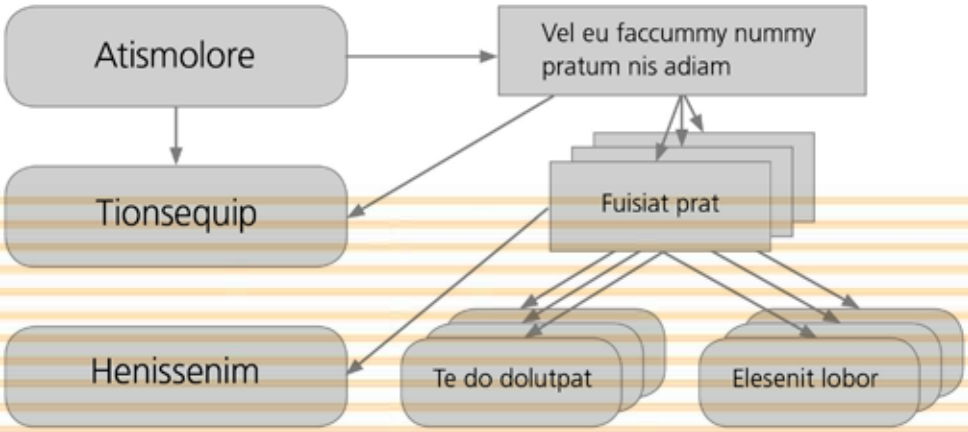
document de travail

07.1001
NOVEMBRE 2007

Connaissance, théorie de l'information et hypertextes : histoire d'une lecture sélective

Luc-Olivier Pochon & Alain Favre



irdp
Institut de recherche et de documentation pédagogique

Connaissance, théorie de l'information et hypertextes : histoire d'une lecture sélective

Luc-Olivier Pochon & Alain Favre

IRDP
Faubourg de l'Hôpital 43
Case postale 556
CH-2002 Neuchâtel

Tél. (+41) (0) 32 889 86 18
Fax (+41) (0) 32 889 69 71

E-mail: documentation@irdp.ch
http://www.irdp.ch

Cette publication est également disponible sur le site IRDP :

http://www.irdp.ch/

TABLE DES MATIÈRES

1. Introduction

Préambule

A la fin des années 1980 et au début des années 1990, nous étions engagés l'un dans l'analyse et l'archivage d'images, l'autre dans des travaux concernant l'enseignement « intelligemment » assisté par ordinateur[1]. Il nous est apparu qu'en amont de ces applications manquait une base théorique qui permettrait de jauger les connaissances apportées, gérées, transformées dans l'interaction des hommes et des machines. L'idée même de mesure de connaissance, et une vague idée de relation avec la notion d'information, était aussi dans nos préoccupations. L'histoire de ce travail à la recherche de principes fondateurs est l'objet de ce document. S'il s'agit plus d'une histoire que d'un aboutissement, c'est que l'entreprise, menée comme « à côté », s'est avérée plus inattendue que prévu, ceci pour deux raisons. La première est que le nombre de pistes possibles est innombrable. Plus qu'à un manque de travaux, comme nous l'imagions un peu naïvement, c'est face à une abondance d'articles que nous nous sommes trouvés confrontés. Mais en même temps, cette masse de documents était composée de travaux très disparates : esquisses de questions pertinentes mais peu abouties, techniques prometteuses cantonnées à des applications très spécifiques, etc. La deuxième raison est l'arrivée du Web[2]. Celui-ci, en même temps qu'il offrait les ressources disponibles en ligne, faisait éclore une multitude de travaux liés à notre sujet. Nous avons commencé à crouler sous le nombre des nouveaux chantiers. Nous étions entraînés dans la dynamique que nous entendions observer de l'extérieur.

Afin de capitaliser notre parcours, nous avons régulièrement enrichi un « thema »[3] sur le site de l'IRDP. Il nous a semblé utile d'en faire la synthèse. C'est le but de ce document qui, de ce fait, est souvent conjugué au passé. Ce travail de synthèse est aussi l'occasion de choisir les pistes prometteuses dans plusieurs domaines d'application. Avant de présenter la façon dont nous pensons les décrire à travers la structure de ce document, nous commencerons par préciser les questions initiales.

[1] Avec une collaboration dans le cadre d'un projet WBO, mené sur le concept I³ (images, intelligence, interaction).

[2] Le Web n'a pas seulement « révolutionné » les modes de communication, mais il a également accéléré l'essor de travaux concernant l'indexation des documents et, de façon plus générale, l'architecture documentaire à partir du « virage » SGML (Marcoux, 1996).

[3] http://www.irdp.ch/thema/htx-info.htm

Les questions initiales

Avec l'élargissement des réseaux informatiques et la banalisation de leur emploi dès la fin des années 1980, les facilités d'accès à l'information n'ont cessé d'augmenter en même temps que l'offre connaissait une croissance sans précédent. Dans ce processus, le statut de la connaissance devenait de plus en plus incertain, sachant qu'à un accroissement de l'information ne correspond pas forcément une augmentation des connaissances ; « abondance d'informations, disette de connaissances » pour reprendre une expression pessimiste (Foray, 1997)[4]. Dès lors, nous nous sommes posé la question de comprendre comment apprécier, mesurer, voire quantifier ce que ce flux sans cesse croissant d'informations pouvait apporter à la connaissance. Dans un premier temps, il s'agit de préciser le sens de ces termes.

De ce point de vue, la formule (l'équation fondamentale) proposée par Brookes (1980a) a fourni un point de départ :

$$K[S] + \Delta I = K[S + \Delta S] \quad (1.I)$$

Cette formule sous-entend que la connaissance K est liée à une « structure » S[5] et que cette structure s'enrichit par l'apport d'un élément d'information ΔI. Cette formule présuppose que l'information est mesurée dans la même unité que la connaissance. Elle offre principalement la nature de la dépendance entre information et connaissance sans nécessairement préciser le contenant de cette connaissance. Cette relation pourrait aussi bien représenter l'activité d'un sujet apprenant (K étant la connaissance organisée dans son cerveau) par accès à de l'information (on parlera dans ce cas du *modèle du cogniticien*) qu'à l'enrichissement d'une base de données (on parlera dans ce cas du *modèle du documentaliste*).

Avant d'aborder ce problème du support, il peut sembler préférable de prendre une notation évolutive, comme le font les chimistes :

$$(\Delta I + K[S]) \rightarrow K[S + \Delta S] \quad (1.II)$$

Un nouvel examen demande de s'interroger sur la façon dont cette formule évoque la modification de structure. Il semble préférable de parler de nouvelle structure en laissant ouvert la nature de l'opération, sans évacuer notamment un « recuit » de l'ancienne structure, et donc de noter K[S'] à la place de $K[S + \Delta S]$. L'apport d'information provoque une réorganisation de la structure et modifie par là même la connaissance associée. En poursuivant la réflexion, il s'agit de préciser la notion de structure. Les réseaux sémantiques se révèlent de bons candidats, mais les applications

[4] Sans oublier que l'information n'est pas « *context free* », ce que le discours sur les facilités de l'accès à l'information ne prend pas en compte (pour une discussion un peu polémique à ce propos, voir Postman [1993]).

[5] « *Knowledge is a linked structure of concepts* » (Brookes, 1980a).

pratiques montrent que cette option est vraisemblablement trop exigeante. Il s'agit donc de considérer dans le modèle la possibilité d'intégrer des informations peu structurées. Ce double aspect de la connaissance, structuré et amorphe, nous a conduit à proposer l'hypertexte comme support de base à la connaissance. L'hypertexte possède les deux exigences : des entités en relation pour la partie structurée, les entités (les nœuds de l'hypertexte) pour intégrer des informations qui ne participent pas forcément de la même logique d'organisation que la structure qui les contient. Par contre, ces informations peuvent participer à l'alimentation de la structure générale. Elles peuvent aussi à leur tour posséder leur propre structure (aspect fractal).

L'équation du modèle devient : $(\Delta I + K[H]) \rightarrow K[H'] \ (1.III)$

A partir de cette reformulation de la relation de Brookes, nous nous sommes attachés à caractériser la structure d'un hypertexte comme un « ensemble d'unités d'information reliées entre elles par des liens qui donnent naissance à la structure. »

Domaine des applications

Pour ce qui concerne les modèles, dans le sens donné par les logiciens[6], nous garderons aussi bien celui du cogniticien que celui du documentaliste. Chaque fois, que l'on voudra sortir du cadre formel pour y intégrer un interprétant, c'est-à-dire des partenaires humains, va se poser la délicate question du choix du système de référence dans l'appréciation des quantités d'information et de connaissance. Dans ce travail, deux intentions complémentaires s'épaulent pour guider notre progression : l'hypertexte vu par le documentaliste et l'hypertexte vu par le cogniticien[7].

Pour le premier (le documentaliste) la connaissance est un ensemble de documents. Il met l'accent sur l'indexation et l'accès à ces documents. Son but est de fournir des clés d'accès qui correspondent aux « concepts » importants. L'approche est celle d'une classification des documents influencée par le domaine concerné (bibliothèque de droit, archives de procès-verbaux, ensemble d'articles consacrés à un sujet[8], etc.). Les concepts importants sont choisis dans les documents en fonction du contexte et permettent de « cristalliser » la structure de l'hypertexte. L'organisation des concepts, et donc en partie la structure de l'hypertexte, est mise à jour par divers procédés aussi bien automatisés que « manuels »[9].

[6] C'est-à-dire des applications concrètes qui « instancient » chacun des symboles de la relation de Brookes.

[7] Mais d'autres modèles seront évoqués : une image avec l'ensemble de ses points et de la relation de proximité, une batterie de tests avec l'ensemble de ses items et la relation de dépendance entre items.

[8] Nous traiterons d'un exemple de ce type avec l'étude de cas « UTOPIA ».

[9] Dans le cadre d'une perception de l'information basée sur les publications, Lenski (2004), après un détour par une approche sémiotique, reprend la relation de Brookes pour initier une interprétation de chacun de ses constituants. Dans son approche, il considère non seulement l'information qui alimente la connaissance, mais également l'information qui est issue de la connaissance : « *information is the result of a problem-driven differentiation process in a structured knowledge base* ».

Le second (le cogniticien) part d'une acception plus large de ce qu'est une connaissance. Son propos est de modéliser un sujet (apprenant). Les concepts qui « décrivent » l'unité d'information, ainsi que les concepts référents, sont plus difficiles à définir parce que les unités d'information sont difficilement traduisibles à un niveau symbolique.

Cette deuxième approche est surtout considérée parce que le cogniticien s'intéresse plus volontiers à la structure de l'hypertexte (de la connaissance) avec notamment les emboîtements descendants (chaque « document » peut-être lui-même un hypertexte) et ascendants (un groupe de documents devient une unité d'information). Ce cas fournit également un cadre plus approprié à la question de la détermination de la part de la structure qui émerge « spontanément » et de la part due à des interventions extérieures, notamment par le choix et la désignation de concepts[10]. Ce problème est largement évoqué par Turkle (1995) à propos des deux approches, symbolique et « associationniste », rencontrées en intelligence artificielle.

Dépassant le cadre de cet article, ces considérations peuvent demander de s'interroger sur les idées de symbiose homme-machine exprimées par des pionniers tels que Licklider (1960), Engelbart (1962), Winograd & Flores (1990). Ces aspects seront brièvement évoqués en conclusion et reliés aux travaux abordant spécifiquement l'émergence des structures (morphogenèse et autopoïèse).

Du point de vue pratique, les travaux dans ce domaine sont nombreux, liés à la prolifération de données électroniques (*datamining, knowledge management*, etc.). Le choix de l'hypertexte comme modèle nous a conduit naturellement à prendre en considération le Web sur lequel se concentraient de plus en plus de travaux concernant les hypertextes, notamment les études liées aux moteurs de recherche et celles menées sous les auspices du « Web sémantique » (Berners-Lee, Hendler & Lassila, 2001).

En ce qui concerne les applications, nous privilégions le domaine de l'enseignement, ce qui rejoint des préoccupations pratiques, mais surtout, permet d'orienter le travail sur le sujet connaissant. Cela met à disposition tout le domaine qui se dédie à l'apprentissage humain et apprentissage de machine. D'autres approches spécifiques existent également. Outre les travaux actuels concernant les normalisations (LOM, IMS, SCORM, etc.)[11], une longue tradition existe à propos de la constitution de bases de connaissances, ou comme cela est parfois nommé, de viviers de connaissance (Vidal & al, 1998 ; Forte, 2002). Les problèmes identifiés concernent la granularité des « objets » pédagogiques, la fabrication des « agrégats » à partir de segments d'unités thématiques, la classification et la hiérarchie des objectifs, des contenus et des modèles d'apprentissage, la modélisation des interactions face à la machine.

[10] Dans le premiers cas, les requêtes de recherche peuvent également influencer la structure de l'hypertexte.
[11] Qui correspondent à l'approche DOM préconisée par Lenski (2004).

Plan du document

L'article reprend notre cheminement qui peut se décomposer en trois étapes.

Etape 1 : structure d'un hypertexte, l'approche « exacte »

Il nous est apparu naturel de baser notre travail sur un type particulier d'hypertexte construit à partir de deux relations « document × concept », et à propos de ces hypertextes, de parcourir les coefficients locaux « classiques » et d'en définir de nouveaux. Cette construction est présentée dans la partie 2. La partie 3 décrit un exemple qui nous a servi à tester quelques-uns des outils mis au point. Ce travail nous a également permis de nous habituer à différentes façons de « simplifier » la structure d'un hypertexte. Mais si certaines notions (par exemple, celle de quotient), sont issues de théories algébriques, il nous est vite apparu qu'une approche algébrique « exacte » ne peut pas être utile pour des hypertextes qui n'ont a priori aucune raison de suivre des structures simples (on pense ici à l'idée de la décomposition canonique d'un hypertexte en une famille d'hypertextes bien répertoriés).

La partie 4 propose une méthode liée au parcours aléatoire d'un hypertexte pour obtenir une caractérisation globale sous la forme d'une matrice de « circulation » S, qui est *grosso modo* la matrice d'adjacence du complété transitif du graphe induit. Cette matrice prend en compte l'hypertexte à partir duquel le graphe induit a été construit. Les parties 5 et 6 montrent comment la connectivité de l'hypertexte et la relation de voisinage peuvent être étudiées à partir de S avec différents degrés d'approximation. La partie 7 compare les méthodes classiques basées sur la décomposition spectrale à l'approche pragmatique utilisant S. La partie 8 termine cette première étape en comparant les informations que l'on peut extraire de S pour approcher les coefficients classiques de structure d'un hypertexte (compacité, linéarité, etc.).

Etape 2 : structure d'un hypertexte, les approximations

Les parties suivantes s'intéressent à préciser la structure en rappelant la décomposition papillon des graphes orientés (partie 9). Afin de posséder une famille d'hypertextes de référence, la construction d'hypertexte par agrégation est proposée (partie 10). La partie 11 examine la structure de différents hypertextes des deux types. La partie 12 s'intéresse à préciser les travaux effectués (notamment concernant les « communautés ») à propos du grand hypertexte constitué par le Web.

Etape 3 : les questions pour continuer

La partie 13 résume les différents secteurs auxquels notre modèle a été confronté. La partie suivante est prospective. Elle élargit le propos en offrant un panorama des théories et techniques permettant de parler de structures de l'hypertexte. Chacune d'elle pourra ultérieurement être

étudiée plus spécifiquement. La partie 15 montre que le modèle construit peut être utile à d'autres applications, notamment dans le domaine de l'éducation et de la formation.

La conclusion précise les pistes ouvertes quant à l'usage de la matrice S, ou de ses variantes, comme des éléments de la formule de Brookes.

2. Hypertextes de type « RD »

Définition

Dans un premier temps, nous proposons de considérer des hypertextes construits à partir de deux relations[12]. La première associe des descripteurs à chaque unité d'information. Les descripteurs sont pris dans un ensemble de concepts. La deuxième relation associe des référents aux unités d'information. Ils sont pris dans le même ensemble de concepts. Ainsi les unités d'information sont doublement étiquetées, par des descripteurs et par des référents[13]. Les liens entre les unités d'information sont induits par ces deux relations. Une unité d'information **u** en relation avec le concept référent **c** aura des liens sur toute unité d'information qui possède le concept **c** comme descripteur. La composition des deux relations donne naissance à un multigraphe orienté dont les sommets sont les unités d'information. Un graphe dual relie les concepts. En langage matriciel, si la première relation est donnée par D et la deuxième par R, la matrice du graphe (graphe pondéré et orienté) sous-jacent est $G_U = RD^T$. Ultérieurement, on comparera la structure de tels hypertextes à celle de ceux obtenus par « agrégation » selon un procédé décrit par Barabasi (2002). Il est également possible de considérer le graphe dual, celui des concepts dont la matrice est $G_C = R^TD$.

Un cas particulier d'hypertexte de type RD est celui où tous les concepts sont à la fois référents et descripteurs de chaque unité d'information à laquelle ils sont associés. Cela revient à poser R=D.

Des coefficients locaux (absolus ou relatifs) peuvent être définis par comptage des liens entrants et sortants. La plupart sont bien connus (valence, disponibilité, etc.) (Balpe, 1990 ; Balpe, Lelu, Papy & Saleh, 1996 ; Pochon, 1993).

Aspect algébrique

L'ensemble des hypertextes considérés est donné par les couples (R,D) avec R et D matrices de même dimension à coefficients dans {0, 1}[14]. Quatre opérations peuvent être définies :

1) L'addition « max 1 »[15], composante à composante, entre deux couples de même dimension. Cela revient à l'ajout de concepts dans les unités d'informations.

[12] http://www.irdp.ch/thema/htxt-col.htm

[13] Sans préciser la localisation des ancrages.

[14] Il serait possible, sans changer le formalisme, de donner d'autres pondérations à l'intersection d'un concept et d'une unité d'information.

[15] 0 + 1 = 1 + 0 = 1 + 1 = 1

$$(R_1, D_1) \oplus (R_2, D_2) = (R_1 +_1 R_2, D_1 +_1 D_2)$$

Les quatre matrices ont les mêmes dimensions.

2) La juxtaposition[16] qui représente l'ajout de nouvelles unités d'information.

$$(R_1, D_1) \oplus_J (R_2, D_2) = \left(\begin{pmatrix} R_1 \\ R_2 \end{pmatrix}, \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} \right)$$

Les quatre matrices ont le même nombre de colonnes (les mêmes concepts).

3) La superposition qui représente l'ajout de nouveaux concepts.

$$(R_1, D_1) \oplus_S (R_2, D_2) = \left((R_1 : R_2), (D_1 : D_2) \right)$$

Les quatre matrices ont le même nombre de lignes (les mêmes unités d'information).

4) La prise du quotient (« quotientage ») qui peut concerner les unités d'information ou les concepts que l'on regroupe en classes (sur-concepts). Du point de vue matriciel cela revient à regrouper (addition « max 1 ») des lignes ou des colonnes.

On peut également considérer des sous-hypertextes (en ne prenant en compte qu'une partie des concepts, que des sur-concepts ou encore qu'une partie des unités d'information) et reconstruire l'hypertexte original par superposition ou la juxtaposition de ces parties. L'hypertexte initial peut aussi être « approché » par des hypertextes construits sur la base de sur-concepts.

Quelques familles de documents structurés

Pour situer le graphe associé à ces hypertextes, il est possible de considérer la famille dans laquelle il s'insère :

DI, Documents indexés : un document indexé est un document auquel sont associés un certain nombre de mots-clés. A un document correspond un vecteur d'index qui est un élément de l'ensemble $\{0, 1\}^c$ (c est le nombre de mots-clés). Une famille de documents est représentée par une matrice « document × concept » (chaque ligne est le vecteur associé au document). Les unités d'information de notre modèle ont deux index qui peuvent être pris séparément ou conjointement.

GO, graphes orientés : on considère simplement l'existence d'un lien d'une unité d'information vers une autre sans s'intéresser à la façon dont le lien a été défini. Dans notre construction, cela revient à considérer G_U (cas pondéré) ou à dichotomiser G_U (les valeurs non nulles sont mises à 1).

GN, graphes non orientés : on s'intéresse aux unités liées sans considérer le sens du lien. Les hypertextes obtenus par agrégation entreront directement dans cette catégorie. Dans notre construction, cela revient à considérer $G_U + G_U^T$ (cas pondéré) ou à dichotomiser cette dernière matrice.

A noter que si les liens sont typés, des sous-graphes peuvent être définis (avec les mêmes nœuds) en se limitant aux liens d'un type donné. Dans le cas des graphes induits par la construction RD, chaque concept peut définir un type. D'autres types plus grossiers peuvent être également définis en regroupant les concepts en famille (typage de concepts).

MO, multigraphes orientés : la construction RD donne lieu à des graphes de ce type. Dans ce cas, le multigraphe est la superposition des graphes (de type GO) créés pour chacun des concepts (G_U est la somme des matrices de chacun de ces graphes). Chaque concept définit un sous-graphe (qui n'est en principe pas un multigraphe). Les concepts peuvent être regroupés en familles (typage en « sur-concepts »). Deux sortes de multigraphes avec mêmes sommets que le graphe initial peuvent être définis : ceux qui ne considèrent que les concepts d'un seul type (projection) et ceux qui sont construits à partir des sur-concepts (quotient).

On peut constater les inclusions suivantes :

RD $\subset$ **MO** : la construction RD génère un multigraphe orienté. De fait, on a l'égalité. Tout multigraphe orienté peut se mettre sous la forme d'un graphe RD engendré à partir de deux relations « document × concept ».

GN $\subset$ **GO** $\subset$ **MO** : un graphe non orienté est un cas particulier d'un graphe orienté qui lui-même est un cas particulier de multigraphe.

Relations d'équivalence entre sommets d'un multigraphe

Il est possible de définir une relation d'équivalence ou de similarité (**S**) entre les sommets d'un multigraphe. Selon **S**, sont en relation les sommets qui ont des liens de mêmes types (définis par les mêmes concepts ou les mêmes types de concepts). Pour les multigraphes orientés, cette relation se décline en trois variantes (liens en entrée, en sortie, tous sens confondus). Pour cette relation, les sommets du multigraphe sont vus comme éléments de DI.

Une relation d'équivalence de « même localité » ou « même voisinage » (**L**) peut également être définie entre les sommets d'un graphe (GN). Selon **L**, sont en relation les sommets ayant des liens sur les mêmes sommets. A nouveau, cette relation se décline en trois variantes pour les graphes

[16] Les deux termes « juxtaposition » et « superposition » ont été introduits sur la base d'une première représentation graphique (voir : http://www.irdp.ch/thema/htxt-alg.htm).

orientés (GO). Cette relation peut être appliquée aux divers stades de la complétion transitive du graphe original sur lequel elle induit des graduations dans la localité. On parlera de localité régionale[17] (**LR** en précisant si nécessaire le degré de régionalité) et de similarité globale (**LG** pour le complété transitif).

Finalement, on rappelle la relation d'équivalence classique de connexité (**C**) sur les sommets, donnée par l'existence d'un chemin (orienté dans le cas des graphes orientés) entre deux sommets.

Dans le cas des graphes non orientés, **L** implique **C** (deux sommets ayant même localité sont connectés). De façon générale, la relation **L** implique trivialement **LR** qui elle-même implique **LG**. Pour les graphes associés à des hypertextes de type RD, **S** implique **L** (deux nœuds similaires sont forcément liés aux mêmes noeuds) et même **C** sur le graphe non orienté correspondant.

Ces diverses relations, liées à certains coefficients (ordre des sommets), permettront de caractériser les hypertextes considérés. Il faut encore ajouter les caractéristiques globales classiques dans l'étude des graphes : le nombre de composantes connexes, la compacité (densité des liens), le diamètre du graphe, le nombre de circuits (recherche des sous-arbres maximum), etc.

Trois approches pour caractériser la structure d'un hypertexte

En définitive, nous utiliserons trois approches pour caractériser la structure d'un hypertexte. La première concerne la topologie globale des graphes associés (connexité, compacité, etc.).

La deuxième approche est locale. Elle consiste à considérer le degré de « même localité » (voisinage) ou de similarité (ou de quasi-similarité en affaiblissant la relation **S** définie ci-dessus qui, rappelons-le, induit une relation de même localité sur les graphes associés à des hypertextes de type RD) entre les unités d'information en fonction du nombre de descripteurs et de référents que deux unités d'information ont en commun (de façon duale, un coefficient de similarité peut aussi être défini pour les concepts[18]). Cette approche généralise l'analyse classique des documents largement explorée par Salton (voir notamment Salton & al, 1996). A partir de cette similarité, on peut considérer un hypertexte « quotient » et une structure d'ordre entre les unités d'informations.

Finalement, la troisième façon présente un cas intermédiaire entre le local et le global. Elle consiste à repérer le nombre de liens (éventuellement classés par type et ceci dans les trois perspectives locales, régionales ou globales) pointant ou issus de chaque unité d'information (recherche d'unités centrales, de « sources » ou de « puits », etc.).

[17] Terme inspiré du vocabulaire utilisé en analyse d'image.

[18] La classification simultanée des observations et des variables est un sujet délicat (voir http://www.irdp.ch/methodo/stat-dbl.htm)

Il reste le problème de la recherche des concepts à partir des documents. A ce niveau peuvent intervenir tout d'abord des analyses automatiques[19] des documents pour en extraire des concepts et un regroupement (typage) des concepts à partir d'une classification des documents[20]. D'autres organisations « plus sémantiques » peuvent être envisagées, des hiérarchies (avec la notion de subsumation) décrites par Woods (1997) à l'utilisation d'ontologies plus sophistiquées. Cet aspect ne sera pas abordé dans cet article. On se limitera à montrer (dans le cas d'une simple catégorisation de concepts) que cette organisation des concepts a une répercussion sur la description de la structure de l'hypertexte.

[19] Par exemple le Latent Semantic Indexing - LSI (Berry, Dumais & Shippy, 1995)

[20] Voir à ce propos la méthodologie appliquée pour l'analyse d'articles de presse par Berney & Pochon (2000). Des techniques sophistiquées existent pour réutiliser et faire évoluer des *thésaurus* pour des grandes quantités de documents (Roux, 2004).

3. Etude de cas : UTOPIA

Un hypertexte qui peut servir d'exemple (en partie créé pour cet usage) provient de l'ouvrage réalisé lors du départ à la retraite de Jacques-André Tschoumy, directeur de l'IRDP. Cet ouvrage, *Des utopies à construire* (Cornali & Weiss, 1996), est une mise en réseau de réflexions d'auteurs sollicités à s'exprimer, sur un mode personnel ou scientifique, sur une citation de Jacques-André Tschoumy. Il s'agissait encore pour les auteurs de mettre leur propos en rapport avec d'autres auteurs, d'autres textes, d'autres documents.

Les thèmes abordés étaient divers : la coordination scolaire, le droit à l'éducation, la citoyenneté européenne, la langue maternelle, l'éducation interculturelle et le plurilinguisme, la formation des enseignants, l'éducation aux médias et aux nouvelles technologies. Par ailleurs, un comité de rédaction a sélectionné un certain nombre de concepts, regroupés en champs conceptuels (par exemple le champ *c_altérité* contient les mots-clés: *autre*, *altérité*, *allophone*), présents dans les textes afin de créer un index au sens classique du terme. Ce premier travail fournit les éléments de la structure de présentation écrite des textes, et rend possible une « navigation » intertextuelle par des renvois et un index de mots-clés.

L'ouvrage a paru sous forme d'un livre et d'un CD-ROM. Une version HTML adaptée à l'Internet[21] complète le panel des principaux supports disponibles à l'époque.

Ces données ont été utilisées pour proposer des exemples de calcul des différents coefficients locaux[22] et pour illustrer diverses façons de « réduire » un hypertexte (quotient, décomposition, etc.). Le noyau central des données servira également à discuter l'approche *a posteriori*, c'est-à-dire la classification « automatique » des unités d'information (mais réalisée manuellement dans ce cas). On verra également sur cet exemple comment la structure d'un hypertexte se ramène de proche en proche à une structure plus simple.

La figure 1 présente la structure de l'hypertexte telle qu'il a été construit[23]. Les unités d'information sont les suivantes : liste des champs sémantiques, table des matières, un document par champ sémantique contenant les mots-clés qui font partie de ce champ, introduction générale, bibliographie, une introduction par partie, 27 textes d'auteurs (ou articles), de nombreuses citations et notes. Les concepts utilisés sont classés en 7 catégories (sur-concepts).

[21] http://www.irdp.ch/utopies/utopies.htm
[22] http://www.irdp.ch/thema/htxt-uto.htm
[23] Mais ce schéma a été établi après coup. Une partie de la structure a été induite par les outils de PAO utilisés.

- mot-clé : c'est la catégorie la plus spécifique de cet environnement. Elle est constituée de 80 mots-clés (appelés concepts dans le cas particulier!) ;
- champ : cette catégorie sert au regroupement des mots-clés en 11 champs sémantiques ;
- auteur : les noms des 27 auteurs constituent une catégorie de concepts ;
- citation : chacune des 151 citations donne lieu à un concept de la catégorie ;
- note : chacune des 13 notes donne également lieu à un concept de cette catégorie.
- doc: cette catégorie est constituée de 9 concepts permettant de caractériser des documents particuliers: liste des auteurs (auteurs), bibliographie (biblio), introduction (intro), introductions à chacune des 6 parties de l'ouvrage (intro1 à intro6).
- outil: cette catégorie comprend deux concepts, l'un permet de repérer la table des matières (tdm) et l'autre la liste des champs sémantiques (*champs*).

Cette classification des concepts est basée sur la fonction des liens créés par les concepts et non sur la sémantique de ces derniers. Ainsi chaque champ conceptuel ne constitue pas dans ce regroupement une catégorie de mots-clés, contrairement à ce qui pourrait se faire par subsumation (Woods, 1997). La structure de l'hypertexte qui sera dégagée à partir de ce typage des concepts devra donc correspondre à l'organisation *a priori*.

Les liens héritent de la classification des concepts. Dans la figure 1, les liens sont représentés par des flèches dont le graphisme varie en fonction du type, selon la convention suivante : mot-clé = flèche noire, mince ; champ = grise, large pointillée ; auteur = noire, large ; citation = grise, mince ; note = noire, large, pointillée ; doc = noire, mince, pointillée ; outil = grise, mince, pointillée.

On notera que les liens « aller » ne sont pas forcément du même type que les liens « retour » (voir Auteurs, Texte d'auteur). Certains liens sont omis (sur la table des matières, sur la liste des champs) pris en charge par le système de navigation. Les liens multiples (une ancre, plusieurs cibles) sont représentés par des flèches avec une pointe double.

La figure montre bien la correspondance entre les types de liens (concepts) et des types d'unités d'information.

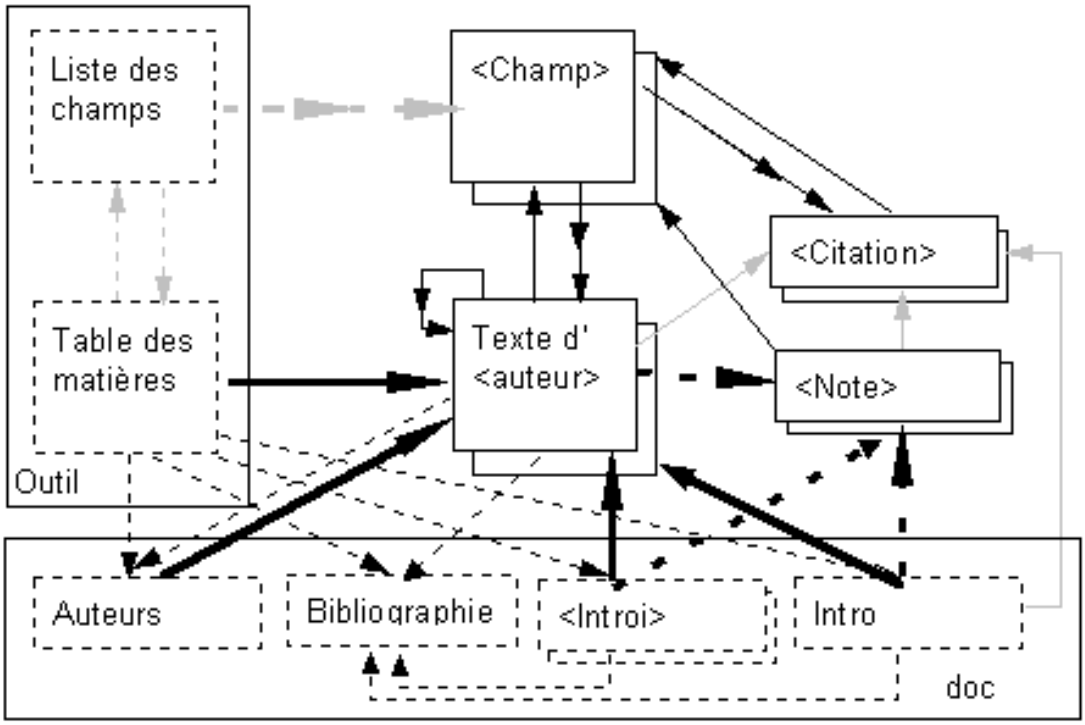


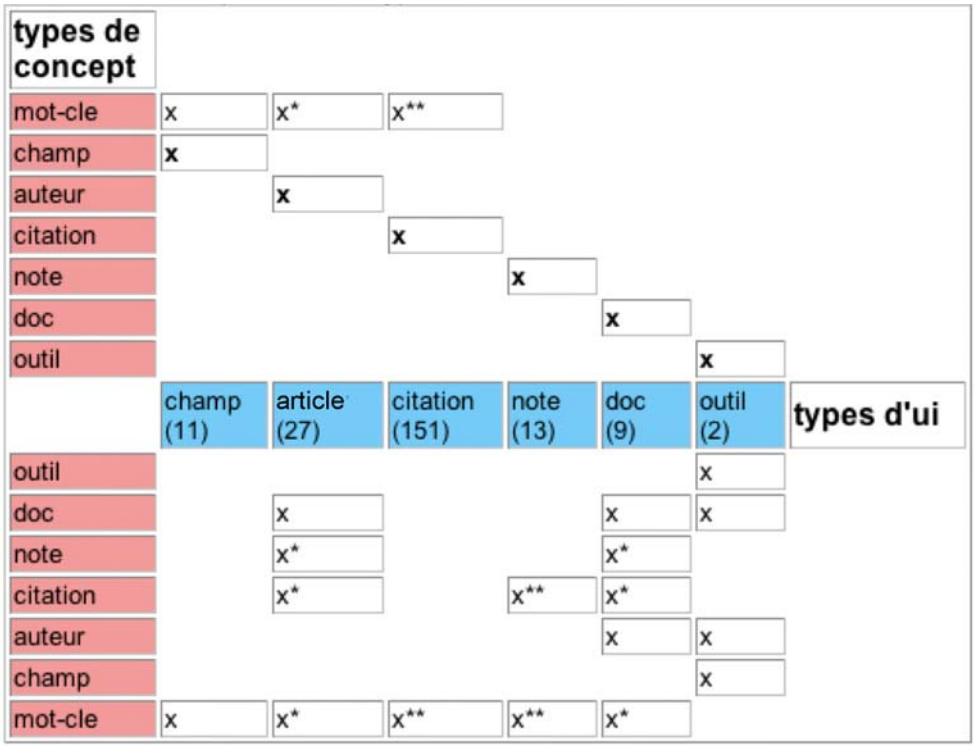
Figure 1. *La structure de l'hypertexte « UTOPIA »*

Cet hypertexte peut se décomposer de la façon suivante, par « juxtaposition »[24] (les articles sont les textes d'auteurs) :

$$H = H|_{champ} \oplus_J H|_{article} \oplus_J H|_{citation} \oplus_J H|_{note} \oplus_J H|_{doc} \oplus_J H|_{outil}$$

Le tableau de la figure 2 montre la correspondance entre la catégorisation des concepts et la catégorisation des unités d'information qui en dépend. Ce tableau doit être vu en deux parties séparées. Une ligne centrale (axe horizontal) énumère les types d'unités d'information d'UTOPIA. L'axe vertical du dessus énumère les concepts descripteurs, l'axe vertical du dessous énumère les concepts référents.

[24] La notation $H|_{article}$ représente l'hypertexte restreint aux unités d'informations constituées des textes d'auteurs (articles) en gardant l'ensemble des descripteurs et des référents. Du point de vue matriciel, il est constitué d'une partie des lignes de R et de D. Le terme juxtaposé provient d'une première représentation graphique (voir la page http://www.irdp.ch/thema/htxt-alg.htm).



types de concept							
mot-cle	x	x*	x**				
champ	x						
auteur		x					
citation			x				
note				x			
doc					x		
outil						x	
	champ (11)	article (27)	citation (151)	note (13)	doc (9)	outil (2)	types d'ui
outil						x	
doc		x			x	x	
note		x*			x*		
citation		x*		x**	x*		
auteur					x	x	
champ						x	
mot-cle	x	x*	x**	x**	x*		

Figure 2. *Les catégories de concepts et la catégorisation associée des unités d'information*

Ce tableau permet de reprendre le problème de la classification *a priori* ou *a posteriori*. L'étoile (*) permet de signaler qu'il peut y avoir des unités d'information du type *a priori* qui ne possèdent pas le concept associé. La double étoile (**) indique que la plupart des unités d'information de la catégorie ne possède pas le concept associé *a priori*. Une classification automatique, *a posteriori*, créerait donc des groupes supplémentaires à moins qu'une certaine « marge d'erreur » ne soit autorisée. Une classification *a posteriori* aurait pu, par exemple, faire figurer la préface dans la classe « doc » plutôt dans la classe « auteurs ». Le même problème se pose à propos des « articles » sans référence bibliographique.

Une décomposition plus grossière serait donnée en regroupant les trois premières parties. La première composante, qui concerne le contenu, pourrait faire l'objet d'une analyse plus détaillée, liée cette fois-ci à la sémantique des mots-clés.

$$H = H\big|_{texte-indexé} \oplus_J H\big|_{note} \oplus_J H\big|_{doc} \oplus_J H\big|_{outil}$$

Cette décomposition est à rapprocher d'une décomposition classique des liens (Lowe & Hall, 1999) qui distingue les liens de structure, les liens de référence et les liens d'association. Cette classification des liens peut se répercuter sur des parties de l'hypertexte

$$H = S + R + A$$

Dans le cas d'UTOPIA, la partie S est constituée des liens créés par les concepts de type « doc » et « outil », la partie R rassemble les liens de type « note » et la partie A résulte de la réunion des liens de type « champ », « auteur » et « citation ».

Cette décomposition paraît difficilement réalisable de façon automatique sans hypothèses supplémentaires. Ainsi, dans le cas des hypertextes de type RD, on peut trouver facilement, s'ils existent, les concepts qui sont à la fois référent et descripteur de documents avec lesquels ils sont en relation[25]. Les liens associés correspondraient à la partie A. De même, la partie R pourrait être constituée des liens conduisant à des unités d'information terminales.

[25] Dans le modèle RD, on peut aussi décider de distinguer les « vrais » concepts et traiter à part les concepts « utiles ». On pourra aussi décomposer les hypertextes de façon à distinguer différents types de liens.

4. Indicateur de structure globale

Définition de la matrice de circulation

Il s'agit de caractériser le graphe de l'hypertexte par des valeurs qui en « mesurent » la morphologie à partir des coefficients locaux définis précédemment.

Notre approche, qui ultérieurement sera comparée à d'autres approches, se base sur une idée probabiliste[26]. Le but est de représenter la « probabilité » d'atteindre un certain type d'information, c'est-à-dire un concept **c**, à partir d'une unité **u** donnée[27]. La procédure est la suivante : choisir au hasard un référent de **u**, à partir de ce référent choisir au hasard l'unité d'information reliée et procéder de même jusqu'à atteindre **c**.

Si l'unité **u** possède **c** comme référent on donnera comme probabilité d'atteindre **c** via ce référent la valeur 1/Di*(u). A cette valeur viendra s'ajouter la possibilité d'atteindre **c** via les unités d'informations liées. En formule :

$$p_u(c) = \frac{1}{Di*(c)} \left(\delta_u(c) + \frac{1}{2} \sum_{\substack{c' \in D*(u) \\ c' \neq c}} \sum_{u' \in U(c')} \frac{p_{u'}(c')}{\mathrm{Re}(c')} \right) \quad (4.I)$$

Dans cette formule, $\delta_u(c)$ vaut 1 si **c** est un "référent" de **u** et 0 sinon.

D*(u) : ensemble des concepts référents dans **u** ;

Di*(u) : ouverture de **u**, nombre de concepts référents dans u (#D*(u)) ;

U(c) : ensemble des unités d'information dont **c** est descripteur ;

Re(c) : rendement ou valence de **c**, nombre d'unités d'information décrites par **c** (#U(c))

Pour tenir compte du fait que l'information obtenue via d'autres unités d'information est moins accessible (et pour assurer la convergence de la suite en cas de boucle[28]), on multiplie la somme par un facteur ½. Sous forme matricielle : on pondère les coefficients de R par les totaux marginaux. On fait de même pour les colonnes de D. Ces deux matrices sont notées $\overline{R}$ et $\overline{D}$. Puis on forme :

$$\overline{G} = \overline{R}\overline{D}^T \quad (4.II)$$

[26] http://www.irdp.ch/thema/htxt-st1.htm

[27] Un autre scénario qui conduit au même résultat est de calculer la probabilité de rejoindre une unité v à partir de u selon le même procédé aléatoire.

[28] pour éviter une sorte d'effet « Larsen ».

On constitue ensuite la matrice du graphe du complété transitif en introduisant le facteur ½ qui tient compte de l'éloignement et assure la convergence de la série.

$$\overline{R} + \frac{1}{2}\overline{G}\overline{R} + \frac{1}{4}\overline{G}^2\overline{R} + \ldots = \sum \frac{1}{2^n}\overline{G}^n\overline{R} \quad \text{(4.III)}$$

Ce qui peut s'écrire : $S\overline{R}$ où S est la matrice d'adjacence du graphe complété avec la pondération de ½ introduite. Cette matrice sera prise comme « mesure » de la structure globale de l'hypertexte. Elle vaut :

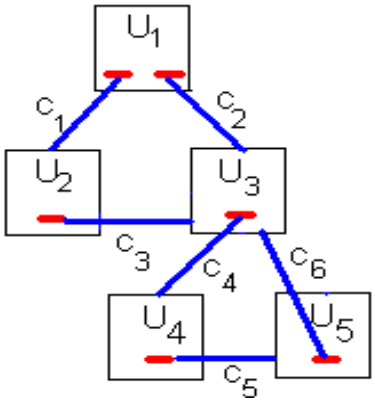
$$S = \sum \left(\frac{1}{2}\overline{G}\right)^n = I + \frac{1}{2}\overline{G} + \frac{1}{4}\overline{G}^2 + \ldots = \left(1 - \frac{1}{2}\overline{G}\right)^{-1} \quad \text{(4.IV)}$$

On désignera S par le terme de matrice de circulation[29].

Il faut préciser encore que si la somme infinie 4.IV est éventuellement calculable pour d'autres matrices d'adjacence, c'est l'utilisation du modèle RD (à partir de deux relations « document × concept ») avec les matrices R et D qui permet d'en donner une certaine interprétation (et assure la convergence). Il n'en reste pas moins que l'usage de la partie droite de la relation pourrait faire l'objet d'applications au-delà de notre modèle particulier. Partant d'un hypertexte quelconque de graphe G, on peut définir D et R tels que $G = G_U = RD^T$[30].

Etude d'un exemple

La matrice de circulation sera calculée pour l'hypertexte donné par la figure 3.



[29] A ne pas confondre avec les matrices de transition des chaînes de Markov. S est une matrice d'adjacence du complété transitif pondéré du graphe initial. Cette matrice représente un condensé de la connaissance contenue dans l'hypertexte et constitue un pas vers une application de la formule de Brookes. D'autres matrices pourraient être imaginées qui faciliteraient le calcul de l'information. Par exemple : S = – k log (1-G/2) possède ne nombreux avantages, notamment une valeur nulle sur la diagonale pour des unités d'information ne faisant pas partie d'un circuit (voir l'annexe 2). Mais on utilisera parfois plus simplement S – I.

[30] http://www.irdp.ch/thema/htxt-equ.htm

Figure 3. *Un hypertexte constitué de cinq unités d'information et de six concepts avec une « boucle ».*

Dans ce cas, on a les matrices associées suivantes :

$$R=\begin{pmatrix}1&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}\quad D=\begin{pmatrix}0&0&0&0&0&0\\1&0&0&0&0&0\\0&1&1&0&0&1\\0&0&0&1&0&0\\0&0&0&0&1&0\end{pmatrix} \quad \text{(4.V)}$$

La matrice de circulation donnant la structure est:

$$S=\begin{pmatrix}1&0.250&0.429&0.214&0.107\\0&1&0.571&0.286&0.143\\0&0&1.143&0.571&0.286\\0&0&0.286&1.143&0.571\\0&0&0.571&0.286&1.143\end{pmatrix} \quad \text{(4.VI)}$$

Le coefficient s_{ij} représente une certaine probabilité d'arriver sur l'unité j en partant de l'unité i. Cette matrice offre de nombreuses informations. Notamment les valeurs supérieures à 1 dans la diagonale indiquent des unités d'informations se trouvant dans des circuits.

La matrice qui nous a servi de base dans la définition est :

$$S\overline{R}=\begin{pmatrix}0.5&0.5&0.250&0.429&0.214&0.107\\0&0&1&0.571&0.286&0.143\\0&0&0&1.143&0.571&0.286\\0&0&0&0.286&1.143&0.571\\0&0&0&0.571&0.286&1.143\end{pmatrix} \quad \text{(4.VII)}$$

Le coefficient de la ligne i et de la colonne j est à considérer comme une certaine éventualité de rencontrer le concept j en partant de l'unité i, ceci dans un parcours infini, mais où la longueur du parcours diminue cette éventualité. La première ligne montre la grande accessibilité de c1 et c2 (chacun sur u1). Puis c3 est plus difficilement accessible (il faut passer par c1). Par contre c4 redevient plus accessible (2 chemins y conduisent), etc.

Définition du foncteur K

Chaque concept, famille de concepts, classe d'équivalence de concepts donne lieu à une matrice S. Cette association est de type fonctoriel. Ce foncteur **K** est à considérer comme une étape dans notre quête d'une quantification de la connaissance en introduisant ce foncteur dans la formule

fondamentale S = **K**[H][31]. Les opérations sur S, mesure de la connaissance, sont héritées, à travers ce foncteur, des opérations algébriques définies sur les hypertextes.

[31] Le symbole S utilisé par Brookes est ici remplacé par H, c'est la structure et nous avons utilisé S comme mesure S = **K** (H).

5. Entre localité et globalité : étude de la connectivité

Le but de ce paragraphe est de présenter les éléments de structure locale qui donnent des informations sur l'aspect global (la troisième approche citée page 12). Il s'agit principalement de visualiser les différents types d'unités d'information, les unités centrales, les sources et les puits. La matrice d'adjacence est utilisable pour donner le nombre « d'entrées » et de « sorties » d'une unité d'information. Elle permet de décider si une unité peut être atteinte directement à partir d'une autre. La matrice S (ou une matrice S « tronquée » comme nous le ferons ultérieurement) permet de compléter cette information en tenant compte de la totalité du graphe. L'utilisation d'un facteur amortissant ½ (et des tronquages) constitue une solution intermédiaire entre l'usage de G (matrice d'adjacence) qui ne permet que de comptabiliser les contacts directs, et le complété transitif qui rapproche artificiellement les unités d'information[32].

La procédure adoptée est la suivante. Sur la base de S, elle consiste à passer en revue les unités d'information[33] et, pour chaque unité **u**, à comptabiliser le nombre de valeurs nulles sur la ligne (nombre d'unités d'information non atteintes depuis u) et le nombre de valeurs nulles sur la colonne (nombre d'unités d'informations ne conduisant pas à u). Ces nombres permettent de définir des points (x_u, y_u) que l'on peut représenter sur un diagramme cartésien que nous nommerons diagramme de connectivité.

La figure 4 montre une telle représentation cartésienne des unités d'information d'un hypertexte avec M = #U – 1 (avec #U nombre d'unités d'information). Chaque unité d'information est placée en fonction de ses « coordonnées » (x_u, y_u) où : x_u représente le nombre d'unités d'information non dépendantes de **u** et y_u le nombre d'unités d'information ne conduisant pas à u. Les **puits** sont les unités d'information sans descendant. Les **sources** sont les unités d'information sans « ancêtre ». Les unités d'information situées sur l'axe des x ($y_u = 0$) sont des unités d'information qui proviennent de toutes les autres. Les unités d'information situées sur l'axe vertical ($x_u = 0$) sont celles qui conduisent à toutes les autres. Les unités centrales sont celles qui, de façon équilibrée, ont « beaucoup » d'ancêtres et de descendants (hubs).

[32] On rappelle que ce procédé ne peut s'appliquer qu'aux graphes définis à partir de la double relation « document-concept ». Toutefois, tout hypertexte peut se mettre sous cette forme. Les résultats obtenus par ce procédé seraient à comparés à ceux obtenus par d'autres techniques (voir aussi la partie 8).

[33] Ce qui revient à parcourir la diagonale de la matrice S ou d'une version « tronquée » de S.

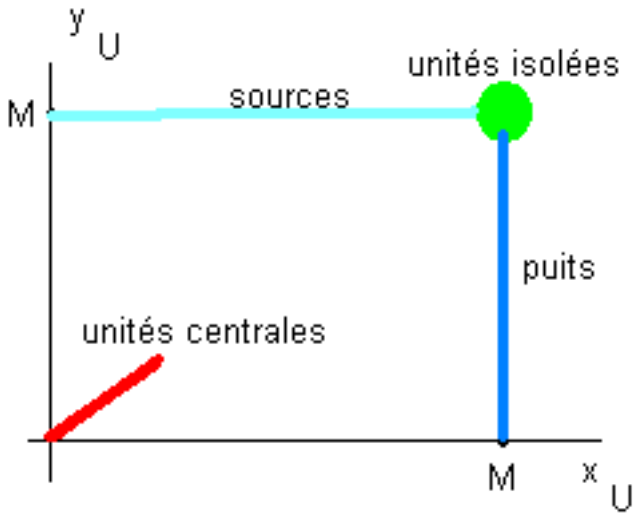


Figure 4. *Diagramme de connectivité*

Rappelons qu'il est toujours possible de considérer le cas dual et de définir le diagramme de connectivité des concepts. En particulier, on peut distinguer :

- les concepts « source » ou racine : ils marquent les thèmes généraux, ils ne servent jamais (rarement) de référent,
- les concepts « puit » ou terminaux : ce sont des concepts qui ne servent jamais de descripteur, ils n'ont « momentanément » aucune description ou définition à disposition,
- les concepts isolés : ce sont des termes marqués mais qui ne servent ni de descripteur, ni de référent,
- les concepts centraux : souvent référés, souvent descripteurs, ils constituent la « substance » de l'hypertexte.

Le cas d'UTOPIA

L'exemple d'UTOPIA permet d'illustrer ce qui précède tout en mettant en évidence, sans provoquer de grande surprise, que l'approche « exacte » ne peut convenir à des hypertextes complexes. Il s'agit de tolérer une certaine erreur. Plutôt que de viser à rechercher des composantes connexes, il s'agira de distinguer des zones à forte concentration de liens reliées de façon très faible au reste de l'hypertexte. L'étude de la connectivité est un pas allant dans ce sens. On notera $\tilde{S}_\alpha$ la matrice S tronquée à la valeur de coupure α[34] et $\hat{S}_\alpha$ (notée également ST par la suite) la matrice où les valeurs

[34] Correspond à une opération de régionalisation en analyse d'images.

non nulles sont mises à 1 (simplifie les calculs et accentue la structure). Ces calculs se réalisent facilement de façon automatique (programme pour MathLab disponible).

La figure 5 représente le diagramme de connectivité pour UTOPIA.

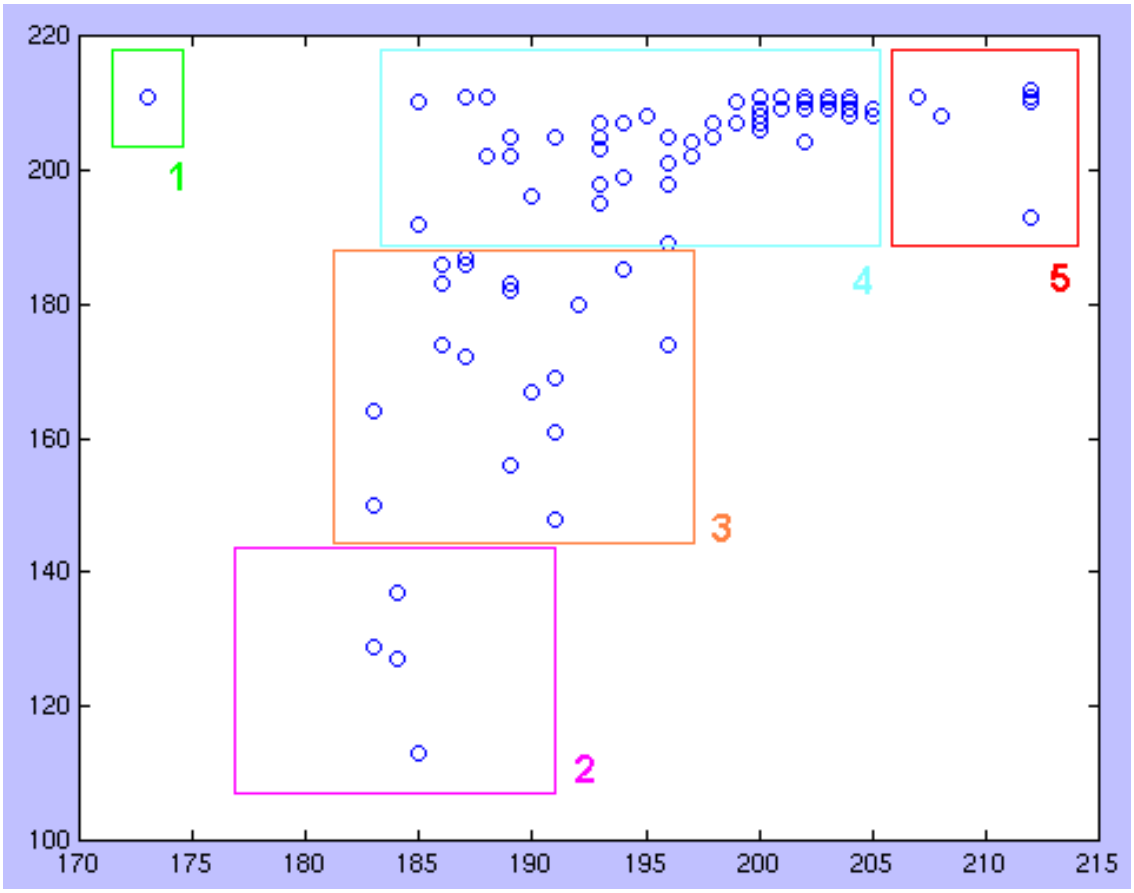


Figure 5. *Diagramme de connectivité de UTOPIA sur la base de* $\hat{S}_{0.01}$

La zone 1 contient deux « sources universelles, en particulier la table des matières *tdm* de coordonnées (173 ; 211). La zone 2 contient 4 unités centrales qui sont des textes qui évoquent plusieurs thèmes dont *cardinet* (185 ; 113) qui est une synthèse. La zone 3 est constituée de 18 unités assez centrées. Il s'agit de 14 textes d'auteurs et de 4 champs sémantiques, les 4 plus centraux qui sont ainsi mis en évidence. La zone 4 contient 35 unités qui sont caractéristiques de sources partielles (10 textes d'auteurs, 6 champs, les introductions, quelques notes et citations). La zone 5 contient les unités quasi isolées dont la majorité sont des notes ou des citations.

En principe, le diagramme ne tient compte que de la connectivité. Il peut toutefois donner des renseignements, comme c'est le cas pour UTOPIA, sur la structure des unités d'information et sur la structure du graphe sous-jacent.

6. Relation de même localité ou de voisinage

Le travail précédent classait les unités d'information en fonction du nombre de connexions, indépendamment des concepts et des relations entre les unités. Dans ce paragraphe, une première approche « naïve » est effectuée pour rendre compte de l'ensemble des unités d'information ayant même voisinage (relations **L**, **LR** et **LG**).

L'algorithme utilise la matrice ST (structure tronquée et dichotomisée) et la symétrise (entrées et sorties pourraient être utilisées séparément). Il se base sur le fait que plus la corrélation entre deux lignes est élevée et plus les unités sont analogues. Il procède de la façon suivante :

- les unités sont triées par nombre de connexions décroissantes,
- la première unité sert à fabriquer le premier groupe composante en ajoutant toutes les unités fortement « corrélées » (coefficient cos),
- la première unité faiblement corrélée sert de base à la deuxième composante,
- on teste les unités suivantes par rapport aux deux composantes et, le cas échéant, on fabrique de nouvelles composantes.

Cet algorithme est efficace, mais il n'est pas symétrique dans le sens où il privilégie les premières composantes formées. Il pourrait être amélioré en choisissant la composante la plus proche pour chaque unité d'information ou en réitérant le procédé. Cela revient à utiliser des algorithmes d'analyse hiérarchique classique.

Avec UTOPIA, cet algorithme (avec tout d'abord une coupure à 0.01 comme ci-dessus puis une valeur de corrélation limite à 0.3) fabrique 28 familles constituées de 2 unités d'information ou plus.

La première famille contient 34 unités. Ce sont principalement les textes d'auteurs, la liste des auteurs et les champs principaux. La deuxième famille contient 33 unités : quelques textes[35], quelques champs et de nombreuses citations. La troisième famille contient encore 29 unités dont l'introduction et de nombreuses citations. Les 14 familles suivantes sont constituées de 5 unités (citations et notes). Il y a encore 5 familles constituées de 4 unités (citations et notes), puis 5 familles constituées de 3 unités (citations) et finalement une famille composée de 2 unités: l'introduction générale (unité 194) et une note. Finalement, 10 unités sont considérées comme uniques.

[35] Le lien entre les relations **S**, **L** et **C** (page 11) et la constitution particulière de UTOPIA (tous les concepts donnés par les mots-clés des textes sont à la fois descripteurs et référents) expliquent le fait que les composantes sont constituées d'unités de même type.

7. Acquis intermédiaires issus de la théorie des graphes

En se contentant d'analyser le graphe, comme cela a été fait dans le paragraphe précédent, en le symétrisant parfois, on bénéficie de toute « l'artillerie » développée pour l'étude des graphes non orientés et orientés.

Les résultats classiques de la théorie des graphes utiles pour notre propos ont principalement trait à la recherche des valeurs propres du « laplacien combinatoire L »[36]. La multiplicité de la valeur propre 0 donne le nombre de composantes connexes. En cas de connexité, la deuxième valeur propre donne une information sur la connectivité globale du graphe.

Dans cette démarche, travailler avec la matrice d'adjacence ou une matrice complétée ne change pas les résultats en ce qui concerne la connexité. Cela peut par contre modifier d'autres caractéristiques. Un essai a été mené dans le cas de l'hypertexte UTOPIA[37]. Mais le travail n'a pas été poussé plus avant. Il aurait fallu comparer le travail effectué à partir de G à celui qui prend S (ou une version tronquée) comme élément de départ[38]. Un algorithme simple qui semble efficace existe pour trouver des composantes « presque » connexes à partir $\hat{S}_\alpha$.[39] Ce procédé serait à comparer à des résultats récents, notamment la proposition de Q Ding, He & Zha (2001) et Q Ding, He, Zha & Simon (1988).

Notre sentiment à propos de ces diverses approches est que les techniques issues des théories sur les graphes sont particulièrement calibrées pour des graphes « réguliers » et d'étendue raisonnable. Dans le cas des hypertextes, l'approche pragmatique, avec des algorithmes simples et robustes à partir de $\hat{S}_\alpha$ semble mieux adaptée aux questions posées sachant que cette technique impose toutefois des limites à la grandeur des hypertextes traités et demandera des traitements préalables pour définir les « morceaux » à traiter (voir § 12).

---

[36] $L = K_A - A$ avec A matrice d'adjacence (symétrique ou rendue symétrique par addition de la transposée) et K_A matrice diagonale donc chaque élément vaut la somme des lignes de A.

[37] on trouve 213 valeurs propres de L "tronqué" à 0.1 et dichotomisé. Calcul de G barre puis S, ST 0.1, S+S', L. Ces valeurs propres (entre parenthèse leur multiplicité) sont tout d'abord : 0 (127) , 1 (3), 3 (3), 5 (2), 6, 7 (2), 8 (2) ; puis 73 valeurs sont non entières variant de 1.155 à 20.217. Cet hypertexte contient donc beaucoup d'unités d'information assez isolées (les notes et références).

[38] Dans le cas de S le procédé ne semble toutefois pas réaliste puisqu'il demande deux opérations coûteuses : inversion et recherche de valeurs propres.

[39] On considère l'unité la plus connectée (somme de la ligne et de la colonne correspondantes de $\hat{S}_\alpha$ maximum) puis de proche en proche les unités connectées avec un nombre de connexions supérieur à une valeur de seuil sont coagulées à l'ensemble. Le processus se répète à partir de l'unité ne faisant pas partie de la composante précédente et présentant un maximum de connexion.

8. Les indicateurs topologiques « classiques »

Les coefficients de structure classiques sont calculés à partir de la matrice des distances converties D = [d_{ij}] où d_{ij} est le plus court chemin entre les sommets i et j et un « grand » nombre, K (constante de conversion, souvent K vaut le nombre de noeuds), s'il n'y a pas de chemin du sommet i au sommet j (Botafogo, Revlin & Schneiderman, 1992)[40]. C'est une manière de réaliser une matrice du graphe complété transitif. Par rapport à S = [s_{ij}], on a $d_{ij} \approx 1/s_{ij}$. S est en quelque sorte la matrice des anti-distances.

Le premier coefficient classique est celui de **centricité**. Les informations que l'on peut obtenir à partir de S, $\tilde{S}_\alpha$ ou $\hat{S}_\alpha$ rendent la même information (voir les figures 4 et 5)[41].

Pour l'hypertexte défini par la figure 3, les valeurs obtenues sont les suivantes :

Valences de sortie

Coefficients classiques (relative out centrality)	13.14 ; 5.75 ; 4 ; 4 ; 4
Coefficients basés sur S	1 ; 1 ; 0.857 ; 0.857 ; 0.857
Coefficients basés sur $\hat{S}$	4 ; 3 ; 2 ; 2 ; 2

Valences d'entrée

Coefficients classiques (relative in centrality)	2.3 ; 2.97 ; 18.4 ; 13.14 ; 10.22
Coefficients basés sur S	0 ; 0.25 ; 1.857 ; 1.357 ; 1.107
Coefficients basés sur $\hat{S}$	0 ; 1 ; 4 ; 4 ; 4

Le deuxième coefficient est celui de **compacité**. La compacité est donnée de façon standard par (Max - sd) / (Max - Min) avec :

* Max = $K(N^2 - N)$ (graphe à N noeuds sans liens)
* Min = $N^2 - N$ (graphe complet à N noeuds)
* sd = somme des coefficients de la matrice des distances converties.

[40] Voir http://www.irdp.ch/thema/htxt-std.htm

[41] Les coefficients classiques sont du type 1/(somme de coef), ceux que nous définissons du type somme (1/coef)

Il est possible de créer les deux hypertextes extrêmes dans le modèle RD. Le cas Max est donné par l'absence de concept. La matrice S associée est l'identité et la somme correspondante est nulle (c'est donc un minimum noté Smin). Le cas Min est donné par la présence de tous les concepts (en nombre k = nombre de noeuds) comme descripteurs et référents de chaque noeud (unité d'information). Les matrices D (descripteurs) et R (référents) valent U, matrice dont tous les coefficients valent 1. La matrice associée du graphe est donc kU. Les matrices pondérées des descripteurs et des référents valent U/k[42].

Cela conduit à définir un coefficient de compacité : sc/k où sc est la somme des coefficients de la matrice de circulation sans la diagonale.

Avec l'hypertexte de la figure 2 on trouve :

Cas classique : K = N = 5 ; Min = 20 ; Max = 100 ; sd = 57 ; compacité = 43/80 = 0.54

Avec S : k = 5 ; Smax = 5 ; sc = 4.57 ; compacité = 0.91

Les informations sont comparables. A noter que dans le cas de l'hypertexte donné par les référents et descripteurs, plusieurs hypertextes différents peuvent conduire à un graphe réduit identique. Il y a aussi plus d'une façon de réaliser un hypertexte dont le graphe associé est un graphe complet.

Coefficient de stratification : ce coefficient capture le degré de linéarité de l'hypertexte. De façon classique, pour un nœud, on définit : le **statut** (nombre de liens en sortie), le **contre-statut** (nombre de liens en entrée). Le **prestige** est alors la différence (statut - contrestatut).

On définit pour un hypertexte le **prestige absolu** (somme des valeurs absolues du prestige de chaque noeud). On définit également le « *linear absolute prestige* – LAP » qui est le prestige absolu d'un hypertexte linéaire avec N nœuds.

La **stratification** (stratum) est le rapport du prestige absolu au LAP.

On peut suivre le même algorithme pour la matrice de circulation, en sachant également que plusieurs hypertextes peuvent conduire au même graphe associé. De même, il n'y a pas qu'un seul hypertexte « linéaire » de référence.

On prend dans ce cas pour l'hypertexte linéaire la matrice R valant 1 dans la « diagonale ». Quant à D, il s'agit de la matrice avec des 1 dans la sous-diagonale.

Dans le cas standard, un hypertexte linéaire possède un « stratum » de 1. Cette valeur est maximale. S'il est possible d'atteindre n'importe quel noeud à partir de n'importe quel noeud, le stratum est nulle.

[42] http://www.irdp.ch/thema/htxt-std.htm

Dans notre cas, la valeur du LAP est minimum. Ce qui conduit à définir le stratum comme LAP / prestige absolu. Les valeurs obtenues sont comparables.

Avec l'hypertexte de la figure 2 on a :

Cas classique : N = 5 ; LAP = 30 ; prestige absolu = 24 ; stratum = 0.8

Avec S : N = 5 ; LAP = 21/8 ; prestige absolu = 3.5 ; stratum = 0.75

En définitive, les coefficients obtenus avec la matrice S de circulation fournissent des informations comparables aux coefficients classiques. Il resterait à comparer les deux méthodes pour un hypertexte quelconque que l'on mettrait sous forme RD.

9. Topologie d'un hypertexte

Cette topologie est celle des graphes orientés dont on rappelle ici deux éléments : la structure papillon et les notions, héritées des travaux concernant le Web, d'autorité et de hub.

La structure papillon

L'ensemble des sommets (unités d'informations) d'une composante connexe peut se décomposer en 6 classes (Barabasi, 2002) qui sont représentées sur la figure 6.

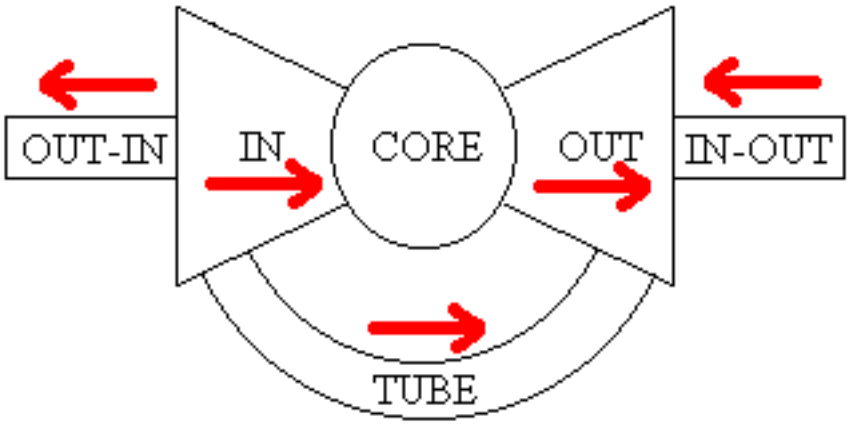


Figure 6. *Les 6 classes de sommets d'un graphe orienté*

CORE : Pour tout couple de sommets x, y de cet ensemble, il existe un chemin de x à y et un chemin de y à x. Cet ensemble CORE n'est pas déterminé de façon unique. Un CORE peut être construit à partir de chaque sommet. Appartenir à un même CORE constitue une relation d'équivalence. La décomposition d'un hypertexte en CORE peut dans certains cas constituer une analyse de structure intéressante.

IN : Tous les sommets x tels qu'il existe un élément y de CORE avec un chemin de x à y.

OUT : Tous les sommets x tels qu'il existe un élément y de CORE avec un chemin de y à x.

TUBE : Les sommets accessibles depuis IN qui rejoignent OUT

OUT-IN : Les sommets accessibles depuis IN qui ne rejoignent pas OUT

IN-OUT : Les sommets non atteignables depuis IN qui rejoignent OUT

Stratification du CORE

Il est possible de définir: $K^+_n(u)$, ensemble des unités d'information que l'on peut atteindre depuis u en parcourant au plus n arêtes.

De même: $K^-_n(u)$ est l'ensemble des unités d'information à partir desquelles il est possible d'atteindre u en parcourant n arêtes au plus.

$$K_n(u) = K^+_n(u) \cap K^-_n(u)$$

Les ensembles sont imbriqués : $K_n(u) \subset K_{n+1}(u)$. $K(u) = \cup K_n(u)$ est le CORE généré par u.

Pour rechercher le CORE d'un graphe orienté (de l'hypertexte) lorsque le nombre d'unités est important, la procédure suivante (en notation Matlab) peut être proposée :

1. recherche de la fermeture transitive GB = G * (I-G/n)^(-1); [43]
2. univaluation : GB = GB & GB;
3. GB(i,:)&GB(:,i) = v_i représente le vecteur « caractéristique » du CORE engendré par l'unité d'information i (v_i (j) = 1 signifie l'unité j est dans CORE)
4. pour chaque Ui, recherche du nombre d'unités (t(i)) qui peuvent être atteintes et qui peuvent l'atteindre : t(i) = sum(GB(i,:)&GB(:,i))

Ce processus donne donc la taille du CORE engendré par chacun des sommets du graphe (rappel si Cx est le CORE engendré par x et si y est dans Cx alors Cx = Cy).

Un même processus peut être utilisé pour déterminer $K_n(u)$ à partir de l'uniévaluation de GB + GB^2 + ... + GB^n. L'annexe 3 propose un exemple de calcul.

Autorités et hubs

Une « autorité » est une page (ou un site) référencée par de nombreuses autres pages. Un « hub » est une page (ou un site) référençant de nombreuses « autorités ». Les tandems « autorité-hub » constituent l'ossature du Web[44].

Dans cette définition, la notion de « hub » est subordonnée à la notion d' « autorité ». Une autre acception serait de prendre « autorités » et « hubs » comme des notions duales. Cette distinction

[43] En prenant n > || G ||, cela permet d'avoir une série (G/n)i convergente qui représente la matrice de la fermeture transitive pondérée d'une façon *ad hoc* (sans importance vu l'étape ultérieure). Sans pondération, est-il possible de travailler avec des séries formelles (les composantes négatives de 1/(I - G) correspondant aux séries non convergentes) ?

serait à préciser dans des cas pratiques pour juger de son utilité. La définition ne semble pas symétrique. Par contre le procédé de calcul issu de la théorie permet de définir « autorités » et « hubs » de façon indépendante. Kleinberg (1997) indique que les « hubs » correspondent aux unités d'information liées aux composantes maximales du vecteur propre associé à la valeur propre maximale de la matrice d'adjacence multipliée par sa transposée. Pour les « autorités », il faut prendre le produit de la transposée par la matrice d'adjacence (voir annexe 1).

[44] On pourrait aussi distinguer les « autorités », « les portails » (appelés hub précédemment) et les « hubs » qui sont à la fois portails et autorités.

10. Hypertextes construits par « agrégation »

Pour avoir un élément de comparaison, un autre type d'hypertexte, construit de façon locale à la manière du Web, a été considéré. Un simulateur a été développé dans ce but. Il permet de créer des unités d'information, chacune caractérisée par :

- un numéro d'ordre ;
- son degré de « fitness » qui représente une certaine qualité intrinsèque de l'unité. Le degré de fitness est un nombre aléatoire compris entre 0 et 1;
- le nombre de liens émis depuis cette unité d'information ;
- son type qui donne la nature de l'unité d'information (domaine traité).

Le nombre de liens et le type sont attribués au hasard selon des distributions décrites dans les deux tables :

- dis_lnk donne la distribution du nombre de liens émis par une nouvelle unité d'information (distribution normale dans l'exemple) ;
- dis_*typ* donne la distribution des types (distribution croissante dans l'exemple).

Les liens sont fabriqués selon un coefficient d'attirance qui dépend des types respectifs de la source et de la cible, du degré de « fitness » de la cible et du nombre de liens de la cible. Ceci à partir d'un germe, de deux germes ou plus.

L'annexe 1 traite de la création et d'une analyse complète d'un exemple.

11. Simulations

Quelques hypertextes ont été créés par agrégation ou selon le modèle RD. L'annexe 3 résume les essais effectués[45].

Le premier hypertexte[46] est obtenu par agrégation (avec les paramètres ci-dessus) et comporte 1000 unités d'information. La distribution des liens « entrants » (fréquence du nombre d'unités d'information présentant un nombre de liens rentrants donné) suit une distribution en puissance (prévu par la théorie) (voir annexe 3). Le nombre de liens sortants est plus faible et la distribution semble suivre la loi donnée par *dis_lnk*. Le graphe comprend près de 5000 arêtes. Parmi ses caractéristiques, on relève une grande composante connexe de 949 unités (et toutes les autres sont des unités isolées), un CORE maximum de 617 unités d'information. Il y a quelques autorités mais pas de hubs.

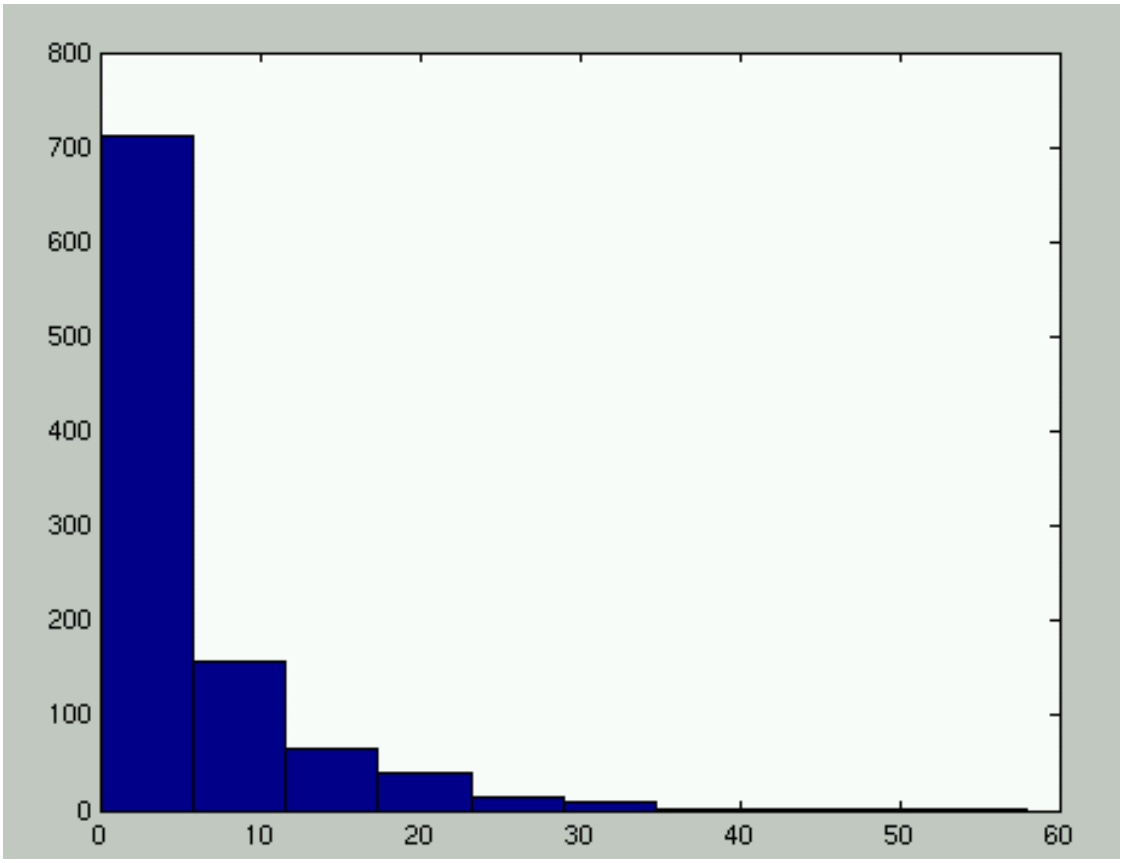


Figure 7. *Distribution en loi de puissance*

Le deuxième hypertexte[47] est obtenu par agrégation en utilisant deux germes (4 unités formant deux composantes). Il est également constitué de 1000 unités. Ces caractéristiques sont : une composante connexe de 943 unités ; une deuxième composante connexe non triviale (15 unités) ; un CORE de 608 unités.

Un troisième hypertexte[48] est obtenu par agrégation en utilisant deux germes (de 4 unités d'information formant deux composantes). Constitué de 500 unités, sa structure a été étudiée après qu'il eût été reconstruit sous forme RD. Son diagramme de connectivité est donné par la figure 8. Il a été obtenu en prenant une valeur de coupure de 0.001. On rappelle que les unités de droite sont des puits, celles du haut des sources. L'unité d'information 76 de coordonnées (398, 41) est un hub.

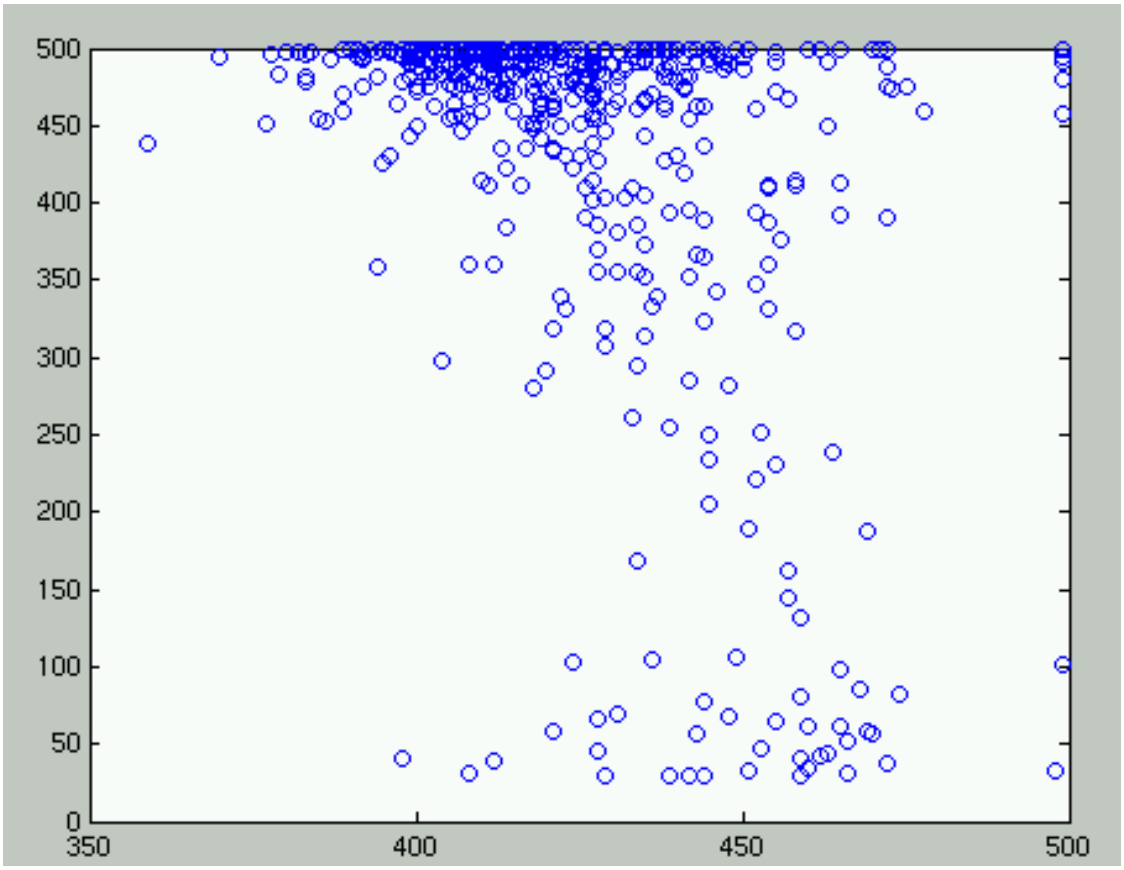


Figure 8. *Diagramme de connectivité d'un hypertexte obtenu par agrégation et reconstruit sous la forme RD*

[45] Les programmes sont écrits pour Matlab. Ils sont à disposition sur le site : http://www.irdp.ch/thema/htxt-info.htm. Ils sont en cours de réécriture pour R.
[46] http://www.irdp.ch/thema/htxt-s31.htm
[47] http://www.irdp.ch/thema/htxt-s32.htm
[48] http://www.irdp.ch/thema/htxt-s32.htm

Les trois hypertextes qui suivent[49] sont de type RD. Ils sont construits de la façon suivante. Les paramètres sont : n (nombre total d'unités d'information) ; c (nombre total de concepts) ; m (nombre maximum de référents par unité d'information) ; k (nombre maximum de descripteurs par unité d'information). Pour chaque unité on choisit au hasard le nombre de concepts référents (de 0 à m selon une distribution uniforme), puis on attribue ce nombre de concepts référents pris au hard parmi les c possibles à cette unité. On fait de même pour les descripteurs.

Les graphes obtenus sont par essence des multigraphes. Ce sont, vu le leur construction, des graphes aléatoires « isotropes ». C'est-à-dire que la probabilité qu'un lien existe entre deux sommets x et y est constante. Cette propriété d'isotropie entraîne que la distribution des unités d'information en fonction du nombre de liens « rentrants » ou « sortants » suit une loi binomiale dont il est possible de calculer le paramètre[50]. Les graphes associés présentent un haut degré de connectivité (voir annexe 3). Le nombre d'arêtes (en tant que multigraphe) est très important (plusieurs dizaines de milliers).

Les derniers hypertextes simulés[51] sont également de type RD. Mais la répartition des liens n'est plus homogène. L'ensemble des concepts, de même que celui des unités sont subdivisés et des densités différentes sont attribuées à chaque bloc. La distribution des unités d'information en fonction du nombre de liens « rentrants » ou « sortants » semble également suivre une loi (proche) binomiale. Contrairement à l'hypothèse faite, la connectivité reste élevée. Le nombre d'arêtes (en tant que multigraphe) reste très important.

Cela a conduit à construire un deuxième exemple, avec une attribution plus marquée de certains concepts à certaines unités d'information mais avec un résultat comparable. Il apparaît donc, sauf dans des cas très artificiels, que les hypertextes obtenus par ce procédé sont relativement « connectés ». Il serait intéressant de continuer dans cette voie en liaison avec les résultats des travaux de Erdös et Renyi concernant les graphes isotropes aléatoires (cités par Kauffman, 1993). Il est aussi intéressant de noter l'aspect très régulier du diagramme de connectivité (voir annexe 3).

[49] http://www.irdp.ch/thema/htxt-si2.htm
[50] http://www.irdp.ch/thema/htxt-si2.htm
[51] http://www.irdp.ch/thema/htxt-s21.htm

12. Le Web

Le Web a engendré un nombre de travaux importants. On peut en distinguer trois types. Tout d'abord la recherche des communautés. Le deuxième type concerne les services de liens qui s'apparentent de plus en plus aux moteurs de recherche. Finalement les théories des réseaux sociaux (*Social Network Theory* – SNT), appliquées aux communautés « virtuelles », utilisent également des outils qui sont issus de la théorie des graphes.

Les communautés

La caractéristique d'une communauté du Web est d'être un « sous-web » fortement interconnecté et faiblement relié au reste du Web. Plusieurs techniques permettent, en principe, de déterminer ces communautés. La première fait appel à l'algorithme « *maxflow-mincut* » de la théorie des graphes (Flake, Lawrence & Giles, 2000 ; 2002). Une autre technique procède par une exploration systématique (largeur d'abord ou profondeur d'abord) des liens (Gibson & Kleinberg, 1998) enfin une technique dite « méthode spectrale » se fonde sur l'analyse de la matrice d'adjacence et de ses vecteurs propres (Q Ding, He & Zha, 2001).

Des « *link services* » aux moteurs de recherche

Un domaine classique traité par la communauté hypertexte concerne les services de liens. Dans plusieurs implémentations d'hypertextes, les liens constituent une base de donnée « indépendante » des documents (Carr, Hall & De Roure, 1999 ; Carr, Bechhofer, Goble & Hall, 2001 ; Vaughan-Nichols, 2003). S'agissant du Web, ce domaine s'est décliné sous la forme de « services » Web. Ces recherches semblent avoir été supplantées par celles liées aux moteurs de recherche. Le site http://del.icio.us maintient toutefois l'idée d'une base de liens gérée et entretenue manuellement.

Social network theory et hypertexte

Issue de l'étude des liens sociétaux cette théorie est appliquée à l'étude des « communautés virtuelles ». Son intérêt est de mettre à disposition des outils informatiques, basés également sur la théorie des graphes (avec quelques résultats originaux), qui permettent d'étudier les hypertextes « virtuels » que constituent les parcours des utilisateurs. Il existe des systèmes qui permettent d'explorer et de représenter, avec différents niveaux de finesse, la structure de graphes, dont les versions actuelles sont de plus en plus performantes (Dousset & Karouach, 2005). Sortant du cadre de cette étude, rappelons ici les recherches liées aux « grammaires de circulation » qui constituent des outils permettant d'évaluer les services éducationnels (Pahl, 2000).

13. Résumé des techniques parcourues

Un bref résumé permet de présenter les différentes techniques évoquées ou utilisées dans l'analyse de la structure de l'hypertexte.

Hypertexte de type RD comme ensemble de documents indexés : dans ce cas les techniques classiques de classification (analyse hiérarchique) sont à disposition. On peut traiter descripteurs et référents conjointement ou séparément. Cela permet de dégager les unités d'informations similaires et, dans le cas des hypertextes de type RD, de même localité. L'extraction des concepts à partir des contenus n'est pas abordée ici. La technique évoquée (LSI), ne permet pas de distinguer ce qui est référent et descripteur. Son champ d'application se limiterait à la partie A des liens (voir p. 18).

Hypertexte comme graphe non orienté (GN) : en perdant l'orientation et la multiplicité des liens, considérer un hypertexte comme graphe non orienté permet d'approcher la structure selon divers points de vue :

- *Etude de la localité (relation **L**) à partir de la matrice d'adjacence* : La régionalité et la globalité (***LR*** et ***LG***) peuvent également être abordées à partir des complétés successifs du graphe. Lorsque le graphe est déduit d'un hypertexte de type RD, ces études peuvent être menées à partir de la matrice S de circulation symétrisée (tronquée ou non).
- *Etude de la connectivité :* cet indice est donné par l'ordre des sommets (nombre de liens de chaque nœud). Il est également possible de différencier une connectivité locale, régionale et globale.
- *Structure globale :* à partir de la matrice d'adjacence (et du laplacien combinatoire), il est possible d'étudier le degré de connexité. Il est également possible de chercher le sous-graphe en arbre maximum (ce qui revient à chercher le nombre de circuit) et de calculer la compacité du graphe. L'exhibition des sommets centraux (issus de l'étude de la connectivité) est aussi une indication de la structure globale.

Si les liens sont typés, la structure du graphe peut résulter de la réunion d'étude sur des sous-graphes.

Hypertexte comme graphe orienté (GO) : La prise en considération de l'orientation ajoute un degré supplémentaire à la connaissance de la structure de l'hypertexte. En reprenant chacun des points évoqués précédemment, nous avons :

- *Etude de la localité (**L**) :* elle peut être effectuée (de même que l'étude de la régionalité ou de la globalité) selon les liens entrants, sortants ou la conjonction des deux. Lorsque le graphe est déduit d'un hypertexte de type RD, ces études peuvent être menées à partir de la matrice S de circulation (tronquée ou non).
- *Etude de la connectivité :* la connectivité peut être envisagée en « entrée », en « sortie » ou comme conjonction des deux.
- *Structure globale :* à partir de la matrice d'adjacence, il est possible de dégager la structure papillon. Il est également possible de déterminer le nombre de circuits orientés et de calculer les coefficients de compacité et de linéarité. L'exhibition des sommets centraux (issus de l'étude de la connectivité), des sources et des puits est aussi une indication de la structure globale. Dans le cas des graphes issus des hypertextes de type RD, la matrice de circulation S permet de réaliser ces opérations de façon aisée. Notamment, les unités d'information participant à des circuits sont repérées par des valeurs diagonales supérieures à 1 (voir page 21).

Si les liens sont typés, la structure du graphe peut résulter de la réunion d'étude sur des sous-graphes.

Hypertexte comme multi-graphe orienté (MO) : Dans les cas des hypertextes de type RD, l'étude de la similarité d'unités d'information revient à l'étude de la localité (**L**).

Dans le cas des graphes de type RD, il y a deux façons de réduire l'étude de l'hypertexte à des composantes :

- en restreignant les liens à ceux engendrés par un concept ou une famille de concepts (à un type de concept). Au niveau matriciel, cela revient à ne considérer qu'une colonne ou une partie des colonnes des matrices R et D.
- en utilisant comme concept des familles des concepts primitifs (structure quotient). Au niveau matriciel cela revient à grouper (addition « max 1 ») l'ensemble des colonnes représentant des concepts de même type en une seule colonne.

14. Les théories utilisables

Cette partie, en élargissant le propos de la précédente, passe en revue tous les domaines qui paraissent utiles à notre quête. Certains ont été évoqués précédemment, d'autres restent à explorer.

Théorie des graphes : le premier domaine, relativement basique, concerne la théorie classique des graphes. Elle propose des outils puissants (méthode spectrale) qui permettent de décrire la topologie du graphe, notamment de détecter des composantes connexes ou presque connexes. Ces techniques ont été étendues par des outils permettant d'aborder des classifications dans le cas de grands graphes, le Web, par exemple. Nous les avons amplement explorées.

Méthodes de classification : les algorithmes d'analyse de données peuvent concerner des méthodes de « *clustering* » « aveugles » ou des techniques liées davantage à la structure du graphe (Stein, Meyer zu Eissen & Wissbrock, 2003) et à la sémantique des contenus. Plusieurs travaux ont été développés dans le cadre de recherches concernant la vision. Une image numérique est un vaste tableau de points de différentes couleurs. Une opération élémentaire pour la compréhension d'une image est de définir des régions significatives de l'image (*region growing*). Les techniques utilisées sont liées aux techniques de groupement.

Analyse reconstructive : il s'agit de méthodes générales pour la détection de variables « latentes » (Zwick, 2002). Deux approches sont proposées : l'approche statistique et l'approche informationnelle. Ces techniques permettent de découvrir dans un ensemble de variables en relation quels sous-ensembles sont indépendants (ou presque) les uns des autres. Ces techniques peuvent servir d'alternative aux analyses hiérarchiques traditionnelles. Dans le cadre des structures « document × concept », elles pourraient permettre de regrouper des concepts en concepts plus généraux et de séparer des groupes de concepts[52] (voir annexe 3).

Homologie ensembliste : il est vraisemblable que cette théorie pourrait être utilisée pour caractériser des hypertextes en prolongeant les travaux sur la théorie des graphes. Toutefois, à l'image de cette dernière, elle semble mieux adaptée à des travaux théoriques sur des ensembles avec certaines régularités. Les temps de calcul semblent également être prohibitifs dans le cas de grands ensembles (Babson, Barcelo, De Longueville & Laubenbacher, s.d. ; Barcelo, Kramer,

[52] Un ensemble de documents indexés, vu comme un ensemble de vecteurs à composantes dans {0,1}, définit une relation que l'on peut décomposer.

Laubenbacher, & Weaver, s.d.). A noter que le formalisme CAT[53] (et la Q-analyse de Atkin) (Valencia, 1998) font le lien entre ce domaine et le suivant.

Représentation des connaissances : ce point regroupe différentes techniques introduites en intelligence artificielle qui ont largement diffusé dans d'autres domaines : la programmation déclarative, la programmation par objet. On retiendra tout particulièrement la notion de réseau sémantique, relativement centrale, à laquelle on peut également relier les travaux actuels sur les ontologies (pour un résumé voir Pochon, 2006). Ces travaux peuvent être utiles dans la classification des concepts, classification qui permet de « filtrer » la structure d'un hypertexte.

Travaux généraux concernant les hypertextes : ces travaux sont nombreux. Ils peuvent concerner des aspects quantitatifs (statistiques, degré de connexion, etc.) et qualitatifs (notamment des aspects de navigabilité). On retiendra également les travaux s'attachant à des définitions formelles de l'hypertexte et ceux concernant les systèmes évolutifs et « adaptatifs » (Ohene-Djan, 2000). Un travail systématique pourrait avoir lieu qui passerait en revue les modèles[54], bien que ceux-ci concernent souvent davantage l'infrastructure des systèmes hypertexte (système de navigation, historique, etc.) que la structure de l'information.

Travaux liés au Web : toutes les techniques de classification se sont vues appliquées à la recherche dans le Web. En particulier, la théorie des graphes a été appliquée à la définition des « communautés Web » en utilisant des algorithmes classiques (*min cut-max flow*, par exemple). Ce type d'application présente deux aspects nouveaux. D'une part la quantité de données est importante. D'autre part, il s'agit d'un système dynamique. Ces deux traits font que la métaphore hypertexte pour une approche de la connaissance gagne à être étudiée en liaison avec ces retombées particulières. La revue « *IEEE computer* » (novembre 2002) consacre un dossier à la « *Web Intelligence* » qui fait le tour de l'état de situation, des problèmes techniques aux aspects sociaux. Les travaux concernant les moteurs de recherche et les bases de liens sont à envisager dans la constitution des concepts et des liens.

Simulation : Les systèmes complexes sont difficilement appréhendables. Les simulations permettent d'en faire une certaine « botanique » et de constater différents types d'émergence de structures. Certaines peuvent être relativement globales et quantitatives (les automates cellulaires), d'autres plus qualitatives.

[53] *Combinatorial Algebraic Topology*.

[54] Pour mémoire, rappelons les modèles d'architecture des systèmes hypertexte : Hypertexte abstract Machine (HAM) de Campbell & Goodman ; Treillis model de Stotts & Furuta ; the Dexter model (Halasz & Schwartz, spécification en langage Z) ; Formal Model (B.Lange, spécification écrite en VDM) ; Tower model (De Bra, Houben & Kornatzky, object-oriented) ; Layman's Hypertext (LHT)

L'analyse implicative : des didacticiens français ont mis au point la théorie de l'implication statistique (Gras, 1996) afin de représenter, voire d'analyser, les liens de dépendance entre diverses compétences d'élèves confrontés à des tâches de résolution de problème. Cette théorie serait à étudier attentivement comme complément ou alternative dans le cadre de l'usage du modèle RD dans son application.

Liens avec les utilisateurs : il y a de multiples raisons de s'intéresser aux travaux liés aux utilisateurs. Tout d'abord, il existe des procédures humaines qui peuvent être utilisées pour modéliser ou évaluer des procédures automatiques de classification et de structuration (Stein, Meyer zu Eissen & Wissbrock, 2003). Les indexations mutualisées (http://del.icio.us) sont également des travaux relevant de cette catégorie. Le deuxième domaine est celui de l'analyse des réseaux sociaux qui partage des méthodes avec l'analyse de la structure des graphes. Toutes les conséquences de cette parenté avec les techniques mises en œuvre restent encore à évaluer. Finalement, la troisième raison relève de la logique de l'usage : comprendre comment dans l'alchimie des interactions homme-machine émergent des techniques nouvelles.

15. Usages du modèle dans la formation

La structure RD permet de décomposer en deux temps l'approche des « espaces de connaissance » (Falmagne & al, s.d.). Pour cela on considère un ensemble de tâches (les documents), et un ensemble de savoirs et de compétences (les concepts). Chaque tâche est doublement analysée : les savoirs élémentaires nécessaires à son accomplissement (les référents) et les savoir globaux mis en évidence par l'accomplissement de la tâche (les descripteurs). « L'hypertexte » qui en résulte met en évidence la hiérarchie des tâches. Il en résulte un réseau de compétences qu'il serait intéressant de mettre en contact avec celui obtenu par l'analyse implicative (Gras, 1996).

De façon plus globale, les documents peuvent aussi être des chapitres de cours, et les concepts être les notions et techniques à étudier. L'étude de la linéarité de « l'hypertexte » engendré, permet par la suite de choisir un parcours de la matière.

16. Conclusion

Notre quête concerne la calibration de la connaissance. Nous sommes partis de la formule fondamentale de Brookes et avons choisi de prendre comme structure un hypertexte. Nous avons envisagé un type d'hypertexte particulier (type RD) engendré à partir de deux relations « document × concept ». Cette conclusion va aborder ce que nous avons appris : caractéristiques des hypertextes de type RD ; méthode pour « quantifier » leur structure ; rapport aux autres hypertextes. Par ailleurs, les questions ouvertes seront précisées : problème des concepts ; calibrage de l'information dans l'évolution de l'hypertexte.

Caractéristiques des hypertextes de type RD

Les hypertextes purement de type RD sont des hypertextes relativement compacts (grande densité de liens). Ils ne conviennent évidemment pas pour modéliser des constructions qui se feraient, à la manière du web, par ajout « aléatoire » de liens. Par contre, ils ont quelque parenté avec la philosophie des moteurs de recherche.

Ils peuvent être caractérisés par une matrice de circulation S, facilement calculable, qui ajoute de l'information à la matrice d'adjacence, à partir de laquelle il est possible de dériver de nombreuses caractéristiques, dont les coefficients de structure classique (centralité, compacité, linéarité).

De façon « artificielle », tout hypertexte peut se mettre sous la forme d'un hypertexte de type RD (non unique). L'apport de ce passage reste à étudier.

La quantification de la structure, la matrice de circulation

La matrice de circulation constitue un palier à notre quête. Nous proposons comme implémentation de la formule de Brookes le foncteur **K** qui associe à une structure H (notée S dans Brookes) la matrice de circulation S que nous avons tout d'abord définie par un calcul d'inverse ; mais la version corrigée (S – I) ou la formule logarithmique pourrait parfois être mieux adaptée (page 20 et annexe 2). Il resterait à étudier de façon expérimentale le faisceau de matrices constitué par les matrices liées aux diverses transformations (réduction en composantes) possibles de l'hypertexte. Notamment : la somme, la juxtaposition, la superposition et le « quotientage »[55].

[55] Par exemple étudier les propriétés de l'opération $S_1 \oplus_J S_2$ avec $S_i = \mathbf{K}(H_i)$ définie par $S_1 \oplus_J S_2 = \mathbf{K}(H_1 \oplus_J H_2)$. Le foncteur **K** se décompose en trois partie. Une partie simple se traduit par des opérations matricielles élémentaires. La deuxième partie, calcul de $\overline{G} = \overline{R}\overline{D}^T$, introduit une sorte de somme pondérée des constituants. La troisième partie, calcul du log, ne se laisse pas apprivoiser aussi simplement.

Dans le cas du « quotientage » selon les concepts, lorsque les concepts peuvent être regroupés en classes, il est possible de considérer un hypertexte plus « grossier » avec les mêmes unités d'information. Un autre changement d'échelle consiste à affiner ou à résumer les contenus des unités d'information.

Dans tous ces cas, il reste également à imaginer comment tirer de l'information entre les différents S et les opérations de « troncages ».

Usage d'autres hypertextes

Notre technique ne s'adresse qu'aux hypertextes (structures d'information) de type RD. Cette structure recouvre toutefois un champ d'application assez large. Poussé à l'extrême, il rejoint le domaine des réseaux neuronaux. Dans le cas des structures d'information qui ne remplissent pas le critère, il est possible de créer des matrices R et D (en passant par la matrice d'incidence B) de sorte à retrouver la matrice d'adjacence à partir de R et D, selon le procédé présenté à la page 20. Dans ce cas il est vraisemblable que seule la partie « association » A de la décomposition A+S+R soit prise en compte (voir page 18).

A la recherche des concepts

Ce travail relève des problèmes d'indexation pour lesquels de nombreuses procédures sont disponibles depuis les travaux « historiques » de Salton (Salton & al., 1996), les définitions de hiérarchies (Woods, 1997) et d'ontologies plus sophistiquées. Les techniques « d'enrichissement » (Roux, 2004), qui consistent à ajouter de nouveaux concepts à d'anciens documents à partir de nouveaux documents, sont aussi à considérer. Les processus automatiques posent le problème du niveau de contrôle « humain » souhaité dans la fabrication des catégories du tout automatique (Google) à la constitution préalable de thésaurus (Yahoo) en passant par des processus mixtes (http://del.icio.us).

Evolution et information

Divers processus d'évolution peuvent être considérés. Un premier processus est lié à l'ajout de nouvelles unités d'information en gardant les mêmes référents. Ce processus correspond à une opération déjà décrite. Du point de vue des informations, les autres processus sont « adiabatiques », ils consistent en des réorganisations internes. Parmi ceux-ci on peut considérer l'ajout ou la réorganisation des concepts. Des unités d'informations peuvent « naître » de nouveaux concepts qui iront enrichir le réseau sémantique. Un autre processus consisterait à regrouper ou décomposer des unités d'informations.

Ces processus adiabatiques demanderaient de mettre en perspective l'information de la structure (**K**(H) = S) et l'information contenue dans les unités d'information. Une étude attentive de ces changements d'état internes permettrait peut-être de définir une « équivalence » entre ces deux types d'information. Cette perspective demande certainement d'introduire une entité d'observation extérieure (modélisée par l'utilisateur de l'hypertexte, par exemple) par rapport à laquelle l'information pourrait être définie.

Remerciements à Corinne Martin et Floriane Pochon pour leur relecture attentive.

Bibliographie

Généralités et aspects psychosociaux

Engelbart, D.C. (1962). *Augmenting human intellect : a conceptual framework*. Memlo Park : Stanford research institute (AFOSR-3233).

Foray, D. (1997). *Abondance d'informations, disette de connaissances : conférence du cycle "L'Homme et le Temps".* Neuchâtel : Université de Neuchâtel.

Licklider, J.C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics, HRE-1*, 4-11.

Postman, N. (1993). *Technopoly : the surrender of culture to technology*. New York : Vintage Books.

Turkle, S. (1995). *Life on the screen : identity in the age of internet*. New York : Simon & Schuster.

Winograd, T. & Flores, F. (1990). *Understanding computer and cognition : a new foundation for design*. Norwood, NJ : Ablex.

Théorie de l'information

Brookes, B.C. (1980a). The foundations of information science. Part I : Philosophical aspects. *Journal of information science, 2*(3/4), 125-133.

Kauffman, S.A. (1993). *The origins of order : self-organization and selection in evolution*. New York : Oxford University Press.

Lenski, W. (2004). Remarks on a publication-based concept of information. *In Proceedings of the satellite conference to the ECM 2004 on new developments in electronic publishing of mathematics, Stockholm, June 2004* (pp. 119-135). [Sl.] : [s.n.].

Watanabe, S. (1969). *Knowing and guessing. A quantitative study of inference information.* New.York: John Wiley and Sons.

Analyse documentaire

Berney, J. & Pochon, L.-O. (2000). *L'Internet à l'école : analyse du discours à travers la presse*. Neuchâtel : IRDP (00.5).

Berry, M.W., Dumais, S.T. & Shippy, A.T. (1995). *A case study of latent semantic indexing*. Knoxville : University of Tennessee, Computer science department.

Marcoux, Y. (1996). *Place de SGML parmi les nouvelles architectures documentaires, Technologie SGML 1996, Château Laurier, Ottawa*. (http://tornade.ere.umontreal.ca/~marcoux/ottawa/marcoux.htm, page consultée en octobre 1998).

Roux, C. (2004). *Indexation automatique pour le traitement des corpus hétérogènes : atelier EGC 2004* (consulté en janvier 2007).

Salton, G., Allan, J., Buckley, C. & Singhal, A. (1996). Automatic analysis, theme generation, and summarization of machine-readable texts. In M. Agostini & A. Smeaton (Eds), *Information retrieval and hypertext* (pp. 51-73). Boston : Kluwer Academic Publishers.

Stein, B, Meyer zu Eissen, S. & Wissbrock, F. (2003). On cluster validity and the information need of users. In M.H. Hanza (Ed.), *3rd IASTED International conference on artificial intelligence and applications (AIA 03), Benalmádena, Spain, September 2003* (pp. 216-221). [S.l.] : [s.n.].

[In the field of information retrieval, clustering algorithms are used to analyze large collections of documents with the objective to form groups of similar documents. Clustering a document collection is an ambiguous task: A clustering, i. e. a set of document groups, depends on the chosen clustering algorithm as well as on the algorithm's parameter settings. To find the best among several clusterings, it is common practice to evaluate their internal structures with a cluster validity measure.

A clustering is considered to be useful to a user if particular structural properties are well developed. Nevertheless, the presence of certain structural properties may not guarantee usefulness from an information retrieval standpoint, say, whether or not the found document groups resemble the classification of a human editor.

The paper in hand investigates this point: Based on already classified document collections we generate clusterings and compare the predicted quality to their real quality. Our analysis includes the classical cluster validity measures from Dunn and Davies-Bouldin as well as the new graph-based measures Λ (weighted edge connectivity) and $\overline{\rho}$ (expected edge density). The experiments show interesting results: The classical measures behave in a consistent manner insofar as mediocre and poor clusterings are identified as such. On real-world document clustering data, however, they are definitely outperformed by the expected edge density $\overline{\rho}$. This superiority of the graph-based measures can be explained by their independence of cluster forms and distances.]

Woods, W.A. (1997). *Conceptual indexing : a better way to organize knowledge*. [S.l.] : Sun Microsystems Laboratories (Technical report series TR-97-61).

Techniques de base, méthodes de classification, théorie des graphes, analyse reconstructive

Barabasi, A.-L. (2002). *Linked, the new science of networks*. Cambridge, MA : Perseus Publishing.

Q Ding, C.H., He, X., Zha, H. & Simon, H.D. (1988). *A min-max cut algorithm for graph partitioning and data clustering* (rapport interne).

[La technique usuelle de Max-Min Cut (MaxFlow MinCut) sépare des portions de graphes peu connectées. Mais elle résulte souvent dans une décomposition « biaisée » contenant des sous graphes de tailles très diverses, en particulier de petits sous-graphes. Pour éviter ceci on a imaginer les notions de Standard-Cut resp. Mormalised-Cut. Ce n'est pas suffisant. Par ailleurs, la recherche de liens de « cut » est NP-Complete. On utilise alors la direction principale (Ordre de Fiedler) de la matrice (pondérée) d'adjacence. On examine les points successivement le long de cette direction. L'article introduit une métrique qui permet d'identifier rapidement les points « proches » des cuts. Cette technique permet une recherche accélérée des composantes dans l'hypertexte.]

Q Ding, C.H., He, X. & Zha, H. (2001). *A spectral method to separate disconnected and nearly-disconnected web graph components*. New York : ACM Press.

[En utilisant la matrice dite « Laplacien » d'un graphe, les valeurs propres nulles permettent de déterminer les composantes connexes d'un graphe. Le Laplacien est définit sur un graphe non-dirigés. C'est la matrice d'adjacence (symétrique) et sur la diagonale on met la multiplicité de chaque nœud, changée de signe. On s'attache aux valeurs propres nulles. Théorème : il y a autant de composantes connexes que de vecteurs propres (indépendants) associés à la valeur propre 0. Dans l'article, il est démontré que ces vecteurs propres montrent des fonctions à escalier après ré-ordonnancement des valeurs. Il est démontré que ces différentes marches de l'escalier correspondant à des composantes « presque déconnectées ».]

Zwick, M. (2002). *An overwiew of reconstructability analysis : proc. 12th International world organisation of systems and cybernetics*. [S.l.] : [s.n.].

[Approche de la modélisation de relations discrètes multi-variables. Diverses techniques sont présentées ou signalées pour décomposer une relation multi-variables en « blocs » aussi indépendants que possibles. S'applique aux relations données par des matrices de co-occurrences (information-theoretic) et aussi par des occurrences de cette relation (set-theoretic). La 2ème partie (set-theoretic) pourrait être utile pour décomposer une matrice « mot-document », à suivre.]

Réseaux sémantiques & Ontologies

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web. *Scientific American, may*, 35-43.

Pochon, L.-O. (2006). *De la possibilité d'usages d'ontologies pour la gestion de contenus mathématiques*. Neuchâtel : IRDP (document de travail 06.1007).

Cluster & communautés web

Dousset, B. & Karouach, S. (2005). *Manipulation de graphes de grande taille pour l'étude des réseaux d'acteurs et des réseaux sémantiques : journée sur les systèmes d'information élaborée, Ile Rousse, 2005* (http://isdm.univ-tln.fr/PDF/isdm22/isdm22_dousset.pdf, page consultée en janvier 2007).

Flake, G.W., Lawrence, S. & Giles, C.L. (2000). Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150-160). New York : ACM Press.

Flake, G.W., Lawrence, S. & Giles, C.L. (2002). Self-organisation of the web and identification of communities. *IEEE Computer, 35*(3), 66-71.

[Une communauté Web est définie comme un sous-graphe du web qui a plus de liens en interne que vers l'extérieur de la communauté. Le papier présente un algorithme basé sur le théorème de « Maxflow Mincut ». Il définit comme germe des communautés des sites de références auxquels on accroche des sources (factices) à énorme débit. Une procédure en deux étapes permet une approximation correcte d'une communauté. Cet article peut être utile pour effectuer le regroupement des unités d'information dans un hypertexte. Mais aucune « chosification » n'est fournie. C'est-à-dire qu'aucun concept (ou descripteur) ne ressort de l'isolement d'une communauté. La méthode ne suggère que d'utiliser le site de référence utilisé comme index.]

Gibson, D. & Kleinberg, J. (1998). Inferring web communities from link topology. In *Proc. 9th ACM Conf. On Hypertext and Hypermedia*. [S.l.] : [s.n.].

[Présente l'algorithme HIT. Une communauté est caractérisée par un « centre » de pages « d'autorités » reliées par des pages « hub ». Il en ressort un type naturel de thèmes hierarchisés que l'on peut découvrir à partir de la topologie des liens.]

Kleinberg, J.M. (1998). *Authoritative sources in a hyperlinked environment : Proc. 9th ACM-SIAM Symposium on discrete algorithms*. [S.l.] : [s.n.].
Also appears as IBM Research Report RJ 10076, May 1997.

[Surtout intéressant par la technique préliminaire d'extraction d'une partie finie du web avant d'appliquer des techniques standard. Définition des hubs et des autorités.]

Hypertextes

Balpe, J.-P. (1990). *Hyperdocument*. Paris : Eyrolles.

Balpe, J.-P., Lelu, A., Papy, F. & Saleh, I. (1996). *Techniques avancées pour l'hypertexte*. Paris : Hermès.

Botafogo, R.A., Revlin, E. & Schneiderman, B. (1992). Structural analysis of hypertexts : identifying hierarchies and useful metrics. *ACM transactions on information sytems, 10*(2), 142-180.

Cornali, I. & Weiss, J. (éds). (1996). *Des utopies à construire*. Neuchâtel: : Institut romand de recherches et de documentation pédagogique (IRDP) ; Le Mont-sur-Lausenne : Loisirs et pédagogie (LEP) (livre et CD-ROM, version Internet adaptée par Floriane Pochon, http://www.irdp.ch/utopies/utopies.htm).

Lowe, D. & Hall, W. (1999). *Hypermedia and the web*. New York : John Wiley & Sons.

[Utile notamment pour caractériser les parcours des usagers (par exemple ceux qui naviguent par menu et ceux qui naviguent par association). voir aussi Bourquard (les remontées au menu).]

Ohene-Djan, J.F. (2000). *A formal approach to personalisable, adaptative, hyperlink-based systems*. London : University of London, Goldsmiths College, Department of Mathematical and Computing Sciences.

Enseignement et formation

Falmagne, J.-C., Cosyn, E., Doignon, J.-P. & Thiéry, N. (s.d.). *The assessment of knowledge in theory and in practice* (Science_Behind_ALEKS.pdf sur http://www.aleks.com/).

Forte, E. (2002). La cyberformation : vaincre la distance, le temps et l'espace. *Bulletin HEC, 64*, 27-28.

Gras, R. (1996). *L'implication statistique : nouvelle méthode exploratoire de données*. Grenoble : La Pensée sauvage.

Pahl, C. (2000). The evaluation of educational service integration in integrated virtual courses (http://odtl.dcu.ie/wp/2000/odtl-2000-06.html).

[Caractéristiques globales et de parcours]

Pochon, L.-O. (1993). *Hypertextes pour apprendre*. Neuchâtel : IRDP.

Vidal, P., Cardinaels, K., Duval, E. & Forte, E. (1998). A knowledge pool system of reusable pedagogical elements. In C. Alvegard (Ed.), *CALISCE 98, Göteborg, juin 1998* (pp. 54-62). Göteborg : Chalmers University of Technology.

Homologie

Babson, E., Barcelo, H., De Longueville, M. & Laubenbacher, R. (s.d). *Homotopy theory of graphs*. (graph-Homotopy.pdf) (janvier 2006).

Barcelo, H., Kramer, X., Laubenbacher, R. & Weaver, C. (s.d.). *Foundations of a connectivity theory for simplicial complexes* (atheory_final.pdf) (janvier 2006).

Valencia, V. (1998). *Introduction au formalisme CAT pour la représentation spatiale des connaisance*. Paris : LRI, ura 410 ; Université Paris-sud. (intro-CAT.ps, consulté : novembre 2001).

Link services, moteurs de recherche

Carr, L., Hall, W. & De Roure, D. (1999). The evolution of hypertext link services. *ACM computing surveys, Decembre*.

Carr, L., Bechhofer, S., Goble, C. & Hall, W. (2001). Conceptual linking : ontology-based open hypermedia. *WWW10, May*, 2-5.

[Description de différents « link services » et essai de classifications « open x richesse des metadata ».]

Vaughan-Nichols, S.J. (2003). Seeking web search technologies. *IEEE Computer, August*, 19-21.

[la lutte entre les moteurs de recherche (enjeux retombées)]

Autres documents consultés

Généralités et aspects psychosociaux

Brockman, J. (Ed.). (2002). *The next fifty years : science in the first half of the twenty-first century*. London : Weidenfeld & Nicolson.

Brookes, B.C. (1980b). The foundations of information science. Part II : Quantitative aspects : classes of things and the challenge of human individuality. *Journal of information science, 2*(5), 209-221.

Brookes, B.C. (1980c). The foundations of information science. Part III : Quantitative aspects : classes of things and the challenge of human individuality. *Journal of information science, 2*(6), 269-275.

Brookes, B.C. (1981). The foundations of information science. Part IV : Information science : the changing paradigm. *Journal of information science, 3*(1), 3-12.

Minsky, M. (1988, 1e ed. 1985). *Society of mind*. New York : Simon & Schuster.

Shank, R.C. (2002). Are we going to get smarter ? In J. Brockman (Ed.), *The next fifty years : science in the first half of the twenty-first century* (pp. 206-215). London : Weidenfeld & Nicolson.

Techniques de base, méthodes de classification, théorie des graphes, analyse reconstructive

Caelli, T. & Kosinov, S. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE transactions on pattern analysis and machine intelligence, 26*(4), 515-519.

[Une méthode de simplification de graphes à but de comparaison. Elle est basée sur la décomposition spectrale dont on ne prend que les premières dimensions. A lier à la méthode utilisée pour tronquer la clôture transitive.]

Cellier, F.E. & de Albornoz, A. (s.d.). The problem of distorsions in reconstruction analysis.

[Problème technique d'amélioration d'un algorithme de reconstruction.]

Favre, A. & Pochon, L.-O. (2003). *About entropy* (http://www.irdp.ch/thema/entrouk.pdf).

Gonzales, J.A., Holder, L.B. & Cook, D.J. (2001). *Graph-based concept learning*. [S.l.] : American association for artificial intelligence.

[Deux techniques principales pour l'apprentissage de concepts: Graphs-based ou Inductive-Logic-Programming. Parmi les approches par graphes : « Graphes conceptuels » ou « Treillis de Galois ». Les graphes conceptuels sont des réseaux sémantiques, connus. L'espace de recherche d'un treillis de Galois consistent dans dans l'ensemble des généralisations possibles fournies par un ensemble d'entrainement. La méthode présentée dans le papier (SubDue) est un outil de data-mining qui regroupe des catégories présentées par des exemples et un techniques d'apprentissage de relations. Utile pour la généralisation de mots pour obtenir des concepts.]

Johnson, M.S. & Zwick, M. (2000). State-based reconstructability modelling for decision analysis. In J.K. Allen & J.M. Wilby (Eds), *Proceedings of the world congress of the systems sciences and ISSS 2000*, Toronto : International society for the systems sciences.

[L'analyse reconstructive est une méthode qui permet de repérer des structure dans un phénomène multivarié. Cet article fait le lien avec la méthone dite "K-systems analysis" (Jones) et donne 2 exemples.]

Kolmogorov, V. & Zabih, R. (2004). What energy function can be minimized via graph cuts? *IEEE transaction on pattern analysis and machine intelligence, 26*(2), 147-159.

Lam, W., Keung, C. & Liu, D. (2002). Face and gesture recognition : descovering useful concept prototypes for classification based on filtering and abstraction. *IEEE transactions on patterns analysis and machine intelligence, 24*(8), 1075-1090.

Willett, K. & Zwick, M. (2002). *A software architecture for reconstructability analysis : Proceedings of the 12th World organization of systems and cybernetics and the 4th International institute for general systems studies workshop*. Pittsburgh : [s.n.].

Zwick, M. (s.d.). *Control uniqueness in reconstructability analysis*. Portland : State University (OR 97207).

Zwick, M. (2001). Wholes andparts in general systems methodology. In G. Wagner (Ed.), *The character concept in evolutionary biology*. [S.l.] : Academic Press.

Réseaux sémantiques & Ontologies

Barnes, J. & Robertson, J. (2002). The use of ontologies in drug. *Bioinformatics world, Automn*, 8-10.

[Dans la recherche de nouveaux médicaments, la documentation concernant chaque molécule est essentielle. Une grande partie du travail des chercheurs consiste à mettre en relation diverses sources d'information pour en tirer des nouvelles conclusions sur l'effet de tel ou tel produit. D'où l'importance d'une classification performante de cette documentation. L'intérêt d'un tel travail est l'aspect évolutif de la base de donnée, aussi bien au niveau du contenu que de la structure.]

Chakrabarti, S., Agrawal, R & Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organising large texte data-base into hierarchical topic taxonomies. *VLDB Journal, 7*, 1631-1678.

[Définition d'ontologies par "statistical pattern recognition". La techniques est fondées sur des mots "features" et des mots "noise". La hiérarchie est définie en changeant les étiquettes de ces mots. L'algorithme est fondé sur une matrice 3D (mot, document, concept) et une méthode d'agrégation. Tout ceci suppose une densité de probabilité calculées sur les éléments de la matrice.]

Fensel, D. (2002). Ontology based knowledge management. *IEEE Computer, november*, 56-59.

Ketterlin, A., Dondainas, N. & Raffy, M. (s.d.). *Classification d'ensembles et distance de Hausdorff*. Paris : CNRS, Laboratoire de Sciences de l'Image, de l'Informatique et de la Télévision (UPRES-A 7005).

[classification de nuages de points.]

Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (2002). The web and social networks. *IEEE Computer, november*, 32-36.

[les enjeux, de « l'Ontology Based Knowledge Management » ; l'image des sociétés reflétée par le web ; sociologie de la création de contenu (utile pour fabriquer des moteurs intelligents)]

Lindley, C.A., Kumar, V.R., Irrgang, R. & Robertson, J.R. (révision 2004). *An evaluation of information retrieval methods and semantic network processing for automatic link generation in hypermedia systems* (http://www.aset.org.au/confs/iims/1994/km/lindley.html, page consultée le 23 mars 2004).

Menzies, T. & Hu, Y. (2003). Datamining for very busy people. *IEEE Computer, November*, 22-29.

[Information apportée par différence plutôt que par communauté.]

Nishida, T. (2002). Social intelligence design for the web. *IEEE Computer, november*, 37-41.

Quillian, M.R. (1968). Semantic memory. In M. Minsky (Ed), *Semantic information processing* (pp. 216-260). Cambridge, MA : MIT Press.

Sowa, J.F. (2002). *Semantic networks* (http://www.jfsowa.com/pubs/semnet.htm, révision août 2002, page consultée le 23 mars 2004).

Cluster & communautés web

Rangarajan, S.K., Phoha, V.V, Balagani, K.S. & Selmic, R.R. (2004). Adaptive neural network clustering of web users. *Computer, April*, 34-40.

[Présentation d'un algorithme basé sur ART1 pour faire une analyse cluster (système neuronal). L'exemple traité concerne la classification d'internautes en fonctions des sites visités.]

Hypertextes

Bourquard, (1998). Prof'Expert : une expérience d'enseignement assisté par ordinateur dans le cadre d'une formation pour adultes au Centre de formation professionnelle du littoral neuchâtelois. *Dossiers de psychologie, 53*.

Rada, R., Zeb, A., You, G.-N., Michailidis, A. & Mhashi, M. (1991). Collaborative hypertext and the MUCH system. *Journal of information science, 17*, 191-196.

Enseignement et formation

Koper, R. & Manderveld, J. (2004). *Educational modelling language : modelling rusable, interoperable, rich and personalised units of learning* (http://www.eml.nl).

[Présentation synthétique du modèle EML en UML et XML. Offre de cours détaillée et « normalisée » (voir aussi « profetic »)]

McClelland, M. (2003). Metadata standards for educational resources. *Computer, November*, 107-109.

[Bref survol de la problématique]

Homologie

Degtiarev, K.Y. (2000). An approach to analysis of large-scale systems structures : a joint use of fuzzy set theory and algebraic topology methods. In *The 9th Turkish symposium on artificial intelligence and neural networks, Izmir, June 21-23 2000* (pp. 149-156).

Popovic, J. & Hoppe, H. (1997). Progressive simplicial complexes. In *Proceedings of the 24th annual conference on computer graphics and interactive techniques* (pp. 217-224).

Annexe 1 : Utilisation du simulateur et étude d'un cas simple

Le simulateur (Matlab[56]) permet de créer des unités d'information, chacune caractérisée par :

- un numéro d'ordre ;
- son degré de « fitness » qui représente une certaine qualité intrinsèque de l'unité. Le degré de fitness est un nombre aléatoire compris entre 0 et 1 ;
- le nombre de liens émis depuis cette unité d'information ;
- son type qui donne la nature de l'unité d'information (domaine traité).

Le nombre et le type de liens sont attribués au hasard selon des distributions décrites dans les deux tables :

- *dis_lnk* donne la distribution du nombre de *links* émis par une nouvelle unité d'information (distribution normale dans l'exemple) ;
- *dis_typ* donne la distribution des types (distribution croissante dans l'exemple).

Les liens sont fabriqués selon un coefficient d'attirance qui dépend des types respectifs de la source et de la cible, du degré de fitness de la cible et du nombre de liens de la cible.

La matrice *uis* contient l'ensemble des unités d'information (une par ligne), les colonnes donnant le numéro, le degré de fitness, le nombre de liens et le type.

r est la matrice d'adjacence. rij = 1 signifie qu'il y a un lien de l'unité d'information i vers l'unité d'information j.

La commande *load simul* initialise les variables (y compris 2 unités d'information).

Le commande *agreg (13,5)* ajoute 13 unités d'information, puis elle choisit 5 unités d'information à partir desquelles des liens sont établis selon le même algorithme que précédemment. La matrice obtenue est :

[56] Les fonctions existent également pour R. Une différence de performance est néanmoins perceptible. La création d'un hypertexte de 1000 unités prend quelques secondes avec Matlab, plusieurs minutes sous R.

```
      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
 r =

1     0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
2     1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
3     0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
4     0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
5     0   0   0   1   0   0   0   0   0   0   0   0   0   0   1
6     0   0   0   1   1   0   0   0   0   0   0   0   0   0   0
7     0   1   0   1   1   1   0   0   0   0   0   0   1   0   1
8     0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
9     0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
10    0   0   0   0   1   1   1   0   0   0   0   0   0   0   0
11    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
12    0   0   0   0   1   1   1   0   0   0   0   0   0   0   0
13    0   0   0   0   1   1   0   0   0   1   1   0   0   0   1
14    0   0   0   0   0   0   1   0   0   1   0   1   1   0   1
15    0   0   0   0   0   0   1   0   0   0   0   1   1   0   0
```

La commande *tri_r* permet d'obtenir une statistique. Pour chaque unité d'information, on calcule l'ordre du sommet, le nombre de liens entrants (descripteurs) et le nombre de liens sortants (référents).

UI	ordre	nb de descr.	nb de réf.
7	10	4	6
5	8	6	2
15	8	5	3
13	8	3	5
6	6	4	2
12	5	2	3
10	5	2	3
14	5	0	5
4	4	3	1
2	3	2	1
11	1	1	0
1	1	1	0
8	1	0	1
3	1	0	1
9	0	0	0
tot.	66	33	33

On voit que les unités 7, 5, 15 et 13 constituent des unités centrales. Ce qui se note également sur la figure A1.1.

Les unités 14, 8, 3 n'ont pas de liens entrants (ce sont des sources), les unité 1 et 11 sont au contraire des puits (pas de liens sortants). L'unité 9 est isolée. Le travail ici est « local ». Le cas « régional » sera abordé ci-dessous.

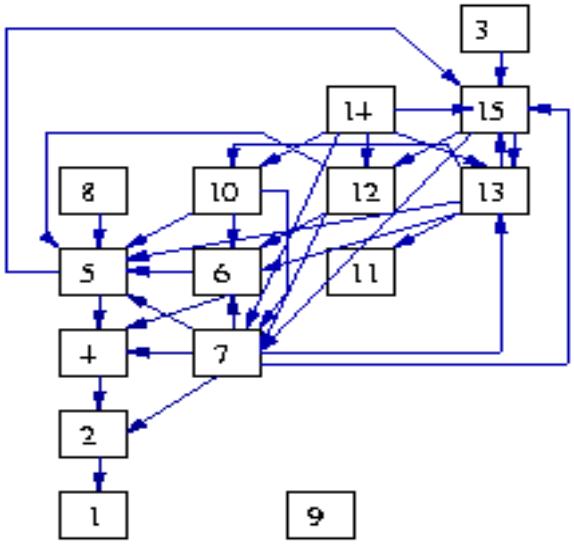


Figure A1.1. *Représentation graphique de l'hypertexte*

Analyse

Le but est de trouver la structure "papillon" de l'hypertexte. L'unité d'information 7 est un candidat pour la création du "CORE". En calculant la clôture transitive r + ... + r^15 (mod 2) on trouve toutes les unités d'informations qui lui sont liées.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
6	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
7	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
8	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
13	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
14	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1
15	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1

On observe qu'il n'est pas possible de revenir sur 7 à partir des unités d'information (1 2 4) qui sont à exclure du CORE. Finalement celui-ci se compose des unités (5 6 7 10 12 13 15).

Finalement:

- IN = (3 8 14)
- CORE = (5 6 7 10 12 13 15)
- OUT = (1 2 4 11)

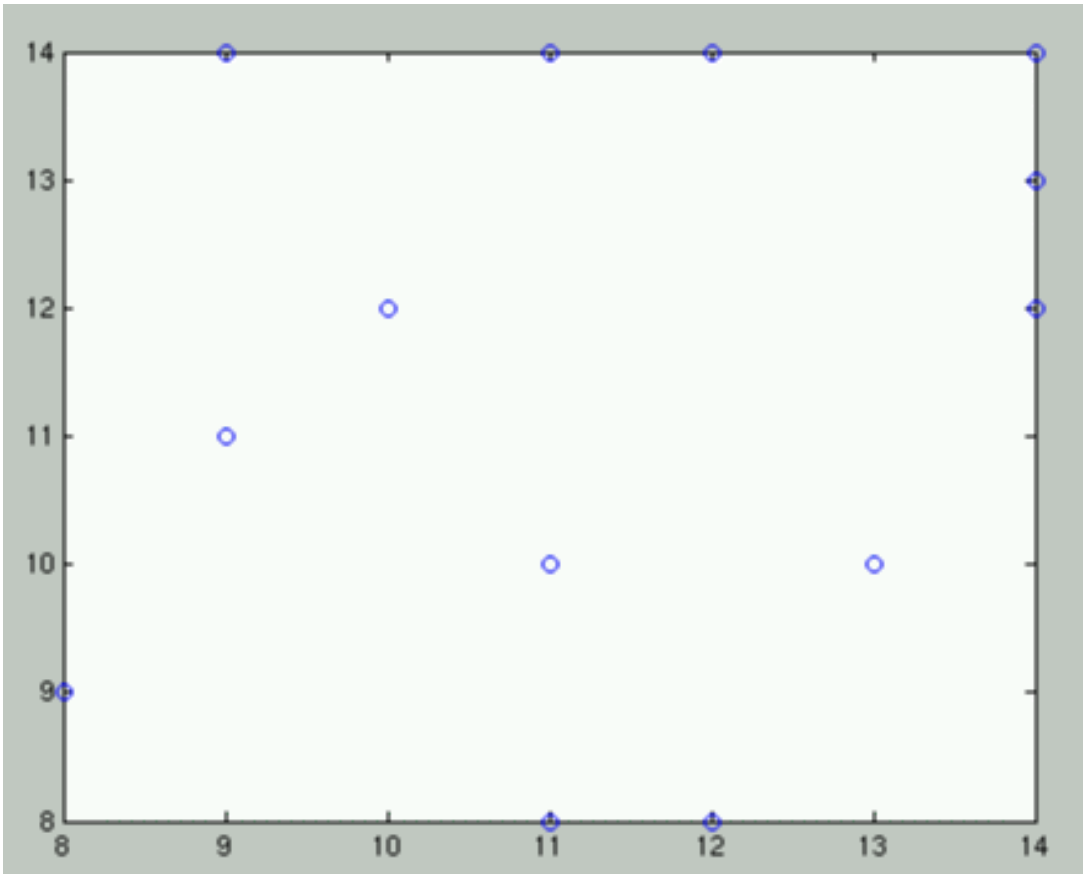


Figure A1.2. *Diagramme de connectivité de l'hypertexte*

Il est également possible d'utiliser la normalisation donnée par la matrice S après avoir reconstruit les matrices D et R. Cela fournit l'analyse « régionale ».

L'introduction d'une coupure à 0.1, montre la structure représentée dans la figure A.1.2.

Les valeurs numériques obtenues sont les suivantes :

UI	7	15	5	13	4	6	10	12	14	2	8	3	1	11	9
DX	8	11	11	9	12	11	10	10	9	13	11	12	14	14	14
DY	9	8	8	11	8	10	12	12	14	10	14	14	12	13	14
Total	17	19	19	20	20	21	22	22	23	23	25	26	26	27	28

où DX représente un manque en référents (la valeur 14 indique un puits) et DY un manque en descripteur (la valeur 14 indique une source).

Les connectivités locale et régionale sont relativement comparables.

Utilisation du laplacien combinatoire

La commande *l=lap(r)* calcule le laplacien combinatoire de r dont la version « pleine » (*l1=full(l)*) est :

```
 1, -1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0
-1,  3,  0, -1,  0,  0, -1,  0,  0,  0,  0,  0,  0,  0,  0
 0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, -1
 0, -1,  0,  4, -1, -1, -1,  0,  0,  0,  0,  0,  0,  0,  0
 0,  0,  0, -1,  8, -1, -1, -1,  0, -1,  0, -1, -1,  0, -1
 0,  0,  0, -1, -1,  6, -1,  0,  0, -1,  0, -1, -1,  0,  0
 0, -1,  0, -1, -1, -1,  9,  0,  0, -1,  0, -1, -1, -1, -1
 0,  0,  0,  0, -1,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0
 0,  0,  0,  0, -1, -1, -1,  0,  0,  5,  0,  0, -1, -1,  0
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0, -1,  0,  0
 0,  0,  0,  0, -1, -1, -1,  0,  0,  0,  0,  5,  0, -1, -1
 0,  0,  0,  0, -1, -1, -1,  0,  0, -1, -1,  0,  7, -1, -1
 0,  0,  0,  0,  0,  0, -1,  0,  0, -1,  0, -1, -1,  5, -1
 0,  0, -1,  0, -1,  0, -1,  0,  0,  0,  0, -1, -1, -1,  6
```

Les valeurs propres de l1 sont : 0 ; 0 ; 0.57 ; 0.83 etc. Il y a donc deux composantes connexes, comme déjà observé (l'unité isolée 9 et le reste).

Le vecteur propre correspondant à la valeur propre 0.57 est :

```
 1   0.82082
 2   0.35583
 3  -0.25851
 4   0.062867
 5  -0.07151
 6  -0.050681
 7  -0.017781
 8  -0.16496
 9   0
10  -0.073295
11  -0.22962
12  -0.076119
13  -0.099539
14  -0.08544
15  -0.11206
```

Il nous indique que, outre l'unité 9 isolée, les unités C1 = {1, 2, 4} (coefficients positifs) forment un paquet relativement peu connecté au reste C2 = {3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15}. C'est une approximation qui minimise le coefficient cut(C1,C2) * (1/#C1 + 1/#C2)

Approche alternative

La théorie (Kleinberg, 1997) indique que les "hubs" correspondent aux unités d'information liées aux composantes maximales du vecteur propre associé à la valeur propre maximum de la matrice d'adjacence multipliée par sa transposée. Pour les "autorités", il faut prendre le produit de la transposée par la matrice d'adjacence.

Dans notre cas r'* r et r * r' ont pour valeur propre maximum 14.6. Les vecteurs propres correspondants sont v1 et v2:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
v1	0	.15	0	.24	.52	.43	.32	0	0	.21	.12	.15	.29	0	.43
v2	0	0	.11	.04	.18	.20	.54	.13	0	.34	0	.34	.45	.37	.20

Les unités 5, 6, 15, 7 et 13 sont des « autorités ». Les unités 7, 13, 12, 10 et 14 sont plutôt des « portails ». Les unités 7 et 13 sont des « hubs ». A noter que les composantes d'un des vecteurs pour les « puits », les « sources » et les unités isolées (unités 1, 2, 3, 8, 9, 11 et 14) sont nulles.

Annexe 2 : H, S et le foncteur K

Le foncteur K a été programmé dans le système R. Un hypertexte est défini par sa structure (R,D).

```
> H
$R
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1    0    0    0
[2,]    0    0    1    0    0
[3,]    0    0    0    1    0
[4,]    0    0    0    1    1
[5,]    0    0    0    0    0
[6,]    0    0    0    1    0

$D
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    1    0    0    0    0
[3,]    0    1    1    0    0
[4,]    0    0    0    0    0
[5,]    0    0    0    1    1
[6,]    0    0    0    1    0
```

Ki [H][57] est donné par la valeur $(1\text{-}GB/2)^{-1} - I$ avec GB = RB * DB' ([58])

```
> Ki(H)
     [,1] [,2]  [,3] [,4]  [,5]  [6,]
[1,]    0 0.25 0.375    0 0.125 0.125
[2,]    0 0.00 0.500    0 0.167 0.167
[3,]    0 0.00 0.000    0 0.333 0.333
[4,]    0 0.00 0.000    0 0.417 0.167
[5,]    0 0.00 0.000    0 0.000 0.000
[6,]    0 0.00 0.000    0 0.333 0.333
```

Kl [H] est donné par la valeur - log (1-GB/2)

```
> Ki(H)
     [,1] [,2]   [,3] [,4]   [,5]   [6,]
[1,]    0 0.25 0.3125    0 0.0127 0.0127
[2,]    0 0.00 0.5000    0 0.0501 0.0501
[3,]    0 0.00 0.0000    0 0.2751 0.2751
[4,]    0 0.00 0.0000    0 0.3875 0.1375
[5,]    0 0.00 0.0000    0 0.0000 0.0000
[6,]    0 0.00 0.0000    0 0.2751 0.2751
>
```

Les opérations algébriques sur les hypertextes sont données par les fonctions : sumH, juxH, supH. Quant aux opérations quotients, elles sont données par quotU et quotC.

[57] Ki[H] pour reprendre la notation de Brookes, qui devient la fonction R : Ki(H). A noter que dans la pratique c'est Ki(H) – I qui est souvent le plus utile.

[58] Le B de GB, RB et DB est mis pour « barre ».

Annexe 3 : Table des simulations

Les différentes simulations effectuées sont résumées dans le tableau A.3.1.

No	Procédé	Paramètres	Conclusion
1	Agrégation	1 germe, 15 ui http://www.irdp.ch/thema/htxt-si3.htm	Exemple pour présenter les manipulations http://www.irdp.ch/thema/htxt-sim.htm
2	Agrégation	1 germe, 1000 ui http://www.irdp.ch/thema/htxt-si3.htm	Distribution des liens « sur » en loi de puissance. Une grande composante connexe (949 ui) et des éléments isolés. CORE de 617 ui. http://www.irdp.ch/thema/htxt-s31.htm
3	Agrégation	2 germes, 1000 ui	Résultat similaire (composante connexe de 943 ui, CORE de 608 ui). Par contre on trouve une deuxième composante connexe non triviale de 15 ui, et le reste isolées..
4	RD	1 bloc, 1000x100 Densité D 10, Densité R 20	Distribution des liens « sur » et « de » selon loi binomiale (résultat théorique et pratique). Connectivité élevée (> 40). Une composante connexe de 999 éléments. CORE de 960 ui. http://www.irdp.ch/thema/htxt-si2.htm
5	RD	1 bloc, 1000x1000 Densité D 10, Densité R 20	Chute vraisemblable de la connectivité (>2). Mais reste relativement dense. Une composante connexe de 995 éléments. 5 ui isolées CORE de 933 ui. http://www.irdp.ch/thema/htxt-si2.htm
6	RD	1 bloc, 1000x1000 Densité D 5, Densité R 10	La diminution de la densité ne modifie pas beaucoup la topologie. 20 ui isolées. CORE de 772 ui. http://www.irdp.ch/thema/htxt-si2.htm
7	RD	2 x 3 blocs, 500 x 300 Densités D : 2 3 5 / 5 3 2 Densités R : 3 7 10 / 10 7 3	Contrairement à l'hypothèse, l'hypertexte reste relativement « monobloc ». Connectivité > 20 CORE de 996 ui. Une diminution de la densité ne change pas la structure (par contre la mise à 0 fait apparaître comme il se doit un découpage en 2 hypertextes). http://www.irdp.ch/thema/htxt-s21.htm
8	RD	2 x 3 blocs, 500 x 300 Densités D : 0 2 5 / 5 2 0 Densités R : 0 3 10 / 10 3 0	CORE de 920 ui.
9	RD	2 x 3 blocs, 500 x 300 Densités D : 2 3 5 / 5 3 2 Densités R : 10 7 3 / 3 7 10	Cette configuration possible ne semble pas devoir conduire à un résultat différent. A nouveau, les concepts sont répartis et assurent une bonne connectivité.

***Tableau A.3.1**. Quelques costructions d'hypertextes*

A propos de la loi de puissance

La distribution des liens « sur » dans le cas de la simulation No 2 est la suivante :

x	3	9	15	21	27	33	39	45	51	57
f	711	156	65	40	14	8	2	2	1	1
f(z>x)	1000	289	133	68	28	14	6	4	2	1

En prenant les logarithmes, on obtient les valeurs suivantes représentées dans la figure A3.1 qui montre une distribution de Pareto de coefficient –2.3 :

ln x	1.10	2.20	2.71	3.04	3.3	3.5	3.66	3.81	3.93	4.04
ln F	6.91	5.67	4.89	4.22	3.33	2.63	1.79	1.39	0.69	0

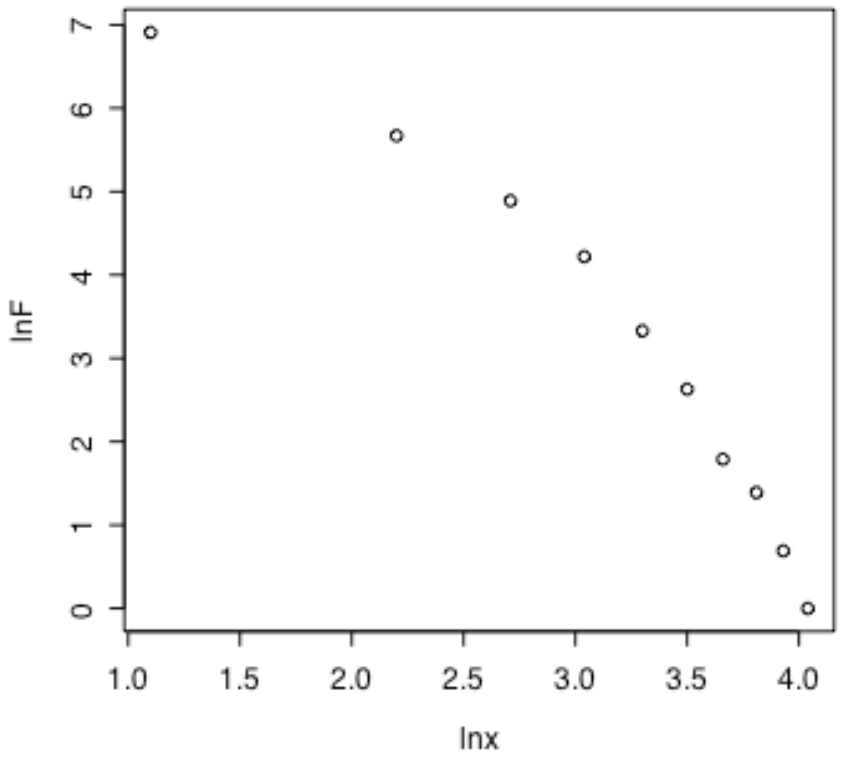


Figure A3.1. *Représentation double logarithmique de la distribution des liens « sur »*

Calcul de #K

Dans le cas de la simulation 8, la distribution de #K3(u) a été étudiée. Elle est donnée par la figure A3.2. La valeur de #K3(u39) (= 917) fait de u39 un véritable hub. La valeur minimum non nulle est : #K3(u508) = 127.

Pour le même hypertexte, le diagramme de connectivité présente une certaine homogénéité (figure A3.3).

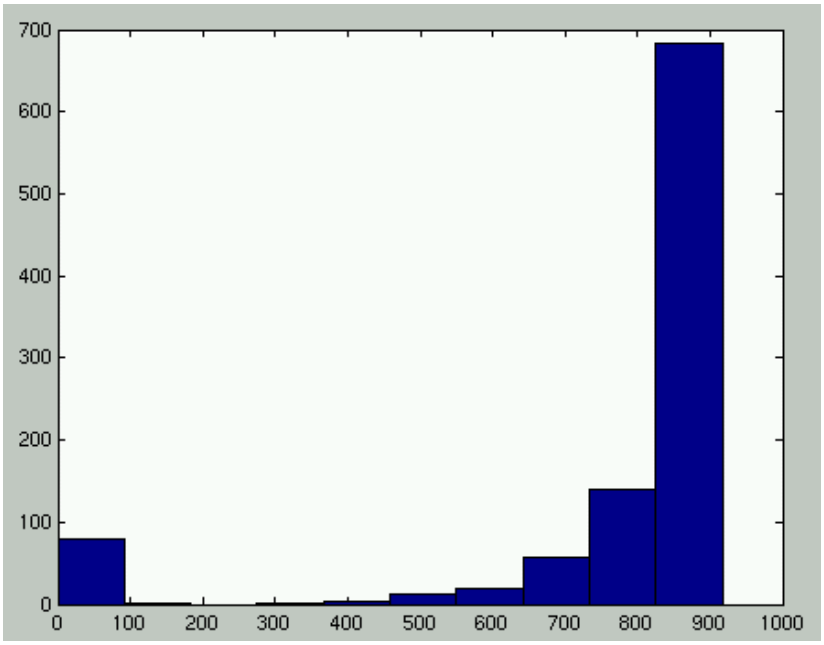


Figure A3.2. *Distribution de #K3(u)*

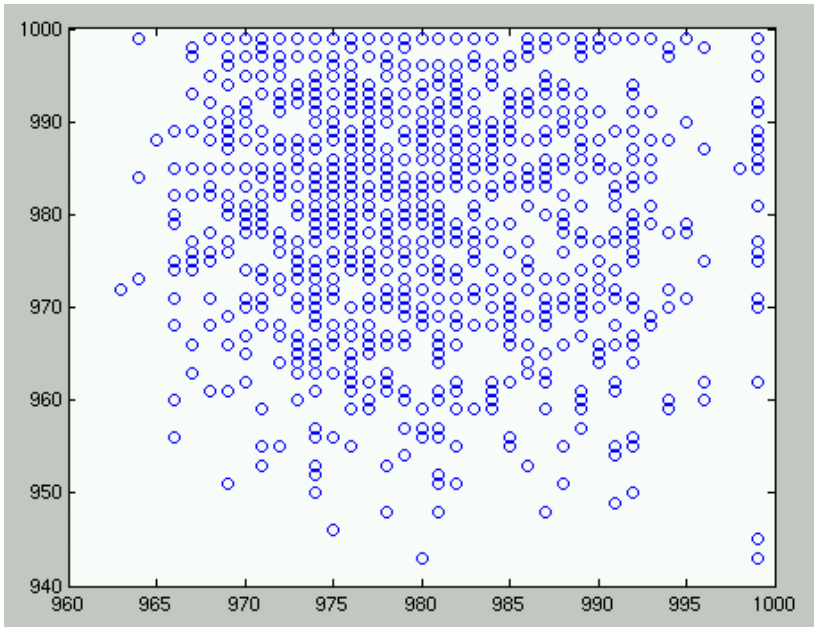


Figure A3.3 *Diagramme de connectivité d'un hypertexte de type RD construit par blocs avec la valeur de coupure 0.01.*

Annexe 4 : Analyse reconstructive

Deux versions sont proposées par Zwick (2002). Dans la version ensembliste-probabiliste, on considère un ensemble R de n objets et une application $\mathbf{I} = (i_m)$ de E sur K^m où K est un ensemble fini de valeurs (par exemple : K = {0, 1} correspond à une indexation classique des objets E). On définit :

$$n_\kappa = \#\mathbf{I}^{-1}(\kappa) \text{ avec } \kappa \in K^m$$

où n_κ est le nombre d'objets de E qui ont le profil κ[59]. On note $p_\kappa = n_\kappa/n$ la fréquence relative associée.

On calcule $U(\mathbf{I}) = -\sum p_\kappa \log p_\kappa$ (U est mis pour uncertainty)

On peut calculer une valeur de U de référence, par exemple en supposant « l'indépendance » statistique des « variables » i_m ce qui conduit à calculer avec des effectifs pondérés. Notation : $U(i_1:i_2: \ldots :i_m)$.

On peut considérer des cas intermédiaires, par exemple $\mathbf{I'}$ comme composition de deux relations : $\mathbf{I'} = \mathbf{I}_1 : \mathbf{I}_2$. Les valeurs théoriques de $\mathbf{I'}$ sont obtenues par projection (totaux marginaux) et reconstruction (en prenant les variables communes à $\mathbf{I}_1$ et $\mathbf{I}_2$ comme références). Et on calcule à nouveau $U(\mathbf{I}_1 : \mathbf{I}_2)$.

On définit la transmission : $Ti = T(\mathbf{I}_1 : \mathbf{I}_2) = U(\mathbf{I}_1 : \mathbf{I}_2) - U(\mathbf{I})$. Il s'agit de l'erreur du modèle $\mathbf{I}_1 : \mathbf{I}_2$ ou des contraintes perdues dans $\mathbf{I}_1 : \mathbf{I}_2$.

Quant à $Tt = T(i_1:i_2: \ldots :i_m) = U(i_1:i_2: \ldots :i_m) - U(\mathbf{I})$, il s'agit de la transmission de référence (par rapport au modèle indépendant).

Tt – Ti représente les contraintes capturées dans $\mathbf{I}_1 : \mathbf{I}_2$. Donc l'information capturée par le modèle que l'on peut normaliser par rapport à Tt : I = 1- Ti/Tt

Pour aider à comprendre les notations, voici un exemple avec 3 variables (i_1,i_2, i_3), K = {0, 1}

[59] Lorsque n = 2, il s'agit des effectifs des cases d'un tableau croisé.

Le dénombrement des profils peut se faire à l'aide d'un tableau.

	$i_1 = 0$		$i_1 = 1$	
	$i_2 = 0$	$i_2 = 1$	$i_2 = 0$	$i_2 = 1$
$i_3 = 0$	$n_{(0,0,0)}$	$n_{(0,1,0)}$	$n_{(1,0,0)}$	$n_{(1,1,0)}$
$i_3 = 1$	$n_{(0,0,1)}$	$n_{(0,1,1)}$	$n_{(1,0,1)}$	$n_{(1,1,1)}$

Dans la version informationnelle, on considère R comme relation (les pondérations ne sont pas prises en compte). L'entropie de Hartley est utilisée $\log_2(\text{Card}(R))$. C, produit cartésien des images des constituants de R, est la référence. A nouveau, on considère les projections, les opérations de reconstruction, puis la chaîne $R \to [Q \otimes S] \to C$. Pour le calcul des transmissions, on considère U(C) et U(R). La différence T(C) = U(C) – U(R) (toujours positive ou nulle) correspond à « l'information » ou la « structure » ou les « contraintes » que contient R par rapport à C. On définit ensuite la transmission entre deux étapes de décomposition : $T(Q,S) = U(Q \otimes S) - U(R)$.

La différence T(C) – T(Q,S) est l'information contenue dans $Q \otimes S$ que l'on peut normaliser par rapport à T(C) : I = 1 – T(Q,S)/T(C).

De fait, ce schéma fait usage des notions classiques d'entropie et entropie conditionnelle. Dans le cas informationnel, on retrouve les notions utilisées par Watanabe (1969).