Indexation et classement en bureautique

Jean Hameon

THESE

présentée à

l'Institut National Polytechnique de Grenoble

pour obtenir le grade de
DOCTEUR DE 3ème CYCLE
« Génie Informatique »

par

HAMEON Jean

INDEXATION ET CLASSEMENT EN BUREAUTIQUE

Thèse soutenue le 30 Janvier 1981 devant la commission d'examen

G. VEILLON	Président
J.C. CHUPIN C. DELOBEL N. NAFFAH	Examinateurs

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Année universitaire 1979-1980

Président : M. Philippe TRAYNARD

Vice-Présidents : M. Georges LESPINARD
M. René PAUTHENET

PROFESSEURS DES UNIVERSITES

MM.	ANCEAU François	Informatique fondamentale et appliquée
	BENOIT Jean	Radioéléctricité
	BESSON Jean	Chimie Minérale
	BLIMAN Samuel	Electronique
	BLOCH Daniel	Physique du Solide - Cristallographie
	BOIS Philippe	Mécanique
	BONNETAIN Lucien	Génie Chimique
	BONNIER Etienne	Métallurgie
	BOUVARD Maurice	Génie Mécanique
	BRISSONNEAU Pierre	Physique des Matériaux
	BUYLE-BODIN Maurice	Electronique
	CHARTIER Germain	Electronique
	CHERADAME Hervé	Chimie Physique Macromoléculaires
Mme	CHERUY Arlette	Automatique
MM.	CHIAVERINA Jean	Biologie, Biochimie, Agronomie
	COHEN Joseph	Electronique
	COUMES André	Electronique
	DURAND Francis	Métallurgie
	DURAND Jean-Louis	Physique Nucléaire et Corpusculaire
	FELICI Noël	Electrotechnique
	FOULARD Claude	Automatique
	GUYOT Pierre	Métallurgie Physique
	IVANES Marcel	Electrotechnique
	JOUBERT Jean-Claude	Physique du Solide - Cristallographie
	LACOUME Jean-Louis	Géographie - Traitement du Signal
	LANCIA Roland	Electronique - Automatique
	LESIEUR Marcel	Mécanique
	LESPINARD Georges	Mécanique
	LONGEQUEUE Jean-Pierre	Physique Nucléaire Corpusculaire
	MOREAU René	Mécanique
	MORET Roger	Physique Nucléaire Corpusculaire
	PARIAUD Jean-Charles	Chimie - Physique
	PAUTHENET René	Physique du Solide - Cristallographie
	PERRET René	Automatique

.../...

MM.	PERRET Robert	Electrotechnique
	PIAU Jean-Michel	Mécanique
	PIERRARD Jean-Marie	Mécanique
	POLOUJADOFF Michel	Electrotechnique
	POUPOT Christian	Electronique - Automatique
	RAMEAU Jean-Jacques	Chimie
	ROBERT André	Chimie Appliquée et des matériaux
	ROBERT François	Analyse numérique
	SABONNADIERE Jean-Claude	Electrotechnique
Mme	SAUCIER Gabrielle	Informatique fondamentale et appliquée
M.	SOHM Jean-Claude	Chimie - Physique
Mme	SCHLENKER Claire	Physique du Solide - Cristallographie
MM.	TRAYNARD Philippe	Chimie - Physique
	VEILLON Gérard	Informatique fondamentale et appliquée
	ZADWORNY François	Electronique

CHERCHEURS DU C.N.R.S. (Directeur et Maître de Recherche)

M.	FRUCHART Robert	Directeur de Recherche
MM.	ANSARA Ibrahïm	Maître de Recherche
	BRONOEL Guy	Maître de Recherche
	CARRE René	Maître de Recherche
	DAVID René	Maître de Recherche
	DRIOLE Jean	Maître de Recherche
	KAMARINOS Georges	Maître de Recherche
	KLEITZ Michel	Maître de Recherche
	LANDAU Ioan-Doré	Maître de Recherche
	MERMET Jean	Maître de Recherche
	MUNIER Jacques	Maître de Recherche

Personnalités habilitées à diriger des travaux de recherche (décision du Conseil Scientifique)

E.N.S.E.E.G.

MM.	ALLIBERT Michel	
	BERNARD Claude	
	CAILLET Marcel	
Mme	CHATILLON Catherine	
MM.	COULON Michel	
	HAMMOU Abdelkader	
	JOUD Jean-Charles	
	RAVAINE Denis	
	SAINFORT	C.E.N.G.

.../...

MM. SARRAZIN Pierre
SOUQUET Jean-Louis
TOUZAIN Philippe
URBAIN Georges — Laboratoire des Ultra-Réfractaires ODEILLO

E.N.S.M.E.E.

MM. BISCONDI Michel
BOOS Jean-Yves
GUILHOT Bernard
KOBILANSKI André
LALAUZE René
LANCELOT Françis
LE COZE Jean
LESBATS Pierre
SOUSTELLE Michel
THEVENOT François
THOMAS Gérard
TRAN MINH Canh
DRIVER Julian
RIEU Jean

E.N.S.E.R.G.

MM. BOREL Joseph
CHEHIKIAN Alain
VIKTOROVITCH Pierre

E.N.S.I.E.G.

MM. BORNARD Guy
DESCHIZEAUX Pierre
GLANGEAUD François
JAUSSAUD Pierre
Mme JOURDAIN Geneviève
MM. LEJEUNE Gérard
PERARD Jacques

E.N.S.H.G.

M. DELHAYE Jean-Marc

E.N.S.I.M.A.G.

MM. COURTIN Jacques
LATOMBE Jean-Claude
LUCAS Michel
VERDILLON André

Je tiens à remercier ici :

- Monsieur Gérard VEILLON, professeur à l'INPG, qui a accepté de présider le jury de cette thèse,

- Messieurs Claude DELOBEL, professeur à l'USMG, et Najah NAFFAH, directeur du projet-pilote burotique KAYAK à l'ADI/INRIA, qui ont bien voulu juger ce travail,

- Monsieur Jean-Claude CHUPIN, directeur de l'antenne de Grenoble du Centre de Recherche CII-HB, qui m'a accepté au sein de son équipe, y a toujours facilité mon intégration et a accepté la responsabilité de ce travail.

Je voudrais également remercier Monsieur Louis BOLLIET, professeur à l'IUT B, d'avoir autorisé, il y a quelques années, l'inscription en DEA d'un étudiant attardé et lointain, ainsi que tous ceux qui, à l'IMAG, ont toujours réservé un excellent accueil à un travailleur émigré ...

Edouard ANDRE, ingénieur au Centre de Recherche CII-HB, puis maintenant au CNET Lannion, m'a toujours accordé un soutien amical et confiant. Son enthousiasme et sa rigueur ont été des facteurs déterminants de mon intérêt pour la recherche. Je suis heureux de le remercier ici pour les efforts et le temps qu'il m'a consacrés.

Marie-Laure MARCON a su déchiffrer un manuscrit mal écrit et utiliser avec virtuosité un certain TTX 80, je l'en remercie vivement.

Je citerai enfin ceux sans qui ce document n'aurait probablement jamais vu le jour :

- Brigitte BOGO, qui a toujours assuré efficacement le gîte et le couvert d'un visiteur souvent envahissant,

- Gilles BOGO, pour le travail que nous avons effectué ensemble, et pour les tâches d'intendance souvent peu agréables dont il s'est chargé.

Leur hospitalité et leur disponibilité sont grandement responsables de la concrétisation de ce travail ; qu'ils trouvent ici l'expression de ma reconnaissance.

Jean HAMEON

TABLE DES MATIERES

A Hélène et Jean
A Edith

INDEXATION ET CLASSEMENT EN BUREAUTIQUE

AVANT-PROPOS

Ce travail a débuté comme devant être une étude sur la spécificité de la coopération des bases de données documentaires, par rapport à la coopération des bases de données en général, et plus particulièrement par rapport à la coopération telle qu'elle est définie dans le projet POLYPHEME.

On s'est aperçu assez rapidement que si les structures de données semblaient plus simples lorsqu'elles étaient documentaires, la coopération de tels systèmes devait aller au-delà de la simple "juxtaposition" ou du simple "habillage" de systèmes existants, et que les problèmes posés dépassaient les compétences "traditionnelles" des "gens du logiciel".

Dans le domaine des bases documentaires, la tendance n'étant ni à la décentralisation (cf. "gros" serveur français), ni à la coopération (cf. réseau documentaire européen), ce travail s'est orienté vers la bureautique, comme champ d'expérimentation de nature documentaire.

Ce travail doit beaucoup à un ouvrage de G. SALTON [SAL 75], une des "bibles" de la documentation automatique, et à un ouvrage de C.D. PAICE [PAI 77], approche plus pédagogique de ce domaine.

On trouvera deux parties principales dans ce document :

- Une première, consacrée à un survol ou "état de l'art" de la documentation automatique et de la bureautique.

- Une seconde, où l'on présentera des éléments sur la nature de la documentation automatique en bureautique, et des propositions pour la réalisation de tels systèmes.

INTRODUCTION

QUELQUES DEFINITIONS

Un système d'information a pour mission de fournir des services d'informations à une population d'usagers.

De quelle information parlons-nous ?

G. SALTON nous dit [SAL 75] : "l'informatique est définie comme la science du traitement de l'information. Avec cette définition, un calcul est simplement un procédé particulier de traitement de l'information, implémenté comme une transformation de nombres ou d'autres entités mathématiques. Dans toutes les formes de tâches de traitement de l'information, il en existe une, particulière, pour qui le terme information prend le sens littéral et direct de connaissance acquise ou d'intelligence. Les éléments à traiter ne sont plus réduits à des symboles mathématiques, mais ce sont des mots, des phrases, des images, des livres, des documents...".

La représentation de cette information est le plus souvent la forme écrite, sur des supports de natures différentes (livre, journal, note, ...), que nous désignons par le terme document. Si des documents forment la "matière première" d'un système d'information, cette information devient donnée documentaire, et un tel système système documentaire.

Une seconde approche des systèmes documentaires peut être faite en distinguant deux types de systèmes d'information [SAL 75] :

- le premier type a pour but de répondre à des demandes spécifiques de données, par des réponses spécifiques, contenant (autant que possible) seulement les données demandées. Les systèmes que nous appelons systèmes de gestion de bases de données (SGBD) en font partie.

- Les utilisateurs du second type de système d'information ne sont pas intéressés par des faits précis, mais plutôt par un "rapport sur l'état de l'art" sur un sujet particulier. La sortie de tels systèmes consisterait normalement en une collection de documents ; en fait, le système informe de l'existence ou de la non-existence de documents sur le sujet donné, et la sortie est le plus souvent un ensemble de références à des documents.

Une troisième approche des systèmes documentaires est celle qui est désignée en anglais par "Information retrieval" ou "Automatic information retrieval" dont la traduction littérale est "recouvrement d'informations". Deux définitions complémentaires sont :

- "Le propos de "Information retrieval" est simplement d'obtenir des réponses pertinentes à des questions" [PAI 77].

- " "Information retrieval" est un domaine concerné par la structure, l'analyse, le rangement, la recherche et le recouvrement de l'information" [SAL 68].

Le mot information étant pris comme autonyme de fait ou de donnée, et désignant des documents tels que définis précédemment.

Plus généralement, le terme documentation automatique est souvent utilisé pour définir l'automatisation d'une bibliothèque ou d'un centre de documentation, c'est à dire pour désigner l'ensemble des opérations nécessaires à la gestion.

G. SALTON y distingue, par ordre décroissant de popularité, les opérations suivantes [SAL 75] :

- traitements comptables, gestion des stocks,

- suivi des acquisitions,

- contrôle de la circulation des ouvrages (prêt, inventaire, diffusion,...),

- génération et maintenance de listes de divers types (catalogues, listes d'acquisition, listes par sujet, index, bulletin, dictionnaires,...),

- travail de recherche proprement dit (recherches bibliographiques, recherches sur des sujets donnés, recherches récurrentes du type diffusion sélective de l'information (SDI) où l'usager est averti périodiquement des nouvelles acquisitions sur un sujet donné...),

- analyse de l'information (indexation, catalogage, résumé, classification).

Si beaucoup de centres de documentation ont automatisé leurs opérations comptables, stocks, listes, acquisitions, prêts, beaucoup moins possèdent des services de recherche automatisés et très peu se sont aventurés dans l'automatisation des tâches de classification, catalogage, indexation, production de résumés.

Notre définition de la documentation automatique sera la suivante : ensemble des méthodes et outils permettant de retrouver, dans une collection de documents, ceux répondant à des questions posées par des usagers.

OU ? QUAND ? COMMENT ?

Certains aspects de la documentation automatique existent dans des bibliothèques ou des centres de documentation, c'est à dire se sont développés dans des lieux ou organismes où le nombre de documents était devenu tel qu'une gestion manuelle était devenue difficile, impossible ou inacceptable.

Il semble que ce soit aujourd'hui les seules applications ; pourtant, la "gestion du grand nombre" n'est pas le seul intérêt de ces méthodes que l'on emploie quotidiennement.

Pour une très petite collection, la mémoire humaine suffit. Dès que la collection grandit, ce n'est plus le cas, et il faut employer ce qui est généralement appelé classement.

Prenons l'exemple de la correspondance personnelle d'un individu : il reçoit des lettres, des factures, des relevés de banque, des feuilles de paye, etc... Il est probable qu'il organisera sa collection de documents dans un ensemble de dossiers qui s'appelleront "à répondre", "à payer", "banque" ... Si la masse des documents croît encore, il peut hiérarchiser et la recherche se fera sur un critère : on pourra trouver dans le dossier "banque" une chemise par numéro de compte.

Prenons maintenant l'exemple d'une petite bibliothèque : il se pose le problème du rangement des ouvrages sur les rayonnages ; on peut les ranger par sujet, ou bien par liste alphabétique d'auteurs, ou bien par ordre d'arrivée, on bien par format, etc... Mais, il faut choisir : alors on choisit un seul critère pour le rangement, par exemple les grands thèmes ; à l'intérieur de chaque grand thème on classe par ordre d'arrivée, et on offre des fichiers (des vrais ! avec des fiches !), ou des listes par ordre alphabétique d'auteurs, par date d'arrivée, etc... renvoyant au rangement du document. L'utilisation de n listes permet de faire n classements, c'est à dire de simuler la mise à disposition et l'utilisation de n exemplaires de chaque ouvrage.

On a bien l'impression que pour que la documentation automatique se développe, il faut qu'il y ait une situation d'urgence : urgence d'une gestion correcte, urgence de satisfaction de besoins de documentation, ... et qu'en fait devant de petites collections de documents ou devant des besoins plus modestes, la documentation automatique n'a pas été ressentie comme quelque chose d'utile.

Il faut sans doute nuancer ce jugement. Les méthodes sont applicables à toute collection de documents, le mot document étant pris au sens large. Songeons, par exemple, aux collections de disques, cassettes, photos, films, etc... Si la documentation automatique ne s'est pas plus développée, c'est sans doute pour les raisons suivantes :

- la non-disposition de matériel et/ou le coût élevé du traitement
- le coût de la saisie
- la difficulté de la caractérisation (indexation).

Cette situation risque aujourd'hui de changer, principalement à cause de (ou grâce à) l'apparition d'une informatique individuelle ou quasi-individuelle. Le coût du matériel, le coût du traitement, ont baissé de façon extraordinaire et vont continuer à le faire. On peut espérer que la disponibilité permanente de ces outils pour un individu ou un groupe d'individus, va faire que la saisie, ou plus généralement, toutes les opérations consistant à "entrer des données", ne vont plus être considérées comme des points de passage obligés et coûteux, pour effectuer un traitement, mais comme une chose "naturelle" pour inscrire des faits sur un support, au même titre que l'écriture manuelle.

A partir du moment où des documents existeront sous cette forme, c'est à dire manipulables par une machine, alors des besoins de gestion de ces ensembles de documents seront à satisfaire, par le simple fait de la disparition des manipulations "physiques" ; il faudra alors mettre en oeuvre les méthodes de la documentation automatique.

L'exemple le plus significatif, aujourd'hui, de cette apparition de matériel quasi-individuel, sont les machines dérivées des machines à écrire "électroniques" que l'on appelle machines de "traitement de textes", et qui ne sont pas autre chose que des minis ou micros ordinateurs possèdant cette fonction.

Cette évolution s'inscrit dans le cadre plus large de ce qu'il est convenu d'appeler "bureautique", et c'est dans ce cadre que s'inscrit ce travail.

1. LES METHODES DE LA DOCUMENTATION AUTOMATIQUE

Ce chapître n'a ni pour but, ni la prétention d'exposer la totalité des méthodes employées dans ce que nous avons appelé la documentation automatique. Il présente simplement un bref aperçu de définitions, de méthodes, d'outils, qui nous semblent les plus importants ou les plus caractéristiques de la documentation automatique et des problèmes posés. Nous reviendrons sur certains aspects dans d'autres chapîtres, lors des besoins d'approfondissements.

1.1. CLASSEMENT, CLASSIFICATION, INDEXATION

Ces termes sont souvent utilisés comme synonymes, dans la littérature, et, s'il est vrai que les méthodes employées pour ces opérations se ressemblent, il nous semble important pour l'instant de les distinguer car elles correspondent à des intentions différentes [PAI 77].

Le but principal du classement est de décider "où mettre" un objet. Dans une bibliothèque les documents traitant de sujets proches sont rangés ensemble. Naturellement, la connaissance du schéma de classification employé indiquera à l'usager "où regarder" pour trouver les informations qu'il cherche, mais s'il n'est pas familier avec son organisation, il ne saura pas "où commencer à regarder". Le but initial du classement est d'organiser le rangement de l'information, mais pas de faciliter sa recherche.

Le but de l'indexation est de faciliter le fait de trouver l'information. Essentiellement, cette opération consiste à représenter chaque document par un ensemble d'indicateurs reflétant son contenu, c'est à dire l'ensemble des faits, notions ou idées qu'il renferme. Nous définissons cet ensemble d'indicateurs comme une structure de descripteurs appelée descriptif [ABH 79]. Les descripteurs ont également d'autres appellations comme termes (d'indexation), mots-clés, concepts,... La fonction de cette opération n'est pas de conserver la totalité de l'information contenue dans le document original, mais seulement cette partie de l'information qui permettra de retrouver le document [TRY 71]. Elle n'a pas non plus pour but de condenser l'information (pour des problèmes de gain de place, par exemple), mais de transformer cette information afin de la posséder sous une forme permettant en parti-

culier d'apprécier la similarité ou la proximité d'informations différentes : une opération de recherche est une opération de comparaison entre l'information véhiculée par la question posée et l'information contenue dans chaque document de la collection.

Citons pour mémoire quelques schémas de classification, dont les plus anciens remontent à l'antiquité, et dont le but avoué est de classer correctement toutes les connaissances humaines [TRY 71] :

- l'arbre de PORPHYRE (233-304) qui est une classification binaire, la classification de BACON (1605) et d'A. COMTE (1851-1854).

- La classification décimale de DEWEY, proposée en 1875, qui est devenue la classification décimale universelle (CDU), utilisée aujourd'hui dans de nombreuses bibliothèques.

- Les schémas de classification multiple (dont fait partie la CDU), et les classifications à facettes (depuis 1952).

L'inconvénient majeur de ces systèmes de classification est leur prétention à l'exhaustivité, qui rend leur évolution très difficile. Aussi, au lieu de chercher à classer plus ou moins systématiquement les documents dans un système universel, on s'est orienté vers la caractérisation du contenu du document par l'utilisation de descripteurs appartenant à un vocabulaire plus ou moins spécialisé.

C'est en ce sens que l'on peut dire que le classement est devenu indexation.

Le classement d'un document implique de l'examiner et de le "mettre dans" la ou les classes considérées comme les plus appropriées. Chaque classe est représentée par un nombre, un nom, ou plus généralement un descriptif. Ainsi, en classant un document, on lui associe un ou plusieurs descripteurs, autrement dit un descriptif.

C'est en ce sens que classement et indexation sont souvent utilisés comme synonymes.

1.2. DIFFERENTES METHODES D'INDEXATION

L'indexation d'un document peut se faire manuellement ou automatiquement.

Indexation manuelle :

La méthode "classique" pour réaliser une telle indexation est d'effectuer en partie ou en totalité les opérations suivantes [SAL 75] :

- des descripteurs sont choisis pour représenter l'information contenue dans le document,
- un poids est éventuellement attaché à chaque descripteur, reflétant ainsi son importance présumée,
- chaque descripteur peut être identifié comme un indicateur d'un certain type, ou jouant un certain rôle (action, propriété, objet...),
- des relations entre les descripteurs peuvent être spécifiées telles des relations de synonymie, d'inclusion, etc...

Ce travail est toujours effectué par du personnel spécialisé, qui peut être aidé dans cette tâche par un ensemble de moyens tels que :

- des dictionnaires, ou thésaurus pour identifier tous les descripteurs possibles,
- des notes définissant le sens ou l'interprétation des descripteurs disponibles,
- des indications sur la manière de relier les termes entre eux,
- des informations statistiques reflétant les indexations déjà faites.

Une bonne indexation étant considérée comme essentielle, puisqu'à la base des performances futures du processus de recherche, son coût en personnel est élevé : les indexeurs doivent non seulement posséder une connaissance poussée du vocabulaire d'indexation, mais également des caractéristiques de la collection de documents et des caractéristiques des questions : une indexation effective doit refléter le type de questions qui seront posées.

Si, en principe, l'indexation manuelle peut produire des résultats d'excellente qualité, la pratique est souvent bien différente ; pour un même document, pour un même vocabulaire d'indexation, des indexations réalisées par des indexeurs différents, ou par un même indexeur à des instants différents, ont produit des résultats très différents : des expériences ont montré qu'on ne rencontre parfois pas plus de 60 % de descripteurs communs [TRY 71].

Indexation automatique :

Une méthode "classique" consiste à extraire d'un texte qui est soit le texte original du document, soit un résumé, tous les mots y apparaissant, et, pour chacun d'eux, effectuer certaines mesures basées sur leur position dans le texte ou sur leur fréquence d'apparition dans le document et dans la collection. Des mots sont éliminés, et des poids sont éventuellement assignés aux mots restants, sur la base des mesures précédentes.

Des améliorations peuvent être faites en identifiant des groupes de mots comme descripteurs, en ramenant ces mots à des formes normales ou réduites, en ajoutant des mots identifiés comme reliés, soit par des méthodes statistiques sur le vocabulaire de la collection, soit par consultation de dictionnaires, listes de synonymes ou thésaurus.

Toutes ces méthodes sont forcément dérivatives, au sens où les descripteurs sont extraits (dérivés) du texte fourni. Les critiques adressées à ces méthodes sont les suivantes [STE 71] [SAL 75] :

- les descripteurs ne sont pas indépendants des modes particuliers d'expression des langages dans lesquels sont écrits les documents,

- le langage des auteurs varie suivant la période et l'environnement,

- les méthodes statistiques, basées sur des techniques de comptage de mots ne sont pas adaptées à l'analyse de textes.

G. SALTON répond en constatant que la justification d'une technique d'indexation se trouve dans les résultats obtenus, par comparaison avec les méthodes manuelles. Nous ajouterons, pour les deux premiers points, que l'emploi de méthodes de traduction (ou indexation par assignation dont nous reparlerons) limitent ces inconvénients.

Autres méthodes :

Il existe d'autres méthodes d'indexation manuelle, en particulier celles basées sur la création d'un "vecteur de propriétés" où un concept n'est plus représenté par un mot, mais par un ensemble de "points de vue" [IRI 71] [SAL 68].

Dans les méthodes automatiques, ou semi-automatiques, on trouve aussi celles basées sur des méthodes linguistiques d'analyse de textes, et qui sont souvent restées au stade de la recherche (ex: SYNTOL [IRI 71]) ; remarquons que ces méthodes ne fournissent pas un moyen totalement automatique d'indexation mais doivent plutôt être considérées comme des outils d'aide à l'indexation.

1.3. LE MODE DE RECHERCHE

Les descriptifs des documents sont souvent des "listes de descripteurs", c'est à dire des ensembles de descripteurs, sans poids attaché à chacun d'eux et sans autre structure. Dans ces conditions, l'indexation rejoint le classement multiple et les descripteurs peuvent être considérés comme autant d'aspects différents des documents. Ainsi, la formulation d'une question s'obtient en reliant entre eux des descripteurs au moyen d'opérateurs booléens : les documents qui répondent à la question sont ceux dont les descriptifs satisfont l'expression booléenne ainsi formée.

Ce mode de recherche a d'abord été utilisé en mécanographie : il s'agissait de sélectionner les trous d'un ensemble de cartes perforées ("peek-a-boo cards") ou des cartes à encoches ("edge-notched cards") [SAL 68] [PAI 77].

Le principal avantage de cette méthode est la rapidité de recherche dès qu'une organisation adequate de fichier est choisie (fichier inverse, voir annexe F) : on connait alors immédiatement les descriptifs qui possèdent les descripteurs présents dans la question.

Son principal inconvénient est que l'élaboration de la réponse à une question ainsi formulée sépare la collection de documents en deux parties : les documents qui répondent à une question et ceux qui n'y répondent pas ; il est alors impossible d'apprécier les différences de pertinence entre les documents répondant à la question.

D'autres méthodes ont voulu ne pas avoir les principaux inconvénients de la méthode booléenne :

- D'abord, prendre en compte le fait que tous les descripteurs n'ont pas la même importance : le descriptif est alors un vecteur dans un espace de dimension le nombre de descripteurs possibles.

- Ensuite, pouvoir apprécier les différences de pertinence entre des documents qui répondent plus ou moins à la question.

L'approche de toutes ces méthodes est la même : le calcul d'un degré de similarité entre la question et chaque document (ou plus exactement entre le descriptif représentant la question, et le descriptif représentant chaque document) ; les documents répondant à la question sont ceux dont la similarité est jugée satisfaisante (c'est-à-dire supérieure à un certain seuil), et/ou les n premiers documents de similarité la plus forte.

Différentes mesures ont été proposées, basées sur l'idée de la comparaison du nombre de propriétés partagées et du nombre de propriétés différentes ; si X et Y désignent deux descriptifs, et $|X|$ la norme du descriptif X, la plus évidente et la plus simple est celle qui correspond au nombre de descripteurs partagés dans le cas où les descriptifs sont des vecteurs binaires, soit $|X \cap Y|$.

Le principal inconvénient de ces méthodes est qu'elles nécessitent des temps de calcul important, puisqu'en principe, il faut calculer la similarité pour tous les couples descriptif-question/descriptif-document, et cela peut limiter énormément les avantages d'une recherche en temps réel. En fait, certaines organisations de fichiers permettent de minimiser ces inconvénients (cf. expérimentation projet SMART [SAL 71]).

1.4. LE PROCESSUS DE RECHERCHE

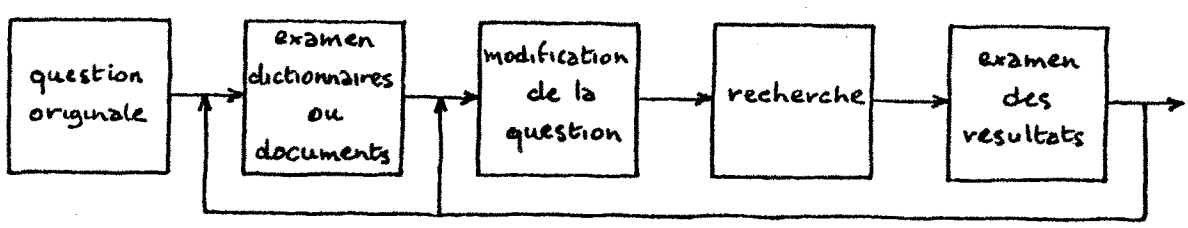
Indépendamment de la méthode choisie pour mesurer la similarité, il est bien rare que la question originale, formulée par l'utilisateur, puisse aboutir à un résultat satisfaisant. En fait, l'opération de recherche est une suite d'opérations, et c'est pour cette raison que nous l'appelons processus de recherche : la fin de ce processus est décidée par l'utilisateur quand il juge les résultats obtenus satisfaisants.

On peut distinguer deux méthodes de recherche [SAL 71] :

- La première nécessite un bon choix de descripteurs, obtenu principalement par la consultation de thesaurus, de dictionnaires,... par l'obtention d'informations statistiques sur l'apparition des descripteurs dans la collection, voire par la consultation de documents considérés comme pertinents (on cherche alors des documents semblables).

- La seconde nécessite de pouvoir examiner précisément les résultats obtenus, et donc d'avoir accès aux différents "attributs" des documents tels le titre, un résumé, le texte complet... La modification de la question peut être faite soit par l'utilisateur, soit par le système, lorsque l'utilisateur lui a communiqué les documents considérés comme pertinents : le principe est de maximiser la différence des mesures de similarité question-document entre ceux déclarés pertinents et les autres.

En fait, lorsque ces services sont présents, l'usager les utilise, et le processus de recherche peut être représenté par le schéma suivant :



1.5. LES LISTES DE DESCRIPTEURS

Les listes de descripteurs jouent un grand rôle, autant en documentation manuelle qu'automatique, et aussi bien lors des opérations d'indexation que de recherche.

L'indexation consiste à créer une structure de descripteurs qui va être associée au document : il peut être utile de posséder la liste des descripteurs à employer, et/ou la liste des descripteurs déjà employés.

Le processus de recherche nécessite la formulation d'une ou plusieurs questions : il est là aussi utile de connaître la liste des descripteurs utilisés ou utilisables.

Plusieurs termes existent et sont employés pour désigner de telles listes de descripteurs ; nous proposons d'éclaircir un peu les différentes définitions trouvées dans la littérature :

Thesaurus : c'est sans doute le plus employé

- "un ensemble organisé de descripteurs...relations...niveau sémantique" [TRY 71],
- "dictionnaire des synonymes" [SAL 68],
- "classification du vocabulaire" [SPA 71].

Il nous semble qu'une bonne définition serait de considérer un thesaurus comme la représentation d'un vocabulaire organisé de descripteurs ; on peut y trouver des descripteurs, des mots, du texte, reliés par des relations d'équivalence (entre mots et descripteurs), des relations hiérarchiques, associatives, ou catégorielles (entre descripteurs et descripteurs), et des relations de définitions (entre descripteurs et textes) [CHA 78].

Leur représentation se fait souvent sous forme de liste alphabétique des descripteurs, ou par représentation de "l'arbre hiérarchique", et quelquefois sous forme graphique ("schéma fléché", tableau graphique).

Lexique ou dictionnaire : c'est un ensemble de termes (mots ou descripteurs), sans autre structure, et présenté sous forme de liste alphabétique.

Un index conventionnel est l'outil de base de la recherche manuelle. Un index est principalement une liste alphabétique d' "entrées", chacune d'elles étant associée à des références vers l'information cherchée.

On trouve des index de termes (sujets), des index d'auteurs à la fin de nombreux livres. On trouve dans les bibliothèques et les centres de documentation des index des auteurs, des titres, des descripteurs, etc... ; ces index sont aussi appelés catalogues.

Beaucoup de ces index ont été édités automatiquement ; une variété d'entre eux consistant à lister l'ensemble des mots (considérés comme significatifs) d'un ensemble de textes, où chaque mot est associé aux contextes dans lesquels il apparait, connait un certain succès. On les appelle des "listes permutées" ou KWIC-index (Key-Word-In-Context et leurs variantes : KWOC, ... Out-of-Context, KWIT,... In-Title, KWAC, ... And ...). Basés sur la même idée, existent également des KLIC (Keys-Letters-In-Context) consistant à offrir une liste permutée sur l'ensemble des lettres de chaque mot du vocabulaire considéré (utilisés principalement sur des noms de formules chimiques).

1.6. EVALUATION, PERFORMANCES

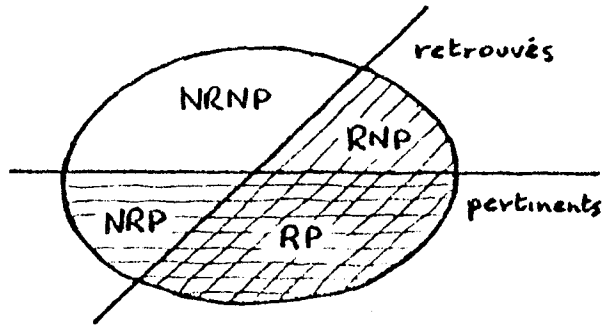
Il semblerait évident qu'un bon système serait celui qui, pour une question, retrouverait le maximum de documents pertinents pour cette question. Cette affirmation est vraie mais incomplète : il faut que, dans le même temps, la réponse à une question fournisse le minimum de documents non-pertinents.

D'autres critères de satisfaction existent (temps de réponse, type de recherche, qualité des sorties, etc...) mais ne nous intéressent pas ici.

Les mesures les plus courantes permettant d'apprécier les performances d'un système sont le rappel et la précision définis comme suit :

Soient les ensembles de documents suivants :

RP = documents retrouvés, pertinents
NRP = documents non retrouvés, pertinents
RNP = documents retrouvés, non pertinents
NRNP = documents non retrouvés, non pertinents

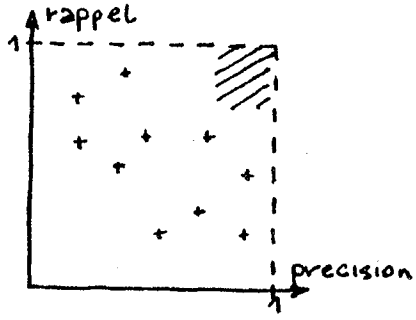


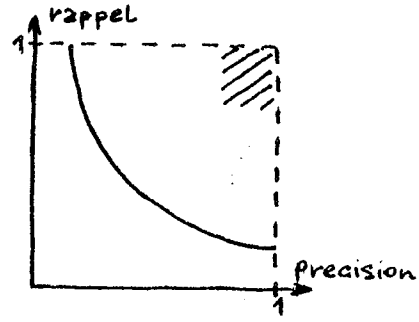
Le taux de rappel est le rapport RP/RP + NRP, qui représente la proportion de documents pertinents retrouvés par rapport au nombre total de documents pertinents.

Le taux de précision est le rapport RP/RP + RNP, qui représente la proportion de documents pertinents retrouvés par rapport au nombre de documents retrouvés.

Remarquons ce que nous mentionnions précédemment : un taux de rapppel maximum peut être obtenu en donnant toute la collection en réponse à une question ; il est clair que le taux de rappel seul n'est pas un élément suffisant de mesure des performances d'un système.

Pour un couple question/réponse particulier, on obtient un couple de valeurs rappel/précision, et pour un ensemble de question/ réponse, on obtient la figure ci-dessous à gauche ; certains résultats sont excellents, d'autres non, la zone idéale étant celle hachurée.





Si l'on fait la moyenne de ces différentes valeurs, on obtient une courbe d'allure générale indiquée ci-dessus à droite qui montre une caractéristique générale des systèmes documentaires et la difficulté d'avoir simultanément des taux de rappel et de précision satisfaisants. Il faut choisir ce que le système doit favoriser, et ce choix influencera les méthodes d'indexation et de recherche.

2 - DES SYSTEMES ET DES EXPERIENCES EN DOCUMENTATION AUTOMATIQUE

De nombreux centres de documentation, publics ou privés, existent aujourd'hui et tentent de satisfaire les besoins des utilisateurs. Ces centres de documentation sont construits autour de logiciels documentaires : nous avons vu précédemment la ou les définitions de la documentation automatique et les problèmes posés ; nous allons maintenant examiner comment les logiciels documentaires répondent à ces problèmes.

Nous mentionnerons ensuite les principales études, travaux ou recherches qui ont été effectués dans ce domaine, et dirons quelques mots concernant les besoins et travaux apparus depuis que l'utilisation des réseaux de transmissions de données est passée dans le domaine pratique.

2.1. LES SYSTEMES EXISTANTS

"En Juin 74, le BNIST (Bureau National de l'Information Scientifique et Technique) et la DIELI (Direction des Industries Electroniques et de l'Informatique) ont décidé de confier au CXP (Centre d'Experimentation des Packages) la réalisation d'une étude sur les logiciels et systèmes documentaires disponibles en France. Le STI (Service Technique Informatique) de l'IRIA était chargé d'assurer le suivi technique du contrat" [ADBS 76].

Cette étude a été publiée en Mars 76 [ADBS 76] et n'a pas été remise à jour depuis cette date ; elle permet toutefois de se faire une idée sur ce qui existe aujourd'hui.

On trouve dans l'introduction : "On parle beaucoup de bruit et de silence, de philosophie d'indexation, de méthodologie de construction de thesaurus et d'évaluation des systèmes documentaires. Alors on s'adresse à l'aspect subjectif de la documentation qui travaille sur un concept flou et ambigu. Dans l'étude sur les logiciels documentaires, seul le côté objectif, c'est-à-dire le programme d'application a été examiné. Un bon logiciel documentaire ne pourra jamais constituer un bon système documentaire s'il s'appuie sur un édifice en mauvais état, à savoir le thesaurus et l'indexation. Cet aspect est volontairement absent de l'étude car il pose à lui seul des problèmes d'une très grande complexité qui sont étudiés par ailleurs et sont encore du domaine de la recherche (construction automatisée de thesaurus, indexation automatique, évaluation de performances des systèmes documentaires)".

Ces propos situent bien les limites de l'étude, mais aussi les limites des systèmes existants, puisqu'à part un produit qui mentionne l'utilisation possible d'un autre programme en amont, aucun "marchand" n'offre de services sérieux d'indexation : il semble bien que les producteurs de logiciels aient considéré qu'il s'agissait là des problèmes des clients. Il ne faut pas sous-estimer la part de ce travail effectué par les documentalistes, mais on peut être surpris de trouver dans l'avant-propos de [CHA 78], un conseiller à la Commission des Communautés Européennes affirmer : "la construction d'un langage documentaire est une chose sérieuse qu'on ne peut confier à un ordinateur...".

L'enquête a d'abord consisté en un inventaire des logiciels documentaires existants actuellement en France, et a permis d'en dénombrer plus d'une trentaine. On peut les répartir en trois classes suivant leur origine :

- On trouve d'abord des produits figurant au catalogue des constructeurs de matériel informatique, par exemple RIRMS et STAIRS (IBM), MISTRAL (CII-HB), GOLEM (SIEMENS)...

- On trouve ensuite les gros produits programmes écrits par des organismes désireux de se doter d'un matériel documentaire important ; on peut citer SABIR (Institut Gustave Roussy), SAFIR (CNET), TITUS (Institut Textile de France), PASCAL (CNRS), "chaîne IFP" (Institut Français du Pétrole)...

- On trouve enfin des produits dont le champ d'action commercial ainsi que les possibilités techniques sont en général très spécialisés ; ce sont par exemple CORA et MARABOUT(ROUSSEL-UCLAF).

L'étude s'est ensuite poursuivie par l'examen plus approfondi de onze produits qui sont sur le marché français et qui peuvent être considérés comme étant à "vocation générale". Ces produits sont GOLEM, "chaîne IFP", MISTRAL, PASCAL, RIRMS, SABIR, SAFIR, SATIN (URADCA Aix-Marseille), SPLEEN (CNRS), STAIRS et TITUS.

Les principales conclusions qui nous intéressent concernent la représentation des documents, l'indexation et le mode de recherche. On ne parlera pas des problèmes de format de saisie, comptabilité, gestion, etc,...

Les principaux résultats sont les suivants :

- La représentation des documents se fait uniquement en texte libre et/ou par liste de descripteurs ; on trouve quelquefois des indicateurs de rôles ou de liens pour relier des descripteurs (chaîne IFP, SATIN) ; une mention particulière doit être accordée à TITUS qui utilise une représentation à l'aide de phrases à syntaxe pré-déterminée dans un but de traduction automatique.

- L'indexation est soit manuelle, soit automatique. La manière automatique, lorsqu'elle existe, consiste à comparer les mots d'un texte à un dictionnaire de mots-vides, ou bien à extraire d'un texte des mots précédemment mis en valeur (manuellement). Seul, le produit PASSAT utilisé en amont de GOLEM semble offrir une méthode "évoluée" d'indexation automatique dérivative.

- La recherche se fait toujours par une expression booléenne de descripteurs (et, ou, sauf), avec éventuellement des possibilités de troncatures de ces descripteurs, opérateurs de comparaison, ou pondération (chaîne IFP). Lorsqu'il existe une représentation en texte libre, on trouve des possibilités de recherche dans le texte, avec les opérateurs de présence/absence, de troncature, de co-occurrence, de distance...

On peut également mentionner :

- Qu'il existe généralement des outils permettant la gestion de thesaurus, lexiques et autres dictionnaires.

- Que la recherche retrospective se fait en traitement par lots, mais de plus en plus en conversationnel.

- Qu'il existe souvent une possibilité de Diffusion Sélective de l'Information.

Notre conclusion sera que ces différents logiciels documentaires peuvent être considérés comme de "super" systèmes de gestion de fichiers, ce qu'ils étaient à l'origine, et ce qu'ils restent en partie : on parle toujours de "fichiers documentaires". L'exemple le plus significatif est sans doute le produit FIND d'ICL, mentionné dans l'étude précédemment citée, donc considéré comme logiciel documentaire, et qui n'est en fait défini "que" comme un logiciel général d'interrogation de fichiers. On peut aussi les considérer comme de "sous" systèmes de gestion de bases de données, "sous" dans le sens où ce sont des systèmes spécialisés dans la gestion de structures de données particulières.

Ces logiciels sont employés dans des centres de documentation généraux, dont des exemples représentatifs sont :

- Le Centre de Documentation de l'ESA (Agence Spatiale Européenne), situé à Frascati (Italie), utilisant le produit RECON sur IBM, développé à l'origine par la LOOCKED.

- Le nouveau "gros serveur" français, situé à Sofia-Antipolis, utilisant le produit MISTRAL sur IRIS 80 (CII-HB).

Bien évidemment, (et malheureusement) ces centres ne possèdent pas les documents eux-mêmes, mais simplement des références à ces documents (en général, ils ne possèdent pas non plus d'informations permettant d'indiquer à l'utilisateur à quel (s) endroit (s) il pourrait trouver ces documents). Il est frappant de constater que l'élaboration de ces références est faite par des organismes spécialisés, et que les centres de documentation sont utilisateurs de ces bases, comme ils sont utilisateurs des logiciels documentaires. Il semblerait normal que ces organismes, qui réalisent l'indexation des documents, recherchent ou utilisent des logiciels permettant de faciliter ce travail ; il ne semble pas que ce soit le cas.

Afin de bien montrer ce qu'est la réalité, aujourd'hui, des services offerts par les centres de documentation, nous donnons ci-dessous trois exemples que nous pensons représentatifs.

. Le premier exemple est le centre de documentation de l'IRIA.

Cet organisme possède sans doutes la plus importante bibliothèque française dans le domaine de l'informatique.

Chaque document a été indexé manuellement par une liste de mots-clés et a été associé à une référence, comportant principalement un numéro d'inventaire, une catégorie de support, un titre, une date, un éditeur, et une liste de mots-clés.

Ces références sont gérées par un logiciel s'appelant MAGISTRAL, et capable d'éditer des listes par mots-clés, par auteurs, etc... ces listes peuvent être consultées sur place.

Ce Centre de Documentation possède un service "question-réponse", destiné à renseigner les utilisateurs : son fonctionnement est représenté par la suite d'opérations ci-dessous :

- Réception par lettre ou par téléphone, de la question d'un utilisateur.

- Transformation, par du personnel spécialisé, de cette question en un ensemble de mots-clés reliés par des opérateurs booléens.

- Regroupement de questions pour former un travail passant en "batch" le jour même ou lorsqu'il y aura suffisamment de questions.

- Envoi du listing résultat à l'utilisateur.

Sans vouloir minimiser les services rendus par ce centre, il faut bien reconnaître que "les cordonniers ne sont pas les mieux chaussés": ce centre est l'exemple typique de ce qu'il ne faudrait plus faire.

Son premier défaut est que l'utilisateur ne dispose pas du dictionnaire des termes à employer : il a formulé une question en langue naturelle, et sauf cas exceptionnel, la transformation de sa question ne reflétera pas exactement sa pensée. Son second défaut est que le système n'est pas interactif ; il est donc impossible de mettre en oeuvre une stratégie de recherche comportant consultation de dictionnaire, examen des résultats, re-formulation de la question.

- Le second exemple est celui du centre de documentation de l'ESA.

Il ne possède pas les défauts exprimés précédemment, c'est-à-dire que le logiciel utilisé est interactif et accessible à distance, et représente bien ce qui "se fait de mieux" dans le domaine commercial aujourd'hui.

Les facilités offertes sont les suivantes :

- Services généraux donnant des informations sur le système, les bases gérées, etc...

- Services permettant de sélectionner les bases avec lesquelles l'utilisateur désire travailler.

- Services de gestion de dictionnaires, permettant de visualiser des termes alphabétiquement voisins, des termes reliés par une relation donnée, etc...

- Services de recherche par formulation d'une question, par expression booléenne de descripteurs, spécification de champ de recherche, facilités de troncature, masquage, etc...

- Obtention des résultats sous forme d'un ensemble de documents, obtention du cardinal de cet ensemble, opérations ensemblistes (et, ou, sauf) sur ces résultats.

- Services d'éditions variées.

- Possibilité de catalogage de questions et de conservation des stratégies de recherche.

- Le troisième exemple est celui du centre de documentation du CNRS.

Il possède des services de production de bases de données, employables et employés par d'autres centres (et se présentant sous forme de bandes : les bandes PASCAL), et des services d'éditions de bulletins signalétiques. Il possède également un service de recherche retrospective analogue à ceux présentés précédemment.

Un service important est la Diffusion Sélection de l'Information. Les centres d'intérêts des abonnés de ce service sont représentés par des profils. Ces profils peuvent être soit des profils de groupe, définis par relation avec des utilisateurs représentatifs de besoins identiques, des profils dits "standards", des profils individuels définis par discussion avec des abonnés isolés ou en petit nombre, profils éventuellement confidentiels.

Les résultats sont diffusés aux abonnés environ une fois par mois, sous forme de fiches détachables : chacune de ces fiches signale un document avec tous les éléments bibliographiques significatifs, les mots-clés et un résumé.

2.2. LES RECHERCHES

Les investigations dans le domaine de la documentation automatique, et particulièrement en France n'ont pas donné lieu à beaucoup de travaux ; il suffit de consulter l'Annuaire de la Recherche en Informatique et Automatique édité par l'IRIA pour se rendre compte du faible nombre de recherches et de chercheurs, et ce, tous aspects confondus, c'est-à-dire, aussi bien sur les problèmes d'indexation automatique, analyse de langues naturelles, que sur ceux du traitement de textes, de l'édition, etc... Notons également que l'IRIA possédait vers les années 70 une activité documentaire, aujourd'hui disparue...(voir [IRI 71]).

Par contre, aux Etats-Unis ou en Angleterre, il semble qu'un important travail ait été fait, bien qu'il soit resté méconnu ou ignoré du "grand public informaticien ou documentaliste". Il n'est pas question ici d'établir un catalogue des différents travaux effectués (nous en serions bien incapables), mais de mentionner les "gros projets" réalisés, et dont les résultats sont très souvent cités dans la littérature. Tous ces travaux ont débuté à la fin des années 50, et parmi les noms qui leur sont attachés, ceux de C.W. CLEVERDON, G. SALTON et K. SPARCK-JONES apparaissent souvent [PAI 77].

- Le projet de CRANFIELD, en Angleterre, a vu le jour en 1957, sous la direction de C.W. CLEVERDON. Il y a eu en fait deux expérimentations :

  - La première a consisté en un test comparatif de performances de méthodes de recherches manuelles, utilisant quatre méthodes d'indexation manuelle.

  - La seconde a été une comparaison des performances de différents langages d'indexation, consistant à examiner l'influence de différentes méthodes d'indexation sur les performances de la recherche (utilisation de mots simples, famille de mots avec racine commune, mots + synonymes, mots + mots reliés...).

    Cela a nécessité la disposition d'une collection de 1 400 documents (en aéronautique) dont les auteurs eux-mêmes ont été mis à contribution pour définir des questions relatives à leurs papiers, et donner des informations relatives à d'autres documents considérés comme proches des leurs.

- Le projet de CAMBRIDGE (Cambridge Language Research Unit), toujours en Angleterre, a débuté à la fin des années 50, et a été plus particulièrement approfondi par K. SPARCK-JONES et décrit dans [SPA 71]. L'idée principale est l'utilisation de relations entre descripteurs pour la formulation des questions et l'expérimentation a consisté dans l'examen de l'influence de différentes méthodes de groupement de descripteurs sur les performances de recherche, autrement dit dans l'examen de différentes méthodes de construction automatique de thesaurus.

- Le projet SMART, dirigé par G. SALTON (décrit dans [SAL 71]), a commencé au début des années 60 à Harvard (Etats-Unis) et s'est poursuivi à Cornell University. C'est certainement la plus complète des expérimentations faites à ce jour : elle a consisté à considérer le problème sous tous ses aspects, et en particulier à expérimenter des méthodes d'indexation automatique, des structures de fichiers où les documents sont classés afin de minimiser les temps de recherche ("clustered files"), des méthodes de prise en compte d'informations fournies par l'usager lors du processus de recherche ("relevance feddback"), la construction automatique de thesaurus. L'appréciation de la validité de méthodes a été faite par comparaison avec des indexations manuelles.

Il faut mentionner que ceci s'est fait dans le cadre de la définition et de la réalisation d'un système réel, dont la moindre des qualités n'est pas l'acceptation d'une question exprimée en langue naturelle, comme point de départ du processus de recherche.

Le principal résultat de ce projet est certainement qu'il a démontré que les méthodes d'indexation automatique donnaient des résultats au moins égaux à ceux obtenus par des méthodes manuelles. Sa principale qualité est sans doute de ne pas être resté à un stade d'expérimentation et d'avoir réalisé un système opérationnel, afin de montrer ce qu'il est possible de faire, dans ce domaine, aujourd'hui.

2.3. AUJOURD'HUI, PLUS PRES DE NOUS

Depuis quelques années, et en particulier depuis l'avènement des réseaux de transmissions de données, il existe différents projets en cours de développement.

- Le projet EURONET, "résultat d'une résolution du conseil des ministres de la communauté européenne, pour promouvoir la coopération dans les échanges d'informations scientifiques et techniques" [HIG 77]. Le but principal est de produire un réseau d'information distribué à travers l'Europe permettant à des usages d'accéder à une grande masse d'informations scientifiques et techniques.

  Les principaux services offerts sont les suivants :

  • pour l'usager, la recherche retrospective interactive, et la diffusion sélective,

  • pour les centres participants, la transmission de données pour la création et la maintenance des bases.

  Le moyen de réaliser cela se fait par connexion à un même réseau, ou par inter-connexion de réseaux nationaux existants, de centres documentaires existants. En 1976, il était prévu de connecter une trentaine de centres, ceci représentant environ une centaine de bases, certaines étant dupliquées.

  Les différentes études préliminaires ont en particulier abouti à la définition d'un jeu de commandes standards pour les systèmes documentaires [NEG 76] [NEG 77]. En d'autres termes, et vu sous l'angle documentation automatique, EURONET "n'est qu'un" service d'accès à diverses bases, à l'aide de commandes identiques.

- Plus près de nous, au sein du projet-pilote SIRIUS, quelques personnes ont défini le système idéal qui devrait permettre une interrogation multi-bases grâce à la formulation d'une question dans un langage unique, aussi bien pour les commandes que pour le langage documentaire utilisé, la localisation des documents sélectionnés, la réservation ou la commande des documents retenus [IRI 79].

Différentes actions ou reflexions sont en cours, en particulier :

. La délimitation des problèmes spécifiques des systèmes documentaires répartis par rapport aux SGBD répartis.

. L'étude des problèmes posés par l'hétérogénéité des langages documentaires.

. La réalisation d'un logiciel MISTRAL réparti utilisant des logiciels MISTRAL standards (CERISS Toulouse).

Une "action pilote" de base documentaire répartie a été lancée, ayant pour objectif l'interrogation à l'aide d'un langage de commande unique, d'un langage documentaire unique, des fonds documentaires français traitant de l'informatique. Dans le même esprit, une maquette pour la gestion et l'interrogation réparties d'un catalogue collectif de périodiques a été réalisée en utilisant le SGBDR FRERES (IRISA Rennes).

C'est à peu près tout ce qui se fait d'assez important aujourd'hui. Il serait toutefois injuste de ne pas citer les travaux de quelques équipes s'intéressant à des problèmes particuliers de la documentation automatique.

Il faut citer par exemple :

- le projet PIAF (équipe intelligence artificielle, IMAG, Grenoble),

- les travaux de l'Institut Gustave Roussy, Villejuif (Indexation, évaluation),

- les travaux de l'IMSS, Grenoble (analyse du français),

- le projet SATIN, CNRS Marseille.

3 - LA BUREAUTIQUE. LES DEBUTS

On se propose, dans ce chapître, de donner un aperçu de ce qu'est la bureautique, à travers les différentes définitions qui ont été proposées, à travers les différentes réalisations existantes, et à travers la vision que nous en avons, afin de définir un vocabulaire et un cadre de travail.

3.1. DES DEFINITIONS

Ce mot se veut être la traduction de l'expression américaine "office-automation". Ce mot désigne un sujet très à la mode (certains y mettent un "o"...) et nous nous garderons bien d'en fixer les limites ; laissons à son auteur (L. NAUGES) le soin d'en donner la définition [BUR 78] :

> "La bureautique est concernée par le développement des technologies de saisie, mémorisation, traitement, diffusion de l'information, et par leur emploi pour la gestion des messages formels et des textes dans les organisations".

D'autres, plus précis, y voient (J.P. DE BLASIS dans [BUR 78]) :

- traitement et gestion de textes,
- gestion des communications (messages textuels, téléphoniques, transmission de documents, traitement d'images, graphiques,...),
- gestion d'emploi du temps (agendas, mementos, coordinateurs d'emploi du temps),
- archivage et accès aux documents,

- suivi du processus de secrétariat (prise en compte des messages, planning, déclanchement, suivi, mise à jour de processus),

- aide à la décision ("alerteurs", systèmes interactifs).

Une définition, plus concise, considère la bureautique concernée par :

- la gestion des textes,
- la gestion des communications,
- la gestion du temps,

c'est à dire un ensemble de services que N. NAFFAH répartit en trois groupes [NAF 79] :

- des services individuels (aide à la décision, production de textes, dessins, systèmes d'informations personnels tels répertoires, agendas, archives),

- des services de groupes (messagerie, téléconférence, production de documents...),

- des services partagés (archivage, bases de données, impression, photo-composition...).

Nous ne donnerons pas une autre définition ; nous disions dans [ABH 79] que l'objet d'un système bureautique était de gérer l'information de bureau : nous ajouterons aujourd'hui "en particulier l'information écrite", et seront ainsi assez proches des "prospectus" IBM qui parlent de "système de communication écrite".

Nous pensons que cette information écrite (ou "information textuelle" ou "information non numérique") est souvent désignée par le terme document, et c'est le terme et le concept que nous emploierons. Un document peut être un message, un télégramme, une lettre, un rapport, etc... il peut être composé, manipulé, communiqué, classé, ... par des utilisateurs.

En essayant d'identifier et de préciser les rapports (ou relations) qui existent entre ces ensembles de documents et d'utilisateurs, on distingue :

- les rapports que les utilisateurs peuvent avoir avec les documents,
- les rapports qu'ont les utilisateurs entre eux, au moyen de documents,
- les rapports que les utilisateurs peuvent établir, constater ou rechercher entre les documents.

Ainsi, ces relations sont réparties en trois domaines principaux, désignés, sinon par des termes exacts, du moins consacrés par l'usage :

- Le traitement de textes, ensemble de fonctions permettant la composition et l'édition de textes et de documents.
- Le courrier électronique permettant la communication de documents entre utilisateurs.
- La documentation automatique permettant le classement et la recherche des documents.

Aussi, toujours sans vouloir donner une définition exhaustive de la bureautique, nous pensons que les trois domaines précités en font partie, et qu'ils figureront dans toute définition de la bureautique.

Ajoutons que ce choix est évidemment déduit des activités et préoccupations des auteurs de [ABH 79] et de leur qualité d'informaticiens ; dans tout ce qui suit nous parlerons informatique et laisserons de côté d'autres techniques, telles phoniques, vidéo, photocopies,...

3.2. DES REALITES

Le traitement de textes

La bureautique aujourd'hui c'est d'abord et principalement le traitement de textes. Les informaticiens savent ce que c'est, puisqu'ils se sont dotés depuis longtemps d'un certain nombre d'outils permettant de gérer facilement des bibliothèques de programmes et d'effectuer des modifications et mises à jour de ces programmes. Ces outils se sont développés avec l'apparition des systèmes conversationnels, et ont pris le nom d' "éditeurs de textes" (traduction hâtive du terme anglo-saxon "editor"...). Ces outils ont alors été utilisés pour créer et mettre à jour non plus des programmes mais des textes, d'abord à caractère technique. Puis, le besoin d'outils plus sophistiqués s'est fait sentir, dans le but d'élaborer des textes plus évolués, et des produits tels SCRIPT, TEXT 360, GUTEMBERG sont apparus.

Nous n'insisterons pas sur les définitions possibles du traitement de textes ; on suppose certains de ces produits connus du lecteur ; les fonctions qu'on y trouve donnent une idée de celles qu'on trouve sur les systèmes bureautiques de traitement de textes.

Les matériels que l'on trouve peuvent être répartis en plusieurs classes [BUR 78] :

- Des ordinateurs "classiques" exploités en temps partagé, ou des ordinateurs de gestion avec un logiciel spécialisé (CII-HB IRIS 80 + GUTEMBERG, IBM 32 + GAT 32).

- Des "ordinateurs de textes" ou "systèmes à logique partagée", ou "multipostes avec écran", qui sont des ensembles de postes de travail avec écran, reliés à une unité centrale, possédant imprimante, disques (WANG 30, DEC WS 200).

- Des "monopostes avec écran", qui sont des postes de travail autonomes, avec écran, avec 1 ou 2 disques souples, une imprimante, étant soit cablés, soit programmés (CII-HB TTX 60/80).

- Des "monopostes sans écran", ou "machines à écrire électronique", qui sont des machines à écrire dotées d'un support magnétique de petite capacité (IBM memosphère, Olympia 6010). C'est actuellement la plus répandue.

Les communications

La bureautique, demain, c'est la messagerie électronique ou le courrier électronique.

On connait le téléphone depuis longtemps, son utilisation dans les activités de bureau, et l'utilisation de ses techniques pour ce qui a d'abord été désigné par le terme télé-informatique.

Les outils de communication dont aura besoin la bureautique se répartissent en deux groupes :

- Les outils de communication locale.

- Les outils de communication à vocation générale, services généralement offerts par les PTT.

Quelques expériences de communications locales ont eu lieu, soit par le biais d'études sur les systèmes multi-processeurs (ex : CORAIL), soit par le biais d'études sur les systèmes répartis (ex : DANUBE), soit par les "marchands" de téléphones ou de commutateurs, et ces outils ont pris le nom de "réseaux locaux".

Les outils de communication à vocation générale sont les réseaux de transmission de données dont l'exemple français est TRANSPAC digne successeur du réseau expérimental CYCLADES. A côté, ou par-dessus ces outils, les PTT ont développé ou développent des services dont la hiérarchie est la suivante [BUR 79] :

". Le TELEX pour transmettre, avec un alphabet réduit, un message, sans présentation soignée, sur un réseau spécifique et étendu.

. Le TELETEXT, pour transmettre, avec un alphabet plus large, une correspondance dont la présentation a une grande importance, et donc fournissant des moyens de traitement de textes.

. La TELECOPIE, pour transmettre des informations graphiques non codables par caractères. "

S'il est évident que le Telex ne servira pas pour des applications bureautiques, il est non moins évident que les applications bureautiques réparties géographiquement ne pourront ignorer le Teletext ou la Telecopie ; le Teletext a d'ailleurs été défini dans ce but : une de ses propriétés est d'assurer la similitude des documents émis et des documents reçus ce qui est, sinon bien adapté, du moins conforme aux habitudes prises en courrier "manuel".

Les projets, les recherches

La réalité, les expériences en "grandeur nature" concernent le traitement de textes ; certaines grandes entreprises se sont équipées en services de traitement de textes, et l'on commence à posséder une histoire ; on peut maintenant disposer de chiffres et de faits précis sur ces évènements, c'est-à-dire d'une bonne connaissance de coût du traitement de textes, et des conséquences de l'introduction de ces outils sur la vie des bureaux (on trouvera dans [BUR 78] des éléments intéressants sur ces sujets).

D'autres expériences, en grandeur nature, ont eu lieu ; certaines restent pour l'instant des outils d'informaticiens au service d'informaticiens ; ce sont principalement des expériences de messagerie électronique (MAIL-Arpa, HERMES-Telenet, AGORA-Cyclades) ; d'autres, comme des services de téléconférence (par ex. PLANET) commencent à connaître un certain succès dans le "grand public".

Dans le domaine de la recherche, certains projets ont vu le jour, et leur caractéristique commune est de vouloir couvrir tous les aspects de la bureautique. On citera :

- Le projet OASYS (Office Automation System), "Système informatisé d'aide au processus de secrétariat", démarré depuis l'été 74 à l'Université de Pennsylvanie (décrit dans [BUR 78]).

- Le projet "GMD Textkommunication System" d'un institut de recherche en Allemagne Fédérale dont le but est de faire des investigations dans le domaine de l'introduction de la technologie informatique dans le secteur public (décrit dans [BUR 79-2]).

- Plus près de nous, le projet-pilote KAYAK de l'ADI/INRIA, défini dans ses objectifs et ses moyens, mais dont les contours sont encore flous....

3.3. UNE VISION DES CHOSES

Nous reprenons ici ce qui a déjà été suggéré comme définition de la bureautique au début de ce chapitre, et développé en détail dans [ABH 79], ceci afin de définir précisément le vocabulaire employé et le cadre de travail dans lequel on se place.

Nous avons identifié trois domaines principaux :

- le traitement de textes
- le courrier électronique
- la documentation automatique.

Chaque domaine correspond à un service ; nous dirons que chaque service est une vue partielle du système intégré de bureautique. Bien que chaque service apparaisse relativement autonome, les domaines ne sont pas indépendants : en particulier, l'intersection évidente de ces trois vues partielles est le concept de document.

On définira le système de bureautique, d'abord par l'union des différentes vues partielles ; on complètera ensuite ce schéma en spécifiant les interactions entre ces vues partielles ; une application mettant en jeu l'ensemble de ces relations aura une vue globale ou intégrée du service bureautique.

La méthode d'analyse proposée repose sur la notion de schéma ou de description conceptuelle. L'outil utilisé est le langage Z [ABR 77] ; la description d'un service comporte d'une part la description des objets et des relations entre ces objets, et d'autre part la description des transactions qui permettent de manipuler ces objets et relations.

Le service courrier électronique :

Le service courrier électronique permet de transmettre des documents (messages, lettres, mots, articles...). Le service voit le document comme un récipient, ne s'intéresse pas à son contenu, et n'en connait que l'aspect externe (couverture, étiquetage,...) qui lui permet de l'identifier.

L'unité d'adressage est l'abonné (personne, groupe, entreprise, ...); celui-ci possède une identité (numéro interne, adresse, nom/prénom, raison sociale, ...) qui permet de l'identifier et de le désigner.

- ensembles manipulés et relations

  DOCUMENT = ensemble de documents

  ABONNE = ensemble des abonnés

DOCUMENT origine (1) ABONNE

L'origine d'un document est un abonné qui a des droits particuliers sur ce document (création, suppression, diffusion, ...).

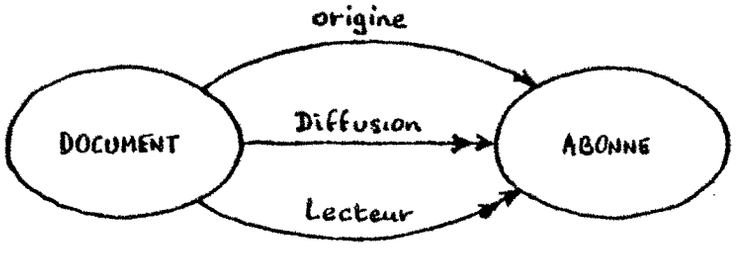
DOCUMENT Diffusion (0,-) ABONNE

La diffusion d'un document est l'ensemble des abonnés auxquels il est ou a été adressé.

DOCUMENT Lecteur (0,-) ABONNE

Un abonné est dit lecteur d'un document s'il appartient à la diffusion de ce document et ne l'a pas encore retiré.

. Illustration



· Transactions

Abonner/Désabonner :	Créer/Détruire un abonné ; le système peut alors/ne peut plus l'identifier. Un abonné ne peut être détruit que s'il n'est plus lecteur ni origine de documents.
Entrer/Sortir :	Créer/Détruire un document ; il en devient ainsi l'origine. Un document ne peut être détruit que s'il n'y a pas d'abonnés lecteurs de ce document. La diffusion disparait en même temps que le document.
Diffuser :	Créer, pour un document, une liste de diffusion composée d'abonnés qui en deviennent alors lecteurs. Seul l'abonné origine d'un document peut le diffuser.
Retirer :	Un abonné lecteur d'un document peut le retirer. Il n'en est alors plus lecteur mais figure toujours sur la liste de diffusion.

Le service documentation automatique

Le service documentation automatique permet d'effectuer un travail de recherche sur un ensemble de documents ; pour cela, un document possède un descriptif, structure de descripteurs.

· ensembles manipulés et relations

DOCUMENT = ensemble des documents

DESCRIPTIF = ensemble des descriptifs

DESCRIPTEUR = ensemble des descripteurs

DOCUMENT indexation (1) DESCRIPTIF

L'indexation d'un document est un descriptif qui en représente la "teneur".

DESCRIPTEUR Référence (0,-) DESCRIPTIF

La référence d'un descripteur est l'ensemble de descriptifs qui comportent ce descripteur.

. Illustration

. Transaction

Entrer/Sortir :	Entrer un document dans le service documentaire, c'est lui associer un descriptif. Sortir un document provoque la destruction de son indexation.

Le service traitement de textes

Un document possède une couverture (ou en-tête) et un contenu (ou corps). Un texte peut citer des documents. Un texte est une structure de mots empruntés à un certain vocabulaire.

. Ensembles manipulés et relations

DOCUMENT = ensemble des documents

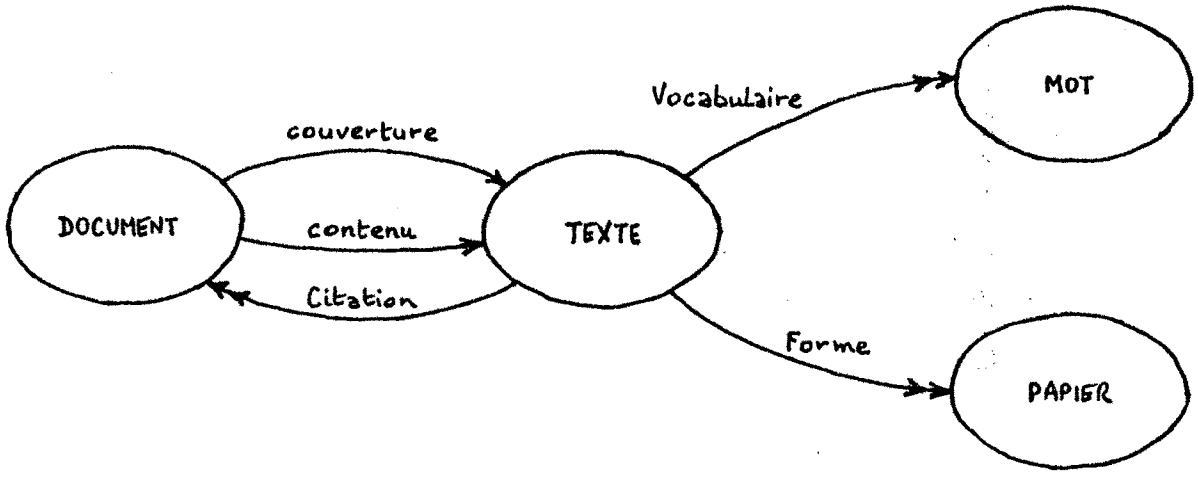
TEXTE = ensemble des textes

DOCUMENT couverture (1) TEXTE

DOCUMENT contenu (1) TEXTE

TEXTE Citation (0,-) DOCUMENT

. Illustration



. Transactions

Composer/Effacer : c'est créer/détruire un texte

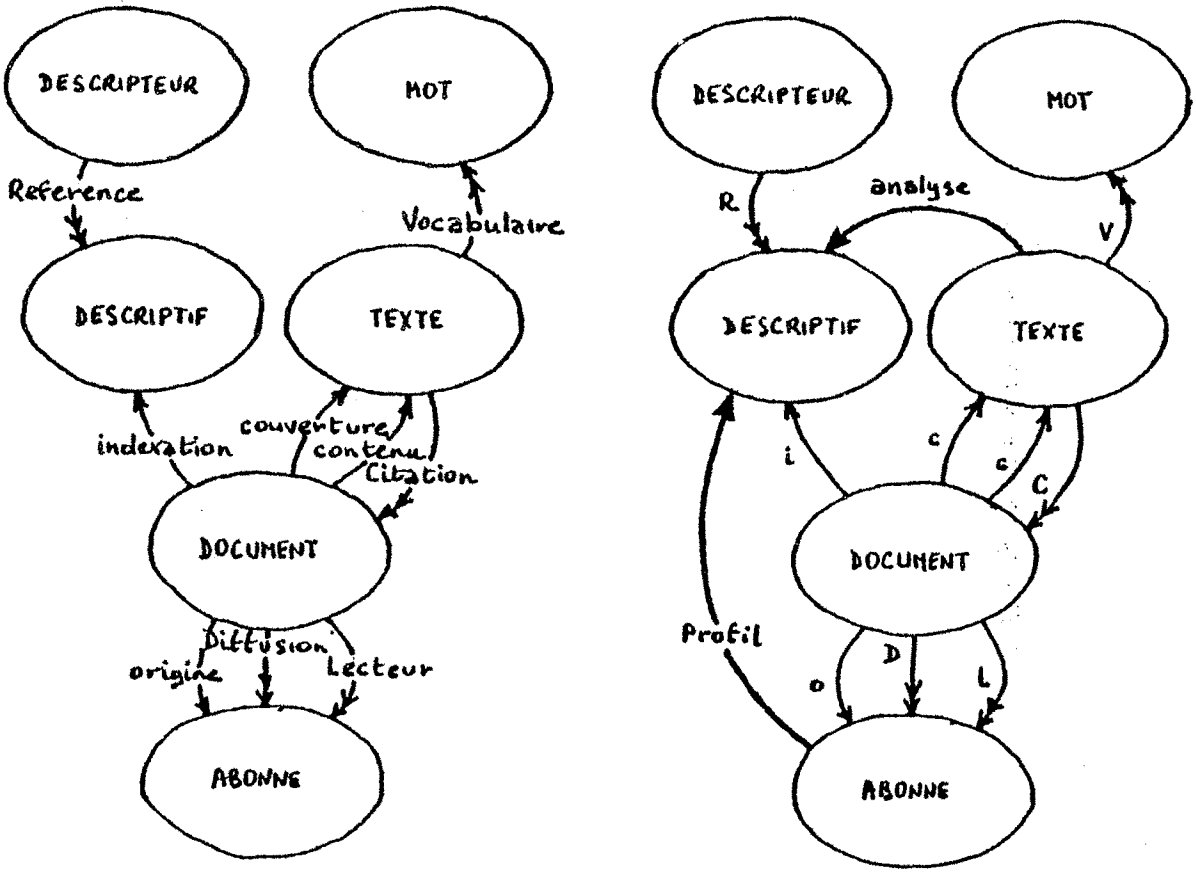
Entrer/Sortir : c'est associer/supprimer un texte couverture et un texte contenu à un document.

Le service bureautique global

Il semble évident que le concept de document constitue l'interdépendance des différents services décrits précédemment ; on définit alors le service bureautique global comme l'union de ces différents services : c'est la figure ci-dessous à gauche.

On peut maintenant enrichir ce schéma par la coopération des différents services, c'est-à-dire par la définition de nouvelles relations et transactions utilisant les concepts des différents services.

On présente ci-dessous, et sur la figure de droite, le schéma du service documentaire "complet" défini de cette façon :



. nouvelles relations

ABONNE profil (0) DESCRIPTIF

Le profil d'un abonné est le descriptif représentant son domaine d'intérêt.

TEXTE analyse (0) DESCRIPTIF

L'analyse d'un texte est le descriptif qui en représente la teneur.

. nouvelles transactions

Profiler : profiler un abonné, c'est lui associer un descriptif.

Analyser : analyser un texte, c'est lui associer un descriptif qui en représente la teneur ; l'opération d'indexation d'un document pourra ainsi être considérée comme une fonction de composition des descriptifs analyse des textes couverture et contenu de ce document.

Diffusion-sélective : diffuser sélectivement un document, c'est le diffuser aux abonnés intéressés, c'est à dire ceux dont le profil est proche de l'indexation du document.

3.4. DES PERSPECTIVES

L'apparition de la bureautique telle que définie dans les paragraphes précédents va certainement d'abord avoir lieu à des endroits où existent aujourd'hui des outils le permettant ; ce sont principalement des centres de recherche s'intéressant à ce problème, possédant des ordinateurs petits ou gros, reliés entre eux ou possédant des réseaux généraux ou locaux.

Il est fort probable que le développement industriel se fera ensuite à des endroits où la manipulation de documents, de textes, de courrier, occupe une place importante, tant quantitative que qualitative dans les activités exercées à ces endroits. C'est la "bureautique de bureau" ou à caractère professionnel.

Il ne faut pas oublier un autre secteur où la bureautique va faire son apparition : c'est celui des usages domestiques ou "grand public" qui va se développer à partir des nouveaux services de communications offerts par les PTT à chaque abonné, ou à partir de la diffusion des ordinateurs individuels.

On peut alors envisager l'introduction de la bureautique de deux façons :

- soit comme une aide aux fonctions de secrétariat telles qu'elles existent aujourd'hui : la bureautique existera là où existe aujourd'hui un secrétariat.

- soit comme un nouvel outil, destiné à offrir à chaque individu ou à des petits groupes d'individus des services de secrétariat.

Dans chacune de ces hypothèses, les services offerts et les documents manipulés ne seront pas les mêmes :

. dans le premier cas, les documents seront plutôt du "courrier", souvent à caractère administratif, (voire bureaucratique...), à usage immédiat,

. dans le second cas, plutôt des "documents de travail", souvent à caractère technique ayant un aspect plus permanent.

Si le premier point de vue n'impose pas de changement dans les méthodes de travail, le second risque de bouleverser ces dernières et de mettre en jeu l'existence même des secrétariats traditionnels.

Nous pensons que la possession de ces outils va également permettre le développement de nouveaux services, particulièrement "documentaires" et ce, à deux niveaux :

. au niveau de la bureautique individuelle (qu'elle soit de "bureau" ou "domestique") on se trouve en présence des problèmes "classiques" de gestion d'une collection de documents.

. au niveau d'une organisation, où la bureautique va être dispersée géographiquement, les différents morceaux étant reliés par des moyens de communication.

Au niveau individuel, que l'on qualifiera de local, cette collection de documents est privée, ou peut être considérée comme telle : seul un ou quelques usagers l'utilisent, et ont, a priori, toute liberté pour la gérer.

Au niveau de l'organisation, que l'on qualifiera de global, il existera certainement une collection ou une sous-collection de documents accessibles à un ensemble d'utilisateurs : cette collection est alors publique, et l'on disposera de beaucoup moins de liberté pour la gérer.

Au niveau local, la collection de documents considérée est centralisée ; au niveau global, la collection est répartie : on se trouve en présence d'un ensemble de données réparties et d'un système documentaire réparti. Un objectif essentiel de la construction d'un tel système est d'assurer la possibilité d'intégration progressive des différents composants du système : c'est le problème de la coopération qui est posé.

Aussi, les solutions choisies pour la gestion d'une collection privée, centralisée, ne pourront ignorer les problèmes posés par la gestion d'une collection publique et répartie.

Les pages qui suivent sont consacrées à l'examen de ces problèmes et à des propositions de solution.

4. DOCUMENTATION AUTOMATIQUE ET BUREAUTIQUE

Le classement n'est pas la moindre des activités de bureau. Des ouvrages y sont consacrés. Il semble même que l'on désigne sous ce terme l'ensemble des opérations de gestion d'une collection de documents.

A. BRAUMAN, dans [BRA 77] le définit ainsi :

- classer quoi ?

  Toujours le courrier, les documents comptables ou de gestion ; souvent de la documentation et des documents techniques ; quelquefois des échantillons ou des objets.

- classer pourquoi ?

  Parce que c'est une obligation légale et réglementaire, parce que c'est une source d'information, parce que c'est un moyen de protection.

- classer où ? classer comment ?

  Selon la fréquence d'utilisation (classement "courant" et classement "dormant"), séquentiellement ("il faut tenir à jour des fichiers de recherche") ou en dossiers ("il faut alors choisir un dossier principal... et mettre dans les autres des renvois...", "...critère autre que le critère de rangement, il faut disposer d'un moyen auxiliaire de recherche...").

On le constate, le but principal du classement est le rangement (ou le stockage, ou l'archivage) et on peut le définir comme le rangement "en fonction des recherches prévues et possibles".

Remarquons dès à présent, que des méthodes automatiques vont permettre de séparer/distinguer les fonctions de "classement logique" (c'est l'indexation), des fonctions de classement "physique" (c'est le rangement).

4.1. QUE FAIRE ?

Prenons l'exemple de l'organisation et de la gestion d'un secrétariat, donné à l'annexe A, et demandons-nous ce qu'il faudrait faire pour "informatiser" cette gestion.

Le minimum de services à offrir serait d'automatiser ce qui existe, c'est-à-dire :

- Implémenter les méthodes qui sont celles du classement manuel : un document est mis dans un ou plusieurs dossiers sur indication de l'utilisateur. Les documents pourraient être ordonnés par date d'introduction dans le système. La recherche se ferait par examen séquentiel des documents contenus dans un dossier, donc en offrant des services de listage des noms de dossier, et des "en-tête" des documents d'un dossier ("en-tête" désignant en toute généralité quelque chose permettant d'identifier le contenu d'un document).

- Faire une structure de données et des outils d'accès à cette structure, pour représenter les livres courrier départ/courrier arrivée (accès par clé = date), les annuaires clients, fournisseurs,... (accès par clé = nom), le remplissage de cette structure de données étant bien entendu manuel.

Ce qu'il serait possible de faire ensuite, serait d'offrir de nouveaux services, simplement par le fait que l'outil dont on dispose permet, sur les mêmes données, d'en faire plus. C'est à dire :

- Faire comme précédemment du classement manuel et représenter les livres, annuaires, etc...

- Mais voir tout cela comme une mini-base de données, gérée par un mini-système de gestion de base de données, c'est-à-dire offrir de nouvelles fonctions d'accès aux documents, par date d'arrivée, de départ, nom d'expéditeur, de destinaire, par identificateur, etc...

Peut-on faire plus ? Une remarque s'impose : si la saisie des données est manuelle, il est peu probable que les documents entiers soient saisis et fassent aussi partie des données. Un tel système, sans doute fortement interactif, demandera à l'utilisateur d'introduire les données nécessaires.

Par contre, si la saisie est automatique (ou plutôt si l'on n'a pas besoin de le saisir à nouveau), c'est-à-dire qu'on dispose de documents dans une forme manipulable par une machine, alors, selon le "formatage" employé, cette machine sera plus ou moins capable d'extraire des documents un certain nombre de choses, afin d'initialiser les structures de données citées précédemment.

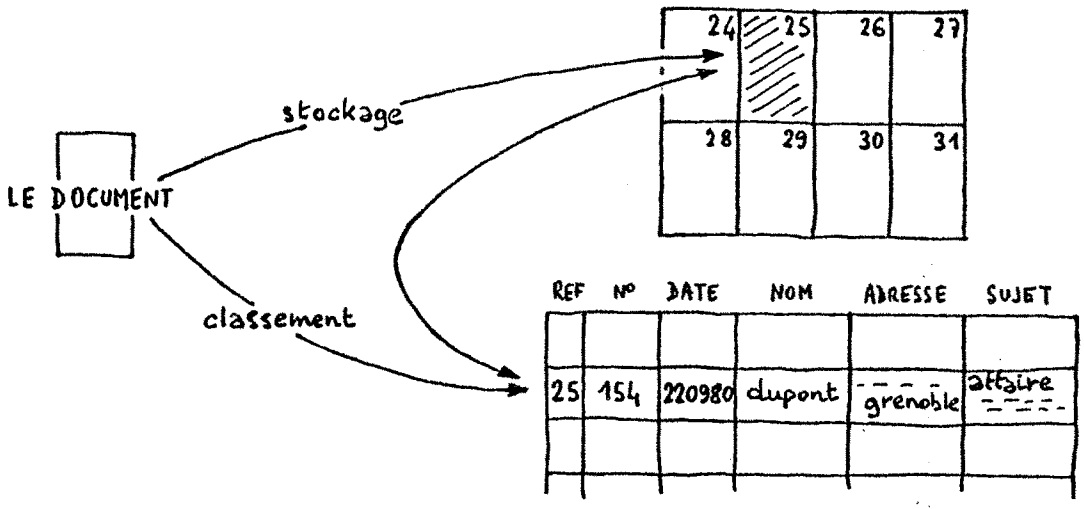
A ce point, les systèmes que l'on vient de suggérer, "ne sont que" des systèmes de gestion de bases de données ; si, par contre, ou si, en plus, on dispose des textes (ou d'une partie des textes) des documents, alors on peut avoir affaire à des systèmes documentaires.

4.2. DES MACHINES POUR CLASSER

Les opérations de classement, ou manuelles, ou telles qu'on vient de les proposer, ont en commun ce qui caractérise l'approche prise dans la conception des bases de données, à savoir la modélisation du monde réel ; on ne peut avoir une appréhension de l'objet, mais on le caractérise par un certain nombre d'attributs. Nous l'appellerons approche base de donnéees.

Ainsi, dans l'exemple du secrétariat cité en annexe, un document est caractérisé par les attributs numéro, date, nom, adresse, sujet, et l'existence d'un document (réel) est caractérisée dans la base de données (que représentent les livres courrier départ/courrier arrivée) par l'existence d'une occurrence de la relation DOCUMENT (numéro, date, nom, adresse, sujet).

Ceci peut s'illustrer de la manière suivante :



Le classement consiste à attribuer quelques "caractéristiques" au document ; ce dernier est rangé/stocké ; les caractéristiques (auxquelles on ajoute une référence vers le document, qui sert de lien entre le classement logique et le classement physique) sont stockées dans la base, et deviennent les seules choses que l'on connaisse du document.

A l'opposé, une autre approche, que nous appellerons approche traitement de textes consiste à considérer le document comme un (ou plusieurs) texte (s) ; dans son acceptation la plus simple, un texte est une chaîne de caractères sur laquelle on peut faire des opérations bien spécifiques (recherche de sous-chaînes).

Certaines sous-chaînes sont appelées des mots et sont considérées comme autant de caractéristiques de documents qui (possèdent des textes qui) possèdent ces mots. Certains caractères ont des rôles de délimiteur, et d'autres sous-chaînes sont appelées des phrases.

Les opérations de recherche que l'on peut effectuer permettent de retrouver les textes qui possèdent un mot ou une portion de mot, une suite de mots, un ensemble de mots, dans une même phrase ou non, etc...

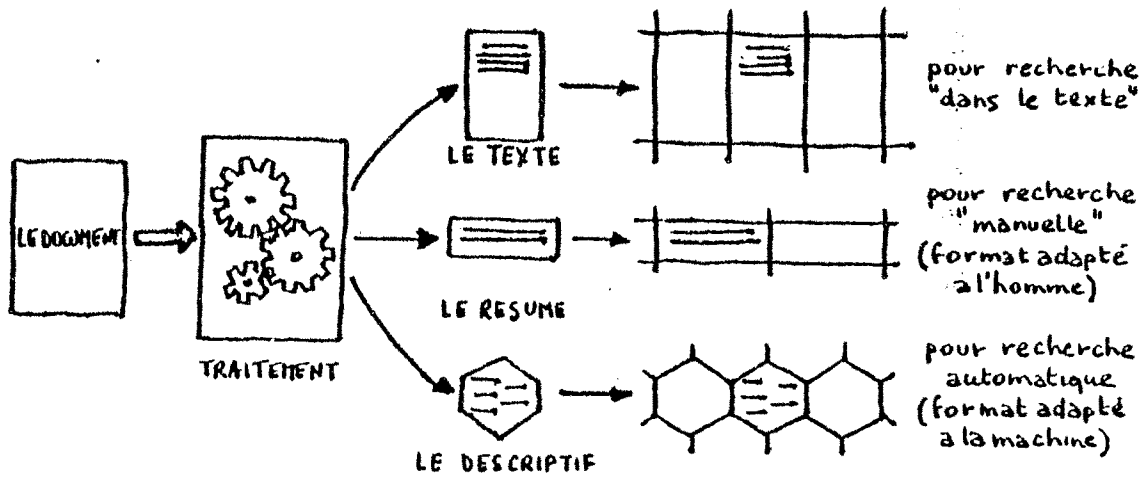
Une troisième approche, dont on pourrait dire qu'elle se situe "entre" les deux précédentes, et que nous appellerons approche documentaire, nous semble différente par la démarche employée ; même si, dans certains cas, les méthodes employées sont les mêmes, elle correspond à une approche différente, qu'il nous semble important de distinguer.

Il ne s'agit plus là de caractériser le document par des éléments qui lui sont souvent extérieurs (numéro, date, ...), mais d'essayer d'en avoir une appréhension totale en tant que véhicule d'information.

Malheureusement, la forme originale du document n'est pas forcément la plus commode à manipuler : on remplace alors le document par une autre entité de maniement plus facile :

- c'est ce qui est fait quand on produit le résumé d'un document,
- c'est aussi le résultat d'une opération d'indexation qui produit le descriptif d'un document,

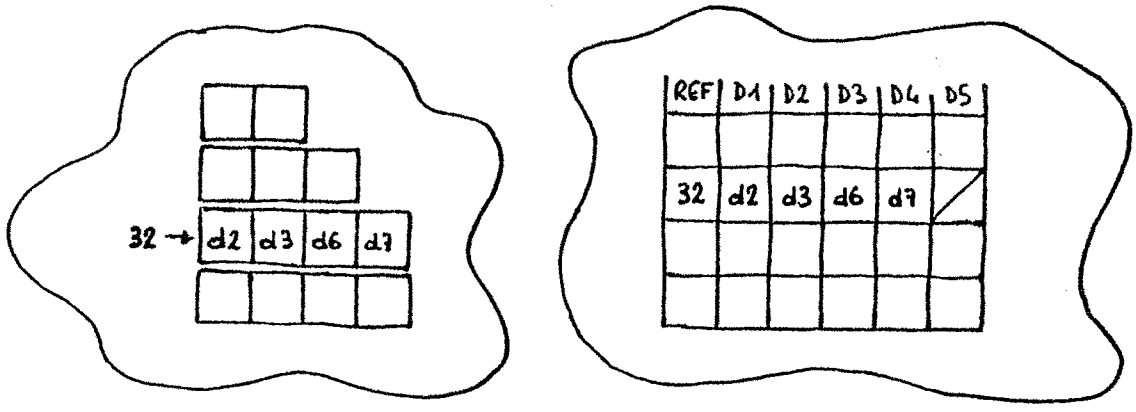
comme illustré ci-dessous :



Remarque 1 :

Souvent, les approches documentaires et bases de données sont confondues, parce que les entités attachées aux documents ont la même forme, et que, par conséquent, les opérations qu'il est possible d'effectuer sur ces entités sont les mêmes : c'est le cas, par exemple, des descriptifs qui sont des listes de descripteurs, et d'une recherche par présence/absence de descripteurs dans les descriptifs.

Exemple :



"BASE DOCUMENTAIRE"
= ensemble de descriptifs

"BASE DE DONNÉES"
= relation document

Dans l'approche documentaire, avec une question q = $\{d2, d3\}$, une opération de comparaison qui est l'intersection non vide, la réponse comprend le document 32 puisque :

$$\{d2, d3, d6, d7\} \cap \{d2, d3\} \text{ n'est pas vide.}$$

Dans l'approche base de données, la réponse à la question comprend le document 32, puisque :

project (select (DOCUMENT, (D1 = d2 $\vee$ d3) $\vee$ (D2 = d2 $\vee$ d3) $\vee$ (D3 = d2 $\vee$ d3) $\vee$ (D4 = d2 $\vee$ d3) $\vee$ (D5 = d2 $\vee$ d3)), NUM) comporte le numéro 32.

Cet exemple montre également que dans certains cas, un système documentaire peut être implémenté sur un SGBD.

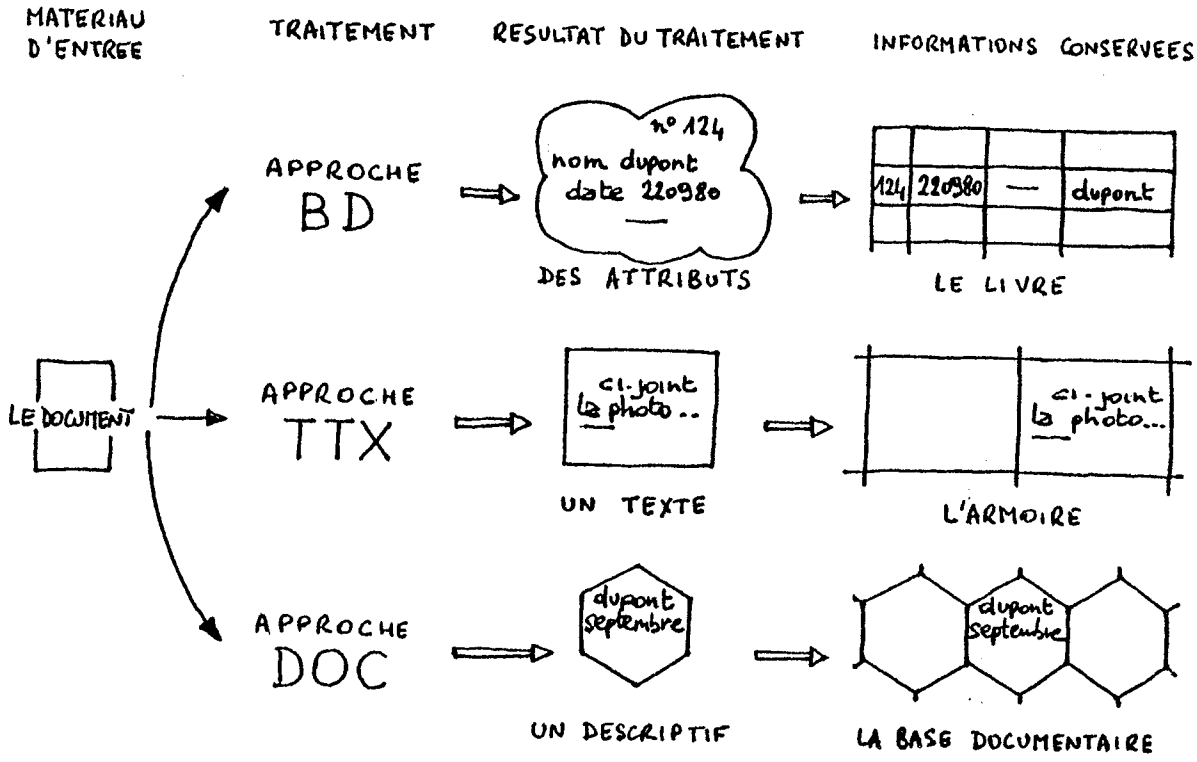
Remarque 2 :

Les approches peuvent également être confondues, ou mal distinguées, du fait d'une mauvaise définition des textes d'un document. Prenons par exemple l'en-tête des documents : pour du courrier, on trouvera des informations semblables à celles qui ont servi à envoyer ce document (c'est évident pour les enveloppes à fenêtre transparente... le nom du destinataire figure sur l'en-tête du document, et non pas sur l'enveloppe) et ne font pas, à proprement parler, partie du contenu du document.

Mais cette en-tête pourra être gérée de différentes manières, suivant la façon dont on la considère ; par exemple :

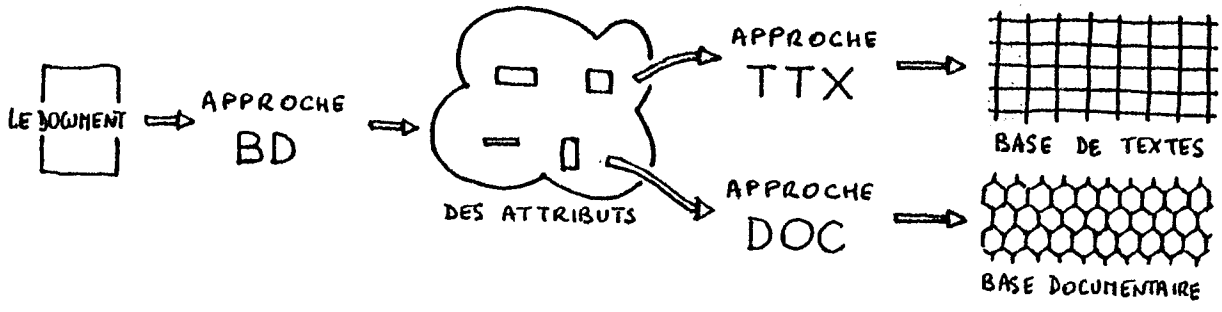
- une approche bases de données la considérera comme un attribut simple,
- une approche traitement de textes la considérera comme une chaîne de caractères,
- une approche documentaire la considérera comme un descriptif (une liste de mots).

Ce qui est important, c'est que, à partir du même matériau de base, le document, ces trois approches fournissent des produits différents, c'est-à-dire des outils de classement et de recherche différents.



4.3. QUE CHOISIR ?

Ces trois approches ne sont pas incompatibles, bien au contraire, et il est probable qu'un "bon" système de classement fera appel à ces différentes méthodes. En particulier, certains attributs obtenus par une approche base de données seront des textes et pourront être traités par une approche traitement de textes ou documentaire.



Les avantages particuliers de chacune de ces approches seront fortement liés à la richesse des structures de données employées, et à la façon dont ces données sont acquises.

Toutefois :

- l'approche base de données favorise la gestion du "contenant" au détriment du "contenu",

- l'approche traitement de textes gère le "contenu" mais n'offre pas véritablement d'outils de classement.

L'approche documentaire semble combler ces insuffisances et offrir des services nouveaux, en proposant une assistance efficace aux fonctions de classement et de recherche des documents. C'est de cette approche dont nous parlons dans les pages qui suivent.

5. INDEXATION, CLASSIFICATION, CLASSEMENT

5.1. DEFINITIONS

Un ensemble d'objets classés est un ensemble d'objets arrangés en classes d'affinités, c'est-à-dire tel qu'un objet classé est "plus ressemblant" aux objets des classes auxquelles il appartient qu'aux objets des classes auxquelles il n'appartient pas.

A ce point, il semble nécessaire de préciser des termes déjà employés mais non définis précisément :

- Une classification est une "distribution systématique en diverses catégories, d'après des critères précis".

- Le classement est "l'action de classer" [petit Larousse].

Ces deux définitions illustrent bien deux aspects complémentaires de méthodes employées en documentation automatique :

- L'aspect classement automatique, dont le but est de proposer, à partir d'une classification existante, une méthode pour classer automatiquement un objet.

- L'aspect classification automatique qui consiste, à partir d'un ensemble d'objets, à déterminer les classes "sous-jacentes" ou "qui existent réellement" dans cet ensemble d'objets.

Classer un objet consiste à l'examiner et à l'assigner à la ou les classes considérées comme les plus appropriées. Faire cette opération de façon automatique implique de pouvoir apprécier une "ressemblance" entre objets et classes.

Déterminer automatiquement une classification nécessite également de pouvoir apprécier des "ressemblances" entre objets.

Pour cela, il est donc nécessaire de pouvoir posséder l'information sous une forme manipulable permettant de faire de telles comparaisons : lorsque les objets sont des documents, c'est le but de l'opération d'indexation qui est d'associer un descriptif à un document. Autrement dit, les méthodes de classement et de classification seront automatiques, si l'on possède, en amont, une méthode d'indexation automatique.

5.2. INDEXATION AUTOMATIQUE

La méthode "idéale" d'indexation automatique n'existe pas : il est probable qu'elle serait (sera ?) fondée sur des méthodes d'analyse des langues naturelles.

Des méthodes d'indexation basées sur l'analyse des langues naturelles existent : elles ne sont pas "idéales" et pas entièrement automatiques.

Des méthodes d'indexation automatiques "simples" existent : l'idée de base est que : "dans un texte, les unités de sens évidentes sont les mots ; il est clair que certains mots ont une forte tendance à apparaître dans les textes traitant d'un sujet donné : ces sont des mots-clés pour le sujet" [PAI 77]. Ces mots-clés sont des mots individuels ou des groupes de mots ou expressions : c'est pourquoi nous avons préféré le terme descripteur à celui de mot-clé.

L'opération d'indexation a pour but de déterminer un descriptif que nous avons défini précédemment comme structure de descripteurs. Dans la mesure où la méthode d'indexation consiste à déterminer les mots significatifs d'un texte, sans prendre en compte la structure de ce texte, il est clair que les descriptifs obtenus seront des listes de descripteurs, éventuellement affectés de poids, c'est-à-dire de vecteurs de dimension le nombre de descripteurs utilisables, où le i-ème élément représente l'importance du i-ème descripteur dans ce descriptif.

Toutes ces méthodes sont forcément dérivatives au moins dans un premier temps : elles consistent à extraire les descripteurs du texte du document à indexer. En fait, il existe deux variantes :

- celle qui consiste à examiner un texte afin d'y reconnaître des descripteurs bien spécifiques,
- celle qui consiste à examiner un texte afin d'y détecter des descripteurs possibles.

La première méthode nécessite de connaître a priori l'ensemble des descripteurs à reconnaître, et pose le problème général de la construction des "vocabulaires d'indexation" ou "langages documentaires". La seconde méthode nécessite de posséder des moyens de sélection de tels descripteurs (la construction plus ou moins automatique de langages documentaires fait appel à des méthodes de classification automatique nécessitant elles-mêmes l'utilisation de la seconde méthode d'indexation).

Le problème reste alors de savoir attacher une "valeur" à ces descripteurs : dans le premier cas afin de représenter leurs importances relatives, dans le second cas, la détermination de cette "valeur" étant la base même du processus de sélection.

Cette valeur dépend évidemment de l'apparition de ce descripteur dans un document donné (et c'est l'idée originale de P. LUHN, vers 1950, le précurseur dans ce domaine), mais aussi de son apparition au sein de la collection toute entière :

- "Les mots importants apparaissent souvent" (P. LUHN).
- "Les mots rares sont plus importants pour la recherche que les mots fréquents" (K.SPARCK-JONES).
- "Un "bon" descripteur doit identifier le contenu d'un document, et en même temps distinguer un document d'un autre" (G. SALTON).

G. SALTON mentionne principalement trois idées, basées sur des calculs statistiques [SAL 75] :

- La première consiste à favoriser les descripteurs apparaissant avec une haute fréquence dans quelques documents et une fréquence franchement basse dans l'ensemble de la collection.

- La seconde prend pour hypothèse qu'un "bon" descripteur est celui qui rend les documents aussi dissemblables que possible les uns des autres.

- La troisième consiste à prendre en compte la satisfaction des utilisateurs quant à l'utilisation d'un descripteur : un "bon" descripteur est celui qui apparait dans une question et dans les documents jugés pertinents.

C.D. PAICE fait état de méthodes non-statistiques, basées sur l'observation de la position dans le texte ; par exemple les mots présents dans la première phrase et la dernière phrase d'un paragraphe, ou les mots suivants d'autres mots identifiés comme prépositions,... [PAI 77].

Pour que ces méthodes soient efficaces, il est souhaitable de faire un traitement préliminaire du texte à indexer, qui consiste à "normaliser" les mots de ce texte, c'est-à-dire d'abord éliminer les mots dont on sait qu'ils sont sans intérêt, puis ensuite ramener à une même forme des mots différents représentant en fait le même concept.

5.3. CLASSIFICATION AUTOMATIQUE

Une collection de documents classés, c'est-à-dire une classification des documents peut aider à la recherche ; le processus de recherche consiste alors à déterminer la ou les classes dans lesquelles peuvent se trouver les documents cherchés, puis à examiner chacun des documents appartenant à ces classes. De cette façon, la définition d'une classification de documents peut être envisagée comme la définition d'un ensemble de question "pré-définies", la réponse à chacune de ces questions étant les documents de la classe correspondante.

Cette idée est la base d'une catégorie d'utilisation des méthodes de classification automatique en documentation automatique : le problème était de minimiser le temps de calcul des similarités (descriptif de) question/(descriptif de) document dans un système documentaire, donc de minimiser le nombre de documents concernés par cette opération. La solution est alors de répartir les documents "semblables" dans différentes classes, non pas tant parce qu'ils traitent du ou des mêmes sujets, mais parce qu'il y a une forte probabilité pour qu'ils répondent à la même question. Chaque classe possède un descriptif "représentant" (ou centre de gravité, ou "centroïde" [SAL 77]). La réalisation de l'opération de recherche consiste alors à calculer les similarités question/document mais seulement pour les documents des classes telles que la similarité question/"représentant" aura été jugée satisfaisante (expérimentation dans SMART [SAL 71]).

Par nature, classer un ensemble d'objets consiste à les répartir en plusieurs groupes ; le processus de classification est essentiellement un processus de division, et le plus souvent, de division par étapes successives. Les principaux problèmes à résoudre sont alors de savoir sur quel (s) critère (s) diviser, et jusqu'à quand (nombre de classes, tailles des classes...).

A priori, la détermination d'une classification de documents ne nous concerne pas, tout au moins pour des raisons de minimisation de temps de calcul, eu égard au volume des données manipulées dans le contexte bureautique. Par contre, les méthodes de détermination de classification du vocabulaire, seront mises en oeuvre pour réaliser des outils d'aide au choix des descripteurs.

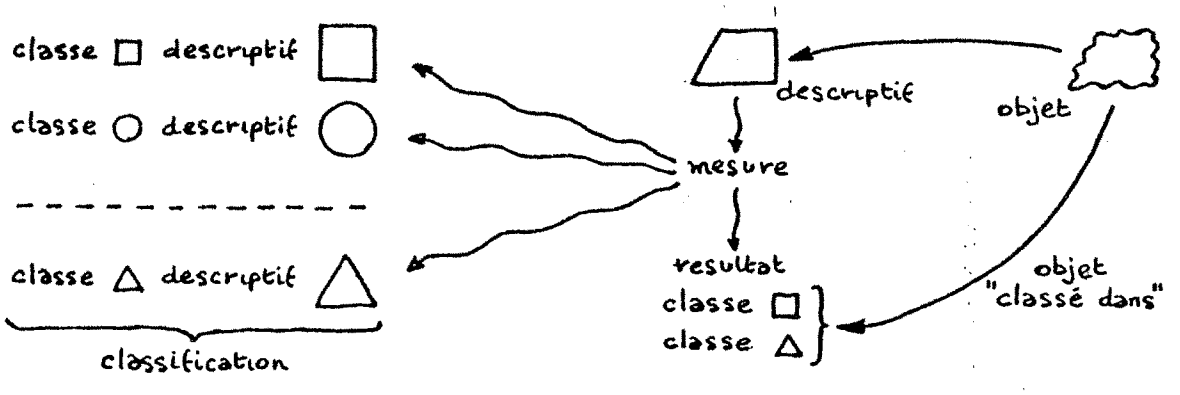
5.4. CLASSEMENT AUTOMATIQUE

En toute généralité, un algorithme de classement nécessite que soit définie une classification, c'est-à-dire un ensemble de classes, chacune d'elles étant définie par un descriptif.

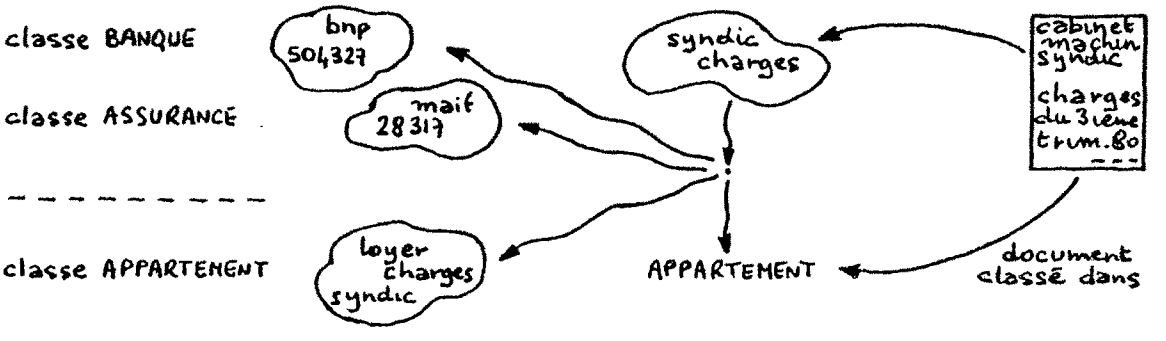
Le classement consiste alors à :

- construire le descriptif de l'objet à classer,
- calculer la similarité entre le descriptif de l'objet à classer, et le descriptif de chaque classe,
- assigner l'objet à la ou les classes pour la ou lesquelles la similarité est jugée satisfaisante.

Ceci peut s'illustrer de la façon suivante :



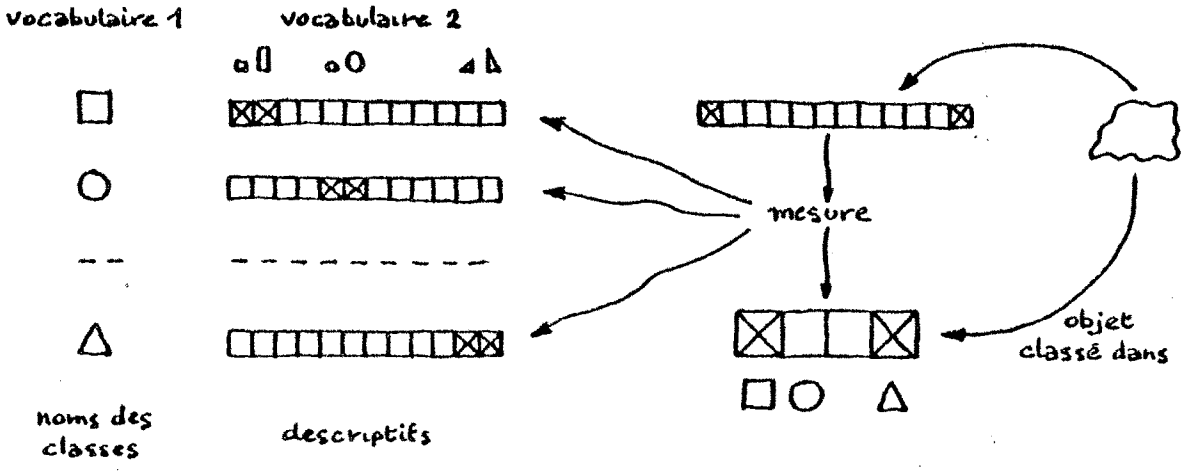
Si les objets sont des documents, considérés comme des textes, les descriptifs des ensembles de mots, et la mesure d'intersection ensembliste, on aura par exemple :



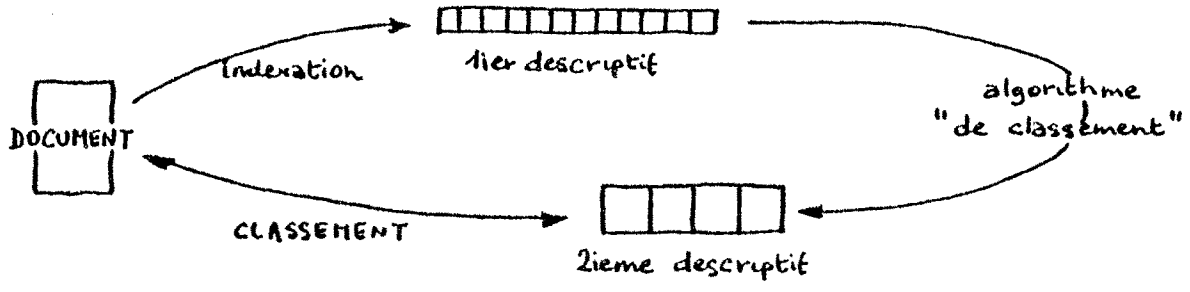
5.5. INDEXATION = CLASSEMENT

Dans le cas de documents, les descriptifs des documents et des classes sont des vecteurs sur un vocabulaire (appelé vocabulaire 1) ; chaque classe possède un nom : les noms des classes forment un second vocabulaire (vocabulaire 2). Le résultat du classement peut être représenté par un vecteur sur ce second vocabulaire.

On peut alors illustrer ainsi le classement d'un objet :



On le voit, l'application de cet algorithme a pour effet d'associer un second descriptif à l'objet. En particulier, si les objets sont des documents, cet algorithme, appelé "de classement" est également un algorithme d'indexation, puisqu'il permet d'associer un descriptif à un document.



Si l'on reprend, en "bon français", l'algorithme de classement, il devient algorithme d'indexation en disant :

- construire le descriptif de l'objet à classer,

- calculer la similarité entre le descriptif de l'objet à classer et le descriptif de chaque classe,

- assigner à l'objet, le nom de la ou les classes, pour la ou lesquelles la similarité est jugée satisfaisante.

Remarquons que l'on aurait également un algorithme d'indexation en disant : "assigner à l'objet, le descriptif de la ou des classes pour la ou lesquelles la similarité est jugée la plus grande" (en ayant bien sûr défini la composition des descriptifs).

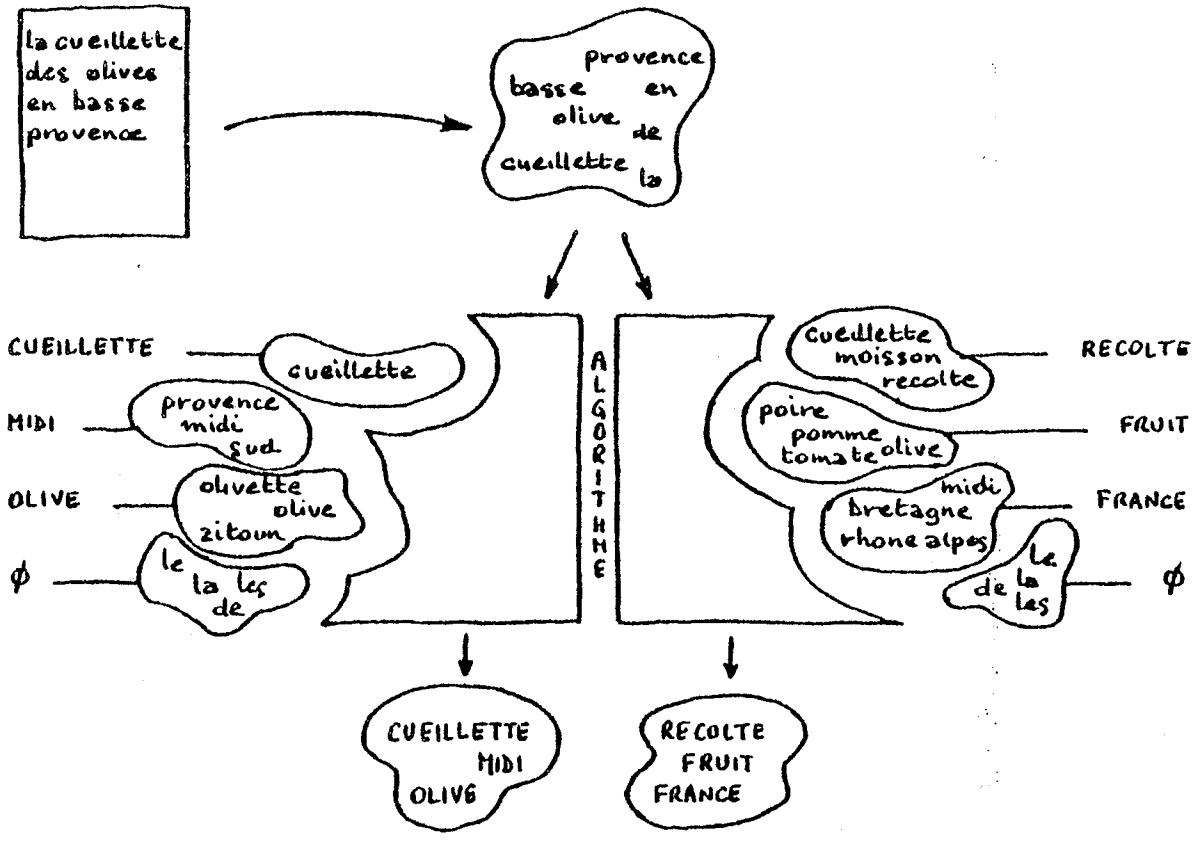
Ces méthodes d'indexation sont dites indexation par assignation (par opposition à indexation dérivative).

Remarquons à nouveau ce que nous avons déjà mentionné, à savoir la confusion entre les termes classement et indexation (par assignation) du fait de la similitude des méthodes employées. On parle du classement lorsqu'on associe des documents aux classes (documents $\rightarrow\!\!\!\!\rightarrow$ classes : on pense au rangement), et l'on parle d'indexation lorsqu'on associe les (noms des) classes au document ((nom des) classes $\rightarrow\!\!\!\!\rightarrow$ documents) ; il s'agit en fait de la même relation entre classes et documents (documents $\leftarrow\!\!\!\!\rightarrow\!\!\!\!\rightarrow$ classes).

Le résultat de la mise en oeuvre d'un tel algorithme va dépendre évidemment de la classification utilisée, mais aussi du mode de calcul utilisé pour calculer la similarité entre le descriptif de l'objet et le descriptif de chaque classe, et de la façon de juger "satisfaisante" la similarité entre les descriptifs.

5.6. REMARQUES

Soit, par exemple, deux exemplaires d'un algorithme de classement utilisant comme descriptif d'entrée tous les mots du texte, comme mesure l'intersection ensembliste non vide, et chacun une classification de vocabulaire : la première représentant des ensembles de mots significatifs et synonymes et la seconde représentant des ensembles de mots spécifiques. Selon la classification employée, les résultats seront différents comme l'illustre la figure ci-dessous :



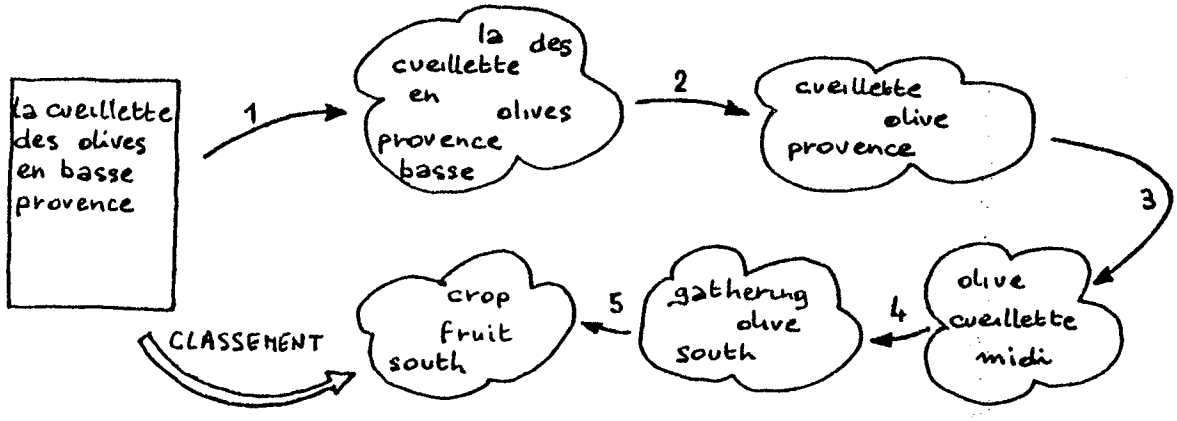
Remarquons là aussi que nous avions tendance, pour l'algorithme 1, à parler d'indexation, et, pour l'algorithme 2, à parler de classement ; il serait peut-être plus simple de parler de projection d'un objet (ici un ensemble de mots) sur une classsification (ici, un autre exemple de mots).

Ce qu'il est plus intéressant de remarquer c'est que, en poussant le raisonnement à l'extrême, le classement d'un document peut se représenter par une succession de projection de descriptif sur des vocabulaires différents.

Par exemple, on pourrait imaginer la suite des opérations suivantes :

- 1 former descriptif 1 = tous les mots du documents
- 2 projeter descriptif 1 sur vocabulaire des mots significatifs → descriptif 2
- 3 projeter descriptif 2 sur vocabulaire de synonymes → descriptif 3
- 4 projeter descriptif 3 sur vocabulaire "traduit" → descriptif 4
- 5 projeter descriptif 4 sur vocabulaire de mots-reliés → descriptif 5

...on l'arrête là...



En poursuivant le raisonnement, on observe que les différentes opérations utilisées en documentation automatique peuvent être représentées par la mise en oeuvre d'un algorithme de classement, ou plus précisément, de la projection d'un objet sur une classification.

Exemple 1 :

Prenons comme classification une collection de (descriptifs de) document (c'est une classification de descripteurs) ; prenons comme objet à classer une question (c'est un ensemble de descripteurs, un descriptif) ; le résultat du classement sont les documents qui répondent à la question (c'est la projection d'une question sur un vocabulaire (d'identificateurs) de documents) : cette opération est une "recherche retrospective".

Exemple 2 :

Prenons comme classification une collection de (descriptif de) questions (c'est une classification de descripteurs) ; prenons comme objet à classer un document (c'est un ensemble de descripteurs, un descriptif) ; le résultat du classement sont les questions auxquelles ce document répond (c'est la projection d'un document sur un vocabulaire (d'identificateurs) de questions) : cette opération est une "diffusion sélective de l'information".

Ces remarques sur l'identité des algorithmes de recherche et le classement, nous suggèrent que la machine qui réalise ces fonctions, doit pouvoir être décrite formellement de façon simple et précise, c'est l'objet du prochain chapître.

6. LA MACHINE CLASSEMENT

On appelle machine classement l'ensemble des fonctions nécessaires à la gestion d'une collection de documents sous l'aspect du classement et de l'approche documentaire précédemment définis.

Les deux caractéristiques qui nous paraissent importantes sont les suivantes :

- La machine classement est une machine individuelle (ou quasi individuelle, c'est à dire dédiée au plus à un groupe de personnes).

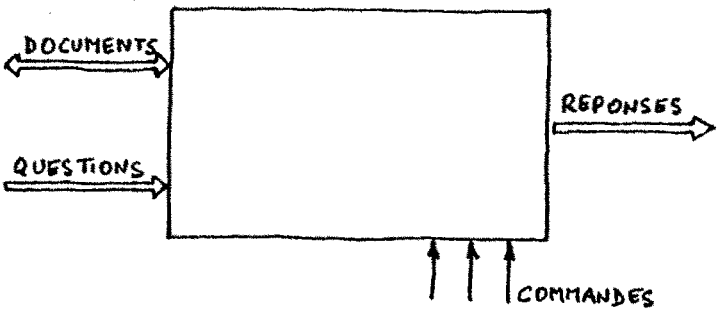
- La machine classement a des possibilités d'interaction avec l'utilisateur.

La première caractéristique est une évidence dans le contexte dans lequel on se situe : la bureautique. La seconde existe parce que c'est possible : l'utilisateur est toujours (ou presque) présent lorsque la machine fonctionne, et parce que c'est souhaitable pour améliorer l'efficacité et pour prendre des décisions que la machine ne pourra pas prendre ou pour donner un complément d'information aussi bien dans le sens homme-machine qu'inversement.

On propose de définir l'architecture générale de cette machine, non pas comme un modèle de structure mais plutôt comme un outil d'analyse des différentes fonctions et des choix pour leur réalisation.

6.1. ARCHITECTURE GÉNÉRALE

Le schéma général est le suivant :



Examinons les différentes entrées/sorties de la machine, soit les documents, les questions, les réponses et les commandes.

- Les documents proviennent d'un service de traitement de texte (saisie manuelle ou non), directement, ou par l'intermédiaire d'un service de courrier électronique. Ils sortent de la machine pour aller vers des services d'archivage ou de destruction (poubelle).

- Les questions, opérations de recherche, proviennent de l'utilisateur et les réponses, résultats de recherches, lui sont destinées.

- Les commandes proviennent de l'utilisateur, d'un autre service ou d'une autre machine.

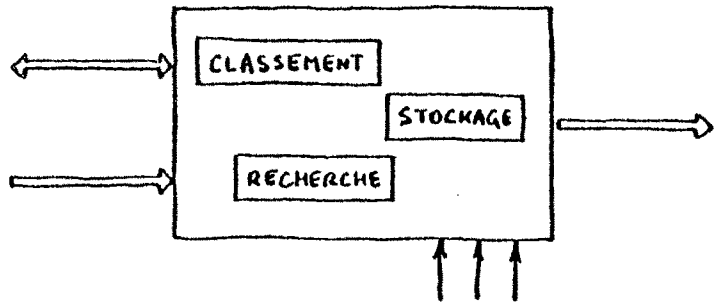
Quelles sont ces commandes ? Elles sont évidemment liées aux fonctions à réaliser. On peut distinguer :

- Les commandes d'entrée et de sortie d'un document.

- Les commandes liées aux fonctions de classement d'un document.

- Les commandes liées aux fonctions d'interrogation et de recherche de documents.

Nous ajouterons qu'il existe certainement des fonctions de stockage. La figure ci-dessous donne le schéma général montrant les fonctions, réparties en trois grandes catégories.



Examinons maintenant ces différentes catégories de fonctions.

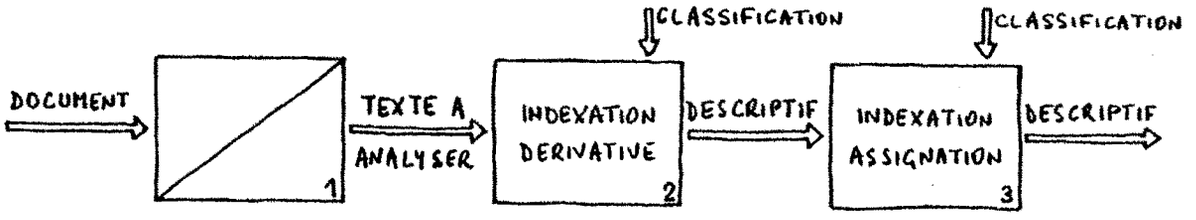
6.2. LA MACHINE A CLASSER

La machine à classer est celle qui assure les fonctions de classement proprement dites. On lui donne en entrée un document et elle fournit en sortie un document classé.

L'entrée peut être de différentes formes : par exemple le document lui-même, ou une référence de document et un texte réduit,... De la même façon, la sortie peut être le document et le descriptif ou une référence de document et le descriptif ou encore le descriptif seul.

A priori, il n'y a pas de commande particulière pour cette machine, l'entrée d'un document dans la machine de classement entrainant son classement.

On a vu que l'opération de classement comportait deux étapes : une étape d'indexation dérivative, suivie d'une étape d'indexation par assignation. L'opération d'indexation dérivative se fait par l'analyse des textes composant le document (on indexe un document, on analyse un texte, indexation = composition (analyse (texte de document)) [ABH 79]). Il doit donc exister une étape préliminaire permettant de passer du document au texte à analyser. On obtient alors le schéma suivant :



Remarque :

La machine (1) peut se décomposer en deux parties, une qui permet de passer du document au (x) texte (s) du document, et l'autre permettant de passer des textes du document au texte à analyser :

document ⟶ texte (s) du document ⟶ texte à analyser

Exemples :

- de machine (1) : texte à analyser = résumé du document
- de machine (2) : descriptif = mots du texte d'une longueur supérieure à trois caractères (classification mots vides/mots pleins)
- de machine (3) : descriptif = les synonymes des descripteurs du descriptif d'entrée (classification = classes de synonymes).

Les machines (2) et (3) sont de même nature : ce sont des machines à indexer ; elles utilisent des classifications. Quelle que soit la méthode utilisée, il est peu probable que celle-ci soit entièrement satisfaisante, soit parce que la méthode n'est pas parfaite, soit parce que le matériau d'entrée est de mauvaise qualité. Il est alors souhaitable que l'utilisateur intervienne aux endroits où il peut le faire. On peut penser à l'algorithme et aux paramètres fournis à ceux-ci.

- Sur l'algorithme :

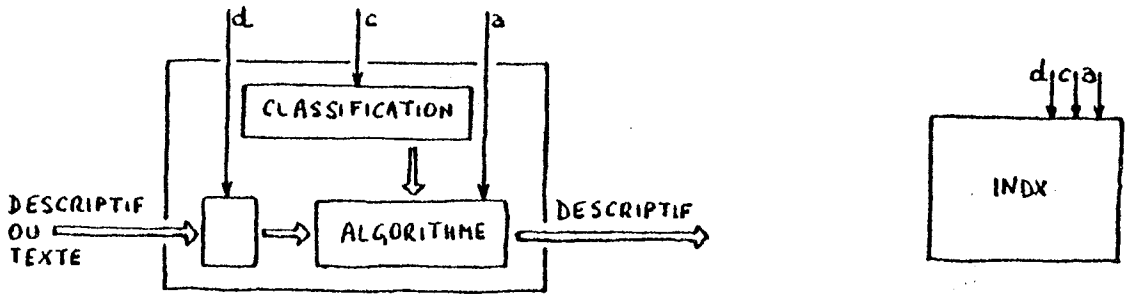
  Il s'agit en fait d'une modification des paramètres d'exécution de l'algorithme (par exemple la valeur du seuil de comparaison).

- Sur les données d'entrée de la machine :

  . Le texte ou les descriptifs : ceux-ci peuvent être utilisés ou insuffisants.

  . La classification : elle n'est pas parfaite et l'intervention de l'utilisateur peut permettre de l'améliorer (mécanisme d'apprentissage).

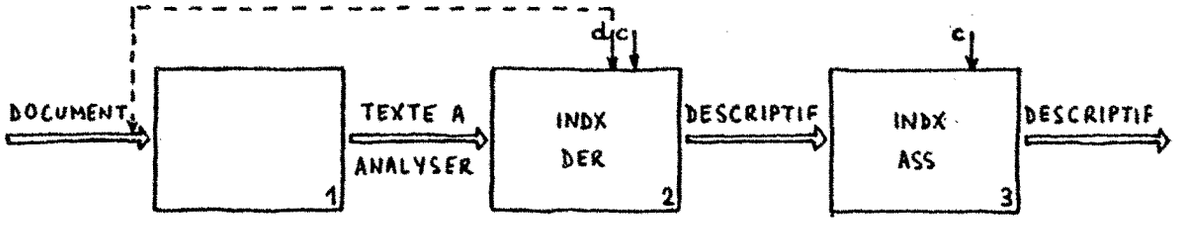
On peut représenter la machine à indexer de la façon suivante :



Ainsi, au sein de la machine à classer, l'intervention de l'utilisateur sur les algorithmes des machines à indexer est impossible par hypothèse. Une méthode a été choisie comme étant la meilleure et son changement n'est pas du ressort de l'utilisateur. On pourrait penser à modifier certains paramètres concernant la mesure de similarité, mais des impératifs de reclassement partiel impliquent d'avoir une méthode stable au cours du temps.

Pour des raisons analogues (reclassement automatique), il serait souhaitable (voire très important) que le processus complet puisse être répété de manière automatique. Donc, toute intervention de l'utilisateur sur le descriptif intermédiaire est à proscrire et d'autre part, il peut être préférable d'intervenir sur le document lui-même au niveau de la machine (1) (par exemple pour pallier à un texte illisible).

Le nouveau schéma est alors le suivant :



6.3. LA MACHINE A CHERCHER

La machine à chercher est celle qui reçoit une question en entrée et délivre une réponse en sortie.

La question peut se présenter sous différentes formes dont les plus courantes sont :

- des mots reliés par des opérateurs booléens,
- oui simplement une liste de mots ("langue naturelle").

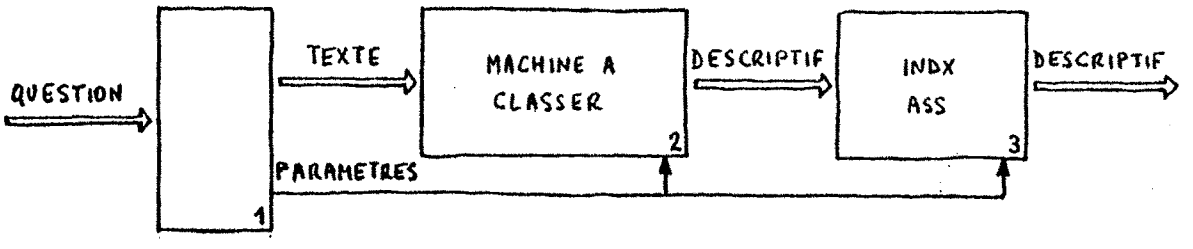
La première forme consiste en fait à donner un descriptif (liste de descripteurs) et l'opération de comparaison (représentée ici comme une liste d'opérateurs booléens) qui est à effectuer. La seconde forme peut être considérée soit comme une expression booléenne de descripteurs ("ou" inclusif implicite par exemple), soit comme un descriptif (liste de descripteurs). On peut alors être amené à fournir certains paramètres à l'algorithme qui effectue la comparaison entre les descriptifs (méthode de comparaison, seuil de succès, taille de la réponse...).

Ce que nous avons appelé "question" comporte donc un texte et des paramètres. La machine à chercher comportera alors une machine permettant de passer de la question aux textes et paramètres.

Le texte de la question doit être analysé, c'est-à-dire qu'il faut, à partir de ce texte, produire un descriptif analogue à ceux des documents classés. C'est le rôle d'une (partie de) machine à classer à celle décrite précédemment.

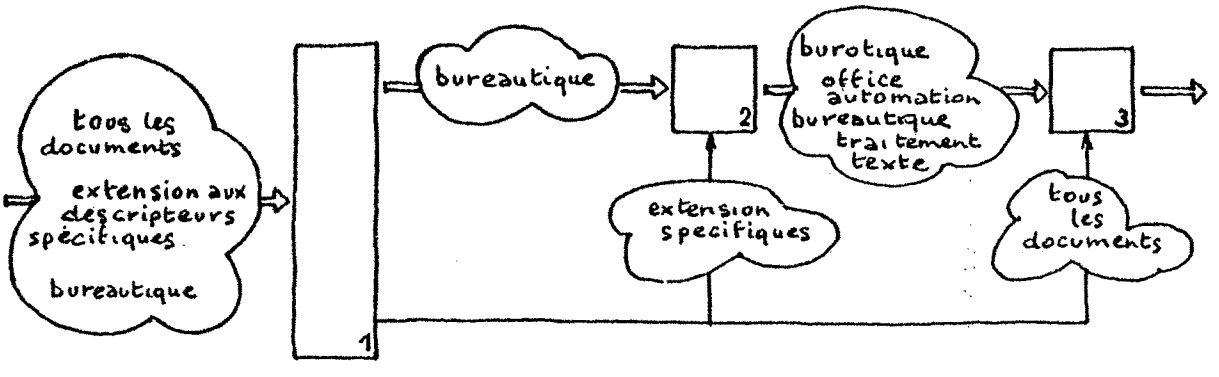
La recherche proprement dite consiste à effectuer des comparaisons entre le descriptif de la question et le descriptif de chacun des documents. Le résultat d'une comparaison indique si la question est proche ou non du document considéré. On obtient ainsi un ensemble de documents plus ou moins proches, c'est-à-dire un vecteur sur le vocabulaire (d'identificateurs) des documents. On retrouve là un algorithme de classement ou d'indexation par assignation : chaque document est une classe et chaque classe possède un descriptif.

Ainsi, la machine à chercher comportera une machine à classer suivie d'une machine à indexer par assignation, et peut être représentée par le schéma suivant :



Exemple :

On désire obtenir les documents parlant de bureautique et de domaines associés. On pourrait alors avoir :



Comme précédemment, l'utilisateur pourra intervenir sur les différentes machines à indexer :

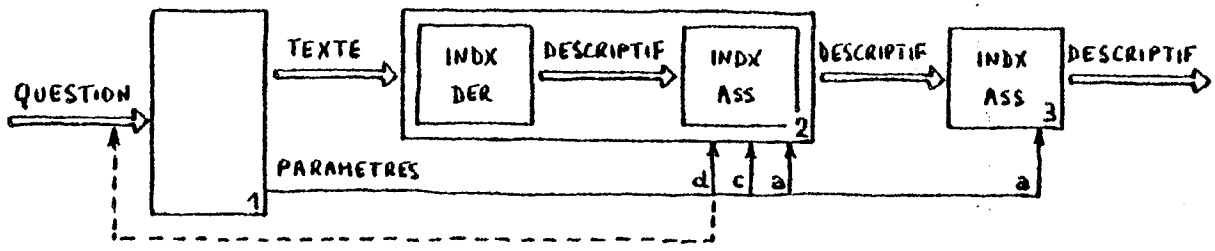
- sur certains paramètres d'exécution des algorithmes tels que la classification à utiliser et la façon de s'en servir (machine 2) ou la façon de calculer la similarité ou le seuil (machine 3).

- Sur les textes ou descriptifs en entrée, en particulier le texte de la question (machine 2).

- Sur les classifications utilisées : évidemment pas sur celle utilisée par la machine 3 (ce sont les descriptifs des documents), mais sans doute et sous certaines conditions sur celle (s) utilisée (s) par la machine 2 afin de mettre en oeuvre la aussi un mécanisme d'apprentissage.

Il faut remarquer que ces interventions seront faites directement sur les différentes machines (par exemple pour la modification d'une classification), ou indirectement par une action sur l'entrée de la machine à chercher, c'est-à-dire par une re-formulation de la question. On retrouve là la notion de processus de recherche : par définition une "bonne" formulation de la question amènera une bonne réponse.

Ainsi, la machine à chercher sera représentée comme ci-dessous :



6.4. LA MACHINE A STOCKER

La machine à stocker est celle qui assure les fonctions de stockage nécessaires à chacune des machines que nous venons de décrire. On trouvera en particulier le stockage des différentes classifications utilisées.

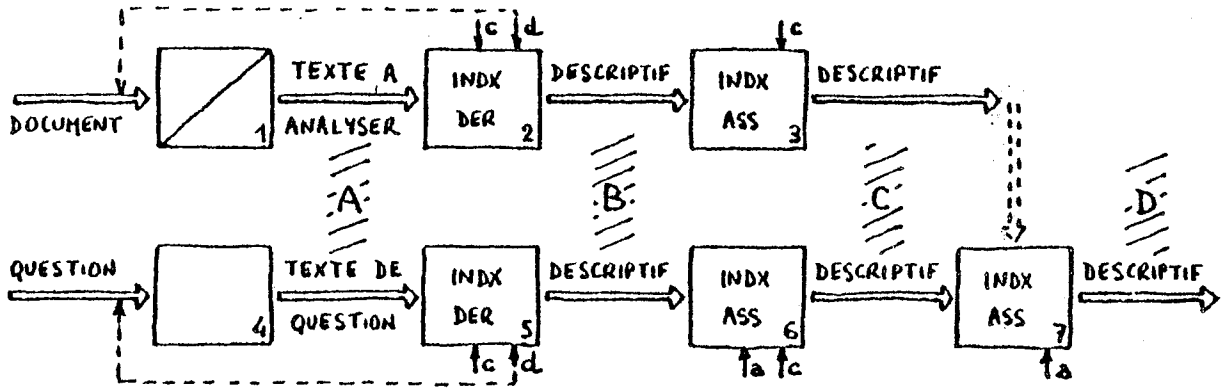
Le fonctionnement de la machine classement ne nécessite pas la disposition de la totalité des textes de documents ; celle-ci a besoin des "textes pour indexer" (partie des textes d'un document utilisé pour l'indexation) et des "textes pour savoir" (partie des textes d'un document permettant à un utilisateur de se faire une opinion sur le contenu du document) ; par exemple, une implémentation particulière choisira le titre comme "texte pour indexer", le résumé comme "texte pour savoir", le document lui-même restant sur support papier.

Les "textes pour savoir" sont utilisés lors de chaque opération de recherche ; les "textes pour indexer" sont utilisés lors d'une opération de classement ; ils doivent être stockés si l'on veut répéter l'opération (re-classement).

Où et comment sera fait ce stockage est affaire d'architecture et d'implémentation de la machine bureautique incluant cette machine classement : mais, d'une manière générale, la machine classement devra disposer de fonctions d'accès à des textes (de documents) et à des classifications (de descripteurs, de documents).

6.5. SCHEMA GENERAL

On peut maintenant donner le schéma général de l'architecture de la machine classement :



Les vocabulaires utilisés sont :

- en A, une langue naturelle
- en B, un vocabulaire de descripteurs
- en C, le vocabulaire "final" des descripteurs (par exemple le nom des classes)
- en D, le vocabulaire d'indentificateurs de documents.

Les classifications utilisées sont :

- pour les machines 2 et 5, une classification de mots de type mot-vide/mot-plein
- pour les machines 3 et 6, une classification de descripteurs (thésaurus)
- pour la machine 7, les descriptifs des documents.

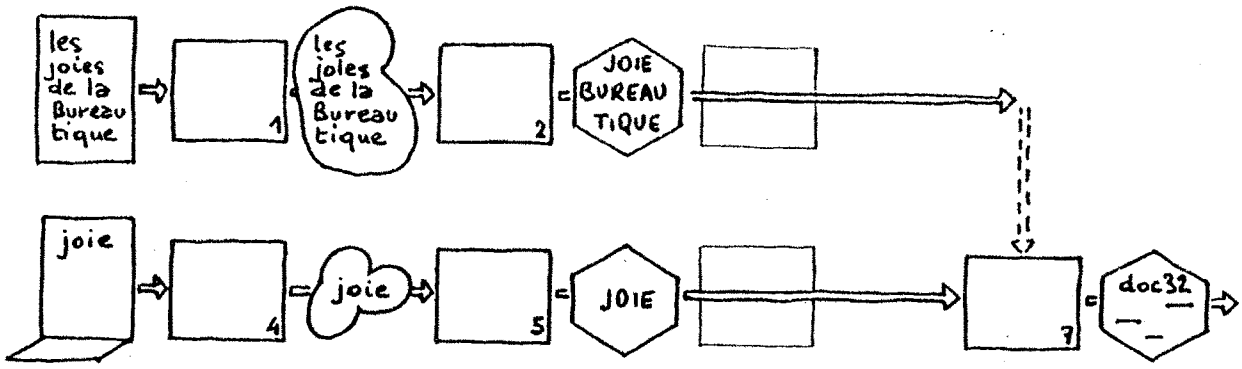
Remarque :

Il apparait clairement sur ce schéma, le rôle d'unification/réduction/ traduction de vocabulaire, de la phase représentée sur les machines 3 et 6. En particulier, cette phase et ces machines peuvent ne pas exister, ou exister en plusieurs exemplaires.

Exemples :

Indexation : tous les mots de texte d'une longeur supérieure à 3 caractères

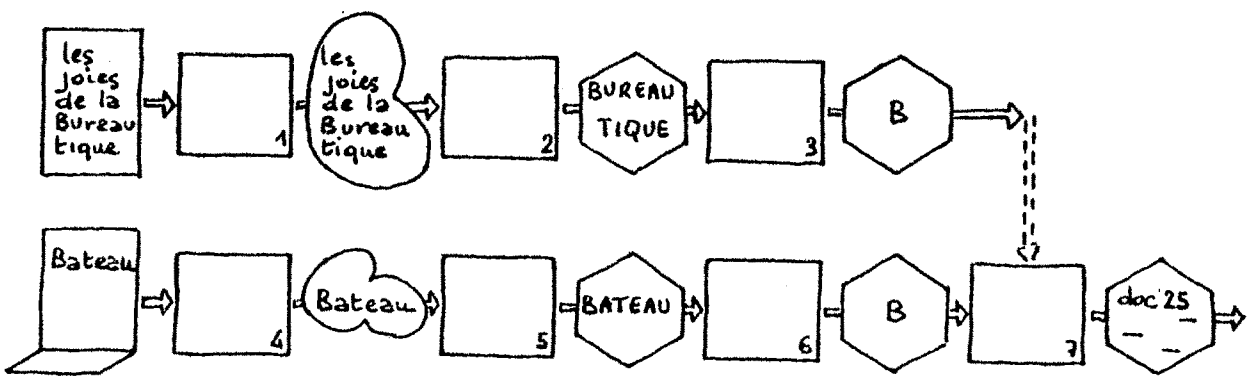
Recherche : les documents possédant un mot.



Indexation : tous les mots en majuscules

Classement : mots commençant par A....Z

Recherche : les documents possèdant un mot



7. PARTICULARITES

On voudrait ici souligner des particularités de la machine classsement en énonçant des hypothèses que nous justifierons et en indiquant des caractéristiques qui influenceront les choix à faire pour sa conception.

7.1. FORME DES DESCRIPTIFS

Les deux fonctions de la machine classement sont le classement et la recherche. Un document est classé une fois, quand il entre dans la machine, quelques fois, si l'on fait du reclassement, et il est concerné par toute opération de recherche. Intuitivement, on peut penser que les temps nécessaires à la réalisation des opérations de classement et de recherche doivent, bien sûr, rester dans les "limites raisonnables" et donc être du même ordre de grandeur.

La machine à classer manipulant un (descriptif de) document et la machine à chercher manipulant tous les (descriptifs de) documents, il s'ensuit que le temps de traitement d'un (descriptif de) document dans la machine à chercher doit être beaucoup plus petit que le temps de traitement d'un (descriptif de) document dans la machine à classer, et qu'en conséquence on devra être très attentif à l'implémentation de la machine qui réalise la recherche, c'est à dire dans la machine 7, sur le schéma général.

En fait, ce raisonnement n'est pas tout à fait vrai. Disons simplement :

- Que les machines 2, 3, 5, 6 et les structures de données qu'elles manipulent seront vraisemblablement très simples.

- Que les phases de traitement 1, 2, 3 (vraisemblablement) et 4, 5, 6 (certainement) seront interactives, et que le temps de réponse apparent est l'intervalle de temps entre 2 messages de la machine (la sortie de messages permet souvent de masquer des temps de calcul...).

- Que la recherche proprement dite (machine 7) nécessite en général, que tout le traitement soit terminé, avant qu'un élément de réponse soit fourni à l'utilisateur.

Ces réflexions nous donnent à penser que la réalisation de la machine 7 doit être telle que le temps de recherche reste raisonnable, et qu'un moyen, sinon le moyen de satisfaire ce souhait, est que les descriptifs manipulés, qui sont des vecteurs, soient des vecteurs binaires.

On peut également ajouter :

- que les machines 3 et 6, "assignent le document à une ou plusieurs classes" ou "assignent au document le nom de ..." : on ne pense pas qu'il soit nécessaire dans le contexte où on est, de pondérer une relation d'appartenance : un document appartiendra ou n'appartiendra pas à une classe.

- que des impératifs de stockage des données nous amèneraient à choisir des vecteurs binaires.

- que les machines 3 et 6 peuvent ne pas être là, ce qui implique que les descriptifs produits par les machines 2 et 5 soient des vecteurs binaires.

- qu'en tout état de cause, des raisons analogues (efficacité, simplicité) nous amèneraient à penser que les machines 2 et 5 doivent manipuler des descriptifs vecteurs binaires.

On ne pense pas que des descriptifs non binaires soient très utiles sur le plan des résultats obtenus. On ne le montrera pas, mais on essaiera de mettre en évidence que, dans le contexte où l'on se place, les descriptifs binaires répondent bien aux besoins de représentation des entités manipulées (questions, documents, classifications) (Remarquons toutefois pour être tout à fait précis que le descriptif résultat de la machine 7, c'est à dire d'une recherche, n'est binaire qu'avec une méthode de recherche de type booléen).

Autrement dit, le fait que les descriptifs soient binaires ne constitue pas une restriction de la machine classement, mais ce fait étant à la fois une hypothèse et le résultat de constatations, il en devient probablement une caractéristique.

7.2. SES QUALITES

On veut parler des qualités en termes documentaires, c'est à dire se poser la question : doit-on choisir des méthodes qui favorisent un taux de rappel élevé ou au contraire un taux de précision élevé.

Dans les systèmes documentaires classiques, l'utilisateur, celui qui effectue des opérations de recherche, est une personne distincte de celle qui a effectué le classement (indexation) du document. Cela a pour conséquence que l'utilisateur connait très peu de choses sur la physionomie de la base (du fait de sa taille), et en particulier ne sait pas ce qu'il cherche ou, plus exactement, n'a que très peu d'idées, a priori, sur la forme (par exemple la taille) de la réponse qu'il va obtenir (notion de processus de recherche).

Par contre, dans un contexte bureautique :

- La taille de la base est très en-dessous de la taille des bases documentaires. Nos estimations et d'autres (voir annexes B et C) nous font penser qu'elle est de l'ordre du millier : de quelques centaines à quelques milliers de documents.

- Le caractère (quasi) individuel de la machine classement fait que celui qui classe est également (ou presque) celui qui cherche.

Cela signifie que l'utilisateur, grâce à sa mémoire personnelle, possède des éléments sur la collection de documents et en particulier, il sait ce qu'il cherche, c'est à dire ce qu'il veut obtenir : le plus souvent un seul document (ou un très petit nombre), et ce document il le connait ou il possède des informations sur lui (ce qui l'aidera à formuler sa question). En particulier, il saura le reconnaître parmi d'autres.

Autrement dit, ce qui est important, c'est que dans l'ensemble des documents constituant la réponse à la question posée, se trouve(nt) le(s) document(s) cherché(s). En termes documentaires, cela implique de choisir les méthodes qui favorisent l'obtention d'un taux de rappel élevé.

7.3. FORME DE LA QUESTION

A priori, la forme de la question dépend du type de recherche qu'on veut mettre en oeuvre : recherche booléenne ou par calcul de similarité. Dans le second cas, la question se présentera comme une liste de mots (presqu'un descriptif) et dans le premier cas, elle risque de se présenter comme des mots reliés entre eux par des opérateurs booléens.

On a déjà mentionné les avantages et inconvénients de la recherche booléenne, mais il faut préciser quel est, en termes documentaires, le rôle ou l'utilité de chacun des opérateurs employés. Si l'on songe à la façon de formuler une question, il est évident qu'on va lier par des "ou" des mots qui, pour la demande considérée, représentent des choses semblables (par exemple "bateau ou navire"), et que l'utilisation de "et" et "sauf" va permettre de "préciser" la question (par exemple "bateau et voile" "est plus précis que" "bateau ou voile").

Si l'on doit favoriser le taux de rappel, les opérateurs "et" et "sauf" ne sont peut être pas nécessaires (même si leur coût d'implémentation est très faible). De plus, la formulation d'une question sous forme d'expression booléenne n'est pas chose évidente, et d'autant moins évidente que les utilisateurs de la machine classement ne sont pas habitués à cette logique.

Si l'on choisit un mode de recherche par calcul de similarité, il ne semble pas non plus nécessaire que le pré-traitement de la question aboutisse à un descriptif non-binaire, c'est à dire que la forme autorisée de la question permette de pondérer l'importance relative des différents mots qui y figurent ; et ceci, d'abord pour des raisons de simplicité de formulation de la question.

Ainsi, on préférera une question se présentant sous la forme d'une liste de mots (qui deviendront des descripteurs), ce qui ne préjuge absolument pas du mode de recherche qui sera choisi, puisqu'on pourra la considérer :

- comme une expression booléenne de "ou" implicites

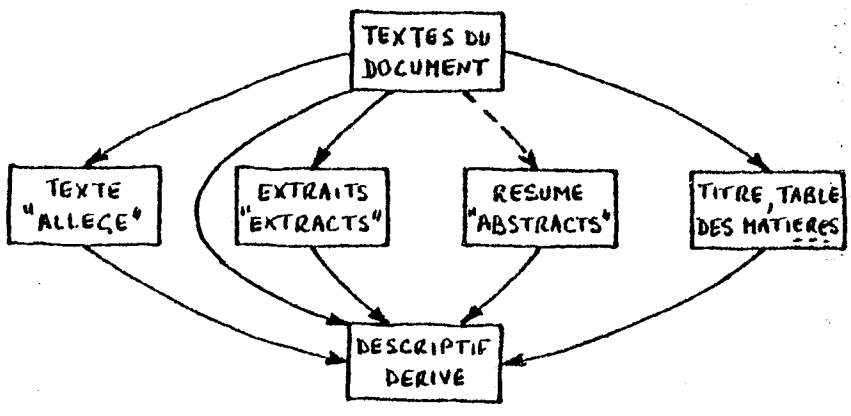
ou

- comme un (presque) descriptif binaire.

7.4. CHOIX DES TEXTES

Indexer un document, c'est à dire lui associer un descriptif, défini par l'analyse des textes du document, nécessite de posséder ces textes ; mais le(s) quel(s) va-t-on choisir comme texte à analyser ?

Les différents produits obtenus à partir des textes d'un document peuvent être représentés par le schéma ci-dessous (d'après [PAI 77]) où un trait plein représente une opération qui peut être automatique, et un pointillé une opération manuelle.



Le matériau de base est constitué par les textes du document :

- On peut les prendre dans leur totalité : autrement dit on a l'identité texte du document = texte à analyser.

- On peut en prendre seulement une partie, soit directement et simplement (ex : couverture, titre,...), soit après une transformation consistant à "déduire" un autre texte des premiers (ex : extracts, résumés,...).

Les documents peuvent arriver par un service de courrier électronique. Il est probable qu'on ne sera jamais (ou du moins avant bien longtemps) dans un environnement où tous les documents arriveront par ce moyen ; même si, au sein d'une organisation, les différents producteurs/consommateurs de documents sont reliés entre eux par des moyens de transmission, il existera longtemps ou toujours des documents n'arrivant pas/ne partant pas par ce moyen.

On retrouve là la coexistence de deux systèmes souvent appelés courrier interne/courrier externe et qui vont devenir courrier électronique/courrier manuel (ce qui ne signifie pas qu'il n'y aura pas de courrier électronique externe).

Si l'on veut que les documents arrivant par courrier manuel (vraisemblablement sur support papier ; le cas de supports directement lisibles, cassettes, disques souples, se ramène au cas du courrier électronique) entrent eux-aussi dans la machine classement, il va falloir les saisir (manuellement) ; ceci peut être long ; c'est toujours une étape qui coûte cher.

Il n'est sans doute pas nécessaire de posséder tous les textes du document pour faire l'indexation ; le document lui-même sera stocké sur un support étranger à la machine classement ; à ce niveau sera entrée dans la machine classement la partie du document "utile" à l'indexation (c'est la méthode classique dans les bases documentaires : seuls sont entrés les attributs qui permettent la recherche et le stockage est constitué par les documents eux-mêmes, lettres, livres, etc...).

Autrement dit, une partie de l'implémentation de la fonction d'accès textes de document textes à analyser est réalisée à l'extérieur de la machine classement. Ce que nous avons appelé document à l'entrée de la machine classement ne sera pas forcément le document entier. Ce qui n'empêche que la décision du choix du texte à analyser aura déjà été partiellement prise.

La diversité des tailles des textes à analyser n'est pas sans importance. L'énoncé "tous les documents sont égaux entre eux" semble être un bon principe. Si on décide de décrire les documents à l'aide des mots des textes "d'égal intérêt", l'importance d'un document va être lié à sa taille, ce qui est en contradiction avec l'énoncé précédent [RIJ 76].

Cela implique donc de prendre en compte la longueur des textes, c'est à dire le nombre de mots, et leurs fréquences d'apparition. Si les textes à analyser sont de longueur sensiblement égale, la fréquence d'apparition d'un mot est une information équivalente au nombre d'occurrences de ce mot. Si les textes sont courts, le nombre d'occurrences d'un mot (dans un texte) est une quantité binaire (voir annexes B et C) et les machines savent bien manipuler ces quantités...

On pense donc que le choix pour les textes à analyser doit se porter sur des textes courts et de tailles sensiblement égales, pour des raisons de saisie et d'efficacité.

Dans le projet SMART [SAL 71], ont été faites les comparaisons d'efficacité entre des titres et des résumés (de document) ; les résultats semblaient meilleurs avec les résumés qu'avec les titres et il était loin d'être évident que des textes plus longs donnaient de meilleurs résultats.

On remarquera qu'on dispose de résumés pour des livres, des articles, etc... et non pas pour les documents de secrétariat (ajoutons que résumer un document est une opération manuelle, par nature non objective, et de même nature qu'une indexation manuelle) ; de plus, les tailles des documents bureautiques sont bien souvent supérieures à celles des résumés.

On doit donc choisir (et on a choisi dans les différentes expérimentations, voir annexes) des textes à analyser qui seront des titres (d'articles), peut être des auteurs, des objets (de lettres : c'est justement fait pour ça...), c'est à dire en tout état de cause, des informations textuelles appartenant à ce qu'on a appelé la couverture de documents.

7.5. REMARQUE DE L'ENVIRONNEMENT

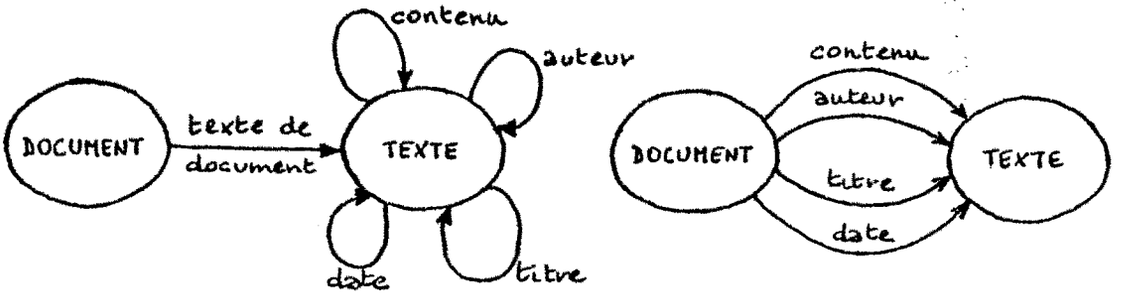
Mais reste un problème : comment, possèdant un document, obtenir les textes de ce document.

On proposait dans [ABH 79], plusieurs façons d'associer des textes et des documents, afin de bien séparer ces deux notions.

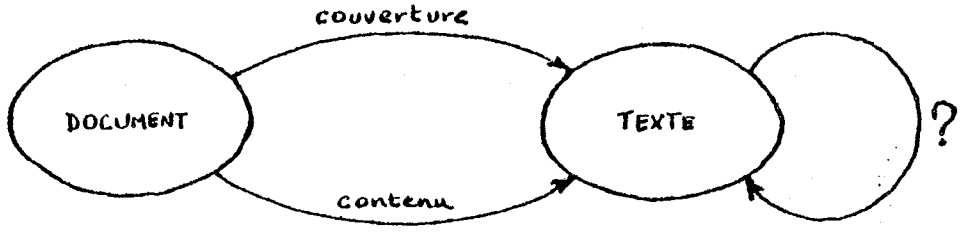
On peut, en particulier :

- soit associer à un document un seul texte (texte probablement structuré),
- soit associer à un document plusieurs textes (textes probablement simples),

comme les figures ci-dessous l'indiquent :



et on avait choisi d'associer à un document, un texte couverture, et un texte contenu, ce qui, pensons-nous, a le mérite de montrer la nature différente de ces textes, sans pour cela préjuger de leurs structures.



Les documents proviennent d'un service traitement de textes (par saisie manuelle ou par saisie automatique), soit local, soit distant (par l'intermédiaire d'un service courrier électronique par exemple) ; le choix de la sémantique des fonctions d'accès permettant d'extraire les textes des documents, n'est pas du ressort de celui qui va les implémenter, mais va dépendre du "formatage" utilisé, des "choses" qu'on pourra retrouver, autrement dit du protocole texte ou plus exactement du protocole document employé.

On peut imaginer des choses très sophistiquées, par exemple :

- exprimer toute la syntaxe d'un texte avec des notions telles que mot, phrase, paragraphe, etc...

- exprimer "le plus de sémantique possible" avec des notions telles que introduction, résumé, titre, nom-propre, etc... ; remarquons ici qu'on exprime à la fois le rôle joué par un texte (ex : introduction, résumé), et les sous-textes que l'on peut y trouver, aussi bien du point de vue de la structure (ex : introduction), que du vocabulaire employé (ex : noms propres).

A l'opposé, on peut imaginer des choses très simples, par exemple, dire qu'un document ne possède qu'un texte, qu'un texte est une chaîne de caractères et que l'implémentation des fonctions d'accès couverture et contenu peut consister à considérer les 500 premiers caractères comme représentant la couverture, et les suivants le contenu du document.

On peut implémenter un protocole texte avec un protocole appareil virtuel ; parce que tout le monde est d'accord sur des notions telles "un mot est une chaîne de caractères non blancs" ou "un caractère "." délimite deux phases", etc... De la même façon, les protocoles Telex, Teletext, voire Telecopie (du nom des services PTT) ne vont pas au-delà de la transmission d'images ou des chaînes de caractères, avec un peu de mise en page, et sont du même niveau que le PAV : dans le meilleur des cas, un document est une suite de textes, un texte est une chaîne de caractères : des notions telles la couverture, le titre, le contenu, n'apparaissent pas dans ces protocoles.

Un exemple de l'insuffisance de ces protocoles est l'absence de moyen simple de connaître la langue dans laquelle est écrite un texte ; les règles de coupure de mots, diffèrent suivant les langues : comment font les éditeurs de textes ?

Un tel protocole ne pourra ignorer ceux déjà existants, il est probable qu'une telle définition consistera alors à définir des caractères sémantiques des textes à partir de caractères syntaxiques.

L'intérêt d'un protocole est qu'il soit utilisé : sa définition doit recueillir un large consensus : elle ne peut être le fait d'un seul individu : il se contentera donc de ces quelques remarques mais en insistant sur l'urgence d'un telle définition si l'on veut assurer la coopération "propre" des différents services bureautiques.

8. DES MACHINES A INDEXER

La machine classement étant clairement définie, il nous faut maintenant examiner les machines la composant, afin d'envisager les différentes possibilités pour chacune d'elles, de mentionner les problèmes posés pour leur conception, et d'indiquer les solutions ayant notre préférence.

8.1. LES MACHINES A INDEXER DERIVATIVES

Ce sont celles qui, à partir d'un texte fournissent un descriptif dérivé.

Au sein de la machine à chercher, il s'agit de transformer le texte de la question en descriptif : la question est déjà presque un descriptif ; son rôle sera donc très simple : vérifier que les mots du texte sont bien des descripteurs utilisables et d'informer l'utilisateur des traitements subis par ce descripteur le cas échéant. Il semble préférable que cette phase soit interactive et que l'utilisateur conserve le contrôle de la formulation de sa question.

Au sein de la machine à classer, le problème est plus complexe.

Cette opération d'indexation se fait généralement en deux étapes :

- Une première qui consiste à "normaliser" le texte fourni.

- Une seconde qui, à partir du texte normalisé, détermine le descriptif.

La normalisation du texte

C'est l'étape où l'on perd le plus d'informations, puisque c'est ici que l'on considère le texte non plus comme une entité d'information au sens le plus général du terme, mais comme un ensemble ordonné de mots. Il s'agit là d'éliminer les mots dont on sait qu'il sont vides de sens (vides de sens en général, ou vides de sens pour l'environnement considéré), et de ramener à une même forme, des mots différents dont on sait qu'ils représentent le même concept.

On peut penser représenter ces fonctions comme des machines à indexer par assignation (classification mots vides/mots pleins, synonymes, etc...). On ne l'a pas fait, d'une part pour rester fidèle à la définition donnée (l'entrée est un texte et non pas un descriptif) et d'autre part, parce que l'ensemble de la fonction indexation dérivative n'est pas (généralement) qu'une suite de machines à indexer par assignation. Il n'en reste pas moins que les méthodes employées seront tout à fait analogues.

Il existe un produit effectuant cette phase de normalisation, c'est PIAFDOC (Programmes Interactifs d'Analyse du Français appliqués à la résolution des problèmes documentaires) développé par l'équipe Intelligence Artificielle de l'IMAG [GRA 78]. On disposera bientôt d'une version sur MICRO1 ; est-ce une machine bureautique ? PIADOC est un gros produit (pas tant par l'importance des programmes, mais plutôt par celle de la grammaire et des dictionnaires utilisés), et peut-être trop sophistiqué pour les besoins qui nous occupent.

On peut effectuer ce travail par des méthodes plus simples.

L'élimination des mots vides peut se faire par la consultation d'un dictionnaire des mots vides (ou "stop-list", ou dictionnaire négatif) ; se pose alors le problème de la constitution d'un tel dictionnaire ; ce n'est pas difficile et il faut prévoir sa mise à jour.

L'implémentation d'un tel dictionnaire, ou plus exactement la définition d'un mot vide peut-être beaucoup plus simple; par exemple, est mot vide, un mot d'une longueur inférieure ou égale à trois caractères. Cette méthode peut sembler simpliste mais les expérimentations faites montrent des résultats satisfaisants (annexe B).

On peut également adopter une méthode intermédiaire consistant à considérer deux dictionnaires d'exceptions de mots vides (mots pleins) qui ont plus (moins) de trois lettres : un mot est vide (plein) s'il a moins (plus) de trois lettres et s'il n'est pas une exception. Là aussi, il faudra prévoir la mise à jour de ces dictionnaires.

La reconnaissance des termes de même famille peut se faire par reconnaissance et remplacement ou suppression des suffixes ; par exemple, ramener les mots à une forme singulière peut consister à supprimer les "s" finals. On peut également dire que les 2 mots représentent le même concept quand leurs 6 premiers caractères sont les mêmes (annexe B), ou lorsqu'ils ont 75 % de caractères consécutifs communs (annexe C) ou, etc... . Toutes ces méthodes favorisent un taux de rappel élevé ; elles ne sont évidemment pas exactes, mais le taux d'erreur restant faible, des algorithmes plus sophistiqués ne nous semblent pas justifiés.

On peut également penser à faire intervenir ici un dictionnaire de synonymes, (un certain type de thésaurus) ; c'est possible, mais il faut bien remarquer que ce n'est pas ici que se situe "traditionnellement" une telle intervention. Son rôle est surtout de réduire le vocabulaire de descripteurs employés, et ainsi, d'unifier les différents "sous-vocabulaires" de descripteurs issus de textes (donc de documents) différents. Autrement dit, l'intervention de ce thésaurus ne se situe pas (généralement) au niveau d'un document, mais d'une collection de documents. C'est une phase d'indexation par assignation.

Le choix des descripteurs

Disposant maintenant du texte normalisé, considéré comme ensemble de mots significatifs, c'est-à-dire comme autant de descripteurs possibles, il faut alors choisir parmi eux, ceux qui formeront le descriptif du document.

On rappelle qu'il existe deux variantes :

- Chercher des descripteurs bien spécifiques

ou

- Détecter les descripteurs possibles.

Chercher des descripteurs bien spécifiques implique de les connaître, c'est-à-dire d'en posséder la liste (dictionnaire des mots pleins ; remarquons que cette étape pourrait se représenter comme une indexation par assignation).

Cet ensemble de descripteurs peut être obtenu a priori, (parce qu'on veut favoriser un bon taux de précision, donc des descripteurs "précis"),ou par l'étude de la ou d'une partie de la collection. Mais, ceci est basé sur l'idée que l'état connu de la collection est une bonne approximation de son état futur. Or, rien n'est moins sûr. Se pose alors le problème de la mise à jour du dictionnaire de mots pleins : il faudra prévoir des outils permettant de le faire, en permettant également le reclassement par suite d'un ensemble de mots pleins non satisfaisants.

Détecter les descripteurs possibles se fait généralement de deux façons : soit sur des critères de position des mots dans le texte, soit sur des critères statistiques.

Remarquons tout de suite qu'avec les textes dont on s'occupe, c'est-à-dire, plutôt des textes courts, plutôt des textes qui sont la couverture ou des morceaux de couverture de documents, on est en présence de textes qui sont plus des groupes de mots juxtaposés les uns aux autres (qu'on peut évidemment appeler phrases), que des textes au sens traditionnel du terme, avec des notions de paragraphes, première phrase d'un paragraphe, introduction,... Il n'apparaît donc pas que ces méthodes soient bien adaptées à nos besoins.

Par contre, on peut penser que des méthodes à caractère plus sémantique, consistant par exemple à détecter les mots suivant les mots "objet", ou "veuillez trouver ci-joint" seraient mieux adaptées. C'est sans doute vrai, mais suppose évidemment que l'on dispose de tout le texte couverture, et surtout que l'ensemble des textes à analyser soit homogène ; or ceci n'est sans doute pas le cas le plus courant : il suffit de constater la variété de documents gérés par un individu ou un secrétariat.

La sélection de descripteurs sur des critères statistiques est beaucoup plus séduisante ; la plupart des méthodes se servent, à l'origine, de la fréquence d'apparition d'un descripteur possible, pour décider de le sélectionner.

La méthode la plus simple est d'en rester là ; d'associer à chaque descripteur possible sa fréquence d'apparition dans le texte, et de sélectionner ceux dont la fréquence est supérieure à un certain seuil, où les n premiers de fréquence les plus élevés, basée sur l'idée que les descripteurs les plus fréquents sont les meilleurs...

On améliore cette méthode en prenant également en compte, la fréquence d'apparition au sein de la collection entière, l'idée de base étant de composer l'utilisation du descripteur pour le texte, et l'utilisation "en général"; la collection dont on dispose représentant le "général".

Différentes mesures ont été proposées : si ft et fc désignent respectivement les fréquences d'apparition dans le texte et dans la collection, on attache au descripteur des valeurs comme ft - fc, ft/fc, ft/ft+fc, log ft/fc [EDM 61].

Là aussi, il faut remarquer que les textes dont on s'occupe ne présentent pas beaucoup de concentration de mots, et que bien souvent, la notion de fréquence d'apparition d'un descripteur possible, pour un texte, est la même que la notion d'absence ou de présence de ce descripteur. Le seul critère restant est l'apparition du descripteur dans la collection.

Autrement dit la décision de choisir un descripteur va dépendre de l'état de la collection ; autrement dit, l'indexation d'un document va dépendre de l'environnement où il se trouve. Ce n'est sans doute pas gênant pour de grosses bases documentaires avec peu de modifications, c'est-à-dire ayant un caractère de grande stabilité : cela semble difficilement acceptable pour de petites bases bureautiques où la collection de documents est sans cesse modifiée (arrivée constante de documents, archivage).

Aussi, pour des textes courts, le plus simple est certainement de prendre comme descripteurs, tous les descripteurs possibles ; pour les textes un peu plus longs, une méthode directement liée à la fréquence d'apparition dans les textes et toute méthode combinée (par ex. tous les descripteurs issus de titre, et ceux issus de résumé apparaissant plus d'une fois, à condition d'avoir en amont un protocole document permettant d'identifier titre et résumé...).

8.2. LES MACHINES A INDEXER PAR ASSIGNATION

Ces machines sont celles qui transforment un descriptif en un autre descriptif ; on les a rencontrées en 3, 6, et 7 sur le schéma général de la machine classement.

Leurs fonctions respectives méritent d'être différenciées, puisque :

- Les machines 3 et 6 peuvent ne pas exister, ou au contraire exister en plusieurs exemplaires ; leur fonction est véritablement la "traduction" de descriptif, alors que,

- La machine 7 est celle qui effectue la recherche proprement dite ; le résultat sera un descriptif, non pas sur un vocabulaire de descripteurs, mais sur un vocabulaire d'identificateurs de documents.

Le principe sur lequel elles reposent est la mesure d'une distance ou similarité entre descriptifs.

Choix de la mesure

De nombreuses mesures ont été proposées et expérimentées, et notre idée n'est pas d'approcher le sujet de façon théorique, mais d'en montrer les principes.

Si X et Y désignent deux vecteurs et ΣXi la somme des éléments du vecteur X, on trouve par exemple, les mesures suivantes :

$\Sigma XiYi \; / \; \Sigma Xi + \Sigma Yi - \Sigma XiYi$ (TAMIMOTO)

$\Sigma XiYi \; / \; \sqrt{(\Sigma Xi^2 \Sigma Yi^2}$) ("cosinus", SALTON)

$\Sigma \min (XiYi) / \min (\Sigma Xi, \Sigma Yi)$ ("overlap", SALTON)

Avec une autre notation, et en s'intéressant à des vecteurs binaires, ces mesures s'écrivent respectivement : $|X \cap Y| \; / \; |X \cup Y|$, $|X \cap Y| \; / \; |X| \; . |Y|$, $|X \cap Y| \; / \min (|X| , |Y|)$.

D'une façon générale, en prenant des vecteurs binaires, comme illustration, une mesure devrait permettre de distinguer les situations suivantes [PAI 77].

- Egalité ex : 111 000 et 111 000
- Inclusion ex : 111 000 et 110 000
- Recouvrement ex : 111 000 et 011 110
- Dissimilarité ex : 111 000 et 000 011
- Complémentarité ex : 111 000 et 000 111

En fait, les expériences montrent que le choix de la mesure n'est pas primordial sur la qualité des résultats, et PAICE [PAI 77] pense que si plusieurs mesures savent distinguer l'égalité de l'inclusion et donnent des valeurs déterminées même si l'un des vecteurs est nul, alors il faut choisir la plus simple.

Deux cas particuliers doivent être mentionnés :

- Lorsque les vecteurs sont de normes sensiblement égales, les mesures proposées par SALTON se simplifient, et à un coefficient près, on retrouve la plus intuitive et la plus simple, c'est-à-dire $|X \cap Y|$.

- Des choix d'implémentation, où l'on effectue la substitution des notions de présence/absence à celle de distance (recherche booléenne, par exemple), peuvent être considérés comme une transformation d'une mesure de distance appartenant à l'intervalle $[0,1]$, en une mesure appartenant à l'ensemble $\{0,1\}$.

D'où l'importance de posséder des descriptifs binaires (voir chapitre 7), de tailles sensiblement égales, et ainsi choisir la plus simple des mesures pour aboutir à des machines à indexer très simples (voir annexes D,E, et chapître 9).

Classifications employées

Examiner la nature d'une classification consiste à examiner les relations qui existent entre les classes et les propriétés (ici, les propriétés sont les descripteurs), la relation entre les classes et les objets (recouvrement ou partitionnement), et éventuellement les relations entre classes et classes (ordonnées ou non).

La machine 7 réalise la recherche : la classification employée est formée par les descriptifs des documents, il est clair qu'un descripteur appartient en général à plus d'un descriptif, que la réponse à une question va comporter plus d'un document, et que les descriptifs jouent tous le même rôle.

Les machines 3 et 6 réalisent un changement de vocabulaire ; contrairement à la machine 7, où la classification employée et le descriptif résultant sont de même nature ; quelquefois les descripteurs et le nom des classes appartiennent au même vocabulaire. Il s'effectue bien une opération de traduction, et parfois de traduction "mot à mot" (et seront implémentées de cette façon).

Les classifications employées peuvent être de deux sortes, correspondant à deux idées différentes de la façon de classer les descripteurs [RIJ 75] : on peut

- soit relier les descripteurs représentatifs d'un même sujet,
- soit relier les descripteurs représentatifs de sujets considérés comme voisins (ou reliés).

Dans le premier cas, on obtient des ensembles de descripteurs interchangeables, c'est-à-dire des classes d'équivalence (par exemple des synonymes), dans le second cas les liens entre les mots seront plutôt à caractère sémantique (par exemple des relations de co-occurrence) ; dans le premier cas le traitement pourra s'effectuer au niveau du descripteur (par exemple remplacement d'un descripteur par le synonyme préférentiel, voir annexe E), alors que dans le second cas le traitement s'effectue au niveau du descriptif (voir annexe D).

Les classifications utilisées par les machines 3 et 6 seront vraisemblablement construites manuellement avec des aides diverses (liste de mots, fréquence d'apparition,...) provenant d'une étude à priori d'une collection de documents semblables (voir annexe D) ou d'informations collectées au fur et à mesure de la constitution de la base (voir annexes D et E).

Ces classifications seront de natures différentes selon le rôle exact assigné à ces machines :

- si c'est un changement de vocabulaire (intervention d'un thésaurus) on aura un partitionnement des descripteurs et recouvrement des documents (voir annexe E).
- si le rôle ressemble plus à un "vrai classement"", un descripteur pourra ou non être partagé entre plusieurs classes selon que l'on voudra ou non favoriser la "ressemblance" entre documents. Ce choix pourra être guidé par le désir d'obtenir ou non une partition des documents (classement simple ou multiple) (voir annexe E).

Ces dernières classifications peuvent être du type mots vides/mots pleins, et on les a déjà rencontrées dans certaines machines à indexer dérivatives. C'est un abus de langage : ces classifications sont telles que certains descripteurs (les mots pleins) sont remplacés par eux mêmes et d'autres (les mots vides) sont remplacés par "rien".

9. DES IMPLEMENTATIONS

On voudrait maintenant dire ce que sera la réalisation de telles machines. On ne proposera pas des spécifications de définition ou de réalisation de la ou d'une machine classement, ce serait l'objet d'une application particulière, mais on indiquera certains aspects ou certaines idées qui devront guider de telles réalisations.

9.1. GENERALITES

Les collections de documents n'ont pas un caractère statique. Dans les bases documentaires "classiques", il arrive toujours de nouveaux documents (ou références de), et la taille de la collection a tendance à croitre indéfiniment. Dans les secrétariats, l'archivage, c'est à dire la sortie des documents du système de classement, se fait régulièrement, souvent à des dates fixes ; l'évolution de la collection de documents suit une allure en "dents de scie", en passant par des minimum où le nombre de documents est proche de zéro. Les collections de documents à caractère plus individuel ont une allure du même genre, dictée par des changements d'activités ou préoccupations du propriétaire ou utilisateur de la collection.

Cette évolution au cours du temps, va se traduire par une modification (un glissement) du vocabulaire (des mots des textes) employé, donc des modifications aussi bien quantitatives que qualitatives du vocabulaire de descripteurs dérivés, et peut impliquer des modifications des classifications utilisées.

A partir du moment où certaines de ces classifications sont modifiées, certaines autres, et en particulier les descriptifs de documents le seront également ; c'est à dire qu'un document indexé ou classé antérieurement à ces modifications risque de se retrouver "mal classé". Cela provoque certainement une détérioration de la qualité des services rendus, et le seul remède est d'effectuer une restructuration de la base, c'est à dire un re-classement de chacun des documents de la collection.

Ceci est impossible à réaliser sur des bases documentaires classiques, à cause de leur taille ; par contre, dans un contexte bureautique, on peut espérer le faire.

Prenons le cas extrêmement simple où une machine à classer produit une base représentée par un fichier inverse (voir annexe F) et où l'indexation se réduit à prendre les 10 mots les plus longs : il faut, pour mettre à jour le fichier inverse 2 x 10 lecture/écriture par document, soit environ 7 secondes sur disquette (20 x 350 ms) et 2 secondes sur cartouche (20 x 100 ms) ; pour 1 000 documents, cela représente respectivement presque 2 heures et plus d'une demi-heure.

C'est un calcul simpliste et un système réel optimisera ses entrées/sorties, ce qui donnera des chiffres peut être plus faibles. Mais cela montre qu'une opération de reclassement est bien sûr toujours possible, mais qu'elle n'est pas immédiate ; le reclassement restera une opération relativement peu fréquente.

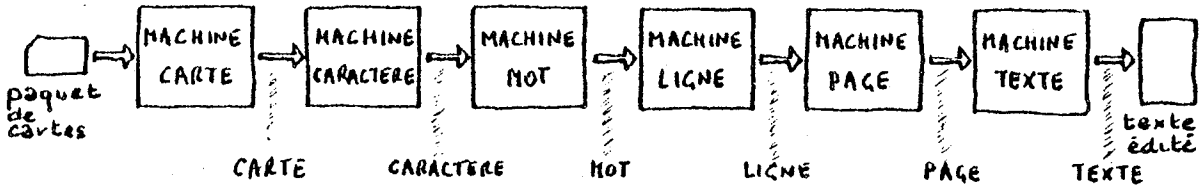
Remarquons ici, au risque de se répéter, que le reclassement est possible parce qu'on dispose d'une fonction d'accès aux textes à analyser, et d'une méthode d'indexation ne nécessitant pas la présence de l'utilisateur.

Si on accepte cela, ce qui est très respectable pour des collections évoluant peu dans le temps, ou dont les documents restent semblables (du même type), alors on peut choisir une méthode relativement "violente" pour la machine à classer, par exemple du type reconnaissance de mots pleins, un certain nombre de documents inclassables indiquant la nécessité de mettre à jour cet ensemble de mots pleins et d'effectuer un reclassement.

Dans d'autres cas (documents trop variés, collection trop changeante,...), ce choix peut ne pas être satisfaisant, voire impossible à exploiter.

9.2. DIFFERENTES REPRESENTATIONS DES TRAITEMENTS

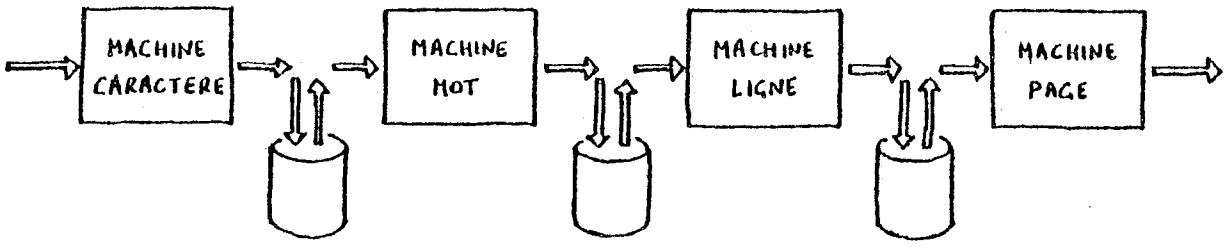
Prenons l'exemple du traitement nécessaire à l'édition d'un texte : on dispose en entrée d'un texte sous un certain format (par exemple un paquet de cartes), et on veut en sortie un texte sous un autre format. Un tel traitement consiste à former des pages avec des lignes, des lignes avec des mots, des mots avec des caractères,..., ce qu'on peut représenter par le schéma ci-dessous, en utilisant le graphisme déjà employé pour la description de la machine classement.



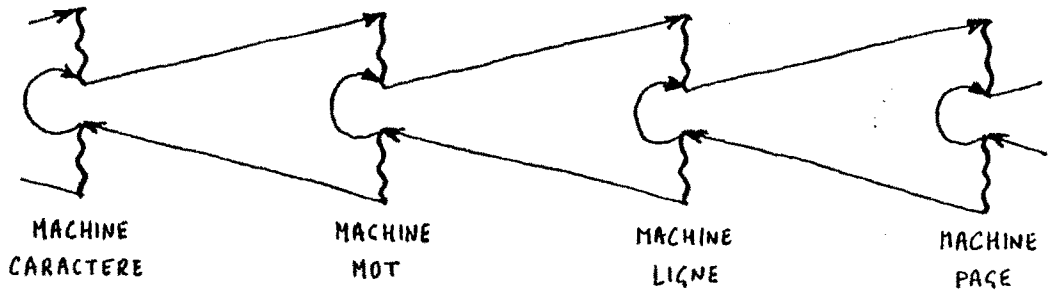
On est en présence d'une suite de machines, chacune d'elle ayant ses entrées et ses sorties, le fonctionnement de chacune d'elle dépendant du fonctionnement de la précédente ; on peut les voir comme un ensemble de processus et une relation d'ordre entre leurs exécutions (schéma producteur-consommateur).

Avec un tel schéma, deux cas extrêmes sont souvent rencontrés, qui correspondent à des implémentations différentes :

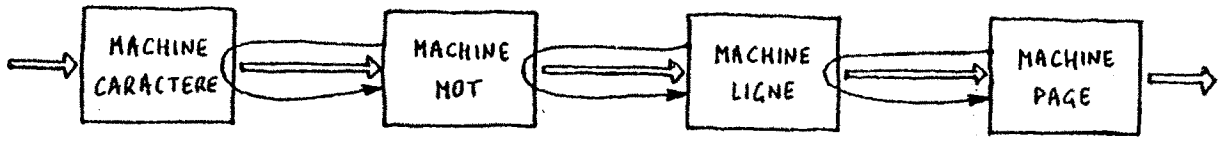
- celui où une machine attend la fin du fonctionnement de la précédente pour commencer le sien ; ceci implique de stocker les sorties de la machine précédente ; on pourra représenter ce cas comme ci-dessous :



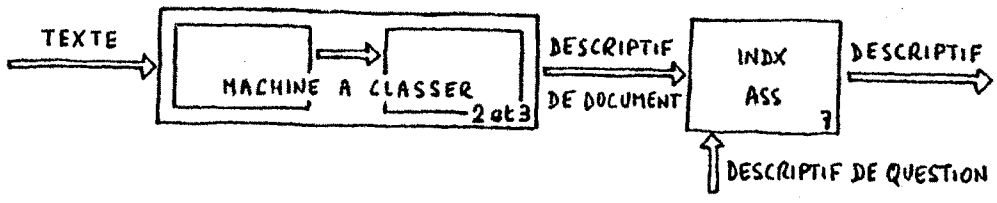
- celui où une machine est directement commandée par la suivante, selon ses besoins : on est en présence de la représentation d'un algorithme ; il n'y a plus qu'une seule machine, qu'un seul processus, son exécution étant une suite d'appels de procédures : on la représente souvent de la façon suivante.



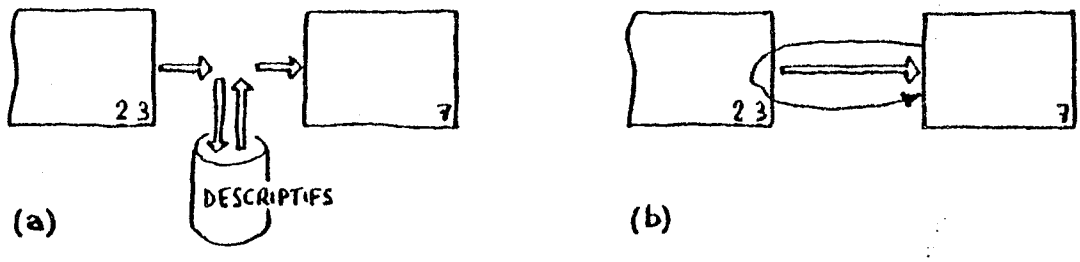
mais on préférera le schéma suivant :



Revenons à la machine classement, la partie recherche est décrite ainsi :



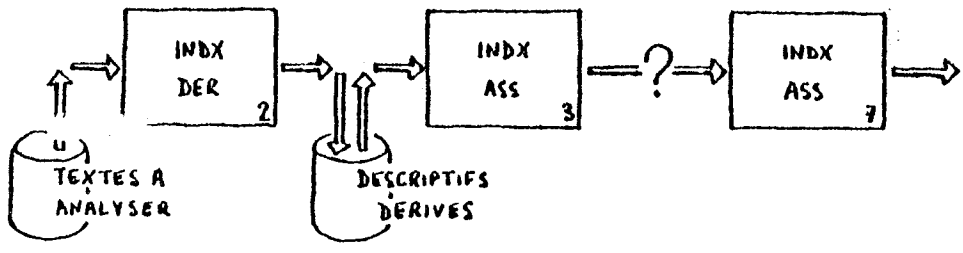
D'après ce qui précéde, cette partie peut être implémentée de plusieurs façons, et en particulier selon les deux schémas ci-dessous :



Cela signifie que sur le schéma (a), la base documentaire est réelle, c'est à dire que chaque descriptif est un objet réel auquel on va accéder, alors que sur le schéma (b) la base est virtuelle, chaque descriptif étant le résultat d'un calcul effectué par la machine à classer.

Le schéma (a) représente l'implémentation classique. Le schéma (b) semble irréaliste : il signifie que chaque opération de recherche va nécessiter le calcul des descriptifs des documents : si ce calcul se fait à partir des textes à analyser, les estimations du temps de reclassement (voir annexe F) montrent que ce n'est guère possible.

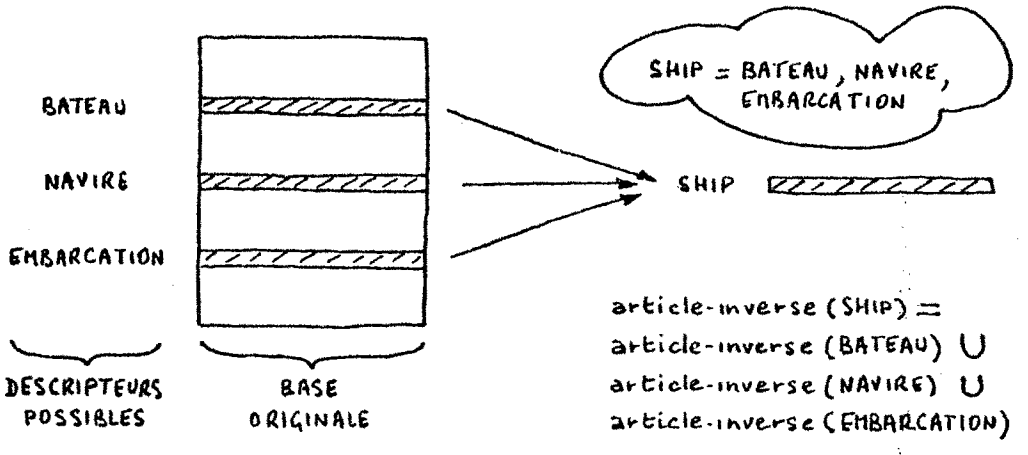
Mais n'oublions pas que la machine à classer comporte une machine à indexer dérivative (machine 2), suivie d'une machine à indexer par assignation (machine 3). Ainsi, on peut envisager le schéma ci-dessous, où les descriptifs dérivés sont des objets réels :



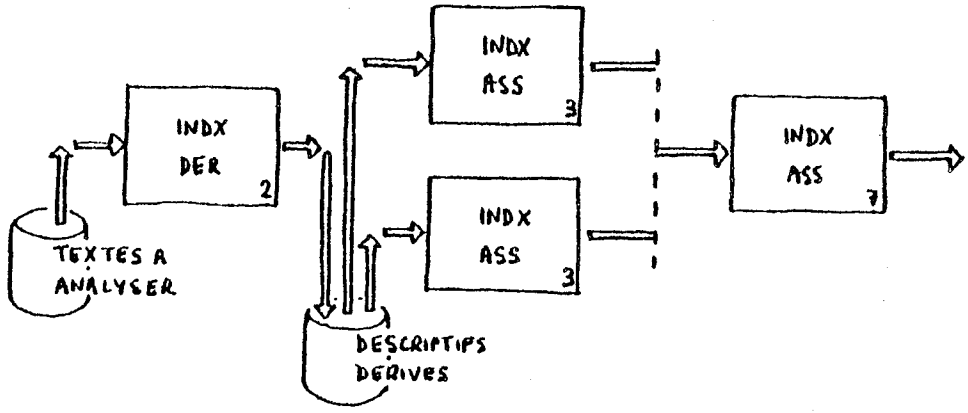
Ceci veut dire qu'un reclassement peut ne pas nécessiter l'accès aux textes à analyser, mais simplement aux descriptifs dérivés ; ceci suppose alors que les descriptifs dérivés aient un caractère objectif et permanent, c'est à dire que l'opération d'indexation dérivative soit faite une fois pour toutes, donc qu'elle soit "sans risque", et indépendante de l'environnement.

Un exemple d'une telle solution, lorsqu'on est en présence de textes à analyser de taille réduite est de prendre comme opération d'indexation dérivative, celle qui consiste à conserver tous les mots non vides des textes.

Le faible coût de calcul de l'opération d'indexation par assignation (machine 3), est mis en évidence par le schéma ci-dessous, où la base originale (ensemble des descriptifs dérivés) est implémentée comme un fichier inverse (voir annexe F), et où l'obtention d'une partie de la base restructurée par une classification de descripteurs synonymes, consiste à effectuer quelques opérations ensemblistes.



A partir de là, on peut envisager plusieurs types de réalisations, selon que la base classée (c'est à dire le résultat de la machine 3) est réelle (la machine 3 a fonctionné et produit une base réelle) ou virtuelle (la machine 3 est une procédure appelée par la machine 7). Dans le premier cas, une opération de reclassement consiste effectivement à produire une autre base, dans le second cas, elle consiste simplement à modifier la classification utilisée. Pourquoi, dans ces conditions, ne pas envisager l'utilisation de plusieurs classifications pour produire plusieurs classements, sur la base originale, puisqu'un classement est simplement le résultat d'un calcul particulier.



Le but de l'opération de classement, dont le résultat est une classification de documents, apparait bien comme ce qu'il est : une vision particulière d'une collection de documents.

Ce type d'implémentation, qui permet d'effectuer des reclassements à faible coût, sera préféré pour des collections de documents inconnus a priori, ou des collections très changeantes. Les bases résultantes pourront être soit réelles, soit virtuelles, mais n'oublions pas la condition de départ : disposer d'une base originale "objective", donc d'une méthode d'indexation "sans risques", cette base étant à l'origine de toutes les autres.

9.3. DES VARIETES DE SYSTEMES

On est maintenant arrivé à un point où l'on peut dire qu'on va certainement rencontrer deux variétés de systèmes, selon que la classification des documents sera représentée par des objets réels, ou sera le résultat d'un calcul particulier. Si dans le premier cas les notions de classe et de classement sont bien présentes, elles semblent beaucoup plus diffuses dans le second cas ; en particulier, le mot reclassement, s'il décrit bien la solution au problème de la mise à jour, dans le premier cas, semble mal adapté à la description d'une opération analogue dans le second cas.

Se pose alors le problème de savoir, pour une application donnée, quelle variété de système choisir. Ce choix sera évidemment guidé par des impératifs techniques, dépendant du matériel employé, tels le temps de recherche souhaité, le volume du stockage, etc... mais également par l'importance accordée au fait de posséder, toujours, une collection bien classée : est-ce qu'on acceptera ou non de posséder une collection qui pourra, à certains moments, être mal classée ?

En effet, il ne faut pas oublier les principales raisons pour lesquelles une collection (bien classée à l'origine), devient mal classée, c'est :

- parce que la collection a évolué,
- parce qu'on voudrait en avoir une autre vision (ou plusieurs simultanément).

Si donc, on accepte une collection non toujours bien classée, alors on peut envisager la méthode "classique" du classement, réalisée d'abord par une indexation dérivative de type reconnaissance de mots pleins, un nombre de classes relativement faible, l'opération de recherche consistant simplement à choisir une classe et lister les documents appartenant à cette classe. On peut y ajouter certains indicateurs ou outils permettant d'obtenir des informations sur la validité des classifications utilisées (documents "inclassables" "centre de gravité" d'un ensemble de documents, descripteurs utiles, distance entre documents,... voir annexe D), mais, en tout état de cause, une mise à jour consistera à redéfinir les descriptifs des classes, donc d'abord le vocabulaire des mots-pleins et les descriptifs dérivés des documents, puis effectuer un reclassement. On ne peut en dire plus pour ce genre de réalisation.

On peut aussi envisager la solution où la classification des documents est le résultat d'un calcul particulier, parce que l'outil le permet, et parce qu'elle apparait beaucoup plus séduisante, du fait de la disparition de la notion de reclassement physique, et par suite de pouvoir posséder facilement plusieurs classifications d'une même collection.

On peut remarquer ici qu'on suit la même évolution qu'en documentation classique où l'on est passé de la notion de schéma de classification à celle d'indexation ; on passe ici de la notion de classification, résultat d'un classement, donc forcément unique et rigide, à celle de classification multiple.

Une telle solution sera d'abord réalisée par une indexation dérivative "sans risques", de type élimination des mots (toujours) vides. L'inconvénient de cette méthode est qu'alors, le vocabulaire de descripteurs dérivés est important : il est évident qu'on va y trouver beaucoup de descripteurs inutiles, et qu'à la main on en choisirait beaucoup moins.

Si la phase d'indexation par assignation n'existe pas (machine 3 sur le schéma général), le descriptif final associé au document est exprimé sur le vocabulaire des descripteurs dérivés (vocabulaire B), et (le descriptif de) la question doit être formulée dans ce vocabulaire ; il faut donc offrir des outils permettant d'aider au choix de ces descripteurs.

Ce problème du choix des descripteurs se pose également dans d'autres cas de figures, et lorsque la phase d'indexation par assignation existe. Par exemple :

- lorsqu'on possède une collection de documents et qu'on voudrait déterminer un "bon" vocabulaire de mots pleins.

- Lorsqu'on possède une classification de documents et qu'on voudrait déterminer les descriptifs qui permettraient, par un processus d'indexation par assignation, d'obtenir cette même classification.

- Lorsqu'on utilise une classification de descripteurs synonymes, le vocabulaire résultant peut lui aussi être important.

Il est intéressant de remarquer que cette évolution de la classification de documents, résultat d'un classement, vers l'indexation (dérivée ou assignée) et les outils d'aide au choix des descripteurs, pour la gestion d'une collection de documents, consiste en fait à passer de la notion de classification de documents a celle de classification de descripteurs, comme outil principal d'aide à la recherche. L'aide au choix des descripteurs pour la formulation d'une question est basée sur l'idée de proposer à l'utilisateur une approximation d'une partition de descripteurs : ceux à utiliser, et ceux à ne pas utiliser. On insiste bien sur le caractère automatique que doivent revêtir ces outils, qu'on ne doit pas confondre avec d'autres, compatibles bien sûr, qui consistent à restituer des données enregistrées (consultation de thesaurus par exemple).

9.4. DES EXEMPLES D'OUTILS

Deux idées, empruntées à la documentation automatique, permettent de réaliser simplement de tels outils :

- une première, consistant à associer à chaque descripteur, une valeur en représentant la "signifiance",
- une seconde, consistant à associer à chaque descripteur, d'autres descripteurs liés au même sujet.

Associer une valeur à chaque descripteur ressemble à la technique utilisée pour l'indexation ; la différence est que pour l'indexation, on associe une valeur aux descripteurs possibles d'un document (et on ne conserve que ceux de plus grande valeur), alors qu'ici, on applique la méthode à l'ensemble du vocabulaire des descripteurs.

Cette valeur pourra être simplement la fréquence d'apparition du descripteur (et on retrouve là ce qui existe dans les systèmes documentaires classiques), ou une fonction de la fréquence d'apparition, mais des choses très sophistiquées ne semblent pas nécessaires (voir annexes B et C). Pour une fonction donnée, on peut imaginer de présenter à l'utilisateur les "meilleurs" descripteurs (par exemple les "100 meilleurs", la centaine étant un nombre de descripteurs permettant l'affichage sur un écran), parmi lesquels il pourra choisir ceux, pour formuler sa question, ou bien ceux qu'il considère "utiles" pour sa vision des choses (par exemple, éliminer des mots vides et faire un nouveau choix des "100 meilleurs").

Evidemment, une précaution à prendre est de s'assurer que les descripteurs ainsi choisis couvrent bien la quasi-totalité de la collection, autrement dit que la quasi-totalité des descriptifs des documents, s'expriment en partie avec ces descripteurs ; on n'aura probablement jamais une couverture totale de la collection : il restera quelques documents, parce que leurs textes seront de mauvaise qualité par exemple (voir annexes B et C) ; la solution sera alors de les réindexer après avoir changé leurs textes, ou encore de les laisser tels qu'ils sont : ils formeront les documents inclassables (de la même façon que dans une gestion manuelle), et on pourra chercher parmi eux (les documents inclassables forment une classe...).

Associer d'autres descripteurs à un descripteur, là aussi de façon automatique, est une méthode utilisée pour la construction automatique de thésaurus, basée sur l'examen de la co-occurrence, c'est à dire sur l'idée que des descripteurs apparaissant ensemble ont probablement trait au même sujet, tout au moins à des sujets non indépendants.

Autrement dit, on possède une classification des descripteurs puisqu'on possède un ensemble de descriptifs, cette classification n'est pas anodine, et l'on peut utiliser ces informations ; ce peut être un hasard si deux descripteurs apparaissent une fois ensemble, ce n'en est plus un si cela se produit une fois sur trois (voir annexes B et C).

Une implémentation particulière pourra présenter les descripteurs par ordre décroissant du nombre de co-occurrences et ce, à différents seuils. Remarquons que la réalisation de cet outils nécessitera la disposition des deux fonctions d'accès entre descripteurs et descriptifs, puisqu'on utilise le chemin descripteur ⟶ descriptif ⟶ descripteur.

Il existe sans doute d'autres outils d'aide au choix des descripteurs, mais il est certain que ceux que l'on vient de suggérer sont efficaces et d'une extrême simplicité de mise en oeuvre.

Mentionnons ici encore, l'importance de posséder une opération d'indexation dérivative "sans risques", permettant donc de disposer des ensembles des descripteurs possibles et de descriptifs "objectifs", les aides au choix des descripteurs étant le résultat de l'exécution d'un algorithme ayant ces ensembles pour données.

A priori, on ne fera pas un choix entre ces différentes méthodes : chacune d'elle a ses avantages, et c'est leur co-existence qui fournira un service efficace. Chacune d'elles produit un service dont le résultat est comme une fenêtre sur l'ensemble des descripteurs ; leur coopération fournit des moyens de cheminement dans ce vocabulaire, comme autant de fenêtres ouvertes sur une même réalité.

On peut sans doute faire l'analogie avec un individu arrivant dans un magasin en libre service et devant effectuer un certain nombre d'achats en fonction d'un certain objectif. Deux cas se présentent :

- ou bien il connait parfaitement la liste des achats qu'il doit effectuer, et il va aller directement dans les rayons où il sait qu'il va trouver ces objets, à l'aide de panneaux qui indiquent l'emplacement de ces rayons.

- ou bien il ne connait pas, a priori, les achats qu'il va effectuer : il va se promener dans l'allée centrale et à la vue des noms des rayons, et plus sûrement à la vue d'un morceau de chaque rayon, il décidera si ce rayon est susceptible de comporter des objets correspondant à son objectif initial.

Le premier cas est celui où l'utilisateur connait le schéma de classification (le rangement du magasin), le second cas est celui où il choisit pas à pas : le nom des classes (le nom des rayons) ou quelques descripteurs (quelques objets d'un rayon) puis les descripteurs reliés (les objets du rayon).

Parmi toutes les variantes de systèmes possibles, on avouera notre préférence pour des systèmes où les classifications de documents ne sont pas représentées par des objets réels, comportant des outils d'aide au choix des descripteurs, complètement automatique, sinon "objectives", du moins se comportant toujours de la même façon, et étant une réponse efficace aux problèmes posés par la modification des collections, et leur partage par plusieurs utilisateurs (vision multiple).

On illustrera l'importance de ce dernier point en parlant de coopération et de répartition de bases bureautiques.

9.5. COOPERATION - REPARTITION

La bureautique, au sein des organisations, sera par nature répartie, et même fortement répartie. Un service bureautique déjà mentionné, le courrier électronique, permettra la communication entre utilisateurs. Chaque utilisateur possède une collection de documents : on peut penser qu'il désirera la mettre, totalement ou en partie, à la disposition d'autres utilisateurs. On est en présence d'une base répartie et on doit offrir des outils de gestion de cette collection répartie.

Ce souci de coopération des bases documentaires bureautiques peut sembler en contradiction avec une caractéristique déjà mentionnée : le caractère individuel de la machine classement. En fait, il faut bien distinguer deux choses :

- les documents eux-mêmes (la collection de documents),
- les descriptifs associés à ces documents (la base documentaire proprement dite).

Ainsi, s'il est évident que la fonction de classement (dont le résultat est un ensemble de descriptifs), est individuelle et en général non partageable puisque représentant une vision particulière de la collection, il est non moins évident que les documents eux-mêmes peuvent être partagés : on prête des documents, on ne prête pas leur classement.

Pour les bases documentaires et les systèmes documentaires classiques, ces problèmes de coopération/répartition n'ont quasiment pas été abordés.

Ce qu'on pourrait prendre pour de la coopération de bases documentaires, c'est par exemple le réseau documentaire EURONET. Ce n'en est pas ; le service offert est principalement un service d'accès à distance (non simultané) de plusieurs bases. A un instant donné, il n'existe qu'une liaison bi-point entre un utilisateur et une base. En fait, ce qu'on savait déjà, le problème de la coopération n'est pas induit par l'éloignement géographique des bases ; on remarque que les serveurs documentaires gèrent, sur un site, plusieurs bases indépendantes : il est nécessaire de choisir une base avant toute interrogation et il n'a jamais été question de faire en sorte qu'il en soit autrement.

Ce qu'on pourrait prendre pour de la répartition de bases documentaires, c'est par exemple le logiciel MISTRAL utilisant des logiciels MISTRAL standards, répartis sur un réseau (CERISS Toulouse). Ce n'est en fait qu'une proposition d'architecture de logiciel, et l'aspect sémantique des bases pouvant être gérées par ce logiciel n'est absolument pas abordé.

Ce qu'on attendrait d'un tel système est de permettre une interrogation "multi-bases, multi-sites", ou cet aspect "multi" est transparent (ou le plus transparent possible) à l'utilisateur, afin de lui donner l'impression qu'il est en présence d'une seule base. S'il n'est pas difficile, parce que ce sont maintenant des problèmes connus et en partie résolus, car rencontrés dans les projets de coopération de bases de données, "d'unifier" les différentes formulations de commandes (aspect syntaxique : modèles de données, langages de description et de manipulation), il est par contre beaucoup plus délicat d'aborder l'aspect sémantique, c'est à dire de construire ou de simuler l'existence d'un langage documentaire unique (ensemble des descripteurs et des relations entre ces descripteurs).

Les solutions qui sont suggérées sont, en particulier :

- la mise en oeuvre de stratégies de recherches multi-bases [FAU 78], par exemple par comparaison de l'indexation de documents identiques dans des langages documentaires différents [KOP 78],

- la mise en oeuvre de techniques analogues à celles utilisées pour l'élaboration de thésaurus multilignes [KOP 78] (lexique intermédiaire, reconciliation de thésaurus).

En règle générale, le premier exemple nécessite d'avoir une structure particulière de la base (duplication de documents) et des moyens de reconnaître cette structure (identification des documents), et le second exemple un gros effort intellectuel. On peut sans doute mentionner ici que, si le problème de la coopération des bases de données documentaires n'a pas beaucoup été abordé, c'est à cause des difficultés rencontrées.

Par contre, dans un contexte bureautique, du fait de sa nouveauté ("on part de rien"), et de la faible taille des collections (on peut re-classer, re-indexer,...), les notions d'indexation dérivative "objective" et de classification dynamique comme résultat de l'exécution d'un algorithme, présentées précédemment vont permettre de réaliser une "vraie" coopération, une coopération "complète" entre bases documentaires bureautiques, sans rencontrer les difficultés mentionnées ci-dessus, pour peu que l'on respecte quelques règles très simples pour l'implémentation.

On peut imaginer un type de réalisation défini de la façon suivante :

- la collection globale des documents est réalisée par l'union des différentes collections locales,

- la base documentaire (ensemble des descriptifs) est réalisée par l'union des différentes bases.

Pour que cette dernière affirmation soit correcte, il faut que ces bases soient sémantiquement analogues, c'est à dire plus simplement que si un même document était indexé sur plusieurs machines locales, les descriptifs résultats seraient identiques (utilisation du conditionnel : ce cas ne doit pas se produire fréquemment puisque la notion d'origine de document (chapitre 3) doit permettre de distinguer les "originaux" des "copies").

Pour que ceci soit réalisé, il suffit que les différentes machines classement locales possèdent les mêmes fonctions "texte à indexer" (machine 1) et les mêmes machines à indexer dérivatives (machine 2) ; même algorithme et même classification. De la même façon, les machines traitant la question (machines 4 et 5) doivent être identiques. On ne mentionne rien sur les machines à indexer par assignation (machines 3 et 6), qui peuvent bien sûr exister au niveau local, mais n'entrent pas en jeu au niveau global, puisqu'à ce niveau aucun classement particulier n'est pris en compte.

Dans ces conditions, il semble évident de mettre en oeuvre au niveau global :

- une classification particulière, résultat d'un processus de calcul,

- des outils d'aide au choix des descripteurs (association d'une valeur, association d'autres descripteurs), décrits précédemment,

implémentés par le lancement sur chacun des sites, de l'exécution des algorithmes mentionnés précédemment.

Ceci ne revient pas à dire que les machines locales doivent être identiques ; seulement les bases, résultat d'une indexation dérivative "objective", doivent l'être, et sans doute doit-on trouver sur chacun des sites des (parties de) machines à chercher identiques. Toute liberté est laissée pour le reste en particulier tous les classements possibles.

L'intérêt évident et extrêmement important d'une telle architecture est de permettre la connexion/déconnexion dynamique de "morceaux" de la base globale. On a présenté ce qui précède selon l'approche ascendante de coopération ; une approche descendante aboutira à la même architecture et permettra de réaliser un système réparti, à contrôle centralisé, de gestion de données réparties, offrant les mêmes services qu'un système centralisé et pouvant fonctionner en mode dégradé.

On n'en dira pas plus, ce n'est pas le but de ces lignes, mais soulignerons l'extrême intérêt de cette approche, permettant la réalisation d'un "vrai" système documentaire réparti.

CONCLUSION

Au long de ce rapport, on a d'abord montré ce que l'on appelle "documentation automatique" et précisé ce qu'était pour nous la bureautique. On a ensuite indiqué comment des méthodes classiques pouvaient être utilisées pour la gestion de collections de documents bureautiques, choisit de s'intéresser à ce qu'on a appelé l'approche documentaire, et à travers la définition du classement, présenté l'architecture de la machine classement.

Ce nom a été choisi car le mot classement semble être clair pour ceux qui gèrent des documents, et que les expressions "documentation automatique" ou "système documentaire" semblent les effrayer... On a défini en fait ce qu'on aurait pu appeler la "machine documentation", en espérant avoir montré que le classement n'en est qu'un aspect particulier.

On a ensuite précisé quelles sont ses particularités, quel environnement lui est nécessaire, et quels sont les choix possibles pour ses différents composants. On a enfin donné des indications sur certains aspects de réalisation.

Arrivé à ce point du travail, on ne se trouve ni en présence d'une théorie de la documentation automatique en bureautique (si tant est qu'il puisse en exister une), ni en présence d'une réalisation d'une machine classement. Cette remarque indique bien, dans quelles directions maintenant orienter la suite de ce travail :

- d'abord réaliser une telle machine en "vraie grandeur", c'est à dire :
  - implémentée sur du matériel bureautique (ou individuel),
  - utilisée réellement, non pas par ceux qui l'ont conçue ou réalisée, mais par des utilisateurs prêts à s'adapter à la bureautique.

Seulement à ces conditions on pourra véritablement apprécier la validité des propositions faites.

- Ensuite, continuer à examiner les travaux effectués dans le domaine de la documentation automatique, travaux relativement anciens (les années 60), et qui semblent être oubliés, peu diffusés, ou ignorés. Un reproche majeur fait à l'époque à ces travaux est que les expériences menées l'ont été sur des collections dont la taille ne correspondait pas à la réalité des bases documentaires ; ce n'est pas le cas en bureautique, et nous sommes persuadés qu'un grand profit peut être tiré de ces travaux.

Seulement à cette condition, on pourra faire évoluer et améliorer les premières réalisations.

Au terme de ce rapport, nous pensons qu'au-delà de la définition des concepts et des objets manipulés, au-delà de la suggestion de quelques idées, on pense avoir défini un cadre de travail pour maintenant développer une classe d'applications bureautiques. Sans avoir la prétention que ce rapport devienne un élément du "petit manuel à l'usage des buroticiens", il offre en particulier une présentation de la documentation automatique trop souvent dispersée dans la littérature, voire inexistante, qui nous semble digne d'intérêt.

On espère avoir montré qu'il existait une approche différente pour offrir des outils de gestion d'une collection de documents, approche qui n'est ni celle du traitement de textes, ni celle du traitement de données, ni, dans une autre direction, celles liées à l'analyse des langues naturelles, mais une approche simple et offrant de réels services.

Toutefois, il ne faudra jamais oublier la nature du problème auquel on s'intéresse ; on empruntera notre conclusion à celle de [PAI 77] : "Le fait qu'il n'est pas possible de poser une question à une machine et de recevoir en réponse la liste des documents se rapportant à la question, et seulement ceux-là n'est pas une mise en accusation des machines : c'est dans la nature du problème de recouvrement d'informations. Pas plus qu'il n'en est une autre de dire que les systèmes informatiques ne font pas mieux que les systèmes traditionnels : le dire est énoncer une évidence sur la difficulté de traiter le problème et sur le peu de choses que nous en connaissons".

Autant dire qu'un souci pédagogique constant devra prédominer dans la définition de produits de cette classe d'applications bureautiques.

ANNEXE A - EXEMPLE DE GESTION DE SECRETARIAT - INTERVIEW

"... le problème principal est le classement... le but est de pouvoir retrouver un ou plusieurs documents... donc il faut classer puisque la mémoire humaine est défaillante... sur demande d'un document, la secrétaire doit pouvoir le donner... éventuellement, quelqu'un d'autre doit pouvoir s'y retrouver...".

"... en général, un document se rapporte à une question ou à une personne...".

"... il existe deux gros bouquins... courrier arrivée... courrier départ...".

"... quand une lettre arrive... un numéro sur la lettre... on écrit sur le cahier le numéro du document, la date d'arrivée, l'expéditeur, le sujet... le sujet n'est pas forcément la même chose que l'objet qui figure sur la lettre et qui est mis par l'expéditeur... le sujet est déterminé par le destinataire... servira à classer la lettre... sera peut être classée plus tard, après la réponse...".

"... quand on envoie une lettre, il existe deux doubles... sur un double, on met un numéro, la date, le destinataire, l'objet, le classement. des doubles va dans un dossier... l'autre dans un gros classeur... classé chronologiquement... qui reflète l'activité du secrétariat..."

"... il existe un petit carnet répertoire avec les noms des destinataires, leurs adresses, les numéros de lettres... pouvoir répondre à la question "qu'est-ce qu'on a écrit à Tartempion "...".

"... tout document est classé... va dans un dossier... un document peut concerner plusieurs dossiers... ou en choisit un et on met une note, une référence dans les autres..."

"... un dossier est identifié par un mot... un numéro...".

"... la détermination des dossiers est liée à l'activité... à l'appréciation de la secrétaire... par exemple un secteur devient important, on ouvre un nouveau dossier...".

"... on retire des dossiers les lettres de 2 ou 3 ans... souvent prétexte à une remise en ordre... paquets... ficelle... ou grenier... détruit au bout de 30 ans... on recherche rarement au grenier...".

Hélène, Août 1979

ANNEXE B - UNE COLLECTION DE DOCUMENTS ADMINISTRATIFS

Collection du courrier d'un secrétariat, unité administrative d'un "Institut des Sciences exactes" au sein d'une université étrangère (Constantine - Algérie). Courrier, comme reflet de l'organisation, à tendance très administrative, voire bureaucratique.

Examen de la collection correspondant à une unité d'archivage, soit une année : nombre de documents ≈ 750.

Méthode de gestion "classique" : partition des documents

- d'une part, en "départ" et "arrivée", puis par origine/destination pour :
  - . les services administratifs (de loin la plus importante)
  - . les autres instituts
  - . les fournisseurs
- d'autre part, selon la nature du support ou de l'origine :
  - . les circulaires (en provenance du ministère)
  - . les télégrammes/telex.

On peut répartir ce courrier en deux classes : courrier interne et courrier externe ; le secrétariat d'un institut ou un service administratif, unités de gestion, sont les unités d'adressage (boîte à lettres).

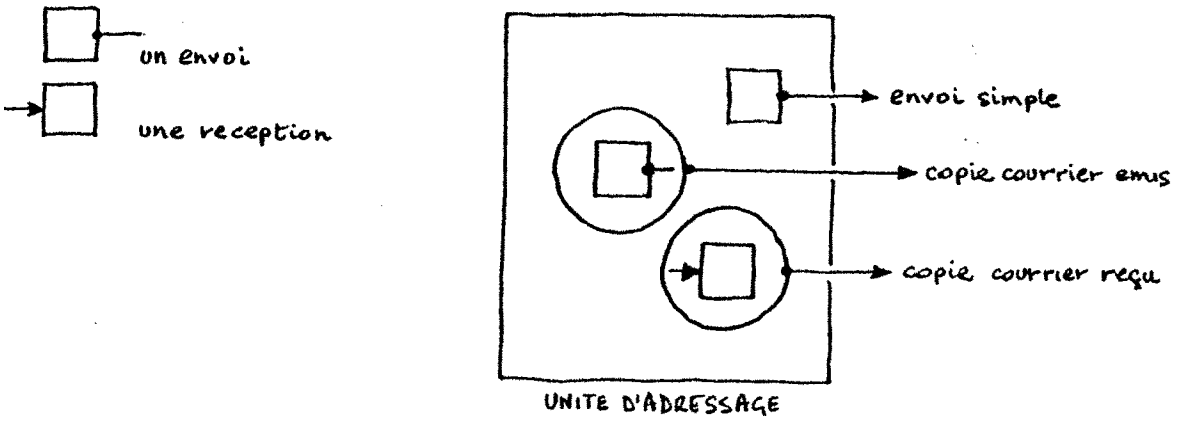
On identifie plusieurs sortes de documents :

- les documents "simples" qui sont :
  - . soit élaborés au sein d'une unité de gestion,
  - . soit élaborés à l'extérieur, l'unité de gestion en assurant simplement une nouvelle diffusion.

- Les documents qui sont une copie d'un courrier émis : l'information transmise est non seulement le document mais également sa diffusion.

- Les documents qui sont une copie d'un courrier reçu : l'information transmise est le document et son origine.

Schématiquement, l'"activité courrier" d'une unité de gestion (équivalente à une unité d'adressage) peut se représenter de la façon suivante :



La distinction de ces différentes sortes d'informations n'est pas indépendante de la présentation de ces informations : on trouve des lettres (simples) et des lettres accompagnées d'un bordereau d'envoi ; ce dernier cas ne peut correspondre qu'à du courrier interne ; on remarque également que l'archivage consiste souvent à ne conserver que le bordeau (le document lui-même ayant été ré-utilisé).

On s'est intéressé à une activité de 6 mois, soit 399 documents, répartis en 174 documents arrivés et 225 documents émis.

On a choisi de saisir, pour chaque document, une partie du texte en représentant le contenu. Pour les lettres, il existe une information "objet...", et pour les bordereaux d'envoi, une zone réservée à la correspondance, où l'on trouve souvent "veuillez trouver ci-joint..."; ce sont ces informations qui ont été saisies. De nombreux documents ne possédaient pas ces informations, et on a choisit de les ignorer. On est resté en présence de 136 documents arrivés et 124 documents émis, soit 260 documents.

On aurait pu choisir de saisir les informations figurant non pas sur les documents, mais les informations inscrites par la secrétaire sur les livres courrier arrivée/courrier départ. Ces informations étaient de meilleure qualité et correspondraient aux informations introduites dans un "vrai" système : l'utilisateur fournit une information textuelle lorsqu'une information analogue ne peut être "prise" dans le document. On s'est placé dans le cas le plus défavorable, l'étude voulant plus être l'étude de la réalité des documents en dehors de toute contrainte ou de tradition de présentation, plutôt que l'étude de textes particuliers produits par un seul individu.

Caractéristiques des documents ainsi saisis :

- 260 documents
- Longueur variant de 1 à 91 mots
- Nombre total de mots : 4400
- Nombre de mots différents : 1051
  (mot = chaîne de caractères alphanumériques comprise entre 2 délimiteurs)

Exemple de documents :

JE VOUS [illegible] TROUVER CI-JOINT PROCES-VERBAL D'EXAMEN + LETTRE D'ACCOMPAGNEMENT.

COPIE DE LA LETTRE DU 3 JANVIER 1978 EMANANT DE L'UNIVERSITE D'ALGER-FACULTE DES SCIENCES-DEPARTEMENT DE CHIMIE ET RELATIVE A LA TENUE D'UN SEMINAIRE DE CRISTALLOGRAPHIE. PROGRAMME DU SEMINAIRE.

VEUILLEZ TROUVER CI-JOINT EMPLOI DU TEMPS DES ENSEIGNANTS DE BIOPHYSIQUE ET PHYSIQUE.

NOTE RELATIVE AUX TITRES DE PASSAGES GRATUITS.

AFFILIATION DU PERSONNEL AU ASSURANCES SOCIALES/REGIMES GENERAL CASORES.

IMMATRICULATION DU PERSONNEL STAGIAIRE ET TITULAIRE NON ENCORE DECLARE A LA CSSF.

REGULARISATION SITUATIO+ AUPRES DE LA MUTUELLE POUR LES PERSONNELS CONTRACTUELLES STAGIAIRES ET TITULAIRES.

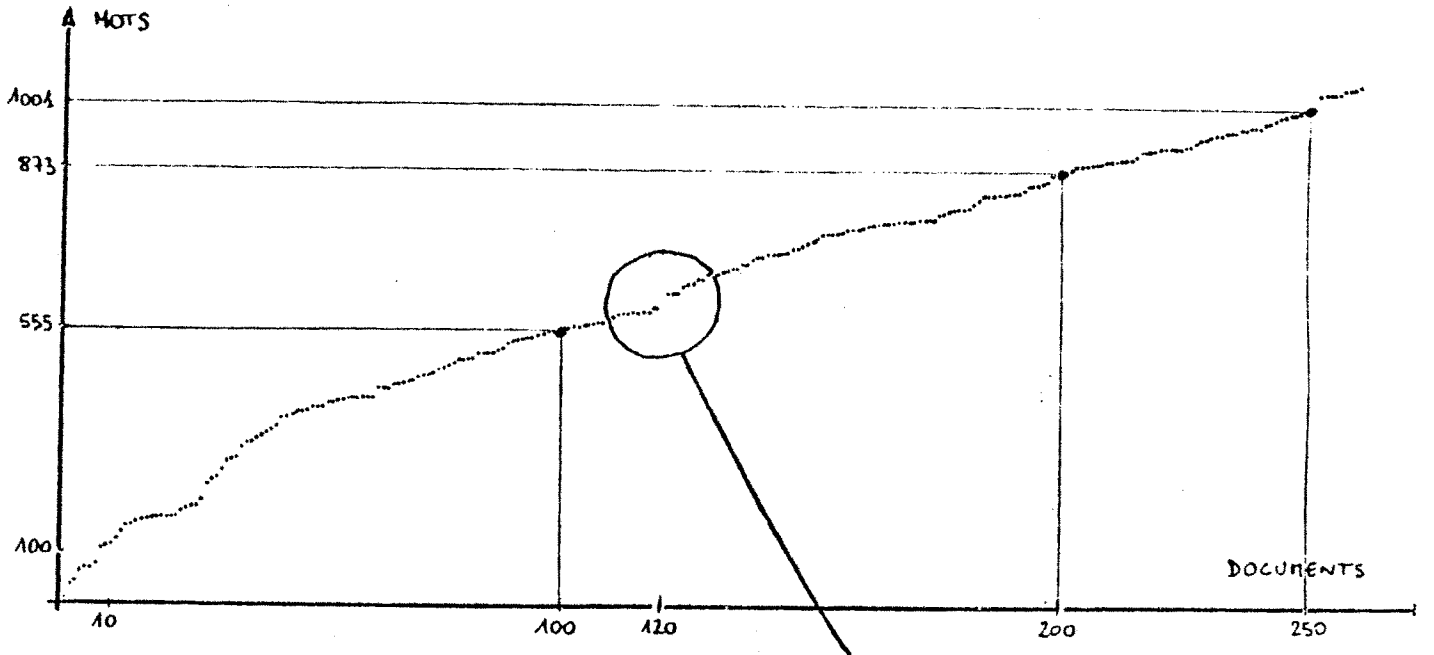
PRESTATIONS ASSURANCES MALADIES POUR PERSONNEL STAGIAIRES ET TITULAIRES DEJA AFFILIE A LA CSSF.

EXTRAIT DU PROCES-VERBAL DU CONSEIL D'UNIVERSITE DU 24 JANVIER 1978

CONGE DE MALADIE

CONGE DE MALADIE

Evolution nombre de mots/nombre de documents :



VEUILLEZ TROUVER CI-JOINT
DEMANDE CONCERNANT MR CLEMENT
GERARD POUR UN CONGE COMPLEMENTAIRE

SOIT TRANSMIS:
LES QUITUS D'INSTITUT, DE LOGEMENT ET

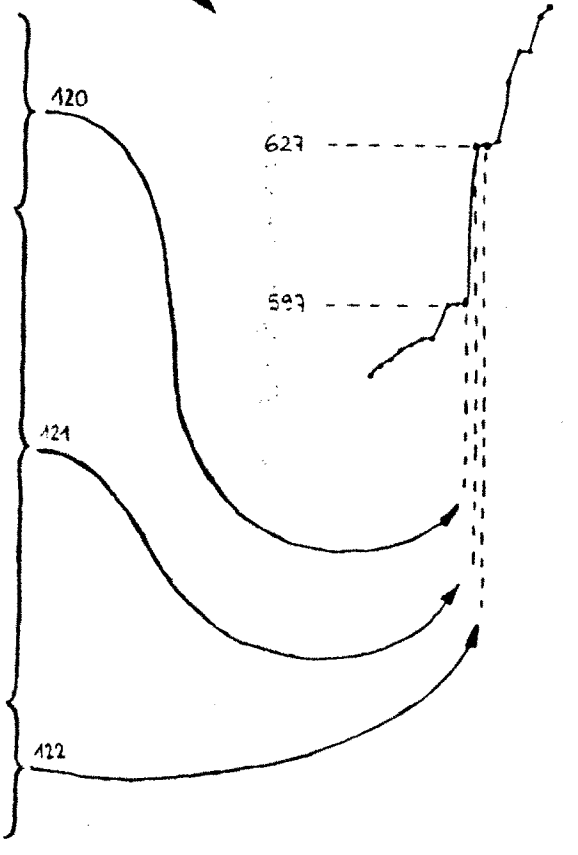
OS BIBLIOTHEQUE DES ENSEIGNANTS DE
L'INSTITUT DONT LES NOMS SUIVENT:
EMERY JOEL

SUITE A VOTRE CIRCULAIRE NO 78/659/SG DU 11
JUIN 1978 RELATIVE AUX VACANCES D'ETE,
J'AI L'HONNEUR DE VOUS TRANSMETTRE LA LISTE
DES BENEFICIAIRES:
POUR LA PREMIERE PERIODE DU 6.7.78 AU 5.8.78 INCLUS
FERCHICHI HASSIBA
BENINI MOURAD
BOUDERSA FARIDA
BOUDERSA HAFIZA
LAHLAH RAHAH
FELLOUS AMOR
CHEMACHEMA GARMI
POUR LA SECONDE PERIODE DU 6.8.78 AU 5.9.78 INCLUS
BOUGHACHICHE HADDI
DERBANI LAMRI
MERROUCHE AKILA
RAIS ABDELHAFID
BENSALEM MOHAMED
ZIAD AMARA
VEUILLEZ AGREER, MONSIEUR LE DIRECTEUR,
L'EXPRESSION DE MES SENTIMENTS CORDIAUX.

VEUILLEZ TROUVER CI-JOINT :
DOSSIER DE RENSEIGNEMENT CONCERNANT
MR EMERY JOEL, MAITRE ASSISTANT A
L'INSTITUT DE PHYSIQUE.

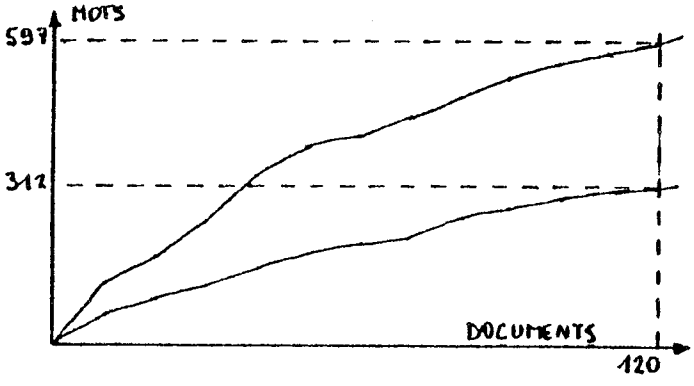
DEMANDE D'ACHAT

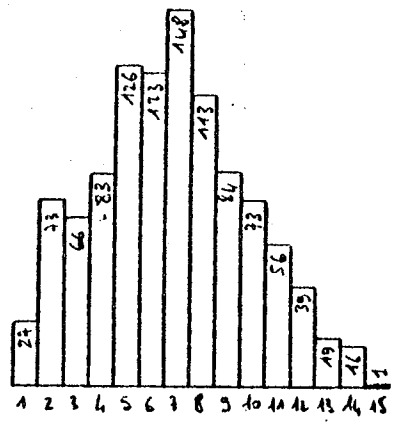
VEUILLEZ TROUVER CCJOINT L'AUTORISATION DE
TRANSFERT DE FONDS DU 21 NOVEMBRE 1978,
CONCERNANT VOTRE FACTURE NO 703 A 711 DE 53.212,00 FF.
VEUILLEZ CROIRE, MESSIEUR, A NOS SENTIMENTS
DISTINGUES



Les documents d'une longueur importante sont pratiquement tous ceux correspondant à des bordereaux d'envoi sur lesquels figurent des listes de noms. Choisir de les éliminer, c'est à dire choisir un système favorisant un taux de rappel élevé, peut constituer par exemple à prendre en compte uniquement les 10 ou 20 premiers mots des textes.

On obtient alors une réduction du vocabulaire (613 mots) et une évolution plus lente.

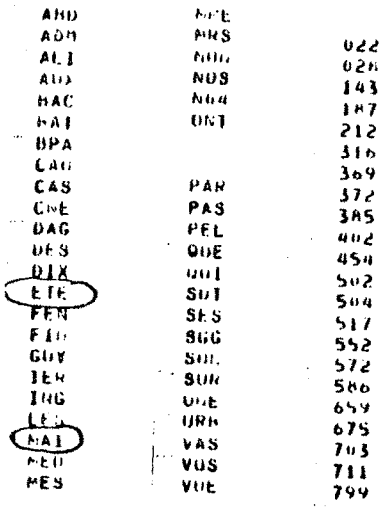




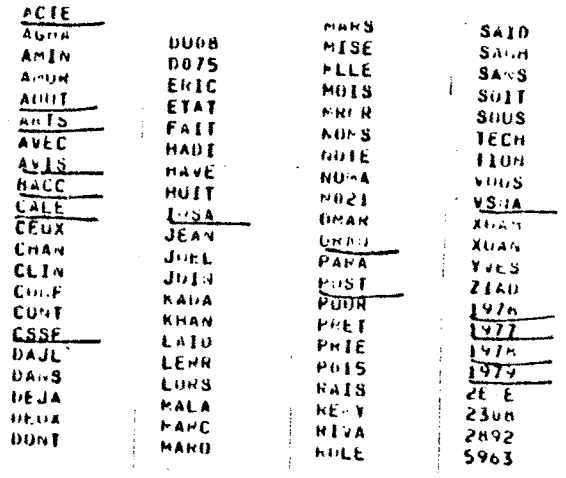
Examen taille des mots/mots-vides

Répartition de la taille des mots

Longueur des mots	1	2	3	4	5
Nombre de mots	27	73	66	83	126
Nombre de mots vides (avis subjectif)	27	73	64	68	83
Taux d'erreur si on considère tous ces mots comme vides	0 %	0 %	3 %	18 %	34 %
Réduction du vocabulaire en supprimant les mots d'une longueur inférieure à	2 %	9,5 %	15,7 %	23,6 %	35,6%



Les mots de 3 lettres



Les mots de 4 lettres

Examen troncature/même concept

Si on considère que 2 mots représentent le même concept s'ils ont leurs premiers caractères identiques →

on perd →

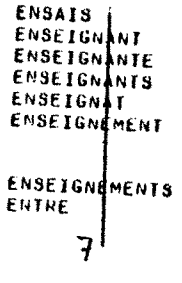
notions, soit →

et on réduit le vocabulaire de →

soit →

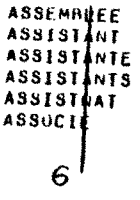
5	6	7	8
14	6	2	2
1,3 %	0,5 %	0,2 %	0,2 %
142	123	93	69
13,5 %	11,7 %	8,8 %	6,6 %

Exemples :



7

Satisfaisant



6

Satisfaisant

5

Erreur

ANNEXE C - UNE COLLECTION DE DOCUMENTS TECHNIQUES

Collection d'articles scientifiques et techniques, commune à 5 personnes du Centre Scientifique CII-HB de Grenoble. A la fin du 1er semestre 1980, il existait 309 documents. Les documents sont écrits en français et en anglais.

La taille des documents originaux varie généralement de quelques pages à quelques dizaines de pages. Ils ne possèdent pas tous des résumés.

Chaque document est rangé dans un dossier dans des classeurs métalliques, et chaque dossier porte un nom.

Les informations suivantes ont été saisies : le titre, le ou les auteurs, la date, quelquefois le lieu, quelquefois quelques mots (clés ?) destinés à améliorer la qualité du titre, et le nom du dossier où est rangé le document. Aussi la taille des informations saisies varie de 10 à 30 mots par document.

Exemple de documents

```
BARTHET MF
      BARTHET KARAOUCH ROTHENBURGER
      APPROCHE DU PROJET D'ACCES A DES SYSTEMES DOCUMENTAIRES REPARTIS
      (A.S.D.D.A) FEVRIER 1978.
BOCHMANN G
      GECSEI J.          A UNIFIED METHOD FOR THE SPECIFICATION AND VERI
      VERIFICATION OF PROTOCOLS.
BERT D
      JACQUET .
      GENERIC ABSTRACT DATATYPES.
      RAPPORT DE RECHERCHE IMAG NOVEMBRE 1977.
BOUCHET P.
      RAPPORT FINAL CONTRAT SESORI 76.050 :
      PROCEDURES DE REPRISE ET FIABILITE DES BASES DE DONNEES.
BENNETT J
      A WEEK OF CYCLADES DATA - AS SEEN BY AN IRISA.
      NOVEMBER 1977.
BACHMAN C
      PROGRESS ON THE PRESENTATION CONTROL / CONCEPTUAL SCHEMA/
      DATA INDEPENDANCE PROJECT.
      25/8/77
BOYD D
      PIZZAPELLO, VESTAL. HONEYWELL
      RATIONAL DESIGN METHODOLOGY.
      WELLMADE.
BOYD D
      PIZZAPELLO.
      INTRODUCTION TO THE WELLMADE DESIGN METHODOLOGY.
      MAY 1978.
BERNSTEIN P
      FOK,LANDERS,GOODMAN,HAMMER,REEVE,ROTHNIE,SHIPMAN,WONG.
      A DISTRIBUTED DATABASE MANAGEMENT SYSTEM FOR COMMAND AND
      CONTROL APPLICATIONS. 1/2 ANNUAL TECHNICAL REPORT.30/4/77.
BOSC P
      CHAUFFAUT A.     INTERROGATION DE FICHIERS REPARTIS SUR
      UN RESEAU DE CALCULATEURS HETEROGENES.
```

Méthode d'indexation :

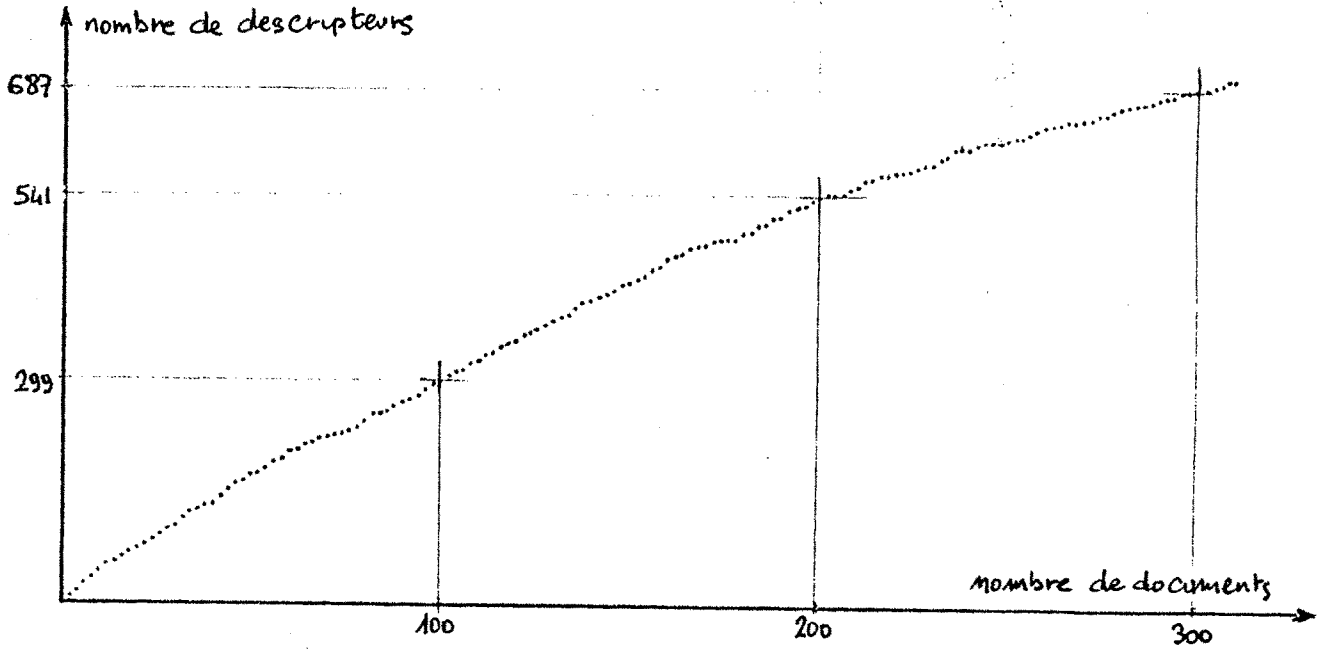
Prendre les 10 mots les plus longs, limités à 12 caractères, supérieurs à 3 caractères, n'appartenant pas à une liste de mots-vides et différents entre eux (moins de 75 % de caractères identiques à la même place dans les 2 mots : au-dessus de 75 %, les deux mots sont considérés comme étant les mêmes) ; ces mots sont proposés à l'utilisateur qui peut en ajouter ou en supprimer, avec les mêmes contraintes.

Exemple de descripteurs :

ARRAGON	ASSIGNMENT
ASTRAHAN	ATOMIC
ATTRIBUTE	AUTOMATION
AUTONOMUS	BABIC
BACHMAN	BAKER
BALTER	BANATRE
BANCILHON	BANKS
BARTHET	BASES DE DON
BASSET	BEKKERS
BELFORD	BELL
BENEFITS	BERKOWITZ
BIBLIOGRAPHI	BIBLIOTHEQUE
BILAN	BILLER
BISKUP	BLASGEN
BOGO	BOUK
BORGEN	BOYCE
BRAGG	BRAHAM
BRIAT	BROOKS
BSC	BTREE
BUILDING	BULLETIN
BUNEMAN	BUREAUTIQUE
BUROTIQUE	BUS
CACH	CALCUL
CALCULATEURS	CALECA
CAP	CAPABILITY

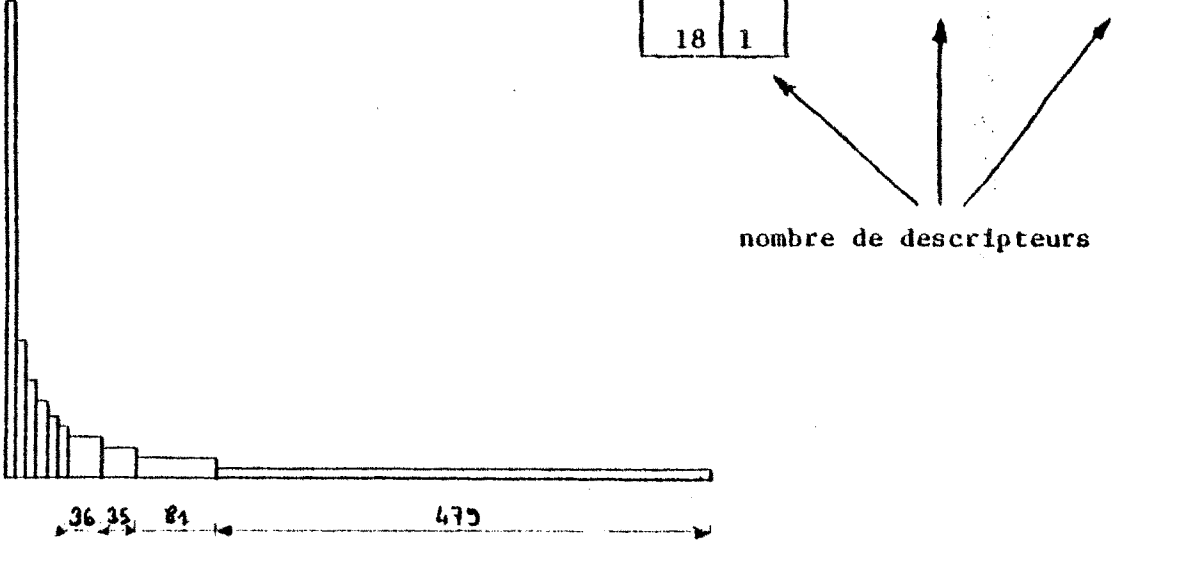
On remarque en particulier que les deux mots BOCHMAN et BACHMAN ont été considérés comme un même descripteur BACHMAN

Evolution du vocabulaire :



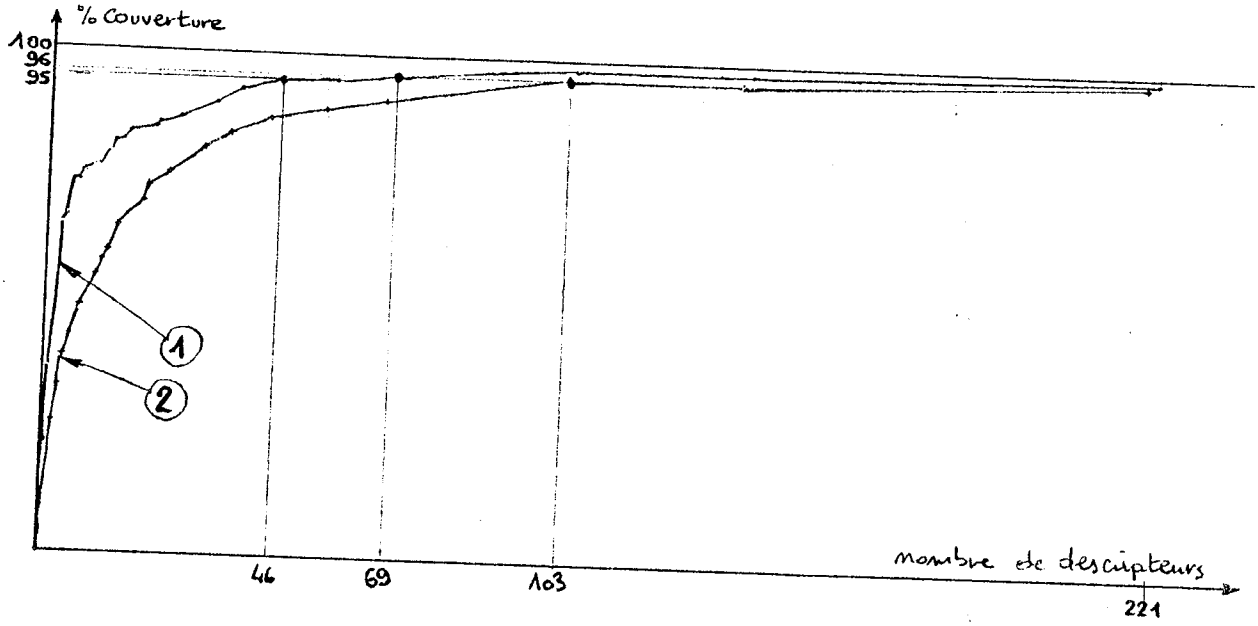
Répartition du vocabulaire :

On indique le nombre de descripteurs qui correspond à chaque nombre d'occurrences. L'importance de certains groupes de descripteurs est représentée par la surface des rectangles sur la figure ci-dessous.



Evaluation de la couverture :

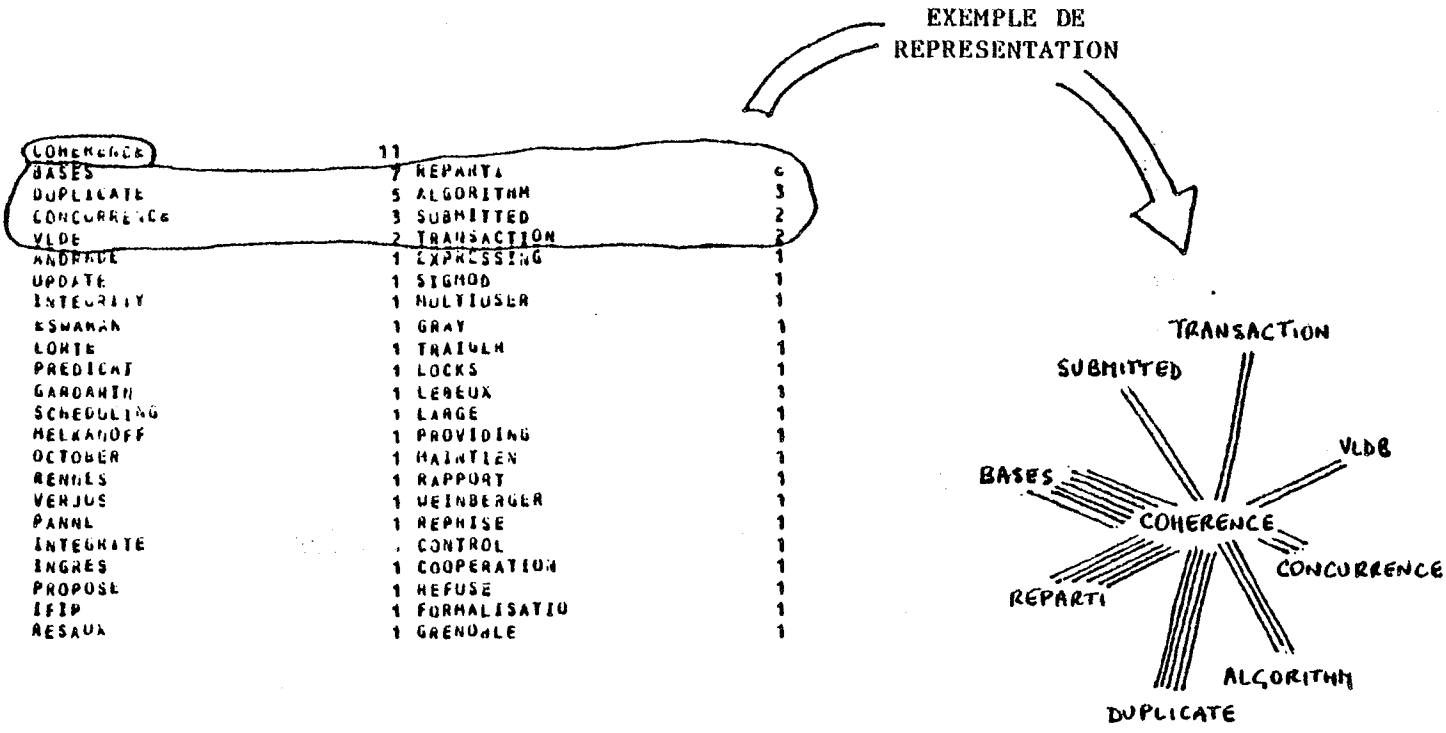
On examine le pourcentage de documents "couverts" par les "n meilleurs" descripteurs.



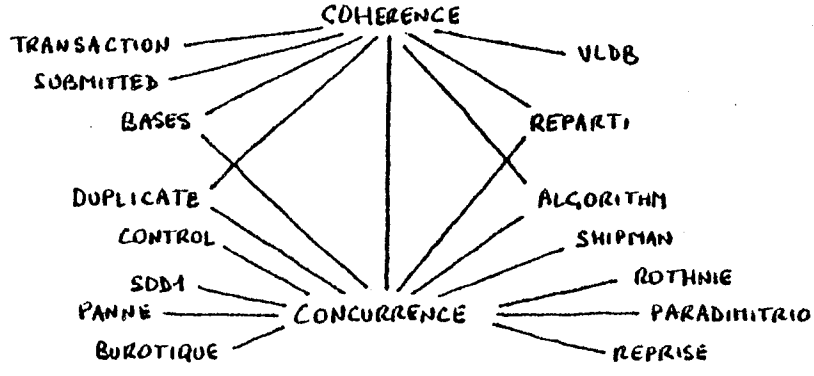
Ces courbes correspondent à l'utilisation des n descripteurs les plus fréquents : pour la courbe 1 , ils ont tous été utilisés ; pour la courbe 2 , les 2 descripteurs les plus fréquents (nombre d'occurrences 116 et 91) n'ont pas été utilisés.

Exemple de co-occurrence :

Liste des descripteurs apparaissant dans les descripteurs où apparait le descripteur COHERENCE.



Exemple : descripteurs liés au descripteur COHERENCE, par obtention des descripteurs co-occurrent de CONCURRENCE (nombre de co-occurrence > 1).



ANNEXE D - UN SYSTEME DE CLASSEMENT

Système de classement à caractère expérimental, en cours de réalisation au département informatique de l'Université de Constantine, pour gérer/classer des documents de secrétariat analogues à ceux décrits à l'annexe B. Réalisation en Fortran sur Mitra 15.

Caractères de la classification des documents

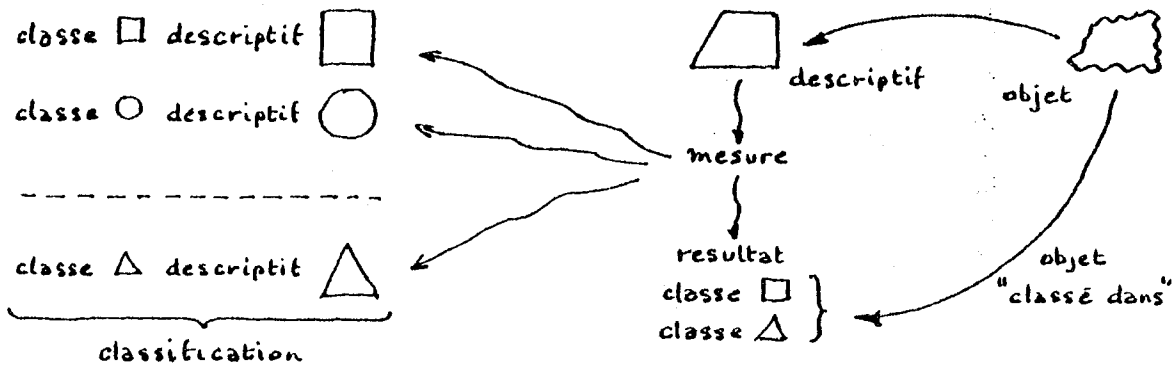
Les différentes classes dans lesquelles seront répartis les documents sont peu nombreuses et choisies manuellement a priori ; on veut arriver à une quasi-partition des documents, le résultat d'une recherche étant l'obtention des documents d'une classe.

Caractères du système

Constatation : certains textes sont de très mauvaise qualité.

Conclusion : il faut profiter de la présence de l'utilisateur lors de l'entré document (c'est lui qui frappe le texte "titre" ou "objet" représentant le document dans le système) pour apporter de l'information complémentaire si besoin est : le critère pour apporter cette information étant l'appréciation de la justesse du classement effectué automatiquement.

Rappel de l'algorithme de classement



Principe du traitement

Un document est classé ou non-classé ; plus précisément, classer un document consiste à l'assigner automatiquement à une ou plusieurs classes, ou bien à l'assigner manuellement à une classe particulière appelée INCLASSABLE.

On peut représenter ce traitement ainsi :

```
tantque  satisfaction-utilisateur = mauvais faire
   début
      proposer-classement
                satisfaction-utilisateur ← réponse-de-l'utilisateur
      si satisfaction-utilisateur = mauvais alors
         choix
             choix 1 : modifier-texte-entrée
             choix 2 : classement-autoritaire-dans-inclassable
                       satisfaction-utilisateur  bon
         fchoix
   fin
```

Ceci montre :

- La mise en oeuvre d'un mécanisme d'apprentissage, dans le sens où ce mécanisme permet d'améliorer la qualité des textes : on peut donc penser que les documents pourront être reclassés automatiquement de manière satisfaisante.

- La nécessité d'un apprentissage/d'une adaptation de l'utilisateur à la machine (méthode relativement autoritaire...), permettant aussi de bien lui faire comprendre qu'il n'y a pas de miracles : de "bons" descriptifs de classes et de "bons" descriptifs de documents amèneront un "bon" classement.

Remarque :

le mauvais classement d'un document peut provenir de mauvais descriptifs de classes. On peut les modifier. Ceci entraîne le déclassement de certains documents. Pour éviter un processus récursif, cette possibilité de modification n'est pas autorisée dans la boucle "tant que" ci-dessus.

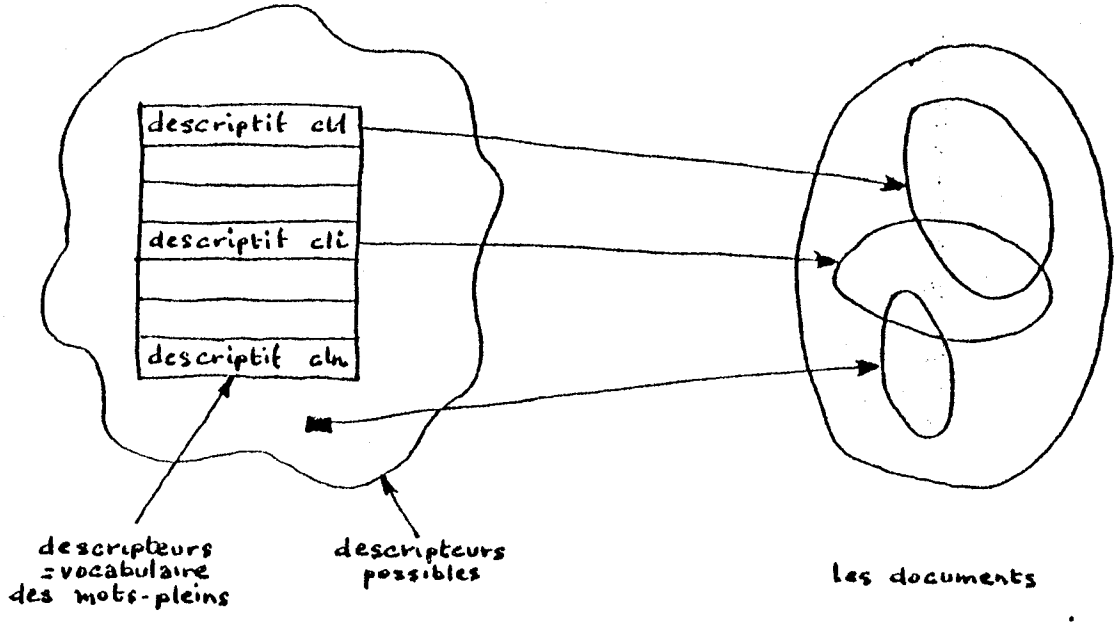
Choix pris

- Texte d'entrée : choisi par l'utilisateur de la même façon que lorsqu'il remplit son livre courrier départ/courrier arrivée), généralement très semblable au "titre" ou à l' "objet" figurant sur le document papier.

- Méthode d'indexation dérivative : les mots d'une longueur ≤ 3 caractères sont vides, les mots ayant leur 6 premiers caractères identiques sont considérés comme étant les mêmes.

- Principe de classement : assignation à une classe par reconnaissance de mots-pleins, c'est-à-dire, assignation à une classe si descriptif-classe $\cap$ descriptif-document $\neq \phi$, ce qui "implique presque" que $\bigcap_i$ descriptif-classe-i $= \phi$, autrement dit que les descriptifs de classes déterminent une partition du vocabulaire des mots-pleins ; en fait, on prend cette propriété comme hypothèse simplificatrice.

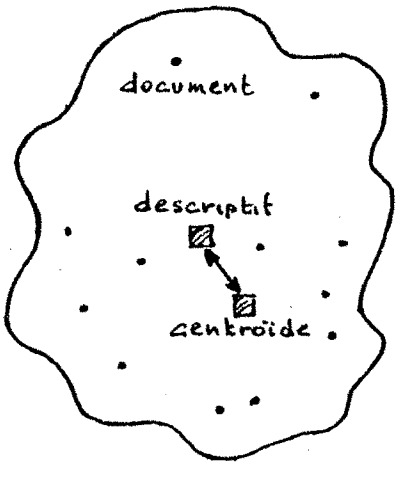
Illustration



Particularités

1) Avec ces hypothèses, lorsqu'on ajoute/retranche un descripteur (mot-plein) au descriptif d'une classe, les documents qui possèdent ce descripteur dans leur descriptif peuvent être/peuvent ne plus être classés dans cette classe. Il faut donc les re-classer, mais ceci nécessite de posséder la relation descripteurs-possibles —— document, c'est à dire tous les descriptifs de documents.

2) Afin d'aider à l'amélioration de la qualité des descriptifs de classes (qui sont déterminés a priori, avec l'aide éventuelle d'outils décrits à l'annexe C), on conserve, pour chaque classe, son "centroïde" ou centre de gravité, défini comme la somme des descriptifs des documents de la classe.



Ceci permet dans un premier temps de visualiser la répartition des descripteurs possible et d'ajuster manuellement les descriptifs des classes, et dans un second temps, de mettre en oeuvre des outils plus puissants pour déterminer les descripteurs possibles les plus discriminants, ceux qui sont inutiles..., le but étant de faire évoluer l'un vers l'autre le descriptif et le centroïde de la classe, leur similarité étant caractéristique d'une "bonne" classification [SAL 75].

La conservation du centroïde de chaque classe (qui, rappelons le, sont peu nombreuses), permet de posséder une bonne approximation de la relation descripteur-possible —— document, et permettre ainsi la conservation d'un classement toujours "à jour".

Ainsi :

- lorsqu'on veut ajouter un mot à un descriptif de classe,

  . si ce mot est déjà descripteur, c'est impossible par hypothèse : un descripteur ne peut apparaître dans plusieurs descriptifs.

  . si ce mot n'est pas déjà descripteur, donc inconnu ou "descripteur possible", il peut devenir descripteur : il faut reclasser tous les documents des classes où ce mot apparaît dans le centroïde (il existe alors au moins un document qui possède ce mot). Ce processus ne peut être long puisque, par définition, pour devenir mot-plein, un mot ne devrait pas être très dispersé parmi des documents de classes différentes.

- Lorsqu'on veut enlever un descripteur à un descriptif de classe, il faut ensuite reclasser les documents de cette classe.

Un document se trouve alors dans un des états suivants :

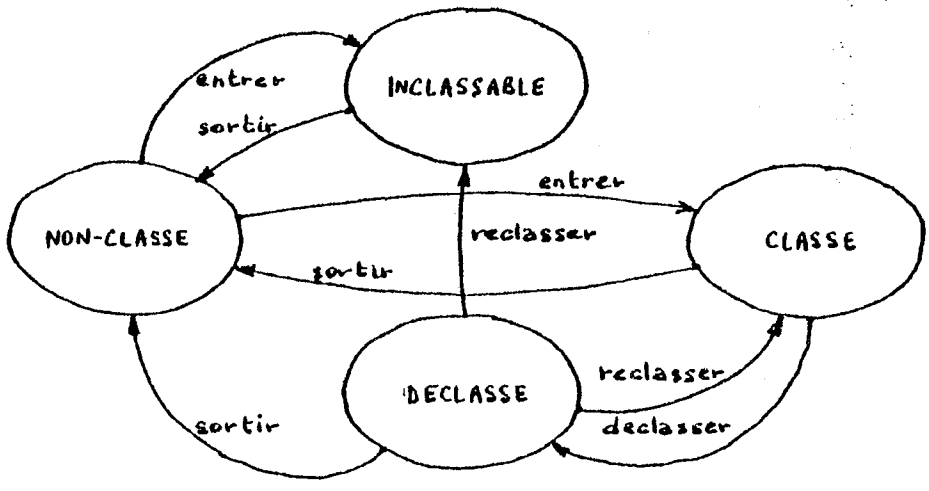
. non-classé : n'est pas entré dans le système,

. classé : est classé dans une ou plusieurs classes (ou dossiers),

. inclassable : n'est pas classé dans un dossier,

. déclassé : a été classé ; est en attente d'être reclassé

que l'on peut représenter ainsi avec des notations évidentes.



Les commandes sont :

- d'entrée/sortie d'un document

  • ENTRER < texte de document >
  entrer et classer un document

  • SORTIR < référence de document >
  sortir un document du système

- de recherche

  • LSTDOS < nom de dossier > | TOUS
  obtenir les documents d'un dossier, ou tous les noms des dossiers

- de gestion de la classification

  • CREDOS < nom de dossier > , < descriptif >
  créer un dossier

  • SUPDOS < nom de dossier >
  supprimer un dossier ; les documents de ce dossier deviennent déclassés

  • MODDOS < nom de dossier > , < descriptif >
  modifier le descriptif d'un dossier ; certains documents peuvent être reclassés

  • RECLAS < référence de document > | TOUS
  reclasser un document déclassé ou inclassable, ou tous les documents déclassés

  • LSTDES < nom de dossier >
  lister le descriptif d'un dossier

  • LSTCEN < nom de dossier >
  lister le centroïde d'un dossier

- de mise au point

  • LSTDOC < référence de document > | TOUS
  lister un document, ou tous les documents de la base

Un scénario

```
?entrer
TEXTE : veuillez trouver ci-joint = dossier de renseignement concer-
nant M. DUPONT, maître-assistant à l'Institut de Physique
DOCUMENT 121
CLASSE DANS INCLASSABLES - D'ACCORD ? oui
?lstdos
TOUS OU NOM : tous
NOM DES DOSSIERS ET VOLUME
ENSEIGNANT 17
MINISTERE 22
.
.
.
DECLASSES 25
INCLASSABLES 3
?lstdes
NOM DE DOSSIER : enseignant
DESCRIPTIF :     ENSEIGNANT   CONTRACTUEL   PROFESSEUR
                 VACATAIRE    AUXILIAIRE
?moddos
NOM DE DOSSIER : enseignant
NOUVEAU DESCRIPTIF : enseignant contractuel professeur vacataire
auxiliaire assistant maître conférence
19 DOCUMENTS DANS CE DOSSIER
?reclas
TOUS LES DECLASSES OU NUMERO : 121
CLASSE DANS ENSEIGNANT - D'ACCORD? oui
```

ANNEXE E - UN SYSTEME DOCUMENTAIRE

"Mini" système documentaire réalisé au Centre Scientifique CII-HB de Grenoble pour gérer la collection de documents décrite à l'annexe C. Réalisation en PL/1, sur IRIS 80, avec une structure de fichier séquentiel.

Exemple d'articles du fichier

```
A004   AND?9Z    08  08  ANDRE E
 CONCEPTUAL /OFFICE      /AUTOMATION /INTEGRATED /VERSAILLES /BOGO        /MANEUN      /BUREAUTIQUE /
 CONCEPTUAL APPROACH TO OFFICE AUTOMATION.
 INTERNATIONAL WORKSHOP ON INTEGRATED OFFICE SYSTEMS.
 VERSAILLES SEPTEMBER 1979. BOGO MANEON

A005   ADI?9Z    10  08  ADIBA M
 POLYPHEME   /DISTRIBUTED /DATABASE    /DESIGN      /IMPLEMENTATI/ANDRADE     /DECITRE     /FERNANDEZ   /NGUYEN      /TOAN
 POLYPHÈME : AN EXPERIENCE IN DISTRIBUTED DATABASE SYSTEM
 DESIGN AND IMPLEMENTATION.
 ANDRADE DECITRE FERNANDEZ NGUYEN-GIA-TOAN.

A006   AND08Z    08  08  ANDRE E
 POLYPHEME   /DISTRIBUTED /EXECUTION   /MONITOR     /DATABASES   /VERSAILLES /MARCH       /DECITRE     /
 POLYPHEME PROJECT THE DEM DISTRIBUTED EXECUTION MONITOR.
 INTERNATIONAL CONFERENCE ON DISTRIBUTED DATABASES .
 VERSAILLES MARCH 1980. DECITRE
```

Aspect logiciel documentaire

Recherche booléenne : la question est une expression booléenne non parenthésée (priorité des opérateurs).

Réponse en 2 parties :

- le nombre de documents répondant à la question,
- listage de ces documents sur demande.

Exemples d'interrogation

```
QUESTION? local&reseau
AUCUN DOCUMENT SELECTIONNE, VERIFIEZ VOTRE QUESTION


QUESTION? reseau
NOUS AVONS SELECTIONNE        24 DOCUMENTS
VOULEZ VOUS LES IMPRIMER ?
O/N?n


QUESTION? reseau!network
NOUS AVONS SELECTIONNE        77 DOCUMENTS
VOULEZ VOUS LES IMPRIMER ?
O/N?n

QUESTION? local&reseau!network
NOUS AVONS SELECTIONNE         1 DOCUMENT
VOULEZ VOUS  L' IMPRIMER ?
O/N?oui
IR16  NICOUD J.D.
BENEFITS OF A WORKING LOCAL NETWORK.
SYSTEMES INTEGRES DE BUROTIQUE VERSAILLES NOV 1979.
```

```
QUESTION? coherence!consistency&reparti!dis;tributed
NOUS AVONS SELECTIONNE        5 DOCUMENTS
VOULEZ VOUS LES IMPRIMER ?
O/N?oui
A@1@  ADIBA M.
ET ANDRADE J.M.  - EXPRESSING UPDATE CONSISTENCY IN DISTRIBUTED
DATABASES. SUBMITTED TO SIGMOD 80.


H@84  HERMAN D.
ALGORITHME DE MAINTIEN DE LA COHERENCE DE COPIES MULTIPLES.
FEVRIER 1979 RENNES  RAPPORT DE RECHERCHE.


P@@4  PARENT C.
INTEGRITE DES DONNEES D'UNE BASE REPARTIE.
DISTRIBUTED DATABASE CONSISTENCY.


S@@3  STONEBRAKER
CONCURRENCY CONTROL AND CONSISTENCY OF MULTIPLE COPIES OF DATA
IN DISTRIBUTED INGRES.


W@1@  WILMS PAUL
ETUDE D'ALGORITHME DE COHERENCE D INFORMATIONS DUPLIQUEES
ET REPARTIES FORMALISATION A L AIDE DE RESEAUX DE NUTT .
RR 160 IMAG FEV 79 . DISTRIBUTED CONSISTENCY CONCURRENCY
```

Aspect système documentaire :

L'outil principal est le thésaurus (dictionnaire de synonymes), construit manuellement par :

- choix d'un ensemble de descripteurs parmi l'ensemble des descripteurs obtenus par indexation dérivative (décrite à l'annexe C),
- répartition en classes d'équivalence (ou considérées comme telles),
- choix d'un descripteur, dit synonyme préférentiel pour représenter chaque classe.

Portion de thésaurus (240 classes) :

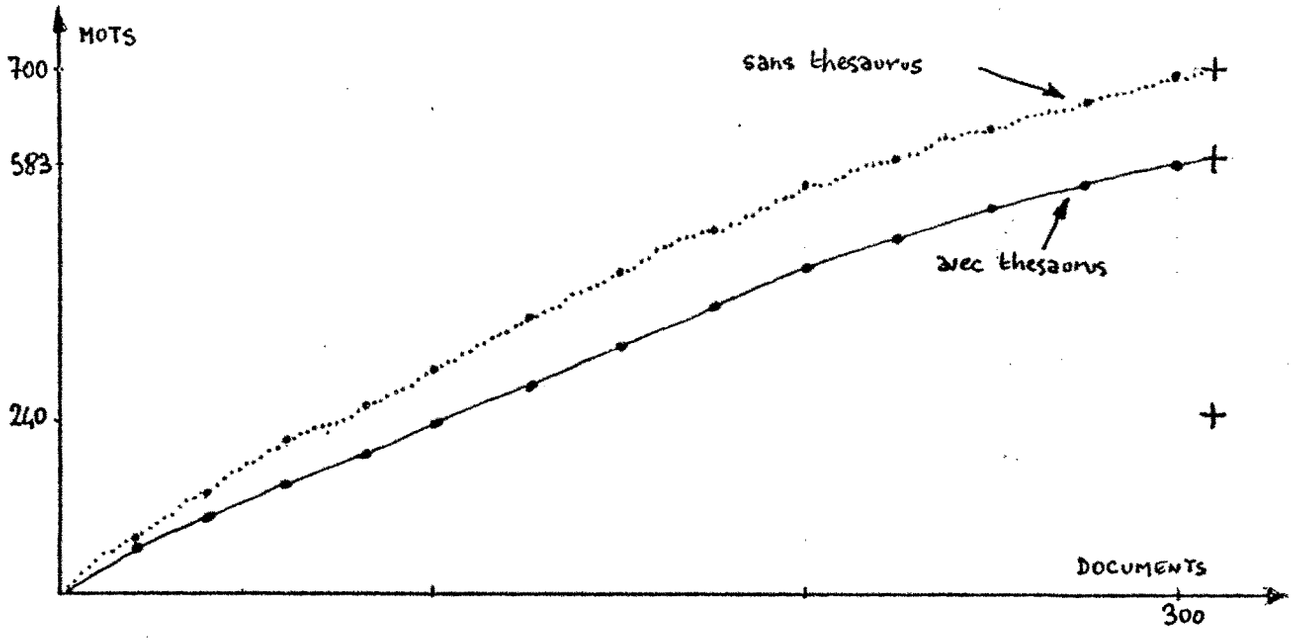
```
APPLICATION APPLIQUE    ;
APPROCHE    APPROACH    ;
ARBORESCENCE;
ARCHITECTUREDESIGN      ;
ARPA        ;
BUROTIQUE   BUREAUTIQUE AUTOMATION  OFFICE      ;
BASES       BANKS       DATA        DATABASE    SGBD        BDMS        ;

BIBLIOGRAPH ;
BIBLIOTHEQUECATALOGUE   DIRECTORY   ;
BTREE       ;
BULLETIN    ;
BUS         ;
CALCULATEUR ;
CALCUL      CALCULUS    ;
CAPABILITY  ;
CARACTERISTI;
CIGALE      PALAET      PAQUET      TRANSPORT   COMMUTATION PAQUETS     ;
CIIHO       CII-HB      ;
CLOCAS      HORLOGE     ;
CLUSTER     ;
CNAS        ;
CODASYL     DBTG        ;
COHERENCE   CONSISTENCY INCONSISTENT;
COMMERCIAL  ;
COMMITMENT  VALIDATION  ;
```

Mise à jour de la base par "restructuration" c'est à dire par production d'une nouvelle base (fichier séquentiel) : chaque descripteur est remplacé par son synonyme préférentiel s'il existe dans le thésaurus (ou sinon reste inchangé).

On remarque que les exemples d'interrogation donnés précédemment ont été réalisés sur la base originale (c'est à dire non restructurée).

L'utilisation du thesaurus permet de réduire le vocabulaire de descripteurs.



Le gain aurait été beaucoup plus important si le thésaurus avait été utilisé comme dictionnaire de mots pleins (suppression des descripteurs n'appartenant pas au thesaurus). Ici, tous les descripteurs de la base originale, et en particulier les noms d'auteurs ont été conservés.

ANNEXE F - COUT DU STOCKAGE - REMARQUES

La taille des collections de documents bureautiques est de l'ordre du millier : de quelques centaines de documents à quelques milliers (voir par exemple annexes B et C).

Stockage des textes :

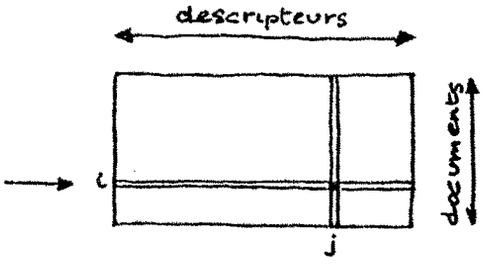
Dans l'exemple de la collection de l'annexe B, le choix qui a été fait est l'équivalence des notions de document, "texte pour indexer", "texte pour savoir" : c'est le titre des documents réels qui a été choisi. On avait prévu 256 octets par document (un secteur sur disque à cartouche), ce qu'on peut évaluer à presque 4 lignes pleines de texte, ce qui, a postériori est trop si l'on s'en tient effectivement seulement aux titres des documents. Si l'activité d'une année correspond à 1 000 documents, l'espace nécessaire au stockage des textes est de 256 x 1 000 = 256 Ko ; chiffre très intéressant à rapprocher de la capacité d'une disquette simple face, simple densité, aujourd'hui ($\approx$300 Ko).

Stockage des descriptifs :

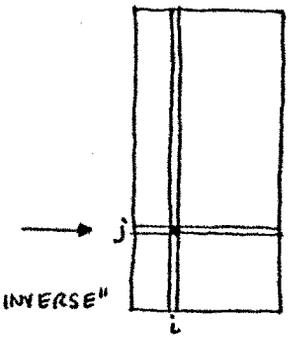
Les descriptifs sont des vecteurs binaires ; une base documentaire est un ensemble de descriptifs qu'on peut voir comme une matrice binaire BD telle que : BD (i,j) = 1/0 $\Leftrightarrow$ le i-ième (descriptif de) document possède/ne possède pas le j-ième descripteur.

Exemple de l'annexe C : 300 documents, 700 descripteurs, 10 descripteurs maximum par document ; il est évident qu'une telle matrice est presque vide.

Une structure pour représenter ces données a eu beaucoup de succès, le fichier inverse, qui est l'implémentation de la fonction d'accès descripteur $\longrightarrow\!\!\blacktriangleright$ descriptif, ou plus exactement descripteur $\longrightarrow\!\!\blacktriangleright$ (identificateur de) (descriptif de) document.



BASE DOCUMENTAIRE "BD"

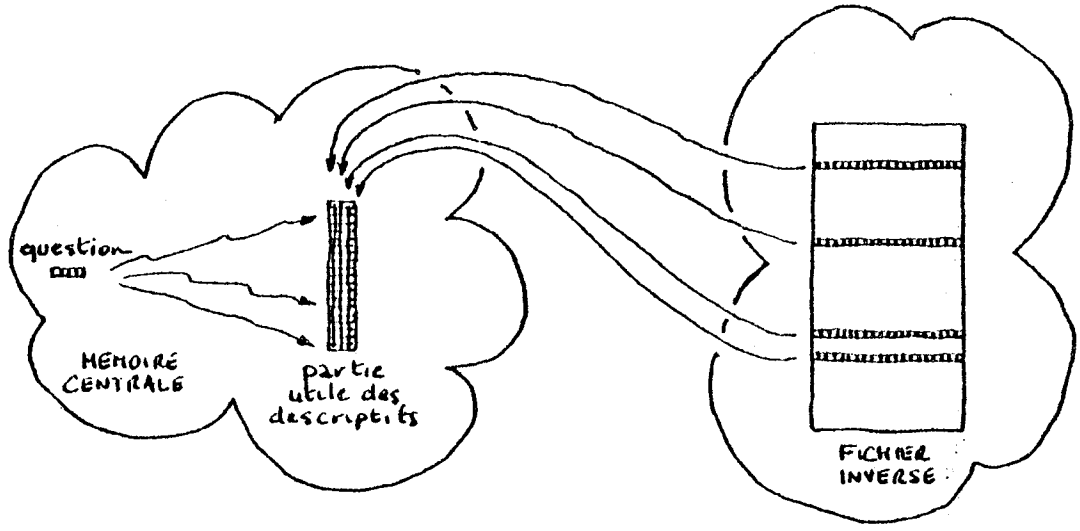


"FICHIER INVERSE"

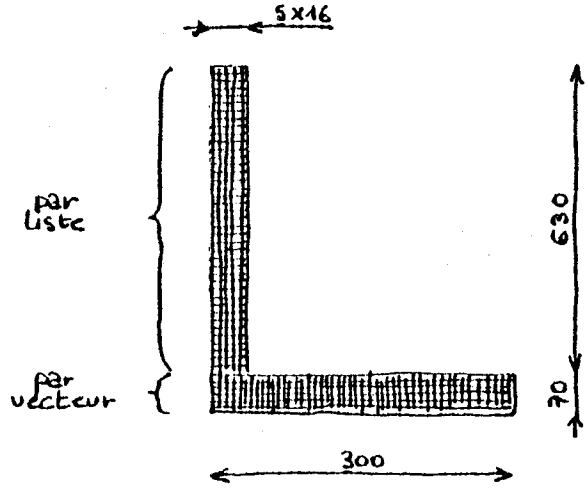
Cette structure n'est pas la panacée universelle ; la mise en oeuvre du "gros" serveur français à Sophia-Antipolis a "parait-il" réservé quelques surprises... voir également [BNT 78]... (taille des données, mises à jour...).

Elle est bien adaptée à la recherche booléenne, c'est évident ; par exemple le traitement d'une question "dj ou dk" consiste à faire l'union des lignes j et k du fichier inverse.

Ceci semble moins vrai pour une recherche par mesure de similarité où il faut effectuer toutes les comparaisons descriptif-question/ descriptif-document. Faire 1 000 opérations n'est pas un problème, faire 1 000 entrées-sorties pour obtenir les descriptifs risque d'en poser. En fait, avec une mesure "judicieusement choisie", seules sont concernées les parties de descriptifs correspondantes aux descripteurs utilisés dans la question. Le nombre de ces descripteurs reste faible et ces parties de descriptifs très courtes (vecteur binaire = chaîne de bits). On peut les charger en mémoire, le nombre d'entrées-sorties pour le faire est fonction du nombre de descripteurs utilisés dans la question, et la comparaison peut s'effectuer très rapidement. Cette structure de fichier inverse semble donc, là-aussi, bien adaptée.



Ce fichier inverse, matrice presque vide, ne sera pas représenté tel quel, mais plutôt comme une suite de listes : à chaque descripteur on fait correspondre la liste des identificateurs de documents qui possèdent ce descripteur. On peut également penser à une structre mixte : par exemple, dans la collection de l'annexe C, (300 documents, 700 descripteurs), seulement 10 % des descripteurs apparaissent dans plus de 5 documents ; en prenant 16 bits pour représenter un identificateur, on peut penser à la structure suivante :

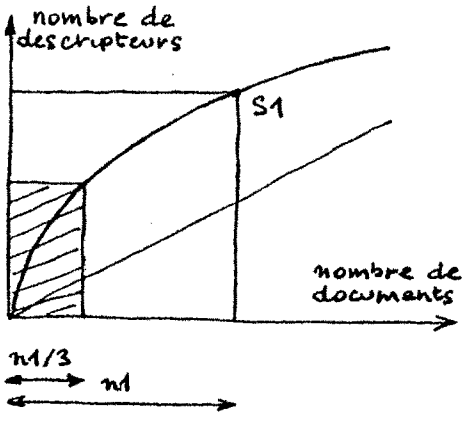


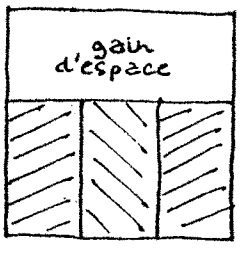
TAILLE = 5x16x630
+ 70x300
= 71400 bits
soit ≈ 9Ko

Calcul évidemment simpliste, mais il est intéressant de comparer le coût de stockage des descriptifs (9 ko) à celui des textes (256 ko).

Remarque :

En passant à un nombre de documents plus important, on peut avoir intérêt à diviser la base en plusieurs morceaux ; l'évolution du vocabulaire a l'allure représentée ci-dessous, et on est dans une situation de type S1 (nombre de descripteurs > nombre de documents) ; le découpage de la base permet un gain d'espace.





BIBLIOGRAPHIE

[ABH 79] Vers un schéma conceptuel de bureautique
E. ANDRE, G. BOGO, J. HAMEON
Congrès bureautique 79 - Grenoble
Mai 1979

[ABH 79.2] La bureautique : un exemple d'approche conceptuelle
E. ANDRE, G. BOGO, J. HAMEON
R.R. n°176 - USMG-INPG
Novembre 1979

[ABR 74] Data semantics
J.R. ABRIAL
IFIP Working Conference - Cargese, Corse
Avril 1974

[ABR 77] Manuel du langage Z
J.R. ABRIAL
Paris
Mai 1977

[ADBS 76] Logiciels et systèmes documentaires - Tomes I, II, III
Association Française des documentalistes et bibliothécaires spécialisés
Les Cahiers de l'ADBS
1976

[ADBS 79] 3ème Congrès National Français sur l'Information et la Documentation - Contributions aux tables rondes
ADBS - ANRT - BNIST
Mars 1979

[ADI 78] Un modèle relationnel et une architecture pour les systèmes de bases de données réparties
M. ADIBA
Thèse USMG - Grenoble
Septembre 1978

[BAC 77] The role concept in database models
C.W. BACHMAN
Note Honeywell
Avril 1975

[BARB 78] A distributed mailbox service
BARBER
EIN
Octobre 1978

[BART 78] Definition d'un langage pivot d'interrogation de systèmes documentaires hétérogènes
M.F. BARTHET
Thèse 3ème cycle - Université Paul Sabatier - Toulouse
Février 1978

[BNT 78] Text file inversion : an evaluation
R.M. BIRD, J.B. NEWSBAUM, J.L. TREFFTZ
4th Workshop on computer architecture for non-numeric processing
Août 1978

[BOB 63] Automatic Document Classification
H. BORKO, M. BERNICK
Journal of the ACM - Vol.10
1963

[BOB 64] Automatic Document Classification. Part II. Additional experiments
H. BORKO, M. BERNICK
Journal of the ACM - Vol.11
1964

[BOC 75] Le système documentaire SATIN
L. BOURRELY, E. CHOURAQUI
CNRS - URADCA - Aix-Marseille
Janvier 1975

[BOG 78] Agora : un système de courrier électronique sur le système de base de données réparties Polyphème
G. BOGO
Rapport de DEA - INPG
Septembre 1978

[BRA 77] Classement pratique
A. BRAUMAN
Editions Hommes et Techniques
1977

[BUC 78] Gutemberg - Manuel d'apprentissage
M. BUCHERON
IRIA
Juillet 1978

[BUR 78] Congrès Bureautique 78
AFCET - IMAG - IRIA
Cours de formation - Grenoble
Mars 1978

[BUR 79.1] Congrès Bureautique 79
Actes des conférences
AFCET - IMAG - IRIA
Mai/Juin 1979

[BUR 79.2] Séminaire international - Systèmes intégrés de bureautique
IRIA
Actes des conférences
Novembre 1979

[CHA 78] Les langages documentaires
J. CHAUMIER
Entreprise Moderne d'édition
Mars 1978

[CIG 80] Les besoins et attente des utilisateurs dans le domaine de la bureautique
CIGREF - Club informatique des grandes entreprises françaises
INRIA - Projet Pilote KAYAK - GAL.2.065
Avril 1980

[CII 79] Systèmes d'informatique
CII-HB
Magazine n° 30
Printemps 1979

[COS 67] L'analyse automatique des documents
M. COYAUD - N. SCOT-DECAUVILLE
Mouton
1967

[DEL 80] Dossier électronique - Le moyen d'exploiter l'information
J.L. DELOBEL
Correlative Systems International
1980

[EDM 61] Automatic abstracting and indexing - Survey and recommendations
H.P. EDMUNSON, R.E. WYLLYS
Communications of the ACM - Vol.4 - N°5
1961

[DUC 73] Le système TITUS II
J.M. DUCROT
Institut Textile de France
Septembre 1973

[EDW 77] Implementation of a new interactive on-line information Retrieval System : 3 RIP
M.E. EDSTROM, M.A. WALLIN
Royal Institute of Technology - Sweden
1977

[EGLT 79] An office communication system
G.M. ENGEL, J. GROPPUSO, R.A. LOWENSTEIN, W.G. TRAUB
IBM System Journal - Vol.18 - N° 3
1979

[ESA 78] Brochures commerciales sur serveur Frascati
ESA - ESRIN
1978

[FAU 77] Proposition pour la réalisation d'un système de catalogage collectif réparti
J.C. FAURE
Note STERIA
Octobre 1977

[FAU 78] Actions possibles dans le domaine de l'informatique documentaire répartie
J.C. FAURE
Note STERIA
Janvier 1978

[GEN 80] Le monde du bureau des années 80 ou les cols blancs face au grand jeu électronique
R. GENTON
Institut R. GENTON
1980

[GRA 78] Definition et utilisation du produit prototype PIAFDOC
E. GRANDJEAN
Equipe Intelligence Artificielle - IMAG
Novembre 1978

[HDA 77] Bases de données documentaires réparties - Modèle d'accès aux données
J. HAMEON - D. D'AGARO
Rapport de DEA - INPG
1977

[HER 77] Hermes message system user's guide
BBN - Cambridge - USA
Janvier 1977

[HIG 77] A view of Euronet
P.L. HIGGINSON
INDRA - University College
September 1977
(Document SIRIUS DOC.E.008)

[IFI 80] Functional model of a computerised message system
IFIP WG6.5 - Ad hoc Group 2
Janvier 1980

[INF 78] The Planet System - User's guide
INFOMEDIA
1978

[IRI 71] L'informatique documentaire
IRIA
Cahier n° 7
Novembre 1971

[IRI 79] Bulletin de l'IRIA
IRIA
Mai 1979

[ISO 78] Reference model of open systems interconnection - V3
ISO
Novembre 1978

[KAR 79] Etude du système accès réparti à Mistral conversationnel
A. KARMOUCH
Rapport fin de contrat - CERISS - Toulouse
Mars 1979
(Document SIRIUS DOC.1.008)

[KOP 78] Compatibilité entre langages documentaires - Travaux réalisés dans le cadre des thesaurus multilignes - Bibliographie
G. KOPF
Note SIRIUS - Sous-groupe documentation répartie
Mars 1978

[KUO 79] Message services in Computer Networks
F. KUO
University of Hawa
1979

[MAR 61] Automatic indexing : an experimental inquiry
M.E. MARON
Journal of the ACM - Vol.8
1961

[MAR 76] The networking of interactive bibliographic retrieval systems
R.S. MARCUS, J.F. REINTJES
MIT - Cambridge - USA
Mars 1976
(Document SIRIUS DOC.E.006)

[MAR 77] Computer interfaces for user access to heterogeneous information retrieval systems
R.S. MARCUS, J.F. REINTJES
MIT - Cambridge - USA
Avril 1977
(Document SIRIUS DOC.E.005)

[MIS 78] Mistral : Manuel d'utilisation
CII-HB
La Documentation française
Octobre 1978

[NAF 79.1] Les services burotiques
N. NAFFAH
IRIA - Projet Pilote KAYAK - SRV.2.501
Mai 1979

[NAF 79.2] La burotique : définition et justification
N. NAFFAH
IRIA - Projet Pilote KAYAK - GAL.2.503
Juin 1979

[NAF 79.3] Analyse d'une application de messagerie
N. NAFFAH
IRIA - Projet Pilote KAYAK - MSG.2.502
Juin 1979

[NAF 79.4] Exemple d'un poste de travail burotique à interface universelle
N. NAFFAH
IRIA - Projet Pilote KAYAK - MEV.2.503
Juin 1979

[NAF 79.5] Les réseaux locaux en burotique
N. NAFFAH
IRIA - Projet Pilote KAYAK - REL.2.508
Juin 1979

[NEG 76] Study to determine the feasibility of a standardised command set for Euronet
A.E. NEGUS
INSPEC
Juin 1976
(Document SIRIUS DOC.E.007)

[NEG 77] Draft Euronet guideline : standard command for retrieval systems
A.E. NEGUS
INSPEC - London - England
Mai 1977
(Document SIRIUS DOC.E.011)

[PAI 77] Information retrieval and the computer
C.D. PAICE
Mac Donald and James - London
1977

[POLY] Notes internes projet POLYPHEME

[POL 75] POLYPHEME : propositions pour un modèle de répartition et de coopération de bases de données dans un réseau d'ordinateurs
Laboratoires CII-ENSIMAG-USMG
R.R. n° 29
Décembre 1975

[POL 79] POLYPHEME : un système de bases de données réparties
CII-HB - ENSIMAG, J.C. CHUPIN, C. DELOBEL, M.ADIBA...
Rapport de fin de contrat
Octobre 1979

[POU 75] Systèmes documentaires informatisés
B. POUSSOT
IREP Grenoble
Août 1975

[RIS 75] Information retrieval
C.J. RISSBERGEN
Butterworths - London & Boston
1975

[SAL 68] Automatic information organization and retrieval
G. SALTON
Mc Graw Hall Book Co
1968

[SAL 71] The Smart Project
G. SALTON
Prentice Hall
1971

[SAL 75] Dynamic information and library processing
G. SALTON
Prentice Hall
1975

[SAL 76] Cataloging software packages for automatic document processing
G. SALTON
Cornell University
Mars 1976
(Document SIRIUS DOC.E.003)

[SAL 78] Generation and search of clustered files
G. SALTON, A. WONG
ACM transactions on database systems
Decembre 1978

[SPA 71] Automatic keyword classification for information retrieval
K. SPARCK-JONES
Butterworths - London
1971

[STE 71] Automatic indexing - A state of the art report
M.E. STEVENS
US Department of Commerce Publication
1971

[SYR 79] Introduction aux systèmes informatiques répartis
INPG Grenoble
Session de perfectionnement - Chamrousse
Septembre 1979

[TOA 77] URANUS : une approche relationnelle à la coopération de bases de données
NGUYEN GIA TOAN
Thèse 3ème cycle - USMG Grenoble
Décembre 1977

[TOM 76] Réseaux d'Information - Etat de la question
A. TOMBERG
CCE - 2nd congrès européen sur les systèmes et réseaux documentaires
(Document SIRIUS DOC.E.001)

[TRY 71] La documentation automatique
J.P. TRYSTRAM
Dunod Economie
1971

[VEB 78] An electronic message system : where does it fit ?
A. VEZZA, M.S. BROOS
MIT Cambridge
Février 1978

[WHI 68] Kwic/360. Keyword in context indexing program for the IBM System/360
P.L. WHITE
Février 1968

[WIL 63] A discriminant method for automatically classifying documents
J.M. WILLIAMS
AFIPS Proceedings - 24
1963

[WIL 77] On-line retrieval - Today and tomorrow
M.E. WILLIAMS
University of Illinois - USA
1977

dernière page de la thèse

AUTORISATION DE SOUTENANCE

VU les dispositions de l'article 3 de l'arrêté du 16 Avril 1974,

VU le rapport de présentation de Monsieur :

- J.Cl. CHUPIN, Ingénieur, Responsable du Centre de Recherche CII-HB - GRENOBLE -

Monsieur Jean H A M E O N

est autorisé à présenter une thèse en soutenance pour l'obtention du titre de DOCTEUR de TROISIEME CYCLE, spécialité "Génie Informatique".

Grenoble, le 16 Janvier 1981

Le Président de l'I.N.P.G.

Ph. TRAYNARD
Président
de l'Institut National Polytechnique