Vers des Systèmes de Découverte et de Filtrage d'Information Documentaire : Quelle Stratégie Faut-il Mettre en Place?

Laurence Favier, Madjid Ihadjadene

CAIS 2000:
DIMENSIONS OF A GLOBAL INFORMATION SCIENCE
Canadian Association for Information Science
Proceedings of the 28th Annual Conference
Table of Contents

ACSI 2000:
LES DIMENSIONS D'UNE SCIENCE DE L'INFORMATION GLOBALE
Association canadienne des sciences de l'information
Travaux du 28e congrès annuel
Table des matières

Vers des Systèmes de Découverte et de Filtrage d'Information Documentaire : Quelle Stratégie Faut-il Mettre en Place?

Laurence Favier

Université de Bourgogne

Madjid Ihadjadene

Université de Paris X

Résumé

Le problème de l'exploitation de grands gisements d'information, celle de bases de données du type "datawarehouse" dont la constitution se généralise, de catalogues informatisés (OPACs) de bibliothèques, de bases de données spécialisées, d'Internet (en particulier du Web) est l'impossibilité pour l'usager de visualiser l'ensemble des réponses que les systèmes de recherche mettent à leur disposition. Par exemple, de recentes études empiriques effectuées sur les www-Opacs ou sur les moteurs de recherche (par exemple Spink 1999, Ihadjadene1999) montrent que les requêtes des usagers sont pauvrement formulées (moins de deux termes et pas d'opérateurs booléens), ne visualisent pas plus de deux pages web. La richesse des réponses obtenue est ignorée. La tendance de l'informatique documentaire aujourd'hui est de répondre à ce problème en se centrant sur le rôle de l'usager pour filtrer, adapter, personnaliser sa recherche. Deux directions se font jour à ce sujet : soit l'on offre à l'usager un ensemble d'outils pour qu'il construise lui-même son parcours de recherche, soit on pré-calibre sa recherche en fonction d'une connaissance de l'usager obtenue directement à partir de questions qui lui sont posées (diffusion sélective d'information traditionnelle par l'intermédiaire de déclarations de profils, agents "push" sur Internet) ou indirectement par les essais et erreurs de son parcours (agents intelligents par exemple). Dans la première direction

l'usager reste maître de sa recherche : l'ambition du système est de lui proposer des filtres afin qu'il puisse construire un parcours au fur et à mesure de l'évolution de son besoin d'information. Dans la seconde, il reçoit des réponses sans qu'il ait la maîtrise de l'évolutivité de son besoin d'information. Après un rappel des modèles de recherche d'information centrés sur l'usager, nous proposons de mettre en évidence les caractéristiques d'une approche basée sur la première approche (celle de l'usager actif) à partir de deux méthodes facilement adaptables au monde documentaire, l'une destinée aux fonds spécialisés, l'autre aux fonds encyclopédiques caractéristiques des catalogues de bibliothèques. La première repose sur l'exploitation de classifications statistiques (plus précisément d'une méthode d'analyse factorielle appliquée aux mots de textes courts, résumés ou articles de presse, afin de créer des regroupements thématiques de documents et de mots), l'autre sur celle de classifications encyclopédiques (Dewey) et listes d'autorité-matière utilisées comme moyens de filtrage des termes ou des notices bibliographiques. Dans l'un et l'autre cas, on montre comment un système documentaire peut guider un usager, tantôt expert, tantôt "grand-public", dans la découverte des réponses, sans se contenter d'afficher des listes de référence, fussent-elles classées par ordre de pertinence.

I. La sous-utilisation des SRI et la surcharge d'info

Le problème de l'exploitation de grands gisements d'information, celle de bases de données du type "datawarehouse" dont la constitution se généralise, de catalogues informatisés (OPACs) de bibliothèques, de bases de données spécialisées, d'Internet (en particulier du Web) est l'impossibilité pour l'usager de visualiser l'ensemble des réponses que les systèmes de recherche mettent à leur disposition.

Les récents travaux de (Jones,1999), de (Spink,1999) et de (Bruza,1997) sur l'usage des moteurs de recherches, des bibliothèques numériques et des www-Opacs (Ihadjadene 1999[1]) ont montré que les ressources du système sont sous-utilisées et que les outils mis à la disposition de l'usager final pour explorer le nombre élevé de réponses sont insuffisants et inadaptées. Les requêtes des usagers sont pauvrement formulées (moins de deux termes et absence d'opérateurs booléens) et ceux-ci ne visualisent pas plus de deux pages web. La richesse des réponses obtenue est ignorée. Dans le cas du moteur de recherche ALTAVISTA (Silverstein, 1998), 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. Les tactiques élaborées pour réduire ce problème de surinformation sont rudimentaires. Seulement un usager sur deux tente de réduire le nombre de réponses en ajoutant souvent un terme à l'équation d'origine. Les usagers préfèrent changer le contenu de la requête plutôt que de modifier sa structuration logique, ce qui pose le problème de la pertinence des outils logico-analytiques mis à leur disposition et suggère d'autres modalités d'exploration.

II. Les approches du problème

La tendance de l'informatique documentaire aujourd'hui est de répondre à ce problème en se centrant sur le rôle de l'usager pour filtrer, adapter, personnaliser sa recherche. Placer l'usager final au centre des études est devenu l'une des évolutions les plus marquantes ces quinze dernières années en informatique documentaire.

Essentiellement, trois directions se font jour à ce sujet :

1. soit l'on offre à l'usager un ensemble d'outils pour qu'il construise lui-même son parcours de recherche. L'usager reste maître de sa recherche : l'ambition du système est de lui proposer des filtres afin qu'il puisse construire un parcours au fur et à mesure de l'évolution de son besoin d'information. L'information est organisée à la sortie et c'est l'usager qui a la tâche de trier cette masse d'information : c'est l'approche "filtrage d'information".
2. soit l'information est organisée au moment où on la saisit dans le système de recherche : c'est l'approche "métadonnées" et catalogage qui permettent de simplifier la récupération ultérieure de l'information.
3. soit on pré-calibre la recherche de l'usager en fonction d'une connaissance de celui-ci obtenue directement à partir de questions qui lui sont posées (diffusion sélective d'information traditionnelle par l'intermédiaire de déclarations de profils, agents "push" sur Internet, cookies). Il reçoit des réponses sans qu'il ait la maîtrise de l'évolutivité de son besoin d'information.

Nous ne pensons pas qu'il faille opposer ces approches. Concernant les approches 1 et 2, on peut dire que la structuration de l'information (catalogage ou métadonnée) n'est pas suffisante, et qu'il est donc nécéssaire d'inclure des possibilités de filtrage d'information. De même que la qualité des techniques linguistiques ou automatiques adoptées pour le filtrage ne suffisent pas et gagneraient à être enrichies par une meilleure structuration de l'information, d'où l'engouement actuel pour XML. Par ailleurs l'approche *push* répond à un certain type de besoin d'information : celui de la mise à jour de l'information plus que de la recherche *strito sensu*.

Donner un rôle à l'usager dans la recherche d'information, rendre le système interactif, c'est lui permettre d'agir sur la sortie des résultats (approche n°1) grâce à des interfaces utilisateur qui ordonnent les informations trouvées. Nous présentons, à ce sujet, deux expériences issues de deux types de SRI : l'une propose un filtrage par les classifications encyclopédiques (Dewey) et listes d'autorité-matière (Rameau) pour améliorer la recherche dans les catalogues de bibliothèques, l'autre un filtrage par les classifications statistiques (plus précisément des méthodes d'analyse de données textuelles) pour améliorer la recherche dans les bases de données textuelles spécialisées.

Dans l'un et l'autre cas, on montre comment un système documentaire peut guider un usager, tantôt " grand public " tantôt " spécialiste ", dans la découverte des réponses, sans se contenter d'afficher des listes de références, fussent-elles classées par ordre de pertinence.

III. Expérimentations de techniques de filtrage

III.1 Le cas des fonds encyclopédiques : proposition de filtrage par les langages documentaires. Le système Cathie.

Une des solutions au problème de surcharge d'information, consiste à construire des interfaces utilisateur qui regroupent automatiquement les résultats en catégories. Cette méthode (clustérisation) est apparue avec l'introduction du modèle vectoriel. Elle a été constamment améliorée. Plusieurs travaux récents ont permis de donner un support visuel à cette catégorisation des documents. Le professeur Khorflag (1998) a effectué, à l'Université de Pittsburg, un ensemble de recherches pour mettre au point des interfaces facilitant la visualisation de l'information, et par conséquent, le repérage Il a conçu trois prototypes : VIBE (*Visual Information Browsing Environment*), GUIDO (*Graphical User Interface for*

Document Organization) et BIRD (*Browsing Interface for Retrieving Documents*). On peut aussi citer les prototypes suivants :

- TileBars (Hearst, 1995)
- Scatter/Gather (Cutting, 1992)
- InfoCrystal (Spoerri, 1993)

Nous pensons que les interfaces de visualisation sont encore difficile à utiliser pour le grand public. L'interface de recherche d'information restera encore, une interface textuelle. Pollit (Pollit 2000) critique les possibilités de représentation des connaissances par des interfaces 3D. C'est la raison pour laquelle nous pensons que les classifications et/ou les ontologies peuvent jouer un rôle dans le filtrage d'information. La classification hiérarchique de Dewey exemplifie deux fonctions de la classification traditionnelle : la collacation (inclusion) et la partition (exclusion). L'inclusion rapproche les objets et les idées semblables. Mais dans un domaine d'information très vaste, il est tout aussi important d'exclure l'information non désirée qu'inclure ce qui est recherché. La partition peut être opérée en divisant une grande quantité d'information en parties plus petites comme moyen d'isoler la partie qui a la plus grande probabilité d'être pertinente. (Chan, 1995).

Nous avons tenté d'apporter notre contribution à ces problèmes de surcharge en utilisant un certain nombre d'approches permettant d'assister l'usager dans le choix des termes et dans l'élaboration de stratégies de recherches. Le prototype CATHIE (CATalogue Hypertexte Interactif et Enrichi) que nous avons développé tend à remédier en partie à divers déficits que nos analyses d'usage ont mis en évidence. Il associe la richesse des vocabulaires contrôlés, les possibilités de visualisation et de navigation de l'hypertexte et la puissance du modèle probabiliste. Après une requête, l'usager peut exploiter ces quatre stratégies :

- Effectuer une recherche dans le lot de documents trouvés
- Reformuler la question à travers les vedettes matières proposées
- Établir un filtrage thématique des termes et/ou documents
- Afficher la notice et voir les documents similaires.

Nous avons décidé que les vedettes matières issues de la liste RAMEAU doivent être regroupées selon le champs sémantique. Nous effectuons donc deux types de classification. La première concerne un calcul de fréquence des vedettes matières construites (VMC) sur un lot de documents. La seconde consiste à structurer ces VMC selon les domaines. De ce fait, nous faisons intervenir plus de sémantique en établissant une classification des vedettes par domaine. Le catalogue en ligne, perd ainsi une part importante de son opacité.

Après une recherche sur la question "access" et si l'usager décide de spécifier les termes et les réponses relatifs au domaine informatique, CATHIE affiche ces nouveaux dossiers (figure 1).

Figure 1: Filtrage d'information dans CATHIE (exemple 1)

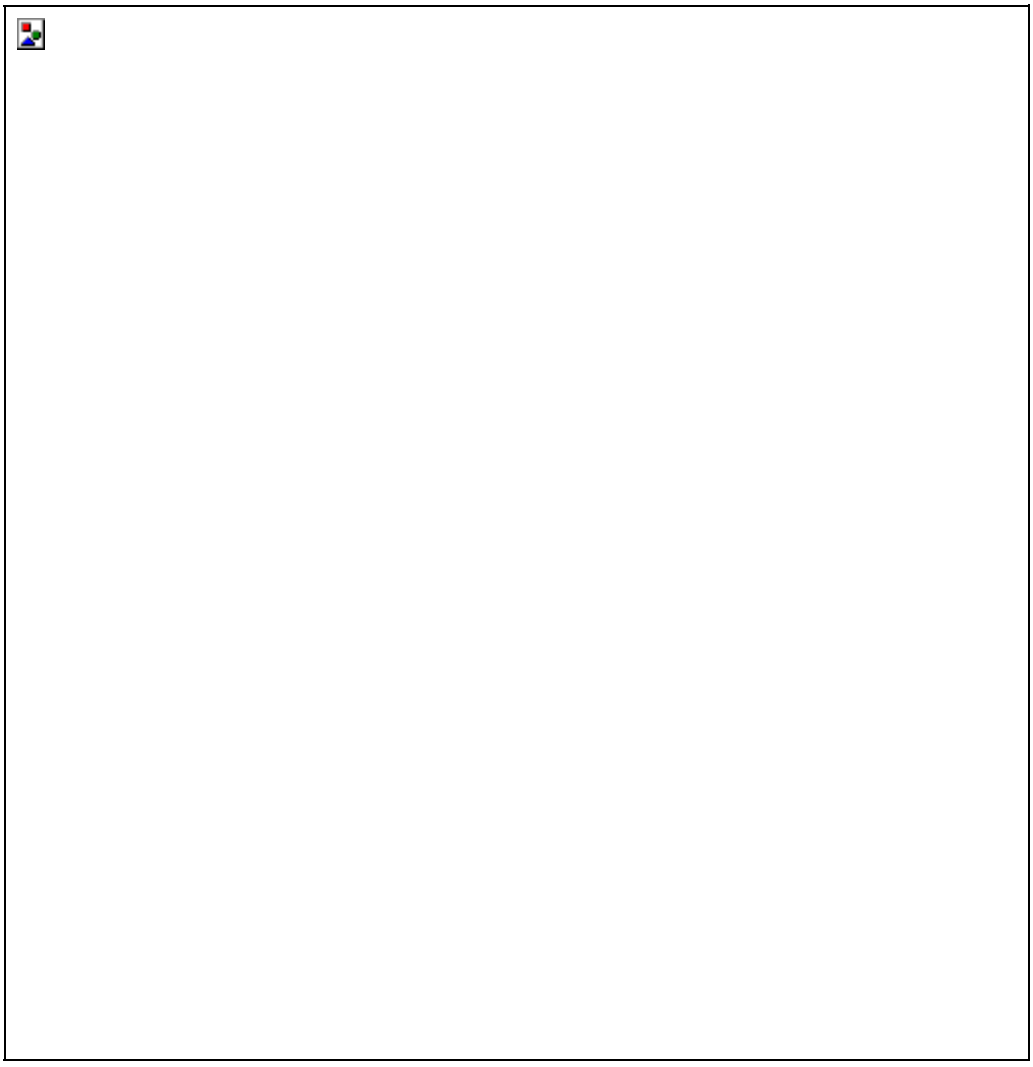
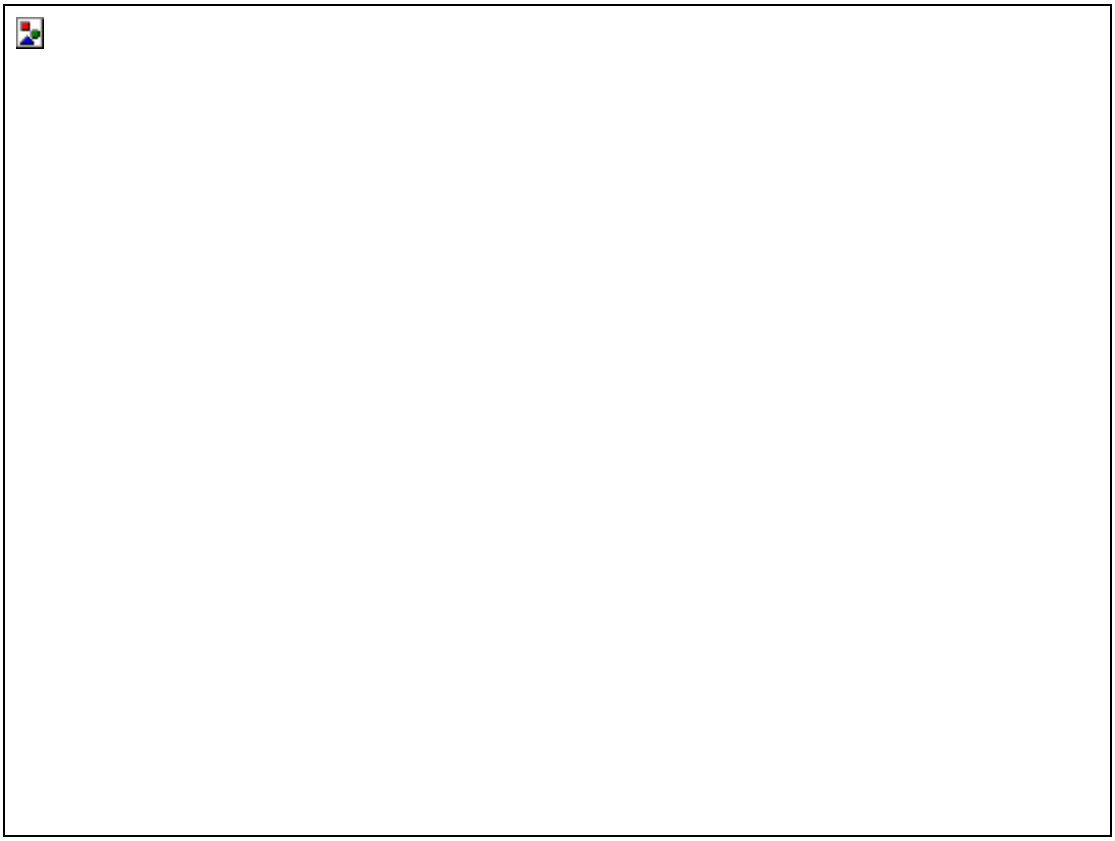


Figure 2: Filtrage d'information dans CATHIE (exemple 2)



- Le prototype CATHIE est pour l'instant encore expérimental et de nombreuses améliorations d'ordre techniques pourraient lui être apportées.
- L'étude de ces nouveaux modèles d'interaction (visualisation de l'information, catégorisation thématique, reformulation interactive, etc.) est encore récent. Il convient maintenant d'examiner leur usages en situation réelle d'utilisation.

III.2 Le cas des fonds spécialisés : filtrage statistique et découverte de connaissances.

III.2.1 Intérêt du filtrage statistique

La surcharge d'information dans le cas des bases documentaires spécialisées ne se pose pas dans les mêmes termes que dans celui des fonds encyclopédiques car le besoin d'information n'est pas le même. Il ne s'agit pas seulement de sélectionner des documents mais d'avoir la connaissance de la structuration thématique d'un sujet : si mon sujet est la micro-injection des polymères et que le système me donne 500 réponses à cette requête, mon objectif *ne sera pas de réduire le nombre de réponses* mais d'en avoir une *synthèse*. Opposant les listes de vedettes-matière (destinées à l'indexation des fonds encyclopédiques) aux thesaurus (conçus pour décrire un domaine spécialisé de la connaissance), M.Hudon (Hudon 1994) montre que l'un a pour vocation la recherche *de documents* dans les collections documentaires quand l'autre a celle de la recherche *d'information* dans les documents (p.37). Le document n'est plus le centre de la recherche.

Nous sommes plus proches de la problématique de la "découverte de connaissances" (knowledge discovery) que de celle de la sélection de documents.

La question qui se pose est de savoir quelle est la méthode la plus appropriée pour donner à l'usager une synthèse "interactive" des réponses à son sujet.

Seules les méthodes statistiques peuvent constituer des synthèses. Le filtrage par les langages documentaires ne me permettra pas d'avoir une connaissance synthétique des 500 réponses à mon besoin d'information. Ils peuvent être utilisés pour filtrer l'information sachant qu'ils classent le connu grâce aux relations "terme associé" "terme générique", "terme spécifique". La synthèse statistique permet d'aller au-delà de la formulation du besoin d'information, vers "l'inconnu".

Pour évaluer l'apport d'information de la synthèse statistique pour l'usager et la manière de l'intégrer dans un système interactif destiné à la recherche d'information spécialisée pour des experts, nous avons élaboré[2] un prototype, qui gagnerait toutefois à être amélioré comme nous l'expliquerons par la suite.

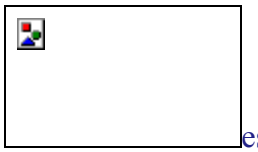
III.2.2 La méthode

Ce système repose sur l'utilisation d'une méthode statistique multidimensionnelle appliquée aux mots des textes des documents. La méthode utilisée est l'analyse factorielle des correspondances (AFC) (Benzécri 1980), une méthode proche de Latent Semantic Analysis (Deerwester S. *et alii* 1990) . L'analyse des correspondances est une technique de description de tableaux croisés (tables de contingence) ou de tableaux binaires de type "présence-absence". Deux ensembles (individus et variables, ou observations et variables, ou mots et textes) sont mis en correspondance sous la forme de tableaux rectangulaires de données numériques avec, à l'intersection de la ligne et la colonne, le nombre de fois que l'élément est présent, ou bien l'indication par 0 et 1 de sa présence ou de son absence. Cette technique de description convient particulièrement au cas des données textuelles (Benzécri, ibid).

On note k_{ij} le nombre de fois où le mot i est présent le document j. Le tableau de données est une matrice X de terme général (k_{ij}) $i = 1,n$; $j = 1, p$ ($p<n$).

Par cette technique chacune des dimensions du tableau de données numériques (table de contingence) permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension. A partir de ce tableau de distances on obtient une représentation géométrique décrivant les similitudes entre les lignes et les colonnes.

La technique utilisée est la décomposition en valeurs singulières.


est la matrice des valeurs propres de X'X. Les valeurs propres non nulles de XX' et de X'X sont égales. X' est la transposée de X.

U est la matrice p x p d'ordre ayant en colonne les vecteurs propres de X'X. Ces vecteurs propres sont orthogonaux et de norme 1.

V est la matrice n x n d'ordre q ayant en colonne les vecteurs propres de XX'. Ces vecteurs propres sont orthogonaux et de norme 1.

Les valeurs propres sont rangées par ordre décroissant 

Si les plus petites valeurs propres sont très faibles et jugées négligeables, on peut limiter la sommation aux premiers termes correspondant aux plus grandes valeurs propres.

Dans IR^p, □ est appelé le □axe factoriel et le vecteur des coordonnées des n points sur cet axe s'écrit □.

Dans IR^n □ est le □axe factoriel et on construit les coordonnées des p points □.

L'objet de l'AFC est de représenter géométriquement les deux ensembles en correspondance (les lignes et les colonnes) de telle sorte que soient respectées les proximités distributionnelles.

A partir du traitement par Analyse Factorielle des Correspondances d'une table de contingence (ici un tableau mots*documents), plusieurs séries de paramètres sont extraites :

- les valeurs propres et les pourcentages d'inertie représentés par chaque axe (les facteurs). Ces valeurs mesurent la "part d'information" représentée par chaque axe.
- les coordonnées de chaque point sur les axes.
- les contributions qui décrivent la part prise par un élément (ligne ou colonne) dans la construction d'un axe factoriel.
- les cosinus carrés qui mesurent la qualité de représentation de chaque élément sur les axes.

Nous avons considéré que l'interprétation des axes devait s'opérer sur la valeur de contribution des mots et sur celle de la qualité de représentation des documents.

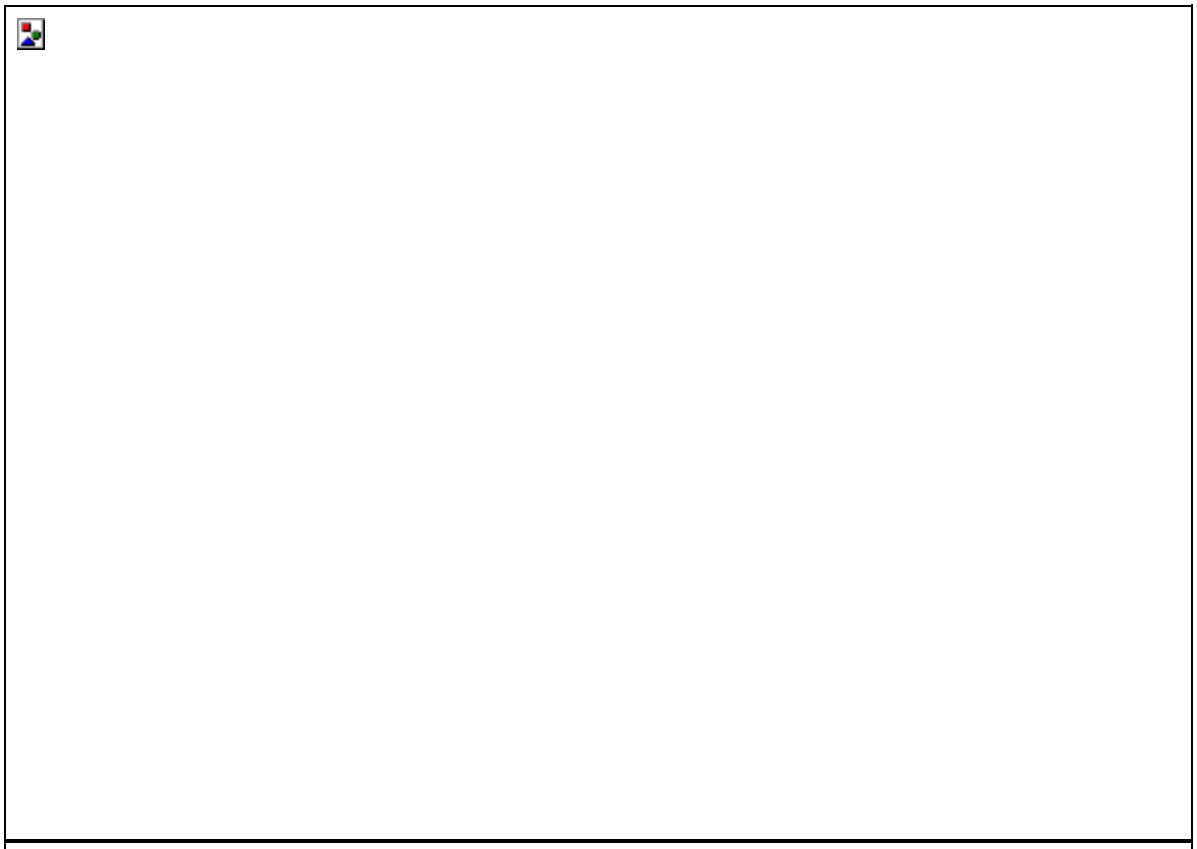
Afin de rendre l'interprétation possible, nous avons édité exclusivement les mots contribuant le plus à la construction des facteurs (seuil □2 fois à la moyenne des contributions pour chacun des axes) et les documents les mieux représentés sur ces facteurs (seuil choisi de 67 pour mille[3])

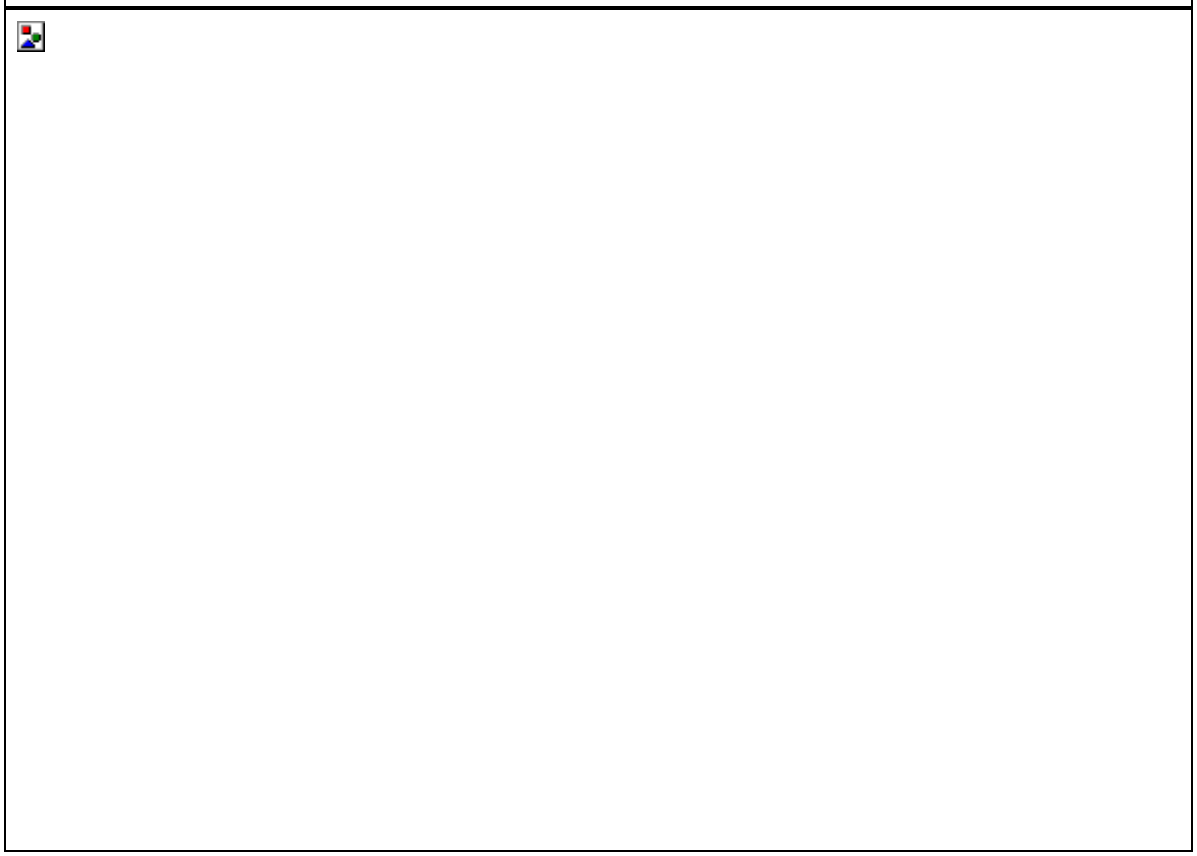
Le résultat de cette méthode appliquée aux documents répondant au besoin d'information de l'usager, est la contruction de deux types de filtres reposant sur une double représentation des axes factoriels : l'interprétation des associations par pôle positif ou négatif de chaque facteur (représentation unidimensionnelle) et celle que peuvent apporter les cartes factorielles (représentation bi-dimensionnelle) mettant en relation chaque pôle d'association.

Nous appellerons " métaclé " les pôles positifs ou négatifs de chaque facteur. La métaclé représentait, pour l'équipe de travail avec laquelle nous avons effectué ces travaux, le niveau d'intégration supérieur au mot-clé : un ensemble de mots-clés décrivant un thème et qui pouvait ensuite être réutilisé pour sélectionner une sous-base à l'intérieur du corpus de départ. Le mot-clé représente le contenu d'un texte; la métaclé caractérise un sous-ensemble de textes.

Nous avons souhaité, pour notre part, ne pas limiter la métaclé aux mots-clés mais l'étendre aux textes associés (c'est-à-dire aux documents) afin de savoir s'il était possible d'interpréter à la fois l'association de mots et l'association de textes en se servant de la première comme d'un guide de lecture pour la seconde. C'est cette double association qui constitue chaque filtre thématique à travers lequel l'usager peut lire les réponses à son besoin d'information.

Figure 3: Visualisation d'un corpus de documents par métacles : Un exemple





III.2.3 Résultats

Pour la réalisation des calculs dont nous présentons ici les résultats, nous avons utilisé le logiciel de calcul d'une équipe Inserm[4] : le logiciel BI (BI@logInserm 1979, 1987, 1993). Ce

logiciel permet le traitement des données numériques et alpha-numériques, le codage de ces données, la mise en œuvre d'outils statistiques, la préparation des données en vue de traitement par d'autres logiciels, en particulier les logiciels d'analyse de données de l'ADDAD (Association pour la Développement et la Diffusion de l'Analyse des Données).

La métaclé permet de passer des mots (l'ensemble des mots de la base) aux concepts (abstraction née de l'association). Elle est d'abord un moyen de laisser le contenu d'un sujet se dévoiler par le seul jeu des associations sans recours à des ressources externes (dictionnaire, thesaurus etc.). Elle est également un moyen de révéler des informations sur des " abstractions ".

Ainsi peut-on mettre en évidence des informations sur le marché des polymères à cristaux liquides par exemple. Or pour sélectionner ce type d'informations dans des bases de données, il n'y a pas d'autres recours que d'utiliser des bases économiques (études de marché) et de choisir dans un thésaurus un réseau de termes appropriés pour décrire la notion de " marché " car aucun descripteur ne peut à lui seul résumer cette notion.

Dans l'exemple suivant nous montrons la métaclé regroupant les informations technico-économiques sur le marché des polymères à cristaux liquides. Le corpus initial regroupe 548 documents sur les polymères à cristaux liquides (*liquid crystal polymer*).

L'interprétation repose sur la cohérence de l'association des documents. La question à laquelle la démarche interprétative répond en premier lieu est : qu'est-ce que ces documents ont de commun ? Pour y répondre, les mots de la métaclé nous donnent un guide de lecture. Dans l'exemple qui suit ils relèvent de quatre registres :

1. les noms commerciaux de matière : *Vectra, Xydar*
2. les noms de société : *Hoechst Celanaese, Dartco*, noms signalant qu'il s'agit de données relatives au marché : *company, applications, products, market, plant, Japan* .
3. le vocabulaire particulier de certains articles commerciaux, lesquels contiennent la phrase " *Brief details are given* " pour signaler que des détails sur les produits signalés sont mentionnés dans l'article intégral.
4. associés à cette dernière catégorie, des mots tels que *resistance, plastics* pour désigner les matériaux.

Figure 4 : le ciblage des données marché sur les polymères à cristaux liquides

Les documents rassemblés sont tous des documents commerciaux.

L'analyse de ces documents, le recoupement des informations et un classement par année montrait que l'on s'engageait vers :

- une augmentation de la production,
- un déplacement de la production vers l'Asie,
- une orientation vers des applications essentiellement électriques et électroniques.

L'article n°00000532, cité in extenso ci-dessus, résume assez bien ce mouvement confirmé par les autres documents dès lors qu'on les envisage ensemble et chronologiquement.

Tel est l'exemple de ce que peut apporter une typologie thématique : non seulement le repérage de thèmes mais, dans le même temps, un ciblage rapide d'informations de synthèse décrivant le thème.

Ce ciblage d'information de synthèse sert de piste au même titre qu'un indice dans une enquête et permet ensuite de refaire éventuellement une recherche d'information de vérification. C'est le point de départ d'un feed-back. Cette typologie peut s'accompagner d'une représentation classique bi-dimensionnelle (deux facteurs orthogonaux) qui est nécéssaire et complémentaire pour élucider certains défauts d'interprétation. Voici un exemple de typologie thématique pour " *Les polymères à cristaux liquides (LCP)* ", à partir d'une base de 548 documents. A chaque classe est associée en général une dizaine de documents, quelle que soit la façon de réaliser l'analyse factorielle. Nous donnons ci-après les principaux thèmes :

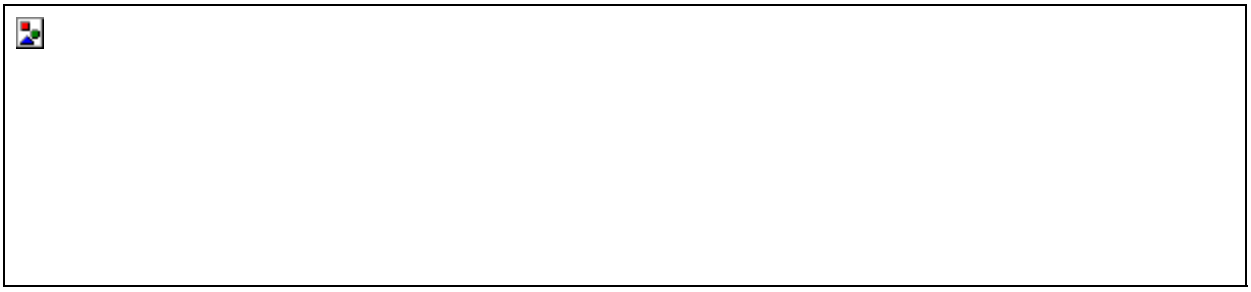
1. Les données commerciales : voir précédemment.
2. Les données matériau : études des propriétés du matériau en fonction de la phase dans lequel il se trouve (smectique, nématique , cholestérique), ces phases se caractérisent par l'arrangement des molécules entre elles. Les polymères à cristaux liquides sont en effet des liquides qui s'arrangent comme des solides : ils ont une structure ordonnée. Leurs propriétés dépendent de cette structure.
3. Propriétés, dans la transformation, des mélanges (" blends ") à base de polymères à cristaux liquides. Les LCP sont des matériaux chers et l'idéal est donc de pouvoir les mélanger avec des matériaux moins onéreux tout en obtenant des propriétés qui demeurent intéressantes.
4. LCP et fibres en particulier fibres de carbone.
5. Le problème des lignes de soudure dans la fabrication des pièces en LCP.
6. Les stratifiés (" laminates ")
7. Les LCP utilisés comme films barrière
8. Organisation en lamelles des films extrudés.

Ce parcours de lecture par métaclés peut être éclairé à la demande par des cartes bi-dimensionnelles. L'avantage des cartographies est d'abord de présenter sous forme synthétique plusieurs thèmes. La seule vision d'un plan à deux dimensions permet de voir quatre thèmes simultanément. La principale difficulté est de représenter tous les points (mots et documents) sur une carte. Les produits logiciels du marché présentent en général des cartes illisibles au premier regard mais qui comportent des fonctions "zoom" permettant d'agrandir le nuage de point et des fonctions hypertexte pour obtenir "le contenu du point". Nous avons, pour notre part, établit des cartes à partir des points sélectionnés dans les métaclés, uniquement pour les points-mots. Pour obtenir les documents il suffit de se reporter à la métaclé.

Dans l'exemple ci-après nous montrons une carte des deux premiers facteurs de l'analyse *Polymères à cristaux liquides*. Cette représentation à deux dimensions permet une superposition des métaclés (facteur 1 positif et facteur 2 positif) ayant l'une et l'autre comme objet la structure de la matière (l'essentiel des documents du facteur 2 positif faisant d'ailleurs partie du facteur 1 positif). Ce thème occupe le cadran droit de la carte. L'avantage de cette représentation à deux dimensions est de mettre en évidence, dans ce quadrant droit, des sous-thèmes que nous avons entourés : les termes ayant trait à la structure du matériau stricto sensu (*chain, nematic, smectic…*), les termes décrivant une propriété physique afférente à la structure (*field, electric* : champ électrique), termes relatifs à la cinétique de polymérisation (*theory, kinetics, polymerisation, transition…*), ceux concernant l'ordre moléculaire (*molecular, order, phase*).

Ces associations de termes ne sont visibles que grâce aux deux dimensions. Elles forment des guides de lecture du même sous-ensemble de documents que ceux mis en évidence dans l'interprétation linéaire (par métaclé). Mais elles permettent une lecture plus fine et plus rapide (les deux métaclés se superposant visiblement) de la même analyse.

Figure 5: Carte thématique



III.2.4 Comparaison entre une synthèse issue d'un groupe d'experts et une synthèse obtenue par AFC sur les mêmes documents

Pour mieux comprendre le type de synthèse effectuée par une analyse factorielle, nous avons comparé un état de l'art rédigé par la documentaliste d'un centre industriel de recherche et de développement du domaine de la plasturgie, au terme d'un travail avec le groupe d'industriels demandeurs de l'étude, et celle effectuée par une analyse factorielle sur les documents sélectionnés par la documentaliste.

Le sujet était : " le recyclage du polyuréthanne " pour lequel 528 documents ont été sélectionnés par la documentaliste. Ces 528 documents ont été lus par le groupe d'industriels, chacun s'étant réparti une partie des documents. Des réunions de mise en commun des résultats se sont ensuite produites et la synthèse finale a été effectuée par la documentaliste. La collecte des documents a été réalisée de façon empirique : pour chaque base de données certains documents ont été gardés d'autres rejetés selon l'intérêt qu'ils semblaient manifester *a priori*. Ils ont été sélectionnés comme dans toute étude documentaire classique, où les requêtes booléennes sur les serveurs de bases de données ont été couplés aux moyens intellectuels d'interprétation, eux-mêmes reposant sur la capacité de lecture des acteurs en présence. De plus, ces documents n'ont pas été obtenus à partir d'une seule et même requête mais tantôt avec le terme de *reclaim* (valorisation au sens de valorisation des déchets) tantôt avec celui de *recycling* (recyclage).

Nous avons pu reprendre ces 528 documents sous le format bibliographique (format comprenant des informations sur l'identité du document et, sur le plan du contenu : le résumé et les mots-clés). Nous n'avons pas pu obtenir les textes intégraux des articles sous forme électronique. Or certains de ces articles avaient été lus par le groupe.

Une seconde difficulté s'ajoute à celle-ci : elle consiste à comparer des synthèses issues de données sélectionnées en fonction d'une connaissance préalable du sujet (les résultats de la synthèse sont, d'une certaine façon, déjà contenus dans la sélection). Malgré ces difficultés, nous avons effectué la comparaison entre l'état de l'art et l'analyse des correspondances à partir du m^me corpus de documents afin de mettre en évidence les mécanismes de la synthèse dans les deux cas.

Le recyclage du polyuréthanne consiste à refaire un nouveau matériau ayant de bonnes propriétés à partir de polyuréthanne usagé. Dans le cas de l'état de l'art obtenu par le travail de groupe, on obtient une typologie des recyclages associée à une définition. Ainsi définit-on les différents types de recyclage mécanique (broyage/pulvérisation, moulage par compression, compression adhésive, utilisation de déchets réduits en poudre comme charge), le recyclage chimique (pétrochimie, chimiolyse).

Dans le cas de l'analyse que nous avons effectuée, les documents sélectionnés sur les métaclés et ceux cités dans l'état de l'art étaient à peu près les mêmes *mais associés différemment*. Seuls les procédés de chimiolyse se sont distingués en tant que thème de la même manière dans l'état de l'art et dans nos analyses. L'état de l'art ne s'est pas caractérisé pas par le fait qu'il contenait des informations nouvelles et singulières que l'analyse statistique aurait occultées mais bien par le type de synthèse qu'il effectuait.

Dans l'analyse statistique ce ne sont pas les types de recyclage qui sont mis en évidence mais des regroupements de documents autour, par exemple, du rôle du phosphate, participant tantôt d'un liant, d'un plastifiant ou d'un agent supprimant les fumées (les fumées issues de la combustion). Ce sont les conditions expérimentales du recyclage qui créent aussi des regroupements sur un même facteur : ainsi les termes de degrés, bar, minutes sont associés avec une forte contribution à la valeur propre et mettent en évidence un ensemble de brevets.

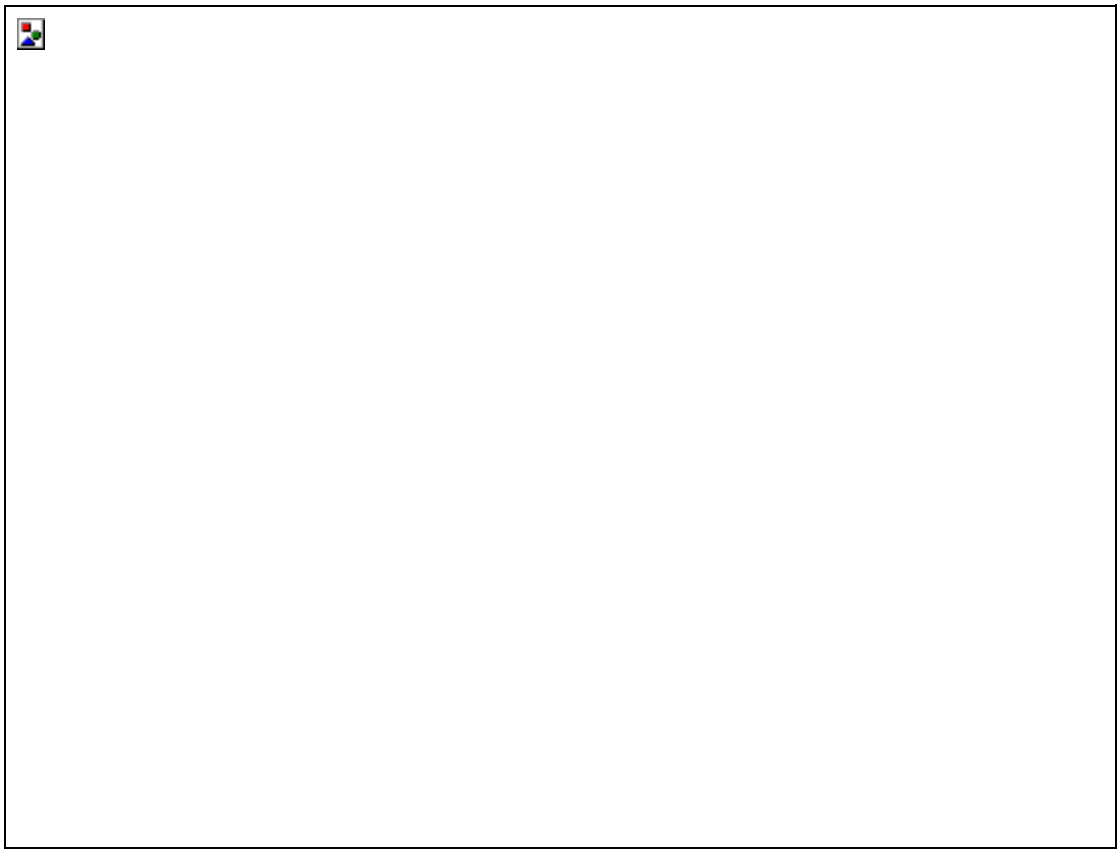
On ne peut considérer qu'une analyse statistique ne mettrait en évidence que le général alors qu'un état de l'art serait plus sensible à la singularité. La réalité est plus complexe. S'il est vrai qu'une analyse statistique ne peut mettre en évidence que des relations et donc occulte nécessairement le singulier, elle ne se cantonne pas néanmoins dans la généralité. Un facteur correspond à un groupe d'éléments qui se discriminent par rapport à l'ensemble. Le point commun, la tendance s'exprime au centre de gravité. Ce groupe d'éléments assemblés sur une partie du facteur pourrait être considéré comme une tendance locale.

Une étude statistique ne permet pas de prendre connaissance d'un sujet comme le fait un état de l'art effectué par un documentaliste ou un groupe de travail. Son intérêt est seulement de faire apparaître des informations issues d'associations non décelables à la lecture. On peut dire qu'elle met en évidence des informations qu'une lecture séquentielle du contenus des textes ne permet pas d'appréhender.

III.2.5 Prototype de système intégrant le filtrage statistique

Voici le schéma général de méthode utilisée (figure 6) :

Figure 6 : Prototype de système intégrant le filtrage statistique



Quatre modes de recherche sont permis par le système :

1. Le corpus de travail contenant tous les documents concernant le sujet d'étude que nous avons appelé "sous-base" est consultable par un moteur de recherche classique (équations logiques). Toutes les bases concernant tous les sujets traités par l'utilisateur sont capitalisées et interrogeables de cette manière. C'est le mode de consultation n°1.
2. Les résultats des analyses factorielles sont visualisables sur le même serveur. On peut entreprendre la connaissance du sujet d'étude en navigant à l'intérieur de la typologie thématique représentée par les différentes métaclés : c'est le mode de consultation n°2. Sur chaque métaclé les associations de mots et documents permettent d'obtenir cette connaissance. Le simple accès à des documents (mode n°1) permet de chercher un document ou mot particulier; le mode n°2 nous donne un aperçu de la façon dont le sujet est traité et nous invite à une première lecture guidée des documents.
3. Grâce au mode n°3, il est possible de passer des résultats d'une équation logique donnés par le moteur de recherche (mode 1) à leur situation dans la typologie thématique. Ces résultats sont des documents dont on peut obtenir la position sur l'un des facteurs de l'analyse. Ainsi le document apparaît "en contexte" sur une métaclé.
4. Le mode 4 permet de naviguer à l'intérieur de l'hyperplan d'un document c'est-à-dire de l'ensemble des métaclés sur lesquelles il est éventuellement présent. Si le document est multithématique, il pourra être interprété selon les différents axes d'interprétations suggérés par les métaclés.

La difficulté de ce système réside dans les techniques de visualisation liée à la représentation multidimensionnelle. Plus une représentation comportera de dimensions moins l'analyse révélera d'ambiguïtés, mais plus elle deviendra difficile à comprendre et à regarder. En d'autres termes on retrouve la difficulté que la représentation géométrique tente de lever… La

représentation linéaire est la seule permettant de lire l'ensemble des points mots et documents sélectionnés et elle est simple à interpréter.

L'idéal est sans doute, pour la recherche d'information dans les bases spécialisées, d'allier filtrage statistique et méthode de visualisation en s'inspirant des travaux réalisés sur la visualisation appliquée à l'analyse de textes (voir par exemple Rockwell et Bradley 1999). Mais si l'outil statistique permet d'envisager une représentation visuelle des réponses à un besoin d'information, il reste à trouver des techniques de représentation, à la fois exhaustives, faciles à comprendre, et dont l'usager aurait la maîtrise afin de reformuler son besoin d'information.

IV. Conclusion

Nous avons voulu montrer, dans ces études, la richesse des techniques de filtrage applicables aux documents numériques. Elles restituent à l'usager un véritable pouvoir d'interprétation de son besoin d'information et des réponses qui lui sont données en lui offrant une boîte à outils pour organiser les informations. Il faut encore ajouter que dans les techniques que nous avons proposées nous avons cherché à gérer la multidisciplinarité et la multi-thématicité des documents.

IV.1 Filtrage d'information et interdisciplinarité dans les fonds encyclopédiques

Les classifications ont toujours eu à résoudre deux problèmes classiques mais qui sont toujours d'actualité : comment éviter le fractionnement d'un thème et comment représenter l'interdisciplinarité ? On ne peut pas donner deux cotes à un livre. Un ouvrage est soit dans une classe X, soit dans une autre classe Y. Le savoir actuel est de plus en plus multidisciplinaire. Or, les grandes classification sont encyclopédiques et présentent chaque discipline académique. Cette structuration s'accommode mal des besoins de transversalité entre les sciences.

Concernant le filtrage, le problème se pose d'une manière différente. Rien n'interdit de séparer l'indexation systématique de la cotation. Un ouvrage peut avoir par exemple deux ou trois indices de classification mais seulement une cote. On peut retrouver un ouvrage dans le sous ensemble "informatique" mais aussi dans le sous ensemble "science cognitives". Ce sont des copies virtuelles. A travers les liens hypertextes, on pourra lier ces copies à la cote.

En ce qui concerne, les documents électroniques, le problème du classement physique ne se pose pas. Il est donc possible de donner deux ou plus indices à ce document, facilitant ainsi une recherche pluridisciplinaire. Tinker (1999) et Beghtol (1999) préconisent aussi d'assigner plusieurs indices de classification à un document . Cette solution est déjà mise en œuvre dans la base bibliographique ITER (http://iter.library.utoronto.ca/iter/index.htm).

IV.2 Multidimensionnalité thématique d'un document dans le cas du filtrage statistique sur fonds spécialisé

On peut non seulement interpréter des associations de mots et de documents (des thèmes) mais aussi ce que l'on peut appeler l'hyperplan d'un document c'est-à-dire l'ensemble des métaclés (et donc des thèmes) sur lesquelles il est représenté. Un document peut être monothématique s'il n'est présent que sur une dimension, mais peut être multithématique : son interprétation peut alors se faire soit par rapport à sa spécificité ou par rapport à

l’ensemble des thèmes qui le constituent. Il s’envisage ainsi à partir de contextes variés constitués de mots et de documents associés différents.

Il en est ainsi pour le document 99, que nous citons ci-dessous (Figure 3). Ce document fait partie de l’une de nos analyses sur les capteurs utilisés dans les procédés de transformation des polymères. Il est représenté à la fois sur le facteur 1 et sur le facteur 4. Il est envisagé sous deux aspects différents : sur le facteur 1 il renvoie au contexte des capteurs impliqués dans la transformation des composites; sur le facteur 4 il fait partie du thème général relatif au contrôle des process (du point de vue des matières : composites, caoutchouc, thermoplastique... ou des process /injection, extrusion).

Les même termes (communs aux deux facteurs) que sont cure et resin sont associés :

- dans le facteur 1 à des noms de matière de *fibre, epoxy* (thème : composite et thermodur) et à model et dielectric montrant que l'enjeu est ici la modélisation des propriétés de ce type de matière lors de la transformation (dans ce document 99, il s'agit du Resin Transfer Moulding process). C'est le thème de l'ensemble des documents de ce facteur, dont nous présentons les titres ci-après.
- dans le facteur 4 à ceux de *control, developments, automation* illustrant le thème du " *process control* " (contrôle des procédés).

HYPERPLAN D'UN DOCUMENT:

Le document suivant s’exprime à la fois sur deux métaclés (facteur 1 positif, facteur 4 négatif) exprimant par là deux thèmes

Figure 7 . Hyperplan d'un document (suite). Les dimensions d'un document

Notes

[1] Voir CAIS 1999.

[2] Les travaux dont nous présentons ici les résultats ont été réalisés dans le cadre d'une collaboration avec Michel Kerbaol, Département Santé Publique, Faculté de Médecine, 35000 Rennes, France.

[3]Sachant que le cosinus carré est égal à 1 et est multiplié par mille pour une meilleure lisibilité.

[4]INSERM : Institut National de la Santé et de le Recherche Médicale. Le logiciel BI a été conçu et développé par M.Kerbaol et A.Josse.

Bibliographie

Beghtol C (1999). *Knowledge domains: multidisciplinary and bibliographical classification systems.* Knowledge organisation, 25 (1/2), 1-12.

Benzécri JP (1980)

Bruza, P.D. and Dennis, S. (1997). Query re-formulation on the Internet: Empirical Data and the Hyperindex Search Engine. In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, 488-499, Centre de Hautes Etudes Internationales d'Informatique Documentaires.

Chan L M (1995). *Classification, present and future*. Cataloging and classification quarterley. 21(2), 5-17.

Clark C (1997). *Relevance ranking for one to three term queries*. In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, 388-400, Centre de Hautes Etudes Internationales d'Informatique Documentaires.

Cutting, D.R., Karger, D.R., & Pedersen, J.O. (1993). *Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections*. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, 126-131

Deerwester S. et alii [1990], "Indexing by latent semantic analysis", *Journal of the American Society for Information Science [JASIS]*, vol. 41, n°6,pp. 391-407

Grefenstette G (1997) . *SQLET : Short query linguistic expansion techniques*. In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, pp 500-509, Centre de Hautes Etudes Internationales d'Informatique Documentaires

Hearst M (1995) . *TileBars: Visualization of Term Distribution Information in Full Text Information Access*, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Denver, CO, 1995.

Hudon M (1994) : " Le thesaurus : conception, élaboration, gestion ", Editions ISIED.

Ihadjadene M (1999a). *Les échecs et la surcharge d'information dans un SRI: examen des tactiques mises en œuvre par les usagers.* Proceedings of the 27th Annual CAIS/ACSI Conference , Turner (eds), University of Sherbroke, 9-11 june, 1999, Canada.

Jones, S. Cunningham SJ (1998). *An analysis of usage of a digital library*. Proceedings of the 2th european conference for digital library, Crete, pp 261-277

Pollit, A.S & Tinker A (2000). *Navigating N-dimensional information space with data and documents through view-based searching*. BCS-IRSG 2000 Meeting , Cambridge (à paraître)

Rockwell et Bradley 1999

Silverstein C, Marais H (1998). *Analysis of a very large Altavista query log*. SRC technical Note 1998-014. Digital, Palo Alto.

Spink, J. Bateman, and B. J. Jansen (1998). *Searching Heterogeneous Collections on the Web: Behavior of EXCITE Users*. Proceedings of the 1998 National Online Meeting, May, New York, 1998

Spoerri, A. (1995). *InfoCrystal: A Visual Information Retrieval Interface*. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, 367.

Tinker A J , Pollit A & Braekevelt (1999). *The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation*. Knowledge organisation, 26 (2): 80-96.

Retour au haut de la page Table des matières