 0 & 0 & 1/5 & 1/5 & 0 & 1/5 & 1/5 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix}$$

On calcule DD^7 (matrice à la page suivante) (car en 7 arcs/arêtes nous avons le temps de parcourir tous les nœuds de la composante S_1 au moins une fois, et un peu plus d'une fois pour ceux de la composante S_9)

On y voit très bien les deux composantes (rouge pour S_1 et bleue pour S_9).

Autrement dit il n'est pas vraiment nécessaire de pratiquer des algorithmes basés sur la recherche de circuits car la matrice DD^k permet de faire les regroupements recherchés.

DD^{7} :

!	0.1785334	0.0711515	0.0711515	0.0950763	0.0946581	0.1187209	0.0946581	0.0946581	0.0563950	0.0358364	0.0266622	0.0358364	0.0266622 !
!	0.1897373	0.0808388	0.0812960	0.1070578	0.0957972	0.1223823	0.0957972	0.0957972	0.0450367	0.0242291	0.0189006	0.0242291	0.0189006 !
!	0.1897373	0.0812960	0.0808388	0.1070578	0.0957972	0.1223823	0.0957972	0.0957972	0.0450367	0.0242291	0.0189006	0.0242291	0.0189006 !
!	0.1901527	0.0802933	0.0802933	0.1045816	0.0978215	0.1258854	0.0978215	0.0978215	0.0438645	0.0229295	0.0178028	0.0229295	0.0178028 !
!	0.1893161	0.0718479	0.0718479	0.0978215	0.1062312	0.1314688	0.1062923	0.1062923	0.0421404	0.0215447	0.0168261	0.0215447	0.0168261 !
!	0.1899535	0.0734294	0.0734294	0.1007083	0.1051750	0.1296367	0.1051750	0.1051750	0.0419412	0.0212009	0.0164873	0.0212009	0.0164873 !
!	0.1893161	0.0718479	0.0718479	0.0978215	0.1062923	0.1314688	0.1062312	0.1062923	0.0421404	0.0215447	0.0168261	0.0215447	0.0168261 !
!	0.1893161	0.0718479	0.0718479	0.0978215	0.1062923	0.1314688	0.1062923	0.1062312	0.0421404	0.0215447	0.0168261	0.0215447	0.0168261 !
!	0.0902320	0.0270220	0.0270220	0.0350916	0.0337123	0.0419412	0.0337123	0.0337123	0.1617770	0.1468807	0.1110079	0.1468807	0.1110079 !
!	0.0716727	0.0181718	0.0181718	0.0229295	0.0215447	0.0265011	0.0215447	0.0215447	0.1836008	0.1692541	0.1278745	0.1693151	0.1278745 !
!	0.0710993	0.0189006	0.0189006	0.0237371	0.0224348	0.0274789	0.0224348	0.0224348	0.1850132	0.1704993	0.1232837	0.1704993	0.1232837 !
!	0.0716727	0.0181718	0.0181718	0.0229295	0.0215447	0.0265011	0.0215447	0.0215447	0.1836008	0.1693151	0.1278745	0.1692541	0.1278745 !
!	0.0710993	0.0189006	0.0189006	0.0237371	0.0224348	0.0274789	0.0224348	0.0224348	0.1850132	0.1704993	0.1232837	0.1704993	0.1232837 !

DD^{35} :

0.1527434	0.0577788	0.0577788	0.0771212	0.0772165	0.0965451	0.0772165	0.0772165	0.0871037	0.0683319	0.0513078	0.0683319	0.0513078
0.1540767	0.0584239	0.0584239	0.0780054	0.0781284	0.0976920	0.0781284	0.0781284	0.0855110	0.0666657	0.0500752	0.0666657	0.0500752
0.1540767	0.0584239	0.0584239	0.0780054	0.0781284	0.0976920	0.0781284	0.0781284	0.0855110	0.0666657	0.0500752	0.0666657	0.0500752
0.1542424	0.0585040	0.0585040	0.0781152	0.0782416	0.0978345	0.0782416	0.0782416	0.0853131	0.0664587	0.0499221	0.0664587	0.0499221
0.1544331	0.0585963	0.0585963	0.0782416	0.0783721	0.0979985	0.0783721	0.0783721	0.0850854	0.0662205	0.0497459	0.0662205	0.0497459
0.1544722	0.0586152	0.0586152	0.0782676	0.0783988	0.0980322	0.0783988	0.0783988	0.0850386	0.0661716	0.0497097	0.0661716	0.0497097
0.1544331	0.0585963	0.0585963	0.0782416	0.0783721	0.0979985	0.0783721	0.0783721	0.0850854	0.0662205	0.0497459	0.0662205	0.0497459
0.1544331	0.0585963	0.0585963	0.0782416	0.0783721	0.0979985	0.0783721	0.0783721	0.0850854	0.0662205	0.0497459	0.0662205	0.0497459
0.1393659	0.0513066	0.0513066	0.0682505	0.0680683	0.0850386	0.0680683	0.0680683	0.1030828	0.0850483	0.0636737	0.0850483	0.0636737
0.1366639	0.0499993	0.0499993	0.0664587	0.0662205	0.0827145	0.0662205	0.0662205	0.1063104	0.0884248	0.0661715	0.0884248	0.0661715
0.1368208	0.0500752	0.0500752	0.0665628	0.0663278	0.0828495	0.0663278	0.0663278	0.1061228	0.0882286	0.0660264	0.0882286	0.0660264
0.1366639	0.0499993	0.0499993	0.0664587	0.0662205	0.0827145	0.0662205	0.0662205	0.1063104	0.0884248	0.0661715	0.0884248	0.0661715
0.1368208	0.0500752	0.0500752	0.0665628	0.0663278	0.0828495	0.0663278	0.0663278	0.1061228	0.0882286	0.0660264	0.0882286	0.0660264

DD^{100} :

0.1482038	0.0555825	0.0555825	0.0741110	0.0741121	0.0926404	0.0741121	0.0741121	0.0925262	0.0740046	0.0555041	0.0740046	0.0555041
0.1482199	0.0555903	0.0555903	0.0741217	0.0741232	0.0926543	0.0741232	0.0741232	0.0925069	0.0739844	0.0554892	0.0739844	0.0554892
0.1482199	0.0555903	0.0555903	0.0741217	0.0741232	0.0926543	0.0741232	0.0741232	0.0925069	0.0739844	0.0554892	0.0739844	0.0554892
0.1482219	0.0555912	0.0555912	0.0741230	0.0741245	0.0926560	0.0741245	0.0741245	0.0925045	0.0739819	0.0554874	0.0739819	0.0554874
0.1482242	0.0555924	0.0555924	0.0741245	0.0741261	0.0926580	0.0741261	0.0741261	0.0925017	0.0739790	0.0554852	0.0739790	0.0554852
0.1482247	0.0555926	0.0555926	0.0741248	0.0741264	0.0926584	0.0741264	0.0741264	0.0925012	0.0739784	0.0554848	0.0739784	0.0554848
0.1482242	0.0555924	0.0555924	0.0741245	0.0741261	0.0926580	0.0741261	0.0741261	0.0925017	0.0739790	0.0554852	0.0739790	0.0554852
0.1482242	0.0555924	0.0555924	0.0741245	0.0741261	0.0926580	0.0741261	0.0741261	0.0925017	0.0739790	0.0554852	0.0739790	0.0554852
0.1480418	0.0555041	0.0555041	0.0740036	0.0740014	0.0925012	0.0740014	0.0740014	0.0927196	0.0742069	0.0556538	0.0742069	0.0556538
0.1480091	0.0554883	0.0554883	0.0739819	0.0739790	0.0924730	0.0739790	0.0739790	0.0927586	0.0742478	0.0556841	0.0742478	0.0556841
0.1480110	0.0554892	0.0554892	0.0739832	0.0739803	0.0924747	0.0739803	0.0739803	0.0927564	0.0742454	0.0556823	0.0742454	0.0556823
0.1480091	0.0554883	0.0554883	0.0739819	0.0739790	0.0924730	0.0739790	0.0739790	0.0927586	0.0742478	0.0556841	0.0742478	0.0556841
0.1480110	0.0554892	0.0554892	0.0739832	0.0739803	0.0924747	0.0739803	0.0739803	0.0927564	0.0742454	0.0556823	0.0742454	0.0556823
0.1480091	0.0554883	0.0554883	0.0739819	0.0739790	0.0924730	0.0739790	0.0739790	0.0925012	0.0739784	0.0554848	0.0739784	0.0554848
A	B	B	C	C	D	C	C	D	C	B	C	B

Avec : A={S_1 } , B={S_2 ,S_3 ,S_{11},S_{13}} , C={S_4 ,S_5 ,S_7 ,S_8 ,S_{10},S_{12}} et D={S_6 ,S_9 }

En étant parti de n'importe quel nœud et en naviguant assez longtemps dans le graphe nous obtenons une probabilité de l'ordre de 0.1480091 d'arriver sur S_1 . Nous pourrions dire que S_1 ne peut pas être "illuminé"/"activé" davantage que 14,8% (1 x 14,8% = 14,8%). Pour les éléments de B cela vaut 5,5% (4 x 5,55% = 22,2%). Pour ceux de C, cela vaut 7,4% (6 x 7,4% = 44,4%). Pour D cela vaut 9,25% (2 x 9,25% = 18,5%).

Pour une puissance de DD moindre, ces valeurs peuvent être soit plus élevées soit quasiment nulles. Les classes les plus importantes (en pourcentage individuel sont A (14,8%), puis D (9,25%), puis C (7,4%) et enfin B (5,55%). Nous pouvons interpréter cela en disant que S_1 est un nœud d'articulation important (qui concentre les chemins: un hub), Viennent ensuite plus modestement S_6 et S_9 . Ces classes de valeurs caractérisent la nature de "hub" d'un nœud mais absolument pas son appartenance à une même composante que ceux de sa classe.

De même, nous poursuivons notre démarche pour représenter visuellement ces résultats en appliquant une analyse en composante principale.

En consultant les coordonnées sur l'axe des x, nous arrivons à repérer les mêmes regroupements, mais nous avons représenté davantage la corrélation sur l'axe des y. Ainsi en termes d'ACP, nous avons gardé les 2 premiers axes factoriels dominants qui permettent de garder la totalité de la dispersion (on parle d'inertie) de tout le nuage des points.

X	Y	Node
-10.314	-0.0028072	S1
3.0949	-0.010477	S2
3.0949	-0.010477	S3
0.4132	-0.010041	S4
0.41322	-0.010341	S5
-2.2685	-0.0098616	S6
0.41322	-0.010341	S7
0.41322	-0.010341	S8
-2.2704	0.019808	S9
0.41145	0.017579	S10
3.0936	0.0098586	S11
0.41145	0.017579	S12
3.0936	0.0098586	S13

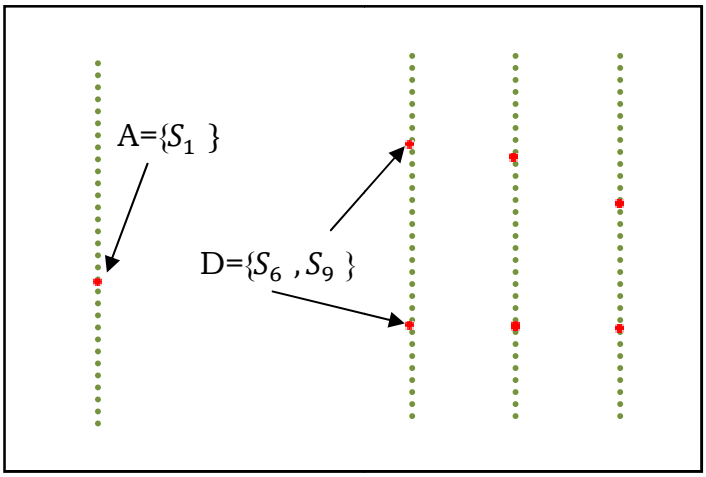


Figure 30. *Copies d'écran des résultats obtenus par notre programme. Vérification des résultats théoriques*

5.6.2 Application à l'univers des mots

Relation de synonymie entre "big" et "fat"

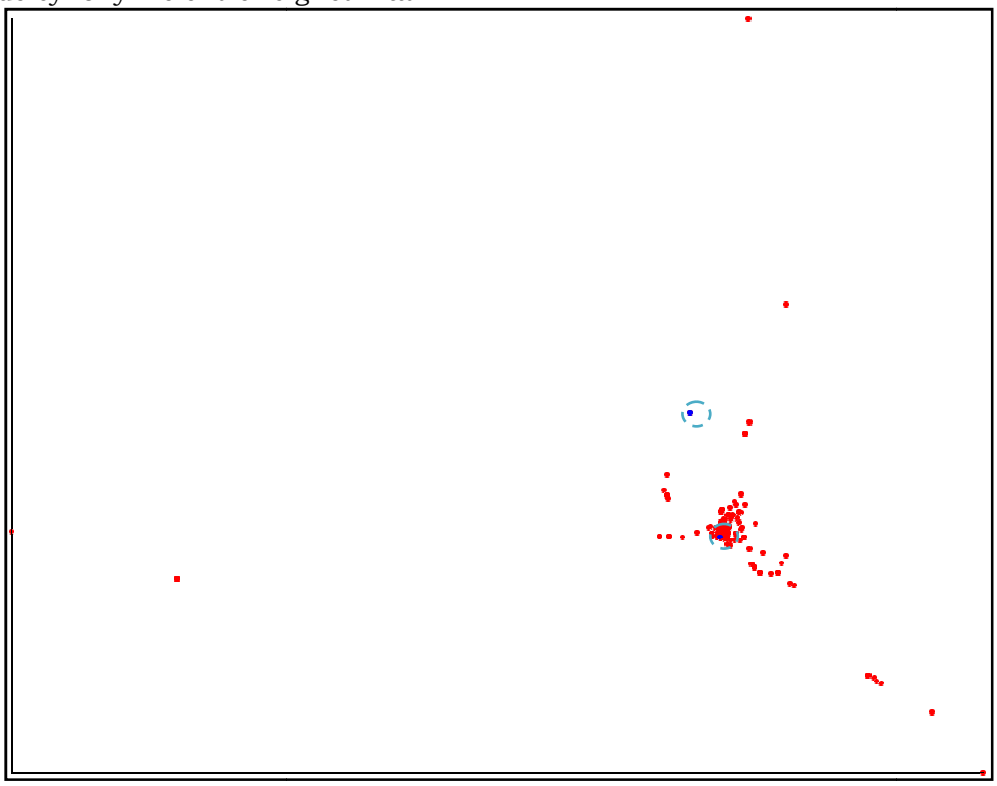


Figure 31. *Nuage de point correspondant à* DD^1

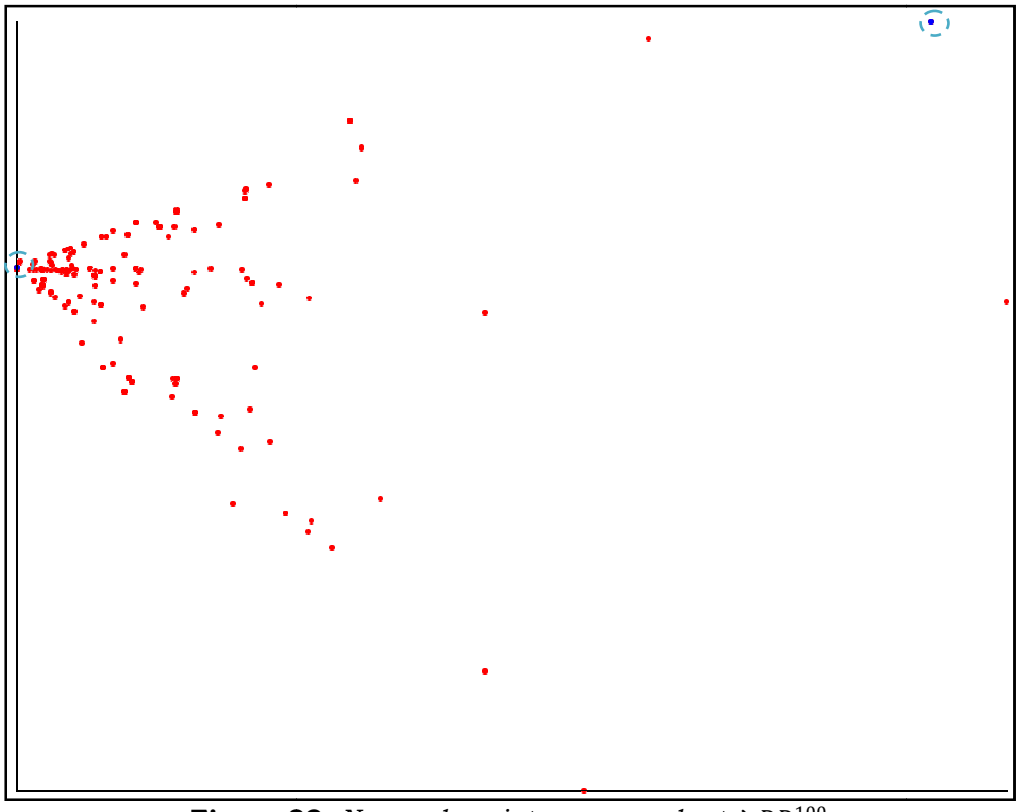


Figure 32. *Nuage de point correspondant à* DD^{100}

Le graphe ainsi formé est constitué de 264 nœuds. La première figure constitue une représentation spatiale de ces nœuds tout en explicitant la relation de synonymie entre eux et en fournissant une approximation visuelle de leur distance sémantique. Les deux mots initiaux sont repérés en bleu sur le graphe. Nous avons décidé délibérément de ne pas afficher les mots sur les points parce que plusieurs points se superposent, ce qui conduit à surcharger la figure et la rendre illisible.

La deuxième figure fournit des informations plus précises. En fait, après avoir parcouru le graphe 100 fois, la figure nous indique un chemin potentiel (généralement le plus petit entre les 2 nœuds en question).

De plus, le nuage de points vérifie que la relation de synonymie est perçue comme un continuum entre les différents mots, ainsi permettant de passer d'un sens à un autre. Outre les résultats discutés auparavant concernant les nœuds concentrateurs, nous remarquons que le nœud (le mot) *fat* est vraiment un mot polysémique vu l'ensemble des nœuds qu'il a pu engendrer.

Selon le nuage aussi, nous remarquons que globalement le mot *fat* a 3 sens différents, chacun correspondant à une direction. En effet, les mots qui sont synonymes et qui constituent un sens commun vont être regroupés ensemble et vont de même être à une distance approximativement égale de tous les autres synonymes ayant un sens différent.

5.6.3 Application à l'univers du WWW

Analyse de lien du site http://antoun.yaacoub.org

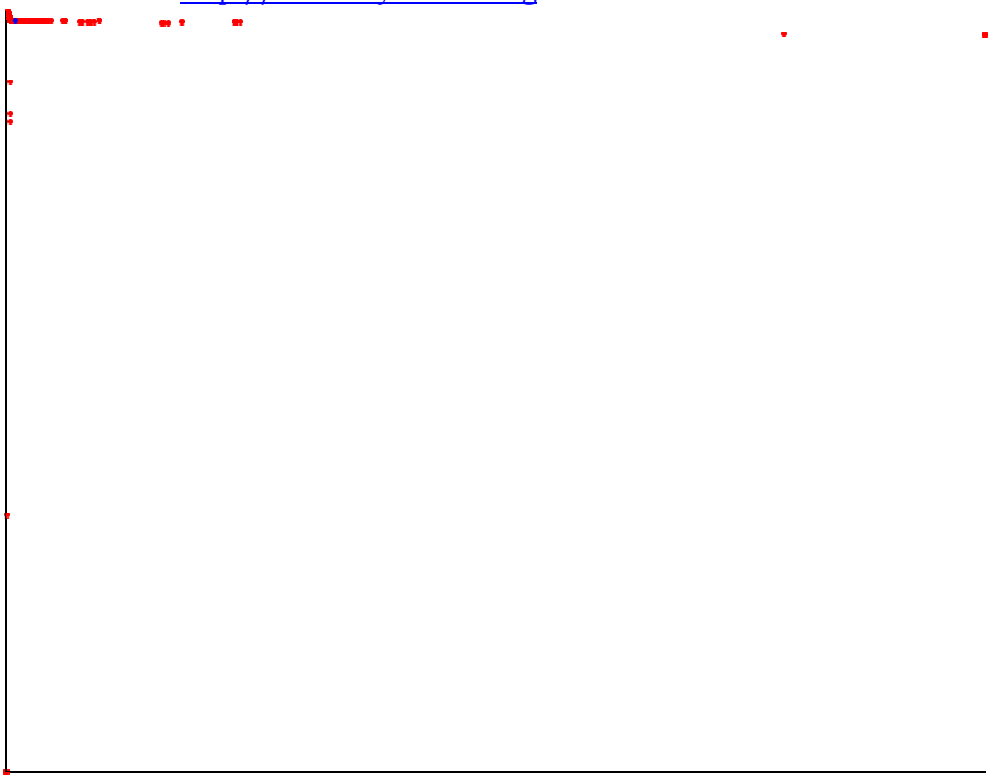


Figure 33. *Nuage de point correspondant à* DD^1

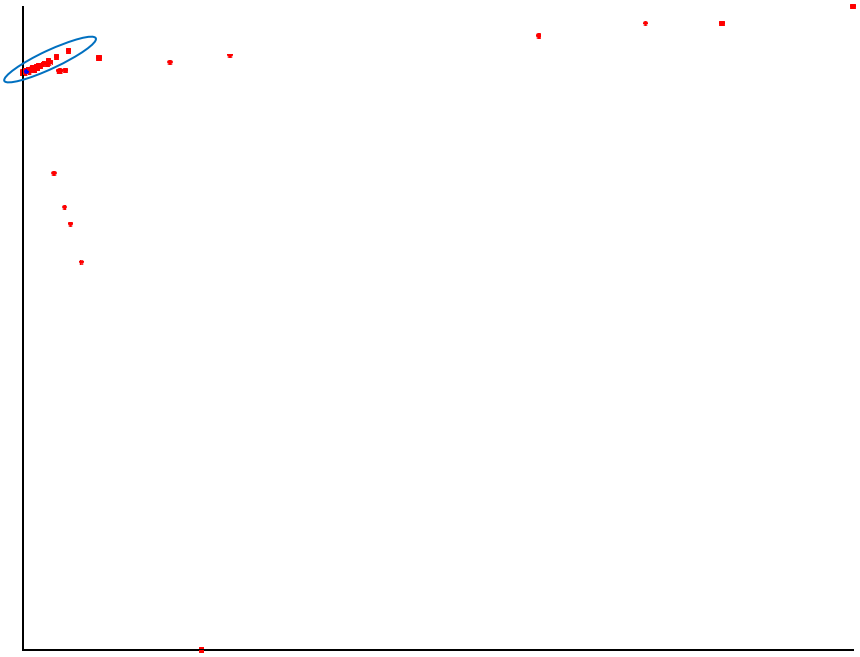


Figure 34. *Nuage de point correspondant à* DD^{100}

Le graphe du site web est constitué de 1637 nœuds ou pages webs.

La première figure constitue une représentation spatiale de la disposition des pages en termes de liens les unes par rapport aux autres. Le nœud initial est repéré en bleu sur le graphe. Nous remarquons que les points dans ce nuage suivent 2 directions orthogonales, ceci relève que les points sur chaque direction n'ont aucun lien avec ceux de la deuxième direction.

La deuxième figure fournit des informations plus précises quant à la relation des pages sur chaque direction. En fait, après avoir parcouru le graphe 100 fois, la figure nous indique que l'ensemble des points disposés suivant la direction horizontale dans la figure une n'entretiennent pas une relation linéaire entre eux (c'est-à-dire que nous ne passons pas d'une page web à une autre en suivant qu'un seul lien). La figure 34 dresse de façon plus exacte la relation en termes d'analyse de liens entre les différentes pages sur chaque direction. Nous arrivons de même à repérer un cluster très dense (repéré par un ovale sur la figure 34). Les pages dans ce cluster sont fortement interconnectées.

Chapitre 6
Conclusion

Le travail effectué dans ce mémoire, s'inscrit dans le cadre du développement d'interface graphique permettant d'étudier plusieurs relations dans plusieurs univers tel que l'univers des dictionnaires et celui des pages web. Or, les graphes induits de ces univers ont en commun des caractéristiques bien particulières. On dit qu'ils sont de type petit monde (Small World - SW). En effet, ces propriétés de SW ne sont vérifiées que lorsque la relation étudiée est la synonymie (dans le cas de l'univers des dictionnaires) ou navigationelle (dans l'univers des pages web). Nous avons proposé une approche qui prend en compte l'aspect SW des graphes en vue de visualiser localement et globalement un graphe RPMH tout en tenant compte de sa structure globale. La solution que nous avons proposée se base sur l'utilisation d'une matrice de transition. Parcourir x fois ce graphe revient à multiplier la matrice de transition x fois. La matrice obtenue en tant que matrice de coordonnées dans $\mathbb{R}^n$ contient une information calculée sur l'ensemble du graphe (la multiplication de la matrice par elle-même renvoie des informations sur la relation de tous les nœuds entre eux) qu'on a pu représenter dans $\mathbb{R}^2$ au moyen d'une Analyse en Composante Principale (ACP) en gardant les 2 premiers axes.

Après avoir explicité les modules implémentés, nous avons énuméré en commentant différent cas de figures d'utilisation des relations sémantiques et navigationelles obtenues à partir de ces modules.

Dans notre contribution, nous avons pu vérifier exactement notre approche en confrontant les résultats théoriques aux résultats pratiques issus de notre interface. Notre interface permet de constituer une représentation spatiale de ces nœuds tout en explicitant la relation de synonymie entre eux et en fournissant une approximation visuelle de leur distance sémantique. De plus, le nuage de points vérifie que la relation de synonymie est perçue comme un continuum entre les différents mots, permettant ainsi de passer d'un sens à un autre en empruntant les nœuds du graphe. Nous avons pu de même repérer sur le graphe des nœuds concentrateurs et décider si un mot est polysémique, et localiser les denses clusters. De plus, une interprétation possible pour les directions empruntées par les points dans le graphe, suggère que le nombre de directions relève du nombre d'acception (sens) du nœud initial et ainsi de sa polysémie.

Outre la capacité de notre outil d'étudier différentes relations dans les univers cités précédemment, il permet de générer des fichiers contenant les résultats calculés (les nœuds d'entrée, la matrice de transition, les coordonnées calculées issues de l'ACP,…) et d'accepter des fichiers dont la structure est conforme à ce qui est proposée (chaque fichier possède une structure particulière qu'il faut respecter). Notre outil est donc dans ce sens « générique » : tout univers satisfaisant ces conditions peut ainsi être étudié par notre outil.

Cependant, nous avons été confrontés à des problèmes purement techniques. En effet, les machines utilisées ne sont pas adaptées à faire des calculs assez puissants et longs et nécessitant de grande capacité de mémoire RAM, ce qui nous a empêchés de vérifier notre approche sur l'univers des textes et de proposer des exemples plus élaborés vu que la taille des graphes explose rapidement.

Les perspectives envisageables de ce travail, portent essentiellement sur trois volets :

- Le premier porte sur le fait de trouver une façon efficace de construire la matrice correspondante au graphe de synonymie issu de WordNet. En effet, pour un mot donné, sa relation de synonymie avec les autres mots en WordNet est fixe (2 mots sont synonymes ou ne le sont pas). Donc chaque mot réserve une ligne dans la matrice (cette ligne contient généralement beaucoup de valeurs 0 et peu de 1) qui est toujours fixe ; elle ne change pas d'une application à une autre. Ce qui suggère de travailler sur une

méthode qui pourra par exemple stocker dans une base de données ces matrices statiques (on pourra de même la calculer pour l'ensemble des lexiques) ou proposer une méthode basée sur les matrices creuses pour mieux exploiter le graphe et accélérer le calcul.

- Le deuxième concerne le couplage de cet outil à un moteur de recherche. Notre outil permet de détecter de très denses clusters (des pages fortement connectées, ou de mots qui sont synonymes entre eux). On pourrait exploiter cet outil pour proposer une méthode d'expansion de la requête, et dans le cas du web de combiner ses résultats avec les notions de *Hubs* et *authorities* [11] pour proposer un nouveau système de recherche d'information.
- Le troisième suggère une exploitation possible de l'outil ainsi construit en élaborant un crawler paramétrable sur différents critères (type de lien/balise, type de parcours, …) appliqué à des documents structurés (pages html, dictionnaire xml, texte, …) [86].

Bibliographie

[1] Eléments de sémantique
John Lyons
Larousse, 1978

[2] Introduction to WordNet: An On-line Lexical Database
George A. Miller, Richard Beckwith, Christiane Fellbaum,
Derek Gross, and Katherine Miller
In: International Journal of Lexicography 3 (4), 1990, pp. 235 – 244

[3] Introduction à la lexicologie. Sémantique et morphologie
Alise Lehmann, Françoise Martin-Berthet
Lettres sup 2008

[4] Grammaire - Tome 1 . Phonologie, Morphologie, Lexicologie
Gardes Tamine, Joëlle
Armand Colin, Cursus, 1990

[5] Principes et méthodes de statistique lexicale.
Muller C.
Paris : Hachette. (1977).

[6] «Une mesure de la distance intertextuelle : la connexion lexicale », Le nombre et le texte.
Brunet É.
Revue informatique et statistique dans les sciences humaines 24 : 81-116. (1988).

[7] Étienne Brunet, « Peut-on mesurer la distance entre deux textes ? », Corpus, Numéro 2, La distance intertextuelle - décembre 2003, 2003, [En ligne], mis en ligne le 15 décembre 2004. URL : http://corpus.revues.org/document30.html. Consulté le 24 juillet 2008.

[8] « Measures of similarity, dissimilarity and distance »,
Gower J.-C.
in Kotz S., Johnson N.-L. & Read C.-B. (eds), Encyclopedia of Statistical Sciences, vol. 5. New York : Wiley, 397-405. (1985).

[9] Cyril Labbé et Dominique Labbé, « La distance intertextuelle », Corpus, Numéro 2, La distance intertextuelle - décembre 2003, 2003, [En ligne], mis en ligne le 15 décembre 2004. URL : http://corpus.revues.org/document31.html. Consulté le 24 juillet 2008.

[10] Xuan Luong et Sylvie Mellet, « Mesures de distance grammaticale entre les textes », Corpus, Numéro 2, La distance intertextuelle - décembre 2003, 2003, [En ligne], mis en ligne le 15 décembre 2004. URL : http://corpus.revues.org/document34.html. Consulté le 24 juillet 2008.

[11] Authoritative Sources in a Hyperlinked Environment.
Jon M. Kleinberg
IBM Research. Report RJ 10076, May 1997.

[12] Introduction to Informetrics,
L. Egghe, R. Rousseau
Elsevier, 1990.

[13] Bibliographic coupling between scientific papers
M.M. Kessler,
American Documentation, 14(1963), pp. 10-25.

[14] Co-citation in the scientific literature: A new measure of the relationship between two documents
H. Small
J. American Soc. Info. Sci., 24(1973), pp. 265-269.

[15] Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace
R. Larson
Ann. Meeting of the American Soc. Info. Sci., 1996.

[16] Life, death, and lawfulness on the electronic frontier
J. Pitkow, P. Pirolli
Proceedings of ACM SIGCHI Conference on Human Factors in Computing, 1997.

[17] HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering
R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, D.K. Gi ord
Proceedings of the Seventh ACM Conference on Hypertext, 1996.

[18] The structure of the scientific literatures I. Identifying and graphing specialties
H. Small, B.C. Grith
Science Studies 4(1974), pp. 17-40.

[19] Analysis of a complex statistical variable into principal components
H. Hotelling
J. Educational Psychology, 24(1993), pp. 417-441.

[20] Principal Component Analysis.
I.T. Jolliffe.
Springer-Verlag, 1986

[21] Subject and Citation Indexing. Part II: The optimal, cluster-based retrieval performance of composite representations
W.M. Shaw
J. American Soc. Info. Sci., 42(1991),pp. 676-684.

[22] Co-citation in the scientific literature: A new measure of the relationship between two documents
H. Small
J. American Soc. Info. Sci., 24(1973), pp. 265-269

[23] Silk from a sow's ear: Extracting usable structures from the Web
P. Pirolli, J. Pitkow, R. Rao
Proceedings of ACM SIGCHI Conference on Human Factors in Computing, 1996.

[24] Spectral Graph Theory
F.R.K. Chung
AMS Press, 1997.

[25] Indexing by latent semantic analysis
S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman
J. American Soc. Info. Sci., 41(1990), pp. 391-407.

[26] Information Retrieval
C.J. van Rijsbergen
Butterworths, 1979

[27] Enhanced hypertext classification using hyperlinks
S. Chakrabarti and B. Dom and P. Indyk
ACM SIGMOD Conference on Management of Data, 1998

[28] Mining the Link Structure of the World Wide Web.
Chakrabarti, Dom, Gibson, Kleinberg, Kumar, Raghavan, Rajagopalan, Tomkins
IEEE Computer, August 1999.

[29] Introduction to Information Retrieval
By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze
Cambridge University Press 2008

[30] Could It Be A Big World After All? The "Six Degrees Of Separation" Myth.
Judith S. Kleinfeld
Society 2002

[31] Algorithmes de routage et modèles aléatoires pour les graphes petits mondes. Emmanuelle Lebhar
INRIA – CNRS : UMR5668 – Université Claude Bernard - Lyon I – Ecole Normale Supérieure de Lyon - ENS Lyon. 2005

[32] On power-law relationships of the internet topology.
M. Faloutsos, P. Faloutsos, and C. Faloutsos
Computer Communications Rev., 29 :251–262, 1999.

[33] On the bias of traceroute sampling ; or, power-law degree distributions in regular graphs.
D. Achlioptas, A. Clauset, D. Kempe, and C. Moore.
In Proceedings of the 37th ACM Symposium on Theory of Computing (STOC), 2005.

[34] Collective dynamics of 'small-world' networks.
Watts, D.J. & Strogatz, S.H.
Nature, 393, 440-442. 1998

[35] On random graphs.
P. Erdös and A. Rényi.
Publicationes Mathematicas 6, 290(7), 1959.

[36] The Small-World Phenomenon: An Algorithmic Perspective.
Jon Kleinberg
Proc. 32nd ACM Symposium on Theory of Computing, 2000.
Cornell Computer Science Technical Report 99-1776 (October 1999).

[37] Emergence of scaling in random networks.
A.-L. Barabási and R. Albert.
Science, 286(509–512), 1999.

[38] Diameter of the World-Wide Web.
Réka Albert, Hawoong Jeong, Albert-László Barabási
Nature, 401: 130-131. 1999

[39] The Small World Web.
Lada A. Adamic
Lecture Notes In Computer Science; Vol. 1696
Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries
Pages: 443 - 452 1999

[40] Power-Law Distribution of the World Wide Web
Lada A. Adamic, Bernardo A. Huberman
Science 287, 2115a. 2000

[41] Global organization of the Wordnet lexicon.
Mariano Sigman & Guillermo A. Cecchi
Proc. National Academy of Sciences USA, 99(3): 1742-1747. 2002

[42] Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes.
Sabine Ploux, Bernard Victorri
Traitement automatique des langues, 39/1, p.161-182, 1998

[45] Hierarchical organization in complex networks.
Ravasz and Barabasi
Phys. Rev. E 67, 026112, 2003

[46] Ballades aléatoires dans les Petits Mondes Lexicaux.
Bruno Gaume
in I3 Vol 4, n°2, 2004

[47] Cartographier la forme du sens dans les petits mondes Lexicaux.
Bruno Gaume
JADT 2006, p 541-465. 2006

[48] Polysémie et calcul du sens.
Fabienne Venant
Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT), 2004

[49] Représentation de graphes par ACP granulaire.
Gaume, Ferré
actes d'EGC 2004 : 4èmes journées d'Extraction et de Gestion des Connaissances, Clermont Ferrand, 20-23 Janvier 2004

[50] Forms of Meaning, Meaning of Forms.
Gaume Bruno, Duvignau Karine, Gasquet Olivier
Journal of Exprimental and Theoretical Artificial Intelligence, 14(1), p. 61-74, 2002

[51] Synonym Extraction Using a Semantic Distance on a Dictionary.
Philippe Muller, Nabil Hathout, Bruno Gaume
Workshop on TextGraphs, at HLT-NAACL 2006, pages 65–72

[52] Etude de la synonymie par l'extraction de composantes N-connexes dans les graphes de dictionnaires.
Awada Ali, Chebaro Bilal
JEL2004, Nantes, France.

[53] Regroupement de synonymes en composantes de sens dans un dictionnaire.
Ali Awada
A paraître

[54] Guest Editor's Introduction IEEE Computer Graphics and Applications
Tamara Munzner
Special Issue on Information Visualization, Jan/Feb 2002

[55] Graphically displaying text
S.G. Eick
In: Journal of Computational and Graphical Statistics, vol 3, pp. 127–142. 1994

[56] Information Visualization.
John Stasko 2004 syllabus for CS7450
http://www.cc.gatech.edu/classes/AY2004/cs7450_spring/ Spring 2004. Retrieved 1 September 2008.

[57] Readings in Information Visualization: Using Vision to Think
Card, Mackinlay, and Shneiderman
1999.

[58] Visualization of search results in document retrieval systems
Zamir, O.
General Examination, University of Washington, 1998.

[59] A taxonomy of visualization techniques using the data state reference model
Chi Ed. H.
INFOVIS'2000, Salt Lake City, pp 69-75, October, 2000.

[60] Visualisation interactive d'information
Hascoët M., Beaudouin-Lafon M.
Revue Information-Interaction-Intelligence (I3), « A Journal in Information Engineering Sciences », 1(1), 2001

[61] Interface Adaptative Pour L'aide A La Recherche D'information Sur Le Web
Max Chevalier
Thèse 2002

[62] TileBars: visualization of term distribution information in full text information access
Hearst M.A.
ACM Conference on Human Factors in Computing Systems (SIGCHI), Denver CO, pp 59-66, May, 1995.

[63] Using categories to provide context for full-text retrieval results
Hearst M.A
Conférence Recherche d'Informations Assistée par Ordinateur (RIAO) : Intelligent Multimedia Information Retrieval Systems and Management, Rockefeller University, NY, October, 1994.

[64] InfoCrystal: a visual tool for information retrieval and management
Spoerri A.
International Conference on Information Knowledge and Management (CIKM), Washington, USA, ISBN 0897916263, pp 11-20, November 1-5, 1993.

[65] Visualization of a document collection: the VIBE system
Olsen K.A., Korfhage R.R., Sochats K.M., Spring M.B.,
Williams J.G.
Information Processing and Management Journal (IPM), 29(1), pp 69-81, 1993.

[66] VR-Vibe: a virtual environment for co-operative information retrieval
Benford S., Snowdon D, Greenhalgh C., Knox I., Brown C.
EuroGraphics, vol. 14(3), pp 349-360, 1995.

[67] Interactive 3D visualization for document retrieval
Cugini J.V., Piatko C., Laskowski S.
Actes CIKM, Rockville MD, 1996.

[68] Visualisation globale de collections de documents sous forme d'hypercube
Mothe J., Chrisment C., Alaux J.
Revue Extraction des Connaissances et Apprentissage (ECA), vol. 4/2001, ed. Hermes Science, ISBN 2746204061, pp 131-142, 2002.

[69] Bead: explorations in information visualization
Chalmers M., Chitson P.
15th International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, pp 330-337, 1992.

[70] Grouper: a dynamic clustering interface to web search results
Zamir O., Etzioni O.
Computer Networks, vol. 31(11-16), pp 1361-1374, May, 1999.

[71] Self-organised formation of topologically correct feature maps
Kohonen T.
Biological Cybernetics, vol. 43, pp 59-69, 1982.

[72] Visualization of search results: a comparative evaluation of text, 2D and 3D interfaces
Sebrechts M., Cugini J.V, Vasilakis J., Miller M.S., Laskowski S.J.
ACM SIGIR, Berkley, pp 3-10, 1999.

[73] The interface of the future
Levialdi S., Badre A.N., Chalmers M., Copeland P., Mussio P., Salomon C.
ACM International Conference on Advanced Visual Interface (AVI), pp 200-205, June, 1994.

[74] Représentation multiples d'une grande quantité d'information
Vernier F., Nigay L.
9ème journées Interaction Homme-Machine (IHM), pp 183-190, Futuroscope Poitiers, France, 10-12 Septembre, 1997

[75] H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space.
Tamara Munzner
Proceedings of the 1997 IEEE Symposium on Information Visualization, October 20-21 1997, Phoenix, AZ, pp 2-10, 1997.

[76] Interactive Visualization Of Large Graphs And Networks.
Tamara Munzner
PHD Dissertation, Stanford University , 2000

[77] Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space.
Tamara Munzner
ACM. Proceedings of the First Annualy mp200 on the VRML Modeling Language, San Diego, CA, December 14-15 1995, special issue of Computer Graphics, ACM SIGGRAPH, New York, pp. 33-38.

[78] Fractal Approaches for Visualizing Huge Hierarchies
Hideki Koike, Hirotaka Yoshihara
Proceedings of the 1993 IEEE Symposium on Visual Languages, pp.55-60, IEEE/CS, 1993.

[79] The Role of Another Spatial Dimension in Software Visualization.
H. Koike.
ACM Transaction on Information Systems, Vol. 11, No. 3,, July 1993.

[80] The Web as a graph: measurements, models, and methods.
Kleinberg, Kumar, Raghavan, Rajagopalan, Tomkins
In Proceedings of the Fifth International Conference on Computing and Combinatorics, Tokyo July 26-28, 1999 (COCOON'99). Berlin: Springer-Verlag, pages1-17

[81] The Structure of the Web.
Jon Kleinberg, Steve Lawrence
Science vol 294 2001

[82] The Small-World Phenomenon and Decentralized Search.
Jon Kleinberg
SIAM News, Volume 37, Number 3, April 2004

[83] The Small-World of Human Language.
Ramon Ferrer i Cancho & Ricard V. Solé
Proceedings of The Royal Society of London. Series B, Biological Sciences. 2001

[84] Kevin Bacon, the Small-World, and Why It All Matters.
Santa Fe Institute Bulletin 1999

[85] How Popular is Your Paper? An Empirical Study of the Citation Distribution.
S. Redner
European Physical Journal B, 4, 131-134 (1998).

[86] Focused Crawling Using Context Graphs.
Diligenti, Coetzee, Lawrence, Giles, Gori
Proc. 26th Intl. Conf. on Very Large Databases (VLDB), 2000.

[87] Scale-free characteristics of random networks: the topology of the world-wide web. Barabasi, Albert, Jeong
Physica A, 281, 2115 (2000).

[88] Cartographie lexicale pour la recherche d'information.
Jean Véronis
Actes TALN 2003, p.265-275

[89] Espaces sémantiques et représentation du sens.
Bernard Victorri
Textualités et nouvelles technologies, éc/arts, 3, 2003

[90] Géométriser le sens.
Fabienne Venant
Les Journées Graphes, Réseaux et Modélisation, ESPCI, Paris. 2003

[91] Représentation géométrique de la synonymie.
Bernard Victorri et Fabienne Venant
Le Français Moderne, 1 (2007)

[92] Polysémie lexicale.
Guillaume Jacquet, Fabienne Venant, Bernard Victorri
Hermès 2005

[93] Semantic memory in semantic information processing
M.R. Quillian
M.I.T. Press, 1968.