

A. LELU

D. ROSENBLATT

Représentation et parcours d'un espace documentaire. Analyse des données, réseaux neuronaux, et banques d'images

Les cahiers de l'analyse des données, tome 11, n° 4 (1986), p. 453-470

http://www.numdam.org/item?id=CAD_1986__11_4_453_0

© Les cahiers de l'analyse des données, Dunod, 1986, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

REPRÉSENTATION ET PARCOURS D'UN ESPACE DOCUMENTAIRE. ANALYSE DES DONNÉES, RÉSEAUX NEURONAUX, ET BANQUES D'IMAGES

[ESPACE DOCUMENTAIRE]

par A. Lelu, D. Rosenblatt

0 Introduction : recherche d'un dialogue avec une base documen-

taire : L'interrogation en langage naturel semble aujourd'hui la voie toute tracée pour le progrès en matière de systèmes documentaires. Nous ne contestons certes pas qu'un degré élevé de tolérance de ces systèmes aux fautes de frappe et d'orthographe, à l'emploi de synonymes, soit d'une utilité essentielle dans la relation entre l'homme et la machine, surtout quand il s'agit de mettre ces systèmes à la portée de couches de plus en plus larges d'utilisateurs. Mais il nous semble que la quête d'un dialogue quelque peu anthropomorphique avec l'ordinateur est prématurée tant qu'un certain nombre de problèmes de *représentation* ne sont pas résolus : comment fournir à l'interlocuteur humain une vue d'ensemble du ou des points de vue éditoriaux inscrits, implicitement ou non, par les gestionnaires et indexeurs du fonds documentaire ? Il faut bien que cet interlocuteur se fasse une idée de ce qu'il peut trouver, de ce qu'il ne trouvera pas à coup sûr, bref il faut qu'il sache à qui et à quoi il a affaire ! Quels outils lui donner pour se repérer tout au long de sa recherche, et la faire progresser à tout moment dans les directions les plus pertinentes de son point de vue ?

0.1 Se représenter une base documentaire et s'y repérer : Une condition sine qua non de toute communication humaine est la nécessité pour chacun de posséder une représentation de son interlocuteur, non seulement de la manière dont il comprend, mais aussi de ce qu'il sait ; représentation qu'il ajuste constamment en cours de dialogue ; personne, sauf les fous, ne parle "dans l'absolu", sans tenir compte de l'autre.

Le dialogue avec un ordinateur n'échappe pas à cette règle : à notre sens comme à celui de beaucoup d'ergonomes, mieux vaut afficher clairement les limites d'un système que d'ennuyer les gens à les découvrir par essais et erreurs - d'où l'angoisse bien compréhensible des utilisateurs occasionnels devant les systèmes muets du type "posez votre question, nous nous chargeons du reste" ...

Autant il est facile de souhaiter que l'utilisateur puisse avoir une vue d'ensemble d'une base documentaire et sache à tout moment se situer dans la "région" qu'il explore, autant réaliser ce souhait n'a rien d'évident : la liste des mots-clés, même structurée en thésaurus, est trop longue, et inévitablement ambiguë ; il est encore plus opaque de saisir quels documents figurent, en gros, et lesquels ne figurent pas [LELU 86].

(*) DGT. Service de la Prospective et des Etudes Economiques.

(**) Laboratoire d'Electronique de l'ESPCI.

Fourtant le tableau de correspondance exhaustif [descripteurs × objets décrits] existe implicitement dans les fichiers documentaires, et les méthodes d'analyse des données ne manquent pas, pour peu qu'elles sachent traiter des tableaux assez vastes, pour en tirer une information synthétique très riche : celle-ci pourra s'interpréter en termes de politique éditoriale du fonds documentaire (choix des documents), en terme de grands faisceaux de sujets traités et de manières de les traiter.

Il existe également de nombreuses voies pour l'exploration "locale" du tableau exhaustif de correspondance, c'est-à-dire de l'environnement immédiat d'un descripteur ou d'un objet décrit ; cette exploration plus fine est à notre sens le complément indispensable de l'information synthétique globale.

Nous passerons en revue un certain nombre de réalisations partielles dans ces domaines ; elles s'inspirent davantage des méthodes d'analyse statistiques des données, de la reconnaissance des formes, d'applications de la théorie des graphes, que de ce qu'on entend généralement par "intelligence artificielle". Nous n'abordons pas les nombreux travaux qui tendent à introduire dans ce dernier domaine des représentations "floues" ; dans tous les cas notre sentiment est que seules des méthodes mettant en oeuvre un auto-apprentissage des données par le système sont adéquates pour résoudre le problème sur un fonds : 1) de grande taille, 2) qui s'enrichit en permanence.

Nous exposerons les points de convergence et de divergence entre analyses factorielles et certains modèles de réseaux cellulaires à auto-apprentissage et à réponse analogique [KOHONEN 84, HERAULT 85]; nous présenterons quelques résultats de simulations de tels réseaux d'automates sur des données contrôlées, puis nous ouvrirons des perspectives sur l'exploitation de données issues d'une base documentaire d'images fixes réelle.

Notre recherche porte sur les types de lois d'apprentissage, sur les lois d'"inhibition latérale" entre neurones, et sur la géométrie de leurs interconnexions, susceptibles d'aboutir à des représentations pertinentes pour un utilisateur de bases documentaires.

0.2 Parcourir une base documentaire et accéder à l'information au moyen des descripteurs, mais aussi des objets décrits :

La faiblesse majeure des systèmes informatiques par rapport au support papier est l'étroitesse de la "lunette" de visualisation que l'on braque sur des masses de données en général considérables. Ceci est aggravé, dans le cas des systèmes documentaires à accès distant, par la lenteur d'affichage (300 ou 1200 bits/s.) ; la conception des logiciels documentaires actuels est bâtie sur ces contraintes structurelles, et il faut tout l'art des documentalistes spécialisés, leur grande maîtrise dans les stratégies de recherche et leur connaissance de toutes les subtilités des langages d'interrogation pour réaliser un fragile et instable équilibre entre le "bruit" documentaire qui submerge l'utilisateur sous des centaines de documents impossibles à passer en revue - et le calibrage idéal d'une ou deux dizaines de documents - dont on n'est jamais sûr qu'il ne laisse pas de côté beaucoup d'éléments pertinents.

Ceci par contraste avec les documents sur papier ou les diapositives disposées en planches qui permettent un feuilletage ultra rapide et un furetage au petit bonheur ; bien sûr ce feuilletage ne peut se faire que selon le seul ordre de rangement physique : alphabétique pour les dictionnaires et les annuaires, par reportage pour les planches contacts ou diapos, etc. - ce qui en constitue la faiblesse réhibitoire.

L'apparition des vidéodisques analogiques, et prochaine des dispositifs d'archivage sur compact-disc et disque optique numérique, dans un contexte de transmissions à 64 kbits/s et 2 Mbits/s, ouvre en définitive des perspectives de simplification de la recherche documentaire : le feuilletage des documents primaires, ou d'images suffisantes de ceux-ci, redevient possible, et cette fois selon un grand nombre de critères de feuilletage, par contraste avec les supports physiques traditionnels.

Dès lors une nouvelle approche de la recherche documentaire nous semble inévitable ; jusqu'à présent l'utilisateur se bornait à communiquer au système des mots-clés ou des combinaisons logiques de ceux-ci : il peut désormais désigner au hasard d'un feuilletage les objets documentaires pertinents à ses yeux ; ceux-ci seront, en termes de documentaire, "ré-injectés en tant que question" et provoqueront une nouvelle réponse du système.

Ce qui nous a amené à développer une idée de symétrie, de dualité entre le rôle des descripteurs et celui des objets décrits ; cette notion, issue de la symétrie formelle entre les lignes et les colonnes d'un tableau de description, est familière aux statisticiens de l'école française d'analyse des données [BENZECRI 1973]. Elle tient dans la proposition : "à chaque objet documentaire est associé un ensemble de descripteurs, à chaque descripteur est associé un ensemble d'objets documentaires".

Nous avons pris le parti de tirer jusqu'au bout les conséquences de cette dualité dans le domaine des bases documentaires de photos, en concevant et réalisant des dialogues totalement symétriques d'accès à partir des mots-clés aussi bien qu'à partir des images ; cet exercice de changement de perspective s'est révélé fructueux sur le plan heuristique, ... mais également encourageant sur le plan opérationnel, comme l'a montré l'évaluation ergonomique de notre maquette ICONOTRON (P.H. DEJEAN, Université de Technologie de Compiègne, Avril 1985) ; les fonctionnalités de cette maquette seront abordées dans notre dernier chapitre.

0.3 Fusion de deux approches dans le logiciel que nous développons :

Dans la suite de nos travaux les approches "représentations synthétiques globale et locale" et "symétrie, dans le dialogue, entre descripteurs et objets décrits" se sont fondues de façon naturelle grâce au concept, lui aussi issu de l'analyse des données, d'espace de représentation commun pour les lignes et les colonnes d'un tableau de description.

Chaque désignation, ou mise en exergue, d'un ou plusieurs éléments (mots-clés ou photos) provoque le calcul d'un environnement local de mots-clés et de photos que l'utilisateur pourra explorer, en se repérant à l'aide de sa "boussole".

Avant de décrire plus précisément ces travaux, un approfondissement des notions de représentation s'impose, de façon à les situer par rapport aux domaines connexes : bases de données relationnelles, systèmes experts, traitement du signal, reconnaissance des formes, etc. .

1 Représentations par règles et représentations d'ensembles

d'observations : Les représentations utilisées dans les applications avancées de l'informatique (SGBD, Intelligence Artificielle) ont la propriété d'être "dures" : une règle ce système expert ou un tableau de relation dans un SGBD relationnel sont vrais de façon absolue dans tout le domaine de l'application, au sens strict et logique du terme;

ils ne tolèrent par définition aucune contradiction, ni aucune marge d'interprétation. Les tentatives faites pour "adoucir" ces systèmes par l'introduction d'ensembles flous (par exemple PRADE 84) semblent nécessiter que "quelqu'un" définisse et introduise des valeurs de probabilité pour chaque relation aussi systématiquement et rigoureusement que les valeurs de vérité dans les systèmes précédents.

Ces règles posées *a priori* façonnent de façon externe les relations de sens entre les concepts d'un micro-monde généralement très étroitement limité à une spécialité (médicale, géologique, ...) et à un type de problème à résoudre. Si l'on plonge ce micro-monde mécanique et rigide dans une pratique humaine et une réalité physique, et que ses réponses ou son comportement ne sont pas adéquats, il faut l'intervention d'experts humains pour ajouter ou changer des règles.

Au contraire, d'autres disciplines du traitement de l'information, comme l'analyse statistique des données et la théorie du signal, sont basées, de par la nature des problèmes qu'elles se posent, sur la notion de traitement d'un ensemble d'observations ; ces observations définissent de façon exhaustive et purement interne le "micro-monde" sur lequel porte l'application - dans un cadre général, posé évidemment *a priori* (choix d'un échantillon de descripteurs et d'objets décrits, d'une métrique, d'une méthode de réduction de dimensions) ; les représentations qui en découlent sont "molles", non figées ; ces représentations peuvent varier de façon continue si on accroît le nombre d'observations, le nombre de variables observées, ou si l'on change le "grain" de l'observation. La qualité qu'on exige d'elles est précisément la stabilité, la "robustesse" au regard de ces changements, bien sûr tant que ceux-ci restent dans les limites raisonnables d'un univers d'observations bien cerné au départ.

Pour résumer, on peut dire que dans le cas des représentations "dures" le micro-monde considéré est défini de façon figée par un ensemble de règles issues de ce qu'on sait de lui dans le vaste monde de la connaissance humaine, alors que dans le cas des représentations "molles", ce micro-monde est défini de façon évolutive par les seules relations observées dans un ensemble d'objets décrits et de descripteurs.

Alors que la classe des problèmes posés consistait, pour les représentations "dures", à alimenter explicitement une prise de décision en vue d'une action, dans le deuxième cas le but est fondamentalement de réquiere l'information, pour en rendre de grandes masses accessibles à l'esprit humain, ou pour la faire transiter par des canaux à débit limité.

Ce qui a d'importantes conséquences techniques : la première approche (dure) a l'avantage de tirer parti de la "puissance du combinatoire", c'est-à-dire qu'un ensemble limité de règles suffit à embrasser une foule de cas spécifiques qu'il n'est pas nécessaire de lister de façon explicite, qualité qu'on trouve, parmi beaucoup d'autres, dans le langage humain ; mais elle a aussi l'inconvénient de cet avantage : il faut limiter de façon drastique le nombre de concepts et de règles à l'oeuvre pour éviter l'"explosion combinatoire" et parvenir à des temps de traitement réalistes. Au contraire l'approche "statistique" se contente de traiter les *combinaisons réellement observées* entre des descripteurs et des observations en très grand nombre, pour en déduire une représentation simplifiée "cadrée" sur les limites des facultés mentales humaines, et également sur des temps de calcul acceptables.

Prenons l'exemple des représentations issues de l'analyse factorielle des correspondances [BENZECRI 73], c'est-à-dire des cartes sur lesquelles sont placés des points représentatifs des variables d'observation aussi bien que des observations elles-mêmes ; elles ont exercé

- et continuent d'exercer - une véritable fascination sur beaucoup de professions confrontées au dépouillement de vastes enquêtes (sociologiques, psychologues, chargés d'études de marchés, ...) : qu'une méthode parfaitement mécanique, une "moulinette" informatique, puisse restituer des rapprochements de données qui aient du sens, et même un sens riche, est un phénomène digne d'émerveillement en un temps où l'utilisation de l'ordinateur est fréquemment synonyme de représentations pataudes et logiques jusqu'à l'ubuesque ! Bien sûr la satisfaction d'obtenir du "quantitatif" à partir de données éventuellement qualitatives au départ, est illusoire puisque ce "quantitatif" s'avère le plus souvent rebelle à toute interprétation autre que qualitative - sauf à s'astreindre à une longue ascèse de retour sur les données initiales (cf. [CIBOIS 80] pour une excellente discussion sur la pratique et l'histoire récente de l'analyse factorielle en France). La boutade de G. TH. Guibaud : "*l'analyse factorielle, c'est comme l'astrologie au Moyen-Age, ça permet à des gens intelligents de dire des choses intelligentes ...*" résume bien les choses au travers - tout compte fait - de son ambiguïté (ça n'est déjà pas si mal de dire des choses intelligentes!). Quoiqu'il en soit, cette technique constitue un remarquable outil de représentation, dont l'usage a prouvé qu'il ne demandait pour son utilisation pas de spécialisation poussée en informatique, ni en statistique mathématique.

Autre exemple issu de la théorie du signal : l'holographie, en réduisant la représentation des signaux à celle de leur auto-corrélation, place en chaque point d'un hologramme une information sur la totalité de l'objet ; quand on casse un hologramme, chaque fragment conserve l'image de l'objet entier, en plus flou. Ce point de vue inhabituel, dans la pratique scientifique et technique usuelle, sur le rapport du tout et des parties, a pu fasciner certains au point d'apparaître comme un paradigme du fonctionnement réel du cerveau [PINSON 85].

Ces méthodes ont en commun de considérer qu'un ensemble d'observations définit une totalité d'éléments en interrelation ; par exemple un descripteur est lié à tous les objets qu'il décrit, donc est lié aussi (au 2-ème degré) à tous les descripteurs présents dans ces objets, donc également à tous les objets indexés par ces descripteurs ... etc. ; l'analyse factorielle présente elle aussi des aspects "holographiques" : amputer plusieurs facteurs extraits ne dégrade pas les données en tout au rien, mais rend la reconstitution de celles-ci plus imprécise.

Dans ces exemples il n'y a pas à définir *a priori* une structure de relations obligatoires au sein d'un ensemble de variables : dans le cadre d'un modèle très général ces méthodes font apparaître ces relations *a posteriori* aux yeux d'un observateur humain ou d'un système technique par l'utilisation des notions d' "apprentissage", d' "adaptativité". Des concepts comme ceux de filtrage adaptatif, d'opérateur de projection sont à même de jeter des ponts entre ces spécialités aujourd'hui largement étanches. En un sens ces méthodes sont plus "mécaniques" que les représentations "dures" de l'intelligence artificielle, alors qu'elles produisent des représentations plus conformes à la faculté d'appréciation et de pondération humaine!

Signalons deux domaines hybrides :

La reconnaissance des formes a pour but la décision automatique en vue d'une action, mais sa matière première est la même que celle de l'analyse des données et du traitement du signal ; c'est pourquoi l'éventail de ses méthodes est particulièrement large, depuis les statistiques descriptives jusqu'aux méthodes les plus "structurelles" ; certains développements centrés sur des considérations d'exploration locale des données, comme la notion de "dimension locale intrinsèque d'un nuage de données" mériteraient à coup sûr d'être transposées dans le cadre des bases documentaires (cf. [SIMON 84] pour un panorama très complet de ces méthodes).

Les méthodes d'analyse linguistique automatique nous concernent moins puisqu'elles ont pour but le traitement du langage naturel, dont nous sommes très loin puisque nous nous plaçons délibérément au départ dans le cas des descriptions par ensembles non structurés, informes, de mots-clés ; elles sont partagées elles aussi entre une tendance statistique, minoritaire, attachée à la collecte et au traitement des observations [GROSS 76, BENZECRI 81], et une tendance formelle très majoritaire, initialisée par les travaux de Chomski sur l'application de la théorie des grammaires formelles aux langues naturelles, et ramifiée en nombreuses spécialités proches de l'intelligence artificielle (réparties selon les niveaux morpho-syntaxique, sémantique, pragmatique).

2 Réalisations existantes dans le domaine documentaire

Parmi les systèmes documentaires dont le "langage de description" des objets documentés ne comporte aucune syntaxe et est réduit à la simple juxtaposition de mots-clés, c'est-à-dire l'immense majorité de ceux-ci, un certain nombre de tentatives existent pour fournir à l'utilisateur une vue globale du contenu documentaire d'une part, des outils de repérage et de navigation locaux d'autre part.

2.1 Structuration globale de l'espace documentaire : Nous ne nous étendrons pas sur les systèmes courants offrant une structuration *a priori* des mots-clés dans le moule rigide d'une ou plusieurs arborescences (thésaurus hiérarchique ou poly-hiérarchique) : leur usager doit nécessairement intégrer un point de vue unique sur la signification des catégories et leur emboîtement, celui du concepteur du thésaurus ; ce qui laisse entiers les problèmes de polysémies, de glissement de sens en fonction des contextes, de conflits entre catégorisation "dans l'absolu" et catégories réellement représentées dans le fonds documentaire, d'unité éditoriale de la base au fil du temps et des changements d'analystes, etc. .

Nous présenterons quelques réalisations, généralement au stade expérimental ou de maquette de laboratoire, qui procèdent de la démarche inverse : présenter une vue synthétique des relations existant entre les mots-clés - ou les unités documentaires - à partir de traitements statistiques sur les fichiers documentaires déjà constitués (liste sans prétention à l'exhaustivité).

- Le dispositif proposé et testé sous forme de maquette "manuelle" par Pascal Sanson dans le cadre de sa thèse "Methodologie mathématique/Informatique relative à la structuration et au classement d'un corpus d'images" (EHESS, 1978, sous la direction de Marc Barbut) [SANSON 78] utilise l'analyse factorielle des correspondances du tableau [mots-clés] x [photographies] pour placer sur les principaux axes factoriels extraits à la fois les mots-clés et les photos. Les tests sur un corpus réel bien que limité (900 photos sur l'Algérie indexées par 80 mots-clés) ont montré qu'un sens évident se dégagait du défilement des photos et des mots-clés dans l'ordre de chaque facteur (par exemple : axe Maghreb/Occident, axe paysages désertiques/présence humaine, ...), en plus de résultats plus "pointus" (place des diverses formes de la guerre et de l'univers marin). Des thèmes de recherche voisins ont été également développés par J.P. Trystram dans le séminaire "documentation image" qu'il a animé à l'EHESS en 1977-78.

- La maquette de banque de petites annonces développée à l'IRISA, à Rennes, [LERMAN, PETER 85] utilise une méthode de classification hiérarchique en parallèle qui permet de dégager un ensemble de classes homogènes et de calibre voisin (compatible avec la consultation par un utilisateur moyen), l'arborescence permettant d'élargir la recherche ; la particularité de cette méthode est de prendre en compte une information extérieure aux seules données et fournie par un expert:

un "pré-ordre" entre couples de modalités de chaque variable, exprimant l'intensité de la ressemblance entre les modalités d'une même variable (par exemple pour la variable "objet de la transaction" : un studio ressemble plus à une chambre qu'un appartement à un pavillon ; ces derniers ont plus de ressemblance entre eux que le couple local/terrain, etc.), ce qui permet d'accepter toute une gamme d'intermédiaires entre variables à modalités purement qualitatives et variables à modalités ordonnées. De ce fait cette maquette est un exemple de transposition au domaine de la documentation automatique de procédures d'analyse de données sous contraintes externes définies *a priori* (ordre ou pré-ordre entre modalités, contraintes de contiguïté géographiques, temporelles, etc. entre objets décrits [LEBART 1969] [CARLIER 85]), préoccupation que nous éliminons d'emblée en première phase de notre recherche.

- La chaîne de programmes LEXIMAPPE, issue des travaux du Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris [CALLON 85] a pour but explicite de tirer des bases documentaires bibliographiques une information sociologique, lexicale et stratégique. Mais l'opération technique qu'elle réalise va dans le sens de nouveaux outils d'aide à la consultation : un traitement des fichiers documentaires permet de dégager, à partir du tableau des cooccurrences entre mots-clés, des graphes de proximité entre termes ; des indices d'inclusion "floue" entre termes permettent de tracer des arborescences lexicales *a posteriori*.

- De nombreux travaux sur le traitement automatique des textes en langue naturelle, orientés vers l'indexation automatique, l'interrogation en langage naturel, ou la recherche lexicologique, ont également des points communs avec notre façon de voir :

. le logiciel SPIRIT pour système documentaire "texte intégral" [ANDREWSKI 73], en fonctionnement opérationnel, entre autres, sur la base de données juridiques du CEDIJ, utilise explicitement une "distance informationnelle" pour pondérer les mots-clés d'un texte indexé, les comparer à une requête formulée en langage naturel, et en déduire une liste de documents proches de la question posée ordonnés par proximité décroissante,

. le prototype DIALECT [BASSANO 81], interface de communication en langage naturel avec une base documentaire classique, fait appel à la dualité descripteurs/objets décrits : un document ou un ensemble de documents, peut être "ré-injecté en tant que question" dans le système documentaire pour élargir la recherche,

. les travaux lexicologiques d'A. Aït Hamlat [AIT HAMLAT 84] utilisant l'analyse factorielle des correspondances sur un corpus de résumés documentaires (rapports de mission des chercheurs d'ELF-Aquitaine) débouchent sur une représentation des mots-clés et des documents dans un espace à 3 ou 4 dimensions principales clairement interprétables (axe énergie/information, axe énergie classique/énergies nouvelles, etc.), qui pourrait être réescomptée en usage documentaire.

2.2 Structuration locale et parcours : Listons tout d'abord quelques systèmes - en fonctionnement opérationnel - utilisant "localement" des listes de cooccurrences :

- Le système ADM (Aide au Diagnostic Médical) du P^r. Lenoir, à Rennes, est une sorte d'encyclopédie documentaire en ligne où chaque maladie est indexée par la liste de ses symptômes [LENOIR 82]. Etant donnée une liste de N symptômes, le système détermine par "halos" successifs les maladies correspondantes aux cooccurrences des N symptômes, si elles existent, puis celles qui en possèdent N-1 parmi N, puis N-2,

etc. ce qui incite à évoquer des maladies rares ou à symptômes trompeurs ; ce qui revient également à lister les maladies proches d'une maladie donnée.

- La commande ZOOM sur le serveur documentaire ESA-IRS a obtenu en 1984 le prix du meilleur produit documentaire [GIRARD 85] ; elle permet, pour chaque descripteur de la base, d'obtenir la liste des descripteurs proches par ordre de nombre décroissant de cooccurrences ; c'est une aide considérable pour les utilisateurs qui découvrent sous leurs yeux des termes auxquels ils n'auraient pas pensé de prime abord.

Remarque : Il est paradoxal de constater que pour ces bases documentaires où le nombre de descripteurs est du même ordre de grandeur que celui des objets documentaires, l'utilisation des cooccurrences n'a été prévue que dans un seul sens : objet vers objets dans le cas d'ADM, descripteur vers descripteurs dans le cas d'ESA-IRS ! Pourtant il serait sans doute très intéressant de connaître les symptômes proches d'un symptôme donné ... Dans le logiciel que nous développons les cooccurrences sont bien sûr utilisées dans les deux sens !

D'autres systèmes opérationnels utilisent, en plus des classiques listes inverses, des structures particulières de fichiers où chaque élément pointe sur un nombre limité de proches voisins, ce qui permet de parcourir, de proche en proche, de longues chaînes d'éléments, et de pouvoir à tout moment changer de chaîne :

- IMEDIA a réalisé sur ce principe les prototypes de banques d'images "Philatel" et "Enluminures de la bibliothèque Ste Geneviève" [KLOSSA 83] ; chaque photo est à l'intersection d'autant "d'axes" de parcours qu'elle comporte de mots-clés ; on peut feuilleter très rapidement les photos correspondant à chaque mot-clé, appeler la "fiche" d'une photo qui nous a frappé visuellement, et repartir en feuilletage sur l'axe d'un nouveau mot-clé ; en outre les photos peuvent être visualisées par "planches" sur plusieurs téléviseurs ; il est possible d'effectuer sa propre sélection de photos dans des "boîtes" qu'on peut alors croiser avec les mots-clés existants. Très en avance sur les logiciels de banques d'images qui se contentent de plaquer une commande "visualisation" sur un système documentaire classique, la partie "navigationnelle" de ce logiciel n'en n'est pas moins d'une simplicité d'emploi extrême ; mais, faute de point d'ancrage et de "boussole", il est facile de se perdre dans ce labyrinthe multidimensionnel, et on peut longtemps tourner en rond dans des "isolats" de photos et de mots-clés fortement liés.

- André Dewèze a mis au point un mode d'exploration extrêmement complet, original et simple d'usage dans le cadre de la base documentaire Merlin-Gerin [DEWEZE 81] ; c'est un mode navigationnel beaucoup moins aveugle, bâti sur l'indexation de chaque document par 1 à 4 mots-clés thématiques, c'est-à-dire par des notions génériques de niveau supérieur.

Le système commence par proposer une vue globale de la base, à travers les quadruplets ou triplets de mots thématiques les plus fréquents ; à partir de ces points d'entrée éventuels, non imposés, l'utilisateur peut :

. choisir un groupe de mots thématiques, puis explorer les documents les plus proches, sémantiquement, des documents indexés strictement par ce groupe,

. naviguer de mot en mot dans un espace local structuré ; en effet le système dégage autour de chaque mot les "lignes de force" locales qui mènent vers tel mot plutôt que tel autre - à partir de concepts issus de la théorie des graphes (le système détermine des "axes

de pyramides", c'est-à-dire des arêtes [-paires] communes à de nombreux tétraèdres[-quadruplets]) ; le système peut même calculer automatiquement les couples de mots en chaîne correspondant au plus court chemin entre deux mots arbitrairement choisis - par exemple relier "pêche sous-marine" à "maçonnerie" !

Nous voyons deux limites à ce système, par ailleurs le plus proche de nos conceptions : il ne semble pas incorporer la dualité complète document/mot-clé (un document désigné n'appelle pas directement, à notre connaissance, les documents proches) ; de par sa nature combinatoire, il requiert d'être alimenté par des notions génériques en nombre relativement limité, et non par des descripteurs spécifiques de plus bas niveau : alors qu'il fonctionne sur le principe d'un "réseau sémantique dégagé *a posteriori*", il exige la constitution d'un "réseau sémantique *a priori*", avec les limites qu'on connaît à tout thésaurus.

3 Analyse des données et réseaux neuronaux

Les algorithmes d'analyse des données dits "à lecture directe", en particulier la méthode d'approximation stochastique en analyse factorielle des correspondances [BENZECRI 69, LEBART 76] et la méthode de classification automatique "k-means" [MAC QUEEN 67] présentent de fortes analogies avec le fonctionnement de certains réseaux de cellules localement connectées à auto-apprentissage, appelés "réseaux neuronaux" [KOHONEN 84]. Nous terminerons en exposant nos propres travaux.

3.1 Facteurs et neurones : L'analyse factorielle consiste à réduire l'espace de représentation d'un ensemble d'observations : on cherche à passer "au mieux" une description de chaque objet par de nombreux descripteurs à une description par quelques valeurs factorielles, c'est-à-dire réduire le nombre de dimensions de l'espace de représentation.

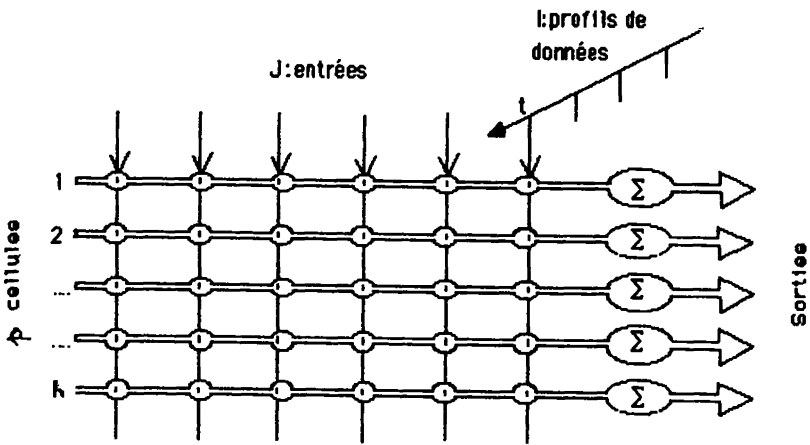
En analyse factorielle des correspondances cet espace permet de représenter simultanément les lignes d'un tableau de données et les colonnes ; chaque ligne est représentée par un point, ainsi que chaque colonne ; la h -ième valeur factorielle $F(i;h)$ du i -ème "profil" de données (par exemple une ligne d'un tableau de nombres contenant les valeurs de variables descriptives d'un "individu" particulier) est donnée, que le profil fasse partie des données analysées ou qu'il soit placé en "élément supplémentaire", par la classique formule de transition entre les valeurs factorielles des lignes et celles des colonnes :

$$\begin{aligned} F(i;h) &= (\lambda h)^{-1/2} \sum \{ (k(i,j)/k(i)) G(j;h) \mid j \in J \} \\ &= \sum \{ (k(i,j)/k(i)) \varphi(j;h) \mid j \in J \} ; \end{aligned}$$

(où le rang du facteur considéré est noté h ; et où $F(.,h)$, $G(.,h)$, $\varphi(.,h)$ désignent respectivement les facteurs usuels sur I et J et le facteur normalisé sur J).

Ce qui invite à considérer les p premiers facteurs comme les réponses de p "neurones" de définition simple (à réponse linéaire ; sans seuil) tous soumis aux lignes successives d'entrées $(k(i,j)/k(i))$ (i.e. à des profils relatifs en ligne) de "poids synaptique" $\varphi(j;h)$, débitant chacun sa sortie $F(.,h)$ constituée par la somme des entrées pondérées par les poids synaptiques (cf. figure).

Pour que ce modèle réalise véritablement une analyse factorielle, il faut que coexistent une dynamique rapide - (la sortie sous l'effet des entrées) - et une dynamique lente : le processus de construction des poids synaptiques convergeant vers les facteurs. Or un tel processus est précisément réalisé dans l'analyse factorielle par approximation stochastique ; méthode que nous rappelons brièvement (cf. J.-P. Benzécri ; in C.A.D. Vol VII n° 4 pp. 387-394 ; 1982).



Partant de valeurs quelconques des poids $\varphi(j,h)$, on remet ceux-ci à jour à chaque entrée d'un individu suivant la formule :

$$\varphi_{i+1}(j;h) = \varphi_i(j;h) + (K_i(j)/C_i(j)) \sum \{(K_i(j')/K_i) \varphi_i(j';h) | j' \in J\} ;$$

dans cette formule l'indice i désigne le numéro de l'individu qui vient d'être introduit ; numéro qui est aussi l'ordre de l'itération ; ou encore "le temps". Le profil de cet individu est noté $\{K_i(j)/K_i\}$ plutôt que $\{k(i,j)/k(i)\}$; afin de souligner le rôle particulier du temps. Et $C_i(j)$ est le total des composantes $K_{i_1}(j)$ des individus introduits jusqu'à présent.

Du point de vue, qui est ici le nôtre, de l'analogie avec les réseaux, il importe de noter que la somme Σ n'est autre que la réponse $F(i;h)$ (cf. *supra*) : on réécrit donc :

$$\varphi_{i+1}(j;h) = \varphi_i(j;h) + ((K_i(j) F(i;h))/C_i(j)) ;$$

la correction apportée à φ rentre ainsi dans le format général :

$$\text{correction} = \text{entrée} \times \text{sortie} \times \text{coefficient} ;$$

on reconnaît la *loi d'apprentissage de Hebb*, bien connue des psychologues et neurophysiologistes, selon laquelle le renforcement d'un lien se fait en proportion à la fois de l'entrée et de la sortie. Cette loi de renforcement des "poids synaptiques" (c'est-à-dire, pour nous des facteurs $\varphi(j;h)$) est utilisée dans de nombreux réseaux de neurones à apprentissage [PALM 85]. Le coefficient de proportionnalité $(1/C_i(j))$, nécessaire pour assurer la convergence de l'algorithme tend à décroître à peu près comme l'inverse du temps (si les individus i se présentent uniformément au cours du temps...).

L'analogie entre approximation stochastique et réseaux de neurones n'est toutefois pas parfaite. Tel qu'on l'a présenté ici, l'algorithme ne peut fournir des facteurs normalisés deux à deux non corrélés : il faut de plus (e.g. toutes les 20 ou 100 introductions déclencher un processus d'orthonormalisation des vecteurs "poids synaptiques" par la méthode de Gram-Schmidt? La question se pose donc, d'introduire dans la structure du réseau des couplages dont les effets soient équivalents à ceux de l'orthonormalisation.

3.2 Apprentissage sur un modèle neuronal et normalisation : Kohonen propose un modèle de neurones (parmi d'autres) où la sortie η_k à l'étape k est définie par le produit scalaire du vecteur entrée x_k et du vecteur poids synaptique m_k :

$$\eta_k = \langle m_k, x_k \rangle ;$$

l'apprentissage des poids synaptiques prenant la forme :

$$m_{k+1} = m_k + \alpha_k \eta_k x_k - \beta_k m_k$$

formules où les minuscules grecques désignent des scalaires ; et les minuscules latines des vecteurs).

Kohonen discute cette loi d'apprentissage : dans le cas général, la norme m tend à croître indéfiniment, mais la direction de m converge vers celle du premier vecteur propre de la matrice de covariance des données x . Ceci vaut en particulier si $\beta = 0$, cas où l'on retrouve la loi de Hebb :

$$m_{k+1} = m_k + \alpha_k \eta_k x_k ;$$

et l'algorithme d'approximation stochastique appliqué à l'analyse en composantes principales. Dans la mesure où l'on n'extrait qu'un seul facteur, le problème d'orthogonaliser ne se pose pas ; en revanche la normalisation peut être obtenue dans le cadre usuel des modèles neuro-naux.

Nous avons construit à partir de l'algorithme de la puissance itérée, l'algorithme stochastique :

$$m_{k+1} = m_k + \alpha_k \eta_k (x_k - \eta_k m_k / \|m\|^2) ;$$

où m est fixée *a priori* comme valeur à atteindre après convergence et :

$$\alpha_k = \alpha_0 \|m\|^2 / (\|m\|^2 + \alpha_0 \sum \{\eta_i^2 \mid i = 1, \dots, k\}).$$

Cette loi d'apprentissage est comprise dans une autre formule de Kohonen :

$$\Delta m = \alpha \eta x - \beta \eta^2 m ; \text{ (en omettant l'indice } k \text{) ;}$$

Kohonen démontre la convergence lorsque α et β sont constants (indépendants de k) ; en particulier, la norme de m tend vers $(\alpha/\beta)^{1/2}$; résultat qui s'accorde avec notre formule ci-dessus.

De nos essais de simulations nous avons retenu que :

- une meilleure convergence est atteinte avec $m_0 = x_1 / \|x_1\|$, qui stabilise très vite $\|m\|$ à la valeur 1 (cf. simulations) ; cette dernière procédure a l'avantage de supprimer l'"inertie" liée à la valeur initiale des poids synaptiques, qui correspondrait à un hypothétique apprentissage passé.

- qu'il existe une "plage optimale" pour α_0 , entre la valeur 0 (pas d'apprentissage) et l'infini (apprentissage chaotique, mais néanmoins convergent, puisque α_0 a une limite finie : $1/k\|\eta_{\text{moyen}}\|^2$) ;

3.3 *Orthogonalisation et inhibition latérale* : Tel quel l'algorithme de Schmidt et Hilbert ne peut se réaliser en faisant appel à des propriétés purement "locales" (dans l'espace et le temps) de chaque "cellule" factorielle. On vient de voir que chaque facteur s'autonormalisait moyennant l'adjonction d'un "terme d'oubli" (de la forme $-\beta \eta^2$). Reste l'orthogonalisation.

Or Kohonen signale (sans le démontrer) que des liens d'"inhibition latérale" entre cellules à auto-apprentissage (feed-back négatif de la sortie d'un neurone sur l'entrée d'un neurone voisin) tendaient à orthogonaliser leurs poids synaptiques, c'est-à-dire à leur faire "reconnaître" des profils d'entrée orthogonaux. Il développe également une méthode simplifiée de simulation de ces inhibitions latérales : pour chaque nouveau profil de données, le neurone qui réagit le plus "autorise" ou non ses voisins à apprendre.

Compte tenu de ce que l'algorithme de Schmidt et Hilbert est une procédure hiérarchisée (en ce sens que le 1-er facteur n'est pas modifié ; le 2-ème l'est seulement en combinaison avec le 1-er ; le 3-ème par rapport aux deux premiers ; etc.) ; nous avons utilisé des "liens d'inhibition latérale" en donnant à notre réseau une structure hiérarchisée : la cellule n° 1 apprend toujours (i.e. est toujours modifiée par la présentation d'une entrée) ; la c. n° 2 n'apprend que si elle-même ou l'une des cellules suivantes a une réponse supérieure à celle de la c. n° 1 ; ... la c. n° h n'apprend que si elle-même ou l'une des cellules suivantes réalise la réponse maxima.

La loi d'apprentissage utilisée est celle décrite par notre formule du § 3.2, avec α_0 paramétrable ; le "temps" propre à chaque neurone, à l'oeuvre dans le coefficient α , est purement local : il ne s'incrémente que lorsque le neurone apprend. En d'autres termes, la somme des η_i^2 n'est effectuée que sur le sous-ensemble des données pour lequel le neurone "apprend".

Nos simulations sur des données réelles sont en premier lieu encourageantes : même si l'orthogonalisation n'est pas parfaite et que le comportement de l'algorithme est sensible à la structure des données et à leur codage, la convergence est rapide ; et il se dégage une structure interprétable au fur et à mesure de la présentation des données.

On peut expérimenter d'autres structures d'interconnexion entre cellules : si chaque cellule "excite" ses proches voisines, l'ensemble d'un réseau de cellules nombreuses disposées géographiquement dans l'espace réalise une "carte" globale des données apprises ; cette carte peut être une projection hautement non linéaire de la structure multidimensionnelle des données sur une portion de surface ; les simulations de Kohonen font apparaître par exemple des circonvolutions quand on "force" une structure de données tridimensionnelle sur une carte à deux dimensions, ou une figure bidimensionnelle sur un réseau à une seule dimension (remplissage de Péano). Mais il faut reconnaître les limites de l'approximation décrite ci-dessus relativement à d'autres types d'inhibition entre cellules .

On trouvera dans [HERAULT 85] des résultats de simulations du comportement de petits réseaux constitués de deux ou trois cellules en inhibition latérale réciproque (et non pas en structure d'interaction hiérarchisée comme les nôtres), suivant un autre modèle d'inhibition latérale par interconnexion physique ; de tels réseaux sont capables de démêler deux signaux mélangés ; du fait de leur structure non hiérarchisée, les axes qu'ils extraient ne sont pas ceux de plus grande variance autour du centre de gravité, comme en analyse factorielle,

mais ceux de covariance minimale, ce qui leur permet par exemple de reconstituer un plan "3 vues" d'une table à partir de deux vues perspectives ...

3.4 Algorithmes de classification adaptative et réseaux neuronaux :

Dans le cas où la règle d'interaction décrite par Kohonen s'applique à un réseau de neurones isolés (c'est-à-dire où seul le neurone ayant répondu le plus fort "apprend", quand un nouveau profil de données se présente) on peut montrer que l'algorithme en oeuvre est une variante d'un algorithme de classification adaptative dit "k-means" [MAC QUEEN 67] procédant par agrégation autour de centres variables, recalculés à chaque nouvel élément ; suivant le schéma

0 - Initialisation : on prend comme centre d'agrégation des k classes souhaitées les k premiers éléments qui se présentent,

1 - Construction des centres de gravité des classes : chaque nouvel élément est affecté au centre dont il est le plus proche (critère de la distance au centre de gravité ; et on calcule le nouveau centre de la classe à laquelle il a été ainsi affecté.

2 - Stabilisation : au temps t , i.e. à l'entrée de l'élément x(i) de rang t , seul un centre est modifié, qu'on peut noter c(i) ; pour ce centre on a :

$$c(i)_{t+1} = c(i)_t + ((x(i) - c(i))_t / \text{card}(c(i))_t) ;$$

où $\text{card}(c(i))_t$ est le nombre des éléments de la cl. c(i), déjà affectés avant i ; comme ce nombre croît linéairement avec le temps (du moins, en moyenne) le coefficient $1/\text{card}$ décroît rapidement, et on peut s'attendre à une stabilisation.

3 - Arrêt : On peut concevoir une source indéfinie d'éléments nouveaux distribués suivant une certaine loi de probabilité : en ce cas il n'y a pas arrêt mais convergence asymptotique ; pratiquement, on fait revenir cycliquement le même jeu d'éléments jusqu'à ce qu'il n'y ait plus de modification d'affectation à un passage ... (ou suivant un autre critère ... : cf. [AGUILAR-MARTIN 81]).

On peut comparer à cet algorithme, le schéma suivant de l'auto-apprentissage d'un ensemble de k neurones isolés.

0 . Initialisation des vecteurs de poids synaptiques m, des neurones, à des valeurs quelconques par exemple égales aux profils des k premiers ensembles.

1 . Construction itérative des poids (vecteurs) des neurones : pour chaque nouvel élément x, on choisit le neurone qui répond le plus fort suivant le critère de maximisation du produit scalaire $\langle m, x \rangle$. On notera que ce critère peut être rendu vraisemblable "physiologiquement" si on admet que plus $\langle m, x \rangle$ est fort, plus le neurone m "se charge" rapidement ; et que se "décharge" le premier neurone qui a dépassé un certain seuil). Le poids de ce neurone est modifié suivant l'équation :

$$m := m + \alpha(t) (x - m).$$

2 . Stabilisation : elle résulte de la décroissance imposée, d'une manière ou d'une autre à la fonction $\alpha(t)$ de chaque neurone.

3 . Arrêt : par exemple quand se termine la présentation d'une longue série d'apprentissage ...

Le parallélisme ainsi apparu entre "k-means" et méthodes de cette famille suggère plusieurs remarques. Si les neurones, au lieu d'être isolés sont répartis dans un espace à une ou deux dimensions grâce à des interactions locales entre neurones voisins - cas que T. Kohonen nomme "associative mapping" - il serait intéressant de tester si l'inconvénient principal des méthodes de classification adaptatives, à savoir leur sensibilité aux conditions initiales, tend à disparaître ; sur ce point Kohonen, tirant la leçon de ses simulations, recommande d'initialiser la structure des interactions dans le réseau avec un large rayon d'interaction entre cellules, et de rétrécir ce rayon avec le temps.

Par ailleurs le calcul de produits scalaires est plus rapide que le calcul de distances euclidiennes.

Tous ces éléments montrent que l'approche "neuronale" ne doit pas manquer d'ouvrir des perspectives intéressantes aux méthodes de classification automatique, secteur commun à l'analyse des données et à la reconnaissance des formes.

En sens inverse, les procédés adaptatifs de création et suppression de classes développés par beaucoup d'auteurs d'algorithmes de classification et de partition à lecture directe [BALL et HALL 73, AGUILAR-MARTIN 81], généralement à partir de seuils sur les critères d'appartenance aux différentes classes, devraient enrichir les modèles de réseaux neuronaux (auto-adaptation de la structure d'interaction et du nombre de neurones aux données).

Par ailleurs certaines études sur des réseaux de modèles de neurones "non linéaires", par exemple [PERSONNAZ 85] [PERETTO 84], ne sont pas sans rapport avec de tels types de systèmes dynamiques : leur théorie fait appel à des notions d'attracteurs, de transitions entre états d'équilibre, de trajectoire dans l'espace des paramètres de contrôle, toutes notions familières aux physiciens.

Enfin, à la rencontre de l'analyse factorielle et de la classification automatique on peut évoquer ici les "rotations" (ex : méthodes VARIMAX et OBLIMAX : cf. [TORRENS-IBERN] 72) ; et l'analyse factorielle typologique [DIDAY 80].

3.5 Nos développements logiciels en matière de banque d'images :

Le moment est venu de décrire brièvement nos propres conceptions : point n'est besoin de s'étendre puisque les pages qui précèdent ne sont qu'un long justificatif aux options que nous avons prises !

Option 1 : extraction d'un environnement local large autour d'un point d'ancrage constitué par un élément quelconque - image ou mot-clé - mis en exergue dans une "zone de référence" ; cet environnement est constitué par :

- une zone de proximité directe, évidente : pour une image ce sont les mots qui la décrivent ; pour un mot ce sont les images qu'il indexe,

- une zone de proximité indirecte, constituée par les éléments de même nature que celui en exergue (les mots proches d'un mot donné, ou les photos proches d'une photo donnée), organisés en couches concentriques de proximité décroissante, selon leur nombre de cooccurrences d'éléments communs.

Option 2 : constitution d'un environnement personnel pour chaque usager ; celui-ci peut stocker dans sa "besace" tous les éléments - image ou mots-clés - qui lui conviennent. Il peut ainsi se constituer

des piles, des aide-mémoire, qu'il conserve sur support magnétique personnel jusqu'à la prochaine cession de travail.

A noter que les fenêtres de visualisation offertes ont toutes la même structure et le même mode d'utilisation, qu'elles soient remplies par des mots ou par des images, ou qu'elles permettent ou non, en fonction du matériel employé, de voir plusieurs images simultanément.

La possibilité qu'on offre ainsi à l'utilisateur d'indexer "transversalement" des milliers d'images selon un paquet de mots-clés, et non plus seulement image par image, permet à la fois un gain de productivité considérable, mais aussi une homogénéité beaucoup plus grande, dans son travail d'analyse et d'indexation documentaire.

Chaque auteur de mots-clés signe ses mots, les explicite par des commentaires, et est seul autorisé à en modifier la forme ou le fond (c'est-à-dire la "liste inverse" des images qui lui sont accrochées).

Les "simples promeneurs", qui n'ont pas l'autorisation d'archiver leurs piles d'images ou de mots, peuvent cependant conserver sur disquette leurs piles personnelles, et créer en temps réel des montages d'images fixes incrustées de textes sur vidéocassette.

Option 3 : représentation des grandes tendances globales et des tendances locales fines discernables à travers l'indexation de la base : c'est là que peut prendre place un "tableau de bord neuronal" quelque peu futuriste qui reste à tester et à valider, contrairement à la plupart des options précédentes - à l'exception de la zone de proximité indirecte - déjà testées sur notre maquette ICONOTRON [LELU 84].

4 Conclusion pour la représentation d'espaces documentaires

L'approche "réseaux neuronaux" nous permet de saisir la parenté profonde entre analyses factorielles et classifications automatiques : une analyse factorielle peut être vue à la limite, comme une classification (à classes recouvrantes) avec un choix particulier du critère de distorsion utilisé, et sous contraintes imposées sur les positions relatives des axes traversant les classes, c'est-à-dire comme une classification avec classes structurées. De multiples types de représentations, hybrides de classification et d'analyse factorielle, sont possibles, selon les critères de distorsion et les structures imposées aux classes.

Nous pressentons que la représentation d'espaces documentaires très fortement multidimensionnels, mais également très "effilochés", "fibreuse" (c'est-à-dire à faible nombre de dimensions locales), puisse se faire de la manière suivante : une carte globale constituée de nombreux neurones spatialement interconnectés et "allumés" en très faible nombre par chaque profil de données ; cette carte serait une projection *a priori* non linéaire décrivant de façon optimale les propriétés des données ; ceci complété par une représentation plus "locale" de type carte factorielle, dans les axes - non nécessairement orthogonaux - définis par les quelques neurones "allumés".

L'utilisateur pourrait alors se situer et se déplacer au sein de la carte globale tout comme des cartes locales, aidé en cela par les possibilités graphiques et les capacités "temps réel" toujours plus poussées des mini et micro-ordinateurs actuels ; l'expérience passée des applications de l'analyse des données, en particulier celles que nous avons examinées plus haut dans le domaine documentaire, permet d'affirmer sans grand risque d'être démenti que les diverses dimensions, globales et locales, de ces cartes présenteront une signification riche et humainement compréhensible - pour peu que le processus d'indexation documentaire ait été réalisé de façon un tant soit peu homogène et contrôlée.

L'aboutissement à court terme de notre recherche sera une "boussole", ou plutôt un "sextant", un outil de repérage plaqué de façon externe sur une base documentaire relativement classique ; à moyen terme, des éléments de syntaxe simples pourraient être pris en compte ; à très long terme, une intégration de tels concepts dans le matériel pourrait remplacer le stockage exhaustif de l'information sur disque (mémoire répartie et active, au lieu de la mémoire passive et centralisée actuelle) ... et résoudre les problèmes de temps de calcul inhérents à la simulation de grands réseaux de cellules sur des architectures informatiques classiques (il faut en effet mettre à jour un tableau de J mots-clés X N cellules à chaque introduction d'une nouvelle unité documentaire).

Par ailleurs ces approches permettent de prendre en compte un problème fondamental des systèmes documentaires et de l'analyse des données : celui de la "granularité" de l'information, c'est-à-dire du calibrage, de la finesse descriptive des mots-clés ; en effet l'expérience des documentalistes et des utilisateurs de bases documentaires enseigne qu'il existe une plage intermédiaire entre d'une part la granularité grossière des grands concepts généraux, qui s'appliquent à la quasi-totalité du corpus, et donc différencient peu les unités documentaires, et d'autre part les mots-clés caractéristiques de très peu de ces unités, presque isolés du restant du lexique, et donc n'informant que très peu sur ces unités (on trouvera dans [SALTON 76] une discussion de cet optimum en termes de théorie de l'information) ; l'expérience des dépouillements statistiques d'enquête enseigne également qu'on ne peut pas se passer de "cadrer" soigneusement le "grain" des catégories utilisées pour l'observation - à coup de pré-enquêtes et d'indispensables programmes de recodification de variables - pour obtenir des résultats significatifs et des analyses de données parlantes. Il n'est pas étonnant que les meilleures méthodes d'analyse, de ce point de vue, soient celles qui présentent la plus grande stabilité sous l'effet de ces regroupements de catégories, grâce à leur propriété d'"équivalence distributionnelle" (fusionner deux éléments dont les profils relatifs sont identiques ne change pas l'analyse, comme en analyse factorielle des correspondances), bien sûr dans des limites raisonnables dont l'appréciation constitue aujourd'hui une bonne part de l'art et de l'expérience du statisticien ...

En conclusion, notre souhait le plus cher serait de réussir l'intégration en un tout cohérent des concepts documentaires nouveaux - ou des concepts anciens, nouveaux dans le domaine documentaire ! - que nous avons exposés ici.

BIBLIOGRAPHIE :**AGUILAR-MARTIN 81 :**

J. Aguilar-Martin, M. Balssa, R. Lopez de Mantaras - "Estimation réursive d'une partition ; Exemples d'apprentissage et auto-apprentissage dans R^n et I^n " - Classification automatique et perception par ordinateur - Publication de l'INRIA Rocquencourt - 1981

AIT HAMLAT 84 :

A. Aït Hamlat - "Analyse des répétitions et indexation automatique des documents" - dans "Les Cahiers de l'Analyse des Données" - Vol. IX, N°2, pp. 173-204 - Dunod - Paris - 1984

ANDREEWSKI 73 :

A. Andreevski, C. Fluhr - "Apprentissage - Analyse automatique du langage - Application à la documentation" Documents de linguistique quantitative N° 21 - Dunod - Paris - 1973

BALL, HALL 73 :

Ball G.H., Hall D.J. in Anderberg M.R. - "Cluster Analysis for Applications" - Academic Press - New York - 1973

BASSANO 81 :

J.C. Bassano, A. Herzallah - "Composante linguistique pour un système du type documentation assistée" dans "Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle" - 1981

BENZECRI 69 :

J.P. Benzécri - "Approximation stochastique dans une algèbre non normée commutative" - Bull. Soc. Math. France N° 97, pp. 225-241 - Paris - 1969 ; cf. également "Les Cahiers de l'Analyse des Données" - Vol. VII, N°4, pp. 387-394 - Dunod - Paris - 1982 (et pp. 395-412 pour un développement philosophique autour de Bergson)

BENZECRI 73 :

J.P. Benzécri et coll. - "L'Analyse des Données" (2 tomes) - Dunod - Paris - 1973

BENZECRI 81 :

J.P. Benzécri et coll. - "Pratique de l'Analyse des Données : Linguistique et Lexicologie" - Dunod - Paris - 1981

CALLON 85 :

M. Callon - "L'établissement de cartes stratégiques de la recherche : l'utilisation des bases de données pour l'aide à la décision technologique et scientifique" - Actes d'INFODIAL-VIDEOTEX 1985, pp. 179-183 - Paris

CARLIER 85 :

A. Carlier - "Analyse des évolutions sur table de contingence : quelques aspects opérationnels" - Actes des 4ièmes Journées Internationales Analyse des Données et Informatique" - INRIA - Rocquencourt - 1985

CIBOIS 80 :

Ph. Cibois - "La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique" - Thèse de Doctorat de 3ième Cycle - LISH - Paris - 1980

DEWEZE 81 :

"Les tableaux sémantiques : essai de modélisation, application à l'indexation et à la recherche documentaire" - Thèse de Doctorat de 3ième Cycle en Science mathématique - Université Claude Bernard - Lyon - 1981

GIRARD 85 :

A. Girard - "Utilisation des stratégies d'interrogation basées sur les co-occurrences de descripteurs" - Actes d'INFODIAL-VIDEOTEX 1985, pp. 22-25 - Paris

GROSS 76 :

M. Gross - "Méthodes empiriques en syntaxe" - dans "Structure et dynamique des systèmes", pp. 149-163 - Maloine - Paris - 1976

HERAULT 85 :

B. Ans, J. Héroult, C. Jutten - "Architectures neuromimétiques adaptatives : détection de primitives" - Actes de COGNITIVA, pp. 593-597 - AFCET/CESTA - Paris - 1985

KLOSSA 83 :

J. Klossa, B. Retailiau, F. Roux "Multimédia et dialogue informatique" - Actes d'INFODIAL-VIDEOTEX 1983, pp. 220-223 - Paris

KOHONEN 84 :

T. Kohonen - "Self-organization and associative memory" - Springer-Verlag - Berlin - 1984

LEBART 69 :

L. Lebart - "Analyse statistique de la contigüité" - Publication de l'ISUP, vol.XVIII, pp. 81-112 - Paris - 1969

LEBART 76 :

L. Lebart - "Sur les calculs impliqués par la description de certains grands tableaux" - Annales de l'INSEE N° 22 23, pp. 255-271 - Paris - 1976

LELU 84 :

A. Lelu - "La machine à classer les images" - Science et Vie hors-série N°149 "La photo et les images synthétiques" - Décembre 1984

LELU 85 :

A. Lelu - "De l'espace urbain à l'espace électronique" - Bulletin de l'IDATE N°20 - Montpellier - 1985

LELU 86 :

A. Lelu - "After the inventors came the editors - a new medium : the interactive screen" - in "Can Information Technology result in Benevolent Bureaucracies ?" - R. Laufer, editor - North Holland - à paraître en 1986

LENOIR 82 :

P. Lenoir - "Vidéotex et banque de données médicales chez le praticien" - Actes de la Convention Informatique 1982, pp. 190-195 - Paris

LERMAN 85 :

I.C. Lerman, Ph. Peter - "Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle" - Actes des 4ièmes Journées Internationales Analyse des Données et Informatique" - INRIA - Rocquencourt - 1985

MAC QUEEN 67 :

J. Mac Queen - "Some methods for Classification and Analysis of Multivariate Observations" - Proc. 5th Berkeley Symp., pp.281-297 - 1965

PALM 82 :

G. Palm - "Neural assemblies, an alternative approach to artificial intelligence" - Springer-Verlag Berlin - 1982

PERSONNAZ 85 :

L. Personnaz, I. Guyon, G. Dreyfus - "Information storage and retrieval in spin-glass like neural networks" - Journal de Physique, Lettres, N°46, pp. L359-L365 - Paris - 1985

PERETTO 84 :

P. Peretto - "Collective Properties of Neural Networks : a Statistical Physics Approach" - Biological Cybernetics, vol.50, pp.51-62 - Springer-Verlag - Berlin - 1984

PINSON 85 :

G. Pinson, A. Demailly, D. Favre - "La Pensée - approche holographique" - Presses Universitaires de Lyon - 1985

PRADE 84 :

H. Prade, C. Testemalle - "Traitement des questions vagues dans une base de données imprécises" - L'informatique Professionnelle, N°27 et 28 - 1984

SALTON 76 :

G. Salton, A. Wong - "Automatic indexing using term discrimination and term precision measurement" - Information Processing and Management, Vol. 12, pp. 43-51 - 1976

SANSON 78 :

P. Sanson - "Méthodologie informatique/statistique et traitements relatifs à la structuration et au classement de corpus photographiques" - Thèse de 3ième Cycle - EHESS - 1978

SIMON 84 :

J.C. Simon - "La reconnaissance des formes par algorithmes" - Masson - Paris - 1984

TORRENS-IBERN 72 :

J. Torrens-Ibern - "Modèles et méthodes de l'analyse factorielle" - Dunod - Paris - 1972