UNIVERSITE CLAUDE BERNARD LYON-I
43, Boulevard du 11 Novembre 1918
69621 VILLEURBANNE

Diplôme d'Etudes Supérieures Specialisées

informatique documentaire

* MEMOIRE DE STAGE
* ~~NOTE DE SYNTHESE~~

Indexation automatique

AUTEUR : DINDOYAL (Devi-Anand)

DATE : juin 1979

Je remercie et assure de ma gratitude M. le Professeur Bouché ainsi que M. Saroul et toute l'équipe du service Informatique de DATA PRESSE qui m'ont apporté une aide précieuse pendant mon stage.

D. A. D

- PLAN -

1- INTRODUCTION.

1.1 Historique

1.2 Le cadre du stage

2- INDEXATION AUTOMATIQUE ...

2.1 Nécessité d'une indexation

2.2 Les objectifs et la méthode utilisée

2.3 Matière première

2.4 Materiels utilisés

3- LES ETAPES DE L'INDEXATION.

3.1 Constitution d'un dictionnaire

3.2 Exploitation

4- CONCLUSION

4.1 Points faibles d'une telle indexation

4.2 Améliorations possibles.

1 - Introduction

1.1 Historique

La vie du Progrès de Lyon est étroitement liée à celle du Dauphiné libéré de Grenoble. En effet en 1966 ces deux sociétés ont signé des accords pour mettre en commun leurs moyens de production tout en restant concurrentes et économiquement indépendantes. Le Progrès est aussi propriétaire des titres de la Tribune (St Etienne), L'Espoir (St Etienne), Le Progrès Soir (Lyon). Le Dauphiné Libéré est propriétaire des titres du Dauphiné (Grenoble), de la Dernière Heure Lyonnaise (Lyon), La Dépêche (St Etienne). Au début il était aussi propriétaire L'Echo liberté qui, depuis, a été remplacé par le Journal Rhône. Alper qui, lui, n'appartient pas au Dauphiné libéré. Ces deux groupes existent toujours mais depuis leurs accords de 1966 ils ont crée des sociétés de service appartenant aux 2 groupes. ces sociétés

de service sont :

- Entreprise Presse no 1 (EP1) qui se charge de la fabrication des journaux (possède donc les rotatives etc...),
- Rhône - Alpes Diffusion (R.A.D.) qui s'occupe de la diffusion (abonnés et dépositaires) et de la comptabilité et facturation de ces journaux,
- Société d'Edition Régionale de Périodiques (S.E.R.P.) qui a été créée pour l'édition des journaux du dimanche ; en effet la loi interdit à un editeur d'éditer 7 jours sur 7,
- Province Publicité no 1 (PP1) qui depuis l'accord de 1972 avec la société Havas est devenue Province Publicité Havas (PPH). Elle s'occupe des petites annonces, de la publicité et de leur facturation.
- Agence d'Informations Générales Locales Economiques et Sportives (A.I.G.L.E.S.), créée sous forme de coopérative ouvrière, elle regroupe les journalistes et correspondants. Les rédacteurs en chef et les secrétaires généraux, eux, ne

dépendent pas de L'A.G.G.L.E.S. Chaque journaliste cependant ne travaille que pour son journal. Les agences régionales (sauf Paris) en revanche travaillent pour les 2 groupes.

- Data Presse, un G.I.E. (groupement d'Intérêts Economiques) qui a été crée en 1973 est chargé de tous les traitements informatiques de ces sociétés. A l'origine, ces traitements informatiques étaient réalisés par une société de service, EURINFOR.

1.2 Le cadre du stage.

Effectuant mon stage à Data Presse, il me semble utile de compléter sa présentation. En 1971 les services d'Eurinfor devenant trop chers, la SOMIPRO (Société d'Etude et de Maintenance de Problèmes de Presse sur Ordinateur) fut chargée de livrer un service informatique clé en main. A l'origine ce service s'occupait surtout de la gestion. Il était mis en place

sur 2 ordinateurs SIEMENS 4004/26 de 16 K mémoire avec chacun 5 unités de disques de 7 millions d'octets et 3 imprimantes connectables sur l'un ou l'autre des ordinateurs. La photocomposition se faisait alors à part.

Puis, en 1975, lors de changements de matériel de fabrication, la CII proposa un soft s'appellant COSY 1000 qui avait pour base un soft de composition de SIEMENS Allemagne, COSY 35, plus un système d'entrée en télétraitement développé par la CII pour Nice Matin. De plus ce soft permet la sortie sur photocomposeuse on-line. La CII en a fait un soft standard pour la presse. En 1976 on commença à travailler en réel. A l'époque ce système était utilisé uniquement pour faire des petites annonces sur un 4004/25 de 256 K équipé de 3 disques de 50 millions d'octets. Un unidata 7730 servait de machine de secours. Il était equipé de 4 disques de 50 millions. Puis le système fut utilisé pour la publicité et à la fin 1976 pour la fabrication

de textes rédactionnels. En 1978 toutes les agences étaient connectées directement sur ordinateur. ces agences possèdent à l'heure actuelle des terminaux MDS qui permettent la saisie on-line et off-line sur disquettes.

Aujourd'hui Data Presse Chassieu est équipé de:

- 3 ordinateurs siemens 7740 dont 2 de 640 K octets et 1 de 448 K;
- 14 unités de disques (50 millions octets)
- 8 dérouleurs de bandes (60 K octet - 1600 BPI)
- 3 imprimantes (1100 lignes/minute)
- 2 photocomposeuses (DIGISET)
- 2 lecteurs de cartes (600 cartes/minute)

Les ordinateurs assurant la production sont connectés à ceux de Veurey (autre centre de production) par des liaisons spécialisées (4800 bits/sec) Sur Chassieu sont connectés 6 autres agences:

- Bourg-en-Bresse	12 terminaux
- Macon	12 "
- Lyon (Rue de la Charité)	12 "
- Paris	3 "
- Avignon	

- Lens - le - Saulniers 8 terminaux
- Avignon 12 "

Jusqu'ici, à Data Presse, l'emphase était mise sur la photocomposition. Maintenant que les problèmes sont résolus dans ce domaine, on se tourne vers la gestion des informations possedées. Les problèmes de la photocomposition étaient celui de l'adaptation d'un soft à un matériel donné - et avec la gestion des inform. ations vont naître les problèmes de l'indexa. tion automatique.

2- Indexation Automatique...

2-1 Nécessité d'une indexation.

Comme nous le savons tous, dans le domaine de la presse, diffuser simplement les informations ne suffit pas, encore faut-il pouvoir les donner en temps voulu c-à-d en même temps, voire avant, ses concurrents. Il arrive quelquefois que l'on ait besoin d'une information antérieure pour rendre un article plus complet - afin de situer le lecteur dans le contexte en lui apportant des rappels. Par information antérieure, il faut comprendre information qui (le plus souvent) a paru dans les journaux. Ceci suppose donc que l'on sache où la trouver. Et c'est là que le bât blesse. Prenons un exemple concret : Un dimanche soir le Progrès reçut la nouvelle selon laquelle le corps d'Alain Colas aurait été retrouvé. On décida alors de publier une sorte de biographie d'Alain Colas. En effet les informations

concernant ce personnage très connu pour ses participations à la course de la transatlantique ne manquaient pas. Malheureusement l'accès à ces informations ne put se faire rapidement et l'on ne trouva aucune bonne photo d'Alain Colas non plus. Pourtant on savait que ces informations existaient mais on ne savait pas où les trouver. En effet au Progrès il n'existe qu'une archive manuelle - donc incompatible avec une recherche rapide, vu le volume d'informations à traiter. ce genre d'incident ne doit pas être très rare, je suppose. D'où ce besoin d'une exploitation automatique dans la recherche des informations et de là à l'indexation automatique il n'y a qu'un pas.

Pour resumé, je dirai que le besoin d'automatiser les recherches se fait sentir principalement dans les deux cas suivants:

- Recherche d'information pour faire des suivis d'articles (regrouper tout ce qui a été antérieurement dit sur un sujet donné);
- Recherche d'informations concernant

une personalité qui a eu un tournant dans sa carrière ou vie (voire décès).

2.2 Les objectifs et la méthode utilisée.

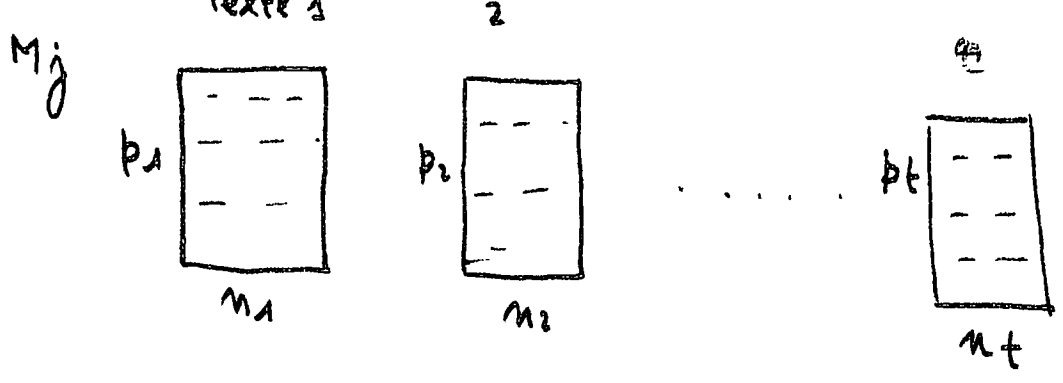
L'objectif principal est d'aboutir à une indexation automatique, c'est-à-dire trouver automatiquement un ensemble d'extraits des mots qui seront susceptibles de caractériser (refléter) un article donné. Il est évident que si l'on veut en arriver là, il faut nécessairement passer par la constitution d'un dictionnaire qui recensera le vocabulaire utilisé par les journalistes. Ce dictionnaire ne va cependant pas contenir tous les mots utilisés dans les articles. En effet, il existe un certain nombre de mots qui sont vides de signification (dans le cas une indexation bien sûr) et qui apparaissent très souvent. Ces mots outils, comme on les appelle, permettent d'articuler les phrases et constituent en général plus de 50 % des mots d'un texte. Dans notre cas nous les avons

limités à 330 et regroupés dans un glossaire avec leur fréquence d'apparition. Ce glossaire comprend entre autre, les différentes formes des verbes avoir et être, les conjonctions de coordinations, des prépositions etc... Nous éditerons à part ce glossaire!

Une fois le dictionnaire constitué nous allons forcément nous en servir pour aboutir à notre indexation automatique Il nous faut donc une méthode qui nous permettra de l'exploiter. Nous avons décidé d'adopter une méthode statistique simple. Pourquoi une méthode statistique? D'abord parce que nous savons qu'un mot est représentatif, à un dégré plus ou moins grand, du texte dans lequel il se trouve et ensuite qu'il existe un lien entre la représentation quantitative de ce mot et la qualité de l'information dont il est le support. Ceci nous amène donc à faire un comptage systématique de tous les mots. Il s'agit alors d'associer à chaque mot (du dictionnaire) un seuil à partir duquel nous pouvons estimer qu'un mot devient significatif. Nous allons maintenant

voir comment calculer ce seuil et comment l'utiliser :

Soient t textes comprenant chacun $n_1, n_2, n_3, \dots n_t$ mots. Considérons le mot M_j qui apparaît $p_1, p_2, \dots, p_t$ fois dans les t textes, les p_i pouvant être égaux à zéro dans le cas où le mot est absent dans le $i^{ème}$ texte.



Dans un premier temps nous calculons pour chaque texte la fréquence relative f_i du mot M_j

Ex. $f_1 = p_1/n_1 \dots f_i = p_i/n_i \dots f_t = p_t/n_t$

puis nous calculons la fréquence relative du mot M_j par rapport à l'ensemble des textes

$$F(M_j) = \sum p_i / \sum n_i \quad i \in [1, t]$$

pour chaque texte on aura un seuil S_i pour le mot M_j donné par $S_i = f_i / F(M_j)$. c'est la

moyenne de ces seuil si que nous allons associer à chaque mot dans le dictionnaire.

Pour l'indexation nous prenons un texte quelconque et nous l'épurons des mots outils. Pour chacun de ces mots (mots outils exclus) nous calculerons sa fréquence relative dans le texte. Soit f cette fréquence relative. ~~Nous allons~~ ensuite rapporter ce f à la fréquence relative $F(M)$ que nous avons calculée lors de la constitution du dictionnaire (Si le mot n'existe pas dans le dictionnaire, nous l'abandonnons). Nous obtenons ainsi un seuil $S = f / F(M)$ pour un mot donné. Nous comparerons ce seuil au seuil moyen du dictionnaire, pour ce même mot et s'il est égal ou supérieur au seuil moyen du dictionnaire nous considérons le mot comme mot-clé. Pour vérification nous comparerons l'ensemble des mots-clés obtenus par cette méthode à celui que l'on obtiendrait par une indexation manuelle.

2.3 Matière première

La matière première consiste en un certain nombre de textes de journaux (plus particulièrement les informations générales), qui ont été saisis sur bande pour la photocomposition. Ces textes sont d'abord épurés des caractères spéciaux qui servent à la photocomposition. Par bande, il y a environ 50 à 100 textes et nous pouvons compter une moyenne de 100 à 200 mots par texte. Nous travaillons sur une dizaine de bandes. A ce niveau il faut remarquer que certains problèmes se posent du fait que nous travaillons sur des textes qui ont été saisis dans un autre but que l'indexation. Par exemple il a fallu ramener toutes les lettres à leur equivalent majuscule quand ce dernier existe dans le jeu de caractère sinon nous les laissons en minuscule pour ne pas perdre l'information. Cette transcription nous permet de ne pas comptabiliser plusieurs fois les mêmes mots sous leurs formes différentes (ex LES, Les, les sont tous ramené à LES). Et lors de la séparation des mots nous étions

obligés de ne pas considérer le point comme délimiteur afin de ne pas éparpiller les sigles aux quatre coins du dictionnaire. ceci nous a amené ensuite à éliminer le point qui se trouve à la fin d'un mot de fin de phrase - donc traitement superflu!

2.4 Materiels utilisés.

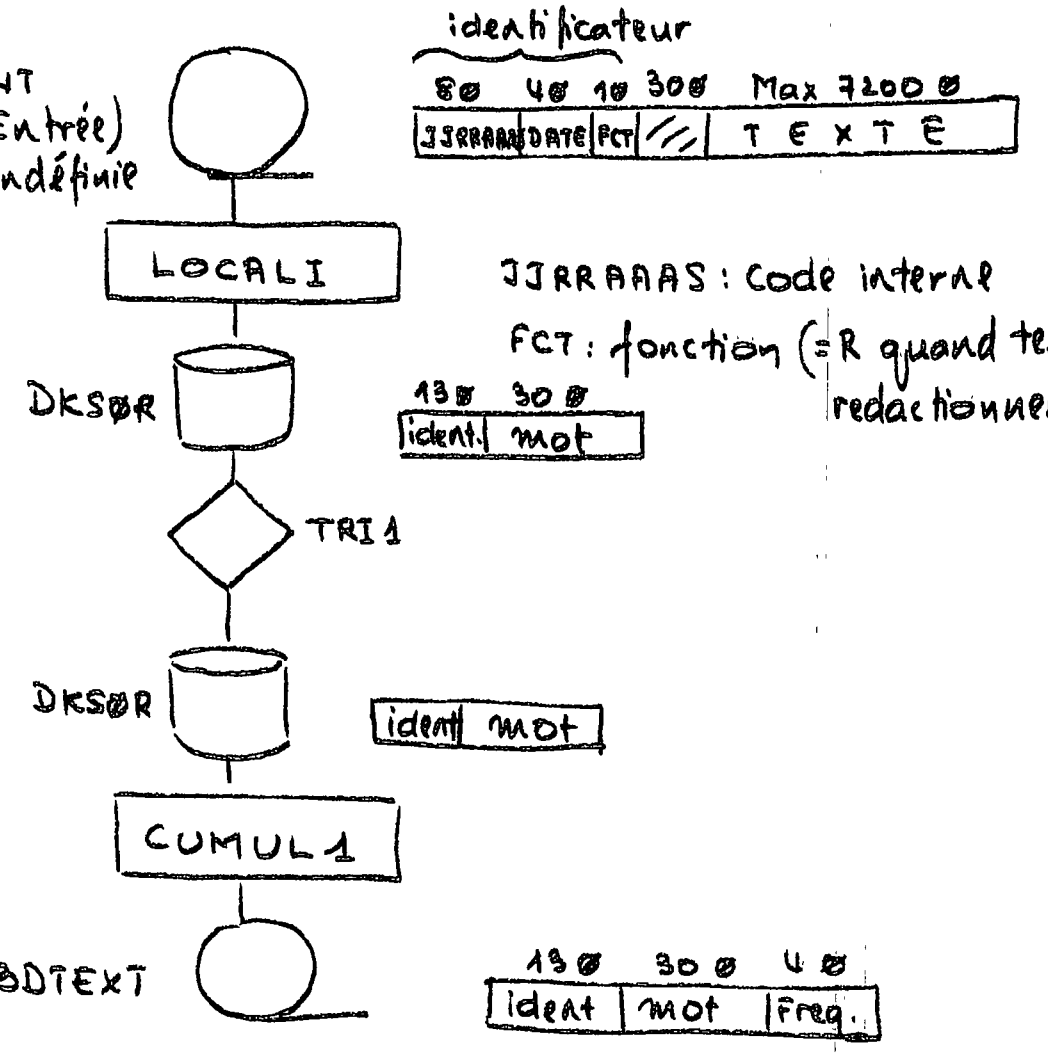
- ordinateur 'siemens' 7740 - taille mémoire 640 K octets.
- unités de disques de 50 millions octets chacune.
- Dévouleurs de bandes - 60 K octets et 1600 BPI
- Imprimante 1100 lignes/minutes
- lecteur de cartes - 600 cartes/minute

(Les langages de programmation utilisés : l'assembleur et le Cobol)

3 - LES ETAPES DE L'INDEXATION

3.1 Constitution d'un dictionnaire

a)



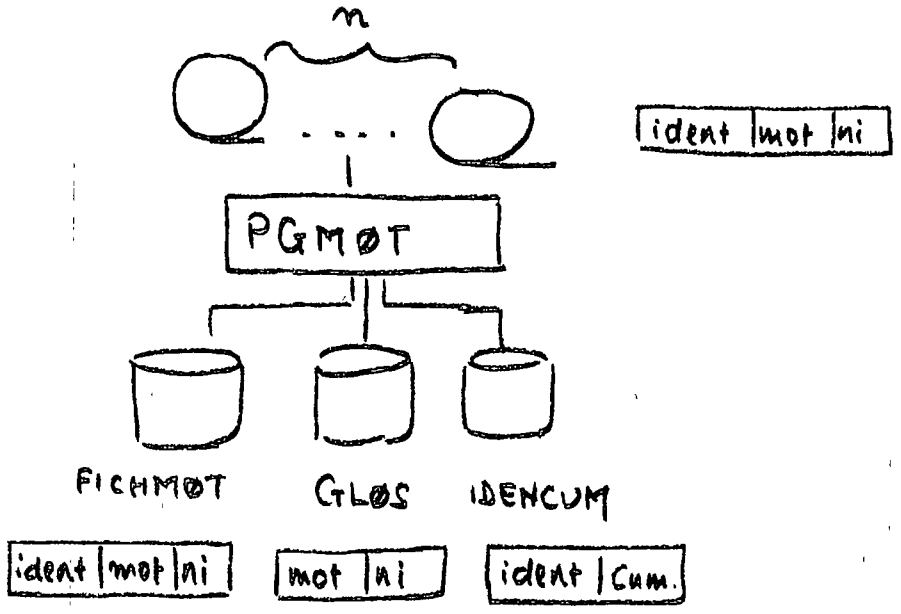
La bande en entrée est déjà épurée des caractères spéciaux qui servent à la photocomposition. Le programme LOCALI a pour fonction principale de localiser les mots en utilisant comme délimiteurs les caractères suivants: ␣ < (+ | ! *) ; ¬ , - > ? / : # ' = " ^

␣ est l'espace normal
- est le tiret (à ne pas confondre avec le trait d'union qui a une configuration hexadécimale différente et est considéré comme délimiteur)

Le point n'est pas considéré comme un délimiteur pour les raisons mentionnées précédemment. Un mot est un ensemble de caractères (délimiteurs exclus) qui se trouve entre 2 délimiteurs.

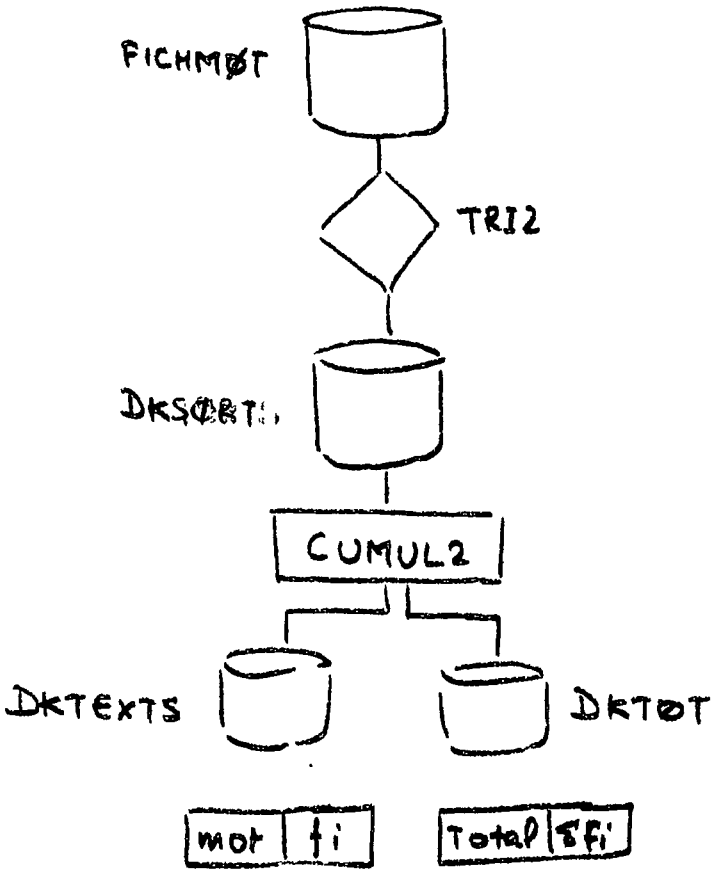
Après un tri sur l'identificateur et sur le mot, nous utilisons le programme CUMUL1 pour faire le cumule des mots qui apparaissent plusieurs fois à l'intérieur d'un même identificateur (un ident. représente un texte). La bande obtenue à la sortie BDTEXT (en réalité on écrase la bande en entrée) sera conservée car c'est elle qui nous permettra, lors d'une recherche, d'associer un mot-clé à un texte. Pour la constitution du dictionnaire nous en aurons besoin d'une dizaine de bandes environ.

b)



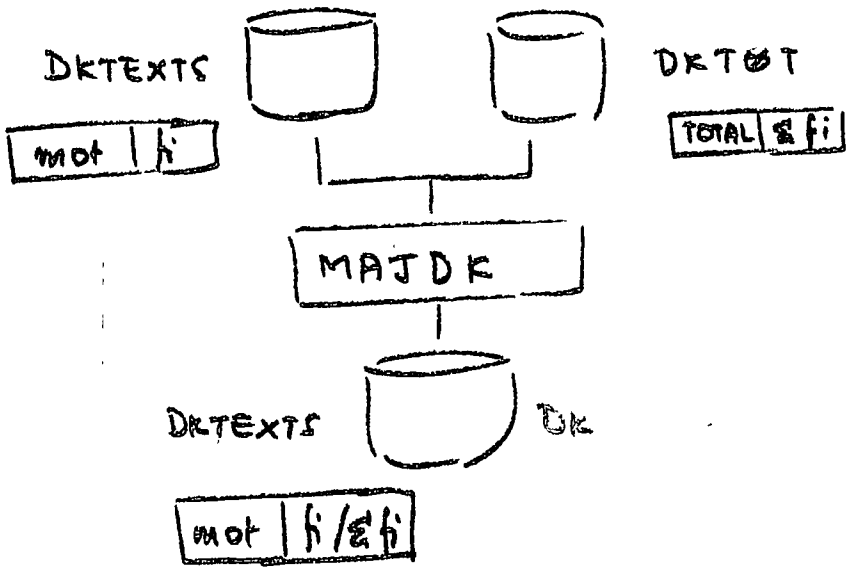
Ce programme nous permet, à partir des bandes obtenues à la fin de la chaine a), de constituer un fichier (FICHMØT) des mots utiles avec leur identificateur et leur frequence. Les mots outils iront directement dans le glossaire (GLØS) avec leur nombre d'apparitions sur l'ensemble des textes (éventuellement on éditera sur imprimante le glossaire). Lors de ce programme nous calculons aussi le nombre de mots par textes (IDENCUM).

c)



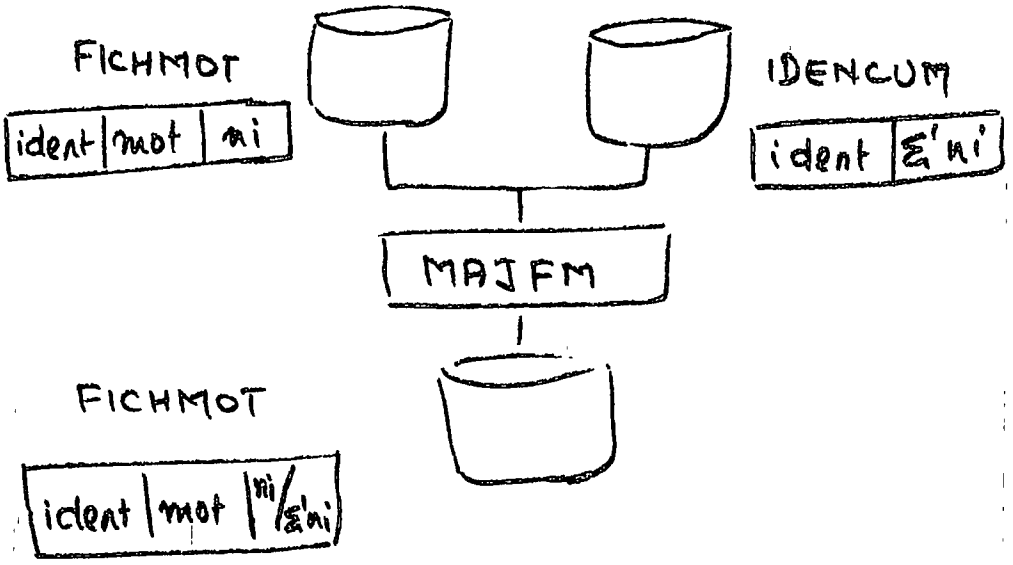
Dans cette chaine nous effectuons d'abord un tri sur les mots et sur les fréquences et ensuite nous regrouperons les mots de façon à avoir leur nombre d'apparitions (fi) sur l'ensemble des textes. ce Cumule est fait par le programme CUMUL2 qui fait aussi le total de tous les mots sur lesquels nous travaillons. Nous mettrons ce total dans un fichier (DKTØT) qui ne contient qu'un seul enregistrement. ceci pour plus de commodités. Le fichier DKTEXTS qui contient les mots et leur fréquence sera conservé. Nous en aurons besoin pour l'indexation.

d)



Lors de ce programme nous allons calculer pour chaque mot la fréquence relative sur l'ensemble de textes, c'est-à-dire, les $f_i / \Sigma f_i$. Nous avons déjà les f_i dans DKTEXTS, nous allons donc mettre à jour ce dernier en utilisant le fichier DKTOT qui, lui, contient la Σf_i.

e)

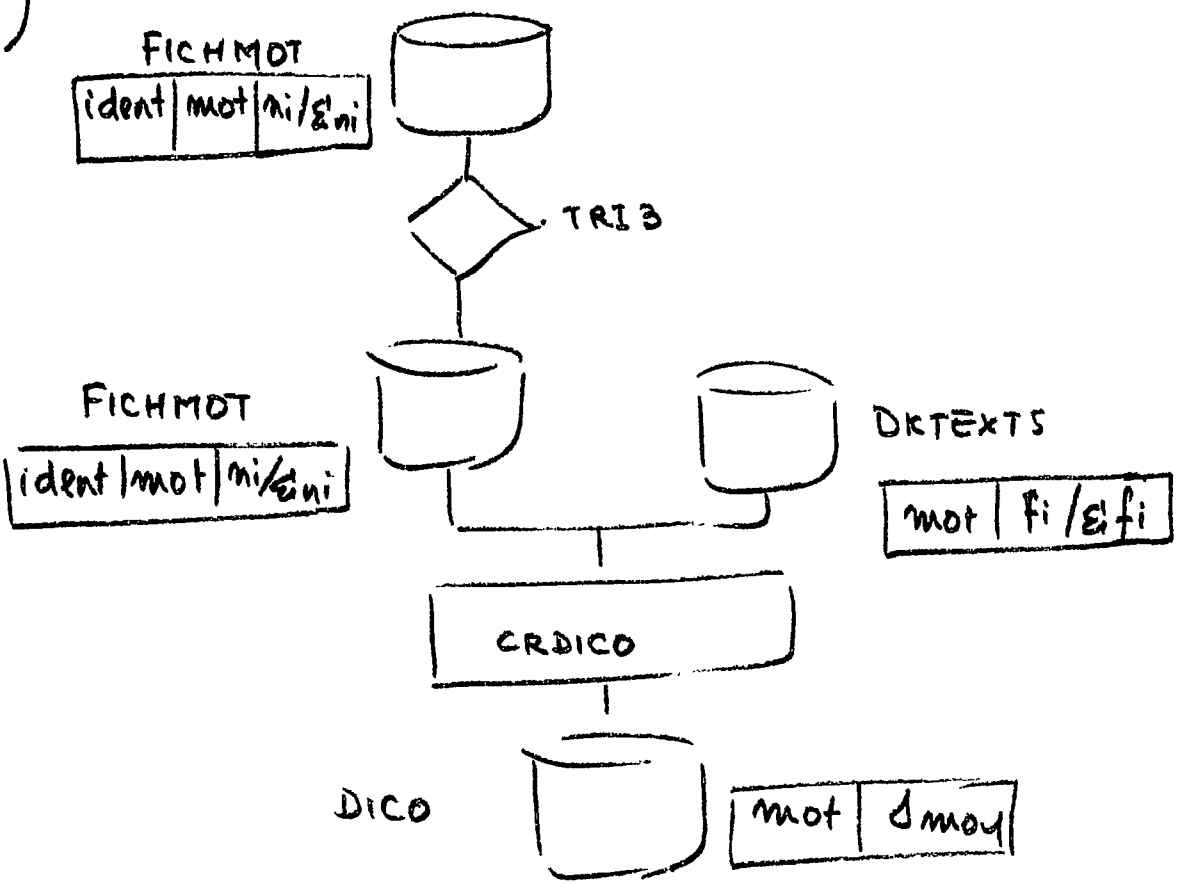


Ce programme va nous permettre, à partir de FICHMOT et de IDENCUM d'obtenir en mettant à jour FICHMOT les $ni/\Sigma' ni$ pour chaque mot. C'est, en effet, la fréquence relative de chaque mot dans un texte que l'on cherche à calculer ici.

Arrivés à ce stade, nous avons toutes les information nécessaires pour calculer les seuils si qui, rappellons-le, seront donnés par la formule suivante:

$$si = ni/\Sigma' ni / fi/\Sigma' fi$$

d)

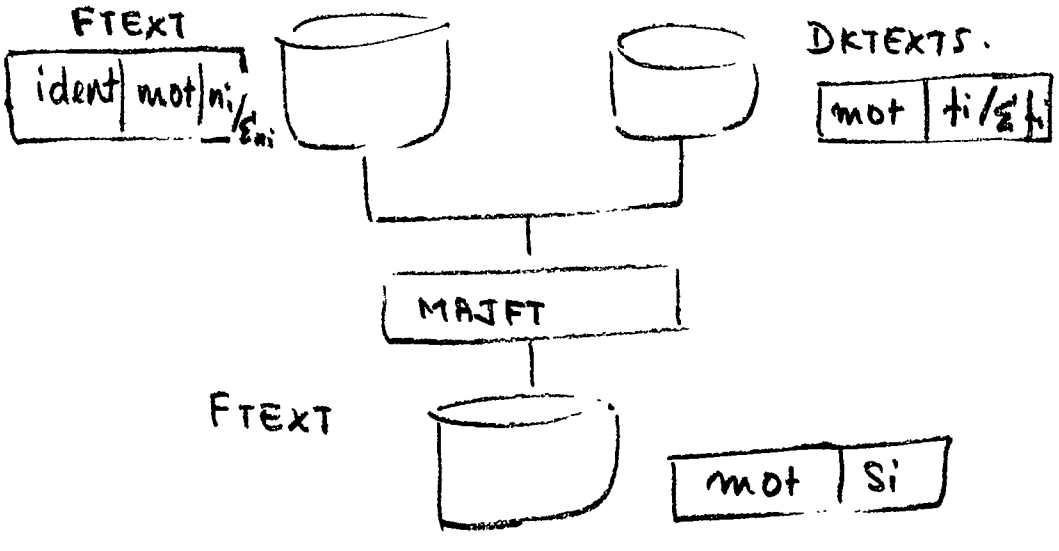


Dans un premier temps nous allons trier FICHMOT de façon à avoir tous les mots dans un ordre alphabétique (on ne tient plus compte des identificateur). Puis en utilisant le programme CRDICO nous allons créer le dictionnaire avec le seuil moyen par mot en rapprochant les deux fichiers: FICHMOT et DKTEXTS. Ce seuil moyen sera la somme des di pour un mot donné divisée par le nombre de textes dans lesquels ce mot se trouve.

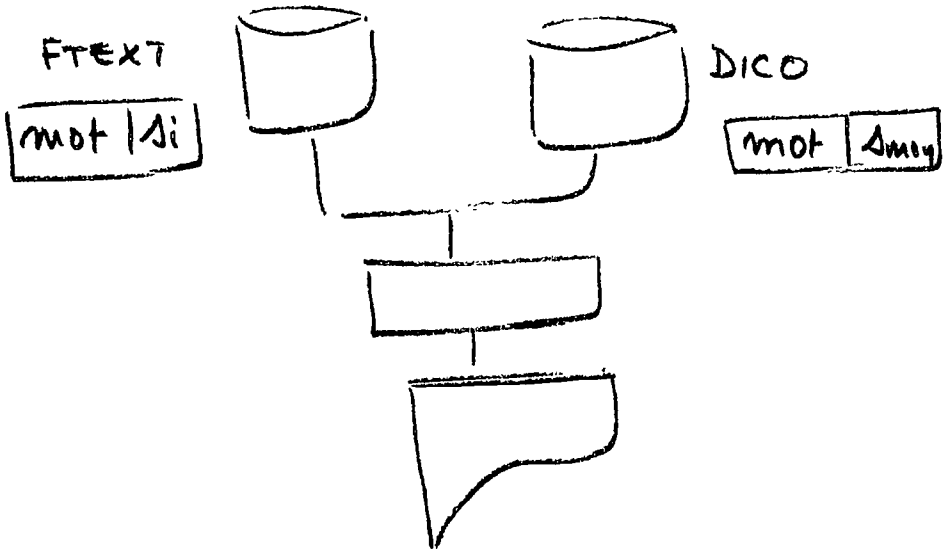
3.2 Exploitation.

Nous supposons bien sûr que notre dictionnaire est valable !!!

Nous prenons un texte quelconque et nous le passons par les chaines a, b et e de la section 3.1. Nous obtiendrons ainsi un fichier trié alphabétiquement avec, pour chaque mot, sa fréquence relative $n_i / \Sigma n_i$. En utilisant le fichier DKTEXTS que nous avions pris soin de conserver, nous pourrons mettre à jour ce fichier en remplaçant les $n_i / \Sigma n_i$ par les seuils S_i.



Puis par ce deuxième programme,



nous allons comparer les Si de chaque mot au seuil moyen du même mot dans le dictionnaire quand le mot n'y est pas absent (sinon on le signalera). On sortira ainsi une liste de tous les mots qui ont leur seuil, calculé pour le texte, supérieur ou égal au seuil moyen du dictionnaire. Ces mots seront considérés comme les mots-clés.

4 - Conclusion

Le problème le plus préoccupant actuellement dans le domaine de la documentation est d'arriver à une indexation automatique qui soit le plus fiable possible. D'ores et déjà la compétition est ouverte à un certain nombre de chercheurs, plus spécialistes cependant en informatique qu'en documentation. Or, pour faire une bonne indexation, il ne suffit pas de savoir quelle démarche scientifique ou informatique suivre, il faut encore savoir ce que l'on veut, i.e à quoi correspondent exactement les besoins des utilisateurs. Dans le cadre de mon stage, ce problème est d'autant plus accru que je travaille sur des informations générales et non spécialisées. Un prétraitement (nous verrons plus loin à quoi il correspondrait) devient inéluctable. Mais ce prétraitement cependant requiert, outre les qualités d'un informaticien, celles d'un linguiste et d'un documentaliste. Car s'il est évident pour un bon documentaliste d'appréhender manuellement une bonne

indexation (il peut juger l'importance d'un mot par rapport au contexte dans lequel il se trouve) il n'est pas du tout évident de le faire automatiquement. Donc compter les mots simplement et déduire par des calculs statistiques à partir de quel seuil il devient significatif, ne suffit vraiment pas.

Tout ce qui a été dit plus haut nous amène à nous demander quelle est donc la validité de notre indexation. Nous allons donc voir, dans un premier temps, les points faibles d'une telle indexation et, dans un second temps, les améliorations que l'on pourrait y apporter.

4.1 Les points faibles

Nous pourrions commencer par critiquer la méthode statistique utilisée. Il est évident que cette dernière est très simple et que la précision d'une telle indexation est loin de parfaite. Mais une autre méthode plus élaborée aurait demandé, d'une part, plus de temps à programmer et, d'autre part, il nous aurait

fallu nous lancer dans des études statistiques plus poussées pour faire le choix d'une bonne méthode. Le temps nous manquait. Quatre mois de stage ne suffisaient pas.

D'autres problèmes surgissent lors de la séparation des mots. S'il est vrai que pendant cette étape on conserve les informations intactes (les accents sont présents même en majuscules, grâce à une typographie riche et les sigles ne sont pas éparpillés aux quatre coins du dictionnaire étant donné que le point n'est pas considéré comme délimiteur) il est cependant regrettable de constater que les syntagmes etc.. eux, sont bel et bien démantelés. Les noms propres sont aussi separés. i.e. nom et prénom ne se trouvent pas ensembles. Ces noms propres ne devraient pas apparaître dans le dictionnaire mais on est obligé de les y insérer car il est très difficile de les détecter; en effet on ne sait pas faire la différence entre le point dans M. Dupont, par exemple, et un point de fin de phrase.

Les Les textes sur lesquels nous travaillons ne faisant l'objet d'aucune analyse morphologique, un certain nombre d'ambiguïtés nous échappent, telles que la polysémie par exemple. En effet on fait pas de différence entre le mot 'son' en tant qu'adjectif possessif et le mot 'Son' en tant que substantif. Le problème est encore plus grave, car 'son' en tant qu'adjectif possessif est un mot que l'on élimine pour mettre dans le glossaire. D'autre part, on comptabilise plusieurs fois les mêmes items parce que ce sont des verbes conjugués différemment ou encore des mots avec un genre ou un nombre différents (problème de désinance)

Un dernier problème (il en existent certainement d'autres mais on ne peut pas tous les recenser) est le problème de concept. Comme on nous apprend à ne pas repéter un même mot trop souvent quand nous rédigeons un article, il se trouve qu'un même concept soit représenté sous différentes formes (Ex: Etude, Projet, Recherche ont la même signification)

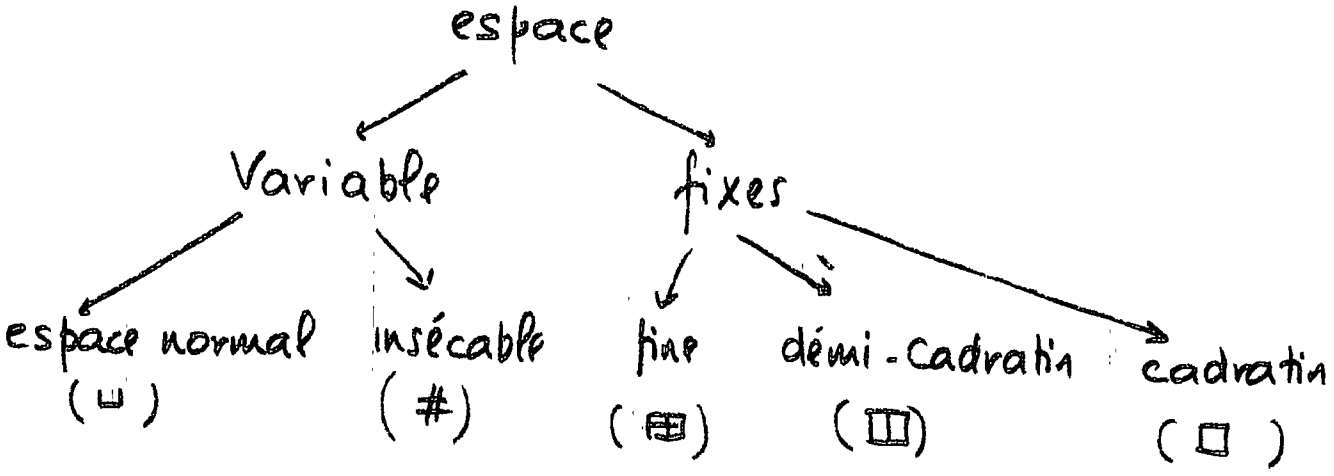
Par conséquent lorsque l'on compte les mots, on perd des informations car il est évident que l'impact d'un terme significatif est alors diminué et qu'on risque même de le perdre.

4.2 Améliorations possibles

En ce qui concerne la méthode utilisée, il est bien évident que si l'on peut en trouver une qui soit plus élaborée, l'idée de base reste la même en revanche. c'est-à-dire qu'il faut passer par le comptage des mots (critère de fréquence). La seule amélioration possible est de trouver une méthode qui donnerait de meilleurs resultats. Cela demanderait du temps pour appréhender les diverses théories statistiques applicables pour faire un choix judicieux.

Pour éviter que des mots qui vont de pair ne soient éparpillés, on peut le faire par une démarche statistiques (calcul des co-occurence etc.) Mais il serait beaucoup plus utile de le faire lors de la saisie. Je m'explique - le jeu de caractères utilisé pour la photocomposition est

très riche; on pourrait donc l'utiliser à bon escient pour l'indexation. En effet il existe différentes formes d'espaces dans ce jeu (configuration hexadécimale différentes) comme nous montre l'arbre ci-dessous:



L'insécable indique à la machine qu'on ne peut pas couper un ensemble de mots lors de la composition pour changer de ligne - De ce fait, la machine va redistribuer les blancs qui existent en tassant les mots ou en les étendant. (Ex on ne pourrait pas separer 15#H#30 à la fin d'une ligne pour avoir 15 à la fin de la ligne et H␣30 sur la ligne suivante; par contre on pourrait avoir 15␣␣H␣␣30 à la sortie.) Les espaces fixes, eux, permettent de ne pas separer un ensemble de mots sans pour cela avoir à redistribuer les blancs à l'intérieur de ce groupe. c'est-à-dire que la distance est fixée entre ces mots une fois pour toute (cette distance est fonction de l'espace

fixe qu'on utilise). Nous voyons donc les avantages que l'on peut en tirer en faisant de ces espaces (autre que l'espace normal) des non-délimiteurs.

D'autre part une analyse morphologique sur les textes nous permettrait d'éliminer beaucoup d'ambiguités. On pourrait par exemple éliminer les désinances pour ramener à une forme unique certains mots (élimination du pluriel, féminin etc...) On pourrait aussi ramener les verbes conjugués à leur forme infinitive. Les avantages d'une telle analyse sont nombreux. Mais encore faudrait-il se lancer dans des études plus approfondies qui demanderaient plus de temps.

En ce qui concerne les problèmes des concepts, le mieux à faire est de se constituer un dictionnaire qui permettra de faire le rapprochement entre les différents concepts!!!

Finalement je pense que de nombreuse améliorations pourraient être faites si, lors de la Saisie, on prenait en compte certains problèmes. Ceci évidemment serait très lourd et difficilement envisageable sur des textes prévus pour la photo-composition. Mais cela n'empêche que l'on

pourrait lever certaines ambiguïtés telles que par exemple certaines polysémies en affectant un signe particulier, si besoin est, aux mots qui vont dans le glossaire.