Prise en compte des besoins des utilisateurs dans la visualisation des connaissances scientifiques : approche dynamique par classification

Pierre Humbert

Prise en compte des besoins des utilisateurs dans la visualisation des connaissances scientifiques de corpus bibliographiques : *Approche dynamique par classification*

Mémoire

Présenté et soutenu le 4 juillet 2006

Pour l'obtention du
Master Recherche
Sciences de l'Homme et de la Société
Sciences de l'Information et de la Communication
Université Nancy 2

Par
Pierre HUMBERT

Sous la direction du Professeur Amos DAVID
Laboratoire Lorrain de Recherche en Informatique et ses Applications

Sous la responsabilité de Claire FRANCOIS et Pascal CUXAC
Service de Recherche et Développement en Ingénierie / INIST

à mon épouse

La rédaction de ce mémoire de master recherche n'aurait pas été possible sans le concours de certaines personnes que je tiens à remercier très sincèrement ici :

M. Amos DAVID, directeur de ce mémoire, que je remercie d'avoir accepté de diriger mes travaux de recherche. Ses conseils avisés m'ont été d'un grand secours ;

Claire FRANÇOIS et Pascal CUXAC à qui j'adresse mes profonds remerciements pour le temps qu'ils ont consacré à encadrer mon travail, à me conseiller, à m'aider à structurer ma pensée et pour avoir relu patiemment chaque partie de ce mémoire ;

Ivana, Véronika, Patricia, Xavier, Dominique, Sylvain et Christophe, du SRDI avec qui il fait bon travailler ;

L'équipe SITE pour m'avoir si chaleureusement accueilli cette année et plus particulièrement Stéphane GORIA pour les pistes de réflexion qu'il m'a inspiré et Sahbi SIDHOM pour l'éclairage qu'il a su m'apporter sur le sujet ;

A mon épouse qui a su me supporter et me motiver durant la rédaction de ce mémoire.

Cette étude a reçu un financement dans le cadre du CPER Ingénierie des Langues du Document et de l'Information Scientifique, Technique et Culturelle (ILD-ISTC) du Pôle de Recherche Scientifique et Technique – Intelligence Logicielle (PRST-IL).

TABLE DES MATIERES

Introduction

C'est en Mésopotamie et en Egypte qu'apparaissent les premiers systèmes d'écriture et avec eux les prémices d'une pensée scientifique. La littérature a permis à cette pensée de traverser les âges, de nourrir la réflexion des philosophes et des scientifiques de l'antiquité à nos jours. De ce fait, depuis ses origines, la science a toujours été considérée comme une « connaissance écrite ». Qui s'intéresse à la science comme objet d'étude doit donc se tourner vers ses écrits.

Au cours du temps des méthodes ont été proposées pour décrire la pensée scientifique et l'activité de la science, méthodes mathématiques et statistiques, fondées sur l'expérience de ceux qui ont longtemps été considéré comme les gardiens du savoir, tour à tour hommes d'église, bibliothécaires, documentalistes... Avec le développement des technologies informatiques, de l'information et de la communication, il est aujourd'hui possible de l'étudier de manière encore plus complète et plus complexe. La puissance des machines permet de rendre compte automatiquement du contenu de milliers de documents ou permet de représenter de vastes réseaux d'hommes et de femmes qui construisent la science d'aujourd'hui et de demain.

Mais si la technologie et les techniques nous permettent de décrire la science de manière de plus en plus poussée, il existe encore bien souvent un décalage entre la représentation statique que l'on fait des connaissances scientifiques et leur nature dynamique et évolutive. En effet, ces connaissances ne sont que le fruit de perpétuels dépassements de la pensée établie, des nouvelles voies permises par les avancées technologiques, etc. Enracinées dans la société des hommes, ces connaissances subissent les phénomènes de mode et les orientations politiques et idéologiques.

Ce mémoire se propose de traiter la représentation de l'évolution des connaissances scientifiques d'un domaine à partir de l'étude automatique des écrits qui le constituent.

Nous menons cette étude dans le cadre d'une collaboration entre le Service de Recherche et Développement en Ingénierie (SRDI) de l'Institut de l'Information Scientifique et Technique (INIST) au CNRS et l'équipe SITE du LORIA[1]. Le SRDI a pour mission d'assurer une activité de recherche dans les domaines des mathématiques appliquées à l'analyse de l'information, du traitement informatique du langage naturel d'importants corpus et de l'application des techniques symboliques et numériques de l'intelligence artificielle, dans le but de l'analyse de l'information.

La problématique de l'équipe SITE, quant à elle, est la modélisation et le développement de systèmes d'informations stratégiques dans le cadre de l'Intelligence Economique (IE). La méthode qu'elle propose est de mettre en évidence le rôle des acteurs dans un contexte socio-économique, de l'information qui constitue un outil de pilotage et de l'exploitation de cet outil pour gérer et définir les stratégies de développement des organismes socio-économiques.

Ce travail s'inscrit dans le cadre du projet « veille automatisée : étude d'un algorithme incrémental de classification automatique », partie du pôle Ingénierie des Langues du Document et de l'Information Scientifique, Technique et Culturelle (ILD-ISTC) du Pôle de Recherche Scientifique et Technique (PRST) Intelligence Logicielle (Contrat plan-état-région).

[1] Laboratoire Lorrain de Recherche en Informatique et Applications

Gershon et Page écrivaient dans un article de 2001 que « *les développeurs de logiciels de visualisation, les designers d'interfaces ont besoin de prendre en compte ce que nous savons sur la manière qu'ont les humains à comprendre, à interagir avec l'information et sur leur système inné de perception. Ils ont aussi besoin d'apprendre comment créer des interfaces utilisateur, des outils de navigation et des méthodes de recherche d'information flexibles et appropriées à chaque type d'utilisateurs, applications et tâches.* » C'est dans cet optique que s'inscrit notre travail : comment rendre la visualisation d'informations utile à l'utilisateur ? Quelles sont les fonctionnalités qui apporteront une valeur ajoutée aux données visualisées ? Telles sont les questions auxquelles nous nous intéressons afin de définir les fonctionnalités de la visualisation des résultats d'une méthode d'analyse de données : un algorithme de classification incrémental.

L'algorithme a comme particularité d'étudier l'évolution des données à l'aide de classifications, c'est-à-dire des regroupements de données proches appelées des classes, effectuées à différents instants. A chacun de ses instants, un journal est créé afin de décrire l'état de chacune des classifications, celui-ci est appelé journal de classes. L'algorithme envisage de tracer l'évolution de thématiques de recherche (les classes) d'un domaine à partir des publications scientifiques de ce dernier. Quelles informations doit-on fournir à l'utilisateur pour étudier l'évolution de thématiques de recherche ? Comment les lui présenter graphiquement de manière efficace ? Ainsi se formulent les questions qui ont animées notre démarche.

Pour apporter des éléments de réponse à cette problématique, nous partons de l'hypothèse suivante : il n'existe pas de représentation de l'information qui réponde à tous les besoins des utilisateurs.

Pour tenter de démontrer cette hypothèse, nous nous sommes donnés comme principe de travail que la prise en compte des tâches et des besoins informationnels des utilisateurs permet à la représentation de leur apporter des résultats utiles. C'est pourquoi, à partir d'un modèle utilisateur nous nous sommes attachés à identifier les besoins auxquels répondent les résultats de l'algorithme et à proposer des informations complémentaires pour les besoins auxquels ils ne répondent pas encore. Pour cela, nous analyserons la structure du journal de classes puis nous tenterons de définir les fonctionnalités de visualisation de ses données à partir de cette analyse.

Pour réaliser cette recherche, il est essentiel dans un premier temps de réaliser un état de l'art sur les trois aspects liés à la problématique :

1. Le premier chapitre place le sujet au sein de la veille scientifique et technologique, contexte de l'utilisateur, élément de notre hypothèse de travail. Ce chapitre sera l'occasion d'entrevoir quelques fonctions, tâches et besoins informationnels d'utilisateurs.
2. Dans un second chapitre nous faisons le point sur les méthodes d'analyse de l'information dans lesquelles l'algorithme de classification incrémental s'inscrit et dont la compréhension nous permettra de saisir à la fois le processus et les résultats.
3. Un troisième et dernier volet de cet état de l'art s'articule autour de la notion de visualisation de l'information afin d'identifier des éléments de réponse au problème de la visualisation des résultats de l'algorithme.

Puis dans une seconde partie, nous voulons connaître quelles sont les fonctionnalités de visualisation de l'évolution de données dans les outils actuels de recherche d'information, d'analyse de données, de veille, etc. Pour cela, nous menons une série d'observations sur ces

outils en essayant d'analyser l'utilisabilité de ces fonctionnalités, puis d'identifier également des éléments de réflexion.

Enfin, au cours d'une troisième et dernière partie, nous tentons de modéliser les utilisateurs potentiels du système en y dégageant leurs tâches et leurs besoins informationnels. Nous nous intéressons ensuite aux données du journal de classe : Permettent-elles de répondre à ces besoins ? Comment peuvent-elles y répondre ? Puis nous cherchons à définir des modes de représentation de ces données au travers la définition d'indicateurs et de fonctionnalités de visualisation adaptées.

Chapitre 1

Veille scientifique et technologique : un cadre d'étude

Nous avons choisi dans notre démarche de situer nos travaux sur la visualisation de l'information dans le cadre d'une démarche de veille scientifique et technologique (VST). De cette façon nous présentons d'abord un cadre applicatif et fonctionnel qui nous permettra ensuite de mieux déterminer les facteurs d'utilisabilité pour les utilisateurs veilleurs. Dans cette partie, nous définissons le processus de veille, lui-même inclus dans un processus plus global d'intelligence économique.

1.1 La veille scientifique et technique : du monde l'information à l'organisation

Face au flot d'informations sans cesse grandissant, les organisations doivent parvenir à une certaine maîtrise de ces informations. Cette maîtrise tient en l'application de processus et de méthodes regroupés aujourd'hui sous le terme générique d'intelligence économique (IE). Veiller, c'est-à-dire être à l'écoute de son environnement, fait partie du processus d'IE, car il s'agit du processus d'alimentation de la réflexion stratégique menée par les organismes.

Martinet et Ribault (1988) définissent la démarche de veille comme une « *attitude organisée d'écoute des signaux provenant de l'environnement de l'entreprise susceptible de mettre en cause ses options stratégiques* ». L'AFNOR la définit comme une « *activité continue et en grande partie itérative visant à une surveillance active de l'environnement technologique, commerciale, etc., pour anticiper les évolutions* » (Norme XP X50-053).

La veille scientifique et technologique quant à elle est définie par Jakobiak (1995) comme « *l'observation et l'analyse de l'environnement scientifique et technologique suivie de la diffusion bien ciblée aux responsables des informations sélectionnées et traitées utiles à la prise de décisions stratégiques* ».

Selon Bouaka (2004), le point de départ de toute démarche de veille est la traduction du problème décisionnel en problème de recherche d'information. Dans cette transition vers la recherche d'information, il n'est pas évident que la requête effectuée par un utilisateur traduise précisément son besoin [David, 2001], la difficulté s'accentue encore lorsque cette requête est effectuée par une autre personne que le décideur, c'est-à-dire un veilleur. Des méthodes pouvant faciliter cela, ont été conçues, comme le Modèle pour l'Explicitation du Problème Décisionnel (MEPD) de Bouaka afin d'optimiser la communication entre le décideur et le veilleur [Bouaka, 2004].

Le veilleur, s'il est un individu distinct du décideur, est en effet l'interface humaine entre la tête de l'organisation et le monde de l'information. Il est en charge de répondre aux attentes et aux besoins informationnels du décideur. Sa méthode consiste en :

- La sélection de sources pertinentes et la validation de celles-ci
- La collecte de l'information
- L'analyse de l'information afin de leur apporter une valeur ajoutée
- La présentation des résultats de sa démarche

La veille, est une méthode raisonnée de collecte et de diffusion d'informations pertinentes, une méthode qui implique deux acteurs : le veilleur, celui qui recherche et analyse l'information, et le décideur, utilisateur final des informations sortantes du dispositif.

L'évolution de l'information dont il est question dans notre problématique est un point important dans la surveillance que pratique le veilleur. Ce point nous amène à nous interroger sur le rôle de la VST dans l'organisation.

1.1.1 Veille et suivi de l'évolution de l'information

Le processus de veille donne une grande importance au suivi des informations dans le temps. En effet, dans la logique de surveillance, il est essentiel de connaître les mouvements de ses concurrents (*e.g.* surveiller leur Communication au travers leurs manifestations sur Internet à l'aide d'outils de surveillance de sites web), la sortie de nouvelles technologies sur les marchés (e.g. surveiller les dépôts de brevets, les travaux scientifiques en cours, les projets à venir, etc.), d'anticiper un événement en détectant une tendance, ce que l'on appelle aussi un signal faible. Jakobiak (1995, p. 110) écrit que « *le suivi systématique de l'évolution technologique des domaines critiques de l'entreprise permet, c'est un des objectifs de la veille technologique, de saisir les opportunités de développement sans perdre de temps* ».

Ansoff en 1975, défini le signal faible comme un type d'information de nature prospective, plus qualitative que quantitative, de durée de vie limitée, rapidement obsolète, de nature imprécise et ambiguë, fragmentaire souvent peu signifiante individuellement (il faut souvent croiser ce type d'information avec une autre pour la rendre utile au décideur) [Vidal, 2004].

La détection d'une émergence, ou d'une tendance peut être interprété comme une forme de signal faible. L'émergence désigne l'apparition d'un phénomène portée à l'interprétation du veilleur par les données et les tendances recouvrent les phénomènes existant qui tendent à se développer davantage,

1.1.2 Rôle de la VST dans l'organisation

Pour l'organisation, la VST répond d'abord à des besoins en terme de gestion des connaissances visées ou établies, puis en terme de processus d'innovation.

Veille scientifique et technologique : un outil de gestion

Du point de vue décisionnel, mener une VST est aussi important que tout autre forme de surveillance (commerciale, concurrentielle, etc.) pour différentes raisons :

- Les décisions peuvent porter sur la politique de recherche et de développement, pour l'élaboration de programme de recherche, pour une programmation des actions de développement des produits, des procédés, etc.
- La surveillance des avancées technologiques et scientifiques joue un rôle important pour réaliser des choix stratégiques en matière de transferts de technologies, c'est-à-dire la procédure par laquelle un acteur (organisme) source ayant créé ou amélioré une part non exploitée de connaissances ou de savoir-faire fait passer cette part à un autre organisme qui compte l'appliquer et l'exploiter commercialement[2]. Ces choix peuvent être liés à la demande pour réaliser une acquisition, ou liés à l'offre pour se positionner en fonction des technologies concurrentes.

[2]Habituellement, le savoir-faire ou les connaissances ne sont pas incorporés, c'est-à-dire qu'ils ne sont pas transférés sous forme de produits ou de pièces d'équipement. Le transfert peut être effectué dans le cadre d'une entente juridique officielle ou encore de façon informelle dans des documents, à l'occasion de conférences, pas des contacts personnels, etc.
http://www.geoconnexions.org/ICDG.cfm/fuseaction/policySupporting.seeFile/id/95/gcs.cfm (consultée le 12/04/2006)

L'analyse de l'IST et notamment la cartographie de la science, dans le contexte de management de la politique scientifique, a un apport proche de la démarche de veille car elle permet, d'après Noyons (2004).

- d'avoir une vue d'ensemble compréhensible
- de faire référence à la situation actuelle
- d'avoir des points de repère pour interpréter les résultats
- de percevoir les dynamiques des structures
- d'identifier les acteurs de la structure
- d'identifier les acteurs de la dynamique

L'observation de l'information scientifique et technique est, selon Noyons, essentielle pour la gestion d'une politique scientifique et technologique.

Veille scientifique et technologique : un outil pour l'innovation

La veille scientifique et technologique, comme la définissait Jakobiak (1995) est une démarche s'intéressant plus particulièrement aux publications et brevets comme source d'informations scientifiques, techniques et technologiques[3].

L'enjeu d'une veille scientifique et technologique est donc de détecter les tendances, de surveiller les découvertes, les innovations afin de positionner l'organisation par rapport à celles-ci. L'organisation peut alors se positionner soit en concurrence, si elle se situe sur le même marché, soit éventuellement en collaboration, si son marché est différent.

Mais si l'on prend le cas d'un service de recherche et développement (R&D), composé de chercheurs, d'ingénieurs et de concepteurs, l'information scientifique et technique, dont ils sont grands consommateurs, est d'importance pour les actions et les choix qu'il a à réaliser. Elle constitue d'une part une matière pour travailler, puisqu'un processus d'innovation se nourrit des travaux déjà menés, d'autre part elle permet à la R&D de déterminer la faisabilité de leurs projets : sera-t-il possible de réaliser ce projet dans des coûts financiers et temporels raisonnables et dans des conditions opératoires acceptables. Enfin, l'IST est prépondérante dans une démarche d'innovation du point de vue juridique à savoir : qu'en est il de la propriété indutrielle ? Avons-nous le champ libre ? Si ce point est négligé, les conséquences s'avéreraient assez sérieuses pour l'entreprise et beaucoup d'énergie aura été dépensé en vain [Jakobiak, 1995].

1.2 Rôle de la VST dans un processus d'intelligence économique

Alors que nous avons vu ce que désignait le terme de veille et plus particulièrement celui de veille scientifique et technologique, nous avions annoncé ce processus de surveillance comme faisant partie d'un processus plus globale appelé intelligence économique.

Qu'est ce que l'intelligence économique ? Notion très à la mode dans les milieux des entreprises et des universités, de nombreuses définitions de l'intelligence économique ont été données depuis de quelques années. Goria (2006) montre que ces définitions de l'intelligence économique sont filles de différentes influences dont il distingue notamment 3 principales : (1) l'influence militaire, (2) l'influence économique et (3) l'influence éthique et numérique.

[3] Jakobiak ([Jakobiak, 95], p.9) distingue le terme d'information technique de d'information technologique sans apporter de définition claire pour l'un et en définissant l'autre comme indiqué pour l'utilisation, la mise en place et la réalisation de matériaux, appareillages, installations à caractère industrielle.

Cette approche présente l'IE sous trois principales perceptions dont nous nous sommes inspirés pour la définir.

1.2.1 L'influence militaire

Le sens du terme *intelligence* dans l'I.E. peut renvoyer au sens du terme anglo-saxon tel qu'on le retrouve dans Central Intelligence Agency (CIA), l'agence de renseignements extérieurs des U.S.A., c'est-à-dire comme l'art de collecter des renseignements, des informations au service d'une stratégie ou d'une volonté d'écoute de l'environnement. Le terme français quant à lui le défini surtout par la faculté de comprendre, ce qui peut occasionner des confusion liées à l'expression *intelligence économique* [Favier, 1998]. La définition du terme anglais pris dans le Robert & Collins (1998) le définit effectivement comme *renseignements* lorsqu'il est associé au terme militaire, comme *naval intelligence*, *military intelligence*, *intelligence Corps*, *intelligence officer*, etc. C'est pourquoi, par ses origines, l'I.E. a souvent été assimilée à l'espionnage [Bulinge, 2002].

Au travers le concept d'intelligence pris dans ce contexte, nous retrouvons la notion de cycle du renseignement, un processus itératif fini décrivant les différentes étapes du renseignement telles que représentées par la figure 1. Alain Juillet (2004) fait d'ailleurs appel à cette notion lorsqu'il définit l'intelligence économique comme « *un concept global qui ajoute à la pratique du cycle du renseignement, son utilisation dans l'aide à la décision et la mise en œuvre de certains types d'actions.* ».

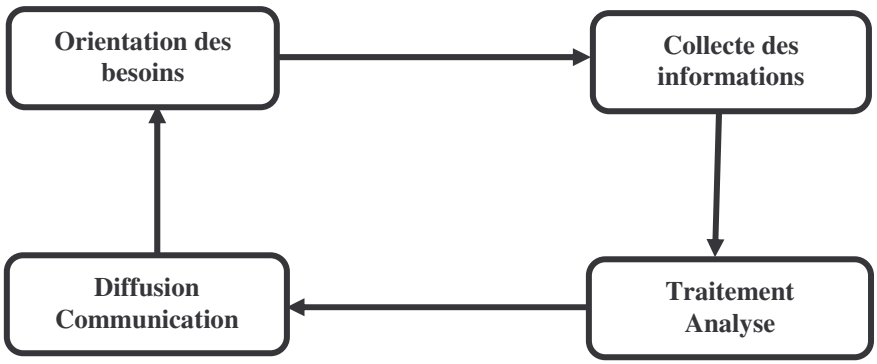


Figure 1 : Le cycle du renseignement

1.2.2 L'influence économique

L'approche économique du concept d'I.E. laisse poindre différents aspects qui viennent compléter les définitions que nous avons déjà citées précédemment.

D'après Goria (2006), contrairement à l'approche militaire, la conception des rapports entre organisations n'est pas traitée de façon aussi radicale. En effet, ils sont plus conçus comme des relations de concurrence plutôt que des relations d'ennemis, les concurrents sont des compétiteurs par rapport auxquels on doit essayer de se positionner de manière la plus stratégique possible. La stratégie de positionnement d'une organisation la pousse parfois à devoir saisir des opportunités selon lesquelles elle est parfois amenée à faire alliance avec ses concurrents d'autrefois et inversement. Ces alliances stratégiques montrent qu'il existe donc une dimension collective à l'intelligence économique, dans laquelle il est important pour les organisations de pouvoir travailler avec des partenaires lorsque l'opportunité se présente.

Cette dimension collective n'est d'ailleurs pas uniquement externe, tournée vers les partenaires, elle est aussi interne. En effet, l'IE est souvent considérée comme un état d'esprit

qu'il faut diffuser dans l'entreprise pour favoriser l'échange d'information en interne [David, 06].

Les définitions « économiques » de l'intelligence économique font souvent appel aux travaux de l'économiste Michael Porter (1980), qui définissent les points pouvant influencer les entreprises dans leur action. Les cinq forces de Porter permettent de décrire l'environnement d'une entreprise et donc les points sur lesquels elle doit se montrer à l'écoute. Selon Martinet et Ribault (1989), *« l'entreprise aura besoin d'identifier et de connaître les éléments clés de ces cinq paramètres qui conditionnent son existence »*. Ils définissent, par rapport à ces cinq forces, quatre types de veille associés (figure 2) :

- **Veille technologique** : les acquis scientifiques et techniques, les produits, les procédés de fabrication, les matériaux, les systèmes d'information…
- **Veille concurrentielle** : s'intéresse aux concurrents potentiels et actuels
- **Veille commerciale** : les clients et les fournisseurs…
- **Veille environnementale** : plus floue, elle concerne le reste de l'environnement de l'entreprise, les événements auxquels l'entreprise doit pouvoir réagir…

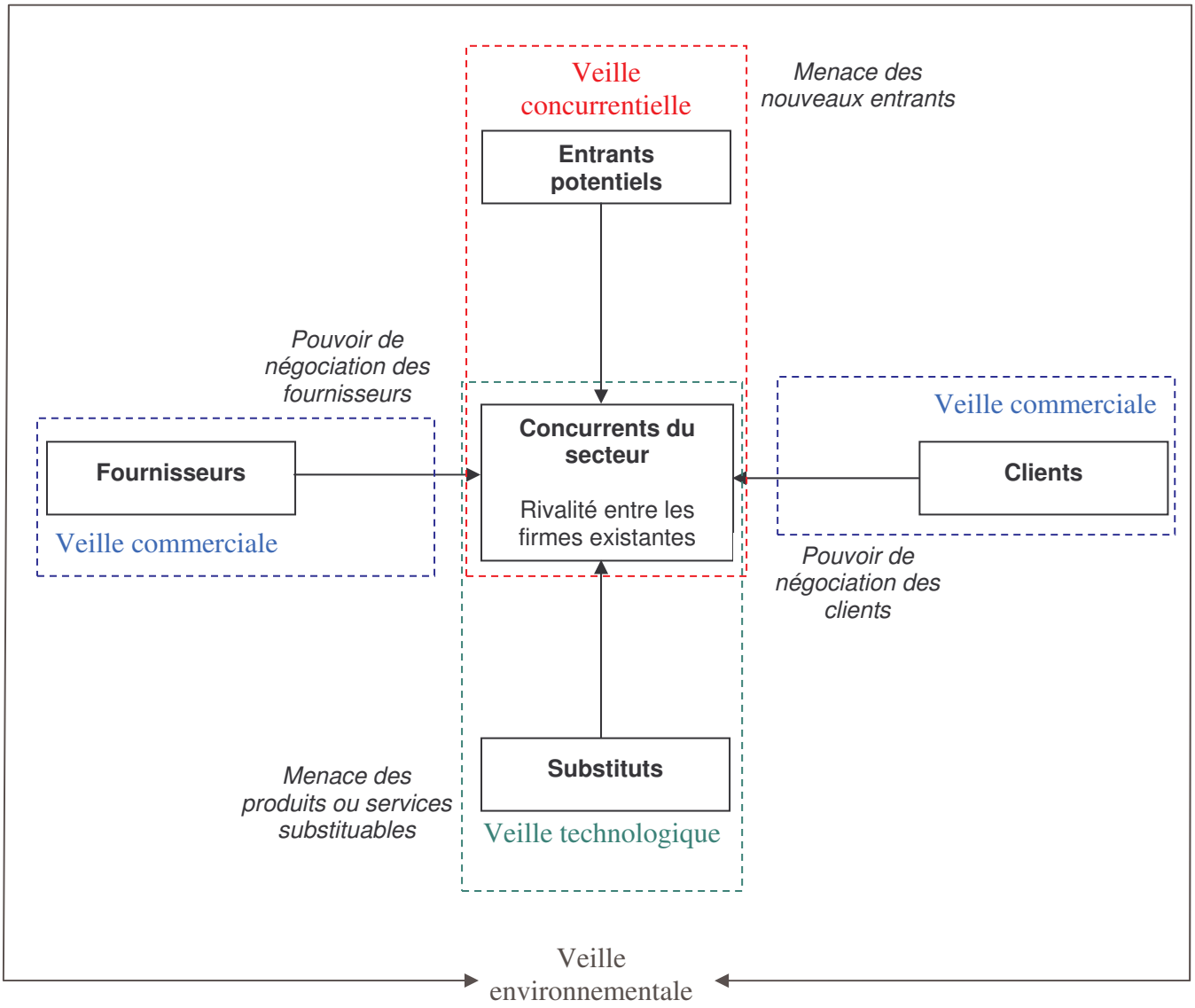


Figure 2 : Les forces de Porter [Martinet et Ribault, 1989]

Dans les aspects économiques de la définition de l'I.E., Goria (2006) mentionne aussi les aspects liés aux techniques d'influences et de lobbying. Ces méthodes utilisées par certaines entreprises visent à faire valoir leurs idées sur les marchés et dans divers sphères d'influence (politiques, institutionnelles, etc.). Elles contribuent plus ou moins directement à défendre leurs intérêts, leur positionnement sur les marchés, etc. Dans cet optique, l'I.E. se définit donc aussi comme un ensemble d'activités de gestion et de création de réseaux relationnels dans le but non seulement d'élargir son champ d'acquisition d'information mais aussi de pouvoir agir sur les différents « *leviers des pouvoirs en place*. »

1.2.3 Influence éthique

Goria (2006) identifie une dimension éthique marquant de plus en plus l'intelligence économique. Cette dimension vient en réaction à une connotation négative liée notamment à ces origines militaires ou encore à ses applications concurrentielles les plus offensives, qui font associer la démarche d'I.E. à l'espionnage économique. Déjà en 1994, le souci d'y apporter une dimension déontologique est présent : comme l'indique le rapport Martre, les actions liées à l'I.E. sont décrites par l'auteur comme « *menées légalement avec toutes les garanties de protection nécessaires à la préservation du patrimoine de l'entreprise…* » [Martinet et Ribault, 94]. Depuis, de nombreux codes et guides de déontologie et de bonnes pratiques ont été écrits comme celui réalisé par la SCIP[4], *Competitive Intelligence Ethics : Navigating the Grey Zone*, mais aussi le *Code de bonne conduite des professionnels de la Veille et de l'Intelligence Economique*, disponible sur le site de la société Cybion[5], dont les signataires s'engagent à respecter les valeurs dans leurs activités. Un autre code de déontologie est celui réalisé en partenariat avec le CIGREF et le CEA-CED[6], intitulé *Déontologie des usages des systèmes d'informations*, publié en janvier 2006.

1.3 Les sources d'informations de la VST

Dans la section précédente nous avons voulu définir l'IE suivant différentes approches afin d'essayer de percevoir ses différentes facettes. Au travers ses définitions, l'enjeu principal reste sans aucun doute la maîtrise des informations de l'environnement d'une organisation. Nous verrons dans cette section ce que nous entendons exactement par *information* et surtout où se trouve cette information, en particulier dans le cadre de la VST, cadre de notre étude.

L'identification des sources d'informations est un élément essentiel dans le processus d'intelligence économique, qui permet de rationaliser la recherche d'informations en ciblant les sources pertinentes. Ces sources sont multiples et méritent d'être décrites afin de situer le cadre applicatif de notre travail dans la démarche d'I.E.

1.3.1 Qu'est ce que l'information ?

D'après Le Coadic (2007), l'information est « *une connaissance inscrite sous forme écrite, orale ou audiovisuelle sur un "support spatio-temporel"(imprimé, signal électrique, onde sonore, etc.). L'information comporte un élément de sens. C'est une signification transmise à un être conscient par le moyen d'un message inscrit sur un support* ».

[4] Society of Competitive Intelligence Professionals. http://www.scip.org/

[5] http://www.veille.net/code_veille_net.pdf (Consulté le 09/03/2006)

[6] CEA-CED : Cercle d'Ethique des Affaires – Cercle Européen des Déontologues. Le CED est une structure du CEA qui est une Organisation Non Gouvernementale, créé en 1993 et regroupant des dirigeants d'entreprises. http://www.cercle-ethique.net/

Nous pouvons aussi définir l'information au travers les concepts de donnée et de connaissance, ce qui nous fournit l'occasion de définir aussi ce qui les distingue les uns des autres, puisque que nous les emploierons de façon régulière dans notre travail.

Selon Goria (2006), il existe une hiérarchie entre ces trois notions. Nous pouvons définir les données comme des éléments qui se présentent à nous hors de tout contexte. Ce sont des « *nombres, mots, événements existants en dehors d'un cadre conceptuel de référence ; en conséquence, et en absence de contexte, les données prises individuellement n'ont pas de signification.* » ([CRL, 2003] p 22.)

D'après la théorie de l'information développée par Shannon et Weaver (1975), les données sont susceptibles de devenir des informations lorsqu'elles sont perçues et interprétées par l'individu si elles apportent des éléments nouveaux à cet individu. Si l'on considère la connaissance comme un processus et le savoir le résultat de ce processus, l'information est donc une donnée qui viendrait en quelque sorte combler les lacunes dans le savoir que possède déjà l'individu.

De ce fait, la connaissance constitue un arrière plan permettant à l'individu de juger si une donnée perçue est porteuse ou non d'un élément nouveau. Elle n'est pas réductible à une accumulation d'informations, mais plutôt à une appropriation d'éléments informationnels, une représentation du réel au travers les relations et les interactions qui caractérisent ses éléments. Les théories constructivistes considèrent la connaissance comme un processus toujours en construction, ouvert aux principes et concepts d'un domaine ou d'une activité (connaissance déclarative), à l'ensemble des actions et des processus mis en œuvre par l'individu (connaissance procédurale) mais aussi aux expériences acquises ou développées par l'individu (connaissance expérientielle)[7].

Pour en revenir plus exactement au concept d'information, dans l'univers des Sciences de l'Information et de la Communication, on distingue 3 types d'information : l'information Blanche, l'information Grise et l'information Noire dont seulement les deux premières concernent la veille. Le tableau ci-dessous, extrait de [Bulinge, 2002], synthétise les différentes caractéristiques de cette typologie.

INFORMATION	***BLANCHE***	***GRISE***	***NOIRE***
Type	Scientifique, technologique, commerciale, juridique, financière, stratégique, personnelle		
Niveau	Tactique, opérationnel, stratégique		
Domaine opératoire	Documentaire, de situation, d'alerte		
Intérêt	Fatal, utile, pertinente	Pertinente, critique	Critique
Accès	Public	Restreint	Strictement limité
Classification	Non protégé	Diffusion restreinte	Confidentiel – Secret
Disponibilité	80%	15%	5%
Acquisition - Exploitation	Légale sous réserve de respecter les droits de propriété.	Domaine juridique non clairement défini. Risques d'ordre jurisprudentiel.	Illégal. L'acquisition relève de l'espionnage. Risques très élevés.

[7] http://www.biblioconcept.com/wiki/index.php?wiki=Connaissance (consultée le 29/03/2006)

Forme	Formelle (texte) ou informelle (conversation, rumeur)		
Sources	Ouvertes	Autorisées - Fermées	Clandestines
Coût	Faible	Faible	Elevé
Rentabilité	Elevée	Très élevée	Faible

Tableau 1 : Tableau synoptique des types d'information [Bulinge, 2002]

1.3.2 Des sources nombreuses et diverses

Nous connaissons désormais différents types d'information et nous savons quel rôle et quelle importance celle-ci a dans le fonctionnement d'une organisation. Cependant, le point de départ de la collecte d'information est la sélection des sources. En effet, ce choix est essentiel pour une prise de décision pertinente et efficace. Pour cela, nous avons voulu évoquer le problème, notamment dans le domaine de la veille scientifique et technologique.

Le web

La première idée venant à l'esprit concernant les sources d'information, est sans doute l'immense réseau de pages web réalisées par des entreprises, des organismes, des passionnés, des médias, etc. En effet, cet univers regorge d'informations et tend à croître de façon exponentielle, ce qui constitue un problème comme nous l'avions déjà évoqué.

Le web met de gigantesque quantité d'informations à disposition, mais celles-ci sont éparpillées, dispersées et veiller sur un sujet particulier à partir du web, contraint le veilleur à croiser des informations de toutes sortes et de toutes origines. On imagine aisément le coût temps / effort que cela peut représenter. Cependant, il existe aujourd'hui des outils offrant un gain de temps non négligeable, comme les « agents intelligents », par exemple. Ces informations sont aussi extrêmement diverses allant de l'information publique et facilement accessible, dite blanche, à l'information protégée, dite noire, selon la typologie vue précédemment.

Les périodiques

Les revues, journaux, publications périodiques diverses constituent une source indispensable en information scientifique et technique. Leur consultation dès la parution permet généralement de capter l'information plus rapidement que sur les bases de données [Jakobiak, 1995], ce qui est d'autant plus vrai désormais, avec le développement de l'Open Access sur Internet. En effet, les archives ouvertes, sont une alternative actuelle au circuit traditionnel de publication scientifique, qui rencontre un certain engouement au sein de la communauté scientifique. Elles se définissent comme *« la mise à disposition gratuite sur l'Internet public, permettant à tout utilisateur de lire, télécharger, copier, distribuer, imprimer, interroger ou accéder par lien au texte intégral de ces articles, les parcourir pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin licite, sans limitations financières, juridiques ou techniques autres que celles indissociables de l'accès et de l'utilisation de l'Internet. La seule contrainte à la reproduction et à la distribution et le seul rôle du droit d'auteur dans ce domaine devraient être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités »* (BOAI, 2002)[8].

[8] Définition donnée par l'Initiative de Budapest pour l'Accès Ouvert (BOAI) en 2002.

Les ouvrages, encyclopédies, thèses

Les livres constituent une source non négligeable pour les études approfondies : ils ne donnent pas les éléments d'information les plus récents sur un sujet, en raison du délai de rédaction et publication, mais fournissent les analyses et les synthèses nécessaires à la connaissance du domaine présenté. Les ouvrages sont utilisés dans la phase recherche, au moment, par exemple, où l'on aborde un sujet nouveau pour lequel il y a lieu d'acquérir des connaissances de base. Ils constituent, avec les encyclopédies, le fonds documentaire indispensable de l'utilisateur de la documentation scientifique et technique.

Les thèses présentent des caractéristiques similaires aux ouvrages mais sont des documents apportant des aspects théoriques et scientifiques, utiles également lorsqu'une connaissance approfondie d'un nouveau sujet s'avère nécessaire. Celles-ci apportent aussi des cas d'application intéressants.

Les brevets

Pour qui veut connaître l'information scientifique et technique, il est indispensable de savoir que la documentation contenue dans les brevets constitue une source extrêmement riche en renseignements. Leur contenu informatif peut être exploité pour la surveillance globale des différents acteurs et des techniques, mais également pour la surveillance de l'activité de la concurrence.

Elle permet aussi de repérer les secteurs protégés où les contraintes juridiques seront les plus fortes.

Saisir les opportunités : l'information informelle

Dans le contexte d'entreprise, surveiller son environnement ne peut se cantonner à une recherche d'information ponctuelle ou récurrente sur les autoroutes de l'information. Dans ce contexte, la veille tend à devenir autant un état d'esprit qu'une pratique, comme nous l'avions défini jusqu'à présent, aux aguets de toutes les informations pouvant alimenter une réflexion stratégique pour l'entreprise [Humbert, 05]. Les sources d'informations peuvent être alors les salons, véritables rendez-vous de professionnels qui échangent, communiquent, exposent, etc. les conférences, congrès et autres forums, les clients avec qui être à l'écoute ne leur sert pas exclusivement mais sert aussi l'entreprise, etc. Nous retrouvons ici toutes les dimensions de la veille de Porter 1980 et de Martinet et Ribeault (1989) telles que nous les avons présenté dans la figure 2. Là aussi on retrouve des éléments de la typologie de l'information, car l'information peut être facilement accessible mais non exploitée (mener un dialogue avec ses clients, par exemple), ce qui revient à de l'information de type blanche, mais elle peut être aussi une confidence, une révélation ou un lapsus sur une information beaucoup plus difficile à accéder car protégée, essentielle, etc., il s'agit là d'information noire.

Les bases de données

Désormais accessibles la plupart du temps par le web, et de façon plus ou moins restreinte, les bases de données constituent de véritables mines d'informations pour le veilleur. Appelées aussi parfois *datawarehouse*, elles regroupent de nombreuses données ou documents spécialisés. Nous nous sommes intéressé aux bases de données utiles à la veille scientifique et technologique, comme les bases bibliographiques ou les bases de brevets.

Les bases de données bibliographiques sont sources de documents primaires (documents) ou de sources de documents secondaires (notices de documents), c'est le cas par exemple des bases alimentées par l'INIST comme PASCAL, base de données bibliographiques, multidisciplinaire et multilingue qui couvre l'essentiel de la littérature mondiale en Sciences, Technologie et Médecine. Cette base contient plus de 15,9 millions de

références bibliographiques depuis 1973. Elle est reconnue comme base de première approche pour toute recherche d'informations scientifiques et techniques ; FRANCIS, une base de données bibliographiques multidisciplinaire et multilingue en Sciences Humaines et Sociales qui contient plus de 2,7 millions de références bibliographiques depuis 1972.

L'ISI, Institut de l'Information Scientifique du groupe Thomson, aux Etats-Unis, met à disposition ce qui est appelé l'« ISI Web of Knowledge »[9] regroupant des bases de données comme le Web of Science, structurant des milliers de notices bibliographiques.

Les archives ouvertes qui se développent fortement sont elles aussi des sources non négligeables d'information scientifique, aisées d'utilisation.

Les bases de données factuelles quant à elle concernent particulièrement des domaines comme la chimie, la médecine, la biologie (génomique) mais aussi parfois l'économie. Elles regroupent des données observables et attestées concernant les éléments du domaine.

Quant aux brevets, le groupe Thomson se place aussi sur ce créneau avec la base de données Aureka[10] en accès restreint. En Europe, c'est l'Office Européen des Brevets qui met librement à disposition ces informations au travers le portail Espacenet[11].

La veille, en tant que processus de recherche d'information et de surveillance de l'environnement est un élément essentiel de l'élaboration des stratégies d'entreprise. Nous avons pu voir son cadre d'application en tant que pratique d'intelligence économique, entrevoir les enjeux et les difficultés qu'elle rencontre. Le veilleur, chargé de mener à bien cette tâche, est amené non seulement, comme nous l'avons vu, à capter les informations pertinentes mais aussi à les analyser. Nous pensons que cette analyse se situe à deux niveaux : lors de sa recherche d'information dans une phase que l'on peut appeler exploratoire et lorsqu'il a à portée de mains suffisamment d'éléments pour les soumettre au décideur afin de leur apporter une valeur ajoutée.

Comment s'y prend-t-il devant tant d'information ? Quelles sont les méthodes qu'il peut utiliser ? Pour ce faire il a à sa disposition des techniques automatiques ou semi-automatiques d'analyse de l'information. Nous allons essayer de rendre compte de ces techniques avec pour double objectif de les présenter tour à tour au lecteur ainsi que de les introduire dans notre cadre de réflexion sur la visualisation de l'information. La visualisation de l'information est en effet une démarche d'analyse de l'information mais elle nécessite au préalable une phase d'analyse qui va construire les éléments à visualiser (classes, termes, etc.) à partir des données : ce qui est visualisé ce situe en sortie du processus d'analyse.

[9] http://www.thomsonscientific.com/frwok/wokdetails/ (consultée le 10/04/2006)
[10] http://scientific.thomson.com/products/aureka/ (consultée le 10/04/2006)
[11] http://www.espacenet.com/ (consultée le 10/04/2006)

Chapitre 2

L'analyse de l'information

Rechercher des informations n'est pas une fin en soit, il est nécessaire d'apporter à ces données une valeur utile à celui qui les recherche ou au demandeur d'informations (le décideur, par exemple). Il existe pour cela des méthodes automatiques ou semi-automatiques d'analyse de l'information, issues des mathématiques, qui permettent de faire émerger d'une masse de données volumineuse, des informations cachées mais utiles pour aider le décideur dans sa prise de décision finale ou le veilleur dans son activité de recherche et de traitement d'information.

Nous procéderons dans cette partie à un exposé des différentes techniques d'analyse de l'information, de l'analyse de données à la scientométrie en passant par la fouille de données et de textes. Nous les présentons dans cet ordre afin d'établir une progression reposant sur la figure suivante :

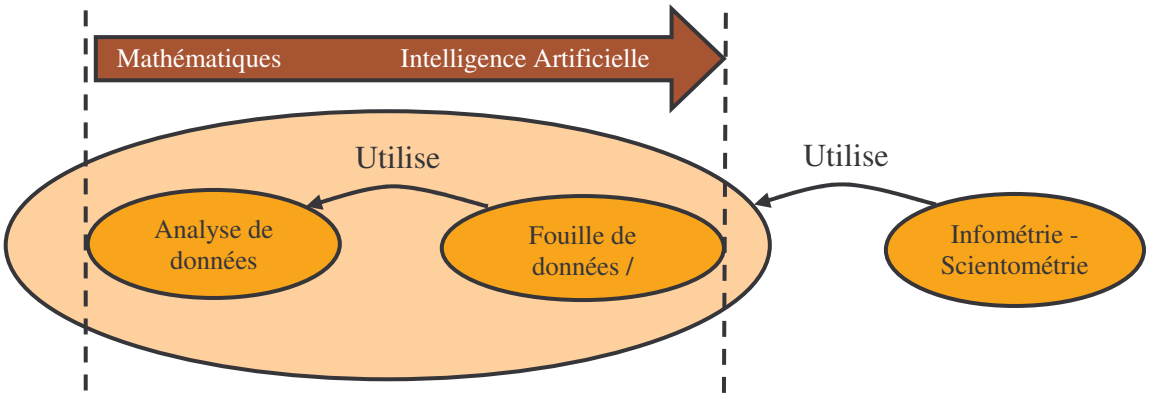


Figure 3 : Démarche de l'étude suivie dans ce chapitre

2.1 L'analyse de données

Les méthodes d'analyse des données permettent une étude globale des individus[12] et de variables[13] en utilisant des représentations[14] graphiques. Elles se distinguent des analyses statistiques traditionnelles dans le fait qu'elles s'appliquent à des grands ensembles d'individus et qu'elles ne sont pas vouées à la vérification d'hypothèses (comme le sont les statistiques inférentielles par exemple) [Bouroche, 1980].

Selon la nature des données et le type de problème, les données peuvent être analysées selon plusieurs points de vue :

---

[12] Dans ce contexte, le terme d'individu désigne un objet concret ou conventionnel sur lequel porte un ou plusieurs caractères pouvant être observés. On parle aussi parfois d'unité statistique, c'est-à-dire un élément simple ou composé (être ou objet), correspondant à une définition donnée, sur laquelle porte une étude statistique.

[13] Les variables sont des choses que l'on mesure, contrôle ou manipule en recherche. Les variables peuvent avoir des relations entre elles. Il existe plusieurs types de variables : continues (valeurs réelles comme un salaire), discrètes (valeurs entières comme le nombre d'enfants), valeurs qualitatives ou catégoriques (numériques, alphanumériques, ordonnées (faible, moyen, fort), non ordonnées), textuelles.

[14] Les individus sont représentés graphiquement par leur caractéristiques (ou variables ou dimensions). Ces variables sont, elles, représentées sur les axes abscisses (x) et ordonnées (y) pour les représentations en 2D, les axes x, y et z pour les représentations en 3D.

- On cherche à **identifier** les ressemblances et les différences entre les individus analysés. Cette analyse se fonde sur la proximité des caractéristiques entre individus. (1)
- On cherche à **construire** des ensembles homogènes d'individus, c'est-à-dire des ensembles d'individus partageant un certain nombre de caractéristiques identiques. (2)
- En décrivant les relations entre les caractéristiques des individus, on cherche à déterminer les corrélations pouvant exister entre elles et ainsi déterminer les liens voire les influences, pouvant exister entre individus. (3)

On distingue traditionnellement pour cela deux types d'approches : (1) les analyses factorielles qui visent à la réduction du nombre de caractères pour un ensemble d'individus et (2) les classifications qui réduisent le nombre d'individus puisqu'elle permettent de travailler sur des regroupement (classes) homogènes d'individus. Le troisième point de vue traduit une démarche commune aux deux autres.

2.1.1 Les analyses factorielles

L'analyse factorielle possède deux objectifs : d'une part réduire le nombre de variables dans un espace de représentation à 2 ou 3 dimensions afin de faciliter leur appréhension par l'esprit humain qui interprète difficilement des représentations graphiques au-delà de 3 dimensions, d'autre part de dégager de la masse de données des structures et des relations entre les variables. Le terme factorielle signifie que la réduction des caractères des individus analysés ne se fait pas par sélection de certains d'entre eux, mais par la construction de nouveaux caractères synthétiques obtenus en combinant les caractères initiaux au moyen de « facteurs ».

On distingue deux principales techniques [Bouroche, 1980] [Vidal, 2004] :

- **Analyse en Composante Principale** (ACP) : méthode de base de l'analyse factorielle, elle réduit le nombre de caractères permettant des représentations géométriques des individus et des caractères, c'est-à-dire de visualiser les données à n dimensions ($n>3$) dans un espace à p dimensions ($p<n$) à l'aide d'une projection de ces données sur les plans définis par les p dimensions. Les composantes principales sont les nouveaux caractères (c'est-à-dire les axes sur lesquels les coordonnées sont calculées) et chacune d'entre elles sont calculées par une combinaison linéaire des caractères initiaux.

- **Analyse Factorielle des Correspondances** (AFC) : L'AFC s'appuie sur la même logique que l'ACP à ceci près qu'elle s'applique à des données qualitatives. Elle fut proposée dans les années 60 par Benzécri[15] (1980) pour l'analyse des tableaux de contingence, c'est-à-dire le croisement de deux caractères nominaux. Elle constitue un des outils les plus puissants pour les dépouillements d'enquêtes et les résultats sont appréciés dans le traitement des données textuelles.

[15] Mathématicien tourné vers la linguistique qui anima le courant d'analyse statistique des données linguistiques dans les années 60.

2.1.2 La classification

On retrouve le terme de classification dans la terminologie des sciences de la documentation désignant les systèmes de classifications de type de ceux utilisés par les bibliothèques [Dobrowolski, 1964], comme la CDU, CDD, LCC, CIB[16], etc. Le processus de classification pour ces dernières, est réalisé de manière non automatique, tandis que les classifications dont il est question ici sont des classifications automatiques, car elles s'effectuent à l'aide d'algorithmes formalisés, c'est le cas par exemple de l'algorithme de classification incrémentale dont il est question dans le cadre de notre réflexion. Les méthodes de classification ou de typologie (dont la science s'appelle la taxinomie) ont pour but de regrouper les individus en un nombre restreint de classes homogènes. L'interprétation des données se porte donc sur un nombre réduit d'individus (par les classes représentatives).

Dans l'univers des classifications, on distingue deux méthodes de classifications :

Les classifications hiérarchiques : Cette méthode produit des successions de partitions de classes à l'extension de plus en plus vaste en lecture ascendante ou de plus en plus réduite en lecture descendante. Les méthodes de réalisation de ces classifications sont relatées dans [Bouroche, 1980] [Bellot, 2004] et [Dobrowolski, 1964] entre autres.

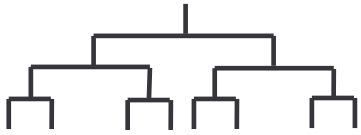


Figure 4 : Classification hiérarchique

Les classifications non hiérarchiques : Celles-ci produisent directement une partition d'un ensemble de *n* individus en un nombre *k* déterminé de classes.

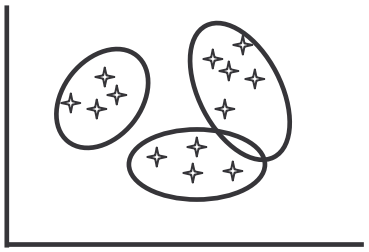


Figure 25 : Classification non hiérarchique

Différentes méthodes, dites de partitionnement, existent afin de construire des classifications non hiérarchiques :

La méthode des centres mobiles : Développé par Forgy (1965), il s'agit de la principale méthode de classification. *k* centres sont placés arbitrairement dans la représentation des données. On considère que les éléments (données) suffisamment proches du centre placé, appartiennent à la même classe dont le noyau est le centre.

[16] CDU : Classification Décimale Universelle ; CDD : Classification Décimale de Dewey ; LCC : Library Classification of Congress : CIB : Classification Internationale des Brevets

La méthode des nuées dynamiques : cette méthode, développée par Diday (1971), est considérée comme une généralisation de la méthode des centres mobiles mais la différence est qu'au lieu de construire les classes autour d'un seul point, on choisi un nombre *q* d'individus représentatifs de la classe, ce sous-ensemble d'individus constituera le noyau de la classe. Les autres éléments de la classe sont captés à l'aide du calcul de leur distance au noyau. Dans les algorithmes issus de ces méthodes, l'effectif des noyaux et le nombre de classes sont entièrement déterminés.

La méthode des k-means : La méthode des k-means (ou k-moyennes) a été introduite par MacQueen (1967). Elle est également une généralisation, de la méthode des centres mobiles vue précédemment. Les centres choisis sont ponctuels mais leur position est modifiée à chaque prise en compte d'un individu. Cette méthode possède donc une dimension adaptative que ne possèdent pas les autres méthodes.

2.2 Data mining et Text mining

Nous distinguons le data mining, appelé aussi *fouille de données*, de l'analyse de données. En effet, l'analyse de données est issue des champs de recherche mathématiques et statistiques tandis que la fouille de données est, elle, issue du champ de recherche sur l'intelligence artificielle (IA) qui utilise des techniques d'analyse de données comme celles que nous avons vu plus haut.

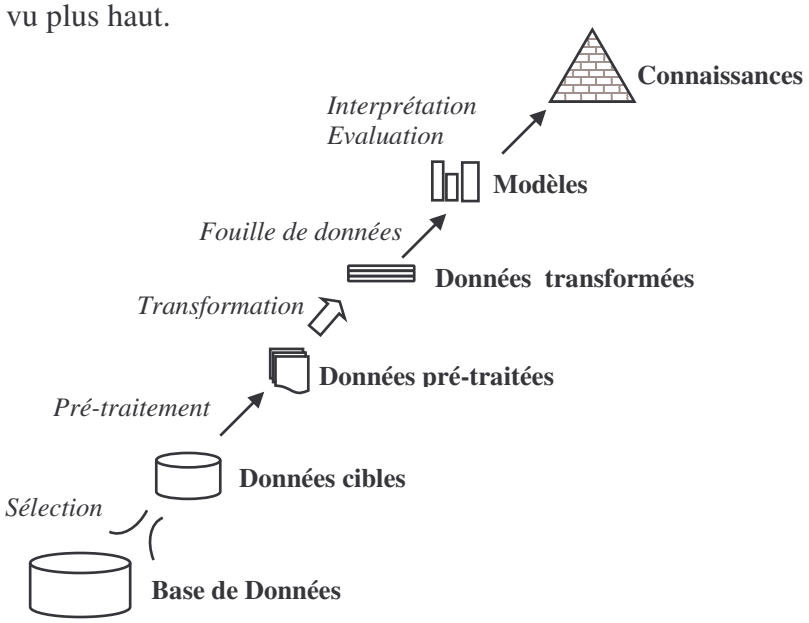


Figure 6 : Schéma global de l'ECBD ([Fayyad, 1996] cité par [Toussaint, 2004])

Toussaint (2004) définit la fouille de données comme une étape faisant partie du processus de d'*Extraction de Connaissances à partir de Bases de Données* (ECBD[17]).

La fouille de données correspond dans le schéma (figure 4), aux tâches de classification, recherche de modèles… et la définition des paramètres appropriés.

L'intelligence artificielle se définie comme « *la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement*

[17] KDD, en anglais : Knowledge Discovery from Databases

critique »[18]. Nous voyons dans cette définition que la notion d'apprentissage, c'est-à-dire la capacité d'un système à s'optimiser en fonction de son environnement, des expériences et des résultats obtenus, est un enjeu essentiel de l'IA sur lequel travaillent les chercheurs du domaine. Cette notion nous permet d'introduire une nouvelle distinction dans les classifications. On parle, en effet, de classifications supervisées et de classifications non supervisées

2.2.1 Classifications supervisées

Ce type de classification fait appel à une intervention humaine en plus d'un apprentissage automatique. En effet, les classes (que nous appellerons *catégories* dans le contexte supervisé, on parle alors de catégorisation automatique) sont établies a priori et contiennent chacune quelques données pertinemment assignées par un individu, il s'agit de la phase d'initialisation. L'apprentissage du système s'effectue, lui, en deux phases : (1) il construit un système de règles à partir des données exemples et (2) il généralise ces règles à toutes nouvelles données se présentant à lui.

Pour illustrer simplement le fonctionnement, prenons par exemple les catégories de l'annuaire *Yahoo!* Et imaginons qu'il soit possible, à l'aide d'un algorithme de classification supervisée, de permettre au système de classer automatiquement tout nouveau document dans la catégorie appropriée, uniquement à partir de règles issues de l'analyse des documents déjà présents dans les catégories existantes.

2.2.2 Classifications non supervisées

Contrairement à la méthode supervisée, la méthode non supervisée n'exige pas de catégories pré-établies, ni de corpus exemple placé par un superviseur, à partir duquel le système établi les règles. Dans ce cas, les classes (on reparle ici à nouveau de classes) sont déterminées par le regroupement de données similaires (à l'aide de méthodes d'analyse de données vues plus haut par exemple). Cependant on ne peut pas parler d'apprentissage puisqu'il n'y a pas de règles déduites d'un ensemble de données de départ, les données sont toutes traitées également lorsqu'elles se présentent.

La classification non supervisée possède un avantage que ne possède pas la méthode supervisée, puisqu'elle part du principe que l'on ne connaît pas le résultat de la classification, elle convient à une démarche exploratoire des données.

L'algorithme de classification incrémentale qui constitue notre domaine d'application, s'inscrit dans cette démarche non supervisée.

2.2.3 Cartes auto-organisées

Les réseaux neuromimétiques (ou réseaux de neurones) sont des techniques de classification fondées sur un processus d'apprentissage qui relève du système cognitif et qui simulent les fonctions neurologiques du cerveau. Grâce à leur capacité de généralisation, les réseaux de neurones sont généralement utilisés dans des problèmes de nature statistiques et perceptives, telles que la classification ou l'évaluation.

Il est possible de classer les réseaux de neurones à partir des types d'apprentissage supervisé et non supervisé vu précédemment. Dans le processus d'apprentissage supervisé, on y trouve les perceptrons[19] monocouche et multicouche. Dans le processus d'apprentissage non supervisé, on trouve les réseaux auto-organisés de Kohonen, appelées aussi cartes SOM (Self-Organizing Map). Ces cartes sont utilisées dans des domaines très variés (visualisation de données, classification de textes, reconnaissance de la parole, reconnaissance de formes,

[18] Définition de Marvin Lee Minsky, *The Society Of Mind*, Simon & Schuster, New-York, 1988.
[19] Système artificiel capable d'apprendre par expérience.

télécommunication, linguistique). Elle permettent non seulement d'aider à comprendre les données mais également de fournir un moyen de les visualiser intuitivement. La carte générée par la méthode SOM est représentée par une grille, c'est-à-dire un quadrillage de nœuds et d'arêtes liant les nœuds.

Les vecteurs des noeuds peuvent être initialisés aléatoirement. Une itération de l'algorithme correspond à calculer, pour chaque individu, la distance qui les sépare de chacun des noeuds du tableau et de l'affecter au plus proche. Après chaque affectation, il faut modifier le modèle du nœud élu ainsi que celui de son voisinage. Ce processus peut être itéré plusieurs centaines de fois. Au final, appliqué à des documents, chaque document est positionné sur un noeud et les mots de poids maximal permettent de les étiqueter (ceux de même étiquette sont alors réunies) [Bellot, 2004].

Ces méthodes viennent enrichir les méthodes statistiques pour la fouille de données. Nous les retrouverons aussi dans des applications scientométriques que nous aborderons un peu plus loin.

Nous avons annoncé le *text mining* (ou fouille de textes) dans la section sur le data mining car, comme son nom le laisse supposer, fouille de textes et fouille de données sont deux processus très proches du point de vue des méthodes, seuls diffèrent les champs d'application : l'un s'adresse à des données essentiellement numériques, l'autre à des données textuelles. La fouille de textes se définit comme « *un processus non trivial qui construit un modèle de connaissance valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts* » [Toussaint, 2004]. L'objectif de la fouille de textes est donc d'extraire des termes contenu dans des textes, grâce aux techniques de traitement automatique des langues (TAL) afin de leur appliquer des techniques d'analyse de données et les résultats de ces analyses, confrontée à des modèles, permettent de construire des connaissances (telles que définies par l'IA, c'est-à-dire comme une représentation formelle et logique d'informations, rendues ainsi interprétables par les machines et permettant à celles-ci d'"agir" [Kayser, 1997]).

Par rapport au processus d'ECBD retracé en figure 4, la fouille de texte rajoute une étape supplémentaire qui est une phase d'extraction d'information, c'est-à-dire principalement une phase de traitement automatique de la langue.

2.3 Bibliométrie, infométrie et scientométrie

Le terme d'infométrie a été adopté en 1987 par la Fédération Internationale d'information et de Documentation (FID)[20] et désigne l'ensemble des activités métriques relatives à l'information, couvrant aussi bien la bibliométrie (documentation) que la scientométrie (information scientifique et technique) [Polanco, 1995].

Alors que la bibliométrie se définit comme « *l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication* » (Prichard dans [Polanco, 1995] p.16), la scientométrie quant à elle se définit comme une spécialisation de la bibliométrie au domaine de l'information scientifique et technique. Cependant, la définition peut être étendue à toute application de méthodes statistiques à des données quantitatives (économiques, humaines, bibliographiques…) caractéristiques de l'état de la science. Price, considéré comme l'initiateur de ce mouvement de recherche, définit la scientométrie comme « *les recherches quantitatives de toutes les choses concernant la science et auxquelles ont peut attacher des nombres* » (Price dans [Polanco, 1995]), il aboutit ainsi à la constitution d'une nouvelle forme de *science de la science*, au-delà de la dimension exclusivement

[20] Pour plus d'informations sur l'historique et les activités de cette organisation : http://people.lis.uiuc.edu/~wrayward/otlet/FIDHIST2.htm

épistémologique. La scientométrie fait donc usage des techniques d'analyse de données et de fouille de données / textes pour répondre à sa vocation métrique.

L'origine de la scientométrie est, avant tout, sociologique, la dimension statistique vient pour appuyer les études dans ce domaine [Courtial, 1990]. En effet, l'objet d'étude de la scientométrie est la dynamique de la recherche à une échelle donnée, les relations entre les acteurs, etc. Dans ce champ de recherche on étudie, par exemple, les *collèges invisibles*, c'est-à-dire des groupes d'élite qui se constitueraient au sommet de la communauté scientifique et autour d'un front de recherche.

Puisque son domaine d'application est la science institutionnalisée à différents niveaux (local, national, international…) et soumise à une politique particulière, la scientométrie a, de plus, pour objectif de répondre à un besoin de gestion voire de management de la recherche. Van Raan [Van Raan, 1988] la définit dans ce cadre comme « *le champs de recherche où l'on utilise les méthodes et les techniques mathématiques, statistiques et de l'analyse de données en vue de rassembler, manipuler, interpréter et prévoir une variété de caractéristiques telles que la performance, le développement et la dynamique de la science et de la technologie* ». Elle acquière dans ce cadre une fonction économétrique[21] [Courtial, 1990].

En se basant sur la dichotomie : recherche fondamentale et recherche appliquée, il existe selon Van Raan [Van Raan, 1988] une distinction dans la finalité de la scientométrie. Dans un contexte de recherche fondamentale, les méthodes quantitatives constituent « *un élément indispensable pour l'avancement de notre compréhension dans l'étude de la science en tant que système complexe de production et d'échange de connaissances* ». Dans le contexte de la recherche appliquée, la scientométrie répond principalement à « *la demande d'indicateurs quantitatifs de la science et de la technologie pour part de la politique scientifique* ».

Ce second cas correspond à la conception que nous pouvons nous faire de cette approche dans le cadre d'un projet de veille scientifique et technique, tel que nous l'avons décrit plus haut : la scientométrie répond à un besoin d'indicateurs quantitatifs de la science et de la technologie pour la prise de décision.

Concernant les méthodes, Courtial considère que la scientométrie « *se fait à partir de des indicateurs qui expriment l'activité scientifique* » ([Courtial, 1990], p.12), il en dégage alors deux types d'indicateurs regroupant différentes méthodes d'analyses : les indicateurs d'activités et les indicateurs relationnels.

2.3.1 Les indicateurs d'activité

Dans cette catégorie, nous regroupons les techniques issues de la bibliométrie concernant le comptage des publications scientifiques par auteur, laboratoire, domaine, etc. ainsi que des brevets par organisme, déposant, inventeur, pays, etc.

Concernant les publications scientifiques, il existe une série de lois bibliométriques, formulées à partir de constats empiriques des phénomènes liés au comptage des publications. Ces lois fondent les méthodes de calcul d'indicateurs de politique scientifique, elles sont, en partie, à la base de calculs d'indicateurs menés par l'Observatoire des Sciences et Techniques (OST), par exemple.

[21] Étude et représentation des phénomènes économiques grâce à l'utilisation des mathématiques et des statistiques. (in GDT, http://granddictionnaire.com/ , consultée le 29/03/06)

Loi de Lotka

Un domaine scientifique se caractérise par un petit nombre de spécialistes publiant beaucoup et un grand nombre de publications occasionnelles. On peut représenter cette régularité observée par Lotka[22] (1926) à l'aide d'une fonction hyperbolique.

Cette représentation se traduit par la fonction $y=1/x^2$, avec la variable *x* pour le nombre d'articles publiés et *y* le nombre d'auteurs ayant publié *x* articles.

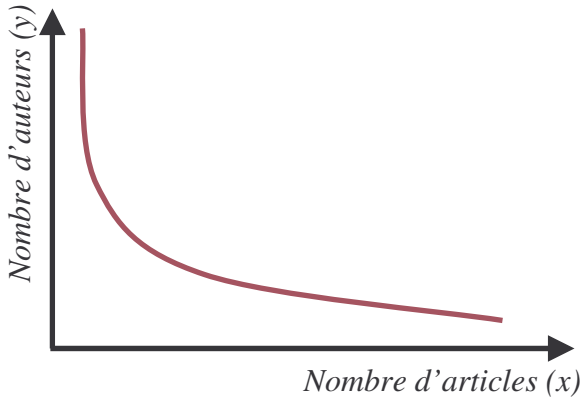


Figure 7 : Représentation graphique de la loi de Lotka

Loi de Bradford

Bradford, en 1930, s'est intéressé à la répartition des articles scientifiques pour un domaine précis, dans les périodiques. Un des problèmes auquel il a été confronté était de sélectionner parmi tous les périodiques d'un domaine ceux qui seraient les plus représentatifs du domaine. Il a pu observer qu'un nombre relativement restreint de revues publie l'essentiel des résultats scientifiques significatifs.

Cette loi, par les préoccupations de son auteur, est au départ à visée plutôt bibliothéconomique[23], mais elle peut être analysée avec des intentions scientométriques et, appliquée sur l'ensemble des publications scientifiques mondiales se citant les unes aux autres, on observe alors une loi du même type ([Courtial, 1990] p.43).

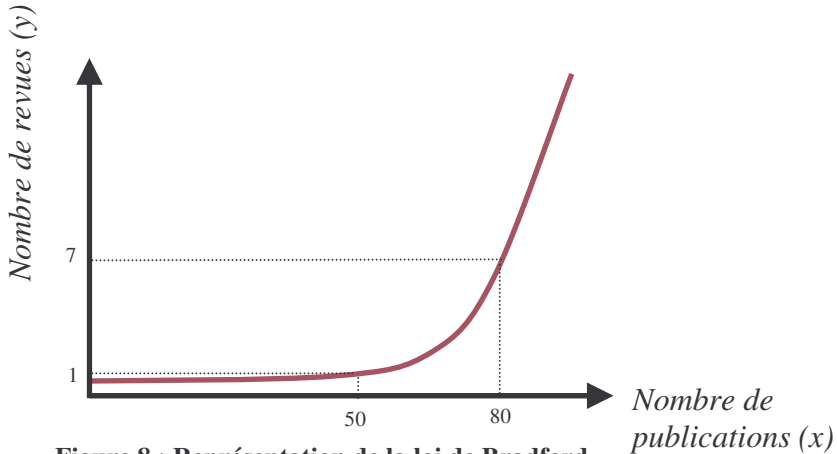


Figure 8 : Représentation de la loi de Bradford

La figure ci-dessus illustre la loi de Bradford. Sur celle-ci nous voyons qu'un petit nombre de revues (axe y) publie un grand nombre de publications (axe x) et plus la quantité

[22] Alfred James Lotka (1880-1949) mathématicien et statisticien américain, théoricien de la dynamique des populations.

[23] S.C. Bradford (1878-1948) était bibliothécaire au Musée scientifique de Londres.

de revues augmente, plus la densité d'article décroît : le nombre de publications se réparti sur un plus grand nombre de revues.

Loi de Zipf

Alors que Lotka a montré une corrélation entre le nombre de publications scientifiques et le nombre d'auteurs, alors que Bradford observe, quant à lui, une corrélation entre le nombre de publications significatives et le nombre de revues qui les diffuse, Zipf[24] s'est intéressé, en 1949, à la relation entre la fréquence d'occurrence d'un mot dans un texte et son rang dans l'ordre des fréquences.

Selon Zipf, on peut observer une constante K résultant du produit entre la fréquence d'occurrence $f(n)$ du terme et le rang n de ce terme, ce qui se formalise sous la forme : $f(n)\times n=K$. Il est évident que plus un terme est fréquent plus son rang est élevé, mais il était moins évident de constater qu'il existe une régularité entre la position de ce terme dans le classement et sa fréquence.

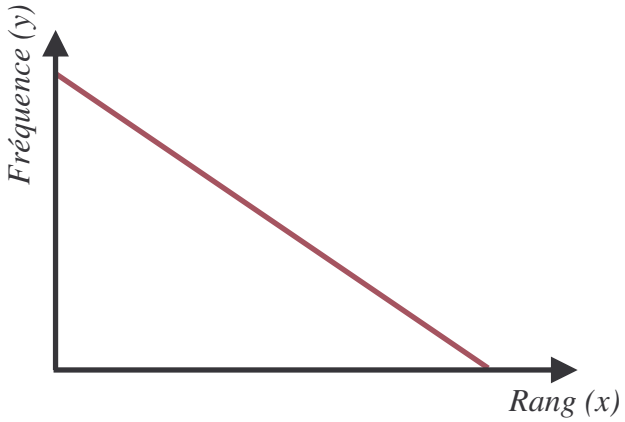


Figure 9 : Représentation de la loi de Zipf

Sur la figure ci-dessus, nous pouvons constater la relation constante entre le rang d'un mot et sa fréquence d'apparition dans le texte.

Distribution des avantages cumulés

Price (1976) développe une théorie probabiliste qui s'inspire, entre autres, des travaux de Lotka, Bradford, Zipf,… Cette théorie, nommée *distribution des avantages cumulés* (DAC), propose le modèle suivant : plus une source produit des éléments (les chercheurs produisent des articles, citent des articles…) plus elle a de chance d'en produire d'avantage et, contrairement à ce qu'on pourrait croire, la non productivité n'implique pas la diminution de ces chances.

Cette théorie présente donc les lois citées précédemment comme des limites du modèle qu'il propose.

2.3.2 Les indicateurs relationnels

Les indicateurs relationnels sont ainsi appelés car ils permettent de traduire des relations entre les éléments étudiés à l'aide des caractéristiques de ces éléments, c'est-à-dire qu'en étudiant des articles scientifiques, par exemple, il est possible d'établir des relations (s'il en existe) entre ces articles, entre les auteurs, les co-auteurs…

[24] George Kingsley Zipf (1902-1950), fut un linguiste et un philologue américain qui étudia la statistique appliquée aux différentes langues..

Callon et al. [Callon et al., 1993] distinguent deux types d'indicateurs relationnels : ceux de première génération et ceux de seconde génération. Les premiers s'attachent particulièrement à la description des relations entre articles, auteurs, organismes, etc. Ils sont basés sur les signatures[25] d'articles, les citations[26], etc. La méthode principale ayant été développée est la méthode des co-citations (*cocitation analysis*) que nous développerons après.

La seconde génération d'indicateurs relationnels est plutôt axée sur le contenu même des productions scientifiques. Ils recouvrent des analyses basées sur la cooccurrence des mots (*coword analysis*), sur la mesure des liens entre mots clés servant à construire des thématiques constituées de termes sémantiquement proches et l'analyse de ces thématiques. La méthode la plus représentative est celle des mots associés

Les co-citations

La méthode est fondée sur une analyse très fine des références que contiennent les articles scientifiques. Une citation prise en dehors du contexte lui donnant sa signification est difficile à interpréter, en revanche l'apparition simultanée de deux citations lorsqu'elle se répète dans un grand nombre d'articles est dotée d'une signification plus précise.

Les travaux les plus représentatifs de cette approche sont ceux d'H. Small et E. Garfield qui fondèrent l'*Institute for Scientific Information* (ISI) dans les années 1960 et développèrent la base de données bibliographiques appelée *Science Citation Index* (SCI) (devenue aujourd'hui l'*ISI Web of Knowledge*), avec l'objectif de mettre au point des indicateurs mesurant la « consommation » des résultats scientifiques. Ils proposent dans leurs travaux d'analyser ces articles en prenant en compte les citations afin de mettre en évidence les relations entre les acteurs. A l'aide de la méthode MultiDimensional Scaling (MDS, méthode mathématique qui permet de représenter dans un espace un ensemble d'individus entre lesquels nous avons une notion de proximité ou de distance), ils constituent ainsi des clusters, c'est-à-dire des regroupements d'articles liés entre eux par leurs références bibliographiques. L'intérêt de cette approche est, comme nous l'avons déjà dit, de retracer les liens entre les auteurs mais aussi de construire des thématiques car *« les articles fondamentaux ou cités dans ces clusters, tendent à partager un thème commun, d'un point de vue théorique ou méthodologique ou les deux* » [Small, 1993].

Les mots associés

L'analyse des mots associés (*coword analysis*) est un outil développé initialement au Centre de Sociologie de l'Ecole des Mines de Paris et à l'INIST / CNRS dans les années 80. Elle s'applique aux document structurés et indexés tels que les articles scientifiques et techniques ou les brevets, mais peut s'appliquer à tout document textuel.

La méthode considère les mots clés comme des caractéristiques des documents et propose de les analyser alors à l'aide des techniques d'analyse de données en partant d'un tableau de type *documents* × *caractéristiques*. A partir de ce tableau (appelé aussi matrice), la méthode permet de construire des réseaux de cooccurrence de mots qui sont à leur tour découpés en classe à l'aide d'une méthode de classification. Les classes résultantes constituent les thématiques que l'on peut dégager de l'ensemble de documents analysés.

Ces thématiques peuvent à leur tour être analysées selon deux critères, à savoir la densité, c'est-à-dire qu'une grande similarité unissant les termes de cette classe, et la

[25] Chaque article scientifique est signé par son auteur qui est affilié à un organisme. L'article peut aussi être co-signé par un ou plusieurs auteurs du même organisme ou d'un organisme différent, ce dernier cas révéle alors une collaboration entre les deux organismes.

[26] La rédaction d'un article scientifique amène son auteur à citer ses sources, les textes sur lesquels il s'appuie dans sa nouvelle argumentation.

centralité, c'est-à-dire que la classe entretient une relation plus ou moins forte (des termes en commun) entre les classes voisines. Une visualisation en diagramme, peut reprendre ces deux critères sous la forme de deux axes et permet de placer les classes les unes par rapport aux autres, comme le fait l'outil SDOC de la plateforme STANALYST[27], qui utilise la technique des mots associés. L'interprétation d'un tel diagramme a été proposé par Callon et al. (1993) par l'introduction d'un diagramme dit *stratégique*.

En plus des méthodes présentées par Callon, nous pouvons rajouter deux nouvelles catégories d'analyses scientométrique que sont les cartes Multi-SOM et les k-means axiales.

Les cartes Multi-SOM

L'approche Multi-SOM [Lamirel, 1995] est une variante des cartes de Kohonen car elle propose de multiples cartes (les cartes SOM de Kohonen ne possédaient que deux dimensions). Il s'agit d'une méthode proposée au LORIA est implémentée à l'INIST pour la découverte et la visualisation de thèmes de recherche. Ce modèle est fondé sur le concept de point de vue, et la mise en œuvre des points de vue sous la forme de cartes auto-organisatrices. Les cartes multiples sont reliées entre elles par un mécanisme original de communication. Le système inclut la capacité d'étiqueter les classes, et également un mécanisme de généralisation.

Les k-means axiales

L'approche des k-means axiales (KMA) est, elle, une variante des k-means. Développée par Lelu, la méthode considère l'ensemble des références bibliographiques comme un nuage de points plongé dans un espace géométrique où chaque dimension correspond à un mot-clé (les axes *w,x,y,z, etc.* sur la figure 10). Elle est caractérisée par une représentation des classes à l'aide de vecteurs (**A**) pointant vers les zones de forte densité du nuage. A l'inverse des techniques de classification non hiérarchiques qui représentent les *k* classes recherchées par leur centre de gravité, les k-means axiales définissent les *k* classes recherchées par *k* demi-axes (**A**…) passant par l'origine de l'espace géométrique ou *k* vecteurs unitaires pointant dans la direction de ces demi-axes.

La position des *k* demi-axes est définie au hasard ou par les *k* premiers documents. Par itérations successives, les axes se positionnent puis se stabilisent dans les zones de forte densité du nuage de documents, élaborant ainsi une classification des documents.

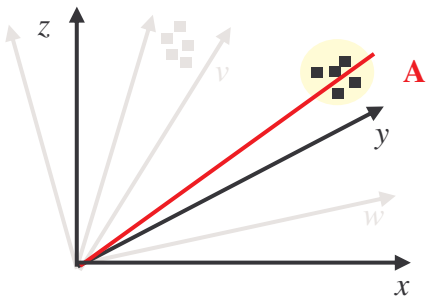


Figure 10 : Illustration des k-means axiales

Les k-means axiales sont associées à la méthode d'ACP afin d'obtenir la carte globale des thèmes (une classe de documents correspondant à un thème). Dans l'application

[27] Pour plus d'information : http://ingenierie.inist.fr/article44.html
Pour accéder (identifiant et mot de passe requis) : http://stanalyst.inist.fr/

NEURODOC[28], autre outil de la plateforme STANALYST illustrant la démarche, une classe est représentée graphiquement par un point.

Nous avons pu voir dans le premier chapitre de cet état de l'art que le processus de veille, inspiré du processus plus global d'intelligence économique, est confronté à une difficulté qui fait sa raison d'être, c'est-à-dire comment extraire des informations pertinentes et utiles au décideur, à partir de grands ensembles d'information et comment leur apporter une valeur ajoutée

Du début de l'état de l'art jusqu'à maintenant, nous avons tenté de montrer le lien existant entre les besoins d'une démarche de veille et les techniques existantes d'analyse de l'information permettant justement d'apporter des outils de plus en plus complets et adaptés à ces grands ensembles.

Nous allons désormais nous pencher sur ce qui fait le cœur de notre problématique. Celle-ci, rappelons la rapidement, cherche à déterminer les fonctionnalités de visualisation de l'information utiles aux tâches d'un veilleur, notamment lorsque cette visualisation représente l'évolution de cette information. Quelles sont les techniques existantes de visualisation ? Quels sont les domaines d'application ? Quels sont les enjeux et l'intérêt de ce mouvement de recherche technologique et scientifique qui tend à se développer depuis une dizaine d'années ? Telles sont les questions auxquelles nous tenterons de répondre dans ce qui va suivre. Nous essaierons aussi de répondre à : qu'a-t-on proposé jusqu'à maintenant pour visualiser l'évolution de l'information ou d'ensemble de données ? Quels dispositifs on été imaginés pour répondre à une démarche couramment appelée « découverte de connaissances » (knowledge discovery) ? Nous nous intéresserons aussi à l'utilisateur, élément de notre triptyque de réflexion : Analyse de l'information – Utilisateur – Visualisation de l'information.

[28] Pour compléments d'informations : http://ingenierie.inist.fr/article44.html

Chapitre 3

La visualisation de l'information

Dans cette partie, notre démarche consiste tout d'abord à définir ce qu'on entend par visualisation de l'information et quels en sont les enjeux. Nous entreprendrons ensuite un état de l'art général sur les différents types de visualisation pouvant être rencontrée, cet état de l'art se focalisera progressivement sur la visualisation des domaines de recherches et notamment sur les manières de visualiser leur modification au cours du temps, ce qui est, rappelons le, un élément essentiel de notre problématique.

3.1 Qu'est ce que la visualisation de l'information ?

Avant de définir ce qu'est la visualisation de l'information, nous pouvons nous interroger sur ce que le terme *visualisation* signifie. Polanco (2002) définit cette notion comme regroupant deux domaines : la visualisation scientifique et la visualisation de l'information.

La première s'applique sur des données physiques (structures ADN, représentation de molécules…), la seconde désigne « *l'utilisation informatisée de représentations visuelles interactives de données abstraites de manière à amplifier la cognition. La visualisation de l'information est un champ de recherche en informatique qui se consacre à la création d'interfaces visuelles riches aidant l'utilisateur d'un système à comprendre et à naviguer au travers d'espaces informationnels complexes.* » [Polanco, 2002].

L'objectif et la difficulté de la visualisation de l'information est de produire des représentations visuelles, donc concrètes, à partir d'éléments abstraits qui n'ont ni forme, ni couleur, ni dimension, ni position, etc. Le problème principal des recherches dans ce domaine est alors de découvrir de nouvelles métaphores visuelles pour représenter l'information et de comprendre quelles sont les tâches analytiques qu'elles peuvent supporter.

La visualisation d'information est donc la présentation graphique de l'information. Son but est de fournir à l'utilisateur une compréhension qualitative du contenu de l'information. L'information peut être des données, des processus, des relations ou des concepts. La présentation graphique peut nécessiter la manipulation des entités graphiques (points, lignes, formes, images, textes, surfaces) et des attributs (couleur, intensité, taille, position, forme, mouvement).

Pour nous qui nous intéressons à l'évolution des thématiques de recherche, la visualisation de l'information, selon Kapusova (2004), combine des aspects de la visualisation scientifique, des interfaces homme-machine (*human-computer interfaces*), de l'exploitation de données (*data mining*), de l'imagerie et des graphiques. Pour Fekete (2005), la visualisation de l'information s'est détachée de trois domaines connexes : l'interaction Home-machine, l'analyse statistiques et la cartographie, et la visualisation scientifique.

3.2 Quels sont les enjeux de la visualisation d'information ?

3.2.1 Adapter l'exploration de grands gisements informationnels

Alors que Polanco (2002) considère la visualisation de l'information comme un *champ de recherche* scientifique et technique, certains auteurs comme Munzner (2003) conçoivent la visualisation comme un *phénomène*, avant d'être une discipline, venant en *réaction* d'un accroissement extrêmement rapide des données accessibles :

« *Le phénomène de la visualisation de l'information est une réaction récente au fait que la quantité de données à laquelle nous avons accès progresse à un rythme qui va au-delà de ce qui se comprend aisément. Ces données explosent même plus rapidement que le taux de croissance de la puissance informatique. Le champ de la visualisation informatique de l'information (ou l'*infovis*) concerne la création d'outils qui exploitent le système visuel humain pour aider les gens à explorer ou à expliquer des données. L'interaction avec une représentation visuelle de données soigneusement conçue peut nous aider à former des modèles mentaux qui nous permettent d'exécuter des tâches spécifiques plus efficacement* » [Munzer, 2003].

Nous voyons donc là le premier enjeu essentiel de la visualisation de l'information qui est d'adapter le mode d'accès aux données à des espaces d'informations qui risquent de devenir inexploitables du fait de la quantité et de la diversité grandissante des informations. Keim (2001) écrit qu'une estimation faite à l'université de Berkley (Californie), en 2001, avait conclu qu'un million de terabytes de données est générées annuellement dans le monde et que 99.997 % de ces données étaient exclusivement disponibles sous forme numérique. Nous commençons déjà à entrevoir comment la visualisation rejoint des problèmes communs à l'analyse de l'information que nous avons présentée précédemment.

La voie poursuivie par ce champ de recherche est donc de modifier la perception de ces données afin d'en améliorer la compréhension. Hearst donne une raison du succès de ce type d'approche : « *Les humains sont fortement habitués aux images et aux informations visuelles. Une représentation visuelle peut communiquer certains types d'information beaucoup plus rapidement et efficacement qu'une autre méthode* » [Hearst, 1999].

3.2.2 Amplifier la cognition

Tous les auteurs écrivant sur la question s'accordent sur le fait que la représentation visuelle des données réduit le travail cognitif[29] nécessaire pour accomplir certaines tâches [Card et al., 1999], [Keim, 2001], [Polanco, 2002], etc.

En effet, comme le montre Card et al. (1999, p.1-2), il suffit de constater qu'il est moins facile et moins rapide de résoudre une multiplication en l'effectuant de tête qu'en la posant sur le papier sous la forme ci-dessous (Figure 11), pour concevoir que la visualisation aide à réfléchir sur des problèmes. S'agissant du fait de reposer sa réflexion sur des représentations visuelles, par exemple, certains auteurs parlent aussi de *cognition externe*, en « opposition » à la *cognition interne* [Card et al., 1999].

34
× 72
68
23[1]80
24 48

Figure 11 : L'utilisation d'une aide externe visuelle accroit la capacité à résoudre cette multiplication

Cette propriété est aussi utilisé dans le domaine de la pédagogie au sujet des images mentales, c'est-à-dire les représentations du monde qui se créent dans l'esprit des individus.

[29] La cognition regroupe les divers processus mentaux allant de l'analyse perceptive à la mémorisation, à l'appropriation dans des schémas ou des concepts, par lesquels nous construisons une représentation opératoire de la réalité à partir de nos perceptions, susceptible en particulier de nourrir nos raisonnements.

Le fait qu'un lieu familier puisse nous être évoqué par une odeur, par exemple, appartient à la catégorie des images mentales.

Card et al. proposent d'identifier 6 principales causes d'amplification de la cognition par la visualisation :

- **Réduction des ressources cognitives mobilisées par l'utilisateur pour traiter et analyser les informations** (interaction élevée avec l'utilisateur, perception menée en parallèle, facilité d'accès à une grande quantité d'information…)

- **Simplification de la recherche d'information** (beaucoup de données dans un petit espace, regroupement de données - par critères par exemple)

- **Augmentation des possibilités de détection de structures** (relations entre données, regroupement significatif…)

- **« Perceptual inference »** : inférer à l'aide de la perception visuelle (certains problèmes paraissent évidents à l'aide d'une représentation visuelle, …)

- **Surveillance des événements** (changements de structures, apparition ou mouvement dans les motifs…)

- **Moyen de manipuler les données**

3.2.3 Construire du sens

Gershon & Pages (2001) indiquent que « *l'objectif d'une telle approche est de permettre à l'utilisateur d'observer, de comprendre et de construire du sens à partir de l'information* ». Cette construction du sens désigne en fait l'interprétation que fait l'utilisateur des données en s'aidant d'une représentation. En effet, si la visualisation permet d'aider à résoudre des problèmes, c'est non seulement parce qu'elle nous donne la possibilité de manipuler ces données abstraites au départ, mais c'est aussi parce qu'elle offre à l'utilisateur des clés d'interprétation qu'il ne pouvait peut être pas percevoir auparavant. La représentation est construite à partir des données qui représentent des phénomènes sous jacents [Hearst, 1999].

3.2.4 Parer à la sous exploitation des Systèmes de Recherche d'Information (SRI)

Selon Favier et al. (2000) et Ihadjadene (1999), les outils de recherche d'information sont sous exploités. En effet, des travaux sur l'usage des moteurs de recherche, des bibliothèques numériques et des www-Opacs menés par Ihadjadene (1999), ont montré que les « *ressources du système sont sous-utilisées et que les outils mis à disposition de l'usager final pour explorer le nombre élevé de réponses sont insuffisants et inadaptées.* ». Favier (2000) ajoute que dans le cas du moteur ALTAVISTA, « *85% des usagers se contentent des dix premiers résultats fournis sur la première page et 78% des requêtes ne sont pas modifiées dans le but de les améliorer.* »

Rejoignant l'argument de Munzner (2003, *cf.* plus haut), une façon de parer à cette sous-exploitation des systèmes de recherche d'information ainsi que celle de tout système d'information contenant un grand nombre d'informations, est de proposer à l'utilisateur final

une approche alternative à la représentation traditionnelle en listes des résultats ([Hascoët, 2004], p. 98).

Sous cet aspect nous retrouvons les interfaces de visualisations de résultats de SRI, comme Kartoo[30] ou encore Webrain[31]. Le moteur Kartoo, en effet, présente ses résultats sous la forme d'une carte de documents dans laquelle les nuances de couleur représentent la pertinence graduée des documents présentés, de plus certaines fonctionnalités guident l'utilisateur pour préciser sa requête. Webrain, quant à lui, se présente sous la forme d'une arborescence hiérarchique et interactive des rubriques d'un annuaire, permettant à l'utilisateur de se situer plus aisément dans la hiérarchie.

Figure 12 : Interface graphique du moteur Kartoo

Figure 13 : Interface graphique du moteur Webrain

3.3 Principes pour la construction de représentations visuelles d'information

3.3.1 La représentation visuelle de l'information quantitative

Les travaux de Tufte (2001) marquent une étape importante dans la réflexion qui s'instaure depuis des années sur la visualisation de l'information. L'analyse d'un grand nombre de représentations graphiques d'information dans [Tufte, 2001], permet à l'auteur de proposer une analyse à partir du type d'idées, c'est-à-dire, d'informations à transmettre à l'aide du graphique, du temps nécessaire au lecteur du graphique pour percevoir ces idées et le rapport entre l'encre et l'espace utilisé par le graphique pour communiqué son message. De cette approche, il en retire un certains nombre de principes.

Les premiers sont les principes d'excellence graphique (*Principles of graphical excellence*) ([Tufte, 2001], p 51) :

L'excellence graphique…

1. consiste en une bonne conception d'une présentation de données intéressantes,
2. consiste à communiquer des idées complexes de manière claire, précise et efficiente (l'efficience, c'est-à-dire, le rapport entre l'effort cognitif du lecteur et l'apport informationnel qu'il en tire),

[30] http://www.kartoo.com/
[31] http://www.webbrain.com/

3. est ce qui donne au lecteur le plus grand nombre d'idées dans le plus court intervalle de temps possible et avec le moins d'encre possible dans le plus petit espace possible.
4. est presque toujours multi variée,
5. requière de toujours exprimer la vérité à propos des données.

Tufte propose aussi d'autres principes réunis sous le titre de principes d'intégrité graphique (*Principles of Graphical integrity*) ([Tufte, 2001], p. 77) :

1. La représentation graphique de valeurs numériques sont physiquement mesurables sur le graphique,
2. L'étiquetage clair, détaillé et complet permet de réduire les déformations graphiques et les ambiguïtés. Les explications du graphique doivent se trouver en dehors du graphique. Les événements important doivent, en revanche, être signalés sur le graphique.
3. Seules les variations dues aux données doivent être montrées, non les variations dues à leur représentation.
4. Le nombre de dimensions utilisées pour représenter une donnée ne doit pas excéder le nombre de dimensions de cette même donnée.

Ces principes, inspirés par de multiples observations, peuvent être intéressants en tant que ligne de conduite à tenir lors de la conception de représentations visuelles de données.

3.3.2 La sémiologie graphique

Dans même que Tufte, Bertin est considéré comme le point de départ essentiel en terme de cartographie d'information. Bertin (1967) s'intéresse à la construction de visualisation par les symboles graphiques. Géographe de profession, ses travaux s'attachent à définir les principes de la construction de cartes. Nous pensons que les connaissances de géographes à propos de la cartographie sont intéressantes dans notre cas.

D'après Bertin, le lecteur de la carte perçoit six variations sensibles attachées aux symboles de la carte :

1. Taille (…)
2. Valeur ( )
3. Grain (…)
4. Couleur (…)
5. Orientation (…)
6. Forme (…)

Il les appelle variables visuelles, « *composantes du système d'expression* ». A chacune d'entre elles se rattache un ou plusieurs niveaux d'organisation du plan, appelées aussi *perception*.

- Variable associative (≡) : *lorsqu'elle permet de regrouper spontanément toutes les correspondances différenciées par cette variable. Selon lui toutes les variables peuvent être associatives. La Valeur et la Taille en revanche ne le sont pas, on les dit dissociatives* (≠).

- Variable sélective (≠) : *lorsqu'elle permet d'isoler spontanément toutes les correspondances appartenant à une même catégorie (de cette variable). Seule la Forme n'est pas sélective*.

- Variable ordonnée (**O**) : *lorsque le classement de ses catégories, de ses paliers et spontané et universel. Grain, Valeur et Taille sont des variables ordonnées.*

- Variable quantitative (**Q**) : *Lorsque la distance visuelle entre les catégories d'une composante ordonnée peut s'exprimer spontanément par un rapport numérique. La Taille est typiquement une variable quantitative*.

Il dresse ainsi le tableau des associations « variables - niveaux d'organisation » que le lecteur trouvera dans le tableau 2 :

	Niveaux d'organisation des variables visuelles			
Dimensions du plan	≡	≠	**O**	**Q**
Taille	≠≡	≠	**O**	**Q**
Valeur	≠≡	≠	**O**	
Grain	≡	≠	**O**	
Couleur	≡	≠		
Orientation	≡	≠		
Forme	≡			

Tableau 2 : Propriétés des variables visuelles [Bertin, 1967]

A partir des éléments de sens que sont capables de véhiculer les variables visuelles, le cartographe est invité à composer les cartes qui communiqueront de façon claire, précise et efficiente, les données et par elles, les idées qu'il souhaite représenter.

Bertin (1967) propose de présenter une théorie de l'image, il met ainsi en évidence la notion d'efficacité (ou prégnance) qu'il défini ainsi : *« Si pour obtenir une réponse correcte et complète à une question donnée, et toutes choses égales, une construction requiert un temps d'observation plus court qu'une autre construction, on dira qu'elle est plus efficace pour cette question. »* ([Bertin, 1967], p.139). Cette efficacité est tributaire du processus de lecture de l'utilisateur, des questions possibles (c'est-à-dire ce que nous appellerons, nous, ses besoins informationnels), de la définition de l'image comme « *forme significative perceptible dans l'instant minimum de vision* », la construction de cette image par l'utilisation des variables visuelles et les limites de l'image comme représentation de la réalité nécessairement simplifiante.

3.4 Visualisation et scientométrie : Maps of science

Dans notre étude sur la visualisation de thématiques, dans un contexte d'information scientifique et technique, nous ne pouvons oublier que la visualisation est étroitement liée à la scientométrie, notamment en rapport aux études des citations dont les représentants sont Garfield et Small, qui parlent alors de « maps of science » (les cartes de la science). Small défini le concept dans [Small, 1999] comme une « *représentation spatiale montrant comment les disciplines, champs, spécialités, articles ou auteurs sont reliés aux autres, représentés par leur proximité physique et leur position relative* ». L'intérêt d'une telle approche de la littérature scientifique est de faciliter notre compréhension des relations et des

développements conceptuels, elle permet de visionner un « état contemporain des connaissances».

Les cartes de la science n'ont pas comme unique fonction de représenter un état des connaissances à un moment donné, elles permettent aussi de prévoir l'avenir de ces connaissances à partir de l'observation de tendances (*trends*), comme l'indique Garfield (1986), « *même si nos cartes ne peuvent pas prédire où les chercheurs iront exactement, elles peuvent servir d'indicateurs. Les changements d'année en année révèlent des tendances et les cartes peuvent donc servir comme outils de prévision*. »

Dans ce domaine nous pouvons signaler, de manière non exhaustive, les travaux de Chen (2006) et de Moya-Anegón et al. (2004) qui ont poursuivi les travaux de Garfield et Small dans ce domaine et qui ont réalisé des outils cartographiques interessants. Le premier a développé *CiteSpace2*, un outil d'analyse et de visualisation de carte des domaines de la recherche scientifique présentant les liens entre publications et la création (Figure 45) de ces liens dans le temps (Figure 46). Les seconds, du laboratoire SCImago de l'université de Grenade (Espagne) ont réalisé un outil appelé *Atlas of Science*. Cet outil appliqué sur le SCI représente les champs de recherche au travers les disciplines et leurs interconnexions, tout cela à partir de l'analyse des citations de publications. La carte est à trois niveaux :

Le premier représente la globalité des champs de recherche d'un pays (Figure 14),

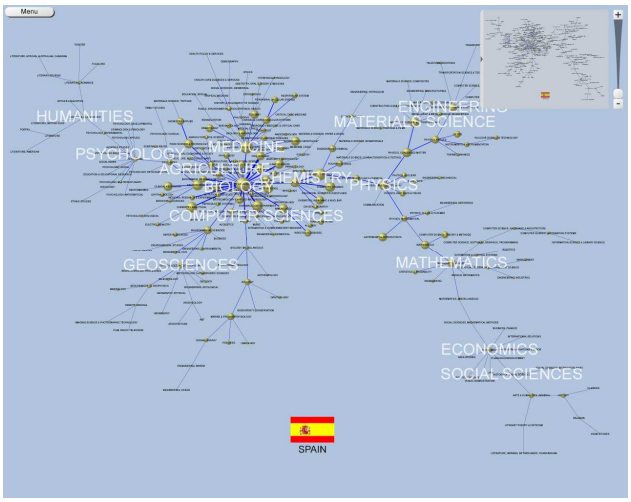


Figure 14 : Vue globale de Atlas of Science

Le second intervient lorsque l'on veut focaliser sur un élément, celui-ci est représenté sous une forme héliocentrique : l'objet sur lequel l'utilisateur porte son attention est placé au centre et les autres objet liés à celui-ci sont placés autour, de manière à créer une représentation en forme de soleil (d'où héliocentrique) (Figure 15),

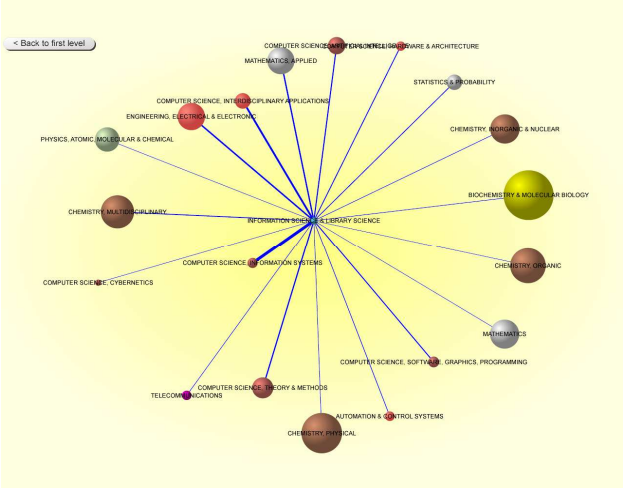


Figure 15 : Détails sur une discipline

Le troisième permet de visualiser sous forme d'un réseau, les sous disciplines, les différents centres d'intérêt présents dans la discipline (Figure 16).

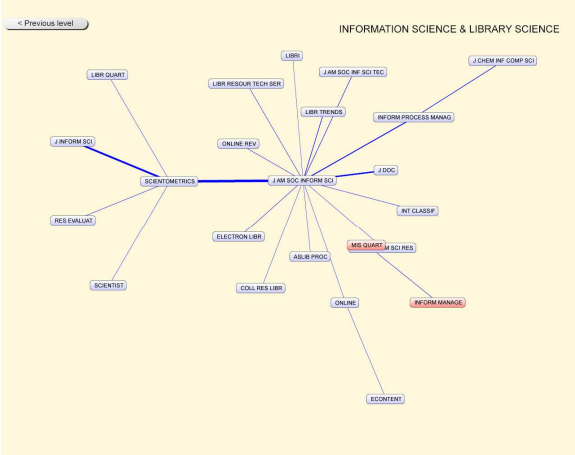


Figure 16 : Détails sur les sous-disciplines

En dehors de la visualisation elle-même, *Atlas of Science* offre de nombreuses possibilités d'accès à des informations sur les documents à l'origine de l'analyse, comme par exemple des statistiques sur les types de documents, les affiliations, etc.

3.5 La visualisation de l'information : approche technique par les données

Il existe de nombreuses typologies des visualisations de l'information. Celle de Keim (2001) propose que les techniques de visualisation soient classées selon trois critères, à savoir le type de données à visualiser, le type de visualisation et les types d'interactions entre utilisateur et visualisation. Ce qui revient à une classification originale à 3 dimensions comme le montre la figure suivante.

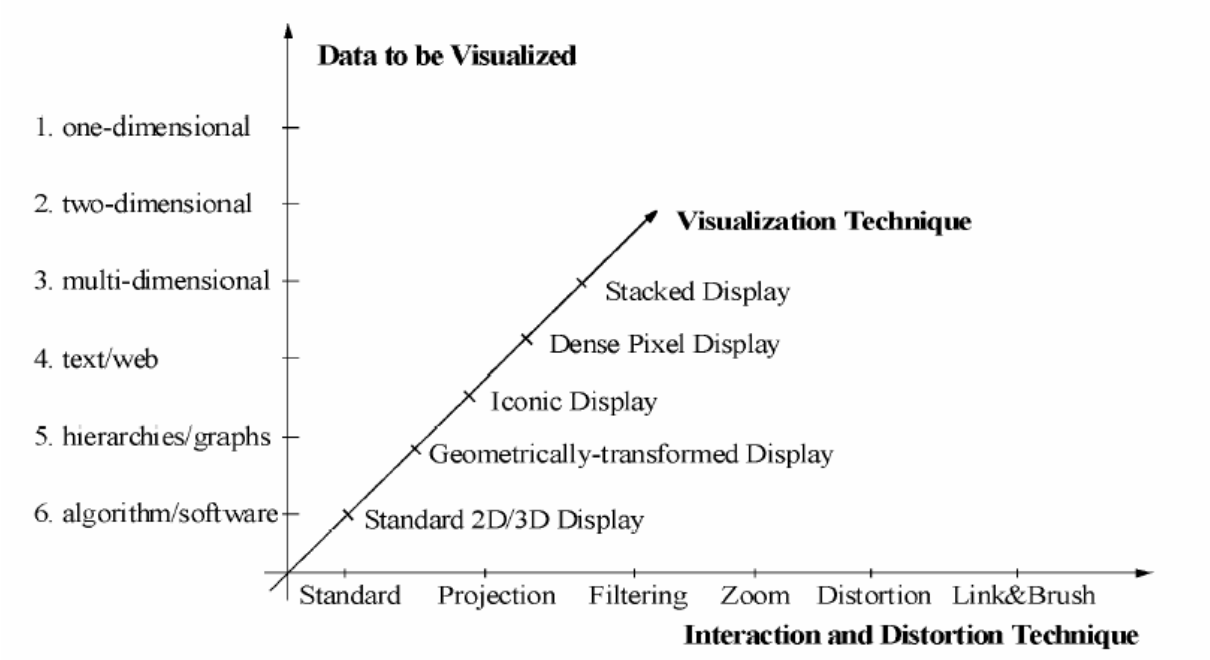


Figure 17 : Classification de Keim

Nous proposons d'établir notre typologie en s'inspirant de cette classification de Keim ainsi que du travail d'Andrews (2002), cependant nous ne reprendrons pas tous les points, car

tous ne nous paraissent pas pertinents. De plus, nous proposerons aussi d'autres catégories qui nous paraissent présenter la visualisation sous un angle plus à propos pour notre travail.

Le développement des technologies informatiques faisant de moins en moins obstacle à l'imagination des chercheurs, ingénieurs, etc. il existe de nombreuses autres modes de représentation que ceux que nous allons présenter, ceux-ci étant les principaux. Nous nous contenterons de dresser un panorama de ces techniques en partant des objets à visualiser et nous fournirons pour chaque type de techniques des exemples et illustrations.

3.5.1 Visualisation de structures linéaires (unidimensionnelles)

Nous désignons par le terme *structures linéaires*, les données de type liste alphabétique, résultats d'une recherche de documents, frise chronologique…

Les utilisateurs des systèmes qui visualisent de telles données souhaiteront parfois rechercher plus loin pour des résultats plus spécifiques et d'autres fois voudront des données globales au sujet des éléments de la liste qu'ils visionnent ou encore voudront comparer un élément particulier dans la liste par rapport à d'autres.

L'approche commune à ces problèmes est de fournir des méthodes pour faire défiler de longues listes jusqu'à ce que l'élément désiré soit atteint. Parfois ceci est accompagné d'un système de commande pour produire des étiquettes pour les données, par page ou par ligne, qui facilitent la navigation. Un résultat typique de recherche sur un moteur de recherche sur Internet type *Google*, illustre ces techniques. De telles visualisations de résultats illustrent également l'insuffisance de telles réponses à de nombreux besoins utilisateurs en étant particulièrement mal adaptées à de larges ensembles de données.

D'autres tâches peuvent s'y appliquer comme énumérer le nombre d'éléments de la liste, trouver les éléments comportant certains attributs (*e.g.* éléments modifiés depuis la dernière fois, …) ou encore voir quels sont les mots les plus utilisés dans le premier chapitre d'Alice au Pays des Merveilles (Figure 18), ou bien encore voir un item avec toutes ses caractéristiques…

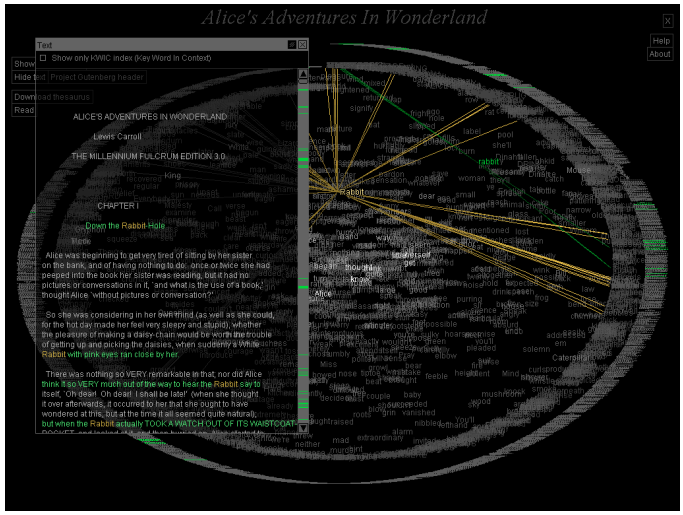


Figure 18 : Le texte intégral d'Alice au pays des merveilles visualisé sur un arc (http://www.textarc.org) .

3.5.2 Visualisation de données bidimensionnelles

Par le terme de données bidimensionnelles, nous désignons toute donnée qui possède deux caractéristiques comme longitude et latitude pour les données géographiques, longueur et largeur pour les photographies, etc. La représentation de données bidimensionnelles se rapporte au placement d'éléments dans un espace à 2 dimensions en fonction de leur position

(gauche, droite…), leur taille, leur distance relative, etc. La représentation de données bidimensionnelles ne doit pas être totalement confondue avec la représentation 2D. En effet, des données tri- ou multidimensionnelles sont parfois représentées dans un espace à 2 dimensions, d'où une ambiguïté et des interprétations à partir des représentations, parfois fausses sur la nature des données.

3.5.3 Visualisation de données multidimensionnelles

Les données possédant trois dimensions sont encore facilement visualisables, notamment à l'aide d'un mode en 3D comme le montre la figure suivante :

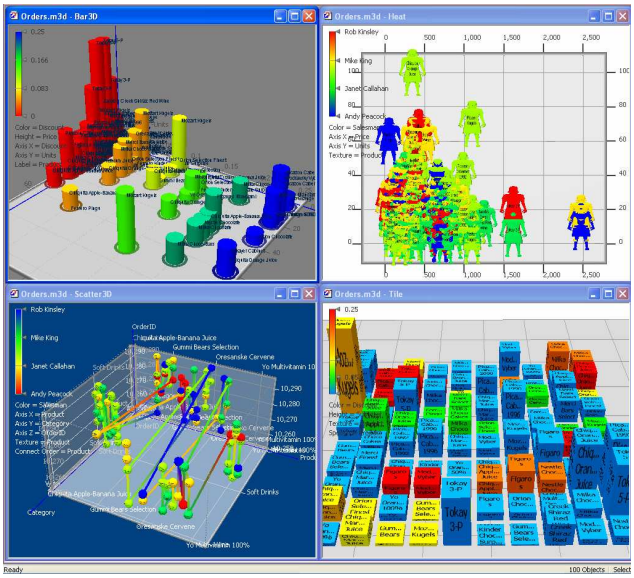


Figure 19 : Affichage en 3D de données (Miner3D)

Au-delà de trois dimensions, la représentation dans un espace devient beaucoup plus délicate, il existe cependant des méthodes que nous allons voir.

Espace 3D augmenté

La notion d'espace 3D augmenté désigne un espace tridimensionnel dans lequel cohabite différentes représentations simultanées des données étudiées. Cette représentation est notamment utilisée en Bourse, par exemple. Il permet d'afficher les données de différentes manières dans un espace unique et réduit permettant d'analyser rapidement la situation.

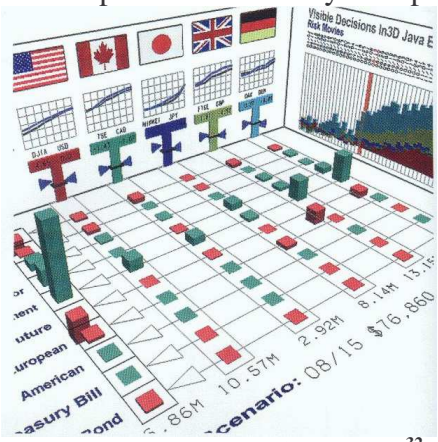


Figure 20 : Espace 3D augmenté en bourse[32]

[32] http://www.infres.enst.fr/~elc/infoviz/toronto-stock/stock3-iv.jpg

L'illustration présente d'abord les données dans un espace en 3 dimensions, puis s'ajoutent d'autres informations sur la représentation, plaquées sur les surfaces. Ce système permet donc à l'utilisateur de visualiser l'ensemble de informations dans un espace réduit.

Table lens

Le principe de la visualisation en « *table lens* » consiste à représenter un tableau de données et à remplacer dans un premier temps les données par des diagrammes afin d'afficher un maximum de données dans un minimum d'espace, puis dans un second temps de permettre de naviguer dans ces données en demandant les détails de certaines entrées de ce tableau.

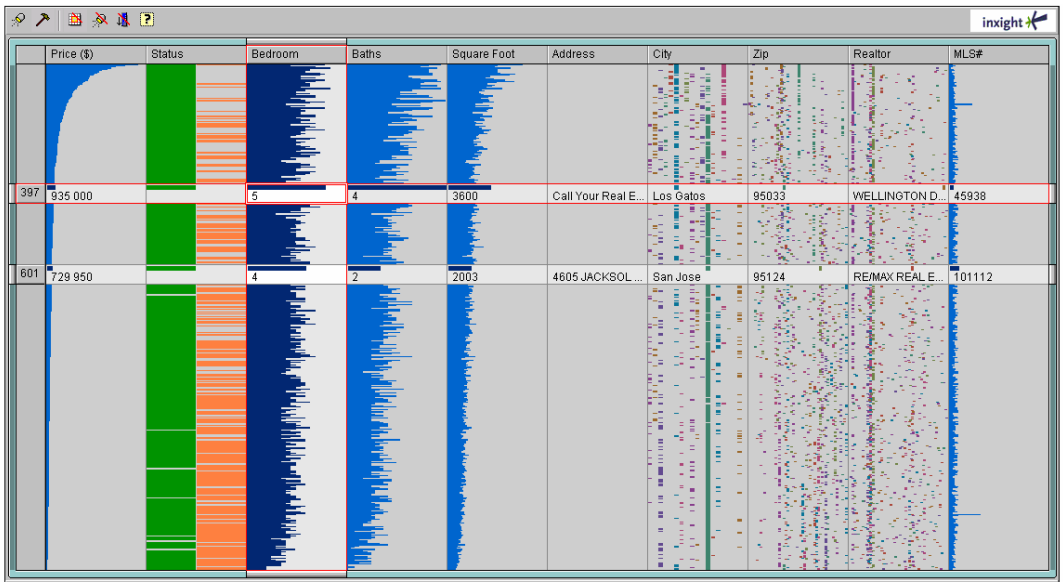


Figure 21 : Exemple de table lens (Vxinsight)

L'illustration ci-dessus nous permet de constater qu'il est possible de visualiser l'intégralité des données sous la forme d'un simple tableau, ces données pouvant être classées par ordre de grandeur pour les nombres, par exemple.

3.5.4 Visualisation de hiérarchies

Les arborescences (*Trees*)

Les hiérarchies, également connues sous le nom de structures arborescentes, sont des collections de nœuds de données où chaque noeud a un parent unique (noeud d'un rang supérieur dans la hiérarchie), mais peut avoir de nombreux enfants de même parent (noeuds de rangs inférieurs dans la hiérarchie). Il est possible d'agir sur un noeud simple, un lien, une collection de noeuds, ou même sur la structure entière.

Les données hiérarchiques sont très diverses et sont produites sous de nombreuses formes. On retrouve ces données dans les taxonomies, les structures d'organismes (organigrammes), la gestion d'espace disque, les généalogies, et des systèmes de classification tels que la classification décimale de Dewey. Les hiérarchies s'utilisent de différentes manières, à savoir : recherche d'un noeud particulier, visualisation un noeud dans son contexte général hiérarchique, examen de la structure et les relations globales de l'arbre, et également recherche des doublons ou des anomalies dans une structure arborescente.

Arborescences hyperboliques

La présentation traditionnelle des hiérarchies s'effectue habituellement à l'aide d'une représentation en 2D où des nœuds-enfant sont placés sous leurs parents. La navigation et la recherche de noeuds spécifiques dans une telle structure peuvent être confuses, désorientantes et assez frustrantes. Des techniques plus récentes de visualisation, comme les arbres hyperboliques, tentent de visualiser le plus de nœuds possible, même si l'arbre complet lui-même ne peut être entièrement visualisé, de façon à fournir des mécanismes pour la navigation et la recherche qui permettent à l'utilisateur de maintenir le contexte de l'arbre entier à l'esprit d'une part et d'autre part afin de réduire le risque de désorientation.

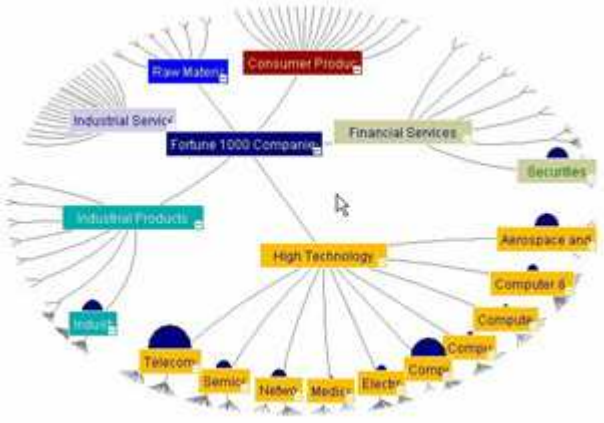


Figure 22 : Arbre hyperbolique (Inxight)

Dans la figure ci-dessus l'utilisateur peut naviguer de nœuds en nœuds, parcourir la hiérarchie ou la descendant ou en la remontant sans perdre le contexte du nœud sur lequel il se focalise.

Les arbres coniques

Cette technique fait partie des systèmes de visualisation en 3D. Un arbre conique, ou *cone-tree*, est une représentation tridimensionnelle d'une hiérarchie dans laquelle on associe à chaque noeud le sommet d'un cône, et on arrange ses fils autour de la base circulaire du cône. C'est ce que nous pouvons voir dans les 2 figures suivantes.

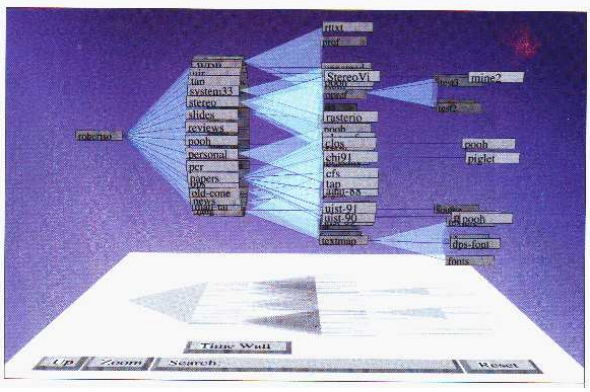


Figure 23 : Arbre conique horizontal (Cat-a-cone, PARC Xerox)

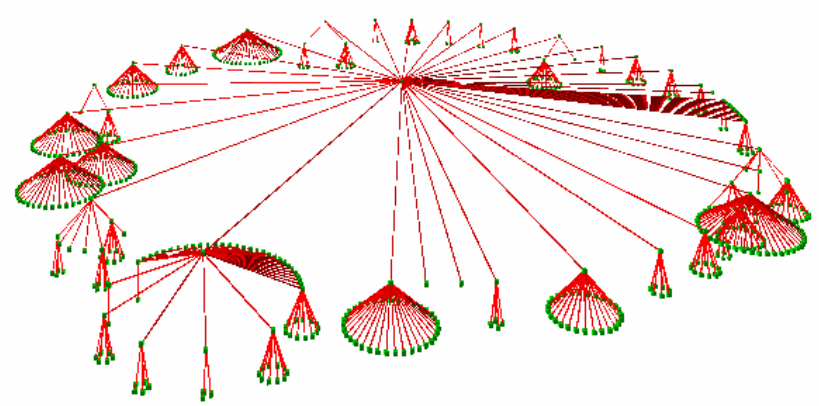


Figure 24 : Arbre conique vertical (PARC Xerox)

Arborescence et « mur en perspectives »

Ce mode de représentation permet de visualiser au centre du mur en perspective (*perspective wall*) la portion de la hiérarchie sur laquelle se focalise l'attention de l'utilisateur et nous permet d'avoir conscience du reste de la hiérarchie. Ce mode de visualisation se rapproche du mode hyperbolique : on focalise sur un élément tout en gardant visible le contexte de cet élément.

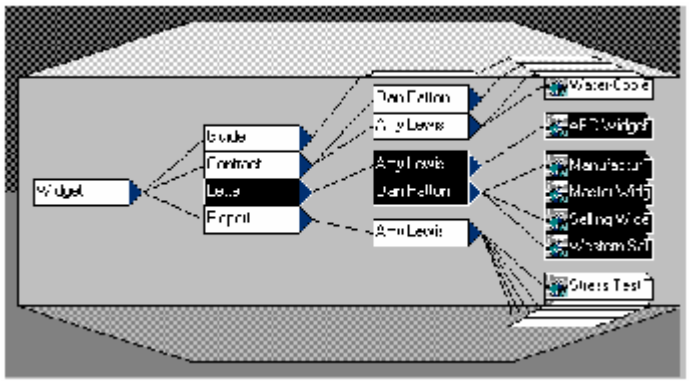


Figure 25 : Une arborescence combinée à un mur en perspective (PARC Xerox)

La notion de mur en perspective est aussi utilisée pour les données temporelles que nous verrons plus tard. Il s'agit ici d'une variation. On peut se demander ce qu'il se passe quand on utilise des hiérarchies profondes. En effet, l'effet de perspective est ici utilisé pour visualiser la largeur de la hiérarchie. Une perspective horizontale supplémentaire pourrait permettre de visualiser la profondeur, mais cela reviendrait à une visualisation hyperbolique telle que décrite auparavant.

Treemap

Le terme *treemap* recouvre les représentations de hiérarchies sous la forme de carte en 2D formant un pavage dont les éléments fils sont imbriqués dans les éléments parents. A la visualisation hiérarchique symbolisée par le rapport d'inclusion, on peut rajouter d'autres dimensions telles que la taille de l'élément fils dans l'élément parent, ou tout autre information signifiée à l'aide d'un code couleur par exemple…

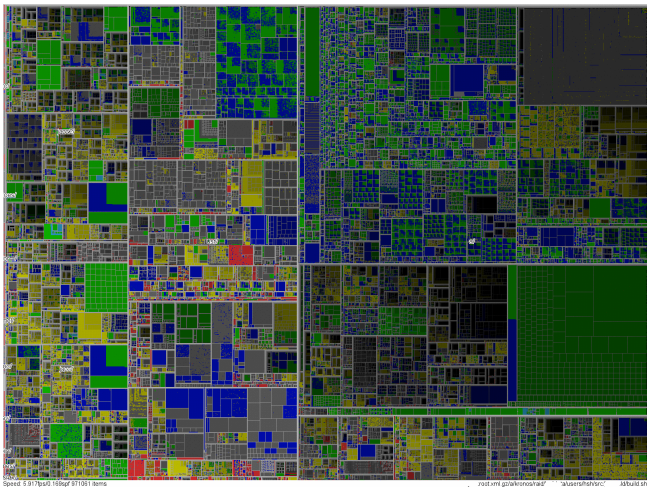


Figure 26 : Visualisation d'un million d'items à l'aide d'une Treemap (MillionVis, Fekete)

La figure ci-dessus montre une représentation standard de *treemap*, avec les frontières (cadres blancs) des différents niveaux hiérarchiques. La sélection d'un cadre permet de descendre d'un niveau hiérarchique et ainsi se rapprocher des éléments terminaux de l'arbre. Les couleurs visibles sur ce système sont des codes attribués par l'utilisateur afin de repérer certaines valeurs.

Visualisation pyramidale

Les informations sont représentées les unes sur les autres. Le socle représente la racine de l'arbre, les éléments sous catégorisés sont accumulés les uns sur les autres selon leur degré de parenté. Cette visualisation est issue des *treemaps*.

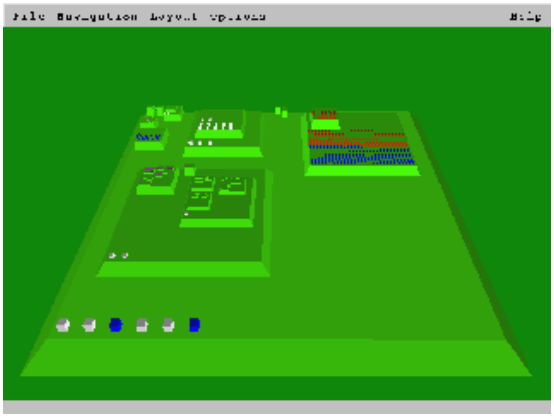


Figure 27 : Représentation pyramidale [Andrews, 97][33]

Visualisation circulaire éclatée (*Sunburst*)

Dans cette métaphore, la hiérarchie est dépliée radialement, le sommet de la hiérarchie étant au centre. Plus on s'éloigne du centre, plus on descend dans la hiérarchie. Les angles et couleurs correspondent à des attributs des données.

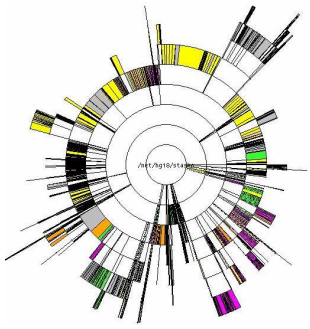


Figure 28 : Sunburst (Georgia Tech, Stasko[34])

[33] http://www2.iicm.edu/keith/papers/vis97/ipyr.html

Visualisation botanique

Cette technique permet de représenter les hiérarchies à l'aide d'un arbre au sens botanique du terme. En effet, le sommet de la hiérarchie se situe à la base de l'arbre (*root*), les éléments sous catégorisés dérivent du tronc par les branches, les éléments finaux sont figurés par les feuilles et les ensemble d'éléments par des fruits.

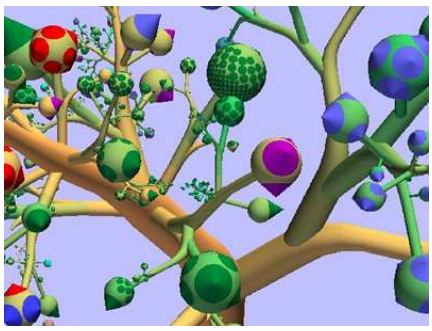


Figure 29 : Visualisation botanique (Eindhoven University of Technology, Kleiberg et al.[35])

3.5.5 Visualisation de réseaux

Le but de la visualisation de l'information en réseau est de faire gagner en perspicacité à la lecture d'une structure qui peut se composer de nombreuses données élémentaires. Les réseaux se composent de noeuds et de liens, un noeud représentant une donnée ponctuelle et un lien représentant un rapport entre deux nœuds. La figure 30 représente les réseaux de voies de communication entre différents lieux géographiques, villes et stations de métro. Un graphique avec peu de points (ou sommets) est facile à dessiner et à comprendre visuellement et actuellement des outils de ce genre sont mis au point pour manipuler de grands ensembles d'information. Ces grands ensembles de données possèdent en effet, des structures invisibles à l'œil nu, enfouies dans les données, des structures au contenu informationnel non négligeable. Déceler une structure ou une hiérarchie parmi un ensemble de points représentant des données n'est pas facile, c'est pourquoi on distingue différentes catégories de réseaux (acycliques, avec racines ou sans racines, orientés et non-orientés…). Ceci permet de développer des algorithmes pour exécuter des tâches sur ces structures telles que trouver les chemins les plus courts ou moins coûteux reliant deux sommets ou traversant le réseau entier.

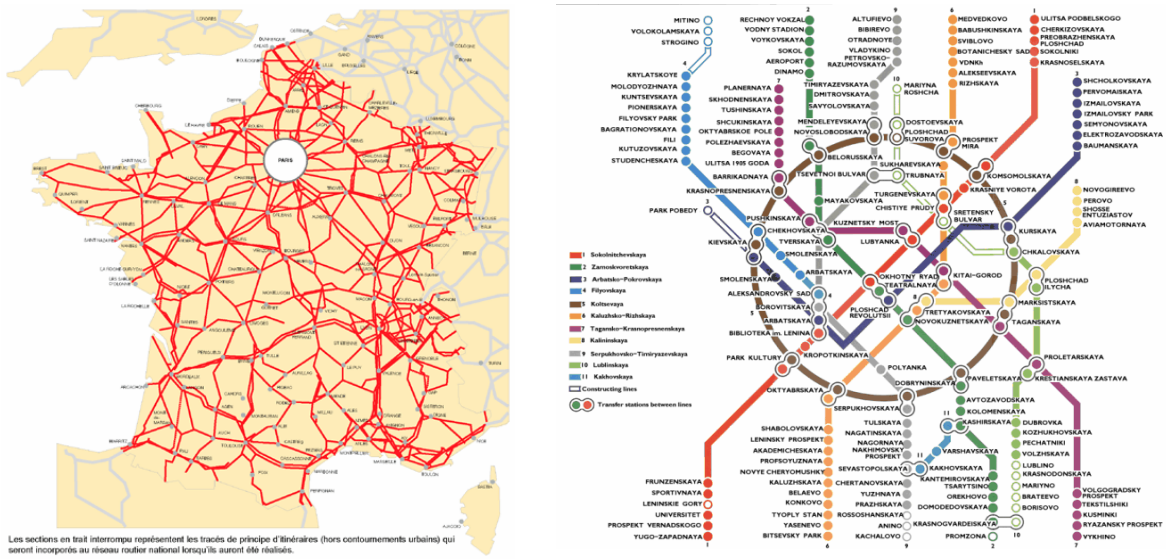


Figure 30 : Deux réseaux : le réseau routier français et le réseau du métro de Moscou

[34] http://www-static.cc.gatech.edu/gvu/ii/sunburst/

[35] KLEIBERG E., Van de WETERING H., and Van WIJK J. J., "*Botanical Visualization of Huge Hierarchies*", InfoVis 2001.

Les domaines d'application de la visualisation de réseau sont les bases de données (à la fois la structure et les éléments d'une base de données), les applications informatiques (raccordement de modules, classes ; raccordement dynamique des processus ; etc.), les réseaux informatiques, le Web mondial à partir des liens hypertextes, les bibliothèques numériques (références, etc.), les Systèmes d'Information Géographique (relations géographiques entre les endroits), les réseaux sociaux, etc...

Réseau de liens hypertextes : cartographier le Web

Ces réseaux, comme celui présenté ci-dessous, représentent les principaux sites pointant sur le site ayant fait l'objet de la requête, ces mêmes sites pouvant eux même pointer vers d'autres sites, etc. suivant le nombre maximum de sommet et le degré de profondeur voulu.

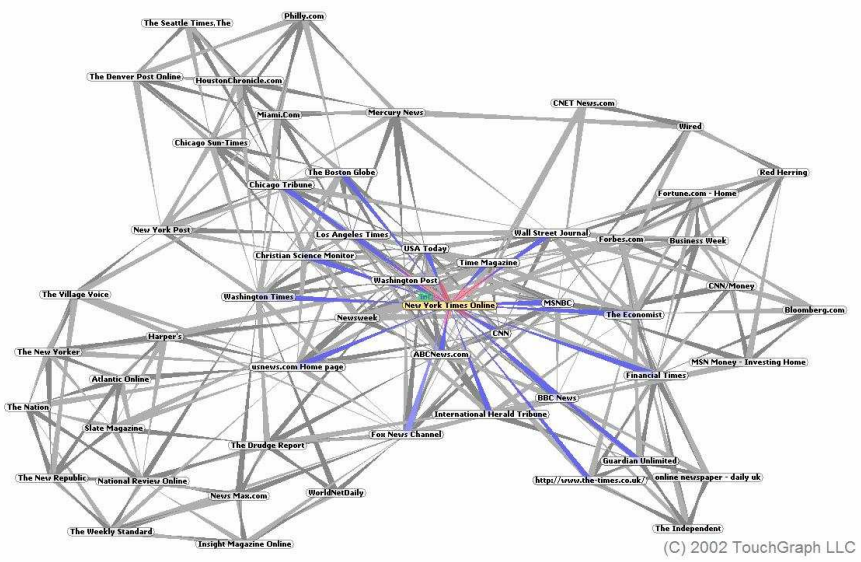


Figure 31 : Réseau de liens hypertextes (Google Touchgaph[36])

Réseau de cooccurrence de termes

A partir de documents structurés (brevets, notices bibliographiques…) ou à partir de documents non structurés (textes bruts), il est possible de représenter les termes extraits à partir des relations de cooccurrence qui existent entre eux (méthode d'analyse des mots associés,…), les sommets représentent les éléments linguistiques (termes, expressions) et les liens les relations pondérées, parfois à l'aide d'une épaisseur de trait significative, parfois à l'aide d'un chiffre indiquant le nombre d'occurrences communes.

[36] http://www.touchgraph.com/TGGoogleBrowser.html [en ligne] (consultée le 3/04/2006)

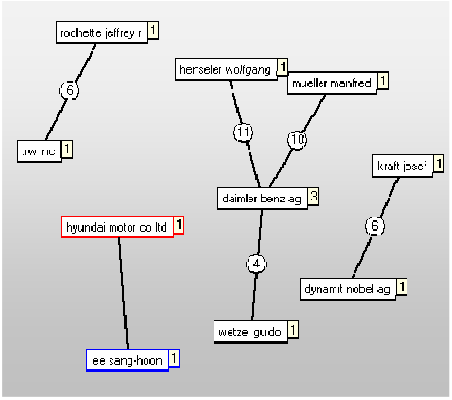


Figure 32 : Réseau de cooccurrence de termes (Matheo Software)

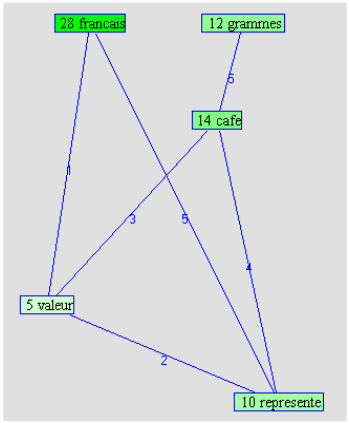


Figure 33 : Réseau de cooccurrence de termes (Wordmapper)

Pour exemple, la carte de Matheo Patent, spécialisé dans l'analyse de brevets, construit des cartes de cooccurrence à partir des différents champs du brevets comme le déposant, l'inventeur, le fabriquant, etc. permettant ainsi de croiser les informations. Ici les arcs sont pondérés par le nombre de documents dans lesquels les termes (sommets) reliés sont en cooccurrence. La seconde carte, tirée du logiciel d'analyse textuelle Wordmapper de la société Grimmersoft, est un détail des classes de mots effectuées par le programme, affichant les cooccurrences des termes, les arcs sont pondérés comme dans le cas précédent, l'occurrence individuelle des termes étant ici indiquée sur l'étiquette du terme.

Réseaux sociaux

Un réseau social désigne non seulement et abstraitement le réseau de connaissances d'un individu, c'est-à-dire toutes les personnes avec qui il est en relation, il désigne de manière plus concrète une carte permettant de visualiser ce réseau en prenant les individus comme sommets du graphes, reliés par des arêtes qui figurent « qui connaît qui ».

Des recherches sont menés afin de construire les réseaux sociaux à partir de l'analyse de pages web personnelles notamment (1) leurs auteurs, (2) les liens entrant, (2) les liens sortant et (4) les listes d'emails permettant de dresser une carte des liens qu'entretiennent les acteurs.

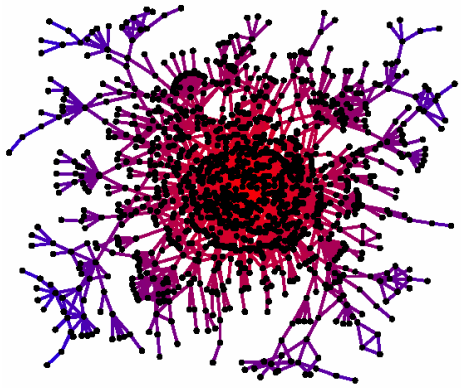


Figure 34 : Réseau social à partir de l'analyse du Web [Adamic et al., 2003]

La figure ci-dessus représente le réseau des auteurs de pages personnelles du domaine « *www.stanford.edu* » liés les uns aux autres par les 4 critères vu auparavant. On perçoit notamment un important cœur de pages en relation très étroite et quelques unes moins connectées aux autres [Adamic et al., 2003].

Cette approche n'est pas sans rappeller la notion de « petit monde » introduite dans les années 60 en psychologie et pour cause puisqu'elle en est probablement l'origine[37]. D'après ce concept, les différents individus seraient connectés les uns aux autres, formant un réseau qu'ils influencent et qui les influencent. En d'autres termes, « je connais quelqu'un qui connaît quelqu'un qui connaît quelqu'un, et ainsi de suite, jusqu'à atteindre n'importe qui sur Terre ».

3.5.6 Visualisation de regroupement de données : les thématiques

Dans le chapitre sur l'analyse de l'information, nous avons parlé de classification, c'est-à-dire l'opération qui consiste à regrouper des données proches ou similaires à l'aide de méthodes automatiques, semi automatiques ou non. L'intérêt d'une telle opération réside bien souvent dans sa visualisation.

Les cartes diagrammes…

A partir des travaux sur les classifications, de nombreuses visualisations de cluster trace une topologie de ces classes à partir des notions de densité et de centralité que nous avons vu notamment au sujet des mots associés. La visualisation est bidimensionnelle, la centralité et la densité assignée respectivement à l'axe des abscisses et à l'axe des ordonnées. Cette technique permet donc d'afficher les classes les unes par rapport aux autres et de pouvoir instantanément les analyser et les comparer. L'outil Wordmapper, place les classes, qu'il détermine automatiquement à partir de l'indexation préalable de texte, en fonction de ces deux axes.

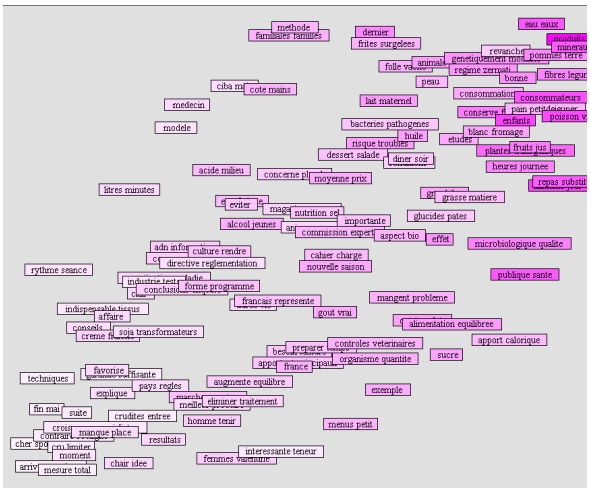


Figure 35 : Visualisation des classes en fonction de leur densité et de leur centralité (Wordmapper)

Il est possible d'y apporter une interprétation à la manière des diagrammes stratégiques de Callon (1993), c'est-à-dire diviser cet espace en quatre quadrant

[37] D'après David Grandadam, Le monde est petit… histoires des réseaux, conférence de l'université L. Pasteur, Strasbourg. 27 octobre 2005.

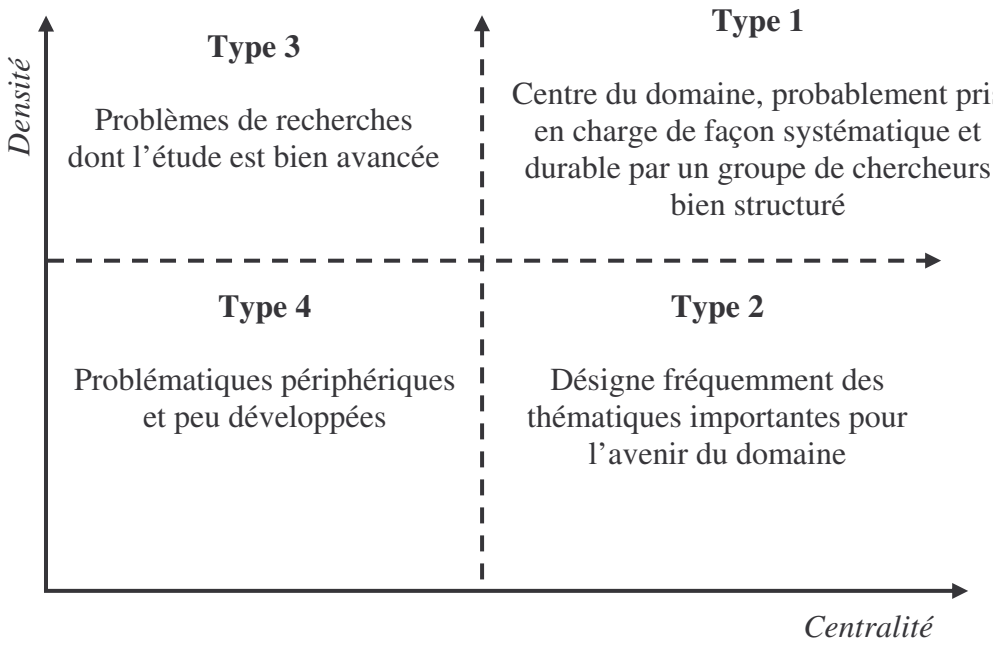


Figure 36 : Diagramme stratégique [Callon, 1993]

Espaces vectoriels

Les approches suivantes, paysages et galaxies de points sont ce que l'on appelle des espaces vectoriels. Selon Nussbaumer (1995) les espaces vectoriels sont souvent utilisés pour caractériser des documents. Les espaces vectoriels sont des espaces pouvant comporter de très nombreuses dimensions représentées par des vecteurs, dans le cas de la représentation de documents, les vecteurs sont les descripteurs. Chaque document est donc un point dans cet espace à *n* dimensions, la similarité entre documents peut être calculée par le produit des 2 vecteurs, pour des informations multidimensionnelles une cartographie en deux dimensions à partir de l'espace multidimensionnel est nécessaire et différentes méthodes existent pour cela, comme la méthode utilisée dans le module Neurodoc de STANALYST qui procède par projection en ACP des clusters résultant d'une analyse par K-means axiales.

Une autre méthode pour les espaces vectoriels fait appel aux cartes SOM de Kohonen. En effet, les données (ou les documents), associées à un vecteur (l'ensemble de leurs caractères), sont positionnées sur le quadrillage du plan. Les vecteurs du plan, situés aux intersections des arêtes, se positionnent par un processus d'apprentissage, se regroupent suivant leur proximité relative. Les collines de la métaphore du paysage ci-dessous, se forment par déformation du maillage initial.

Métaphore du paysage

La métaphore du paysage représente les classes comme des collines, c'est-à-dire des élévations ponctuelles d'un plan proportionnelles au nombre d'éléments que la classe contient.

Des nuances de couleur ou des courbes de niveaux, comme celles figurant sur une carte géographique, permettent à l'utilisateur de percevoir les élévations. Un exemple de métaphore du paysage est celle en 3D du logiciel Vxinsight de Sandia National Laboratories[38], comme le montre l'illustration ci-dessous. Sur celle-ci nous pouvons distinguer les thématiques par leur élévation sur le plan.

[38] SNL, http://www.cs.sandia.gov/projects/VxInsight.html

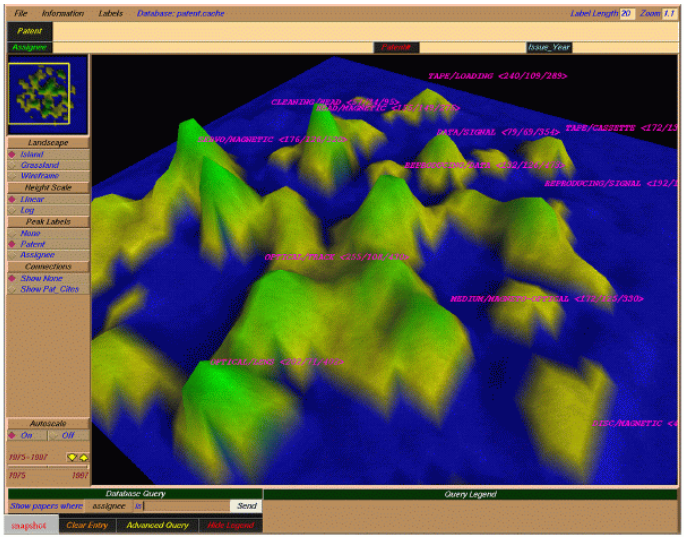


Figure 37 : Métaphore du paysage (Vxinsight, SNL)

Le code couleur employé ici, par exemple, évoque le bleu de la mer pour les zones de basse altitude, le jaune du sable pour les zones intermédiaires entre la mer et la verdure des zones supérieures. Le degré d'élévation est aussi facilement interprétable comme représentative d'une importance particulière. Un des avantages d'une telle représentation est qu'elle est suffisamment intuitive pour permettre à l'utilisateur de saisir rapidement son sens.

D'autres outils permettent cette approche, comme ThemeView du produit IN-SPIRE[39] (Pacific Nothwest National Laboratory), ou encore Themescape de la solution Aureka[40] (Thomson Scientific).

Les galaxies de points

Une autre façon de représenter des regroupements d'éléments, est de représenter chaque élément par un point dans un espace bidimensionnel ou tridimensionnel. Après avoir effectué une classification de ces éléments, des amas de points se forment, que l'on peut appeler galaxies. On peut penser que l'approche est quelque peu similaire à l'approche précédente, mais sur un plan en 2 dimensions. La spécificité de cette visualisation est d'afficher directement le contenu des classes à l'utilisateur, alors que dans les visualisations précédentes, seul l'ensemble était affiché, non les éléments de cet ensemble. Le revers de la médaille étant qu'avec des très grands corpus, la visualisation devient vite totalement illisible.

Comme exemple, nous pouvons citer un outil comme IN-SPIRE, avec son module Galaxy, qui présente cette approche visuelle [Hetzler, 1998].

[39] http://in-spire.pnl.gov/

[40] Aureka, produit de la société MicroPatent, du groupe Thomson, une des principales sources mondiales d'information sur les brevets et les marques du monde entier. http://www.micropat.com/static/advanced.htm

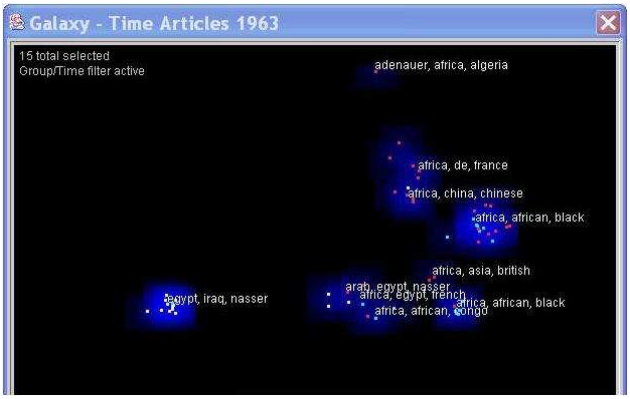


Figure 38 : Représentation en galaxie (IN-SPIRE, PNL)

3.5.7 Visualisation d'informations temporelles

Ayant toujours à l'esprit notre réflexion portant sur la visualisation de l'évolution de thématiques, notre démarche d'état de l'art a été de faire un point sur la visualisation en général, puis un point sur la façon de représenter des données suivant leur type, nous nous sommes ensuite progressivement intéressés aux représentations de thématiques. Pour clore cette approche de la visualisation par les données (3.4) nous allons nous pencher sur la représentation des données temporelles, c'est-à-dire comment celles-ci sont traditionnellement représentées.

Frise chronologique ou *time-line*

Le terme Time-line est ni plus ni moins l'équivalent anglo-saxon du terme français « frise chronologique ». Time-line se réfère donc à une frise chronologique. Cependant, le Times-line possède une acception plus large puisqu'il est repris dans des domaines technologiques pour désigner une représentation qui place des événements parallèles par rapport à un axe temporel absolu. Les Times-lines sont utilisés dans des domaines comme le traitement du son, le montage vidéo, la gestion de projet (diagramme de Gantt), etc. Dans la plupart des systèmes, l'auteur peut placer directement les icônes sur l'axe temporel. De plus la granularité du Time-line peut être adaptée aux besoins.

Les murs en perspective ou « Perspective wall »

Le mur en perspective est une méthode de visualisation bien adaptée pour visualiser des grandes quantités d'informations que l'on peut ordonner selon un critère suivant l'axe des X. Cette méthode semble bien adaptée pour représenter des informations temporelles, dans ce cas le critère utilisé est le temps. L'idée à la base de ce concept est d'utiliser une vue en perspective pour palier les limitations physiques imposées par les tailles des fenêtres d'un système de fenêtrage classique. De cette manière, l'auteur a conscience de la totalité des informations.

La figure 39 présente des murs en perspective, l'un réalisé par Xerox pour visualiser des informations classées chronologiquement. Le mur, plus récent, réalisé par la société Xinsight[41] permet de visualiser les films selon l'ordre chronologique de leur sortie. L'axe vertical représentant les sociétés de production et les couleurs les genres de films.

[41] http://www.inxight.com/products/sdks/tw/ (consulté le 10/06/06)

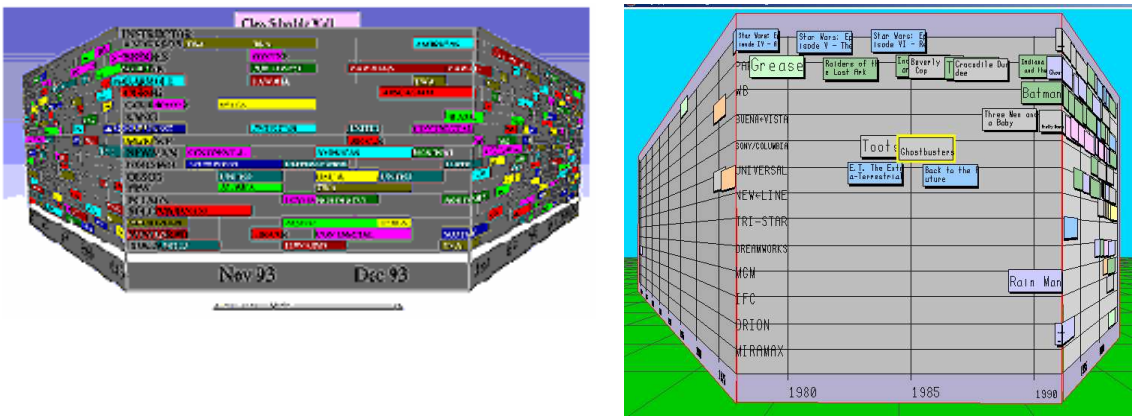


Figure 39 : Perspective walls (PARC Xerox (gauche) et Inxight (droite))

Le système Lifestreams

Lifestream est un programme développé initialement à l'université de Yale (USA). Il utilise une de métaphore de l'espace de travail (*desktop metaphor*42) comme support à la visualisation.

Un *lifestream* est un flux de documents chronologiquement ordonné qui fonctionne comme journal intime de la « vie électronique » de l'utilisateur; chaque document qu'il crée et ceux envoyé par d'autres est stocké dans son *lifestream*. La longueur du flux contient des documents du passé. Se déplaçant du bout de la queue et vers le présent, le flux contient des documents plus récents (document en court ou nouveau courrier électronique) ; d'autres documents (images, correspondance, factures, films, messages vocaux, programmes) sont stockés dans l'intervalle. Au delà du présent et dans le futur, le flux contient des documents dont aura besoin l'utilisateur : rappels, articles de calendrier, listes de choses à faire[43]. Ce système n'est autre qu'une frise chronologique avec un rendu en 3D.

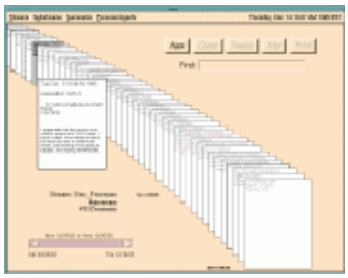


Figure 40 : Lifestreams [Freeman, 1995]

3.6 Visualiser et suivre l'évolution de l'IST

Nous l'avons vu au chapitre sur la veille, suivre l'évolution de l'information pour la détection de tendance est un aspect important des activités de surveillance de l'environnement. Il existe pour cela différentes techniques selon le type d'information surveillée. En ce qui concerne notre travail, nous avons voulu connaître quel est l'état des recherches sur les techniques employées pour retracer l'évolution de l'information issues de bases de données, et notamment comment rendre compte de l'évolution de thématiques. Nous avons choisi de présenter cet état de l'art en fonction de la façon de visualiser cette évolution.

Beaucoup de travaux portant sur ce problème sont orientés par l'analyse de co-citations [Chen, 2006], [Small, 1999], moins sur le contenu des documents. Bien que l'algorithme de classification incrémentale auquel s'associe notre étude soit plutôt dans la seconde catégorie, nous présenterons aussi les travaux issus de la première.

---

[42] Métaphore souvent utilisée pour développer une nouvelle forme d'accès aux documents et d'optimisation du travail dans un environnement multidocuments et donc multifenêtre, dont il n'est pas question ici

[43] http://www.cs.yale.edu/homes/freeman/lifestreams.html (consultée le 12/04/2006)

3.6.1 Evolution dans les « Maps of science »

« *En examinant des séquences chronologiques de cartes, nous pouvons observer de quelle façon la connaissance scientifique avance* » [Garfield, 1986]. Les études menées par Garfield et Small dans les années 80, portant sur la cartographie des champs de recherches scientifiques, se sont penchées sur la question de l'évolution de ces connaissances.

La technique employée alors dans SCI-Map (Science Citation Index – Map) est de comparer des cartes à différents instants et de pouvoir distinguer des mouvements perceptibles. A partir de là, Small (1999) identifie des phénomènes impliqués par les changements observés, tels que ce qu'il appelle une révolution, c'est-à-dire lorsqu'un changement soudain se produit (scission, croissance, etc.), une croissance rapide d'un champs révèle une émergence, une naissance ou encore une dérivation à partir d'autres champs.

3.6.2 Evolution dans les diagrammes stratégiques de Callon

Une approche proposée par Cahlik (2000) consiste à représenter l'évolution des thématiques sur le plan d'un diagramme stratégique (voir [Callon, 1993]). En effet, la proposition consiste à décrire le mouvement des thématiques en fonction du temps au sein des quatre quadrants que nous avons décrits plus haut, permettant d'interpréter facilement la position de chaque thème. Cahlik propose de visualiser cette évolution par périodes à l'aide de graphes orientés représentants les différents états du cluster et des mots clés qui le constitue à ces différentes périodes, les arcs sont pondérés par le nombre de mots clés transitant ou non d'un état à un autre. Sa méthode est implémentée dans un programme nommée LEXIDYN.

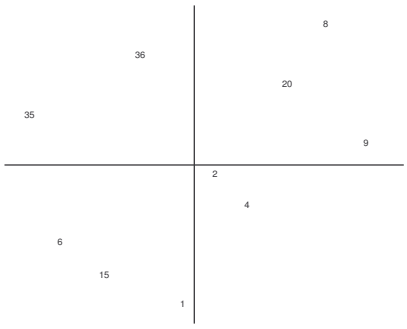


Figure 41 : Diagramme stratégique période 1

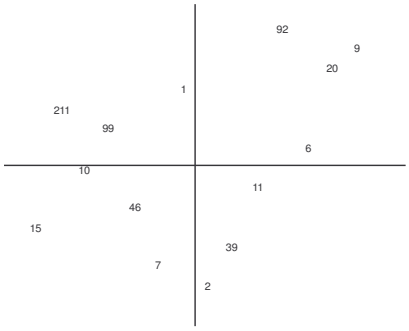


Figure 42 : Diagramme stratégique période 5

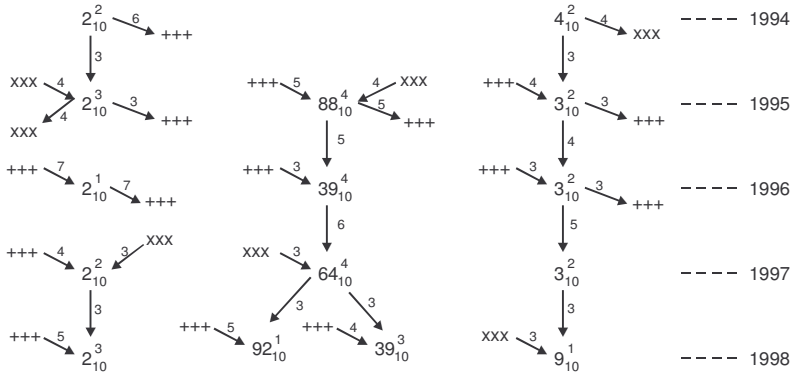


Figure 43 : Informations détaillées sur l'évolution des classes dans les quadrants et évolution des mots clés dans les classes

3.6.3 Tracés simultanés d'évolution sur des axes temporels

Lorsque l'on parle de temporalité, l'idée venant le plus rapidement à l'esprit est l'utilisation d'un axe temporel, idée de base notamment des frises chronologiques, le long de laquelle sont placés les événements.

De ce fait, de nombreuses méthodes de visualisations sont basées sur ce principe. Havre et al. (2002) proposent à travers un outil appelé ThemeRiver, de visualiser l'évolution linéaire des informations à l'aide de la métaphore d'un rivière s'écoulant sur un axe de temps horizontal (voir section suivante).

Erten et al. (2003) proposent une analyse verticale de l'évolution de catégories de documents (classement de l'ACM, Association for Computing Machinery) dont le volume évolue dans le temps. Dans l'outil TGRIP (Temporal Graph dRawing with Intelligent Placement), ces catégories, représentées par des sphères dont le volume est proportionnel et changeant en fonction du volume des catégories, sont reliées entre elles d'une part et d'autre part avec elle-même d'un instant à l'autre (afin de faciliter leur repérage lors d'un changement de période).

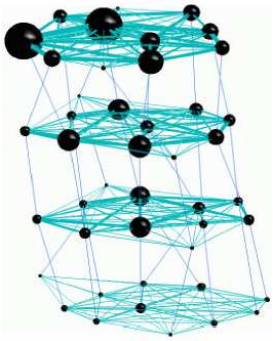


Figure 44 : visualisation de catégories thématiques par [Erten et al., 03]

Chen (2006) propose de visualiser l'évolution de réseaux de citation le long d'un axe temporel (Figure 45). Dans son outil appelé Citespace II Les articles sont représentés par des nœuds liés aux articles cités ou articles citant. Chaque nœud possède une taille caractéristique du nombre de fois qu'il a été cité et d'un jeu de couleurs caractéristique de l'année à laquelle l'article a été cité. Ainsi les nœuds sont constitués de strates temporelles de taille variable en fonction du nombre de citation pour une année. L'ensemble des nœuds est visualisé sur l'axe temporel et permet de distinguer, dans ce cas les fronts de recherche actuels (dominante rouge) des bases intellectuelles plus anciennes sur laquelle se basent les fronts de recherche (dominante verte) ainsi que les dérivations (appelées points pivots, représentés en violet). On remarque cependant que la représentation devient vite illisible, car vite surchargée.

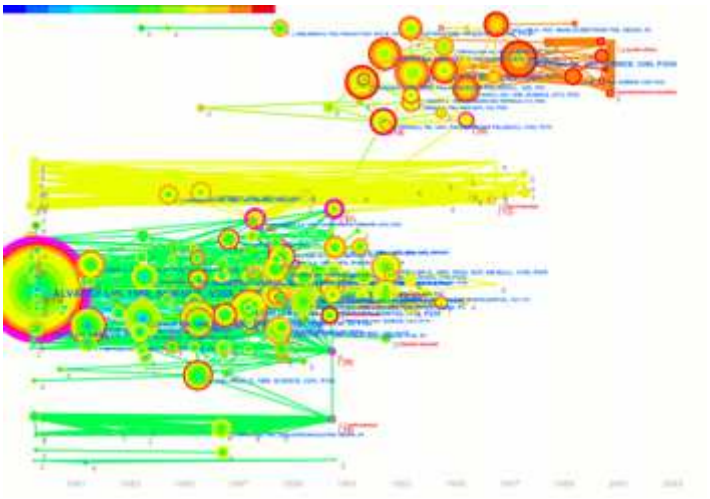


Figure 45 : représentation de réseau de co-citation (Citespace II, [Chen, 2006])

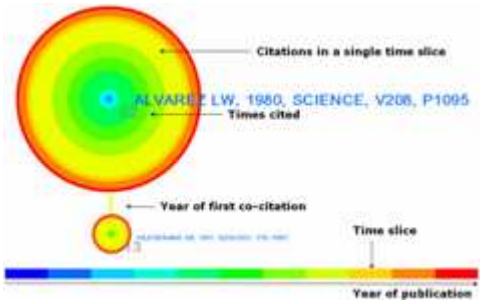


Figure 46 : Vue d'un noeud (article) dans Citespace II

3.6.4 Représentations évoluées sur graphiques et diagrammes

L'utilisation de graphiques et de diagrammes reste une technique très présente. En effet, la lecture de courbes d'occurrence de termes en fonction du temps par exemple, est assez aisée et compréhensible pour l'utilisateur et ces courbes représentent un moindre développement de la part des concepteurs.

Nous avons repéré une variante intéressante, du point de vue visualisation, des diagrammes utilisant la métaphore d'une rivière, appelée ThemeRiver. Développé par le PNNL (Californie, USA), le programme a pour objectif de représenter l'évolution de thématiques issues de l'analyse de discours mais nous pensons que cela peut être élargie à d'autres sources. Le principe est de construire, en suivant un axe de temps horizontal, des histogrammes cumulés de fréquences de termes (représentatifs de thématiques) et de lier ces histogrammes d'un instant à l'autre. Le tout représentant alors le flot d'une rivière de thèmes, s'écoulant et grossissant au gré des occurrences [Havre et al., 2002].

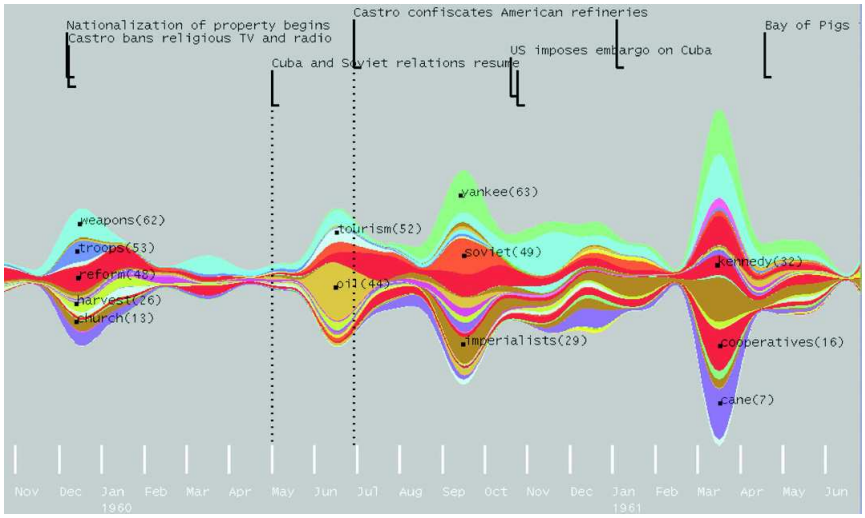


Figure 47 : Etude diachronique de thématiques avec ThemeRiver [Havre et al., 2002]

Une autre représentation est celle de l'outil Gapminder[44]. Ce programme permet de visualiser le développement de nombreux pays sous les axes tels que l'espérance de vie, la

[44] The Gapminder World 2006. http://tools.google.com/gapminder/ [en ligne] (consulté le 7/06/2006)

croissance économique, le nombre d'utilisateurs d'Internet, le budget militaire, etc. en fonction du temps. Ainsi pour chaque pays, il est possible de voir l'évolution de ces axes sur une période donnée. Sur la figure 48 on remarque le mode d'interaction pour cette représentation dynamique : L'utilisateur développe l'évolution à l'aide d'un curseur, il peut revenir en arrière, figer la représentation, etc.

Figure 48 : Représentation dynamique de données par Gapminder

3.6.5 Découper le continuum temporel : les « time slices » ou intervalles de temps

Une autre méthode assez répandue pour permettre à l'utilisateur de percevoir des changements dans les données, consiste à considérer le temps comme une dimension des données (date d'enregistrement, date du documents… cela dépend des données et de la manière de les gérer) et de reprendre la technique de filtrage dynamique vue précédemment, en plaçant des filtres temporels, c'est-à-dire, fixer un seuil minimum et un seuil maximum de temps, entre lesquels les infos datées comprises entre ces deux bornes apparaîtront. L'utilisateur peut alors visualiser les données qui l'intéressent dans une période donnée, voir se construire des regroupements au fil du temps, etc. Pour illustrer cette méthode, l'outil « time slice » de la plateforme *IN-SPIRE* permet de déterminer l'intervalle de temps à étudier, de voir apparaître les documents concernés et de mener une rapide analyse à l'aide de diagrammes sur l'importance des classes suivant leur date. L'outil *Vxinsight* (Figure 37), quant à lui, offre la possibilité de déterminer un intervalle de temps et de manipuler cet intervalle librement afin de percevoir les modifications du paysage. Pour cela l'utilisateur interagit à l'aide d'un curseur (*time slider*) placé le long d'une ligne de temps bornée par la date minimale (MIN) et maximale (MAX) de l'ensemble des éléments visualisés. A l'aide de ce curseur, l'utilisateur détermine son intervalle de temps ([Umin, Umax]), qu'il peut manipuler à sa guise le long de l'axe.

Figure 49 : Moyen d'interaction temporelle utilisé par Vxinsight

L'avantage d'une telle interaction est que l'utilisateur peut agir à sa guise sur la visualisation et sur l'évolution des éléments qu'il visualise, contrairement à une animation simple, elle lui permet de visualiser à son rythme, de revenir en arrière puis en avant, etc. Ce qui peut s'avérer utile pour percevoir des mouvements parfois minimes.

La notion d'intervalle temporel renvoie à deux aspects : celui de l'interaction et celui de la granularité de la visualisation. En effet, l'utilisateur peut agir sur les données à l'aide d'une sélection par critères temporels (interaction dynamique), mais l'expression anglaise, « time-slice » désigne aussi la granularité temporelle de la visualisation, c'est-à-dire la graduation des axes de temps.

3.6.6 Visualiser les changements : les analyses par différences

Une autre façon d'analyser d'une part et de représenter l'évolution des informations d'autre part est ce que nous avons choisi d'appeler l'analyse par différences.

L'approche de Erten et al. (2003) que nous avons déjà vu plus haut, propose une représentation des changements en faisant varier la taille des catégories en fonction du pourcentage de changements calculés d'un intervalle à un autre. Ainsi on perçoit rapidement les catégories ayant subies une évolution ainsi que la nature (croissance ou décroissance) et le degré de cette évolution.

D'un autre point de vue, l'outil développé par Ganascia et al. (2004), EDITE MEDITE (Etude Diachronique et Interprétative du Travail de l'Ecrivain) permet de détecter les transformations, les insertions, les suppressions et les remplacements appliqués à un texte d'auteur littéraire à l'état de brouillon. Sa méthode de comparaison de versions, sans parler de son algorithme, présente les modifications d'une part en surlignant de différentes couleurs les parties modifiées, rajoutée, etc. et en soulignant les déplacements, d'autre part en présentant à l'utilisateur un historique textuel des modifications visible simultanément avec les versions de texte elles mêmes.

Nous avons pu observer aussi cette approche par différentiation sur le logiciel Wordmapper, qui propose une analyse diachronique de l'information contenue dans deux textes. Ici, le logiciel propose de comparer deux textes ou deux corpus de textes et de visualiser les classes émergentes de cette comparaison (qu'il réalise par analyse de cooccurrence), c'est-à-dire les classes qui apparaissent dans l'un des deux textes et qui n'étaient pas présentes dans l'autre. Pour plus d'information nous renvoyons le lecteur à la partie 2 concernant les observations que nous avons effectué sur des logiciels de veille.

3.7 La visualisation de l'information : approche utilisateur

Dans un certaine mesure, le concept de visualisation de l'information peut être considéré comme une approche utilisateur car elle n'a de sens que dans la simplification qu'elle apporte à l'utilisateur dans sa démarche d'exploration et de recherche d'information dans de grands ensembles de données. Cependant, nous nous sommes demandé si les techniques employées répondaient aux besoins des utilisateurs et si elles comportaient des limites. Nous avons voulu savoir aussi comment l'utilisateur peut interagir avec un système de visualisation.

3.7.1 Interface utilisateur et visualisation : techniques d'interaction

C'est l'interaction qui rend possible l'exploitation réelle des vues d'ensembles une fois produite. En effet, l'être humain est particulièrement habile à extraire des informations d'un environnement qu'il contrôle directement et activement par rapport à un environnement qu'il ne peut qu'observer de manière passive [Hascoët, 2004]. Selon l'approche psychologique de la perception, le perception est indissociable de l'action : il faut agir pour percevoir et il faut percevoir pour agir. On parle de couplage (ou boucle) action-perception.

Shneiderman (2005) propose un mantra pour caractériser comment l'utilisateur interagit avec la visualisation : *Overview, Zoom-in and details on demand*. Il existe pour cela des techniques classiques d'interaction, comme les barres de défilement pour naviguer dans le contenu d'une fenêtre, les boites de dialogues pour spécifier les paramètres de la visualisation… mais aussi d'autres techniques pouvant être mises en œuvre [Hascoët, 2004], [Shneiderman, 2005].

Vue d'ensemble et zoom

Les utilisateurs gagnent à percevoir les collections entières de documents ou de données dans leur ensemble. Des études en perception visuelle ont en effet montrée que l'être humain appréhendait mieux les informations en allant du général au particulier, il a une perception d'abord globale d'un scène avant de porter son attention aux détails [Hascoët, 2004][45].

Le zoom est une façon de concilier la vue globale et de permettre aussi aux utilisateurs d'accéder aux détails. Un zoom avant révèle ces détails tandis qu'un zoom arrière révèle le contexte. On peut distinguer 2 types de zoom : le zoom infini et le zoom sémantique. Le premier est un changement d'échelle sur la représentation qui est alors considéré comme une image, ce mode est peu intéressant dans le cas de la visualisation de l'information. Le second en revanche l'est beaucoup plus car chaque niveau de zoom correspond à un niveau différent de la hiérarchie.

Détails sur demande

L'utilisateur peut sélectionner un élément ou un groupe d'éléments et interroger ces éléments sur leurs caractéristiques. La méthode la plus employée est de cliquer sur l'élément en question, les détails apparaissant dans une fenêtre ou un *pop up*[46]. Cette possibilité donnée à l'utilisateur est très importante car elle permet à la visualisation de remplir son rôle simplifiant, au regard de l'utilisateur, la complexité des grands entrepôts de données tout en gardant disponibles les informations concernant les individus en tant qu'ensembles de caractères. La figure 52 permet d'illustrer notre propos. En effet, l'interface se compose ici de 4 éléments fonctionnels : les panneaux de contrôles (pour l'interaction), la légende (pour l'interprétation), la visualisation des données proprement dite et une fenêtre de détails pour un élément (pour l'information). Cette fenêtre est apparue après avoir sélectionné un élément de la visualisation, pour plus d'information à son sujet.

Fish eye

La technique du fisheye (œil de poisson) s'apparente sur certains points à la technique de zoom, puisqu'elle permet de focaliser sur des éléments précis d'un ensemble en donnant à l'utilisateur la possibilité de toujours garder le contexte en vue. Il est en quelque sorte un zoome local, une loupe posée à un endroit précis. Le fisheye peut être confondu avec les arbres hyperboliques qui utilisent la même technique, à la différence que le fisheye peut

[45] REYNOLDS G.S., *Psychology today*, American psychology association, 1977, cité dans [Hascoët, 2004]
[46] Le terme *pop up* désigne une fenêtre contenant un message, apparaissant subitement (*pop*) sur l'écran (*up*)

s'appliquer à tous types de données voire même à tout types d'affichage (fisheye sur un tableau de données, sur un graphe, etc.) alors que les arbres hyperboliques ne s'appliquent qu'à des données hiérarchiques d'une part et d'autre part le contexte n'est pas toujours visible. Leur point commun réside finalement dans le calcul de ce que l'on appelle le degré d'intérêt (DOI : Degree Of Interest).

Les figures suivantes illustrent l'approche fisheye sur deux applications différentes : une liste et un tableur. On voit sur ces deux représentations que le fisheye évolue sous le curseur manipulé par l'utilisateur ce qui lui permet de glisser librement le fisheye sur la représentation.

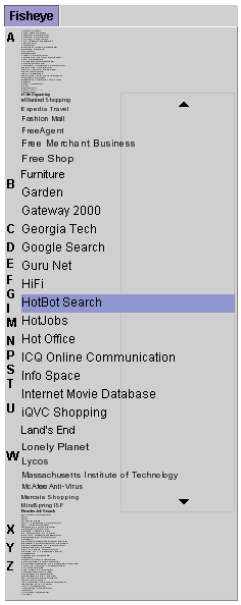


Figure 50 : Fish-eye sur menu déroulant (HCIL, B. Bederson)

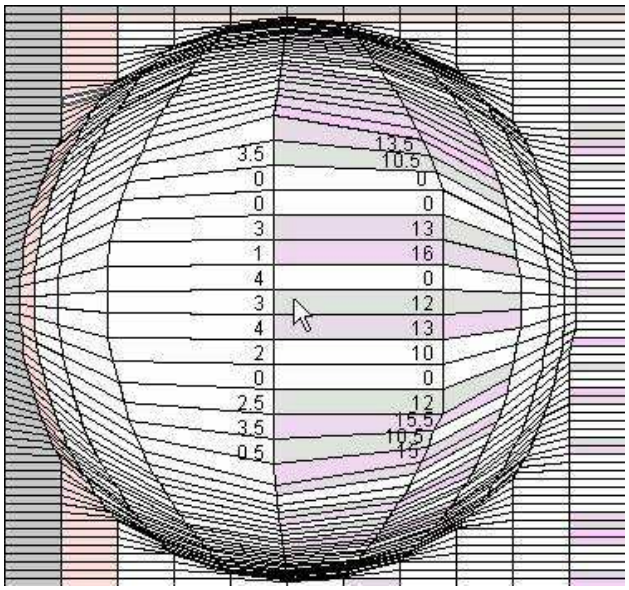


Figure 51 : Fish-eye sur un tableur (Projet FiCell, F.Vernier, IIHM/CLIPS/IMAG/Grenoble1)

Requête dynamique

La requête dynamique possède comme principale caractéristique de permettre à l'utilisateur de filtrer dynamiquement les données qui lui sont présentées sur un espace à deux dimensions. La représentation graphique réagit instantanément aux différentes actions de l'utilisateur qui filtre à l'aide de barres de défilement, de boutons, coches, et autres moyens de contrôles qui agissent sur les couleurs appliquées aux données, sur leur apparition ou disparition, etc. Ce système permet donc de sélectionner les données suivant leurs différents attributs.

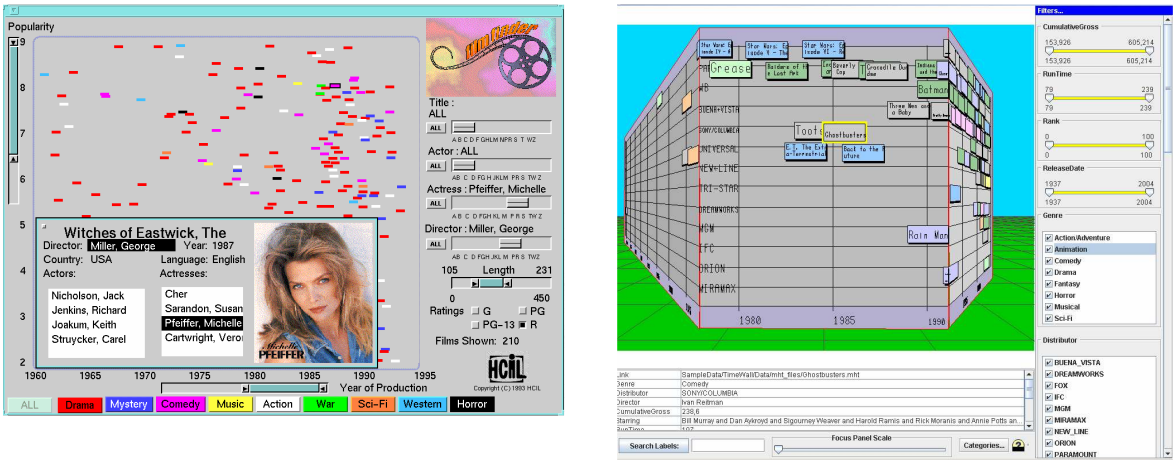


Figure 52 : Exemple de requête dynamique :
Spotfire (HCIL, [Shneiderman, 2005]) et TimeWall (Inxight)

Les figures précédentes présentent un système de requête dynamique très semblable pour deux visualisations différentes. On y distingue les panneaux de contrôle autour de la zone de visualisation qui permettent à l'utilisateur de formuler sa requête en filtrant les éléments visualisés suivants certains caractères : par genres de films, acteurs, récompenses reçues, années de réalisation, durée, etc.

3.7.2 Interface utilisateur et visualisation : tâches utilisateur

En plus de proposer des modes d'interaction à l'utilisateur pour répondre à son besoin de manipuler les données qui lui sont présentées, les systèmes doivent pouvoir répondre aux besoins liés aux tâches de l'utilisateur, mais aussi lui facilitant sa démarche de recherche d'information.

Historisation

Les utilisateurs peuvent conserver un historique des actions menées sur la visualisation afin de permettre de revenir en arrière, de recommencer l'action et d'affiner progressivement sa requête. Il est rare pour un utilisateur qu'une seule action lui fournisse tout ce qu'il souhaite. La visualisation permet une démarche d'exploration des données/informations, cette démarche s'effectue nécessairement en plusieurs étapes et il est important pour lui de garder une trace des étapes franchies.

Exportation d'information

L'utilisateur peut vouloir récupérer des informations qu'il consulte afin de les sauvegarder, de les diffuser ou de les utiliser pour une présentation. S'il les diffuse par exemple, il voudra les mettre à disposition des autres personnes en leur simplifiant la tâche, sans qu'ils aient à manipuler l'outil de visualisation. Pour cela l'outil doit pouvoir permettre d'exporter les données, les cartes, etc.

3.7.3 L'évaluation des techniques de visualisation : les tests d'utilisabilité

Nous venons de le voir, des techniques d'interaction pour les systèmes de visualisation existent et des travaux qui visent à enrichir ces techniques d'interaction homme machine (IHM) se poursuivent. Nous nous sommes demandés s'il existait des limites aux techniques

d'interaction et de visualisation. Atteignent-elles toujours leurs objectifs ? Aident-elles sans faillir l'utilisateur lors de son utilisation ? Facilitent-elles sa perception ? Afin de trouver des éléments de réponse, nous nous sommes penchés sur l'évaluation des systèmes de visualisation de l'information qui pourrait, par sa démarche critique, soulever les problèmes et les limites de l'approche.

D'après Fekete (2005), « *l'intérêt pour l'évaluation de la visualisation de l'information est récent. Depuis une dizaine d'année, le domaine d'où est issue la visualisation d'information (l'IHM) a déjà mis au point plusieurs méthodes d'évaluation […] Elles sont essentiellement issues de la psychologie mais aussi de la sociologie, ainsi que d'autre domaines de l'informatique plus théorique […] Des méthodes de conception et de validation ergonomiques ont essayé d'intégrer la visualisation d'information.* »

Pour Hearst (1998), l'évaluation de ce type d'interface informatique peut se présenter de trois manières :

- On teste l'utilisabilité des systèmes et des représentations (Quels sont les objectifs du système, les techniques employées et les résultats obtenus ?),
- On confronte les systèmes aux interprétations et opinions des experts du domaine de l'information analysée et représentée,
- On mène des expériences contrôlées psychologiquement orientées (l'outil est testé par le participant et on conduit le test de manière à vérifier une hypothèse).

Tricot et Roche [Tricot et Roche, 2006] ont réalisé une expérience mettant l'utilisateur dans une situation de recherche d'information à partir d'un système de visualisation de données hiérarchiques, devant donc y effectuer certaines tâches. L'évaluation est axée sur les aspects suivants : la spatialisation des concepts, l'association d'information aux concepts, l'interaction et la navigation.

Suivant le type de visualisation testé, il en est ressorti les points suivants :

- Concernant les arborescences simples, les utilisateurs ont l'habitude de les utiliser (l'explorateur de fichiers informatiques est, par exemple, une arborescence utilisée quasi quotidiennement) et elles ne posent pas de problèmes tant que le volume de données n'est pas trop important. Au-delà d'une certaine quantité d'information, les arborescences se révèlent relativement inefficaces.
- Les arbres de cônes en 3D ont posé des problèmes à l'utilisateur qui n'a pas l'habitude de naviguer dans un espace en 3D, des effets d'occlusion, par exemple, perdent l'utilisateur et lui demande souvent un effort cognitif plus élevé et une prise en main plus longue.
- Les arbres hyperboliques ont leurs limites aux frontières de la visualisation, lorsque les étiquettes des éléments se chevauchent (problème similaire à l'occlusion du cas précédent). Le comportement de ces étiquettes en bordure pose aussi problème, car elle sont parfois imprévisibles : l'effet de zoom typique de la visualisation hyperbolique accentue les mouvements des objets lointains.

La visualisation de l'information ayant pour objectif de faciliter l'utilisation de grands gisements d'informations, possède donc ses limites. Bien que leur manifestation soit différente suivant le type de visualisation utilisé, les problèmes se posent régulièrement sur la capacité des systèmes à combiner l'aspect technologique, parfois novateur, et l'aspect

utilisabilité, souvent négligé ou inadapté. C'est pourquoi, dans notre étude, il nous a semblé nécessaire de partir de l'utilisateur afin de définir ses besoins et ses attentes du système.

La question des besoins et des attentes utilisateurs nous portent à évoquer les travaux de Bonnel et Chevalier (2006) et de Fekete et Plaisant (2004). Les premiers, à partir d'une étude portant sur ce qu'ils ont appelé les Interfaces Utilisateur d'Information (IUI), sont amenés à conclure que pour qu'un système de recherche rende les meilleurs services aux usagers, il est nécessaire qu'il propose différents Interfaces Utilisateur d'Information. Le choix de l'IUI peut alors être, soit proposé explicitement à l'utilisateur qui choisit en fonction de la tâche à réaliser, soit pris en compte dans l'adaptabilité automatique ou semi-automatique du système par rapport aux données à visualiser (ou aussi par rapport à sa connaissance du profil utilisateur). Plusieurs IUI pour un SRI permettent alors de répondre aux besoins aussi variés que sont les usagers [Bonnel, et Chevalier, 2006]. Les seconds, partent d'une compétition de système de visualisation de l'information organisée chaque année par les universités américaines, nommée *Infovis Contest*, et tirent des leçons de cette compétition au cours de laquelle le meilleur système est récompensé. L'une de ces conclusions est qu'il est possible que, pour des données complexes, une seule visualisation ne permette pas de répondre à toutes les questions et que les réponses ne puissent venir que d'analyses supplémentaires ou d'analyses plus sophistiquées [Fekete et Plaisant, 04].

Ces considérations liées à l'utilisabilité des systèmes de visualisation, nous permettent d'évoquer le travail de recherche de Dâassi (1999) du laboratoire CLIP/IMAG de Grenoble, portant sur la visualisation de données temporelles, dans lequel l'auteur suggère que les techniques d'interaction doivent prendre en compte :

- La nature des données
- La grande taille de l'espace de données à visualiser
- La manipulation des données par l'utilisateur pour spécifier une nouvelle requête à partir de ce résultat (boucle d'interaction que nous avons déjà évoqué)

Il propose alors le schéma d'approche suivant :

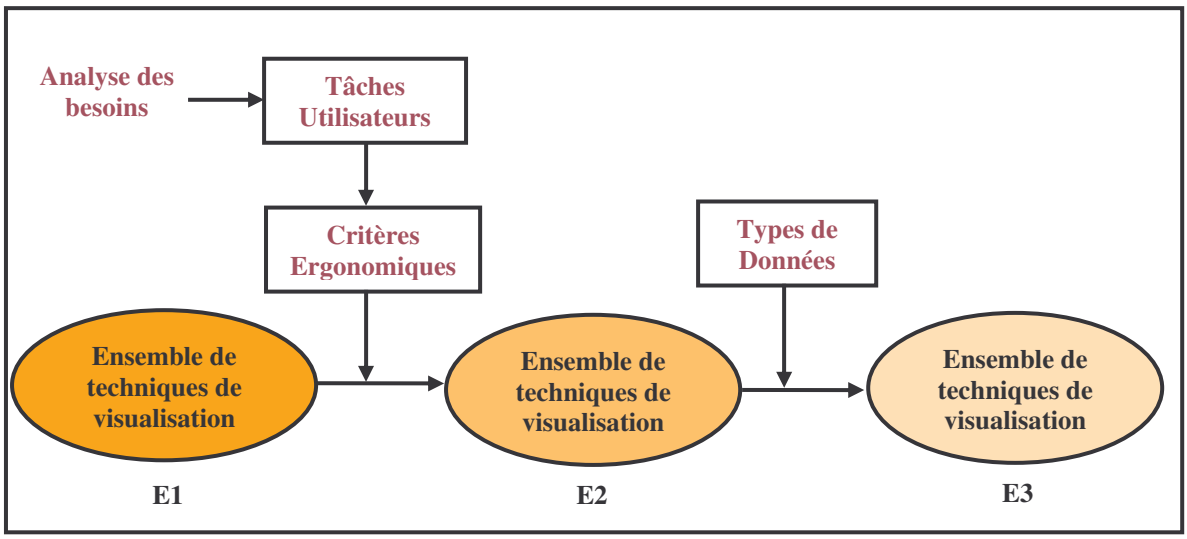


Figure 53 : Approche de Dâassi

Son approche part de l'analyse des besoins des utilisateurs et de la prise en compte de leurs tâches. Comme nous l'avons vu au long de notre état de l'art, de nombreuses techniques

de visualisations existent et peuvent être appliquées sur les données, du moins si elles sont en adéquation avec la nature des données et ce que l'on en attend (ou ce que l'on va en faire), Dâassi propose donc qu'un ensemble de techniques de visualisation (E1) soit constitué, une sorte de bibliothèque. Les tâches utilisateurs identifiées, des critères ergonomiques (navigation, informations affichées, structure de l'affichage…) sont déterminés et un premier filtrage peut être fait sur E1, il reste un sous ensemble E2. Le type de données devant être pris en compte, d'après l'auteur, un second filtrage est effectué sur E2 pour ne laisser que E3 qui désignent toutes les techniques de représentation d'informations qui pourraient correspondre au mieux aux besoins exprimés par l'utilisateur.

3.7.4 La démarche de modélisation de l'utilisateur : vers des systèmes adaptatifs

Nous voyons que l'utilisateur est une dimension centrale dans la réflexion qui s'instaure autour d'un outil de visualisation interactive d'information. De notre côté, nous pensons que cette réflexion doit être menée à l'aide d'une représentation de l'utilisateur qui prenne en compte les informations le concernant, que nous appellerons modèle utilisateur. Simier (2003) indique qu'«*au court de l'histoire des systèmes d'information, l'utilisateur a peu à peu été formaté en regard des systèmes d'information tels qu'on les conceptualisait. Sa représentation s'est ainsi souvent éloignée progressivement de sa réalité individuelle et organisationnelle. Il semble maintenant que relativement aux possibilités des technologies et à la montée en puissance de leurs utilisations, l'intégration d'une représentation poussée de l'utilisateur dans les modèles soit une nécessité, voire un enjeu.* ».

Qu'est ce qu'un modèle et qu'appelle-t-on modéliser ?

Selon un géographe, Hagget (1965), un modèle se défini comme « *une représentation schématique de la réalité élaborée en vue d'une démonstration* ». En effet, il s'agit d'une représentation schématique car il vise la simplification de phénomènes complexes qui implique une abstraction des éléments de ce phénomène, on parle de représentation de la réalité élaborée car il s'agit là d'une construction intellectuelle non d'une réalité, et elle est orientée vers la démonstration car le modèle sert à comprendre des propriétés, à confirmer ou infirmer des hypothèses, à expliquer le fonctionnement d'un processus.

La modélisation est donc, selon Le Moigne (1995) « *l'action d'élaboration et de construction intentionnelle, par composition de symboles, de modèles susceptibles de rendre intelligible un phénomène rendu complexe, et d'amplifier le raisonnement de l'acteur projetant une intention délibérée au sein d'un phénomène : raisonnement visant notamment à anticiper les conséquences de ces projets d'actions possibles* ».

Le modèle utilisateur : comment le déterminer ?

Evidemment, suivant le type de système et de besoins, la modélisation d'un même objet pourra s'effectuer sous différentes approches. Nous pouvons proposer quelques pistes de réflexion sur la manière de construire ce modèle.

Appliqué aux utilisateurs d'un système de recherche d'information, il existe d'après Bouaka (2004) trois catégories de modèle qui correspondent à trois procédés de construction de représentation utilisateur :

- **Le profil** : un profil est renseigné, à chaque fois qu'une requête est effectuée, le profil est appliqué pour ne retenir que les informations pertinentes.

- **Le modèle implicite** : Les préférences de l'utilisateur sont déterminées de manière implicite à partir de ses comportements face aux résultats de sa requête.

- **Le modèle explicite** : Un retour est demandé à l'utilisateur pour spécifier l'intérêt d'un résultat en fonction de son besoin, les préférences sont constituées à partir d'un apprentissage du système.

Pour construire une représentation d'un utilisateur de SRI en quête d'information, Lainé-Cruzel (1999) propose de se poser trois questions essentielles, à savoir :

- **Qui est il ?** (Culture générale, niveau de connaissances dans le domaine),

- **Que veut-il ?** (Volume d'information attendu, caractéristiques générales de cette information),

- **Qu'en fera-t-il ?** (Nature du besoin ou de la tâche qui le conduit à rechercher de l'information)

Qu'elle qu'en soit la manière, il est en effet important de connaître ces informations sur l'utilisateur afin de construire un système adapté à ses besoins.

Les besoins de l'utilisateur : le cas du veilleur

La question de déterminer les besoins des utilisateurs est une question qui requière en général une enquête auprès d'eux sur leurs besoins. Sur ce point, nous nous sommes intéressés aux études déjà menées.

Rousseau et Thil (1997) avancent que devant un corpus d'information, il existe quatre types de besoins du veilleur auxquels doivent pouvoir répondre les systèmes, en particulier les systèmes de visualisation d'information :

- **Besoin d'exploration** : Le veilleur peut naviguer parmi les concepts au travers les représentations visuelles, elles peuvent aussi le guider dans sa réflexion et l'aider à découvrir des indices pertinents ou émergeants sur son domaine de recherche.

- **Besoin de structuration** : les représentations peuvent aider le veilleur à établir un état de l'art de son domaine de recherche. Elles peuvent faire apparaître de l'information qui n'était pas présente dans l'individualité des textes (information endogène) en croisant différents éléments, il lui est possible de construire une vue du champ d'investigation. Qui travaille sur quoi ? Quelles sont les disciplines connexes à celle que l'on étudie ? Quelles sont les thématiques qui émergent et celle qui évoluent ?...

- **Besoin de positionnement** : La méthode consiste à rechercher des valeurs numériques comparables, des indicateurs d'activités.

- **Besoin de prospective** : Niveau d'interprétation des signaux par l'expert qui permet d'envisager l'avenir.

Les techniques d'analyse de données couplées aux systèmes de visualisation de l'information répondent dans une certaine mesure au besoin d'exploration, car à partir d'un ensemble de données, l'analyse permet d'identifier les similarités, dissemblances… entre données et la visualisation révèle de manière visuelle les structures, les motifs, etc., répondant aussi du même coup au besoin de structuration. Un utilisateur en phase exploratoire de sa recherche d'information, peut ainsi trouver des pistes que ce soit pour poursuivre sa recherche ou pour nourrir sa réflexion.

La visualisation de l'information, discipline à part entière, désigne donc un large ensemble de technologies, de techniques et de domaines d'application. Elle tente d'apporter des réponses aux problèmes que suscite l'abondance de données et la nécessité de les analyser pour comprendre les phénomènes qu'elles traduisent.

Nous avons donc pu voir, comme nous l'avions évoqué, que le processus de visualisation de l'information se situait à la fois en aval de l'analyse automatique de données et à la fois en amont d'une démarche d'analyse faite par l'expert qui, dans notre cas de projet de veille automatisée, pourrait être le veilleur lui-même.

Nous avons tenté d'entrevoir la nécessité de prendre en compte l'utilisateur du système. Pour cela, nous avons évoqué tout d'abord les techniques d'interactions entre utilisateur et système, puis nous avons évoqué la nécessité de plus en plus grande de concevoir les systèmes d'information à partir d'un modèle utilisateur qui prendrait en compte ses besoins.

Mais qu'en est il des outils existants ? Comment un veilleur peut aujourd'hui analyser l'évolution de l'information ? Comment est il pris en compte par le système ? Pour tenter d'apporter des réponses à ces questions, nous nous sommes livrés à une série d'observations sur des outils de veille présents sur le marché. Alors que jusqu'ici nous avions voulu garder une approche théorique, à présent nous voulons voir ce qu'il en est sur le terrain, sur ce que tout utilisateur peut être amené à utiliser un jour.

Chapitre 4

Etude des fonctionnalités de visualisation dans les outils de veille

4.1 Les objectifs

Dans notre étude des produits existants, plusieurs objectifs ont animés notre démarche. Le premier objectif était de les parcourir à la recherche d'une interface de visualisation d'information qui correspondrait le mieux à notre cadre de recherche, c'est-à-dire l'algorithme de classification incrémental. Autrement dit, l'interface recherchée doit comporter une représentation de classes/catégories, et éventuellement gérer l'évolution de ces classes dans le temps.

Le second objectif est lié à la nature même de la plupart de ces logiciels, dédiés à la veille. Nous avons pensé qu'il pouvait nous aider à spécifier les besoins des utilisateurs qui nous occupent ou en tout cas de voir quels sont les dispositifs mis en place pour tenter de répondre à ces besoins.

4.2 La méthode

4.2.1 Choix des outils

La première étape de notre démarche a été de se constituer une liste d'outils parmi lesquels nous pourrions sélectionner ceux qui nous intéressent. Pour réaliser ce corpus, nous avons essentiellement repris un ensemble de fiches produits constitué par Claire François, responsable du SRDI, à l'occasion d'une précédente étude menée entre 2003 et 2005 sur les outils de veille. L'ensemble comprend une cinquantaine d'outils aux orientations fonctionnelles diverses puisqu'on y retrouve des moteurs de recherche, des agents de surveillance de l'environnement, des plateformes logicielles, des outils de gestion documentaire, d'analyse textuelle, d'analyse de données, etc.

D'autres sources d'informations nous ont guidé dans l'élargissement de ce premier ensemble, notamment le guide publié par le CIGREF mais aussi la typologie des outils de veille réalisée par Aref Jdey[47] sous la forme originale d'une carte conceptuelle, d'après les travaux de Gilles Balmisse[48].

4.2.2 Critères de sélection

Pour faire correspondre les outils observés à notre problème qui porte sur la visualisation d'un algorithme de classification incrémentale, nous avons identifié deux points essentiels, d'une part les possibilités de visualisation offertes par les outils et d'autre part leur possibilité en termes d'étude d'évolution ou de suivi temporel de l'information.

A partir de ces critères nous avons constitué un tableau à double entrée de la forme

Outil × {Catégorisation, Classification, Cartographie, Diachronie}.

[47] http://vtech.canalblog.com/archives/2005/10/31/946969.html (consulté le 24/01/06)
[48] http://www.gillesbalmisse.com/ (consulté le 24/01/06)

- Par catégorisation nous entendons tout traitement qui affecte des informations ou documents à des catégories prédéfinies de manière plus ou moins automatique.
- La classification désigne quant à elle les processus d'analyse automatique de données qui calcule les classes à partir des caractéristiques des informations ou des documents.

- La cartographie désigne la fonction de construction de cartes à partir des informations, voire, de façon élargie, toute représentation graphique élaborée de l'information.

- La diachronie désigne la fonction de prise en compte de l'évolution d'une classe, d'une catégorie ou de données.

Le remplissage de ce tableau a donné lieu à une première sélection après consultation des fiches produits, des documents commerciaux, sites web, etc. Seuls les produits présentant au moins une des caractéristiques, ont été retenus.

Nous avons retenus 35 outils, qui proposent pour la plus grande majorité des fonctions de catégorisation ou de classification, ces fonctions étant souvent confondues sous le terme de clustering ou même classification. Dans les outils choisis, la catégorisation se présente sous la forme d'un plan de classement thématique établi en amont d'une recherche d'information ou à l'aide d'une taxinomie toute faite de type Dmoz[49], par exemple.

Du côté des classifications, il n'est pas toujours facile de connaître les techniques employées par les outils qui les pratiquent. Néanmoins nous avons pu constater que les analyses des cooccurrences des termes sont les analyses que nous avons le plus rencontré en terme de classification, face aux analyses par k-means entre autres.

Concernant la cartographie, celle-ci est quasi systématiquement associée aux techniques de classification, en représentant les classes, en traçant les réseaux de cooccurrence, etc. D'autres cartographies, notamment pour les analyses textuelles, offrent la possibilité de représenter les contextes de mots. Les outils d'analyse de données comme Miner3D, proposent quant à eux de visualiser ces données dans des espaces à 2 ou 3 dimensions.

Enfin l'aspect diachronique, c'est-à-dire le suivi de l'information dans le temps, est assez peu représenté. Si l'on associe l'analyse diachronique avec les fonctionnalités de visualisation, il en ressort un ensemble d'approches assez peu originales car limitées à des représentations en diagrammes, histogrammes, etc.

Une fois notre tableau constitué, nous avons ordonné les outils en fonction d'un critère d'intérêt que nous avons défini comme une combinaison de fonctionnalités qui nous semblent les plus importantes. Ce critère est illustré dans le tableau 3.

	+	--
1	Classif et/ou Categ + Carto + Diachro	Classif et/ou Categ
2	Classif et/ou Categ + Diachro	
3	Carto + Diachro	
4	Diachro	
5	Classif et/ou Categ + Carto	
6	Carto	

Tableau 3 : Critère de sélection des outils à tester parmi l'ensemble des outils du sous-corpus

[49] http://www.dmoz.org/

Le tableau permet de constater que la fonctionnalité essentielle recherchée est la diachronie. La seconde est la cartographie. Ainsi les outils de rang 1 sont plus intéressants à observer que ceux de rang 6. Ceux de la seconde colonne sont exclus de nos observations. A partir de ce critère nous avons établi une liste d'outils classés par priorité.

En concertation avec les membres de l'équipe concernés, c'est-à-dire C. François et P. Cuxac, nous avons identifié les 12 outils du tableau 4 qui peuvent potentiellement répondre à nos attentes, chacun ayant été classé suivant 3 degrés de priorité ou d'importance. Nous avons défini cette priorité par l'observation rapide des fonctionnalités des outils de rang 1 à 6 du tableau 3, nous avons jugé de l'intérêt de certains outils, de certaines approches parmi d'autres. Prenons AMI Market comme exemple, celui-ci présente des fonctionnalités de diachronie et de visualisations, mais ces dernières étant assez éloignées de ce que nous recherchions, nous l'avons placé en seconde priorité.

Priorité		
A	B	C
Tetralogie Pericles Autonomy Intellixir Wordmapper Miner 3D	AMI Market *Global Finder* VS 2000 Viz server	Kaliwatch Lingway

Tableau 4 : Listes des outils sélectionnnés

Nous avons sélectionné le logiciel Tetralogie, outil d'analyse de données développé par l'IRIT de Toulouse, la plateforme de veille Pericles de la société Datops, le module visualisation de la plateforme Autonomy IDOL Server de la société Autonomy, le logiciel d'analyse de données Intellixir de la société du même nom, l'outil d'analyse textuelle Wordmapper de l'éditeur Grimmersoft et Miner3D, outil d'analyse de données de la société du même nom. En second rang, la plateforme AMI Market Intelligence de l'éditeur Albert a été sélectionnée, ainsi que Global Finder développé par Knowings, VS2000 de Ayonis et VizServer de la société Inxight. Enfin la plateforme Kaliwatch de Arisem et Lingway par la société du même nom ont été retenue pour une éventuelle observation.

A ce stade, s'est posé un problème lié notamment à l'absence de version de test de certains produits, ce qui a impliqué que certains d'entre eux, même parmi les outils de première importance, n'ont pas pu être manipulé comme nous l'aurions voulu, ce qui a été le cas d'Autonomy, AMI Market, Global Finder, VS2000 et Kaliwatch, autrement dit : la moitié d'entre eux ! Pour palier partiellement à cette lacune, nous avons contacté les éditeurs pour leur demander des explications sur le fonctionnement de leur produit. Nous avons obtenu des retours pour AMI Market, Kaliwatch et Autonomy, les autres n'ayant pas répondu à nos appels intéressés.

Nos premiers résultats d'observations révèlent qu'étant donnée notre sélection portée sur certains outils uniquement à partir de la documentation commerciale, celle-ci indique que l'outil effectue les opérations qui nous intéressent, mais malheureusement, après observation, celui-ci ne se montre pas toujours à la hauteur de nos attentes.

4.2.3 Tests et observations des outils

Dans un premier temps, notre démarche a été de construire une procédure d'observation normalisée au travers la constitution d'une grille d'observation indiquant les fonctionnalités sur lesquels nous allions nous pencher, à savoir :

- La classification : quelle est la/les méthodes proposée(s) par le système ? Quels sont les éléments classifiés ? Comment les visualise-t-on ?
- La cartographie : quels sont les éléments représentés ? Comment sont ils représentés ? Quels sont les fonctionnalités et moyens d'interaction donnés à l'utilisateur ?
- La diachronie : quel est le principe d'analyse utilisé pour l'évolution ? Est-elle visualisée, si oui comment ? et là encore, comment l'utilisateur interagit-il avec la visualisation ?

Nous n'aborderons pas avec précision les aspects classification qui nous ont plutôt aidé à comprendre les cartographies réalisées par les logiciels. En effet, selon que l'on visualise des classes ou des catégories, par exemple, le résultat visuel et l'interaction peuvent être différents,

Nous présenterons donc les aspects cartographiques observés, en identifiant en chacun les points pouvant nous intéresser. Nous verrons ensuite les aspects liés à l'évolution de l'information, afin de voir comment le problème est abordé par ces logiciels.

A propos de l'aspect cartographique

Pour représenter des classes ou des catégories, plusieurs types de techniques sont employés.

Wordmapper est un logiciel de traitement de textes qui effectue des calculs de cooccurrence entre termes, applique la méthode des mots associés (*cf. 2.3.2.*) et dresse des cartes à partir de ces analyses. La représentation graphique de classes s'effectue à l'aide d'une étiquette de classe placée sur un diagramme en fonction de la densité de la classe et de sa centralité (figure 54). Cette représentation sur deux axes permet à l'utilisateur de savoir par exemple, quelles sont les thématiques importantes du texte.

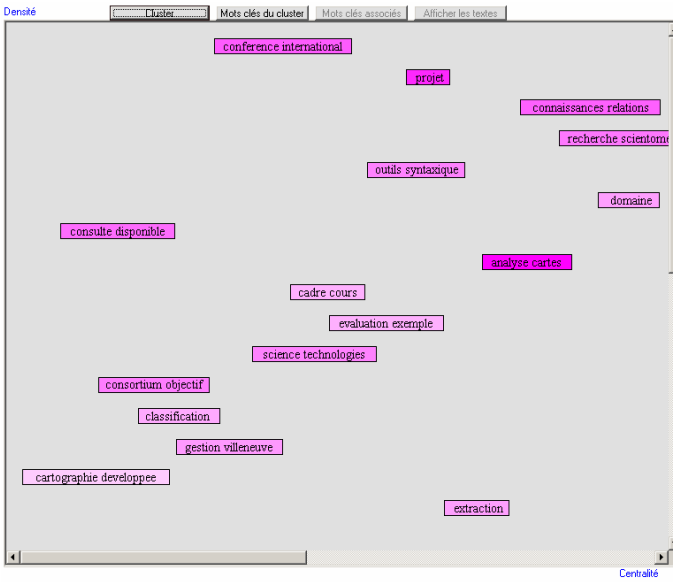


Figure 54 : Carte des classes (Wordmapper)

Les classes sont constituées à partir du calcul de cooccurrence des termes d'indexation de documents. La carte est à deux niveaux : le premier présente les classes comme le montre la figure 54, le second niveau permet à l'utilisateur de voir le contenu des classes sous forme d'un réseau de cooccurrence. Il permet aussi de représenter les associations externes, c'est-à-

dire les liens entre classes, puisque les éléments communs à plusieurs classes sont mis en évidence.

Nous retrouvons la démarche d'analyse par cooccurrence sur l'outil SDOC de la plateforme Stanalyst, l'approche est basée également sur les mots associés. C'est pourquoi nous pensons qu'elle ne nous apportera pas beaucoup d'éléments nouveaux sur la représentation de classes. Nous pouvons nous demander comment faire évoluer cette représentation dans le temps. Wordmapper a pour cela une approche particulière que nous étudierons dans la section suivante.

Dans le même registre, le logiciel Intellixir, en dehors des histogrammes, des courbes, etc., construit des cartes de cooccurrences à partir des caractéristiques des données analysées. Intellixir est un outil d'analyse de données qui offre à l'utilisateur la possibilité de croiser entre elles de multiples données. Des cartes dynamiques (c'est-à-dire dont le positionnement peut être manipulé par l'utilisateur pour écarter les sommets les uns des autres, étirer le graphe, etc.) sont générées (figure 55), ou l'on peut y voir la cooccurrence à l'aide de liens pondérés et l'occurrence par un indice appliqué au terme. La carte en figure 55 par exemple, représente le résultat du croisement d'auteurs avec leurs affiliations. L'utilisateur peut interagir sur le zoom et la rotation verticale de la visualisation et filtrer les données affichées à l'aide de seuils numériques pour les valeurs d'occurrence et de cooccurrence.

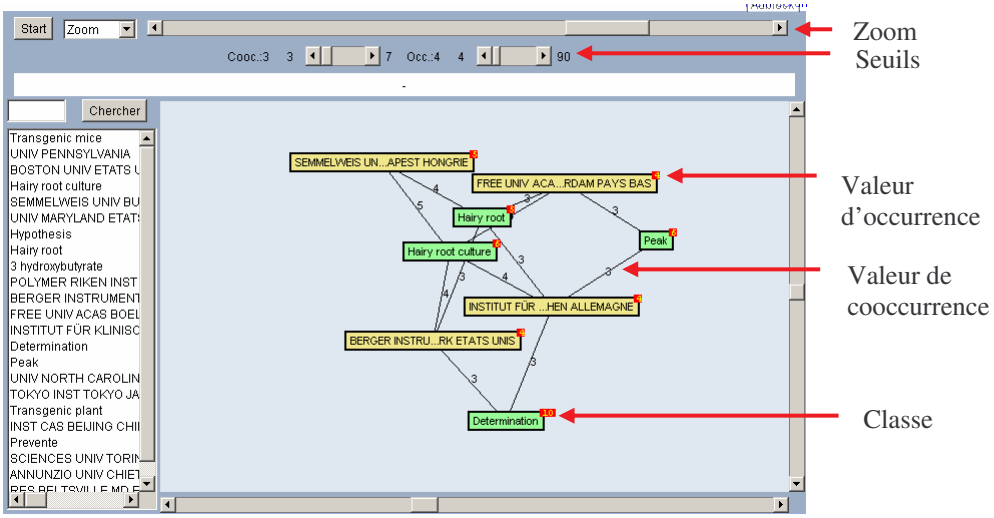


Figure 55 : Carte de cooccurrence (Intellixir)

De même que pour Wordmapper, l'approche d'Intellixir par le tracé de réseaux de cooccurrence ne nous apporte pas nous plus beaucoup d'informations nouvelles sur la manière de représenter les classes. Mais l'intérêt d'Intellixir se situe sur la manière envisagée de représenter l'évolution des ces graphes, question que nous aborderons dans la section suivante.

Miner3D est un outil de visualisation de données. Il permet de représenter visuellement les éléments d'un tableau de données dans des espaces 3D ou sur des plans 2D. L'approche de Miner3D permet d'explorer visuellement toutes les données à partir des représentations, mais aussi à partir de fonctions statistiques, de classifications, de requête dynamique… Pour ce qui est des représentations, les espaces 3D et les plans prennent comme dimensions les attributs des données. Chaque axe est configurable, c'est-à-dire qu'il est possible de modifier la valeur des axes et ainsi accroître les points de vue sur les données. En plus du paramétrage des axes, il est possible de visualiser des informations complémentaires à

l'aide de codes couleur, de formes des items affichés, de sons associés à des valeurs, etc. Tout ceci a pour but de donner à l'utilisateur la possibilité de croiser un maximum d'informations sur un minimum d'espace. Miner3D est intéressant du point de vue graphique, car il offre de nombreuses possibilités de représentation (figure 56).

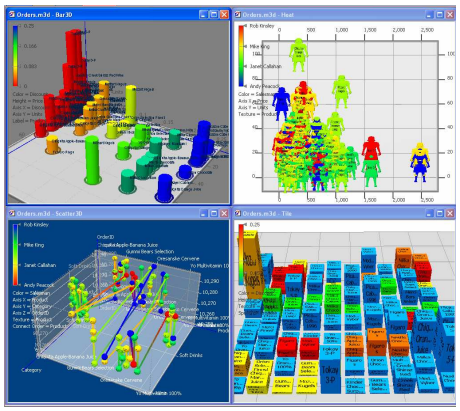


Figure 56 : Différents mode de visualisations de Miner3D

Il est intéressant de voir comment peuvent être employés les éléments graphiques couleurs, formes et sons, par exemple. Ils permettent ainsi d'afficher plusieurs informations sur un espace réduit. L'usage de la représentation 3D présente aussi un intérêt même si elle possède aussi quelques limites. Le fait de pouvoir se déplacer autour des données, ce qui est indispensable pour ce type de représentation, permet d'accroître les points de vue de l'utilisateur sur ces données : selon qu'il se trouve au dessus de l'espace, en face ou sur la gauche, les axes ne donnent pas tous les mêmes informations. Elle augmente en contrepartie les risques de confusion, de désorientation de l'utilisateur vis-à-vis des données.

Cependant Le logiciel n'effectue pas d'analyse de l'évolution des données. Rien n'est prévu pour cela. La seule possibilité envisageable est associée aux classifications puisqu'il est possible d'exporter le résultat des ces opérations. Quelques manipulations sur ces exportations à différents instants, permettraient, en réintégrant les données ensuite, de représenter les classes de données à différents instants. Mais nous pensons que cette analyse serait coûteuse et sommaire.

La plateforme IDOL Server de la société Autonomy propose des fonctionnalités d'indexation, de classification et de catégorisation, que nous n'avons pas eu la possibilité de tester, mais qui présente un intérêt du point de vue cartographique. Les modes de visualisation proposés par la plateforme sont des cartes de classes (figure 57), des représentations 3D de classes (figure 58), des réseaux et un spectrographe servant au tracé de l'évolution de classes que nous étudierons un peu plus en détail dans la section suivante. La carte 2D est un mode de représentation proche de ce que nous avons déjà pu voir au cours de notre état de l'art sur les techniques de visualisation, assimilé aux paysages de densité. La carte en 3D en revanche est proche des modes de représentation de Miner3D. On peut donc y faire la même remarque, à savoir qu'il est intéressant du point de vue perception visuelle, de voir représenter les données dans des espaces tridimensionnels et de pouvoir naviguer parmi elles. L'inconvénient est le même, mais à un degré supérieur pour Autonomy : l'utilisateur manque complètement de repères et à part visualiser des groupes de points (ce qui peut être l'unique but d'une telle représentation) aux propriétés communes ou proches, l'information qu'elle apporte n'est pas rendue facilement accessible à l'utilisateur.

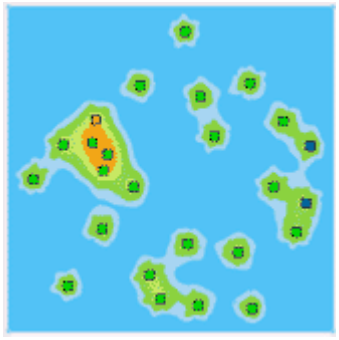


Figure 57 : Carte 2D des classes

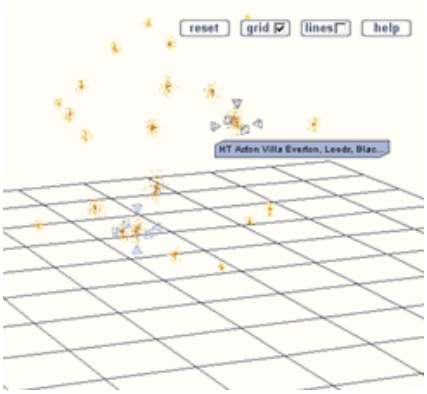


Figure 58 : Carte 3D des classes

Enfin, le logiciel Tétralogie présente, quant à lui, un certain nombre de modes de représentation. Tétralogie est un logiciel de traitement de donnés qui offre de multiples possibilités de croisement d'informations. Il intègre des méthodes d'analyse factorielle (ACP, AFC…), des méthodes de classification (centre mobiles, classification ascendante hiérarchique…). En rapport à ces méthodes d'analyse, le logiciel propose des modes de représentation tantôt en 3D, tantôt en 2D ou encore en 4D animés de manière interactive ou automatique (rotation, zooms, sélection de points, choix des couleurs, etc.). D'un point de vue général, que cela soit en 2, 3, ou 4D, la visualisation des données de Tétralogie sont des représentations en nuage de points (Figure 59) au sein desquels il est parfois difficile de se retrouver, en tant qu'utilisateur non expert du système.

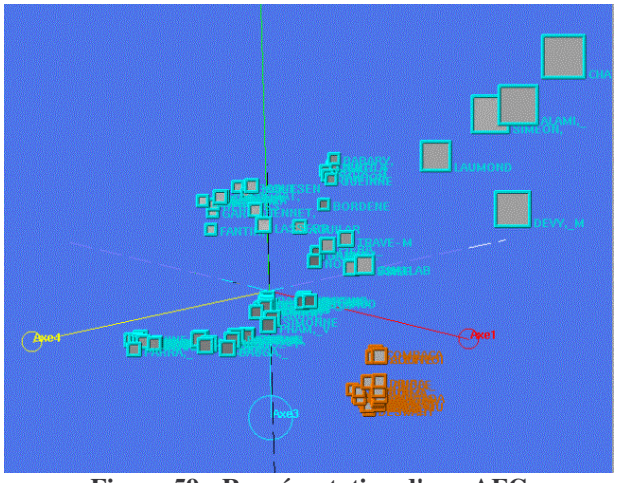


Figure 59 : Représentation d'une AFC

A propos de l'aspect diachronique

L'étude de l'évolution admet plusieurs approches d'après les outils que nous avons observés.

La plupart des outils de veille comme le montrent les figures suivantes utilisent des représentations en courbes et en histogrammes basés sur l'occurrence ou la fréquence des termes. LexiQuest Mine, par exemple, représente la fréquence du terme dans le corpus (figure 60). Intellixir représente l'évolution globale de termes sur un intervalle de temps donné à

l'aide d'un facteur de croissance calculé en fonction de l'évolution des fréquences de ces mêmes termes. Ceux-ci sont ensuite triés dans l'ordre décroissant de ce facteur (figure 61).

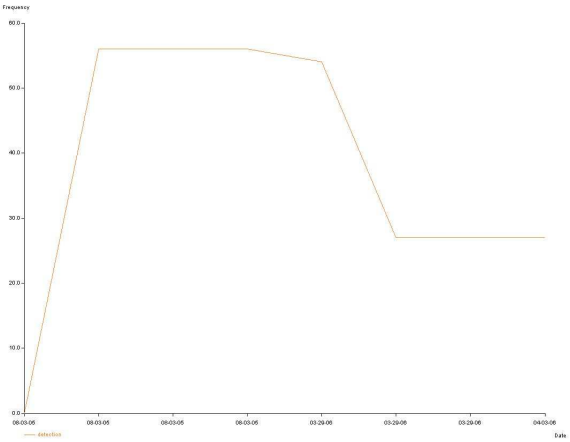


Figure 60 : Courbes des fréquences par SPSS LexiQuest Mine[50]

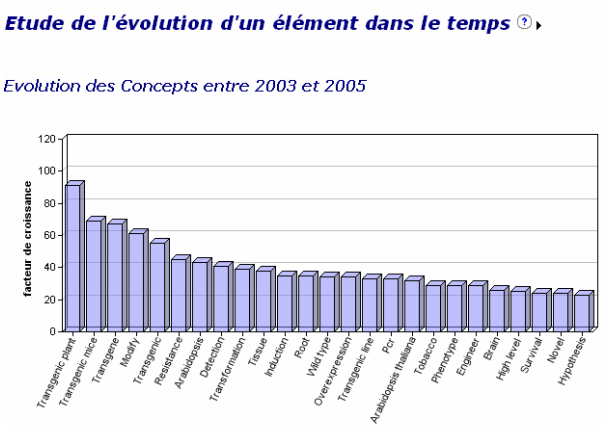


Figure 61 : Diagramme de d'évolution de termes par Intellixir

Le logiciel Wordmapper, du point de vue de l'analyse de l'évolution de l'information, offre la possibilité de comparer deux versions d'un même texte et de faire ressortir les « thématiques émergentes » d'un texte à l'autre. Pour ce faire l'étude se fait par la comparaison des thématiques d'un texte (figure 62) puis de l'autre (figure 63), le résultat étant la représentation en diagramme des classes apparaissant dans le second texte et qui n'était pas dans le premier (figure 64). Une telle approche est intéressante dans le principe puisqu'elle permet de faire ressortir une information essentielle qu'est l'apparition d'éléments nouveaux dans un texte à différents instants. Cependant, il est impossible pour l'utilisateur de voir l'historique de cette évolution.

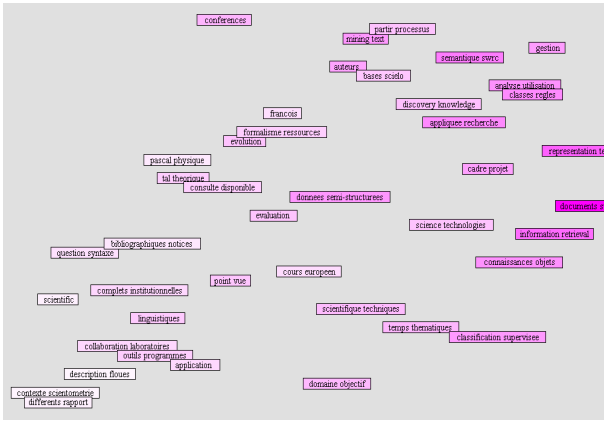


Figure 62 : Carte de l'analyse standard d'une première version d'un texte

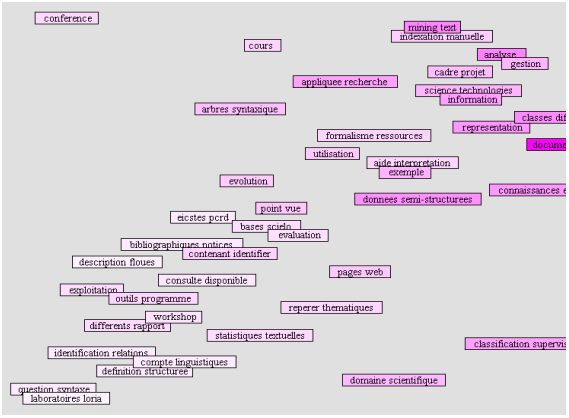


Figure 63 : Carte de l'analyse standard d'une seconde version, enrichie, du même texte

[50] http://www.spss.com/lexiquest/lexiquest_mine.htm

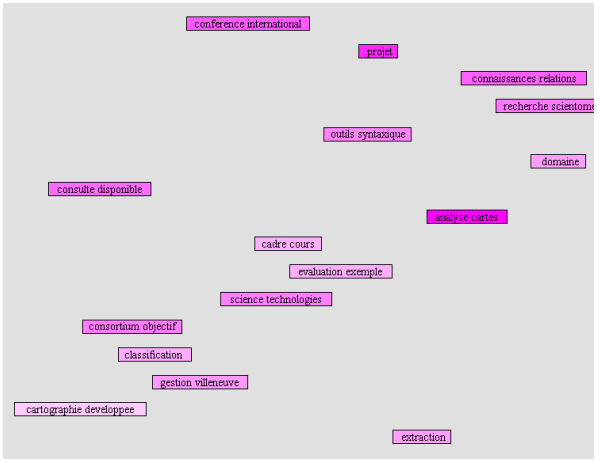


Figure 64 : Carte résultant de la comparaison entre les deux versions. Apparaissent sur cette carte, les nouvelles classes de la seconde version.

Intellixir propose, lui, divers approches pour mener une analyse diachronique de l'information. Les données traitées par le produit peuvent être structurées, c'est-à-dire obéissantes soit à une structure de champs comme une notice bibliographique avec les champs « auteur », « titre », « année de publication », etc. soit à une structure XML[51] décrivant les différents types d'informations présents dans le document. La représentation diachronique de l'information par Intellixir se fonde sur l'utilisation des champs de type « date » des données qu'il traite. Ce qui permet au programme de représenter l'évolution date par date (année par année, mois par mois, etc.) à l'aide de courbes et d'histogrammes. Intellixir permet aussi à l'utilisateur de pouvoir suivre le développement de réseaux de cooccurrence, il est possible pour lui d'interagir avec la visualisation à l'aide de curseurs fixant intervalles de temps basés sur les dates des données (de la même façon qu'avec les seuils de cooccurrences de la figure 55). Ainsi, pour des documents par exemple, il est possible de limiter l'affichage aux documents dont la date est inférieure à 1999, 2000, puis 2001, 2002, etc. faisant alors évoluer le réseau jusqu'à la date la plus récente (figures 65, 66, 67, 68). L'opération inverse est possible aussi en réalisant la démarche à l'envers. Le graphe ci-dessous présente l'évolution d'équipes de recherche de 1999 à nos jours, à partir d'un corpus de documents.

[51] Extensible Markup Language

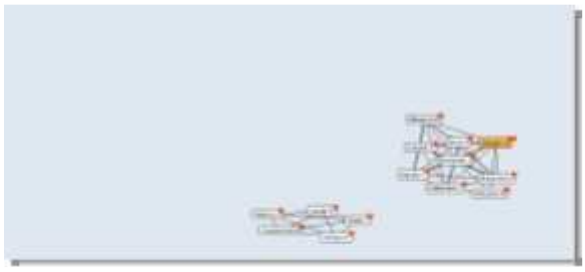


Figure 65 : 1999, les premières équipes

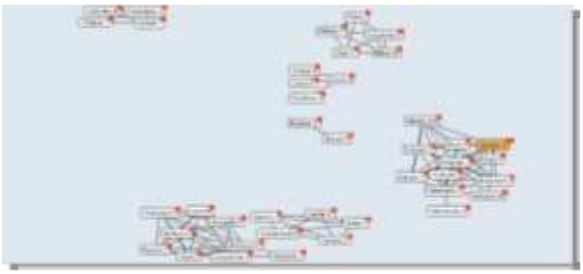


Figure 66 : 1999-2000, apparition de nouvelles équipes

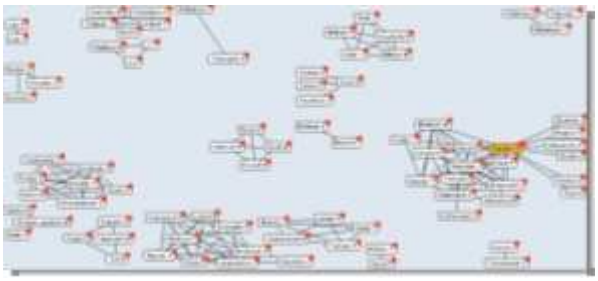


Figure 67 : 1999-2001, de nouvelles équipes rejoignent celles déjà présentes

Figure 68 : 1999 à nos jours, le réseau s'agrandit encore

Nous avons eu quelques informations concernant le logiciel de veille AMI Market, qui présente une visualisation de données intéressante et qui a l'intérêt de proposer une analyse diachronique de l'information un peu originale. Cette évolution concerne les informations de presse, l'outil se propose de montrer l'évolution de l'importance d'une information (ou plus précisément d'un terme) au fil du temps à l'aide d'un diagramme. Les axes de ce diagramme représente le temps pour l'axe des abscisses et l'importance de ce terme sur l'axe des ordonnées, c'est-à-dire là où le terme a été relevé dans le texte. En effet, si le terme est statistiquement plus présent dans le corps du texte, il sera placé proche de la graduation « mention », s'il est plutôt présent dans les titres des articles et journaux, il sera alors proche du degré « article ».

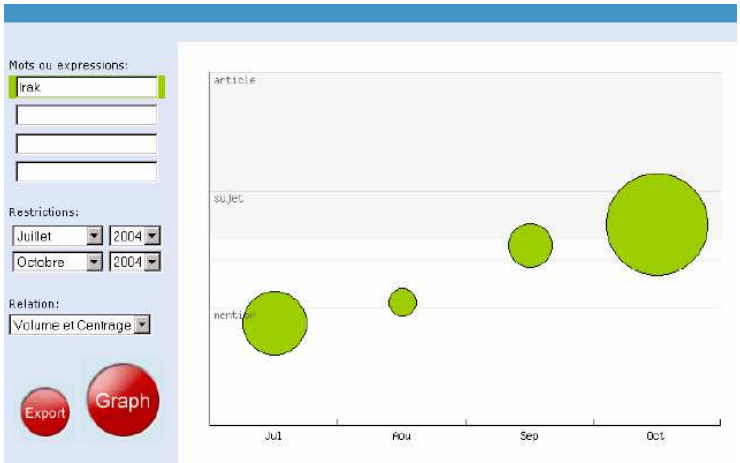


Figure 69 : Visualisation diachronique d'AMI Market

Nous pouvons observer le résultat avec la figure 69. Le diamètre des points correspond au nombre de documents dans lesquels ce terme est présent. Ce mode de représentation propose donc à l'utilisateur une double lecture, l'une sur la fréquence des termes (diamètre),

l'autre sur « l'importance » de ce terme, avec la dimension temporelle, trois dimensions sont ici représentées.

La question de l'évolution est aussi envisagée par la plateforme Tetralogie. En effet, nous avons pu tester qu'il était possible de retracer le parcours, appelé *trajectoire* dans l'outil, d'un élément parmi les autres. Prenons par exemple un ensemble de publications scientifiques sur plusieurs années, il est possible de représenter visuellement le parcours d'un auteur en fonction d'autres auteurs, par exemple, à l'aide d'une trajectoire ordonnée à l'aide de chiffres, c'est-à-dire que s'afficheront, sur une même représentation, les places qu'a tenu successivement l'auteur parmi les autres auteurs.

Pour ce faire, il est nécessaire de mêler les résultats de plusieurs analyses qu'il faudra superposer sur une même représentation. Nous trouvons cette approche intéressante car elle permet de bien retracer, au sens propre du terme, le parcours d'un élément. Malheureusement, l'approche ne peut se faire que dans un contexte statique. En effet, si l'auteur se déplace parmi un paysage d'éléments eux-mêmes en évolution, retracer la trajectoire de l'un d'entre eux sans faire évoluer les autres n'a pas de sens, il est nécessaire de retracer le parcours suivi par tous les éléments.

Un autre mode de représentation est présenté dans Tetralogie basé sur un contexte géographique, c'est-à-dire que les données sont associées à un fond de carte sur lequel elles s'affichent. Il est alors possible de visualiser l'évolution de ces données en les associant à des données géographiques, cette évolution se présente sous la forme de couleurs variant au fil des valeurs prises par ces données.

Enfin, nous avons évoqué précédemment les visualisations graphiques proposées par IDOL Server d'Autonomy. La plateforme dispose d'une visualisation en spectrogramme qui retrace l'évolution des catégories (figure 70). En termes de volume, de taux d'utilisation et de contenu. Sur l'axe vertical sont disposées les catégories et l'axe horizontal représente l'axe du temps. Le nombre d'éléments contenus dans la catégorie est visualisé de deux manières : l'épaisseur du trait et les nuances de couleurs. Le contenu des catégories peut évoluer aussi de manière qualitative, c'est-à-dire que leur contenu peut diverger vers d'autres lignes thématiques, signifiant alors qu'une partie du contenu s'éloigne du sens de la thématique et se reporte vers une thématique plus proche. Evidemment, cette représentation graphique pose quelques problèmes, notamment liés à la disposition des catégories : Comment expliquer que les catégories qui se scindent, reportent la partie scindée sur une thématique proche ? Les catégories se placent elles en fonction des dérivations ? Comment interpréter la position des catégories sur l'axe vertical ? Etc.

Figure 70 : Le spectrogramme d'Autonomy

Chapitre 5

Analyse et définition des fonctionnalités de visualisation nécessaires pour les utilisateurs à l'analyse diachronique d'un domaine

Introduction

Dans le cadre de la construction d'un algorithme incrémental de classification, nous nous penchons sur les premiers résultats obtenus afin d'y déterminer les informations utiles à l'utilisateur pour analyser l'évolution de l'information scientifique. Quelles fonctionnalités de visualisation peuvent être proposées pour rendre compte de l'évolution de thématiques de recherche pour un domaine donné ? La définition des éléments informationnels et les représentations de ces éléments peut s'enrichir d'une définition des utilisateurs du système.

Nous partirons donc d'une démarche de modélisation utilisateur, c'est-à-dire déterminer les caractéristiques de l'utilisateur afin d'adapter le système à l'utilisation qu'il en fait. Deux méthodes sont possibles : soit la conduite d'une enquête auprès des utilisateurs afin de déterminer leurs besoins, méthode privilégiée par les projets à long terme, soit la conduite d'une réflexion à partir d'un **stéréotype** d'utilisateur, méthode plus rapide mais moins fiable que la précédente qui demandera une enquête a posteriori pour validation. Nous avons choisi la seconde méthode pour mener notre travail.

A partir d'une réflexion sur l'utilisateur, nous allons tenter de modéliser les informations que l'on peut tirer du journal de classes, c'est-à-dire l'ensemble des données qui décrivent les thématiques à chaque instant de leur évolution. Pour cela nous utiliserons le modèle utilisateur que nous aurons réalisé et, à partir des besoins, nous tenterons de déterminer les informations devant figurer dans le journal pour répondre à ces besoins.

Enfin, à partir des informations identifiées comme devant figurer dans le journal, nous réfléchirons sur la manière la plus adaptée de les représenter pour l'utilisateur, au travers l'examen de différentes techniques de visualisation de l'information.

5.1 Présentation de l'algorithme de classification incrémentale : GERMEN

5.1.1 Objectif de l'algorithme

L'objectif principal de cet algorithme proposé par Lelu et al. (Lelu et al., 2004 et 2006) est de rendre compte « en temps réel » de l'évolution de thématiques de recherche par une analyse cumulative des publications de recherche.

5.1.2 Fonctionnement de l'algorithme

La méthode proposée est une classification automatique non supervisée, c'est-à-dire qu'elle ne s'effectue qu'à partir des données, sans intervention humaine, aucune classe n'est préétablie.

Elle est basée sur la notion de densité d'un nuage de points. Les points étant les données analysées, chaque point possède un voisinage de points plus ou moins proches, ce voisinage constituant la densité de ce nuage. La proximité relative des points entre eux est fondée, elle, sur la similarité des données analysées.

Selon Lelu et al., pour rendre compte avec exactitude des évolutions temporelles il est nécessaire que le système se conforme à 5 exigences.

Pour avoir une base stable, la classification doit

1. Être indépendante de l'ordre de présentation des données : quel que soit l'ordre dans lequel les données seront analysées, le résultat sera le même.
2. Etre indépendante des conditions initiales, c'est-à-dire, contrairement au k-means, par exemple, pour lesquels l'initialisation du processus s'effectue à partir des données ou aléatoirement ce qui engendre des résultats à chaque fois différents, l'algorithme incrémental donne les mêmes résultats quelque soit l'exécution.
3. Impliquer un minimum de paramètres afin d'alléger les choix offerts à l'utilisateur pour la paramétrage de l'algorithme.

S'ajoutent à ces exigences de classification, les contraintes de l'incrémentalité, c'est-à-dire :

4. Rectifications des frontières entre classes, apparition de nouvelles classes.
5. Indépendance du résultat de la classification de l'ordre des données présentées antérieurement, tout en découlant des données antérieures par un historique pouvant faire l'objet d'interprétation.

L'incrémentalité de l'algorithme réside dans le fait que les données sont ajoutées aux précédentes et les classes constituées par les analyses antérieures sont amenées à évoluer en fonction des nouvelles données. De plus, le choix a été fait d'effectuer l'analyse de manière cumulative, c'est-à-dire que toutes les données sont conservées au fil du temps.

Les évolutions des classes s'effectuent par modifications locales des voisinages, des densités et des noyaux de classes. L'algorithme construit progressivement et en la mettant constamment à jour, une structure de donnée comportant pour chaque point la liste des ses voisins, sa densité et son numéro de chef de classe. A chaque arrivée d'un vecteur (un document à n dimensions, n étant le nombre de descripteurs de ce document), on calcule les changements de densité induits dans son voisinage puis les changements de chef de classe induits. Les classes, par la modification de leur densité, peuvent éclater en plusieurs autres classes, voire s'émietter en de multiples petites classes, plusieurs d'entre elles peuvent fusionner en une, d'autres apparaîtront, ou se développeront, ou encore mourrons faute de recevoir de nouveaux documents et d'interagir avec d'autres classes.

Les classes produites sont dites non-strictes, c'est-à-dire qu'elles admettent de partager des documents avec d'autres classes. Leur population est donc de deux catégories : les documents stricts n'appartenant qu'à une et une seule classe, et les documents multivalents appartenant à plusieurs classes. On dit aussi que les classes sont recouvrantes, comme le montre la figure ci-contre.

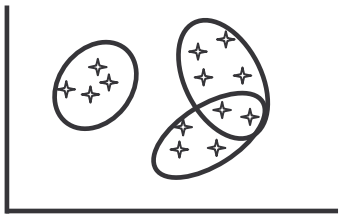


Figure 71 : Une classification non hiérarchique et des classes recouvrantes

Pour expérimenter le fonctionnement de l'algorithme, une série d'expériences a été menée à partir de la base documentaire PASCAL de l'INIST, avec la constitution d'une base nommées « 10 ans de géotechnique dans la base PASCAL ». Cette base recense les documents extraits de la base PASCAL se rapportant au domaine de la géotechnique[52], sur une durée de 10 ans.

Une première interface, appelée Classotron, a été testée pour tenter de rendre compte des phénomènes. Elle se base sur la formation de noyaux, c'est-à-dire les classes, et les liens entre les noyaux, c'est-à-dire les documents multivalents. On obtient alors un graphe non orienté dans lequel les noyaux sont les sommets et les documents multivalents les arêtes entre sommets.

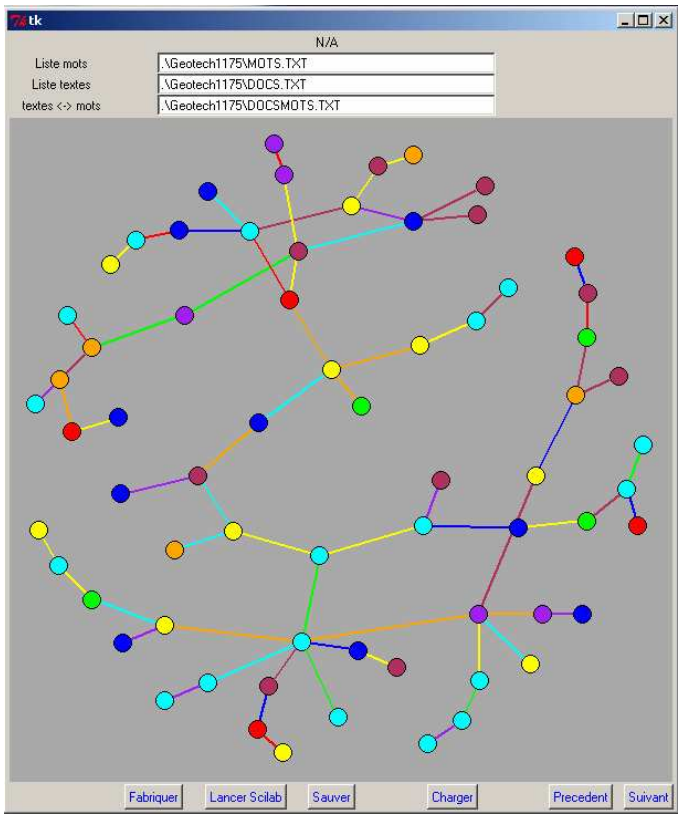


Figure 72 : L'interface *Classotron*

5.1.3 Objet d'application de l'algorithme

L'algorithme s'attache à calculer l'évolution de thématiques de recherche à l'intérieur d'un domaine. Celui-ci se définit comme notre univers de travail, comme le domaine géotechnique, par exemple.

Comment s'organise ce domaine ? Il contient des thématiques, comme la construction de structures souterraines dans le domaine géotechnique, qui sont déterminées à partir du contenu des documents. Ce dernier est exprimé par le vocabulaire d'indexation que sont les descripteurs des documents (les mots-clé). A partir de ces descripteurs il est aussi possible de découvrir les concepts utilisés dans les thématiques. Une thématique se caractérise aussi par un réseau d'auteurs affilié à un organisme de recherche (privé ou public).

[52] Discipline des Sciences de la Terre qui étudie la subsurface pour en faciliter l'aménagement. Elle englobe l'ensemble des activités liées aux applications de la mécanique des sols, de la mécanique des roches et de la géologie de l'ingénieur. Elle se penche sur l'interaction entre les terrains et les ouvrages environnants, dans leur réalisation ou leur exploitation.

5.2 Modélisation des utilisateurs, des usages et des besoins informationnels

« *Anyone who needs to make someone else understand something can then choose the most appropriate genre and medium for the information problem, and audience* » ([Gershon et Page, 2001], p37). C'est cette idée rapportée notamment par Gershon et Page, qui a animé notre démarche de modélisation des résultats de l'algorithme de classification incrémental. Ainsi nous avons choisi de prendre comme point de départ la définition de l'ensemble d'utilisateurs susceptibles d'utiliser le système. Cet ensemble nous permet de définir les besoins auxquels le système doit pouvoir répondre et les données qui lui sont nécessaires pour y répondre.

5.2.1 La démarche générale suivie.

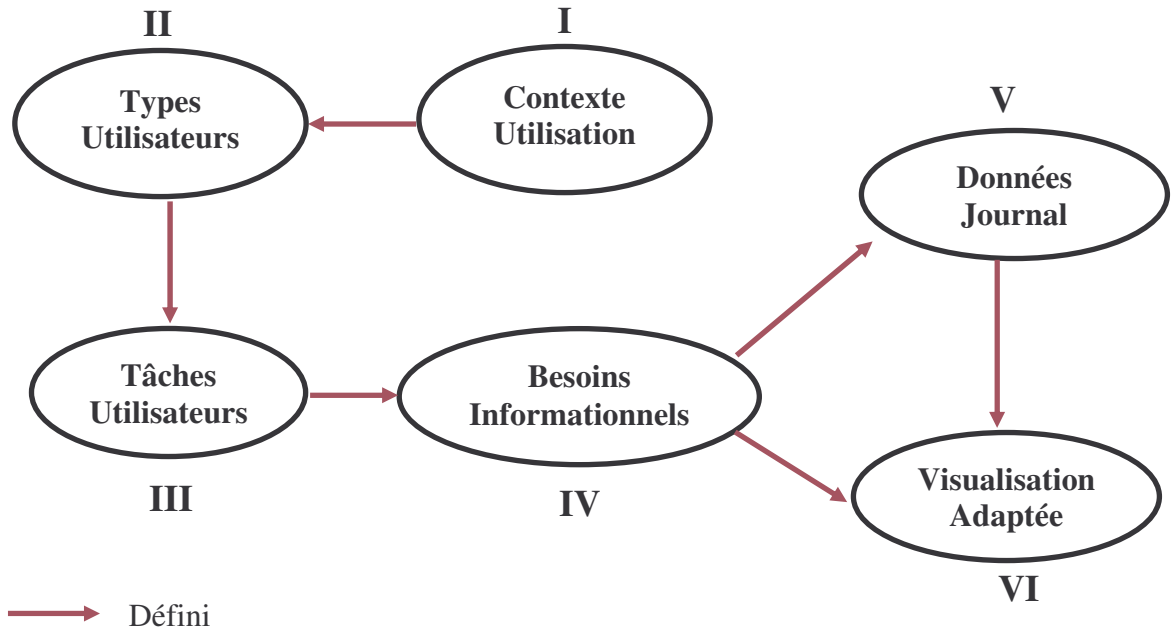


Figure 73 : Démarche générale de modélisation

La figure 73 décrit la démarche de modélisation que nous suivons à présent :

I. Pour délimiter la population des utilisateurs potentiels, il est nécessaire de déterminer un contexte d'utilisation, nous proposons de le définir à partir des données analysées et de réduire la population appartenant à ce contexte en vérifiant l'adéquation entre les résultats attendus du système et les tâches de ces utilisateurs.

II. Les différents types d'utilisateur auxquels nous avons à faire, c'est-à-dire ceux susceptibles d'être intéressés par les informations qu'apporte l'algorithme, sont définis à partir du contexte d'utilisation.

III. On peut déduire des tâches à partir du type d'utilisateur, tâches spécifiques à une fonction ou fonction impliquant des tâches. Les tâches sont les actions que l'utilisateur doit accomplir pour remplir son rôle. Elles se déterminent en fonction des objectifs et des enjeux de chacun.

IV. A partir des tâches de l'utilisateur, nous pouvons déterminer des besoins informationnels spécifiques (ce que l'utilisateur doit faire et de quelle(s) information(s) a-t-il besoin pour le faire ?)
V. L'analyse des besoins permet de filtrer les informations adaptées. Il nous faut par conséquent les identifier, les recueillir et les structurer sous la forme d'un journal.
VI. Enfin la visualisation obéit à 2 contraintes : elle est définie par le type d'information (certaines informations nécessitent une visualisation adaptée à leur nature) et par le besoin lui-même, le rendu devant s'y adapter. L'utilisabilité de la visualisation dépend de la prise en compte de ces 2 contraintes.

5.2.2 Définition des types utilisateur

Le premier critère qui permettra de déterminer les utilisateurs auxquels nous allons nous intéresser, est défini par le type de données analysées. Dans notre cas, ces données sont des notices bibliographiques de publications scientifiques.

Ce type de données permet de fixer un contexte d'utilisation : les notices dont il est question se rattachent à un contexte scientifique.

Le second critère est défini par la confrontation entre les besoins des utilisateurs potentiels et les résultats attendus de notre algorithme de classification. Autrement dit, nous nous posons la question suivante : dans le monde scientifique, qui a besoin de connaître l'évolution des thématiques dans un domaine scientifique donné ?

Nous avons ainsi identifié 5 types d'utilisateur (Types-U) susceptibles d'être intéressés par la représentation de l'évolution de thématiques de recherche : (1) le gestionnaire ou administrateur, (2) le responsable scientifique, (3) le veilleur, (4) l'historien des sciences ou l'étudiant et (5) le chercheur.

Le gestionnaire : que nous appellerons aussi administrateur de la recherche, est chargé de gérer administrativement la recherche pour un organisme public comme le CNRS ou tout organisme privé. Il gère le budget, assigne les crédits aux divers projets ou unités de recherche. L'administrateur peut être un scientifique ou ne pas l'être. Dans le cas ou il ne l'est pas, il peut être amené à devoir consulter un expert afin d'obtenir un avis éclairé sur la situation scientifique et prendre les meilleures décisions.

Le responsable : Par le terme *responsable* nous désignons toute personne chargée de mener une politique pour un service de recherche, une unité de recherche ou une équipe de recherche, c'est-à-dire les directeurs de recherche, les directeurs de laboratoire, les responsables de service de R&D, etc. Pour accomplir cette tâche il a besoin d'avoir une vue globale du champ dans lequel il pilote les chercheurs ou les ingénieurs afin d'orienter sa politique vers des applications, des collaborations, des thématiques prometteuses en terme de résultats, mais aussi, pourquoi pas, en terme de valorisation. Afin d'améliorer la visibilité des ses travaux il a besoin de pouvoir accéder à des informations sur les journaux, les congrès, etc. faisant autorité sur une thématique. Le responsable peut être amené à consulter des experts ou recruter des chercheurs de divers champs de recherche, pour cela il s'intéressera aux différents acteurs et cherchera à percevoir leur activité, leur dynamique, voire leur impact dans le domaine.

Le veilleur : Le veilleur est chargé de surveiller l'environnement à l'écoute d'information pouvant alimenter une stratégie. Par stratégie nous entendons la planification d'actions possibles à mener, pour atteindre un but, en fonction des circonstances extérieures ou du

comportement des autres agents intéressés par cet objectif (les concurrents). Le veilleur travaille généralement pour un demandeur d'information (le décideur), mais la fonction veille n'est pas réservée à une personne et n'est pas toujours subordonnée à une demande, puisque nous pensons que tout type d'utilisateur peut être amené à accomplir cette fonction à son compte. Sa tâche, dans notre contexte, est de détecter les « signaux faibles », c'est-à-dire des indices qui laissent présager des opportunités à saisir, d'un point de vue scientifique, technologique, commercial ou « rayonnant » (l'émergence des nanotechnologies, par exemple, permet de confirmer la position d'une entreprise déjà dans le domaine, si ce n'est pas le cas, il est peut être encore temps pour elle d'intégrer cette thématique de recherche et développement dans son activité avant que le marché ne soit saturé par d'autres nouveaux entrants). A charge du veilleur d'identifier les tendances à venir à partir des dynamiques actuelles, à l'aide d'indicateurs d'évolution, de l'analyse de la situation globale, etc.

L'étudiant et l'historien des sciences : Ce type d'utilisateur désigne toute personne voulant mener une étude descriptive ou analytique sur un domaine scientifique, comme la construction d'un état de l'art. Il voudra retracer l'évolution de ce domaine ou des thématiques. Il cherchera, par exemple, les origines de tel champ de recherche, voudra retracer les étapes de cette évolution, cherchera à identifier les conséquences de comportement des thématiques afin de comprendre l'évolution qui lui est présentée.

Le chercheur : Le profil du chercheur est à mi chemin entre l'étudiant qui cherche à mettre à jour son état de l'art et s'intéresse à l'évolution des idées et le responsable qui oriente une politique de recherche et peut être amené à chercher des collaborations. Comme le responsable, il peut également vouloir améliorer la visibilité de ses travaux et identifier les lieux de publications essentiels.

Ces types d'utilisateur sont des modèles types et ne correspondent pas exactement à des personnes, mais désignent plutôt des rôles ou des facettes. Plusieurs rôles ou facettes peuvent être attribués à une personne.

5.2.3 Définition des besoins informationnels

L'algorithme permet de retracer l'évolution de thématiques dans le temps. Suivre l'évolution peut correspondre à deux types d'utilisation :

- Utilisation **rétrospective** : L'utilisateur souhaite des informations sur l'évolution des thématiques, des informations utiles à la description de l'histoire d'un domaine.

- Utilisation **prospective** : L'utilisateur souhaite des informations afin d'envisager des futurs possibles. Ces prévisions n'obéissent pas à des modèles prévisionnels, elles sont déduites par l'utilisateur en fonction du comportement des données au fil du temps. Cette utilisation fait écho aux considérations de Rousseau et Thil (1997) (*cf. 3.7.4*) sur le besoin prospectif du veilleur et à celles de Garfield (1986) au sujet de l'étude des cartes de la science comme outil de prévision (*cf. 3.4*).

Ces deux catégories reflètent d'une part un besoin de comprendre, d'expliquer des phénomènes à l'aide de l'étude de leur évolution, d'autre part un besoin de réduire

l'incertitude de l'avenir en essayant de construire des prévisions à partir des données de l'évolution.

Alors que nous venons de déterminer 5 modèles-type d'utilisateur (Etape II) en essayant de décrire les tâches qui sont les leurs (III), nous allons tenter de recenser, sous forme d'un tableau, les tâches et les besoins informationnels (IV) qui en découlent en les filtrant à l'aide des informations apportés par les résultats de l'algorithme incrémental (V).

RETROSPECTIF *Observer les dynamiques* *Retracer l'historique*		
Types-U	**Tâches**	**Besoins**
Etudiant **Historien des sciences**	Mener une étude rétrospective Dresser un état de l'art Retracer l'histoire d'un domaine	1. Avoir une vue globale des thématiques du domaine. 2. Identifier les étapes de l'évolution d'une thématique 3. Revenir aux origines
Gestionnaire / Administrateur	A. Gestion et répartition des financements. B. Recherche d'experts	1. Avoir une vue globale des thématiques du domaine. 2. Identifier les thématiques mortes 3. Identifier les thématiques les plus dynamiques[53] 4. Observer la dynamique des acteurs.
Responsable	A. Rechercher des collaborations. B. Faire appel à des experts	1. Avoir une vue globale des thématiques du domaine. 2. Observer la dynamique des acteurs. 3. Identifier l'« impact » des acteurs.
Chercheur	A. Mettre à jour son état de l'art B. Mener une étude rétrospective C. Rechercher des références	1. Avoir une vue globale des thématiques du domaine. 2. Identifier les étapes de l'évolution d'une thématique 3. Identifier les auteurs spécialistes 4. Identifier les journaux spécialisés 5. Identifier l'impact de ces auteurs

Tableau 5 : Usages rétrospectifs

Le tableau 5 rend compte des besoins des utilisateurs dans une démarche rétrospective.

L'étudiant et l'historien des sciences : Pour mener un état de l'art ou comprendre l'évolution des champs de recherche, ce modèle utilisateur a besoin d'identifier les différentes étapes significatives de l'histoire d'une thématique. Il voudra aussi identifier les origines des travaux sur tel ou tel sujet, afin de remonter aux références fondatrices d'une thématique.

Le gestionnaire ou administrateur : Afin de répartir justement les financements parmi l'ensemble des équipes et des projets, il doit identifier les thématiques dynamiques auxquelles il devra apporter un soutien financier de même pour les thématiques en déclin afin

[53] Les thématiques vivantes, c'est-à-dire qui se développent, fusionnent, éclatent, apparaissent

Les thématiques qui fusionnent ou éclatent doivent elle être considérées comme vivantes, étant données que l'une des thématiques en jeu peut être morte ?

La fusion intervenant à t+1, si l'une des thématiques est morte à t, elle réintègre à t+1 une dynamique impliquée par les autre parties fusionnantes. Pour l'éclatement à t+1, nous pensons que toute thématique peut encore être considérée comme vivante, mais sera considérée comme morte à t+2 au vue de son manque de vitalité (absence de rentrée de documents, pas de fusion ou d'éclatement).

de ne pas risquer de perdre des compétences sur une thématique particulière. De plus un administrateur s'il doit faire appel à un expert doit pouvoir identifier les acteurs les plus à même de répondre à sa demande d'expertise.

Le responsable : Pour rechercher des collaborations, l'étude des thématiques et de la dynamique des acteurs est importante. Cette étude peut se faire de manière rétrospective, afin de connaître le parcours des acteurs et construire leur profil. L'étude de l'impact de ses acteurs est aussi un moyen de sélectionner le meilleur candidat. Ceci est aussi envisageable pour sélectionner un expert.

Le chercheur : Le travail d'étude de l'existant du chercheur peut demander d'identifier les différentes étapes franchies par les thématiques d'un domaine. Dans son travail de recherche de référents sur une thématique particulière, l'identification d'auteurs spécialistes, ainsi que les journaux de référence sur la question et l'impact de ces journaux, peuvent l'assister dans sa tâche.

PROSPECTIF *Détecter les émergences pour alimenter une stratégie* *Déduire des tendances*		
Types-U	**Tâches**	**Besoins**
Responsable	A. Construire et mener une politique de recherche. B. Rechercher des collaborations C. Valoriser la recherche D. Elargir ou améliorer sa visibilité	1. Avoir une vue globale des thématiques du domaine. 2. Voir les thématiques en développement. 3. Avoir une vision transversale 4. Observer la dynamique des acteurs 5. Identifier les journaux incontournables
Chercheur	A. Rechercher des collaborations B. Elargir ou améliorer sa visibilité C. S'inspirer de travaux d'autres thématiques (vision transversale)	1. Avoir une vue globale des thématiques du domaine. 2. Identifier les auteurs spécialistes 3. Identifier l'impact de ces auteurs 4. Identifier les journaux spécialisés 5. Avoir une vision transversale
Veilleur	A. Recherche d'experts et leurs collaborations B. Détecter les champs porteurs C. Détecter les émergences D. Détecter les tendances	1. Avoir une vue globale des thématiques du domaine. 2. Identifier les auteurs spécialistes 3. Identifier l'impact de ces auteurs 4. Etre alerté sur les évènements 5. Identifier les thématiques en développement 6. Identifier l'apparition de nouveaux termes 7. Avoir une vue transversale

Tableau 6 : Usages prospectifs

Le tableau 6 rend compte des utilisations prospectives, c'est-à-dire qui visent une action à venir.

Le responsable : En charge de mener une politique de recherche, il doit posséder une vue globale des thématiques de son domaine de recherche afin de se positionner parmi elles. Lle fait de pouvoir aussi percevoir les thématiques en développement lui permet de construire une stratégie visant à améliorer sa visibilité. De même qu'une vision transversale de la problématique de son équipe peut lui permettre d'élargir la portée de sa politique de

recherche. Enfin, pour être visible, le responsable peut identifier les journaux essentiels et ainsi avoir pour objectif de faire publier les travaux qu'il encadre dans ces journaux.

Le chercheur : Dans le même ordre d'idée que pour le responsable, le chercheur devra identifier les auteurs spécialistes à la recherche de collaborations, faire son choix au regard de l'impact que possède les auteurs. Enfin, en terme de visibilité, l'identification des journaux spécialisés est essentielle d'une part, d'autre part il est intéressant pour lui de pouvoir identifier des moyens de s'inspirer de travaux menés sur d'autres thématiques, à l'aide de concepts communs à d'autre thématiques.

Le veilleur : Usage prospectif par excellence, le veilleur observe la dynamique des thématiques et en dégage des tendances pour tenter d'anticiper l'avenir. L'identification du développement des thématiques sera une source d'information sur les champs plus porteurs que d'autres. Le veilleur a besoin d'être alerté sur les événements principaux de l'évolution de l'ensemble des thématiques, notamment sur l'apparition de nouveaux descripteurs. Cette fonctionnalité permet de révéler des signaux faibles, que le veilleur peut ne pas pouvoir interpréter s'il ne peut pas consulter un expert du domaine. Enfin, il peut aussi vouloir mener une veille sur un terme ou un concept particulier sur l'ensemble de thématiques, pour cela il a besoin d'une vision transversale de ces thématiques centrée sur sa requête.

En résumé, nous pouvons donc distinguer 8 besoins informationnels présentés sur la figure ci-dessous :

Figure 74 : Les besoins informationnels identifiés

La figure 74 permet de voir l'ensemble des besoins, mais avant d'aller plus loin, nous pensons qu'il peut être utile pour le lecteur d'avoir une vue synthétique de la répartition de ces besoins en fonction des types utilisateur. Le tableau 7 reprend donc les informations des tableaux 5 et 6 de manière synthétique et condensée.

X Retro.
X Pro.
X Pro. & Retro.

	Dynamique des thém.	Vue globale	Etapes évolution	Dyn. Acteurs	Dyn. Journaux	Impact Auteurs	Impact Journaux	Vue transversale	Apparition de termes	Alerte
Etudiant / Historien des sciences		X	X							
Gestionnaire / Administrateur	X	X		X						
Responsable	X	X		X		X	X	X		
Chercheur		X	X	X	X	X	X	X		
Veilleur	X	X						X	X	X

Tableau 7 : Répartition des besoins informationnels

5.3 Analyse du journal de classes aux vues des besoins informationnels

Le journal de classes est un fichier généré automatiquement au cours de l'exécution de l'algorithme, dans lequel sont présentes les données qui traduisent l'état de l'ensemble des classes[54] à différents instants de leur évolution, c'est-à-dire les classes et le contenu de ces classes.

Dans ce journal, le document est représenté par un numéro. Il est associé à un numéro de classe, c'est-à-dire la classe a laquelle il appartient et possède une densité dans cette classe. Un journal de la même forme est créé à chaque instant de l'évolution.

Nous pouvons tenter de formaliser le journal (**J**) en fonction de l'ensemble des documents qu'il contient (**Doc**), l'instant de sa création (**t**) et l'ensemble des classes (**Classes**), sous la forme :

$$J_t = \{Doc_t, t, Classes_t\}$$
$$J_{t+1} = \{Doc_t + Doc_{(t+1)}, t+1, Classes_{(t+1)}\}$$
$$\dots$$
$$J_{t+n} = \{Doc_{(t+(n-1))} + Doc_{(t+n)}, t+n, Classes_{(t+n)}\}$$

Un journal, tel qu'il est conçu aujourd'hui, possède une portée équivalente à un pas de temps, comme l'illustre la figure 75. Lors de la classification à t_1 (C_1), le journal (J_1) créé ne prend en compte que les modifications survenues entre t_0 et t_1 et ainsi de suite. Retracer l'évolution complète, nécessite de rassembler les journaux successifs.

[54] Cf. Définition d'une classe (§ 2.1..2)

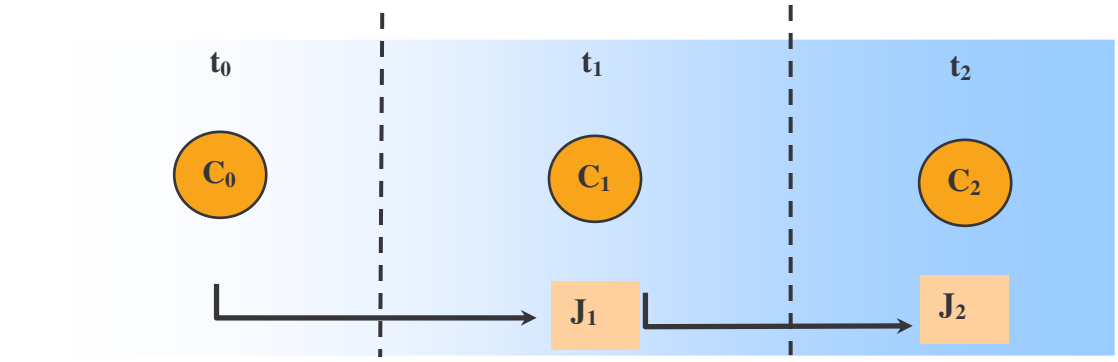


Figure 75 : Représentation de la portée des journaux

Une expérimentation a été menée et sur la figure 76 nous présentons un extrait du journal de classes réalisé à cette occasion. L'extrait présenté est un mode de représentation du journal, tel qu'il est généré par l'outil, construit à partir des données du journal.

Pour mener l'expérience, l'analyse a été menée sur un corpus de notices bibliographiques issues de la base PASCAL. La requête a été formulée de manière à ne sélectionner que des documents traitant de Géotechnique. Le premier ensemble de notices représente des documents de 2003. Un second ensemble composé de documents de 2004 a ensuite été traité par l'algorithme. L'intervalle de temps a été réduit à une année afin de simplifier l'expérience, c'est-à-dire que l'ensemble des documents n'a pas été traité de manière progressive, mois par mois par exemple.

```
**********************************************
! N° chef du noyau à jour : 751①  effectif strict : 8 ②!

...vient des noyaux anciens :
 ③
499  >   1.
751  >   5.

...et des nouveaux docs (de 2004) :
④
  2.

Origines (incluent les docs multivalents)
 ⑤
479  >   4.
741  >   4.
751  >   14.
```

Figure 76 : Journal d'une classe extrait du journal de classes

Il s'agit, dans cet extrait, du journal de la classe **751** en 2004, le numéro correspond au titre du document ayant le score le plus élevé dans la classe. Il est alors le chef de la classe (appelée noyau dans le journal).

Les informations que nous retrouvons sur la figure 76, sont les suivantes :

- ① Nom de la classe = 751 (*Loading models of tunnel supports based on stress measurements*)
- ② Nombre de documents contenus strictement dans la classe **751** en 2004 : 8 documents
- ③ Origine des documents de **751** en 2004 : 5 documents n'ont pas changé de classe puisqu'ils sont issus du noyau **751** à t_0 et 1 document est venu de la classe **499**.
- ④ Apport de nouveaux documents stricts à la classe : 2 documents de l'ensemble de 2004 se sont ajoutés à la classe **751**.
- ⑤ Nom des classes ayant fourni un certain nombre de documents à **751**, ces documents pouvant être des documents appartenant strictement au noyau (5 dans ce cas) ou pouvant appartenir à plusieurs noyaux simultanément (multivalents). Ces derniers sont au nombre de 9 (14-5=9). Les classes **479** et **741** ont chacun fourni 4 documents multivalents à **751** (il est possible qu'il s'agisse, au moins en partie, des mêmes documents).

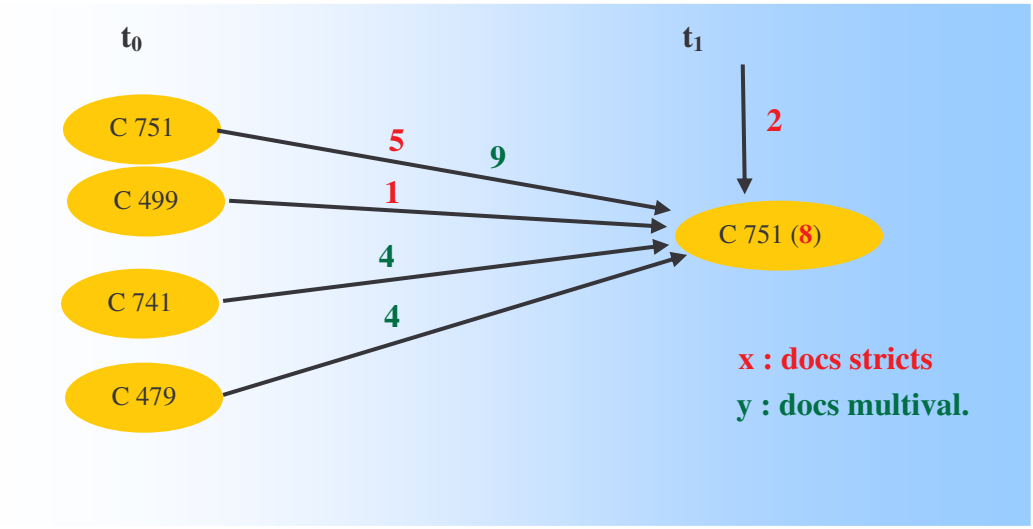


Figure 77 : Représentation graphique de l'évolution de la classe 751

Pour analyser l'évolution des classes, deux points de vue peuvent être identifiés : (**A**) soit on se place à t_0 et on analyse l'évolution de la classe à t_1, (**B**) soit on se place à t_1 et observe comment s'est construite la classe à partir des données de t_0.

Figure 78 : Sens de lecture des données du journal

La granularité de l'évolution

Représenter une évolution demande de s'intéresser à l'unité de mesure de cette évolution. Cette granularité permet de mettre en relation des représentations du temps de différentes finesses. A notre avis, il faut distinguer la granularité de l'analyse (chaque fois que l'algorithme est exécuté) et la granularité de la représentation, car elles sont indépendantes. Le premier cas désigne le rythme de création des journaux de classes, le second désigne l'exploitation de ces journaux, indépendamment de leur création.

Nous nous intéressons pour le moment au premier cas. Le rythme de création des journaux peut s'effectuer de 3 manières :

- Soit en fonction du flux de documents : un journal est créé à chaque fois qu'un nouveau document s'ajoute
- Soit à volume constant : un journal est créé lorsqu'une quantité maximale de documents est atteinte.
- Soit en fonction du besoin de l'utilisateur : L'étude des auteurs, par exemple, ne demande pas une analyse tous les mois, mais plutôt tout les un ou deux ans.

Maintenant que nous avons pris connaissance des informations disponibles dans le journal, essayons d'indiquer à quoi peuvent servir ces données en tant que réponse aux besoins informationnels que nous avons exprimés auparavant et voyons, si cela est nécessaire, comment enrichir ce journal.

5.3.1 Examiner la dynamique des thématiques

Nous avons pu voir que le gestionnaire avait besoin de connaître les thématiques qui ont déclinées et celles qui se sont développées, que le responsable tout comme le veilleur s'intéresse aux thématiques en développement, etc. Il s'agit donc pour eux de repérer la dynamique des thématiques, de les comparer les unes aux autres.

Fusions de thématiques

Il est possible, à la lecture du journal, de constater que 2 noyaux à t_0 ont fusionné en un noyau à t_1. Les informations sont déjà présentes dans le journal à t_1. Il rapporte le nom des classes impliquées dans la fusion et le nombre de documents passant d'un noyau à un autre.

```
! N°chef du noyau à jour: 745  effectif: 22 !

 ...vient des noyaux anciens :

479    >    11.
1503   >     6.
```

Fusion !

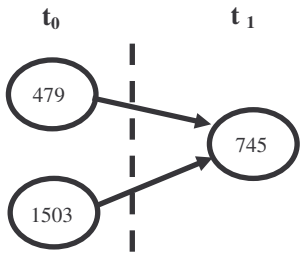


Figure 79 : Fusion de 2 thématiques dans le journal

Eclatement de thématiques

Pour décrire un éclatement, il est nécessaire de décrire la classe à partir de t_0, c'est-à-dire n'ayant pas encore subit la modification. Ces informations sont disponibles dans le journal de t_1.

```
! N°chef du noyau ancien: 1439  effectif: 7 !

 ...se reporte sur les noyaux actuels :

1439   >     4.
1442   >     3.
```

Eclatement !

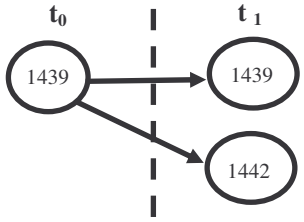


Figure 80 : Scission d'une thématique dans le journal

L'extrait ci-dessus est issu d'une autre version du journal qui effectue l'analyse selon l'approche **A** vue plus haut.

Développement de thématiques

Un noyau qui se développe est un noyau dont le volume de documents croît et qui reçoit plus de documents qu'il n'en perd. Aussi pour connaître les noyaux qui se développent, il est nécessaire d'avoir des informations sur l'effectif des documents entrants et l'effectif des documents sortants.

Un noyau qui se développe est un noyau qui ne subit donc pas d'éclatements (perte de documents en faveur d'un autre noyau) mais peux subir une fusion (récupération des documents d'un autre noyau).

Les informations concernant l'arrivée de nouveaux documents dans le noyau sont déjà présente dans le journal. Il est aussi possible de comptabiliser le nombre de documents ayant quitté le noyau pour un autre noyau. Quant à représenter ce développement, nous verrons dans la section suivante quelle méthode nous pouvons utiliser.

Déclin de thématiques

Une thématique décline lorsqu'elle ne reçoit plus de documents nouveaux. Le nombre de documents qu'elle contient reste stable et le restera jusqu'à ce qu'une autre thématique attire les documents qui lui restent. On parle alors d'émiettement de la thématique, lorsqu'elle tombe dans un état de quasi déliquescence. Là encore, il est possible de déduire l'information à partir des données du journal.

Apparition de thématiques

Trois cas de figure pour qu'une thématique apparaisse : celle-ci apparaît ex-nihilo, c'est-à-dire qu'elle n'est constituée que de nouveaux documents, ou bien elle est le résultat d'un éclatement, c'est-à-dire qu'une partie de sa population de documents appartenait à d'autres classes auparavant. Elle peut être aussi le résultat de l'agrégation de documents isolés, c'est à dire des documents seuls, n'appartenant à aucune thématique, résidus de thématiques ultérieures. Là encore les informations sont déjà disponibles dans le journal.

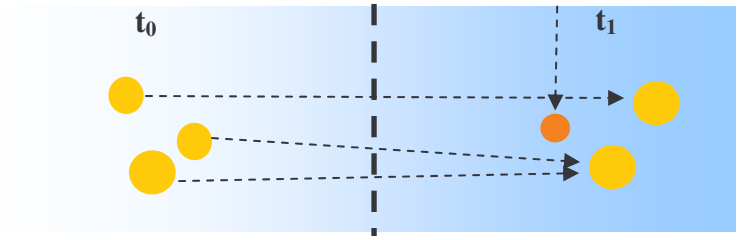


Figure 81 : apparition de thématiques ex-nihilo, uniquement à partir des nouveaux documents

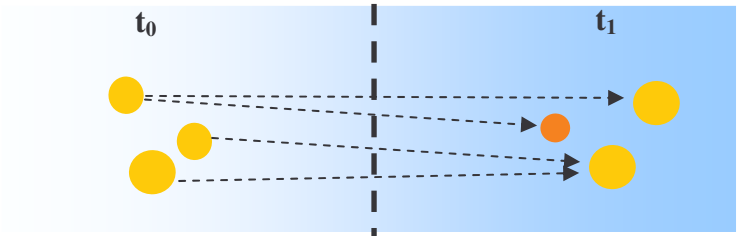


Figure 82 : Apparition de thématiques suite à une scission

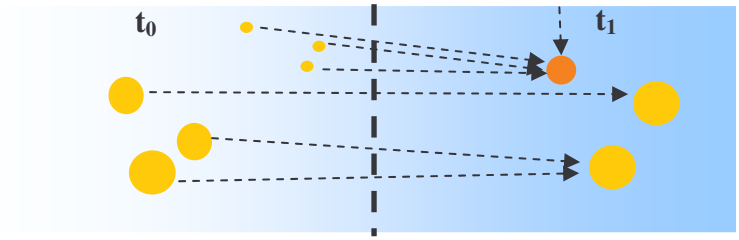


Figure 83 : Formation d'une nouvelle thématique à partir de documents isolés

5.3.2 Avoir une vue globale

Posséder une vue globale de la situation est une exigence de stratège. Elle offre en effet la possibilité d'embrasser d'un seul coup d'œil toutes les informations utiles au responsable pour la prise de décision, elle se doit d'être la plus étendue possible.

Avoir une vue globale peut être aussi considéré comme un point de départ de toute exploration de grandes quantités de données [Shneiderman, 2005]. C'est pour cela qu'elle doit être offerte à tout utilisateur.

Le journal de classes renferme toutes les informations nécessaires pour représenter la situation simultanée des thématiques à un moment donné, mais aussi de présenter l'évolution simultanée des thématiques. Mais pour obtenir une vue globale il est nécessaire de construire une visualisation qui donne la possibilité, nous l'avons vu, d'alléger les charges cognitives de l'utilisateur et donc de lui présenter davantage d'informations simultanément. Afficher par exemple une carte des thématiques ou une représentation graphique des évolutions des classes.

Nous pensons cependant que la vue globale dans notre contexte suscite une contrainte : elle doit être large et structurée à la fois. En effet, nous pensons que représenter simultanément toutes les informations sur l'évolution des thématiques à l'utilisateur est difficile car cela implique une grande quantité d'informations auxquelles s'ajoute la dimension temporelle qui nous oblige à représenter des informations sur ces informations (celles concernant leur évolution). Nous verrons dans la section 5 ce que nous pouvons imaginer pour tenter de répondre à cette contrainte.

5.3.3 Retracer les étapes et remonter aux origines

Visualiser l'évolution d'une idée, d'un ensemble de thématiques de recherche est une opération riche en enseignements pour qui veut comprendre l'état actuel de cet ensemble. Nous avons pensé que l'étudiant était le modèle type de celui qui cherche justement à comprendre et à retracer l'histoire de ce qu'il étudie. Le chercheur quant à lui appréciera de pouvoir remonter aux origines, à la recherche de références fondamentales.

Pour cela, nous pensons que le journal de classe renferme les informations nécessaires pour tracer un retour dans le temps puisqu'il est capable de retracer l'évolution, il est possible de rendre les informations contenues par le journal dans le sens inverse. En effet, à chaque exécution il réalise un cliché de chaque thématique, il suffit alors de remonter l'ordre des journaux suivant un pas de temps défini par l'importance des changements par exemple. Sur ce dernier point, nous pensons en effet que l'utilisateur n'a pas besoin de visualiser certaines étapes si celles-ci n'apportent aucune information nouvelle, distinguer des étapes revient à créer des repères discrets et significatifs dans le continuum de l'évolution.

L'utilisateur s'intéressant aux origines d'une classe sera limité d'une part par la date la plus ancienne du corpus. Si le corpus est constitué de documents datés à partir de 2000, il ne pourra pas remonter plus loin. D'autre part, il est limité par le domaine qu'il a défini lors de la sélection des documents source. En effet, si les origines de la thématique qu'il étudie se trouvent au-delà des frontières du domaine qu'il a défini, les origines seront invisibles et inaccessibles. L'enjeu est donc de constituer un corpus de document suffisamment large, ce qui requière de faire attention au changement de vocabulaire impliqué soit par l'évolution des termes, soit par le changement de contexte.

5.3.4 Observer la dynamique des acteurs

La dynamique des acteurs, c'est-à-dire les associer à des thématiques de recherche, l'observation de leurs mouvements dans l'évolution des thématiques, etc. représente une information utile pour un responsable en quête de collaborations ou pour un administrateur à la recherche d'un expert sur une question.

En l'état actuel de l'algorithme, le journal de classes ne prend pas en compte les informations liées aux auteurs. En effet, celui-ci est basé sur une étude du contenu des documents en prenant comme paramètres les termes d'indexation contenus dans les notices, les autres champs ne rentrent pas du tout en ligne de compte. Nous proposons d'enrichir le modèle du journal en y ajoutant le contenu des divers champs auteurs afin de pouvoir disposer des ces informations pour chaque classe, ainsi il sera possible de savoir qui travaille sur quoi mais aussi et surtout, de savoir qui a travaillé sur quoi, et donc quelles sont les thématiques sur lesquelles un chercheur donné a pu se pencher au long de sa carrière. En calculant les réseaux d'auteurs à partir des co-auteurs, il sera également possible de connaître les collaborations actuelles et passées des auteurs du domaine étudié.

Nous proposons aussi d'ajouter le contenu des champs « affiliation » afin de permettre à l'utilisateur de s'intéresser à l'historique d'un auteur au travers divers cadres de travail. Cela permettra aussi de retracer l'évolution des laboratoires, équipes de recherches, etc. au milieu de ces thématiques afin de voir comment évolue leur centre d'intérêt.

Enfin, nous proposons d'intégrer aussi les données des champs « pays d'affiliation » de manière à pouvoir accéder au même type d'informations que ci-dessus, mais en rapport aux pays.

5.3.5 Observer la dynamique des revues

De façon similaire à notre approche précédente sur les auteurs, l'observation de la dynamique des journaux, c'est-à-dire comment ceux-ci se positionnent sur les champs de recherche, quelles sont les revues spécialisées sur tels ou tels aspects, etc. constitue une information utile pour le chercheur et le responsable.

Pour cela, les données ne sont pas présentes dans le journal actuel. Nous proposons donc d'ajouter au journal le contenu des champs « revue » des notices, afin de pouvoir les positionner sur la carte des thématiques et suivre leur évolution parmi celles-ci.

5.3.6 Percevoir l'impact des auteurs et des revues

Les informations sur l'impact, c'est-à-dire l'influence qu'a pu exercer un auteur ou une revue sur le chemin parcouru par une thématique, sont des informations très intéressantes pour un responsable devant piloter une activité scientifique ou à la recherche d'experts, un chercheur en quête de visibilité ou de collaborations ou encore un historien des sciences cherchant à justifier l'évolution d'une ou plusieurs thématique(s).

Malheureusement, le calcul de l'impact ne peut s'effectuer qu'à l'aide d'une analyse des citations des documents. Or nous ne possédons pas ces données dans la base PASCAL. L'unique calcul pouvant être fait au sujet des auteurs est le nombre de publications par auteur. Cependant, ce nombre n'est pas représentatif de l'impact que peut avoir un auteur. Celui-ci peut très bien publier beaucoup et avoir peu d'influence. C'est pourquoi nous disons que le calcul ne peut s'effectuer qu'avec une analyse de citations. Appliqué à une base comme PASCAL, l'algorithme ne peut pas calculer cet impact, mais il est possible d'imaginer qu'avec des données d'autres bases comme le *Web of Science (ISI)*[55], *Citeseer*[56] par exemple, il est possible alors de déterminer quel auteur a eu un certain impact sur une thématique ou sur un domaine entier, et comment cet impact a pu évoluer.

5.3.7 Avoir une vision transversale

Le chercheur dans son travail de recherche ou le responsable dans la conduite de sa politique de recherche ou sa volonté d'élargir sa visibilité, voudront peut être voir ce qu'il se passe dans les autres thématiques et étudier la possibilité de s'inspirer de travaux menés à l'intérieur de celles-ci, dans leur travail.

Il est possible de traduire cette notion de transversalité, de deux manières : soit par les documents communs à plusieurs thématiques (documents multivalents), soit par les termes partagés par ces thématiques. Ses données sont accessibles en sortie de l'algorithme.

5.3.8 Repérer l'apparition de nouveaux termes

L'apparition de nouvelles thématiques est déjà un point plein d'enseignements pour un veilleur. Nous pensons que cela n'est cependant pas un indicateur suffisamment précis pour détecter l'apparition de nouveautés. Pour affiner cet aspect, nous avons pensé que le veilleur devait pouvoir repérer les termes apparaissant pour la première fois. Car il est possible que l'apparition de nouveaux termes passe inaperçu si ceux-ci se font attirer dans une thématique et que ces termes n'apparaissent que tardivement, une fois que des termes similaires viennent les rejoindre, occasionnant du coup un éclatement des thématiques.

[55] http://scientific.thomson.com/products/wos/
[56] http://citeseer.ist.psu.edu/

Dans cet optique nous devons disposer d'informations sur les nouveaux termes, ceux-ci sont présents sur les notices et l'algorithme se base sur eux pour construire sa classification. Il serait donc possible de les récupérer, de les comparer, etc. Il s'agit alors d'alerter le veilleur de l'arrivée de nouveaux termes.

Cependant, cette notion de nouveaux termes pose un problème essentiel : comment distinguer une réelle apparition d'un terme, d'une simple modification de vocabulaire d'un terme existant déjà auparavant mais ayant changé au gré des modes, des tendances, ou des contextes d'utilisation.

Nous pensons que ce point admettra difficilement une réponse définitive. Si le terme identifié comme nouveau est une dérivation d'un terme existant totalement équivalent, il est probable que ce nouveau terme est changé de contexte. Il est possible aussi qu'il n'ait pas changé de contexte, seule l'instant de son énonciation est différent, ce qui s'apparenterait de près ou de loin à une nouvelle tendance d'utilisation. Ce nouveau terme peut être une nouvelle équivalence d'un terme existant, mais en l'utilisant les auteurs n'ont-t-ils pas voulu enrichir l'ancien terme par l'ajout de nouveaux aspects sémantiques ?

Par conséquent, il peut être dangereux de laisser le système évaluer cela automatiquement, nous pensons que la détermination d'une règle d'équivalence entre un nouveau terme et un ancien, revient à l'utilisateur capable d'interpréter, de juger, si cette équivalence est valide ou non.

5.3.9 Être alerté

Pour le veilleur, la fonction d'alerte est essentielle pour que lui soit signalées les modifications, les tournants importants. Le système doit pouvoir lui pointer les événements important sans qu'il ait besoin de les déduire des phénomènes observables.

Pour cela, nous pensons qu'il serait utile, en sortie de la comparaison des cartes à différents instants, de faire ressortir les événements pour les signaler immédiatement au veilleur. Nous verrons dans la partie visualisation comment faire ressortir les événements, comment créer une alerte.

5.3.10 Conclusion partielle

Il semble que l'essentiel des informations nécessaires à l'étude de l'évolution soient déjà présentes dans le journal que nous pouvons exprimer, pour un instant *t*, par le formalisme suivant :

```
<journal>
   <date/>
   <classe>
       <N°chef>
       <nom_chef>
       <document>
              <N°doc/>
              <Densite/>
              <Titre/>
              <Auteurs/>
              <Affiliation_auteurs/>
              <Pays_affiliation/>
              <Source/>
              <Descripteurs/>
       </document>
   </classe>
</journal>
```

Au regard des besoins informationnels des utilisateurs, nous avons pu identifier les nouvelles données à intégrer dans le journal et dans la visualisation, que sont les auteurs, les affiliations, les pays associés et les sources des documents. Cet apport ne change rien à la manière de fonctionner de l'algorithme, il est uniquement un apport informationnel dont nous devrons tenir compte lorsqu'il faudra déterminer les fonctionnalités de visualisation.

5.4 Fonctionnalités de visualisation des informations utiles à l'utilisateur

Ce qui nous occupe à présent est de trouver des techniques qui permettent de visualiser une information en évolution. Nous présentons ici quelques propositions pouvant alimenter une réflexion sur le sujet. Quelques unes de ces techniques ont déjà été aperçues lors de notre état de l'art et lors des observations menées sur les outils de veille. Nous tenterons de déterminer des techniques adaptées à partir de ce qui existe :

- A nos données (dans notre cas, des descripteurs, des noms d'auteurs / affiliations / pays / journaux, des densités de documents, des effectifs de classes),
- A nos types utilisateurs, c'est-à-dire à leurs tâches et à leurs besoins informationnels

Nous pensons en effet, à l'instar de Dâassi (1999) et Bonnel et Chevalier (2006), que les techniques de visualisation sont contraintes par ces deux paramètres. Toute la difficulté est donc de trouver l'adéquation entre ceux-ci.

5.4.1 Visualisation globale de la dynamique des classes et de l'historique de leur évolution

Retracer la dynamique de classes, ce qui désigne, rappelons-le, les phénomènes de fusion, éclatement, développement, déclin et apparition de thématiques, présente de nombreuses difficultés graphiques. Les réflexions qui vont suivre concernent les besoins utilisateurs « Examiner la dynamique des thématiques », « Avoir une vue globale » et « Retracer les étapes », auxquels elles permettent d'apporter des éléments de solution.

5.4.2 Une représentation dynamique pour visualiser une dynamique des classes

La métaphore du paysage (figure 84, 87) est à notre avis le mode de représentation la plus efficace pour cela. En effet, elle permet une vue globale de toutes les thématiques en suivant l'évolution temporelle de manière interactive par affichage successif des paysages de chaque instants, comme l'effectue le logiciel Vxinsight (*cf. 3.5.6.*).

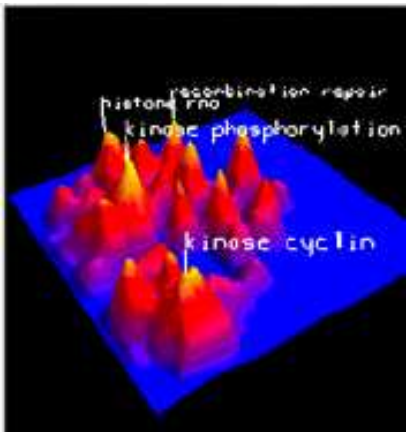


Figure 84 : Paysage de classes (Omniviz)

Approche cartographique

L'approche cartographique, au sens géographique du terme, place les éléments de la carte suivant une position (coordonnées) bien déterminée. Nous envisageons de traiter de cette manière les liens qui existent entre certaines thématiques de recherche, créés par les documents multivalents.

En effet, il est intéressant pour l'utilisateur de voir ces liens : plusieurs classes proches partageant une certaine quantité de documents, si l'utilisateur aperçoit un développement similaire de ces classes, il peut en conclure qu'il existe une interaction et une dynamique entre ces classes. De même, il est intéressant de pouvoir voir les thématiques aux sujets proches.

Nous proposons de représenter ces documents multivalents sur la carte dynamique des thématiques. Cette représentation prend la forme d'une distance relative au nombre de documents partagés, entre les classes. Ainsi, les classes qui partagent beaucoup de documents seront proches et celles qui n'ont rien en commun seront éloignées. Cette relative proximité des classes, fixe leur position et c'est en cela que cette approche est cartographique.

Densité de données dans la représentation

Avoir une vue globale et exhaustive de l'ensemble des données à visualiser est un atout, d'une part pour l'utilisateur cherchant à tirer une information à partir de la globalité, et d'autre pour l'utilisateur qui s'apprête à explorer un univers important dans lequel il sera amené à se repérer. Cependant se pose le problème de savoir à partir de quel point une vue globale sur les données devient plus encombrante qu'enrichissante. Pour qu'une représentation graphique reste efficace il est nécessaire qu'elle soit lisible et cette lisibilité dépend du nombre de classes à représenter, de leur dynamique et les besoins des utilisateurs.

Selon Tufte (2001, p. 162), la densité de données dans un graphique se définit de la manière suivante :

$$\textbf{Densité} = \frac{\textbf{Nombre d'entrées dans la matrice}}{\textbf{Aire du graphique}}$$

Cependant il est difficile de déterminer a priori l'intervalle de densités entre lesquelles le graphique reste lisible. Dans notre cas, le nombre d'entrées est le nombre des classes à visualiser. Le principe de l'algorithme est que ce nombre est amené à varier au cours du temps. Si l'on veut garder une densité constante et acceptable pour l'utilisateur, il est

nécessaire de faire évoluer l'aire du graphique en fonction du nombre de classes à visualiser. Mais il est clair qu'au-delà d'une certaine limite, même en ayant une densité acceptable, la visualisation sera surchargée et l'utilisateur ne pourra pas tirer beaucoup d'information à partir de la représentation, dans ce cas, la globalité s'obtient au détriment de la lisibilité et vice versa.

Si l'on veut conserver une vue globale malgré un trop grand nombre d'éléments présents sur la carte, la carte va perdre de sa lisibilité pour l'utilisateur. En effet, pour représenter le plus grand nombre d'éléments dans l'espace d'un écran d'ordinateur, il sera nécessaire de prendre de la distance (changement d'échelle).

Si l'on veut conserver à la fois la globalité et la lisibilité (c'est-à-dire rester à une distance acceptable de la représentation), l'utilisateur perd la globalité du graphique. Il sera en effet obligé de se déplacer sur la représentation pour percevoir son ensemble.

Les éléments de solution pouvant être apporté à ce problème sont la présence d'une vue globale toujours à la portée de l'utilisateur, comme le fait le logiciel GoogleEarth dont est issue la figure 85 et la déformation du plan.

Figure 85 : Une vue globale aise à se repérer lorsque l'on accède aux détails (GoogleEarth).

Avec GoogleEarth, l'utilisateur parcourt la carte sur les axes Nord-Sud, Est-Ouest et aussi en altitude. Il lui est toujours possible se référer à la vue globale lorsque la vue détaillée ne lui permet plus de savoir exactement où il se trouve.

L'autre solution consiste à déformer la carte à ses frontières, comme le font les représentations hyperboliques ou en « fish eye » (*cf. 3.5.4, 3.7.1*). De cette manière l'utilisateur peut accéder au contenu détaillé sans que cela ne s'effectue au détriment de la globalité et de la lisibilité. Sur la figure 86 nous avons réalisé, à l'aide d'un montage, ce que pouvait donner une déformation de la carte appliquée à la figure 85. L'utilisateur perçoit l'information détaillée tout en gardant son contexte utile pour se repérer et voir toutes les données simultanément.

Figure 86 : La déformation du plan n'isole pas les détails de leur contexte.

Animation et interactivité de la représentation

La dynamique d'une représentation peut s'effectuer de deux manières : interactive ou animée. Le second cas est intéressant lorsque l'utilisateur veut présenter à une autre personne, l'évolution de son corpus de documents. Le premier cas est la solution la plus adéquate pour permettre à l'utilisateur de comprendre la représentation (*cf. 3.6.5.*). En effet, en faisant glisser un curseur (Time slider), il visualise à sa guise les frontières des classes qui se transforment, les apparitions et les disparitions, etc. La carte est dynamique mais reste sous le contrôle de l'utilisateur qui, à tout moment, peut revenir en arrière, figer la carte, etc.

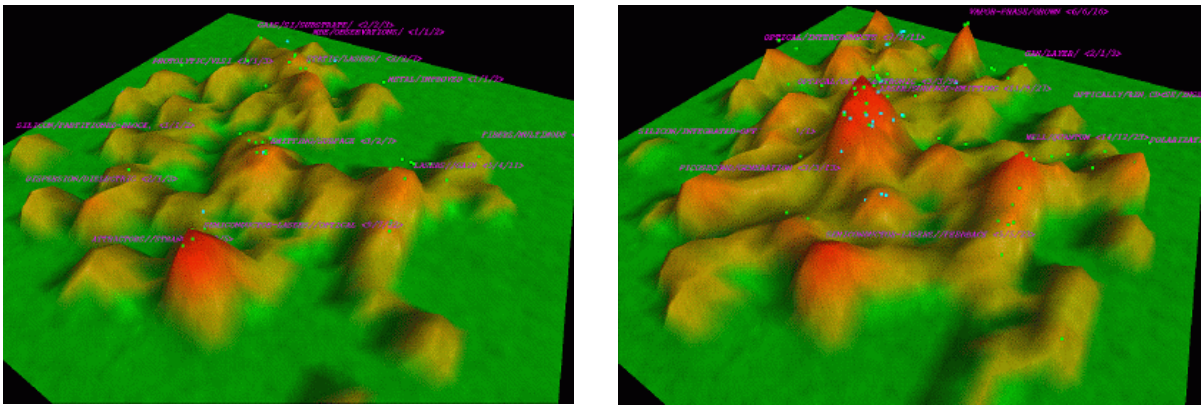


Figure 87 : Evolution de la carte des thématiques sur 10 ans (Vxinsight)

Cependant, une telle représentation n'apporte pas toutes les informations nécessaires à nos utilisateurs. Elle ne permet pas, par exemple, de rendre compte de l'origine des modifications. En effet, si un des sommets de la figure 87 s'élève progressivement, comment savoir d'où viennent les documents ? Comment déterminer leur origine ? Cette carte ne

permet pas non plus de distinguer avec précision les thématiques les plus dynamiques, celles qui se développent plus que d'autres, ou à l'inverse celles qui déclinent dangereusement. Cela dépend de l'ampleur du développement : une thématique se développant beaucoup et rapidement sera sans doute plus visible qu'une autre ayant un développement moins marqué. Pour le savoir, il est nécessaire d'avoir une approche quantitative plus fine pour confirmer l'évolution même minime d'une classe. De plus, cette représentation a des limites inhérentes aux représentation 3D : d'une part certains éléments peuvent cacher d'autres éléments et d'autre part, lorsque la carte atteint une certaine taille, les phénomènes de perspectives tendent à fausser les mesures.

Nous voyons donc que cette représentation ne permet pas de visualiser toutes les informations. Dans notre cas, il est nécessaire d'entrer dans une démarche de comparaison des classes. Ces comparaisons sont de deux natures : nature quantitative (développement des thématiques) et nature qualitative (fusion, éclatement, apparition de thématiques).

5.4.3 Visualiser l'évolution quantitative à l'aide d'un indicateur de développement

Comment rendre compte du développement des thématiques ? Comment comparer l'évolution des effectifs des classes ? Comment déterminer qu'une thématique se développe plus qu'une autre ?

Une représentation comme le système TGRIP (Figure 88) développé par Erten et al. (2003) permettrait de percevoir les développements et les déclins de thématiques, leur graphe étant prévu pour évoluer en fonction du volume de documents contenu dans les classes. Il s'agit d'une application du principe décrit par Bertin (1967), l'exploitation d'une variable visuelle qui est la taille comme variable quantitative. Cependant, il est parfois difficile de dire avec précision à quel point une classe s'est développée, de plus ce mode de visualisation, n'ordonnant pas les informations qu'elle communique, ne permet pas de savoir quelle catégorie s'est le plus développée sur un intervalle de temps donné.

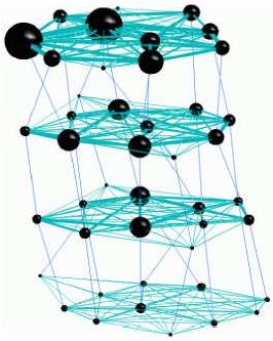


Figure 88 : Visualisation de l'évolution par le *TGRIP*

L'approche graphique que nous pouvons envisager dans notre cas est l'exploitation de diagrammes. L'approche paraît simple mais selon l'un des principes d'intégrité graphique énoncés par Tufte (2001), *le nombre de dimensions utilisées pour représenter une donnée ne doit pas excéder le nombre de dimensions de cette même donnée*. Aux vues de ce principe, nous proposons de visualiser le développement et le déclin des thématiques à l'aide d'une représentation en diagramme.

La première proposition que nous émettons sur cet aspect est donc de rendre compte de la dynamique des classes à l'aide d'un diagramme analysant un indicateur de développement (ID) en fonction du temps. Le temps peut être considéré comme la cause de l'évolution et selon la théorie des graphiques de données de Tufte (2001), pour exprimer la

cause par rapport à l'effet, l'effet sera représenté sur l'axe vertical, la cause sur l'axe horizontal (figue 18).

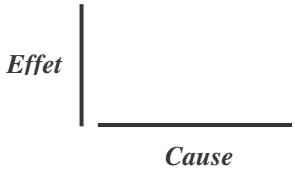


Figure 89 : Représentation graphique des causes et effets

Cette proposition vaut pour visualiser, à la manière de la figure 90, le développement (1) et le déclin (2) de thématiques.

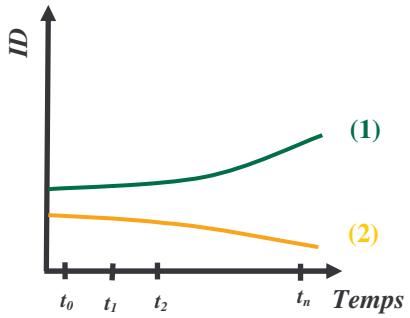


Figure 90 : Diagramme de développement des classes

Cet indicateur de développement est calculé à partir de l'apport de documents entre t_0 et t_1, cet apport (Δ) est la différence entre les documents entrants (E) et sortants (S). Sur le graphique de la figure 90, $ID(t_1)$ = Effectif de la classe à $t_0 + \Delta$, ce qui permet de rendre compte de la variation des effectifs. Cependant, afin de relativiser le développement des thématiques par rapport à ces effectifs, il est nécessaire de ramener le résultat en pourcentage. En effet, un apport de 5 documents, par exemple, n'a pas la même importance pour une classe composée de 10 documents que pour une classe de 50 documents. Une représentation complémentaire, avec un ID exprimé en pourcent, pourra être proposée.

Néanmoins, pour que l'utilisateur puisse comparer les développements des thématiques afin de dire quelle a été la plus forte progression sur un intervalle donné, cette représentation n'est pas la mieux adaptée.

Le système ThemeRiver de Havre et al., (2002), permet une approche quantitative. Nous pouvons en effet imaginer visualiser l'évolution de la quantité de documents d'une thématique à différents instants comme la figure 91. La représentation par accumulation des quantités (les classes sont représentées les unes sur les autres) devrait permettre de comparer aisément les différents éléments du graphique. L'avantage du système ThemeRiver par rapport à TGRIP est que la progression est plus perceptible, d'une part par le jeu des couleurs qui permet de distinguer les classes les unes des autres, d'autre part du fait de la continuité des courbes. En effet, ce que l'on appelle interpolation, c'est-à-dire l'opération mathématique qui effectue une estimation de la position de points en fonction d'autres points, permet de créer un aspect lissé et continu à partir de données discrètes. Cet aspect donne au regard la possibilité de suivre la progression sans à coups.

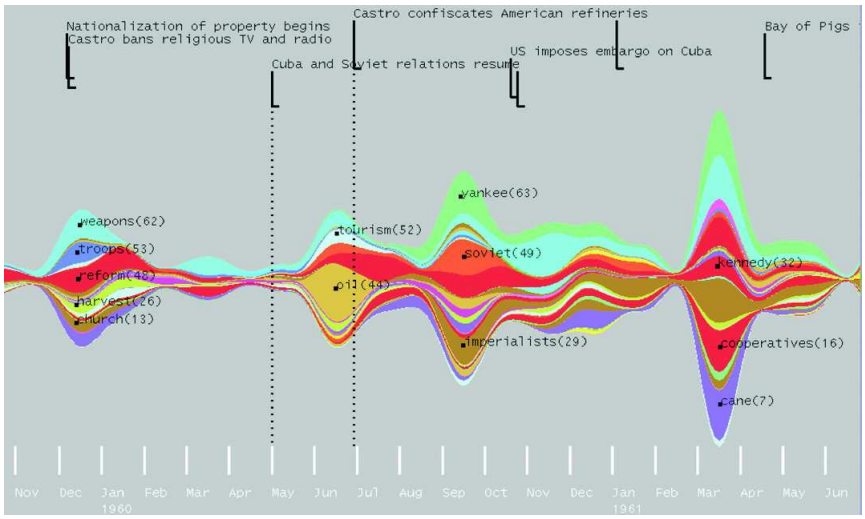


Figure 91 : Visualiser l'évolution de fréquences de mots (quantitatif) avec ThemeRiver

Cependant, afin de mener une comparaison globale de l'évolution des thématiques sur un intervalle de temps donné, nous proposons que l'indicateur de développement soit visualisé sous forme de diagrammes ordonnés, triés par ordre décroissant (figure 92).

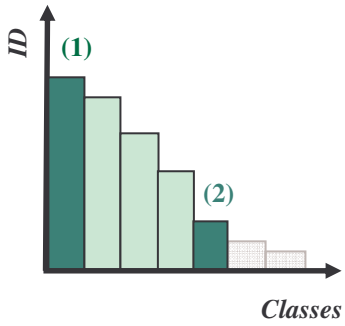


Figure 92 : Développement comparé des thématiques à l'aide d'histogrammes

Ainsi l'utilisateur pourra d'un seul coup d'œil s'apercevoir que sur la figure 92, la classe (1) se développe fortement alors que la classe (2) ne se développe pas beaucoup. ID sera, dans ce cas, la différence entre l'effectif à t_{max} et t_{min}, ceux-ci étant les bornes de l'intervalle de temps.

Visualiser ainsi le développement de thématiques dans une perspective prospective, telle que nous l'avons définie, permet bien de donner à l'utilisateur les moyens de détecter rapidement une tendance.

5.4.4 Visualiser une évolution qualitative à l'aide d'une représentation linéaire

Les représentations de Erten et al. et de Havre et al. sont intéressantes car elles permettent de concentrer les différents instants de l'évolution sur une seule vue. Erten et al. proposent de faire évoluer les sommets en fonction de leur volume, que se passe-t-il lorsque cette évolution n'est plus uniquement quantitative, mais qualitative comme c'est le cas lors d'apparitions, de fusions et d'éclatements, etc. L'enjeu est ici de définir une manière de représenter les différentes étapes de l'évolution, de repérer les ruptures, de permettre à l'utilisateur de reconstituer la généalogie des classes.

Une représentation linéaire (comme le sont les frises chronologiques ou Time Line) semble être une représentation naturelle pour représenter l'histoire de chacune des thématiques.

Elle pourrait consister en un tracé de l'évolution à l'aide d'arcs dirigés d'un état d'une classe à un instant et son état aux instants suivants, comme le montre la figure 93[57].

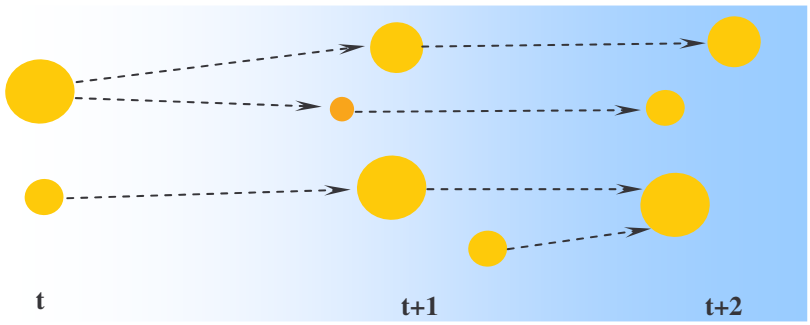


Figure 93 : Représentation linéaire de l'évolution

Cependant cette représentation pose des problèmes lorsque le nombre de classes devient très important. La disposition des classes devant respecter les contraintes de position sur le plan, liées à la représentation linéaire, les arcs se croisent, se multiplient et finalement construisent une représentation globale devenant difficilement exploitable par l'utilisateur.

On peut aussi imaginer utiliser le principe de plans superposés, faire évoluer les classes sur des plans (sorte de clichés de l'ensemble des classes à un instant donné) disposés le long d'un axe vertical, comme le fait le TGRIP.

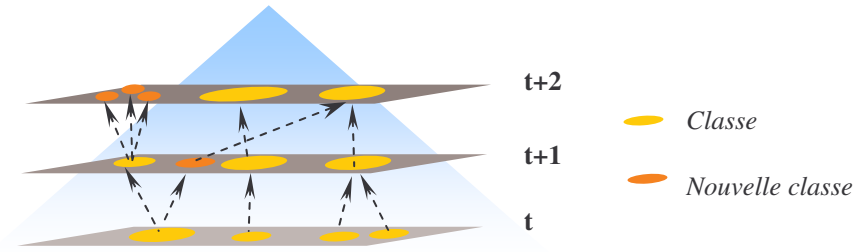


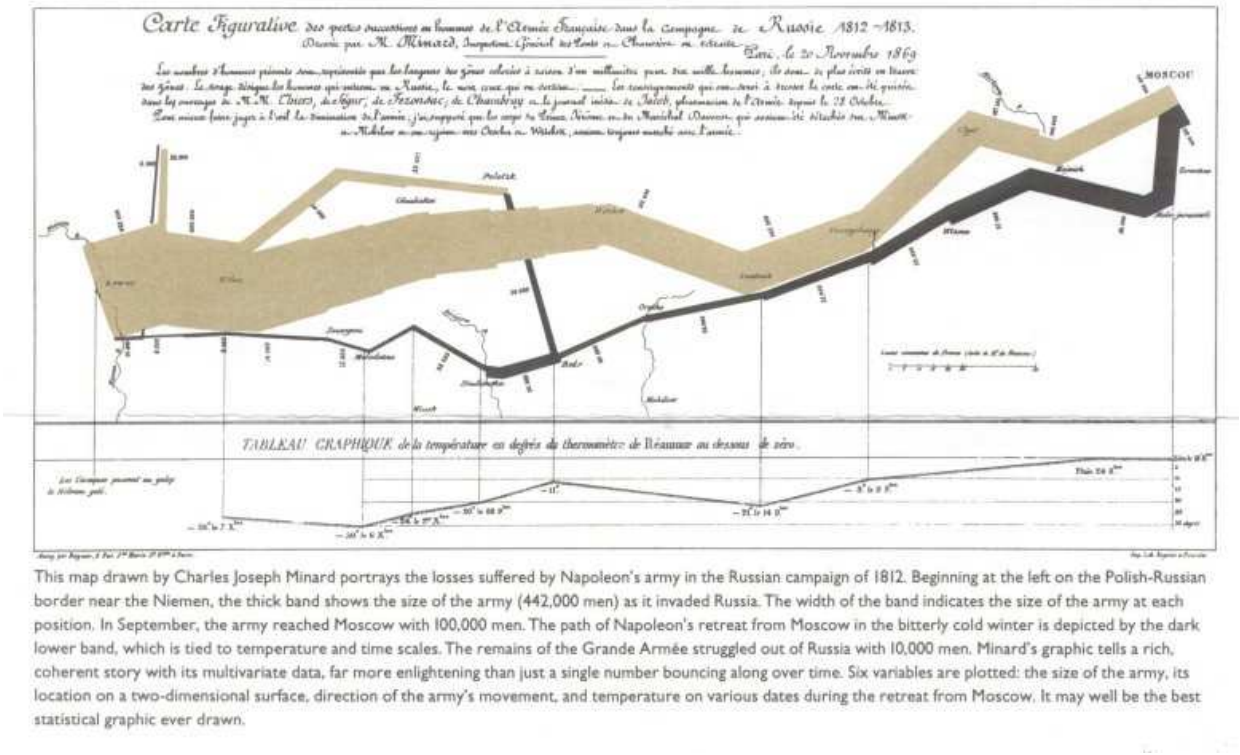
Figure 94 : Evolution sur plans superposés (vertical)

La figure 94 est une représentation simplifiée de l'évolution qualitative des classes à l'aide de plans superposés. Nous pouvons remarquer que cela permet une représentation efficace de la vue globale des différentes classes, en revanche, il est difficile d'imaginer retracer l'évolution sans utiliser de liens (flèches sur la figure 94) entre les différents instants qui permettent à l'utilisateur de suivre cette évolution et de la comprendre. Là encore, l'utilisateur a de fortes chances de se perdre au milieu des liens entre les plans.

[57] Nous employons le dégradé de bleu afin de représenter au lecteur, l'évolution dans le temps.

Les tracés d'évolution

Lorsque Minard[58] voulu retracer le périple de l'armée de Napoléon lors de la campagne de Russie (1812) en représentant le trajet et les effectifs de son armée au fil de ce périple, il construisit la carte de la figure 95 ([Tufte, 2001], pp. 40, 176) :



This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (442,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn.

Figure 95 : Carte figurative des pertes successives en hommes de l'armée française dans la compagne de Russie 1812-1813

On y retrouve d'une part une dimension temporelle puisque la progression marquée géographiquement est temporelle : de gauche à droite pour l'aller et de droite à gauche pour le retour. D'autre part, Minard a su faire figurer une dimension quantitative représentant l'effectif de l'armée au long de sa progression, représentée par l'épaisseur du tracé. Une troisième dimension fait aussi état des températures qu'ont dû endurer les soldats au retour (courbe des température en bas de la carte).

Cette célèbre représentation bientôt bicentenaire nous rappelle le spectrographe conçu pour visualiser l'évolution de classes dans le logiciel Autonomy qui rentre dans le cadre d'une représentation linéaire (figure 96) et possède les même limites que toutes représentations de ce type : graphiquement, que se passe-t-il lorsque les classes éclatent en de nombreuses autres classes ? Que se passe-t-il, par exemple lorsque nombre d'entre elles fusionnent et que la classe produite éclate par la suite ? La visualisation restera-t-elle lisible par l'utilisateur ? Nous avons essayé de retracer en figure 97, ce que pourrait donner une telle technique avec les données de l'algorithme que nous étudions. Chaque ligne correspond à une thématique et ses lignes se rejoignent et divergent pour représenter les fusions et les éclatements. La représentation est sans doute intéressante, on distingue bien les différents phénomènes et les classes impliquées dans l'ordre des événements. On imagine même pouvoir y ajouter une dimension quantitative, comme le fait Autonomy, pour visualiser les effectifs des classes. En revanche, passé un certain nombre de classes et de phénomènes, la visualisation risque de se compliquer au point de ne pouvoir se faire comprendre par le lecteur.

[58]Charles Joseph Minard (1781-1870), directeur de l'école nationale des ponts et chaussées (1830-1836). .

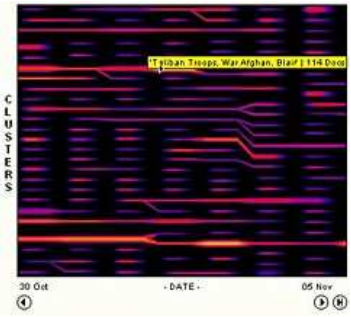


Figure 96 : Représentation linéaire de l'évolution des classes (Autonomy)

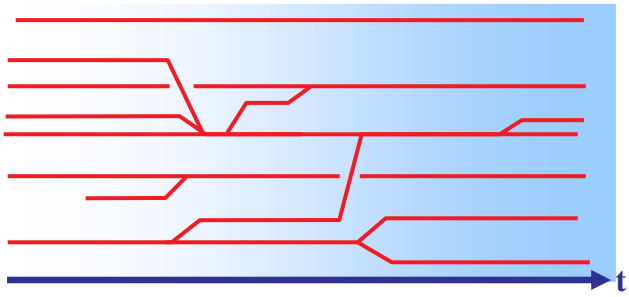


Figure 97 : Spectrographe appliqué à l'algorithme de classification incrémentale

Interaction utilisateur : du global au particulier

A la vue des éléments de réflexion que nous venons de présenter, il semble qu'il est difficile de concevoir une représentation suffisamment claire pour visualiser les étapes de l'évolution de chaque thématique. C'est pourquoi, nous proposons que l'utilisateur puisse interagir avec le système en sélectionnant les thématiques dont il veut suivre l'évolution. Ainsi, le tracé de l'histoire de cette thématique émergera de la masse de tracés des autres thématiques, à l'aide de couleurs nettement distinctes et il lui sera possible de remonter aux origines.

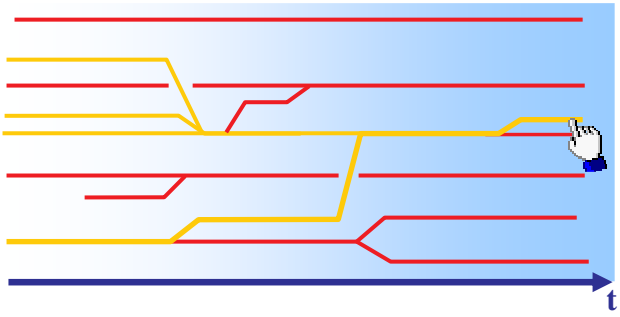


Figure 98 : Sélection d'une thématique et visualisation de son historique

La figure 98 montre comment un utilisateur peut visualiser l'évolution d'une thématique parmi d'autres en la sélectionnant (*cf.* pointeur de la figure). Il peut vouloir remonter les étapes franchies par cette classe, comme le montre la figure 98, il peut aussi vouloir voir comment une thématique va évoluer au cours du temps comme l'illustre la figure 99.

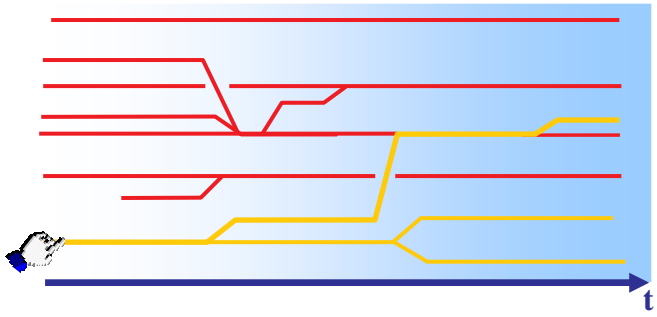


Figure 99 : Sélection d'une thématique et visualisation de son évolution

L'utilisation de la couleur comme variable sélective a été approchée par Bertin (1967). A ce sujet, ce dernier préconise l'utilisation de tons « purs », c'est-à-dire sans mélanges d'autres couleurs, c'est-à-dire une bande très étroite dans le spectre coloré.

Figure 100 : Le spectre des couleurs

Pour faire davantage ressortir les lignes nous proposons par exemple d'employer une couleur de fond qui tranche avec les couleurs utilisées pour représenter l'information[59]. La figure 101, par exemple, utilise le noir comme couleur de fond, faisant ressortir les moindres détails colorés à l'aide de couleurs claires, une nouvelle sélection avec une nouvelle couleur permet une visualisation simultanée de multiples évolutions.

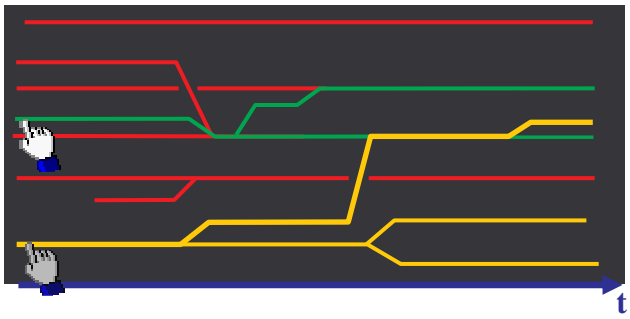


Figure 101 : Le noir fait ressortir les couleurs claires et franches

5.4.5 Alerter le veilleur des changements

Le principe d'une alerte pour un veilleur est de pouvoir être averti des changements qui sont intervenus sans qu'il ait besoin de les suivre lui-même régulièrement. Une alerte est paramétrable, c'est-à-dire que le veilleur doit pouvoir spécifier ce qu'il surveille : les classes et les comportements de ces classes. Il est parfois souhaitable dans une démarche de surveillance, d'être alerté même par un mouvement peu significatif. Nous proposons donc que tous les changements puissent faire l'objet d'une alerte.

Cependant, concernant le développement de thématiques, il est préférable de définir un seuil au-delà duquel le veilleur est alerté. Ce seuil peut être déterminé à partir de l'indice de développement, en prenant en compte la taille de la classe. Seuls les développements importants et rapides peuvent lui apporter une information intéressante concernant une possible émergence.

Graphiquement ces alertes pourraient prendre la forme de symboles apparaissant sur le tracé des classes, à l'endroit ou un nouveau changement est apparu. Les symboles correspondraient aux catégories de changements afin de favoriser l'identification externe, « *qui repose sur les habitudes acquises, la reconnaissance de mots, de formes ou de couleurs* » ([Bertin, 1967], p.140). Ainsi les symboles > pour les fusions, < pour les éclatements et ⊢ pour les apparitions, pourraient être utilisées.

Cependant, toujours d'après Bertin (1967), la forme n'est pas une variable sélective, car elle ne permet pas de répondre à la question « *telle catégorie de symbole, où est elle ?* »

[59] Ce qui n'est pas le cas du dégradé de bleu employé jusqu'à maintenant dans nos illustrations, auquel nous avions attaché une signification de progression dans le temps.

(p.95). Or l'utilisateur se posera probablement la question de savoir « quelles sont les fusions de classes, quelles sont les scissions, etc. ? » Pour répondre à ce besoin de filtrer nous envisageons d'employer la méthode de requête dynamique (*cf. 3.7.1*) basé sur les catégories de symboles qui permet à l'utilisateur de ne voir que les fusions, les éclatements, etc. de manière interactive et dynamique.

Enfin, puisque nous avons fait le choix de donner à l'utilisateur un nombre maximal d'alerte, une fois alerté de l'existence de nouveaux éléments, celui-ci devrait pouvoir trier ce qui a de l'importance et ce qui en a moins. Pour ce faire, nous préconisons d'employer des seuils que l'utilisateur fixera lui-même pour que s'affichent uniquement les éléments importants. Par exemple, nous pouvons dire qu'il est possible qu'un éclatement soit significatif si la quantité de documents qui s'extrait de la classe est d'au moins 20% de l'effectif de cette classe. Nous prendrons soin d'indiquer à l'utilisateur à quoi peuvent correspondre ces seuils, quelles significations on peut donner aux intervalles, etc.

5.4.6 Représenter les thématiques en fonction d'une requête de l'utilisateur

L'idée que nous développerons ici est qu'un utilisateur voudra, par exemple, identifier toutes les thématiques existantes autour d'un concept. Cette fonctionnalité répond notamment au besoin d'avoir accès à une vision transversale. L'utilisateur exprime alors le concept lors d'une requête à l'issue de laquelle le système lui présente toutes les thématiques autour de ce concept. La notion de concept ici peut aussi bien désigner un terme qu'une équation de recherche formulée par l'utilisateur. On peut imaginer utiliser une représentation dite héliocentrique [Moya-Anegon et al., 2004] pour visualiser les thématiques (Th) autour du concept de la requête (R).

Il est possible de calculer un taux d'occurrence (*f*) des termes dans chaque thématique. Ce taux serait l'occurrence du terme dans la classe rapporté à l'effectif global de la classe (strict et multivalent). Des zones concentriques permettraient de rendre compte de la pertinence des thématiques relativement aux autres, comme l'illustre la figure 102.

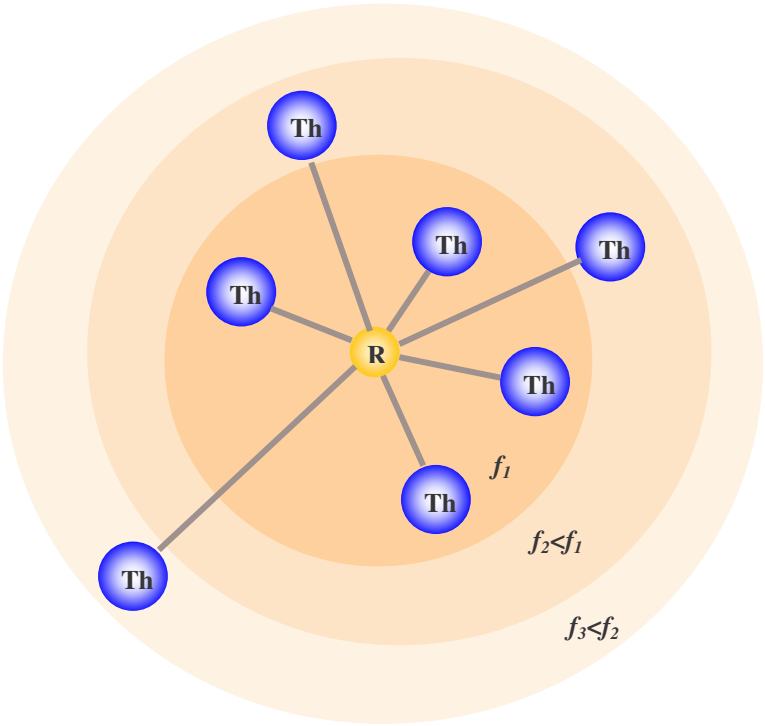


Figure 102 : Visualisation héliocentrique permettant de voir les thématiques associées à une requête

L'optique de faire évoluer ce graphe est intéressante : l'utilisateur peut vouloir observer l'évolution des thématiques ou de leur pertinence en fonction du concept de son choix. A l'aide de la théorie des graphes, il est possible de déterminer si ce graphe, qui est un graphe valué (on attribue aux arcs les valeurs de *f*) d'un instant t_0 et celui équivalent à la même requête à un instant $t_1$sont *isomorphes*[60]. Il est possible aussi de calculer la *distance d'édition*[61] (ou *mesure de similarité*) de ces deux graphes (Sorlin et Solnon, 2003). Les différences observées seront rendues visibles des catégories de modifications (liste des ajouts, des suppression, etc.) ou par des marqueurs sur le second graphe (couleurs par exemple) indiquant les thématiques ayant changé, à l'utilisateur qui pourra alors accéder à leur historique (indice de développement, tracé d'évolution, etc.) sur la période t_0 à t_1.

5.5 Synthèse

Pour conclure cette partie sur la définition des fonctionnalités de visualisation pour l'analyse de l'évolution d'un domaine scientifique, nous présentons, au travers la figure 103, la structure des fonctionnalités que nous proposons.

Cette structure se base sur 4 pôles essentiels aux travers desquels nous proposons de visualiser l'évolution des thématiques de recherche. A partir du journal de classes, nous avons envisagé les quatre approches : globale, quantitative comparée, qualitative et par la requête.

L'utilisateur peut rentrer dans le système par 2 points : Vue globale par la carte (Exploration) ou Requête (Besoin *précis*).

Après avoir formulé sa requête, il peut accéder aux autres points de vue sur la ou les thématique(s) liées à sa requête, pour cela il les sélectionne et indique au système les informations qu'il veut avoir. La relation entre la requête et les autres fonctionnalités de visualisation est à sens unique.

L'entrée par la carte lui permet tout d'abord d'avoir une vue globale des thématiques et de l'évolution de ces dernières. Cette vue lui permet aussi de voir la proximité des classes les unes par rapport aux autres. Il peut sélectionner une ou plusieurs d'entre elles afin d'accéder aux informations complémentaires concernant l'histoire de la ou des sélection(s) (qualitatif) ou leur développement sur un intervalle de temps donné (quantitatif).

Lorsque l'utilisateur se situe sur l'un des 3 pôles, Globale, Qualitatif, Quantitatif comparé, il lui est toujours possible d'accéder aux 2 autres restants. Il est intéressant pour l'utilisateur d'avoir toujours la possibilité d'étudier une, plusieurs ou l'ensemble des thématiques sous différents angles.

A ces 3 dernières fonctionnalités de visualisation, nous fournissons une fonction de veille. Les alertes sont effectuées à l'aide de symboles ou grâce à des seuils déterminés par l'utilisateur lui-même en fonction de ce qu'il veut surveiller.

[60] Deux graphes G et G' sont dits isomorphes si on retrouve dans G' les mêmes arcs et les mêmes sommets.

[61] La distance d'édition entre les graphes G et G' désigne les transformations qu'a subit G pour former G', en termes d'insertion, de suppression et de réétiquetage de sommets et d'arcs.

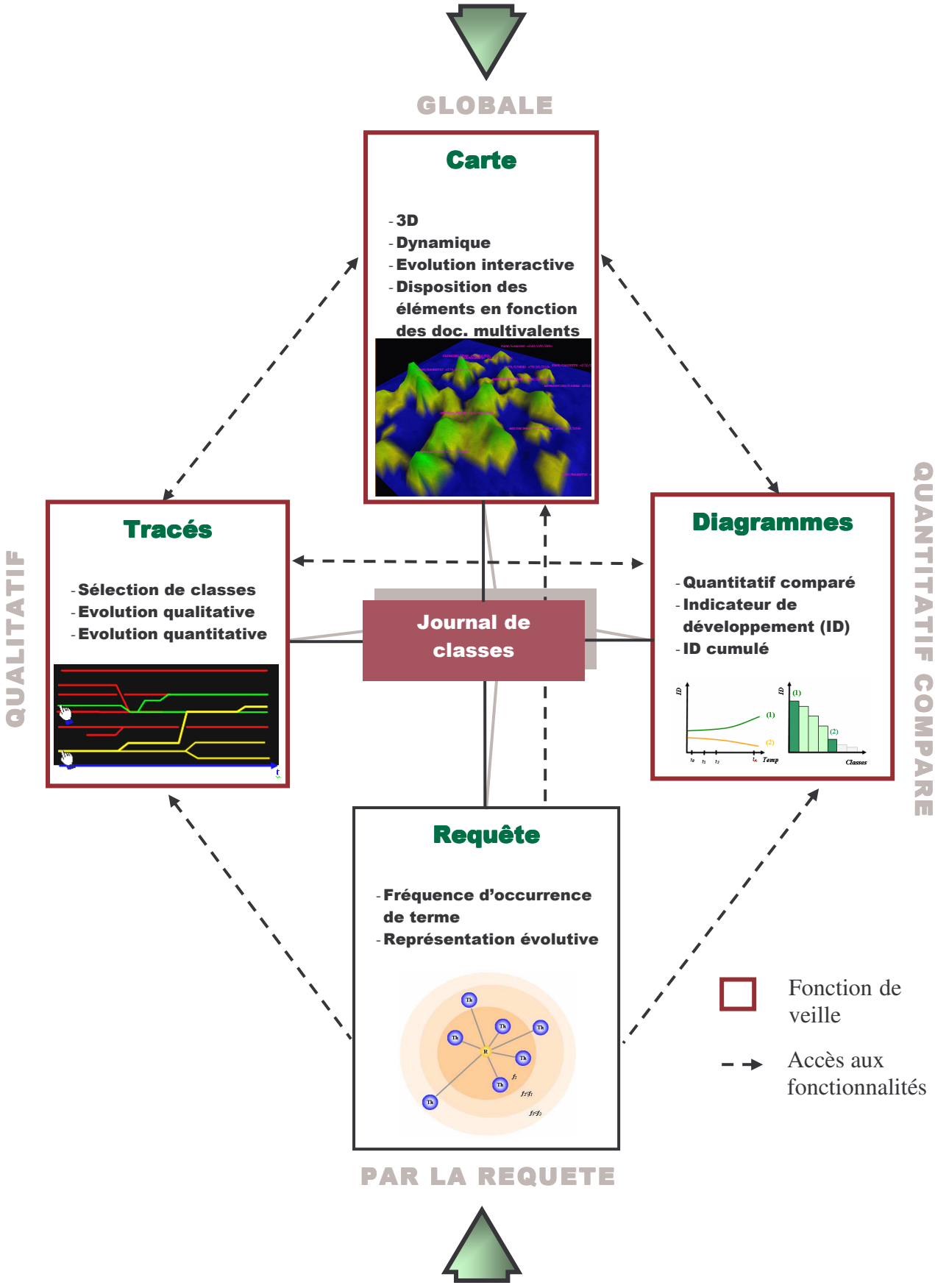


Figure 103 : Structure du système de fonctionnalités de représentation dynamique des connaissances scientifiques

Conclusion

L'objectif de cette étude était de modéliser un dispositif permettant la visualisation de l'évolution de l'information scientifique et technique. Dans ce but nous avons bâti une réflexion autour des questions suivantes : « Quelles sont les informations utiles à l'utilisateur pour pouvoir analyser cette évolution ? Comment les représenter graphiquement à cet utilisateur de manière efficace et efficiente ? »

Dans un premier chapitre nous avons introduit notre projet dans une démarche de veille scientifique et technologique. Nous avons considéré que le suivi de l'évolution de l'IST était source d'informations utiles pour un veilleur. Elle peut en effet alimenter des stratégies d'innovation et la conduite des politiques de recherches des organisations socio-économiques.

Dans un second chapitre, nous nous sommes intéressés aux méthodes d'analyse de l'information auxquelles appartient l'algorithme incrémental de classification. Cette exploration nous a permis de comprendre leur intérêt et leur fonctionnement.

Un troisième chapitre, le dernier de notre état de l'art, a été consacré au domaine de la visualisation de l'information. Nous y avons dressé une typologie des techniques existantes au travers 3 approches : une approche généraliste basée sur la nature des données visualisées, une approche sur la visualisation de l'évolution de l'IST et une approche centrée sur la prise en compte de l'utilisateur pointant notamment sur les techniques d'interaction.

Un quatrième chapitre vient compléter cet état de l'art par l'observation d'outils de veille existants sur le marché afin d'étudier leurs fonctionnalités de visualisation de l'information. Nous avons pu constater que les fonctionnalités de visualisation sont de plus en plus utilisés dans ces outils, l'aspect lié au suivi de l'évolution l'est cependant beaucoup moins.

Enfin, à la lumière des points vus dans les chapitres précédents, nous avons consacré le cinquième et dernier chapitre, à définir les fonctionnalités de visualisation appliquées aux résultats de l'algorithme incrémental de classification.
Cette dernière démarche s'est déroulée en trois temps :

Modélisation des usagers potentiels : Nous avons essayé de définir des stéréotypes d'utilisateurs, déduire les tâches incombant à ces utilisateurs et enfin identifier les besoins informationnels inhérents à ces tâches.

Correspondance entre besoins informationnels et données du journal : La plupart des besoins trouvait leur réponse dans les données du journal. Pour ceux qui n'en trouvaient pas, nous avons proposé des données complémentaires.

Définition des fonctionnalités de visualisation : Nous avons proposé des fonctionnalités de visualisation adaptées aux données et aux besoins. Nous avons discuté ces propositions au regard de leur efficacité et leur efficience, c'est-à-dire en quoi elles permettent de fournir une information à valeur ajoutée à l'utilisateur et quel rapport « effort d'interprétation » / « apport informationnel » peut on leur attribuer ?

Cette discussion nous a ensuite permis de fournir le prototype d'un dispositif de visualisation dynamique des connaissances scientifiques. Celui-ci vient confirmer notre

hypothèse selon laquelle il n'existe pas de visualisation qui soit capable de répondre à tous les besoins informationnels. En effet, pour contenter les principaux besoins de nos utilisateurs, il nous a fallu envisager 5 approches différentes des données et 4 modes de représentation distincts mais complémentaires.

Dans le cadre de notre recherche, ce projet s'est arrêté au stade des propositions. Nous estimons que certaines de ces propositions demanderaient à être approfondies. Le modèle utilisateurs, par exemple, a été constitué à partir d'un stéréotype d'utilisateur, il serait nécessaire d'envisager une enquête plus élaborée pour le valider ou l'invalider. Il serait intéressant de pouvoir expérimenter cette approche plus tard encore, en développant une première interface et en la faisant tester par les utilisateurs désignés dans notre étude. Ainsi nous saurions si notre démarche a atteint son objectif.

D'autre part, notre dispositif est-il réalisable ? Une étude de faisabilité pourrait être menée sur le sujet afin de déterminer s'il est techniquement envisageable. La question se porterait principalement sur les aspects techniques.

Enfin les fonctionnalités de visualisation que nous avons proposées traitent uniquement des aspects liés au contenu des documents, objet de l'analyse mené par l'algorithme incrémental de classification. Quant est-il des aspects contextuels des documents ? Nous avons envisagé, dans notre modèle utilisateur, des besoins liés aux auteurs, aux affiliations de ces auteurs, etc. Or, pour le moment, il n'en est pas question dans nos propositions de visualisations. Pourtant le contexte de ces documents peut constituer une source d'informations très importante dans une démarche de veille : les collaborations entre chercheurs, le positionnement des équipes de recherche sur une thématique, etc. Ceci nous amène à penser qu'une approche dynamique sur le contenu des documents pourrait alors s'ouvrir vers d'autres perspectives tout aussi riches en s'appliquant à la dynamique des organisations pour visualiser leurs rapprochements et leurs dissensions sur des projets, par exemple. Dans une démarche de surveillance d'un environnement concurrentiel, technologique, etc. de plus en plus complexe et changeant, la possibilité de visualiser le contexte d'une organisation et l'évolution de ce dernier, permettrait d'apporter au décideur de nouvelles clés de lecture pour alimenter sa réflexion stratégique.

Bibliographie

ADAMIC L.A., ADAR E., *Friends and Neighbors on the web*, In : Social Networks, 25(3) pp. 211-230, 2003.

ANDREWS K., *Information visualization, Tutorial notes*. IICM, Graz University of Technology, 2002.

BELLOT P., *Classification de documents et enrichissement de requêtes*, In : Méthodes avancées pour les systèmes de recherche d'informations. Dirigé par M. Ihadjadene. Paris : Hermes, 2004. Tome 2.

BERTIN J. *La sémiologie graphique.* Paris : Gauthier-villars, 1967.

BONNEL et CHEVALIER, *Evaluation des Interfaces Utilisateur d'Information*, In : actes du colloque EGC 2006

BOUAKA, *Développement d'un modèle pour l'explicitation d'un modèle pour l'explicitation d'un problème décisionnel : un outil d'aide à la décision dans un contexte d'IE*. Th. doct., Université Nancy 2, 2004.

BOUROCHE J.-M., SAPORTA G., *L'analyse des données*, Paris : PUF, 1989. Que sais-je, n°1854.

BULINGE F., *Pour une culture de l'information dans les petites et moyennes organisations : un modèle incrémental d'intelligence économique*. Th. Doct. Université du Sud Toulon Var, 2002.

CAHLIK T., *Comparison of the maps of science*, Scientometrics, 2000, 19, n°3.

CALLON M., COURTIAL JP, PENAN H, *La scientométrie*, Paris : PUF, 1993. Que sais-je, n°2727

CARD S.K., MACKINLAY J.D., SHNEIDERMAN B., *Readings in information visualization using vision to think*. San Francisco : Morgan Kaufmann publisher Inc, 1999.

LE COADIC Y.-F., *La science de l'information*, Paris : PUF, 2004. Que sais je, n°2873

COURTIAL J.-P., *Introduction à la scientométrie : de la bibliométrie à la veille technologique*. Paris : Anthropos, 1990.

Conseil Régional de Lorraine, *Intelligence économique, un guide pour débutants et praticiens.* Communautés européennes, 2003.

DÂASSI C., *Visualisation de données temporelles*, In : actes de la conférence AFIHM sur l'Interaction Homme-Machine, 22-26 Novembre 1999. Montpellier, 1999. Tome 2. pp. 129-131.

DAVID A., THIERY O., *Prise en compte du profil de l'utilisateur dans un système d'information stratégique*. In : VSST : Veille stratégique, Scientifique et Technologique, Barcelone, 2001.

DAVID A., *La recherche collaborative d'information dans un contexte d'Intelligence Economique*. In : Le système d'information de l'entreprise, 25-26 février 2006, Algérie – Télécom, Alger, Algérie.

DOBROWOLSKI Z., *Etude sur la construction des systèmes de classification*. Préf. d'Eric de Grolier, Paris : Gauthier-Villars, 1964.

DOU H., DESVAIS H., *L'intelligence économique au service du développement industriel*, Paris : Dunod. 1995

ERTEN C., HARDING P.J., KOBOUROV S.G., WAMPLER K., YEE G., *Exploring the Computing Literature Using Temporal Graph Visualization*. Report, Department of Computer Science, University of Arizona. 2003.

FAVIER L., *Recherche et application d'une méthodologie d'analyse de l'information pour l'intelligence économique. Application à un centre technique du secteur de la plasturgie.* Th. Doct. : Université Lyon 2, 1998.

FEKETE J.D., *Nouvelle génération d'Interfaces Homme-Machine pour mieux agir et mieux comprendre.* Habilitation à diriger les Recherches, Research Report 1411, Université Paris-Sud, 2005. (http://www.lri.fr/~fekete/ps/)

FEKETE J.D., PLAISANT C., *Les leçons tirées des deux compétitions de visualisation de l'information*, In : Actes de la conférence Interface-Homme-Machine, IHM'04, Namur. 2004.

FREEMAN E. T., *The Lifestreams Software Architecture*, Ph.D. Dissertation, Yale University Department of Computer Science, May 1997.

GANASCIA J.-G., FENOGLIO I., LEBRAVE J.-L., *EDITE MEDITE : un logiciel de comparaison de versions*. In : Actes du colloque JADT 2004 : 7e journées internationales d'Analyse statistique de Données Textuelles, 2004.

GERSHON N., PAGE W., *What Storytelling can do for information visualization*, In : Communication of ACM, 2001, Vol.44, n°8.

GORIA S., *L'expression du problème dans la recherche d'informations : application à un contexte d'intermédiation territoriale*, Th. Doct. : Université Nancy 2, 2006.

HASCOËT M., *Visualisation d'information et interaction*, In : Méthodes avancées pour les systèmes de recherche d'informations. M. Ihadjadene. Paris : Hermes, 2004. T2.

HAVRE S., HETZLER E., WHITNEY P., NOWELL L., *ThemeRiver : Visualizing Thematic Changes in Large Document Collections*, IEEE transactions on visualization and computer graphics, 2002, Vol. 8, n°1.

HEARST M., *User interface and visualization.* In : Modern Information Retrieval R. Baeta-Yates, B. Ribeiro-Neto (eds), Addison-Wesley, 1999. p. 257-322.

HETZLER B., HARRIS M., HAVRE S., WHITENEY P., *Visualizing the Full Spectrum of Document Relationships*, In : Proceedings of the Fifth International Society for Knowledge Organization (ISKO) Conference, 1998.

HUMBERT P., *Etude de faisabilité pour la mise en place d'une structure de gestion des images au sein de l'entreprise Noremat*. Mémoire de maîtrise en Sciences de l'Information et de la Documentation. Université Nancy 2, 2005.

JAKOBIAK F., *Pratique de la veille technologique*. Paris : les Editions d'organisation, 1991.

JAKOBIAK F., *L'information scientifique et technique*, Paris : PUF, 1995, Que sais-je, n°3015

JAKOBIAK F., *L'intelligence économique*. Paris : les Editions d'organisation, 2004.

KAYSER D., *La représentation des connaissances*. Paris : Hermes. Coll. informatique, 1997.

KEIM D.A., *Information Visualization an Visual Data Minig*, IEEE transactions on visualization and computer graphics, 2002. Vol.7, No.1.

LELU A., et al., Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN, In : Actes du colloque JADT 2006.

LE MOIGNE J.-L., *La modélisation des systèmes complexes*, Paris : Dunod, 1995.

MARTINET B., RIBAULT J.-M., *La veille technologique, concurrentielle et commerciale*, Paris : les Editions d'organisation, 1989.

MOYA-ANEGON F., VARGAS-QUESADA B., CHINCHILLA-RODRIGUEZ Z., CORERA-ALVAREZ E., HERRERO-SOLANA V., MUNOZ-FERNANDEZ F.J., *A new technique for building maps of large scientific domains based on the cocitation of classes and categories*. Scientometrics, 2004. 61(1), 129-145.

NOYONS E.C.M., *Science maps within a science policy context*, In : H.F. MOED et al., Handbook of quantitative science and technology research. Kluwer Academics Publishers, 2004. 237-255.

NUSSBAUMER A., *Hierarchy Browsers, Integrating Four Graph-Based Hierarchy Browsers into the Hierarchical Visualisation System (HVS)*. Th. Doct. Université de Graz (Autriche), 2005.

POLANCO X., Les *sciences de l'information, Bibliométrie, scientométrie, infométrie*. In : SOLARIS, Jean-Max Noyer, Presses Universitaires de Rennes, 1995.

POLANCO X., ZARTL A., *Information Visualization, State of the Art Report*, In : Projet EICSTES, 1999.

POLANCO X., *La notion de visualisation de l'information et le modèle de référence*, In : Actes du colloque « Cartographie de l'information », Paris, ESIEE. 2002.

ROUSSEAU F., THIL J., *Veille et informatique : des besoins aux solutions.* Technologies internationales, nov. 97, n°39.

SIMIER P., *L'intelligence économique et l'utilisateur – acteur au centre du processus de management*, In : Actes du colloque de l'AIM, Grenoble, 2003.

SHNEIDERMAN B., PLAISANT C., *Designing the user interface, Strategies for effective human-computer interaction*, 4e édition. Addison Wesley, 2005.

SMALL H., *Macro-level changes in the structure of co-citation clusters: 1983-1989*. Scientometrics, 1993, n°26, pp. 5-20.

SMALL H., *Visualizing Science by Citation Mapping*, Journal of the American Society for Information Science, 1999. n°50(9), pp. 799-813,.

SORLIN S., et SOLNON C., Similarité de graphes : une mesure générique et un algorithme tabou réactif. LIRIS/CNRS, 2003.

TOUSSAINT Y., *Extraction de connaissances à partir de textes structurés.* Document numérique, 2004, vol. 8, no 3, p. 11–34.

TRICOT C., ROCHE C., *Exploration interactive de bases de connaissances : un retour d'expérience*. In : Actes du colloque RNTI-EGC'06, Cépaduès-éditions, 2006.Vol.1.

TUFTE R. E., *The visual display of quantitative information*, Cheshire : Graphic press, 2001.

Van RAAN A.F.J., *Handbook of Quantitative Studies of Science and Technology*. Amsterdam : North Holland, Elsevir Science Publishers, 1988.

VIDAL, S., *L'approche diachronique dans la détection de signaux faibles : Une illustration dans le bioterrorisme*. Mémoire de DEA, Universités Nancy 2 & Metz, 2004.

Robert & Collins Senior, *Dictionnaire Anglais-Français, Français-Anglais*, 5e édition, HarperCollins Publishers, 1998.

Ouvrages cités non consultés

WEAVER C.E., SHANNON C, Théorie mathématique de la communication, Paris : Retz, 1975.

BENZECRI J.-P., BENZECRI C., *La pratique de l'analyse des données, T1 : Analyse des correspondances, exposé élémentaire*, Dunod, 1980

DIDAY E., *La méthode des nuées dynamiques*, Revue de Stat Appliquée, vol. 19, n°2, pp.19-34, 1971.

FAYYAD U., PIATETSKY-SHAPIROG., SMYTH P., *From Datamining to Knowledge Discovery*, chapitre 1, 1996.

FORGY E.W., *Cluster analysis of multivariate data : efficiency versus interpretability of classifications*. In: Biometric Society Meetings. Riverside, Californie, 1965.

LAMIREL J-Ch, *Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif*. Th. Doct. Université Nancy 1 Henri Poincaré, 1995.

LOTKA, A. J., *The Frequency distribution of Scientific Productivity*, In : Journal of the Washington Academy of Sciences, 1926.

MAC QUEEN, J., *Some Methods for Classification and Analysis of Multivariate Observations*, In : Berkeley Symposium, 1967.

PRICE J. S., *A general theory of bibliometric and other cumulative advantage processes*. In : Journal of the American Society for Information Science, 1976, Vol.39, N°4. p. 292-306.

ZIPF G. K., *Human Behaviour and the Principle of Least-Effort*. Cambridge : Addison-Wesley, 1949.

HAGGET P., *Locational analysis in human geography*, Londres : Arnold, 1965.

Webographie

[Hearst, 98] SIMS 247 Lecture 11 , Evaluating Interactive Interfaces, cours de M. Hearst, Université de Berkeley (Californie), Février 1998. http://www.sims.berkeley.edu:8000/courses/is247/s98/lectures/lecture-02-24.ppt

[Kapusova, 04] KAPUSOVA D., *Visualisation de l'information dans le portail STAF18*, 2004 http://tecfaseed.unige.ch/staf18/modules/ePBLjolan/uploads/proj15/paper%20(et%20dispositif)6.xml (consultée le 17/02/2006).

[Munzer, 03] T. MUNZNER, *Penser par la vision*, Horizon0 : art et culture numériques au Canada, vol.6, 2003. [http://www.horizonzero.ca/textsite/see.php?is=6&file=7&tlang=1 (page consultée le3/02/2006)]

[OLI] OLIVE : On-line Library of Information Visualization Environments [http://otal.umd.edu/Olive/] .

[IVR] Information Visualization Resources on the Web, http://graphics.stanford.edu/courses/cs348c-96-fall/gamma-corrected/resources.html

[INA] Projet VIE, interface dynamique, http://www.ina.fr/recherche/projets/encours/vie/interf/index.fr.html

[VC] visualcomplexity.com. A visual exploration on mapping complex networks http://www.visualcomplexity.com/vc/

TABLE DES ILLUSTRATIONS