مركز للاعلام العلمي والتقني والتحويلات التقنولوجية
CENTRE D'INFORMATION SCIENTIFIQUE ET TECHNIQUE
ET DE TRANSFERTS TECHNOLOGIQUES

ETUDE DE LOGICIELS POUR RECHERCHE DOCUMENTAIRE

Atika LARIBI
Dalila MESSAOUDI

B.T. Doc 3./5.76

SOMMAIRE

I N T R O D U C T I O N

Le stockage d'informations auquel nous assistons depuis plusieurs années devrait aboutir à une consultation. C'est à ce stade que les difficultés apparaissent. Comment faire pour que : étant donné un ensemble sans cesse accru de documents disponibles, la masse totale des documents étudiés par chacun, (dans le temps dont il dispose pour cette activité), corresponde le mieux à ses besoins particuliers. Etymologiquement informatique vient du mot information, elle se doit donc de venir au secours de celle ci: Elle le fait avec ce qu'on appelle improprement la "documentation automatique".

La documentation automatique consiste d'abord à rassembler et à ranger des informations dans les mémoires auxiliaires d'un ordinateur pour en établir un fonds documentaire, ce système permet alors la recherche de ces informations dans le fonds à partir d'une requête formulée par l'utilisateur.

Ces requêtes peuvent être de 2 types :

1 - Périodiques :
L'utilisateur se définit un "profil d'intérêt qui décrit ses besoins. Il lui est alors permis de recevoir régulièrement tous les documents parus susceptibles de couvrir son domaine de recherche. C'est la diffusion sélective.

2 - Ponctuelles :
Relatives à un domaine d'intérêt précis et qui s'adressent au fonds documentaire tout entier. C'est la recherche rétrospective.

Dans les 2 cas la réponse est une liste de documents classés suivant le désir de l'utilisteur et où n'apparaissent que les éléments répondant à son étude.

La chaîne de documentation automatique (ou chaîne documentaire) se divise en plusieurs étapes :

1 - Collecte des documents.

2 - L'analyse de contenu :

Ce n'est pas le document lui-même qui est introduit en mémoire mais son signalement, c'est un ensemble de rubriques qui peuvent être :

a) Un renseignement sans modifications : titre, auteur, éditeur, référence bibliographique etc...

b) Une représentation du contenu: créée par un spécialiste que l'on appelera indexeur. Après analyse du document l'indexeur extrait un certain nombre de mots, dits mots clés, de façon à ce que leur juxtaposition rende compte du sujet traité dans le document. Il peut aussi faire un résumé du document (abstract).

3 - L'exploitation :

C'est la phase automatisée. Elle comprend la mise en mémoire du signalement des documents, le stockage sur des mémoires auxiliaires du fonds documentaire, et enfin la recherche à partir de questions formulées par l'utilisateur. Les questions sont formulées à l'aide de mots clés, selon une logique booléènne_

Exemple : Exploitation du pétrole et du gaz en Algérie.

Algérie ET exploitation ET (pétrole) OU (GAZ)

Les documents sont alors sélectionnés s'ils vérifient l'équation c'est à dire dans l'exemple précédent les documents possédant les mots clés Algérie, exploitation et pétrole seront sélectionnés, de même ceux contenant les mots clés Algérie, exploitation et gaz.

En recherche documentaire il existe 2 grands types de système de recherche:

SYSTEMES AVEC FICHIER INVERSE :

Pour un stock documentaire donné : à chaque mot clé est associé, dans un fichier, dit fichier inverse, une liste de documents qui ont utilisé ce mot.
Lorsque l'on pose donc une question, qui est une succession de mots clés reliés par des opérateurs logiques, on interrogera ce fichier pour chaque mot clé de la question.
Ce fichier doit être à accès rapide donc sera en accès direct.

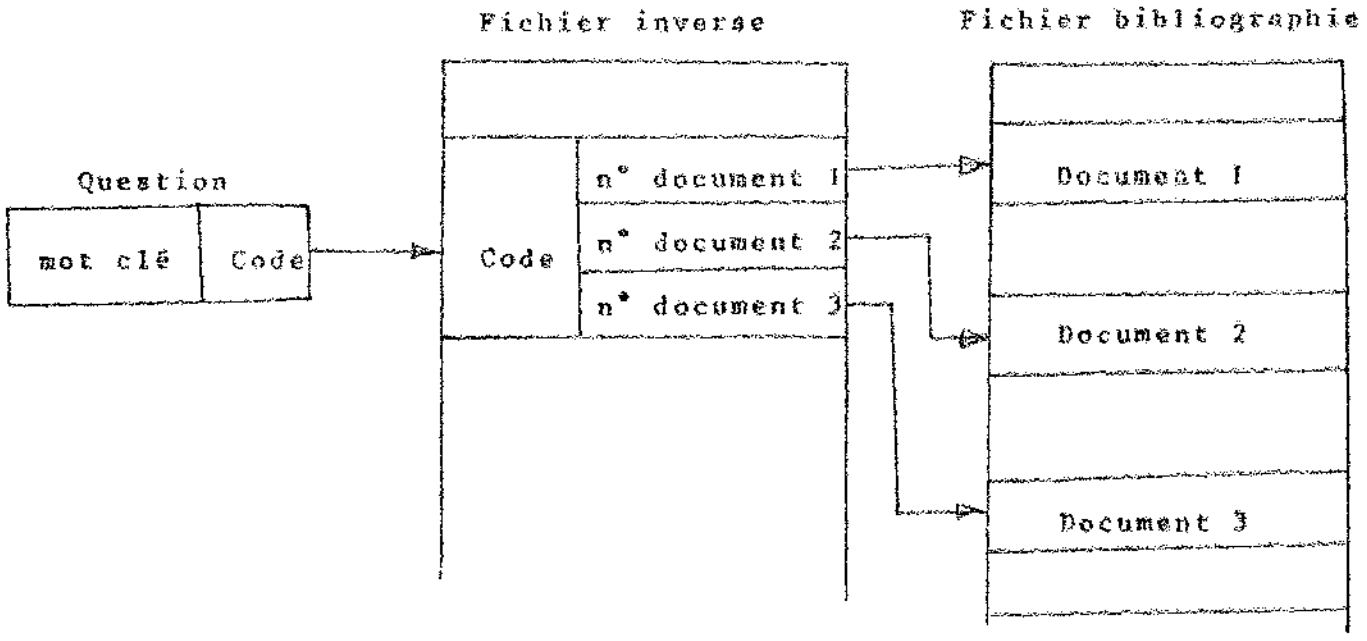

Figure 1

SYSTEMES A ANALYSE SEQUENTIELLE :

Lors d'une recherche on analyse séquentiellement toute la base de données, document par document, et l'on vérifie donc pour chaque document l'équation logique de la question.

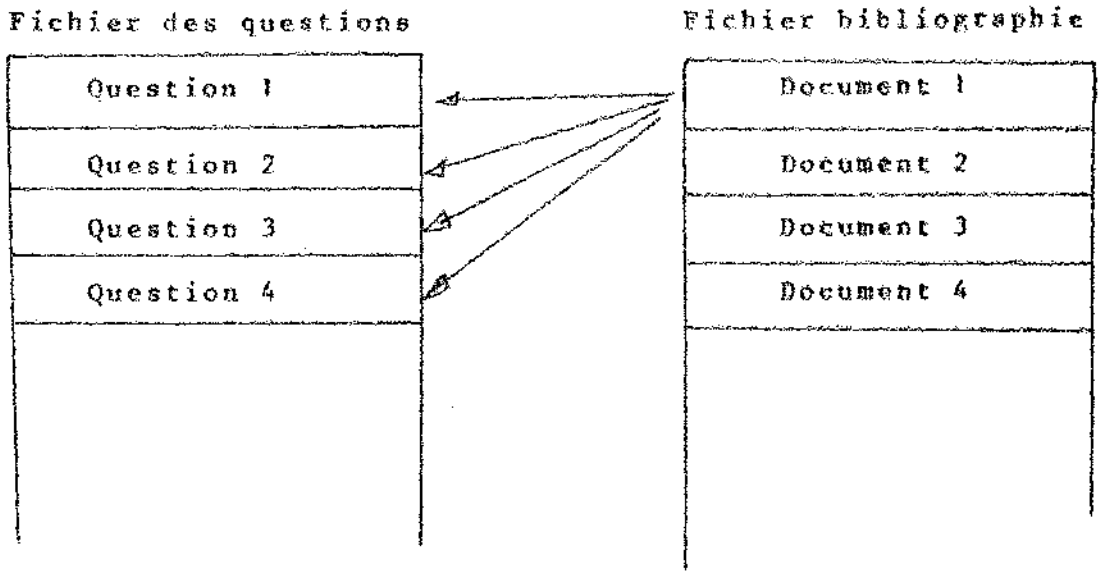

Figure 2

Les systèmes du premier type sont rentables pour les recherches rétrospectives, lorsque l'on ne pose qu'une seule question à la fois. La méthode du fichier inverse étant une méthode trés rapide le temps de réponse et minimisé.
Mai c'est une méthode généralement plus coûteuse en place mémoire que la méthode d'analyse séquentielle, qui peut être elle plus rentable lorsque l'on pose plusieurs questions, (ce qui peut être le cas des profils). En effet une seule analyse de toute la base de données, permet de répondre à plusieurs profils à la fois.
Si l'on posséde un fichier des profils, ce n'est plus le fichier bibliographique (fichier des documents) qui est intérrogé pour la réponse à une question, mais le fichier des profils qui est intérrogé pour savoir si un ou plusieurs profils sont concernés par un document.

Mais la documentation automatique a ses exigences et ne peut être utilisée sans certaines précautions.
La pertinence des documents dépond considérablement de l'analyse et de l'indexation des documents.

En effet, un document mal analysé et mal indexé est perdu.
La recherche documentaire étant une comparaison des caractéristiques d'une question avec celles des documents mémorisés, la réponse donnée par le système n'est pas satisfaisante, si tous les documents susceptibles d'être intéressants, et uniquement ceux-la, ne figurent pas dans la réponse.
Si le contenu du document n'a pas été bien dégagé au niveau de l'analyse, il peut en résulter une perte d'informations, dite silence, ou un excés d'informations impertinentes dit bruit.

Le bruit est le taux de documents parasites qui ne répondent pas à la question posée, et qui sont extraits lors des opérations de sélection.
Ces bruits sont, le plus souvent, dus à l'introduction de descripteurs impropres au niveau de l'indexation. Certains des descripteurs du document, ne reflètent pas le sujet traité dans ce document.

Le silence est le taux de documents pertinents, mais qui ne sont pas extraits lors des opérations de sélection. Ces silences provenant, la pluvart du temps, du manque de descripteurs au niveau de l'indexation du document. En effet si l'indexeur n'introduit pas suffisament de descripteurs, le document ne répondra pas à certaines questions utilisant des descripteurs correspondant à un sujet traité dans l' l'ouvrage, mais n'apparaissant pas dans l'indexation.

En documentation automatique,. obtenir ces deux défauts est inévitable car : diminuer le silence en introduisant un grand nombre de descr-pteurs, augmente le bruit et réciproquement.

Dans les 2 cas cette perte de documents entraine une mise en doute du système par l'utilisateur. Nous voyons donc que quelque soit la perfection du système de documentation automatique, son utilisation, n'est valable que dans la mesure où, l'on fournit à ce système, des documents correctement indexés. Le rôle très important de l'indexation nous apparait alors.
Celle-ci nécessite des spécialistes ayant une expérience poussée dans le domaine qu'ils traitent. En effet, pour pouvoir indexer correctement un document, il faut connaitre parfaitement la matière traitée.

Il existe un certain nombre de grands organismes possédant des indexeurs spécialisés. Ces organismes sont producteurs de documents analysés mis généralement sous forme de bande magnétique. Nous pouvons noter INSPEC, COMPENDEX, Chemical Abstracts Condensates (CAC), MEDLARS; et INIS, AGRIS, DEVSIS, SPINES (les 4 derniers étant dûs à des organismes internationaux).

Mais il est bien entendu, que l'aquisition de ces bandes, ne résoud pas tous les problèmes. Il faut pour les traiter mettre en jeu des moyens importants et des ordinateurs de taille conséquente.

Il existe actuellement sur le marché 2 types de logiciels :

* Les logiciels qui fournissent des bandes et en réalisent le traitement.

* Les logiciels traitant uniquement les bandes fournies par d'autres systèmes.

Un inconvénient de ces logiciels est que, généralement ils sont en partie écrits en langage assembleur et seront liés au choix d'un constructeur.

Parmi les produits fournis par les systèmes de documentation automatique, ceux ayant du succés auprés des utilisateurs sont surtout les résultats des recherches rétrospectives et des diffusions sélectives. Mais il existe aussi d'autre produits tels les index.

Nous avons entrepris l'étude approfondie de deux LOGICIELS documentaire. MISTRAL (V 2) et SPLEEN 2. Le premier entre dans la catégorie des systèmes à fichier inverse. Le second quand à lui utilise une méthode d'analyse séquentielle.
L'étude de SPLEEN 2 ayant été faite, dans le cadre de notre stage effectué au centre de Documentation des Sciences Humaines (CNRS), Nous avons pu étudier les programmes de la chaîne SPLEEN, et les tester sur des données propres.

Quant à MISTRAL (V 2) il nous a été fourni par la CII à Alger. Nous l'avons donc étudié, sans pouvoir approfondir le détail des programmes, qui nous ont été fournis sous forme de LOAD MODULES; nous ne pouvions donc les "examiner" de prés.

Nous avons testé MISTRAL (V 2) avec un stock documentaire de taille restreinte et nous avons récemment généralisé à une base de données géologiques.

En annexe nous donnons un aperçu sur les logiciels RIRMS (IBM), PASCAL (C.N.R.S), l'organisation des banques de données du bureau de Recherches Géologiqueset Minières (France) où nous avons séjourné une semaine et les bases de données SFLFEN).