D.E.S.S. en Informatique Documentaire

N O T E de S Y N T H E S E

' Les modèles théoriques en documentation '

Année 1979 - 1980

Catherine GRIMAULT

S O M M A I R E

Introduction

I Historique

Introduction

De nombreuses théories mathématiques ont été utilisées pour décrire les modèles de recherche de l'information : théories basées sur les ensembles, sur les graphes, sur les probabilités, sur la classification, sur la linguistique, etc...

Les premières approches formelles des problèmes de recherche de l'information datent d'environ vingt ans. Cependant, les mises en oeuvre réussies dans des milieux opérationnels sont relativement récentes.

Dans une première partie nous allons présenter rapidement les idées de base des premiers modèles puis, dans un second temps, nous décrirons plus en détail les principaux modèles étudiés de nos jours.

I Historique

La théorie de Spark Jones, basée sur la linguistique, suppose que les mots reflétant le contenu d'un document sont ceux qui apparaissent fréquemment. Contrairement à cette hypothèse, d'autres chercheurs pensent que les termes rares sont les plus importants. Cette théorie, basée sur une hypothèse contestée, n'est donc guère valable et ne peut être que de peu d'utilité. Ainsi, l'expérience reste la seule façon d'approcher le problème de manière linguistique.

Une autre approche est donnée par la théorie de l'information de Shannon. Selon ce dernier, les termes représentatifs d'un document doivent équifréquents. Par contre, selon Zunde et Slamecka, la distribution de ces termes doit être géométrique. Là encore les avis diffèrent et on a voulu se baser sur l'expérience pour trancher. Finalement les deux hypothèses ont été abandonnées, les résultats des expériences ne concordant pas avec la théorie.

En fait, ce qui semble manquer le plus à ces premières théories est la connaissance de la fonction qu'éxécute le système pour faire sa recherche. Récemment, de nombreuses théories essayant d'y remédier ont été développées. Ce sont en particulier les théories de modèles structurés dont les premiers travaux sont dus à Mooers, Faithorne et Hillman. La préoccupation principale est de savoir si

on peut appliquer ou non l'algèbre de Boole à la recherche de l'information. Hillman, en s'inspirant de Faithorne, a construit une théorie topologique non booléenne : il suppose que certains documents ne peuvent être décrits ni par A ni par non A.

Le modèle stucturé de Soergel a pour but de représenter la structure logique des termes d'index et de relier les documents entre eux. Soergel considérait que son modèle pouvait constituer un préalable pour un modèle fonctionnel mais son idée n'a pas été développée par la suite.

Un modèle simple présenté par Bookstein et Cooper identifie les aspects de structure les plus importants. Il est basé sur un ensemble de termes d'index, un ensemble de formulations de recherche et une fonction d'appariement qui ordonne partiellement les documents en réponse à une formulation de recherche. Cette idée d'ordre est nouvelle et constitue une étape dans la formulation des modèles de recherche.

II Etude de quelques modèles

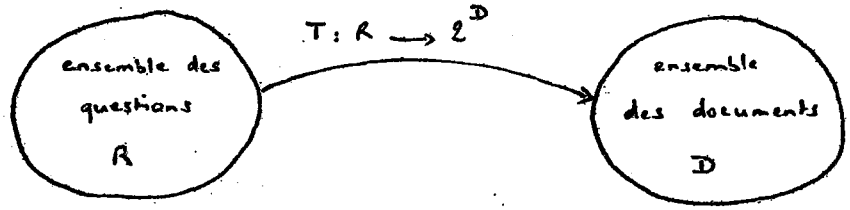
1) Les modèles ensemblistes.

Le choix d'un modèle ensembliste pour représenter un système de recherche de l'information est justifié par le fait que chaque recherche de documents se traite avec des ensembles : l'ensemble des descripteurs et l'ensemble des documents. En fait, les données fondamentales de la recherche sont fournies, dans cette théorie, par les relations entre l'ensemble des descripteurs d'un sujet et l'ensemble des documents correspondants.

Pour commencer, nous nous placerons dans le cas le plus simple constitué :

- d'un ensemble R de questions ou demandes d'informations
- d'un ensemble D de documents que nous choisissons statique, c'est-à-dire qu'il ne varie pas au cours du temps.

On utilise alors pour représenter les opérations de recherche une application $T : R \longrightarrow 2^D$ appelée fonction de recherche. A chaque élément r de R correspond par T un élément, noté T(r), qui appartient à l'ensemble de tous les sous-ensembles de D.



Représentation de l'application T

Afin d'étudier les propriétés de cette application, nous allons définir des relations d'ordre.

- Relation d'ordre dans l'ensemble des questions R :

Chaque question est formulée en choisissant des descripteurs c'est-à-dire en sélectionnant un sous-ensemble de l'ensemble R. La relation d'ordre $\lesssim$ établie pour les questions est donc fondée sur la relation d'inclusion des sous-ensembles de R.

- Relation d'ordre dans l'ensemble des documents D :

Cette relation est la relation d'inclusion des sous-ensembles de D.

La fonction de recherche T qui vérifie la propriété

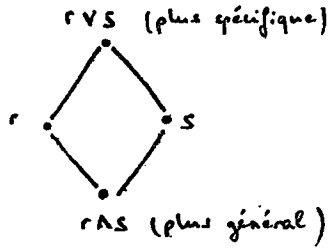
$$r,s \in R \qquad r \geqslant s \Rightarrow T(r) \subseteq T(s)$$

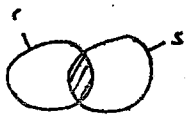
est une fonction inclusive. On dit alors que le système de recherche de l'information est inclusif.

Soit S un sous-ensemble ordonné de R et soit r,s des éléments de S. On définit alors la limite supérieure de r et s, notée $r \vee s$ (se lit r union s) et la limite inférieure, notée $r \wedge s$ (se lit r intersection s). S est un treillis défini par $(S,\wedge,\vee)$ et les opérations union et intersection sont équivalentes respectivement à l'union ensembliste et à l'intersection ensembliste des sous-ensembles correspondants. Comme la fonction de recherche T envoie les éléments de S sur les éléments de 2^D et que les sous-ensembles de D forment un treillis, nous avons pour chaque paire d'ensembles de documents $A,B \in 2^D$:

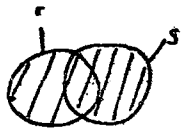
$$A \wedge B = \lim \inf (A,B) = A \cap B$$
$$A \vee B = \lim \sup (A,B) = A \cup B$$

Considérons maintenant les nouvelles questions $r \wedge s$ et $r \vee s$, définies précédemment, formées en combinant des éléments inclus dans les questions précédentes. $r \vee s$ est plus spécifique que r ou s seul et par conséquent, puisque le système est inclusif, on trouvera moins de documents qu'avec r seul ou s seul. De même, puisque $r \wedge s$ est une question plus générale on doit trouver plus de documents qu'avec r ou s seule. Cette situation se représente graphiquement par :





$r \wedge s$



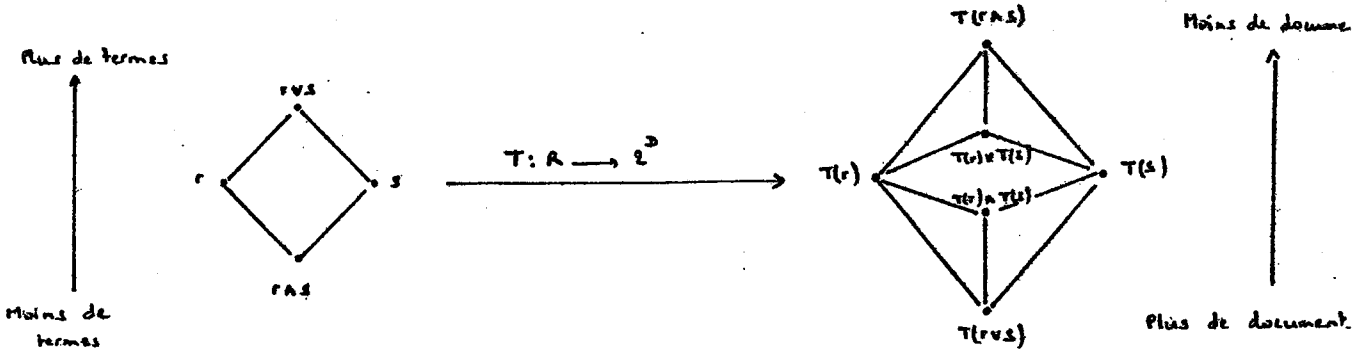
$r \vee s$

On peut maintenant donner une autre définition d'une fonction inclusive :

T est une fonction de recherche inclusive si et seulement si

$$r,s \in R \quad T(r \vee s) \subseteq T(r) \wedge T(s) \subseteq T(r) \cap T(s) \quad \circledast$$
$$T(r \wedge s) \supseteq T(r) \vee T(s) \supseteq T(r) \cup T(s)$$

Cette propriété peut être illustrée par la figure suivante :



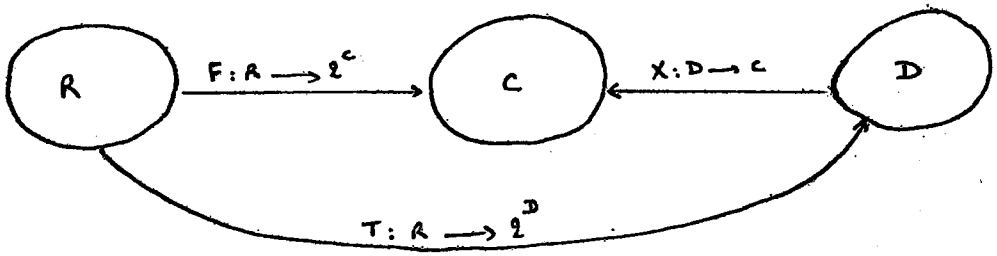
Dans la plupart des systèmes de recherche, les documents physiques ne sont pas les images directes de la fonction de recherche. Le système agit sur des descriptions de documents, par exemple des mots-clés. De plus, l'ensemble des documents peut être différent de toutes les questions possibles. On définit alors une fonction X de l'ensemble des documents D dans l'ensemble des descriptions possibles des documents. X est appelée fonction classification et fait correspondre à chaque élément d de D une valeur X(d) appelée description de d. X permet de définir une relation d'équivalence sur D par : un élément donné d_i est équivalent à d_j si et seulement si

$$X(d_i) = X(d_j)$$

Un système complet de recherche peut donc être établi à partir d'un langage de recherche R et d'un ensemble de documents D tous deux munis des fonctions $X : D \longrightarrow C$ et $F : R \longrightarrow 2^C$. (C désigne l'ensemble des descriptions possibles des documents)
La fonction de recherche $T : R \longrightarrow 2^D$ est alors définie par :

$$T(r) = \{d : X(d) \in F(r)\}$$

Ce modèle peut être illustré par le schéma suivant :



Dans certaines conditions, pour un système basé sur la classification, l'inégalité ⊛ devient une égalité. Dans un système de recherche basé sur la classification où $R = \overline{C}$ et où $T(r) = \{d : r \leq X(d)\}$ on a, chaque fois que $r \vee s$ existe :

$$T(r \vee s) = T(r) \wedge T(s) = T(r) \cap T(s)$$

Si de plus R est un treillis et T une (1,1) application de R dans 2^D alors :

$$T(r \wedge s) = T(r) \cup T(s)$$

En général, on considère un vocabulaire fini de descripteurs et on prend l'ensemble R identique à C. On peut alors définir deux fonctions de recherche :

$$T_1(r) = \{d : r = X(d)\}$$

$$T_2(r) = \{d : r \leq X(d)\}$$

Dans la pratique, T_1 n'est pas une fonction très satisfaisante car pour deux questions similaires (différant par exemple d'un seul terme) on trouvera deux ensembles différents. T_2 est une fonction inclusive.

Jusqu'ici, les identificateurs d'information utilisés étaient des propriétés positives, c'est-à-dire qu'un identificateur était attribué à un document lorsque celui-ci possédait la propriété correspondante. Dans de nombreux systèmes il est pratique d'opérer avec des identificateurs d'information négatifs. Par exemple $\bar{A}$ signifie non rouge si A désigne la propriété 'de couleur rouge'.
On a alors deux cas :

- Soit la propriété négative est considérée indépendante de toute autre propriété existante c'est-à-dire qu'une propriété intitulée non rouge n'est pas le complément de la propriété rouge mais une autre propriété.

- Soit la propriété négative est considérée être le complément logique c'est-à-dire les documents sont indexés soit avec l'identificateur positif soit avec le négatif.

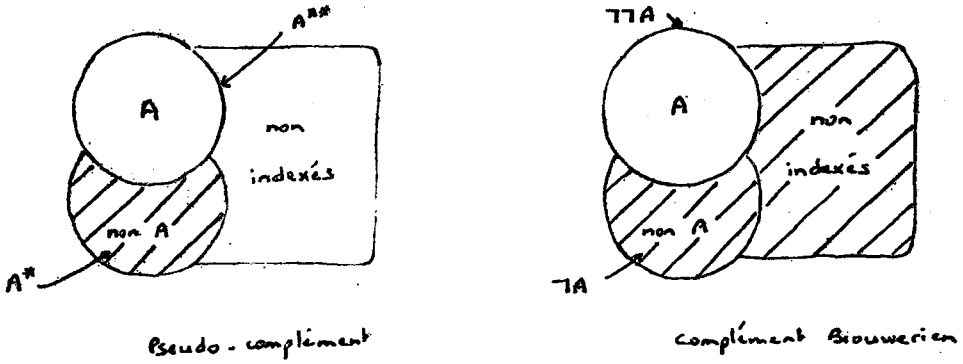
Considérons maintenant une collection de documents dans un système évolutif où de nouveaux documents sont continuellement ajoutés et les vieux effacés. Le modèle théorique ensembliste ne peut s'appliquer à ce système que localement et temporairement pour certaines sous-sections de la collection pour lesquelles aucun changement récent n'aura été enregistré.

Dans cette situation, il existe pour chaque ensemble de documents A deux compléments :

- A^*, appelé le pseudo-complément, défini comme le plus grand ensemble de documents qui ne contient aucun A. A^* est donc constitué de tous les documents connus pour être non A mais ne contient aucun des articles non indexés car ils peuvent être des A.

- $\neg A$, appelé le complément Brouwerien, défini comme le plus petit ensemble de documents contenant tous les non A. Cet ensemble contient tous les documents non indexés car certains d'entre eux peuvent être des non A.

Le double pseudo-complément A^{**} est le plus petit ensemble de documents contenant tous les A. Le double complément Brouwerien $\neg\neg A$ est le plus grand ensemble ne contenant que les A.
Cette différence de complément est illustrée par la figure suivante :



On peut utiliser les deux algèbres définies par ces compléments comme modèle de système de recherche. De par la définition des compléments, l'algèbre du pseudo-complément sera utilisée lorsque l'on désire un rappel élevé tandis que l'algèbre Brouwerienne le sera pour une grande précision. L'imprécision de l'ensemble retrouvé peut être mesurée par la différence des compléments :

$$\text{imprécision de } A = A^{**} - \neg\neg A$$

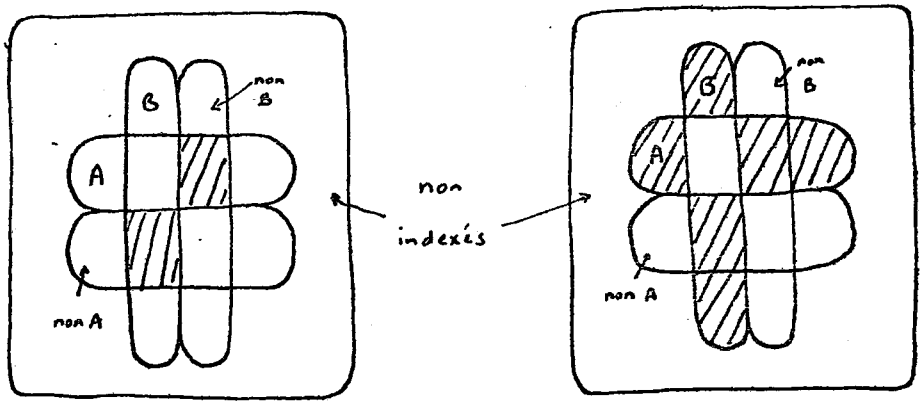
Ainsi, toutes les opérations comprenant les compléments d'ensemble peuvent avoir deux versions différentes. Par exemple, la différence symétrique de deux ensembles, qui mesure la distance entre eux prend les deux formes suivantes :

$$(A + B)^* = (A \cap B^*) \cup (A^* \cap B)$$

$$\neg(A + B) = (A \cap \neg B) \cup (\neg A \cap B)$$

$(A + B)^*$ représente le plus grand ensemble de documents appartenant

à un ensemble mais pas à l'autre tandis que $\neg(A + B)$ représente le plus petit ensemble de documents appartenant à un ensemble mais pas à l'autre. Ces deux types de différence symétrique peuvent être représentées par le schéma suivant :



$(A \cap \neg B) \cup (\neg A \cap B)$ $\qquad$ $(A \cap \neg B) \cup (\neg A \cap B)$

On a aussi eu l'idée d'utiliser comme modèle de recherche un espace topologique d'ouverts superposé sur la structure d'ensembles du début. On définit alors des voisinages et les documents sont attribués aux voisinages pour lesquels ils paraissent les plus proches. Chaque voisinage d'un terme (ou d'un document) est interprété comme ensemble de termes (ou de documents) similaires ou mutuellement pertinents et la collection des voisinages, pour un espace topologique T, est supposée fournir une représentation des relations de pertinence entre les éléments de T. Les voisinages eux-même sont définis pour chaque élément comme des ouverts contenant l'élément donné. On utilise l'algèbre des fermés d'un espace topologique pour des recherches de haute précision et l'algèbre des ouverts ou des voisinages de documents pour des recherches avec un rappel élevé. Dans le premier cas on utilise l'algèbre Brouwerienne, dans le second celle du pseudo-complément.

2) Le modèle vectoriel.

Un document est identifié par un ensemble de mots-clés ou propriétés. Nous pouvons donc le représenter par un vecteur-propriété V de dimension t, le ième élément v_i représentant le poids de la ième propriété. Un poids égal à zéro signifie que la propriété i n'est pas applicable au document représenté par ce vecteur. Un élément positif indique que la propriété est applicable avec un poids égal à cet élément. Dans de nombreuses applications, seuls les vecteurs-propriété binaires sont autorisés : les poids ne peuvent alors prendre que les valeurs 0 ou 1.

Dans ce modèle trois fonctions de vecteurs sont importantes :

$\sum_{i=1}^{t} v_i$ somme des poids de toutes les propriétés incluses dans v. Pour des vecteurs binaires, cette somme est égale au nombre de propriétés différentes de zéro noté N_v.

$\sum_{i=1}^{t} v_i w_i$ produit scalaire des vecteurs v et w. Pour des vecteurs binaires, ce produit représente le nombre d'appariement de propriétés (c'est-à-dire les propriétés qui sont au même rang dans v et w) différentes de zéro, noté N_{vw}.

$\sqrt{\sum_{i=1}^{t} v_i^2}$ longueur du vecteur V.

Pour déterminer la similitude de deux vecteurs-propriété il existe plusieurs mesures de corrélation. La première est le produit scalaire habituel :

$$r_{vw} = \sum_{i=1}^{t} v_i w_i$$

L'inconvénient de cette mesure est qu'elle ne tient pas compte de la fréquence des propriétés apparaissant dans l'un des vecteurs pas plus que du nombre total de propriétés différentes de zéro. Le fait qu'une propriété soit absente dans les deux vecteurs est pris en considération dans la mesure suivante :

$$r_{vw} = \sum_{i=1}^{t} v_i w_i + \sum_{i=1}^{t} \bar{v}_i \bar{w}_i$$

$\bar{v}_i$ représentant le complément de v_i.

En fait, cette mesure ne peut s'appliquer sans ambiguïté qu'aux vecteurs binaires ($v_i = 0 \Rightarrow \bar{v}_i = 1$).

La mesure suivante :

$$r_{vw} = \frac{\sum_{i=1}^{t} v_i w_i}{\sum_{i=1}^{t} v_i + \sum_{i=1}^{t} w_i - \sum_{i=1}^{t} v_i w_i}$$

attribuée à Tanimoto est normalisée, c'est-à-dire que sa gamme de valeurs varie de 0 à 1.
La similitude peut aussi être mesurée par le cosinus de l'angle compris entre les vecteurs v et w soit :

$$r_{vw} = \frac{\sum_{i=1}^{t} v_i w_i}{\sqrt{\sum_{i=1}^{t} v_i^2 \sum_{i=1}^{t} w_i^2}}$$

Une autre mesure définie pour des vecteurs binaires a été suggérée par Maron et Kuhns. Elle suppose l'indépendance statistique des vecteurs et s'écrit :

$$r_{vw} = \frac{(\sum_{i=1}^{t} v_i w_i)(\sum_{i=1}^{t} \bar{v}_i \bar{w}_i) - (\sum_{i=1}^{t} v_i \bar{w}_i)(\sum_{i=1}^{t} \bar{v}_i w_i)}{(\sum_{i=1}^{t} v_i w_i)(\sum_{i=1}^{t} \bar{v}_i \bar{w}_i) + (\sum_{i=1}^{t} v_i \bar{w}_i)(\sum_{i=1}^{t} \bar{v}_i w_i)}$$

En termes de fréquence d'items, cette formule s'écrit :

$$r_{vw} = \frac{N_{vw} N_{\bar{v}\bar{w}} - N_{v\bar{w}} N_{\bar{v}w}}{N_{vw} N_{\bar{v}\bar{w}} + N_{v\bar{w}} N_{\bar{v}w}}$$

Considérons maintenant deux documents spécifiques i et j parmi tous les documents de la collection. L'ensemble des propriétés peut être divisé en quatre parties : N_{ij}, le nombre de propriétés attribuées à la fois à i et à j ; $N_{i\bar{j}}$, le nombre attribué au document i mais pas à j ; $N_{\bar{i}j}$, le nombre attribué à j mais pas à i et enfin $N_{\bar{i}\bar{j}}$, le nombre de propriétés qui ne sont attribuées ni à i ni à j. Il est clair que si t est le nombre total de propriétés on a :

$$t = N_{ij} + N_{i\bar{j}} + N_{\bar{i}j} + N_{\bar{i}\bar{j}}$$

De plus :

$$N_i = N_{ij} + N_{i\bar{j}}$$

$$N_j = N_{ij} + N_{\bar{i}j}$$

Le fait que les propriétés soient statistiquement indépendantes se traduit en termes de fréquence par :

$$N_{ij} = \frac{N_i N_j}{t}$$

Une valeur importante pour la suite est la valeur δ_{ij} définie par :

$$\delta_{ij} = N_{ij} - \frac{N_i N_j}{t}$$

Pour construire une mesure de corrélation basée sur l'attribution d'une propriété donnée on peut utiliser le critère suivant :

1- La mesure de corrélation r_{ij} entre les documents i et j doit être 0 quand $\delta_{ij} = 0$ et doit varier comme δ_{ij} pour un nombre fixé de propriétés t et un nombre de documents n.

2- La corrélation maximum r_{ij} doit avoir lieu lorsque le vecteur-propriété identifiant le document i est contenu dans le vecteur-propriété assigné à j. Cette corrélation maximum est obtenue lorsque $N_{i\bar{j}} = 0$ ou $N_{\bar{i}j} = 0$ ou $N_{i\bar{j}} = N_{\bar{i}j} = 0$.

3- La corrélation minimum doit avoir lieu lorsque le vecteur-propriété identifiant le document i est inclus dans le complément de celui attribué à j ou inversement ou lorsqu'un vecteur-propriété est le complément de l'autre. Formellement ceci se traduit par $N_{ij} = 0$ ou $N_{\bar{i}\bar{j}} = 0$ ou $N_{ij} = N_{\bar{i}\bar{j}} = 0$.

De nombreuses mesures de corrélation ont été utilisées expérimentalement sur plusieurs collections de documents. Sur l'ensemble, cependant, il apparait que les différences de performance résultant de l'utilisation de mesures de corrélation différentes sont de second ordre par rapport aux différences dues aux nombreux types de procédures d'analyse.

3) Le modèle matriciel.

Considérons un ensemble D de n documents et un ensemble A de t attributs ou propriétés utilisés pour identifier ces documents. Un document D_i peut être représenté par un vecteur-propriété

$$D_i = (a_{i1}, a_{i2}, \ldots, a_{it})$$

où a_{ij} représente le poids ou degré d'importance de la propriété A_j dans D_i. On peut ainsi caractériser l'ensemble complet des documents par une matrice C propriété-document de dimension n x t.

$$C = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots\cdots & A_t \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots\cdots & a_{1t} \\ a_{21} & a_{22} & \cdots\cdots & a_{2t} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots\cdots & a_{nt} \end{pmatrix} \end{array}$$

Chaque ligne représente un enregistrement D_i et chaque colonne correspond à l'attribution d'une propriété particulière aux documents de l'ensemble. Lorsqu'une propriété n'est pas attribuée à un document on prend comme poids correspondant 0.

Considérons maintenant une opération de recherche. Si une question Q est représentée par un vecteur-propriété de dimension t

$$Q = (q_1, q_2, \ldots, q_t)$$

où q_i est le poids de la propriété A_i dans la question, on calcule une mesure de similitude r_i entre la question Q et le document D_i par :

$$r_i(D, Q_i) = \sum_{j=1}^{t} a_{ij} q_j \qquad (a)$$

On utilise aussi des fonctions similaires à celles définies dans la méthode vectorielle décrite précédemment :
la fonction cosinus :

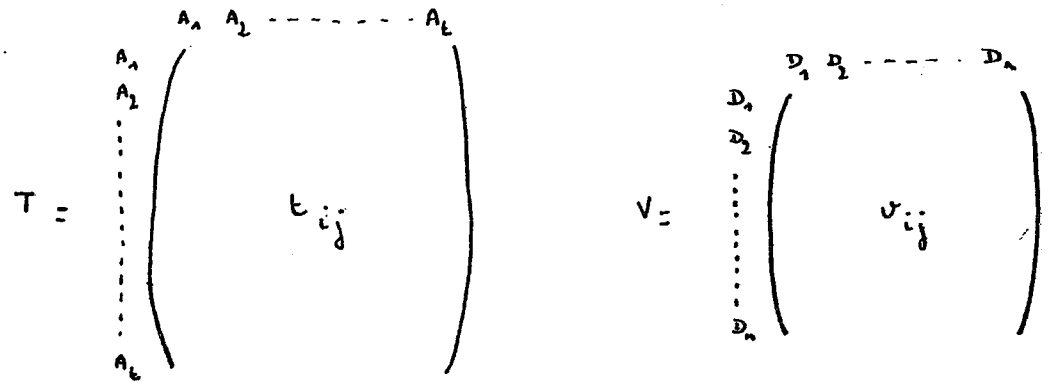
$$r_i'(Q, D_i) = \frac{\sum_{j=1}^{t} a_{ij} q_j}{\sqrt{\sum_{j=1}^{t} q_j^2 \sum_{j=1}^{t} a_{ij}^2}}$$

ou la fonction :

$$r''_i(Q,D_i) = \frac{\sum_{j=1}^{t} a_{ij}q_j}{\sum_{j=1}^{t} q_j^2 + \sum_{j=1}^{t} a_{ij}^2 - \sum_{j=1}^{t} q_j a_{ij}}$$

Mais, ces formules supposent que les vecteurs-propriété sont orthogonaux ce qui n'est pas toujours le cas en pratique. On envisage alors une matrice T de similitude propriété dont le terme général t_{ij} représente le coefficient de similitude entre la ième et la jéme propriété. De même, on peut calculer des coefficients de similitude v_{ij} entre les documents D_i et D_j. On obtient ainsi une matrice V de dimension n x n.

En général, on utilise des mesures symétriques ($v_{ij} = v_{ji}$) et on suppose que tous les éléments de la diagonale sont égaux. Ainsi, seulement $\frac{n(n-1)}{2}$ des n^2 éléments de V représentent des propriétés indépendantes.



En notation vectorielle, l'équation (∞) s'écrit :

$$r = Cq$$

En considérant les similitudes de propriété et de documents, l'équation de recherche devient :

$$r = VCTq$$

Les matrices de similitude T et V reflètent un premier ordre de similitude entre les paires d'attributs et les paires de documents respectivement. Des similitudes d'ordre plus élevé entre des triplets, des quadruplets, etc..., peuvent être aussi utilisées mais, en principe l'utilité pratique doit diminuer rapidement lorsque l'ordre de similitude augmente.

4) Le modèle probabiliste.

D'un point de vue théorique, on peut exprimer la tâche de base de la recherche par trois paramètres principaux :

$P(per)$ probabilité de pertinence

a_1 paramètre de perte associé à la recherche d'enregistrements non pertinents ou étrangers.

a_2 paramètre de perte associé au fait de ne pas trouver un enregistrement pertinent.

Le fait de trouver un terme étranger cause une perte de $(1 - P(per))a_1$ tandis que le rejet d'un terme pertinent produit une perte de $P(per)a_2$. La perte totale est donc minimisée si la recherche d'un document ne se fait que lorsque

$$P(per)a_2 \geqslant (1 - P(per))a_1$$

c'est-à-dire lorsque la fonction g définie par :

$$g = \frac{P(per)}{1 - P(per)} - \frac{a_1}{a_2}$$

est positive ou nulle.

Cependant, cette règle est peu utilisable en pratique car les propriétés de pertinence ne peuvent pas être séparées des autres paramètres du système. On définit alors des probabilités conditionnelles :

$P(x_i/w_1)$ probabilité pour que l'enregistrement contenant x_i soit pertinent pour une question donnée.

$P(x_i/w_2)$ probabilité pour que cet enregistrement ne soit pas pertinent.

D'après la formule de Bayes, la fonction de recherche $P(w_i/x)$ pour le terme x s'écrit :

$$P(w_i/x) = \frac{P(x/w_i)P(w_i)}{P(x)} \qquad i = 1 , 2$$

où $P(x) = \sum_{i=1}^{2} P(x/w_i)P(w_i)$

Si les deux paramètres de perte prennent la valeur 1 ($a_1 = a_2 = 1$) alors la recherche est efficace lorsque

$$P(w_1/x) \geqslant P(w_2/x)$$

c'est-à-dire lorsque $g(x) = \dfrac{P(w_1/x)}{P(w_2/x)} \gtrless 1$

En pratique, les probabilités $P(x/w_i)$ doivent être déterminées en deux temps. Il faut d'abord faire le calcul d'occurence pour chaque terme pris séparément puis spécifier les interactions entre termes. Considérons d'abord le cas simple où les termes sont indépendants. On a alors :

$$P(x/w_i) = P(x_1/w_i)P(x_2/w_i).....P(x_n/w_i)$$

On peut se restreindre au cas où les vecteurs d'information sont binaires c'est-à-dire $x_i = 1$ si le ième terme est présent $x_i = 0$ sinon. L'équation devient alors :

$$P(x/w_1) = \prod_{i=1}^{n} P_i^{x_i} \ (1 - P_i)^{1 - x_i}$$

$$P(x/w_2) = \prod_{i=1}^{n} q_i^{x_i} \ (1 - q_i)^{1 - x_i}$$

où $P_i = P(x_i = 1/w_1)$

$q_i = P(x_i = 1/w_2)$

La fonction g s'écrit alors :

$$g(x) = \sum_{i=1}^{n} \left\{ x_i \log \frac{P_i}{q_i} + (1 - x_i) \log \left(\frac{1 - P_i}{1 - q_i} \right) \right\} + \log \frac{P(w_1)}{P(w_2)}$$

Il reste à déterminer les probabilités d'occurence de chaque terme x_i séparément. Pour cela, considérons une collection échantillon de N enregistrements où R enregistrements sont pertinents pour question Q. On a alors pour un terme x_i le tableau suivant :

	w_1	w_2	
$x_i = 1$	r_i	$n_i - r_i$	n_i
$x_i = 0$	$R - r_i$	$N - n_i - R + r_i$	$N - n_i$
	R	$N - R$	

On peut donc écrire :

$$P_i = \frac{r_i}{R} \qquad q_i = \frac{n_i - r_i}{N - R}$$

La fonction de recherche g(x) devient alors :

$$g(x) = \log \frac{P(w_1)}{P(w_2)} + \sum_{i=1}^{n} \log \frac{1 - P_i}{1 - q_i} + \sum_{i=1}^{n} x_i \log \frac{r_i / R - r_i}{n_i - r_i / N - n_i - R + r_i}$$

D'après le premier terme, on ajoute pour chaque x_i un poids proportionnel à

$$L_i = \frac{r_i / R - r_i}{n_i - r_i / N - n_i - R + r_i}$$

L_i est appelé la limite de pertinence du terme x_i.

Pour utiliser un tel modèle, il faut avoir une information suffisante sur les probabilités d'occurence des termes. Malheureusement, cette information est rarement connue à priori. On utilise alors des distributions théoriques de probabilités en supposant que les données réelles suivent le modèle théorique. Une des plus connues et des plus utilisées est la distribution de Poisson qui est assez facile à traiter mathématiquement.

Si p occurences d'un terme sont réparties dans des textes de longueurs approximativement égales, la probabilité P(k) pour qu'un texte contienne k fois ce terme est :

$$P(k) = \frac{1}{k!} \left(\frac{p}{n}\right)^k e^{-p/n}$$

où $\frac{p}{n}$ représente à la fois la moyenne et la variance de la distribution. p est donc proportionnel à la variance. Ceci a été utilisé dans plusieurs des premiers modèles automatiques d'indexation en notant que les mots de spécialité capables de représenter le contenu de l'information ont tendance à être groupés dans peu de documents tandis que les mots qui ne sont pas de spécialité présentent des caractéristiques d'occurence plus uniformes.

On peut étendre le modèle de Poisson en supposant que sa distribution reflète non seulement l'occurence des mots courants mais aussi ceux de spécialité. Si il n'existe que deux classes de pertinence (un document est soit pertinent, soit non pertinent) on suppose que deux distributions de Poisson caractérisent les occurences des mots de spécialité, la première représentant les termes pertinents, l'autre les non pertinents. Si π et $(1 - \pi)$ représentent les probabilités pour qu'un document appartienne aux classes 1 et 2 respectivement, λ_1 et λ_2 étant les moyennes de distributions de Poisson respectives, la probabilité pour qu'un mot de spécialité donné soit k fois présent dans un document est :

$$P(k) = \pi \frac{e^{-\lambda_1} \lambda_1^k}{k!} + (1 - \pi) \frac{e^{-\lambda_2} \lambda_2^k}{k!}$$

La fonction de décision g s'écrit alors :

$$g(x) = e^{-(\lambda_1 - \lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^k \left(\frac{\pi}{1 - \pi}\right)$$

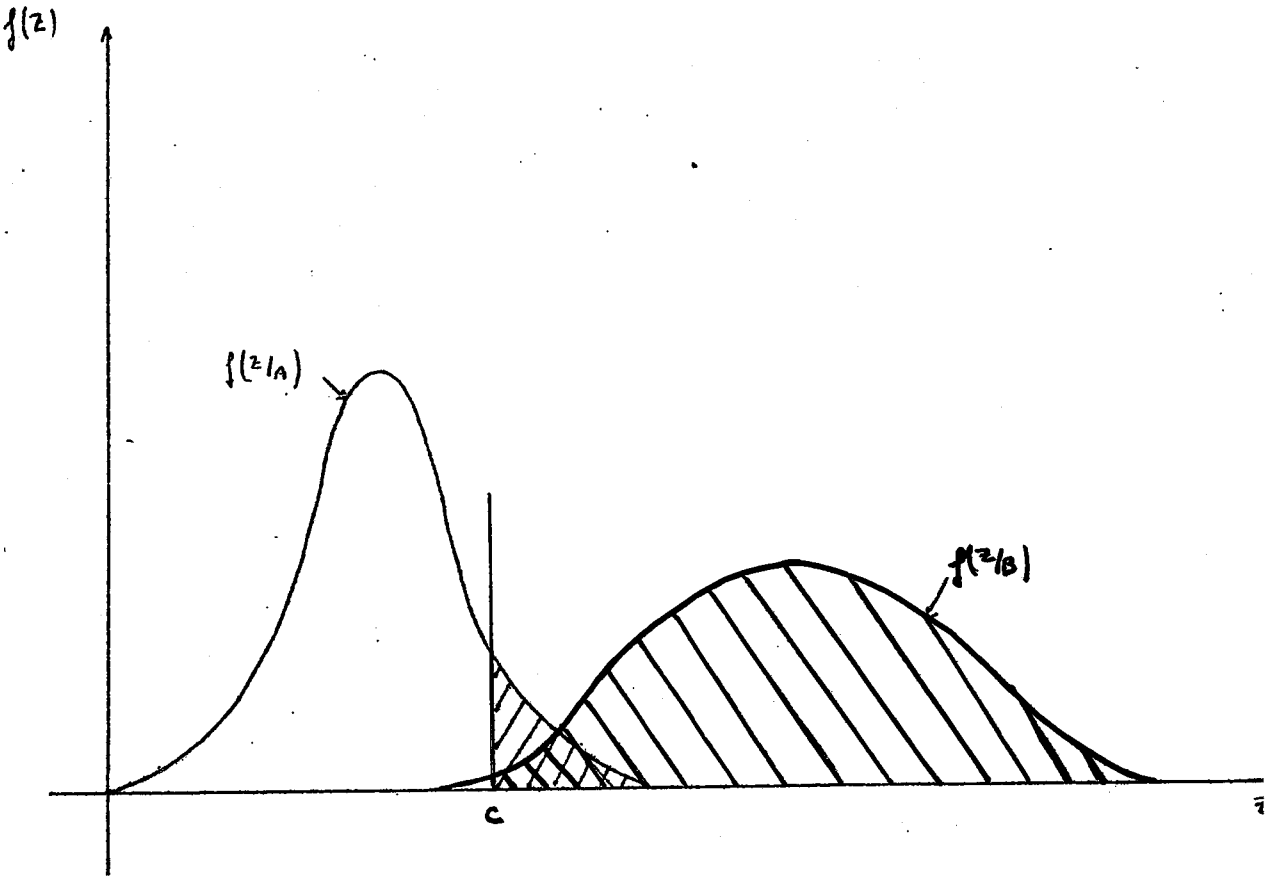
Bien que le modèle double de Poisson ne soit pas parfait (la séparation des enregistrements en deux classes homogènes est irréalisable en pratique), l'idée a été reprise dans de nombreux modèles.

Soit deux populations distinctes d'objets A et B identifiées respectivement dans un environnement de recherche d'information par les enregistrements pertinents et les non pertinents par rapport à une demande quelconque d'information. On calcule alors le coefficient de similitude Z question-enregistrement pour chaque élément de A et de B pour cette question. Le rendement du système est obtenu en mesurant les différences entre les fonctions densité de probabilité respectives f(Z).

articles pertinents $f(Z/B)$

articles non pertinents $f(Z/A)$

Une distribution typique pour les coefficients de similitude est montrée dans la figure suivante :



Fonctions densité de probabilité pour les enregistrements pertinents (B) et les enregistrements non pertinents (A).

On définit une valeur de coupure Z = C. La proportion d'objets B pour lesquels $Z \geqslant C$ est représentée par l'aire comprise entre la courbe $f(z/B)$ et la partie à droite de la coupure. Il en est de même pour les objets A.

5) Le modèle de Swets.

Swets considère que le processus de recherche a deux étapes : en réponse à une question, le système calcule pour chaque document la valeur d'une fonction d'appariement. Le système sélectionne ensuite les documents correspondant aux valeurs les plus élevées de la fonction ou aux valeurs plus grandes qu'un certain seuil. Le modèle est très semblable au modèle de Bookstein et Cooper mais la différence réside dans l'interprétation de la valeur seuil. Pour Swets, cette valeur est fixée à priori, avant que la recherche ne commence tandis que pour Bookstein et Cooper cette valeur est fixée par l'utilisateur selon les résultats obtenus c'est-à-dire qu'elle est fixée à postériori.

Le modèle de Swets dépend d'hypothèses sur la nature des distributions de la fonction d'appariement dans les documents pertinents et non pertinents. La forme la plus simple du modèle de Swets suppose que ces distributions sont normales et de variances égales. Avec cette hypothèse, lorsque l'on trace la courbe de recherche dans des axes appropriés on doit obtenir une droite à 45 degrés. Mais, lorsque Swets testa son modèle il découvrit que l'on obtenait une ligne droite mais qu'elle n'était pas toujours à 45 degrés. Il décida alors d'adopter un modèle plus général pour lequel les distributions sont toujours supposées normales mais sont de variances différentes. Ce modèle donne une ligne droite mais, l'angle est quelconque (l'angle est relié au ratio des variances).

Heine a suggéré que la théorie de Swets soit vue sous deux aspects : une forme faible liée aux moyennes des questions et une forme forte liée aux questions individuelles. Cette suggestion est interessante car elle concorde avec les données expérimentales de Swets qui montrent des variations plus grandes de la pente de la droite quand elle est tracée pour des questions individuelles. On peut expliquer ceci par le fait que les variations de la pente sont liées à la taille de l'échantillon des questions et que la moyenne des questions serait une droite à 45 degrés.

6) Le modèle de dimension N (extension de la méthode indirecte de Goffman)

Dans les modèles de recherche traditionnels, on utilise les fonctions booléennes et on suppose que les documents sont indépendants les uns des autres. Cependant, Goffman a clairement montré que les documents sont interdépendants. Le modèle suivant se propose d'élargir la méthode indirecte de Goffman en utilisant d'autres moyens que les termes d'index pour refléter le contenu des documents. On définit une distance entre les documents basée sur les propriétés des documents eux-même. L'utilisation des termes d'index comme mesure de la proximité entre deux documents X_i et X_j peut être représentée par le segment compris entre X_i et X_j. Ajoutons maintenant une seconde mesure de telle sorte que deux mesures soient utilisées pour déterminer la proximité des documents. En utilisant la formule euclidienne de la distance, nous obtenons :

$$D(X_i, X_j) = \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2}$$

où $X_j - X_i$ est la première mesure et $Y_j - Y_i$ la seconde.
En général, il peut y avoir N mesures et la distance s'exprime alors par :

$$D(X_i, X_j) = \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2 + \ldots + (N_j - N_i)^2}$$

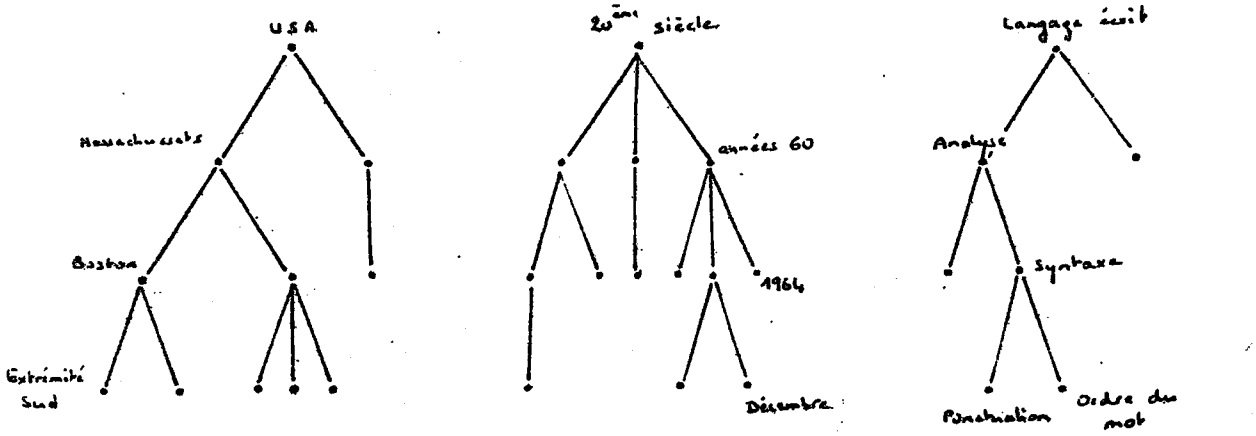
On suppose que les mesures sont linéairement indépendantes et mutuellement orthogonales créant un système de référence orthonormé dans l'espace euclidien de dimension N.
Cette fonction étant basée sur les propriétés des documents calcule donc la proximité entre deux documents à partir du calcul de la distance euclidienne.

Une expérience basée sur quatre mesures de proximité des documents : les termes d'index, les journaux dans lesquels le document apparait, les auteurs, les citations, a montré que les meilleures mesures de proximité étaient celles basées sur les auteurs et les auteurs plus les journaux.

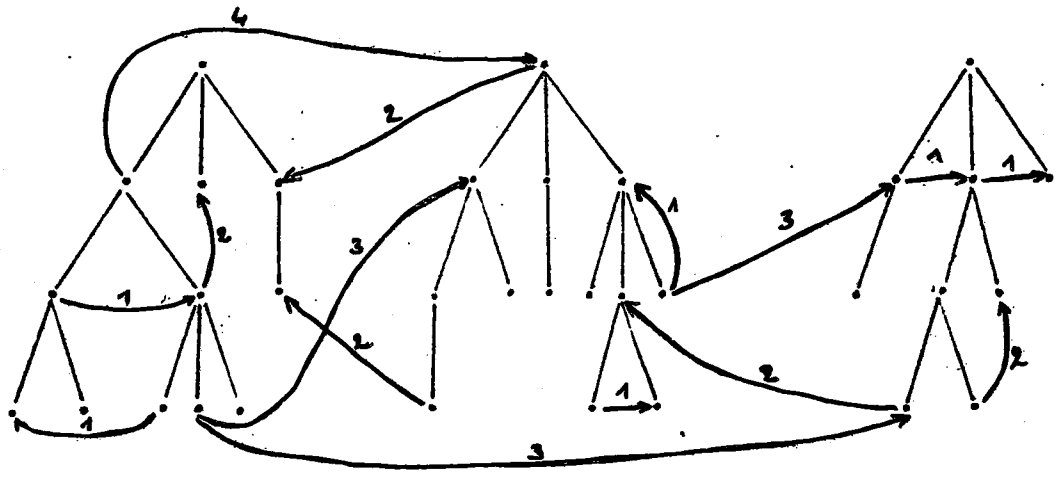
7) Les modèles en graphe ou en arbre.

Le modèle ensembliste n'est applicable que lorsqu'aucune structure n'est attribuée aux ensembles de documents. Mais, en pratique, on impose souvent des relations entre les documents et les modèles résultants prennent l'une des deux formes suivantes :

- Si un seul type de relation est défini entre les documents ou propriétés (relation d'ordre hiérarchique ou de similitude sémantique) alors la structure en arbre est la plus adéquate pour représenter les documents et leurs relations. Les documents ou propriétés sont les noeuds de l'arbre tandis que les relations sont représentées par les branches.
- Si on a plusieurs relations différentes, un ensemble de références croisées est superposé à la structure d'origine en arbre et l'arbre est alors transformé en graphe.



ARBRES



1 → relation de type 1

GRAPHE

En général, les opérations de recherche sont bien plus difficiles à réaliser avec les modèles de graphe qu'avec le modèle vectoriel. Pour cette raison, l'expérience avec le modèle graphe na jamais été beaucoup étendue et en particulier n'a jamais été réalisée avec des graphes de grande taille du type de ceux que l'on trouve en pratique.

8) Modèle utilisant l'âge du document.

La plupart des méthodes sont basées sur un algorithme de recherche ne tenant compte que des similitudes entre questions et documents. Le modèle suivant tient compte de deux variables : la similitude question-document et l'âge du document au moment de la demande. En effet, le vieillissement de la littérature n'est pas négligeable pour deux raisons : soit l'information devient moins précise ou moins pertinente avec l'augmentation d'âge du document, soit elle est toujours valable mais l'utilisateur préfère des documents plus récents qui ont tendance à faire une synthèse des informations plus anciennes. Les utilisateurs, en jugeant certains documents pertinents définissent une probabilité sur les âges de l'ensemble des documents. On dira alors que le vieillissement est positif (respectivement négatif) lorsque l'âge moyen des documents pertinents est plus petit (respectivement plus grand) que l'âge moyen de tous les documents. Ainsi, l'optique est de considérer que le vieillissement est une entité dynamique observable, défini de façon formelle par les distributions de probabilités, qui varie d'une recherche à l'autre et d'un utilisateur à l'autre. La pertinence des documents n'étant percue que par l'utilisateur lui-même l'existence et le signe du vieillissement font partie du profil de recherche de l'utilisateur au même titre que les mots-clés.

a) Les distributions de base.

Le procédé de recherche attribue deux nombres à chaque document de la collection. Le premier est le poids 'sujet' noté x, basé sur le degré de correspondance entre les mots-clés du document et ceux de la question. Le second, appelé poids 'âge', noté t, est l'âge du document au moment de la question.
On prendra par la suite la convention suivante :
i attaché à une variable relie cet article à l'ensemble des documents pertinents lorsque i = 1 et à l'ensemble des documents non pertinents lorsque i = 2.

On suppose que les variables X_i, associées aux valeurs de x, ont des fonctions de probabilité de la forme :

$$f_i(x) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp(-(x - \mu_i)^2 / 2\sigma_i^2) \qquad x \in]-\infty, +\infty[$$

$\mu_i = E(X_i)$ est l'espérance, $\sigma_i^2 = V(X_i)$ est la variance.

Pour les variables T_i associées à l'ensemble des valeurs de t on approxime les fonctions de probabilité par :

$$g_i(t) = \eta_i \exp(-\eta_i t) \qquad t \in [0,+\infty[\qquad (1)$$

Si de plus l'âge des documents que l'on veut retrouver doit être compris entre 0 et A, l'équation (1) devient, après normalisation :

$$g_i(t) = \begin{cases} (\eta_i / (1 - \exp(-\eta_i A)))\exp(-\eta_i t) & A > 0 \quad t \in [0,A] \\ 0 & t \notin [0,A] \end{cases}$$

b) Les méthodes de recherche.

Deux méthodes existent :

- La méthode du sous-ensemble

On retrouve d'abord tous les documents ayant des valeurs x dépassant un certain seuil x_c puis on extrait de cet ensemble tous les documents qui ont des valeurs de t inférieures à un certain seuil t_c.

- La méthode du poids bivariant

On attribue à chaque document un poids bivariant $z = z(x,t)$ ce qui signifie que z est une fonction de x et de t. On choisit z sous la forme linéaire $z = \lambda_1 x + \lambda_2 t$, λ_1, λ_2 étant des constantes et on retrouve les documents pour lesquels $z > z_c$.

Cette méthode parait sujette à extension et on imagine facilement une méthode générale de recherche à partir de n différents types d'attributs de documents $x_1, x_2, \ldots, x_n$ chacun mesurant ou caractérisant la pertinence du document pour une question donnée. On choisierait alors entre deux méthodes : une avec une valeur seuil et une séquence de sous-ensembles, l'autre avec un poids multivariable de la forme $z = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n$ attribué à chaque document et une valeur seuil z_c. De nombreuses variables peuvent être utilisées : le poids linguistique, l'âge de document, sa langue, sa provenance, sa forme littéraire ...

9) Les modèles basés sur la théorie des sous-ensembles flous.

Les modèles mathématiques décrits jusqu'à présent étaient basés sur une logique à deux valeurs : un mot-clé appartient ou n'appartient pas à la description d'un document, un document est soit pertinent soit non pertinent pour une question donnée, etc... En termes ensemblistes, ceci se traduit par le fait que la fonction caractéristique d'un ensemble ne peut prendre que les deux valeurs 0 ou 1. La théorie des sous-ensembles flous, introduite par L. A. Zadeh, est une généralisation de la théorie ensembliste traditionnelle : la fonction caractéristique est remplacée par la fonction d'appartenance qui est une fonction continue, prenant toutes les valeurs comprises entre 0 et 1. Cette théorie est donc basée sur une logique multivaluée avec une continuité de valeurs dans l'intervalle [0,1]. Une définition plus précise d'un sous-ensemble flou s'énonce de la façon suivante :

> Soit $X = \{x\}$ un ensemble.
> Un sous-ensemble flou A de X est l'ensemble des paires ordonnées $(x, \mu_A(x))$, où $x \in X$ et où $\mu_A(x)$ est le degré d'appartenance de x à A, μ_A étant la fonction d'appartenance ($\mu_A : X \longrightarrow [0,1]$).

Formellement :

$$A = \{(x, \mu_A(x)) \ / \ x \in X\}$$

On peut alors définir un système de recherche d'information comme étant un quadruplet :

$$I = \langle D, Q, T, \Psi \rangle$$

où D est un ensemble de documents de cardinal n
Q est un ensemble de questions de cardinal m
Ψ est une application de la forme :

$$\Psi : Q \longrightarrow 2^D$$

Pour une question $q \in Q$, la réponse du système sera l'ensemble $A = \Psi(q) \subset D$.
On définit alors une relation binaire floue R à partir des triplets ordonnés $\langle x, t, \mu_R(x,t) \rangle$ où $x \in D \cup Q$, $t \in T$, $\mu_R(x,t)$ étant une fonction allant de l'ensemble des paires ordonnées (x,t) dans l'intervalle [0,1] qui définit le degré d'importance du terme $t \in T$ dans le modèle de recherche de x, $x \in D \cup Q$.

$$R = \left\{ \langle x,t,\mu_R(x,t)\rangle \ / \ x \in D \cup Q,\ t \in T \right\}$$

Nous pouvons alors définir les sous-ensembles flous correspondants à des documents particuliers et aux recherches $x \in D \cup Q$, noté R_x :

$$R_x = \left\{ \langle t,\mu_{R_x}(t) = \mu_R(x,t)\rangle \ / \ t \in T \right\}$$

Un ensemble flou de niveau α sera un ensemble flou de la forme :

$$R_{x(\alpha)} = \left\{ \langle t,\mu_{R_{x(\alpha)}}(t) = \mu_{R_x}(t)\rangle \ / \ t \in T,\ \mu_{R_x}(t) \geq \alpha \right\}$$

Pour $\alpha = 0$, nous avons $R_{x(\alpha)} = R_x$.

a) <u>Modèle de recherche utilisant la notion de thésaurus flou.</u>

L'expression thésaurus flou désigne un ensemble de termes T qui vérifie les conditions suivantes :

- Il existe sur cet ensemble T une relation de similitude S qui est une relation floue déterminée par la fonction d'appartenance $\mu_S : T*T \longrightarrow [0,1]$ et qui vérifie :

$$\forall\ t \in T \qquad \mu_S(t,t) = 1$$

$$\forall\ t,\ t' \in T \qquad \mu_S(t,t') = \mu_S(t',t)$$

$$\forall\ t,\ t',\ t'' \in T \qquad \mu_S(t,t'') \geq \max_{t'} \left\{ \min \left[\mu_S(t,t'),\ \mu_S(t',t'') \right] \right\}$$

- Il existe un ensemble non vide $T_d \subset T$ appelé ensemble de descripteurs élémentaires et une relation λ-synonyme $S_\lambda \subset T*T$ vérifiant :

$$\forall\ t,\ t' \in T_d \qquad t,t' \in S_\lambda \iff \mu_S(t,t') \geq \lambda$$

$$\forall\ t,\ t' \ T_d \qquad t \neq t' \quad \Rightarrow \quad \langle t,t'\rangle \notin S_\lambda$$

$$\forall\ t \in T - T_d \quad \exists\ t' \in T_d \text{ tel que } \langle t,t'\rangle \in S_\lambda$$

S est par définition une relation d'équivalence.

- Il existe sur l'ensemble des descripteurs élémentaires T_d une relation $G_{T_d}(\beta)$ de généralisation des descripteurs de

niveau β définie par :

$$t\ G_{T_d}(\beta)\ t'\ ,\ t \neq t'\ ,\ \beta < \lambda \iff t \text{ a une signification}$$

sémantique plus générale que t' et $\langle t,t'\rangle \in S_\beta$.

Cette relation est non réflexive, antisymétrique et transitive.

Nous allons maintenant étudier le modèle de recherche. Il est établi en plusieurs étapes :

- Déterminer l'ensemble flou $R_{x(\alpha)}$

- Eliminer de $R_{x(\ )}$ l'ensemble des paires ordonnées

$$\left\{ \langle t', \mu_{R_{x(\alpha)}}(t') \rangle \ / \langle t',t'' \rangle \in S_\lambda\ ;\ t' \in T - T_d\ ;\ t'' \in T_d\ ;\ t',t'' \in \mathcal{D}R_{x(\alpha)} \right.$$

où $\mathcal{D}R_{x(\alpha)}$ est le domaine de l'ensemble flou $R_{x(\alpha)}$ défini par

$$\mathcal{D}R_{x(\alpha)} = \left\{ t \ / \exists \mu_{R_{x(\alpha)}}(t) \langle t,\ R_{x(\alpha)}(t) \rangle \in R_{x(\alpha)} \right\}$$

Le sous-ensemble flou $R_{x(\alpha)}$ ainsi modifié est noté $R'_{x(\alpha)}$ où

$$\mu_{R'_{x(\alpha)}}(t'') = \max \left(\mu_{R_{x(\alpha)}}(t'), \mu_{R_{x(\alpha)}}(t'') \right)\ ,\ t' \in T - T_d, t'' \in T_d$$

- Eliminer de l'ensemble flou $R'_{x(\alpha)}$ l'ensemble des paires ordonnées

$$\left\{ \langle t, \mu_{R'_{x(\alpha)}}(t') \rangle / t'\ G_{T_d}(\beta)\ t''\ ;\ t',t''\ \ \mathcal{D}R'_{x(\alpha)} \right\}$$

L'ensemble flou ainsi modifié est noté $R''_{x(\alpha)}$. L'ensemble flou $R''_{x(\alpha)}$ correspondant à $x \in D \cup P$ est son modèle de recherche.

Sur l'ensemble des documents et demandes d'information, nous définissons la relation $G_0(\alpha)$ de modèle par :

$$\forall x,y \in D \cup P \quad y\ G_0(\alpha)\ x \iff R''_{x(\alpha)} \subset R''_{y(\alpha)}$$

où $R''_{x(\alpha)} \subset R''_{y(\alpha)}$ signifie que l'ensemble flou $R''_{x(\alpha)}$ est un sous-ensemble de lensemble flou $R''_{y(\alpha)}$ de niveau α c'est-à-dire

$$R''_{x(\alpha)} \subset R''_{y(\alpha)} \iff \forall t \in \mathcal{D}R''_{x(\alpha)}\ \ \mu_{R''_{x(\alpha)}}(t) \leq \mu_{R''_{y(\alpha)}}(t)$$

forme q = t + Nt + (t ∧ t) + (t ∨ t)

où t est un élément de T, + est mis pour 'ou', N pour la 'négation', ∧ pour la 'réunion' et ∨ pour l'intersection'.

La fonction d'appariement Ψ est définie par les étapes suivantes :

.S_1 : Si la question est le descripteur t

$$\Psi(q,x) = \begin{cases} \mu_x(t) & \text{si } t \quad \text{SUPP}(x) \\ \max(\mu_x(t_1), \mu_x(t_2), \ldots, \mu_x(t_k) & \\ \quad \text{si } \{t_1, t_2, \ldots, t_k\} \subseteq \text{SUPP}(x) & \\ \quad \text{et } t \; R \; t_i \text{ pour } i = 1,2,\ldots,k & \\ 0 & \text{sinon} \end{cases}$$

.S_2 : Si q est une question

$$\Psi(Nq,x) = 1 - \Psi(q,x)$$

.S_3 : Si p et q sont des questions

$$\Psi(p \wedge q,x) = \min(\Psi(p,x) \; , \Psi(q,x))$$
$$\Psi(p \vee q,x) = \max(\Psi(p,x) \; , \Psi(q,x))$$

.S_4 : Si q est une question quelconque alors Ψ(q,x) est obtenue par des calculs basés sur les étapes S_1, S_2, S_3.

Les règles définies ci-dessus ne sont pas uniques : d'autres fonctions peuvent être déterminées.

Pour la question q, la réponse du système de recherche est un sous-ensemble flou de X dont la fonction d'appartenance est donnée par :

$$\mu_{f(q)}(x) = \Psi(q,x) \qquad \text{pour } x \in X$$

Autrement dit :

$$f(q) = \{(x, \Psi(q,x))\}$$

$\mu_{f(q)}(x)$, degré d'appartenance de l'enregistrement x dans la réponse

De par sa définition $G_0(\alpha)$ est réflexive, antisymétrique et transitive. Elle induit donc un ordre sur l'ensemble des documents et demandes d'information.

La réponse du système de recherche d'information à la demande d'information $p \in P$ est l'ensemble des documents $A = \Psi(p) \subset D$ de la forme :

$$\Psi(p) = \{ d \,/ d\, G_0(\alpha)\, p \ , \ d \in D \}$$

Ainsi, la réponse du système de recherche à la demande d'information $p \in P$ est l'ensemble des documents $d \in D$ qui sont en G_0 relation avec la demande d'information donnée d.

b) Modèle de recherche en arbre.

De même que précedemment, nous considérons qu'un système de recherche est un quadruplet $\langle D, G, T, \Psi \rangle$.

Soit T un ensemble fini de descripteurs (card T = m) et R une relation d'ordre, appelée relation de généralisation, sur cet ensemble T.

Si t_i R t_j, on dit que le descripteur t_i est plus général que le descripteur t_j ou que t_j est plus spécifique que t_i.

L'ensemble T des descripteurs avec sa relation de généralisation G peut être représenté par un graphe basé sur les contraintes suivantes :

- Les descripteurs de T sont les noeuds (ou sommets).
- Chaque branche (t_i, t_j) de G correspond à une relation t_i R t_j. La branche (t_i, t_j) a pour origine le noeud t_i et pour fin le noeud t_j.
- Par souci de simplicité, les branches générées par transitivité ne sont pas représentées (exemple : on ne représente pas t_i R t_j lorsqu'on a déjà t_i R t_k et t_k R t_j).

Nous allons maintenant définir la notion de support d'un sous-ensemble flou, notion qui est utile par la suite.
Le support d'un ensemble flou A de X est un sous-ensemble non flou noté SUPP(A) et défini par :

$$SUPP(A) = \{ x : \mu_A(x) > 0 \}$$

Dans ce modèle nous supposons que les questions sont des combinaisons booléennes de descripteurs c'est-à-dire qu'elles sont de la

trouvée f(q) peut être considéré comme le degré de pertinence du contenu de l'enregistrement x pour la question q. Si pour une question q $\mu_{f(q)}(x_i) > \mu_{f(q)}(x_j)$ on peut dire que l'enregistrement x_i est plus pertinent que l'enregistrement x_j pour la question q. Les valeurs de pertinence définissent ainsi un ordre sur les enregistrements de SUPP(f(q)).

Conclusion

Conclusion

Le nombre important de modèles existant ne permet pas de les présenter tous. Ainsi, tous les modèles basés sur la classification des documents (hypothèse de mise en ordre des documents, hypothèse de rangement des probabilités,....), ceux basés sur l'indexation automatique ainsi que la recherche interactive n'ont pas été traités volontairement. De même, tous les problèmes de rendement ou d'efficacité d'un modèle ainsi que leur mesure ont été ignorés.

Le problème de recherche de l'information est vaste. De nombreuses recherches ont déjà été menées et de nombreuses sont en cours. La plupart des méthodes utilisées actuellement ont une base commune : l'indexation des documents par une liste de mots-clés. Bien que cette théorie ne s'avère pas satisfaisante sur tous les points elle semble pourtant la plus adaptée au problème de recherche de l'information. Le modèle idéal n'est pas encore découvert et la recherche documentaire a encore un grand avenir devant elle.

B I B L I O G R A P H I E

Ouvrage :

- SALTON (Gérard) : Automatic information organization and retrieval . - New-york : Mc Graw-Hill Book Company, 1968 . - (Computer science series)

Articles de revues :

- SACHS (Wladimir M.) : An approach to associative retrieval through the theory of fuzzy sets
in : Journal of the American Society for Information Science, Vol 27 N°2, Mars-Avril 1976, pp 85-87

- CLEVELAND (Donald B.) : An n-dimensionnal retrieval model
in : Journal of the American Society for Information Science, Vol 27 N° 5/6, Septembre-Octobre 1976, pp 342-347

- TAHANI (Valiollah) : A fuzzy model of document retrieval system
in : Information Processing and Management, Vol 12, 1976, pp 177-187

- RADECKI (Tadeusz) : Mathematical model of information retrieval system based on the concept of fuzzy thesaurus
in : Information Processing and Management, Vol 12, 1976, pp 313-318

- RADECKI (Tadeusz) : New approach to the problem of information system effectiveness evaluation
in : Information Processing and Management, Vol 12, 1976, pp 319-326

- RADECKI (Tadeusz) : Mathematical model of time-effective information system based on the theory of fuzzy sets
in : Information Processing and Management, Vol 13, 1977, pp 109-116

- HEINE (M.H.) : Incorporation of the age of a document into the retrieval process
in : Information Processing and Management, Vol 13, 1977, pp 35-47

- ROBERTSON (S.E.) : Progress in documentation : Theories and models in information retrieval
in : Journal of Documentation, Vol 33, N°2, Juin 1977, pp 126-148

- SALTON (Gérard) : Mathematics and information retrieval
in : Journal of Documentation, Vol 35, N°1, Mars 1979, pp 1-29