enssib

**ÉCOLE NATIONALE SUPÉRIEURE DES SCIENCES DE L'INFORMATION ET DES BIBLIOTHÉQUE**

DEA SCIENCE DE L'INFORMATION ET DE LA COMMUNICATION

OPTION 3

SYSTÈME D'INFORMATION DOCUMENTAIRE

NOTE DE SYNTHÈSE SUR

# LE FILTRAGE DES INFORMATIONS


Réalisée par :

***Ramzi ABBES***

Sous la direction de :

***M. Mohamed HASSOUN***


Mars 1999

# TABLE DES MATIÈRES


**INTRODUCTION** ..... 4

I/ LES CARACTÉRISTIQUES DU FILTRAGE DES INFORMATIONS : ..... 5

II/ DESCRIPTION GÉNÉRALE D'UN MODÈLE DE FILTRAGE : ..... 6

**PREMIÈRE PARTIE : LE FILTRAGE LINGUISTIQUE** ..... 7

I/ INTRODUCTION ..... 8

II/ FILTRAGE MORPHOSYNTAXIQUE ..... 8

*II/1 INTRODUCTION* ..... *8*

*II/2 LES GRAMMAIRES HORS CONTEXTE* ..... *8*

II/2/1 INTRODUCTION ..... 8

II/2/2 DÉFINITIONS ET PROPRIÉTÉS ..... 8

II/2/3 LE FILTRE ..... 11

A) DESCRIPTION DU FONCTIONNEMENT DU SYSTÈME ..... 11

B) LE FILTRE SIMPLE ..... 11

C) LE FILTRE COMBINÉ ..... 12

III/FILTRAGE SYNTAXIQUE ..... 12

*III/1 INTRODUCTION* ..... *12*

*III/2/ LES INFORMATIONS LATENTES* ..... *13*

III/2/1 PROBLÉMATIQUE : ..... 13

III/2/2 LE SYSTÈME GLEAN ..... 13

III/2/3 LATENT SEMANTIC INDEXING (LSI) ..... 14

IV/ LE FILTRAGE SEMANTIQUE ..... 15

*IV/1 INTRODUCTION* ..... *15*

*IV/2/ LES EXPRESSIONS CAUSALES* ..... *15*

IV/2/1 L'ORGANISATION SÉMANTIQUE DES INDICATEURS DE LA NOTION DE CAUSALITÉ ..... 15

IV/2/2 LES VERBES INDICATEURS D'UNE CAUSALITÉ PRÉCISANT L'EFFET PRODUIT ..... 16

IV/3/3 LE SYSTÈME COATIS ..... 16

A) PRÉSENTATION DU MODÈLE : ..... 16

B) EXPLOITATION DES RÉSULTATS ..... 17

*IV/3 LA DÉFINITION* ..... *17*

IV/3/1 PRÉSENTATION DE LA DÉFINITION ..... 17

IV/3/2 LE FILTRAGE À L'AIDE DE LA DÉFINITION ..... 18

*V/4 FILTRAGE AUTOMATIQUE DE PHRASES TEMPORELLES D'UN TEXTE* ..... *18*

V/4/1 INTRODUCTION ..... 18

V/4/2 DESCRIPTIONS DU FONCTIONNEMENT DU SYSTÈME ..... 19

**DEUXIÈME PARTIE : LE FILTRAGE SUR LE WEB** ..... 21

I/ INTRODUCTION ..... 22

*II/1/ DESCRIPTION GÉNÉRALE DUN SYSTÈME DE FILTRAGE SUR LE WEB* ..... *22*

*I/2 LES ÉLÉMENTS DU MODÈLE DE FILTRAGE* ..... *23*

1) DISPOSITION DE L'UTILISATEUR ..... 23
2)L'ÉCHELLE DU TEMPS ..... 23
3)LA LIVRAISON DE L'INFORMATION ..... 24
4) LE CONTENUE DE L'INFORMATION ..... 24
II/ LES MOTS CLÉS ..... 24
*II/1 PROBLÈMES AVEC LES MOTS CLÉS* ..... *24*
*II/2 UTILISATION DES MOTS CLÉ* ..... *25*
*II/3 INCONVÉNIENTS* ..... *25*
III/ FILTRAGE SELON LE PROFIL UTILISATEUR ..... 25
*III/1 INTRODUCTION* ..... *25*
*III/2 LE MODÈLE UTILISATEURS* ..... *26*
*III/3 PROBLÈME DE DÉVELOPPEMENT DE MODÈLE UTILISATEURS* ..... *26*
IV/ LES STÉRÉOTYPES ..... 27
*IV/1 INTRODUCTION* ..... *27*
*IV/2 LA MODÉLISATION DES UTILISATEURS PAR LES STÉRÉOTYPES* ..... *27*
*IV/3 INTÉGRATION DES STÉRÉOTYPES DANS LE MODÈLE DE FILTRAGE D'INFORMATION* ..... *28*
VI/3/1 DESCRIPTION DU FONCTIONNEMENT DU MODÈLE ..... 28
V/ LE MODÈLE DE FILTRAGE PROBABILISTE ..... 29
*V/1 DÉFINITION* ..... *29*
*V/2 PRÉSENTATION DU MODÈLE* ..... *29*
*V/3 LES DIFFICULTÉS DE LA MISE EN PLACE* ..... *30*
VI/ UTILISATION DES MÉTHODES DE L'INTELLIGENCE ARTIFICIELLE ..... 30
*VI/1 LES AGENTS INTELLIGENTS* ..... *30*
VI/1/1 DESCRIPTION DU FONCTIONNEMENT DU SYSTÈME ..... 30
*VI/2 MÉTHODE D'APPRENTISSAGE AUTOMATIQUE :LE SYSTÈME INFOS* ..... *31*
VI/2/1 INTRODUCTION ..... 31
VI/2/2 DESCRIPTIONS DU FONCTIONNEMENT DU SYSTÈME ..... 31
**CONCLUSION** ..... **33**
**BIBLIOGRAPHIE** ..... **35**

# INTRODUCTION

Le développement récent des ordinateurs et des réseaux informatiques qui relient un grand nombre de systèmes, a facilité la communication et l'accès aux informations, mais la quantité des informations échangées et produites n'arrête pas de croître, ainsi les utilisateurs du système se trouvent « inondés » d'informations et n'arrivent plus à les gérer, ce qui rend la procédure de recherche d'information, une tâche pénible qui prend beaucoup de temps et d'effort [Morita 94].

Afin de faire face à ce problème et de réduire cette surcharge d'information, Belkin et Croft ont introduit le filtrage des informations. Leur article [Belkin and Croft 1992], a constitué un point de départ et une référence incontournable pour toutes les recherches faisant appel à la notion du filtrage des informations qui l'on suivit.

En fait, le filtrage des informations n'est pas un nouveau concept et il n'est pas limité aux documents électroniques. Nous achetons uniquement certains magasines, car les autres contiennent des informations qui ne nous intéresse pas directement, et dans les magasines que nous achetons nous ne lisons que certains articles. Quand nous lisons un texte standard, nous filtrons les informations que nous devons retenir [Foltz and Dumais1992].

Au fur et à mesure de l'avancement de notre recherche (interrogation des bases de données, recherche sur le web... ), nous nous sommes rendu compte que les travaux sur le filtrage des informations ne sont pas nombreux, du fait que les chercheurs ne se sont intéressés à ce domaine que récemment.

Nous avons remarqué, d'une part, que les concepts du filtrage des informations ne sont pas bien définis, d'autre part, que les limites entre le filtrage, la recherche, le routage et l'extraction des informations ne sont pas très claire chez les auteurs.

Toutefois nous avons pu dégager deux grandes classes utilisées généralement par les auteurs, à savoir, le filtrage linguistique et le filtrage sur le Web.

Avant de présenter ces travaux, nous allons montrer, dans les deux paragraphes qui suivent, les caractéristiques des filtres et le modèle général de filtrage tel qu'ils sont définis par Belkin et Croft [Belkin and Croft 1992].

# I/ LES CARACTERISTIQUES DU FILTRAGE DES INFORMATIONS :

Le filtrage des informations est un terme utilisé pour décrire plusieurs processus nécessitant l'acheminement de l'information d'une source bien définie jusqu'à l'utilisateur qui ont à besoin.

Pour bien définir le filtrage des informations Belkin et Croft [Belkin and Croft 1992] ont présenté les différentes caractéristiques de ce processus :

- Un système de filtrage des informations est un système d'information, conçue pour les données non structurées ou semi-structurées. Ce qui n'est pas le cas des bases de données classique qui nécessitent des données structurées et bien définies.
- Les systèmes de filtrage d'informations traitent les informations textuelles, qui peuvent inclure aussi, le son, l'image et la vidéo.

- Les systèmes de filtrage nécessitent un grand volume de données (des Giga bits d'informations).
- Les applications de filtrage nécessitent un flot de données entrant, transmis par des ressources distantes. Le filtrage, est aussi utilisé, pour décrire le processus d'accéder et de rechercher des informations dans des ressources distantes.
- Le filtrage se base sur la description des préférences d'un usager ou d'un groupe d'usager, appelés aussi profile (le profile représente les besoins à longs termes).
- Le filtrage, consiste aussi à éliminer certaines données du flot des données provenant de sources distantes (non pas l'extraction des données appropriées ).

# II/ DESCRIPTION GENERALE D'UN MODELE DE FILTRAGE :

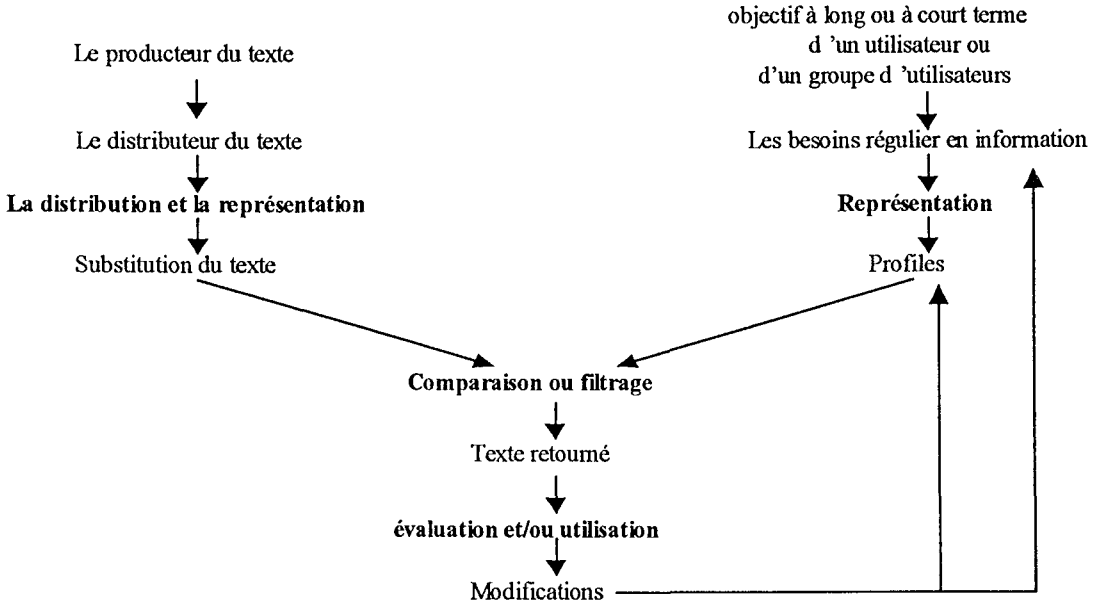



Fig.1 Modèle générale de filtrage d'informations

Ce modèle de filtrage présenté par Belkin et Croft [Belkin and Croft 1992] considère, d'une part, les utilisateurs du système de filtrage (ils sont caractérisés par des besoins relativement stables à moyen et long terme), d'autre part, le circuit de production et de distribution des textes.

L'évaluation des documents se fait en confrontant les documents retournés, par les outils de recherche, et la requête de l'utilisateur (les besoins).

Pour la construction de ce modèle, Belkin et Croft [Belkin and Croft 1992] se sont inspirés de leur modèle de recherche des informations.

# PREMIERE PARTIE
# LE FILTRAGE LINGUISTIQUE

# I/ INTRODUCTION

Le filtrage linguistique est la mise du texte sous une forme particulière en éliminant ou en extrayant un certain nombre d'informations, en ce basant sur des critères linguistiques tels que l'analyse morphosyntaxique (les grammaires), l'analyse syntaxique (les informations latentes) et l'analyse sémantique (l'expression définitoire, les relations de causalité ou les "événements" temporels...).

# II/ FILTRAGE MORPHOSYNTAXIQUE

## II/1 INTRODUCTION

Le filtrage des textes structurés prend de plus en plus d'importance avec l'ampleur d'utilisation que prennent les documents du type SGML. Dans les systèmes de filtrage des informations, une bonne spécification des textes est requise pour l'automatisation du filtrage. C'est pour cela que Kuikka et Salminen [Kuikka and Salminen 1997] ont eu recours à l'analyse syntaxique en se basant sur les grammaires.

En effet, dans les grammaires de modélisation de base de données, une grammaire formelle est considérée comme un schème de données textuelles. Elle représente la structure hiérarchique, l'ordre, l'optimisation, les alternatives et les structures récursives.

## II/2 LES GRAMMAIRES HORS CONTEXTE

### II/2/1 INTRODUCTION

Les grammaires hors contextes sont très utilisées pour représenter les textes structurés. Les documents définis par SGML sont l'exemple typique d'utilisation des ces grammaires.

### II/2/2 DEFINITIONS ET PROPRIETES

Une grammaire hors contexte est une grammaire de la forme G(A, N, P, s) où :

A : un alphabet (ensemble de symboles terminaux)
N : un ensemble de mots appelés "symboles non terminaux" pour la représentation des éléments structurés.
P : ensemble de règles de production.
s : le symbole de départ.

Exemple :

Nous prenons l'exemple de Kuikka et Salminen [Kuikka and Salminen (1997)], d'une base de données des rapports publiés par différentes universités.

(1) base de donnée des rapports --> rapport
(2) rapport --> titre auteurs? éditeur catégorie+ chapitre+
(3) auteurs --> auteur+
(4) catégorie --> informatique | étude d'information |....
(5) chapitre --> titre para* section*

(6) para --> para_du_texte | graphique
(7) section --> titre para+
(8) para_du_texte --> phrase+

fig.2 Grammaire de production pour une base de données des rapports

- La prémisse de chaque règle est un symbole non terminal.

- La conclusion peut être un symbole terminal ou non terminal ou un méta symbole.

*Exemple de méta symbole* : * : 0 ou plusieurs item
+ : 1 ou plusieurs item
| : alternative
? : optimale (ou moins)

- Une production dont la prémisse est un terminal est appelée une *t-production.*

- Une grammaire hors contexte définit un langage formel qui spécifie les différents symboles pouvant être utilisés, et la manière dont ces symboles sont combinés pour construire ce langage.

- La structure hiérarchique de ce langage peut être présentée par un arbre d'analyse [Aho and Ullman 1972]. Dans cet arbre, chaque parent avec son descendant forment une règle de production.

- La racine de l'arbre d'analyse est le symbole de départ.

- Chaque nœud est un symbole terminal ou non terminal.

- Les *parties* sont les nœuds de l'arbre d'analyse annotés par des symboles non terminaux. Ainsi un nœud X contenant une *partie x* est appelé "*contenant de x*".

Le modèle est donc un arbre qui représente la structure de la grammaire d'un type de texte *t* à un niveau de détail fixe.

(a) Chapitre

(b) Chapitre
titre
para*
section*

(c)chapitre
titre
para*
|texte_du_para
| graphique
section*
titre
para*
|texte_du_para
|graphique

fig.4 trois modèles de chapitres générés à partir de la fig1

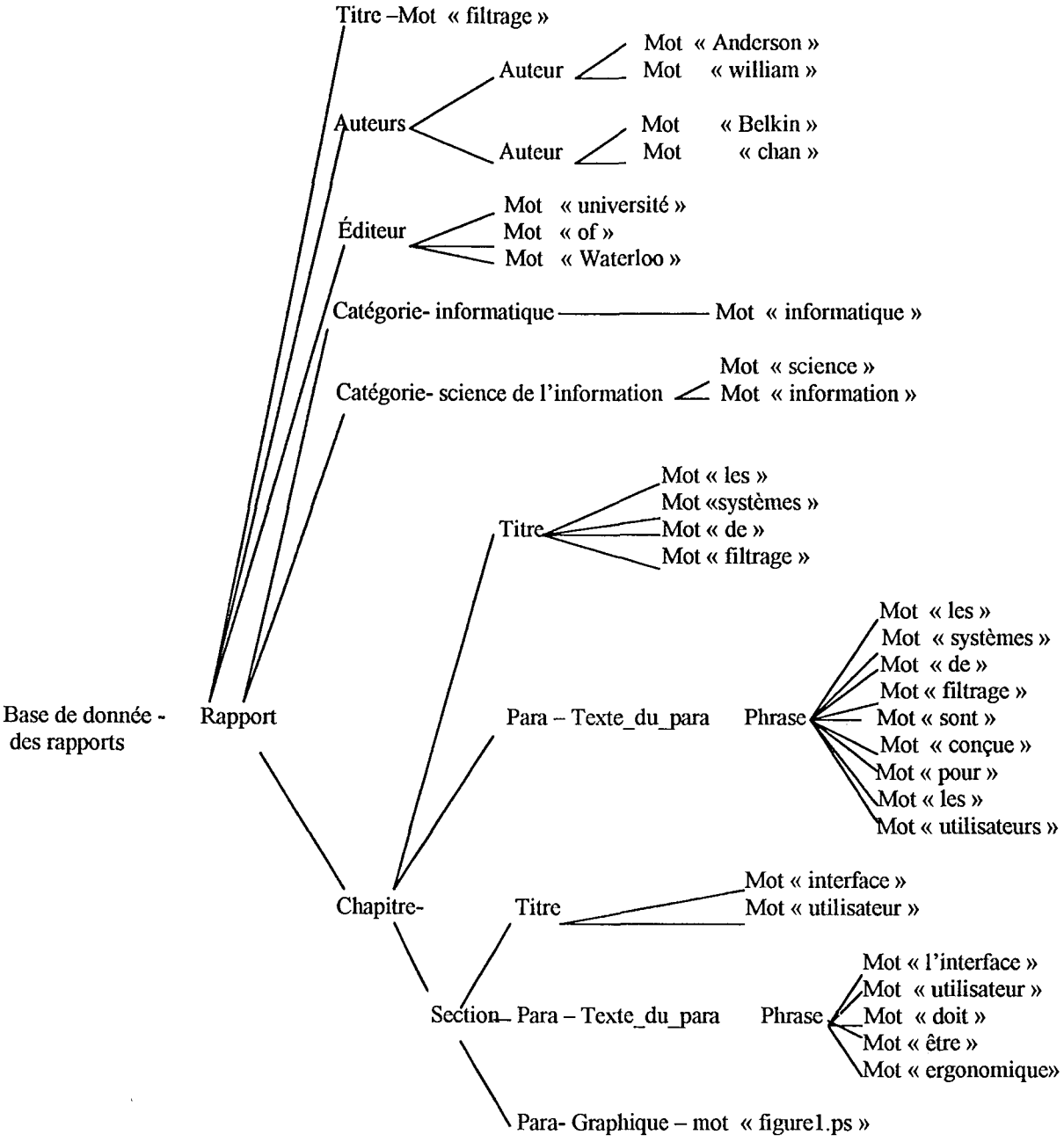



fig.3 Une partie de l'arbre d'analyse de la grammaire de la fig.2

***Le modèle avec contrainte :*** c'est un modèle selon lequel nous ajoutons une contrainte (*c*) à un nœud (*t*) tel que *t{c}* est une propriété.

Les contraintes améliorent :

- l'accès au contenu de certaines parties.
- la position des parties dans un ensemble ordonné.
- le nombre des parties qui vérifient une propriété spécifique.

Exemples:

X auteurs
auteurs+ « Smith » Qty:2.

C'est à dire la propriété {Smith} est vrai pour la partie 'auteur' dans la partie 'auteurs'.

X Chapitre
Titre
Para*
Section*qty:All
Titre
Para+ texte_du_para qty:< 10

Les chapitres dans lesquels toutes les sections ont un nombre de paragraphe inférieur à 10.

Les modèles avec contraintes sont bénéfiques pour spécifier des parties basées sur des conditions ou sur le contenu de la partie, mais ne sont pas efficaces pour définir le contexte de la partie.

Exemple : le modèle avec contraintes présenté précédemment ne permet pas de spécifier les requêtes suivantes « *les titres des articles écrits par 'Smith'* » ou bien « *les rapports dont les titres ou les chapitres contiennent le mot 'SGML'* », ainsi la capacité de ces modèles est limitée.

## II/2/3 LE FILTRE

### *a) description du fonctionnement du système*

Le système présenté par Kuikka et Salminen [Kuikka and Salminen 1997] utilise un modèle à deux dimensions. En premier lieu, il part de la grammaire pour représenter la structure des éléments du texte (selon un degré de détail déterminé), d'une manière hiérarchique. Ce modèle est, en deuxième lieu, augmenté par des contraintes et des annotations.

### *b/ Le filtre simple*

Soit G une grammaire, D un ensemble de symboles non spécifiés dans G (ces symboles vont être utilisés pour annoter les types de contraintes construits dans les modèles). Une annotation spécifie un ensemble de parties.

Exemple de filtre simple
Chapitre
Titre 'filtre'
Para*
Section*1..5
titre 'combiné'
Para* texte_du_para qty:>10.

Ce filtre retourne les 5 premières sections d'un chapitre tel que le titre du chapitre contient le mot "filtre", le titre d'une section contient le mot 'combiné', et la section en question contient plus de 10 paragraphes.

Cette requête peut être représentée à l'aide du filtre simple car les titres de la section et du chapitre appartiennent à des parties différentes, ces parties sont en plus contenues l'une dans l'autre.

Mais avec ce filtre simple nous ne pouvons pas représenter des requêtes du type, «*le titre des publications et des rapports dans le premier auteur est 'Aho' et le second auteur est 'Ullman', et il n'y pas d'autres auteurs avec eux* », car les noms des auteurs se situe au même niveau hiérarchique. C'est pourquoi nous faisons appel au filtre combiné.

*c/ Le filtre combiné*

Reprenons l'exemple précédent
« *Le titre des publications et des rapports dans le premier auteur est 'Aho' et le second auteur est 'Ullman', et il n'y pas d'autres auteurs avec eux* ».

```
                                    Auteurs
Aho_as_first..... auteur+           'aho'&1
                                    Auteurs
Ullman_as_second.....auteur+        'Ullman'&2
                                    Report          Aho_as_first& Aho_as_second
Aho_Ullman_title.....Titre
                                        auteurs?
                                            auteurs+qty:2
Aho_Ullman_publisher.....éditeur
                                        Catégorie+
                                        Chapitre+
```

Le premier filtre s'intéresse à la partie auteur qui contient le mot 'Aho'. Il est la première composante d'une partie 'auteurs' appelée Aho_as_first.

Le deuxième filtre s'intéresse à la partie auteur qui contient le mot 'Ullman'. Il est la deuxième composante d'une partie 'auteurs' appelée Ullman_as_second.

Enfin le troisième filtre spécifie le titre et l'éditeur des rapports qui contiennent une partie des 2 nouveaux types, et dont le nombre d'auteurs est deux.

Avec ce filtre combiné nous avons pu élaborer un filtre d'information efficace en nous basant sur les grammaires hors contextes de l'analyse morphosyntaxique. Dans ce qui suit, nous allons présenter une méthode plus sophistiquée avec plus de précision, qui utilise les informations syntaxiques obtenues à partir de l'étiquetage des textes.

# III/FILTRAGE SYNTAXIQUE

## III/1 INTRODUCTION

La plupart des documents retournés par les outils de recherche d'information classique ne sont pas exploitable par l'utilisateur qui se trouve noyé dans cette masse de documents. Selon Chandrasekar et Srinivas [Chandrasekar and Srinivas 1998], il serait plus intéressant d'élaborer des outils de filtrage faisant appel à des techniques de l'analyse syntaxique dans le

but d'identifier les documents pertinents pour l'usager. En effet, Chandrasekar et Srinivas [Chandrasekar and Srinivas 1998] souligne que les textes cohérents contiennent des informations « latentes » significatives, comme la structure syntaxique et les schèmes du langage, qui peuvent être utilisées pour améliorer la performance des systèmes de recherche et de filtrage d'information.

## III/2/ LES INFORMATIONS LATENTES

### III/2/1 PROBLEMATIQUE :

Nous allons illustrer la problématique par un exemple simple présenté par Chandrasekar et Srinivas [Chandrasekar and Srinivas 1998].

Supposons que nous sommes entrain de chercher des informations sur les rendez-vous officieux aux États Unis "*official appointements*". Nous devons donc chercher des phrases contenant le mot "*appointement*", le résultat de la recherche est le suivant

a- *The Philadelphia Flyers will meet today to appoint a new manager...*
b- *The president appoints judges of the Supreme Court*
c- *Fed Vice Chairman Alan Blinder, a clinon appointee, has a rate cut necessary to keep the economy from slowing too sharply. (NYT)*

Dans ces trois phrases, la seule qui se révèle pertinente est la première alors que les deux autres ne nous intéressent pas. La question qui se pose est : Comment éliminer les phrases non pertinentes ?

L'idée est de trouver des indicateurs syntaxiques qui permettent d'éliminer (b) et (c).

### III/2/2 LE SYSTEME GLEAN

Chandrasekar et Srinivas [Chandrasekar and Srinivas 1998] définissent le filtrage comme étant un accessoire pour la recherche standard. Pour eux, le filtrage consiste à fournir un ensemble d'information supplémentaire (recueillies à partir de la requête de l'utilisateur), afin d'éliminer certains documents qui ne répondent pas directement aux besoins de l'utilisateur.

Principe de fonctionnement de GLEAN

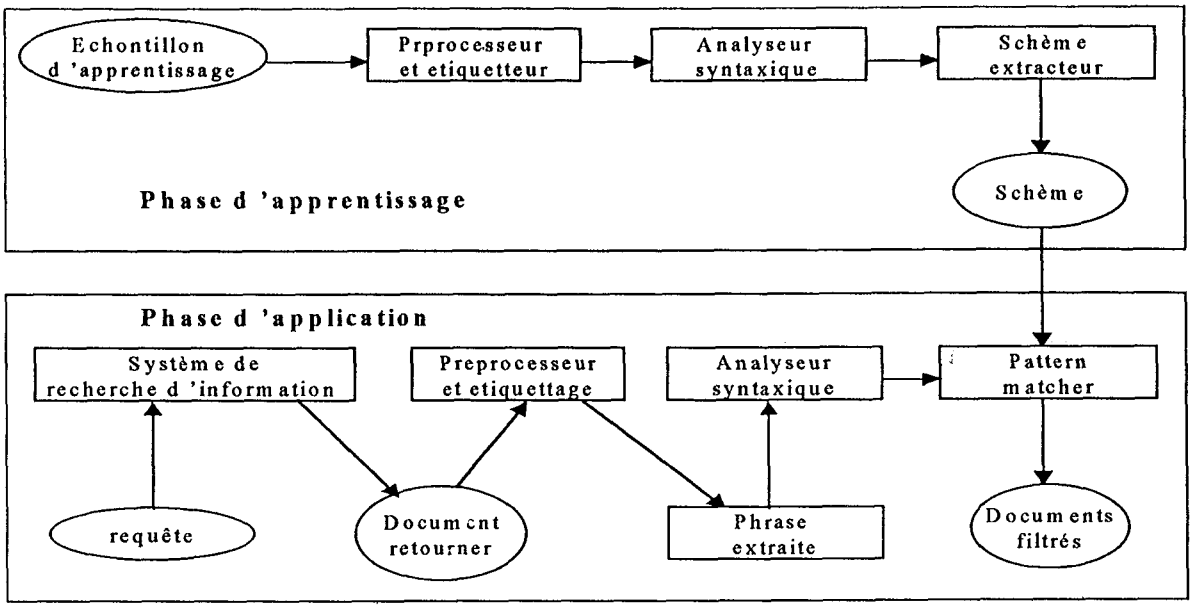



Fig.5 schéma de fonctionnement de GLEAN

L'approche générale à ce problème consiste en deux phases, une phase d'apprentissage et une phase d'application (fig.5) [Chandrasekar and Srinivas 1998].

Dans la phase d'apprentissage, un ensemble de phrases pertinentes (pour le domaine d'intérêt) est sélectionné manuellement à partir d'un corpus de nouveaux textes appelés *échantillon d'apprentissage*.

Dans la phase d'analyse syntaxique, une description syntaxique est associée aux mots des phrases d'apprentissage. Ces descripteurs vont permettre l'identification des régularités contextuelles. Cette étape s'achemine par l'obtention d'un ensemble de schèmes de pertinence du domaine d'intérêt. Ces schèmes peuvent être utilisés comme un filtre dans la recherche.

## III/2/3 LATENT SEMANTIC INDEXING (LSI).

Foltz et Dumais [ Foltz and Dumais 1992], soulignent qu'il existe des informations latentes dans les schèmes d'utilisation des mots de chaque document. Cette structure latente peut être estimée à l'aide de certaines techniques statistiques.

La description des termes, des documents et des requêtes des utilisateurs en se basant sur les structures latentes ( au lieu des surface-level word_chois) est utilisée pour représenter et retrouver l'information.

Le LSI [Deerwester et All 1990] utilise le « singular_value_decomposition (SVD) ». Le SVD considère une « large word by document matrix » et la décompose en un ensemble de *k* facteurs orthogonaux, vecteurs, (à partir des quelles, nous pouvons reconstruire la matrice originale avec des combinaisons linéaires), Par exemple : si deux termes sont utilisés dans des contextes similaires ils vont avoir des vecteurs similaires dans la représentation du LSI.

Avantages de la LSI :

- LSI permet de relier des termes même s'il n'existe pas de lien visible entre ces 2 termes.
- la requête peut retourner des documents même s'ils n'ont pas des mots clés en commun.

Le filtrage des informations en faisant appel à l'analyse syntaxique a donné de meilleurs résultats que ceux qui se sont limités à la morphologie (morphosyntaxe). Désclé affirme que des meilleurs résultats peuvent être obtenus en mettant en place des outils de filtrage intégrant des informations sémantique des textes. Dans ce qui suit, nous allons présenter des méthodes sémantiques de filtrage d'information.

# IV/ LE FILTRAGE SEMANTIQUE

## IV/1 INTRODUCTION

Avant de parler d'analyse sémantique fiable et efficace d'un texte, un grand nombre de problèmes d'ordre syntaxique doivent être résolu [Chandrasekar AND Srinivas (1998)]. Afin d'éviter ces problèmes, certain chercheurs ont proposé des méthodes intégrant des notions sémantiques telles que la causalité [Garcia 1998], la définition [Cartier 1998] [Rebeyrolle and Pery-Woodly 1998] ou l'expression temporelle [Faiz 1998].

## IV/2/ LES EXPRESSIONS CAUSALES

### IV/2/1 L'ORGANISATION SEMANTIQUE DES INDICATEURS DE LA NOTION DE CAUSALITE

Les indicateurs linguistiques, exploités par les systèmes informatiques sont classés dans un modèle sémantique qui organise le lexique verbal de la notion de causalité en français[Garcia 1998].

Ce modèle rend compte de 25 relations causales spécifiques (ou causalités), par exemple */crée/*, */empêcher/*, */faciliter/* ou */pousser_à/*, dont la signification est décrite par des schémas sémentico_cognitifs.

Le modèle linguistique comprend

* La description des 25 valeurs sémantiques de causalités.
* Les indicateurs de la langue (les verbes) qui véhiculent ces notions.
* L'organisation sémantique des relations les unes par rapport aux autres.

Exemple

*Le développement des réseaux amont est engendré par l'accroissement des charges*

Le verbe *engendrer* est indicateur d'une relation causale spécifique qui précise l'effet produit : la causalité /créer/.

Interprétation causale de l'énoncé
[*accroissements des charges*] */crées/* [*développement des réseaux amont*]
ou bien
[*accroissements des charges*] */causer/* {CREATION_DE[*développement des réseaux amont*]}

La décomposition de la signification du verbe *engendrer* ( a l'aide de /causer/ et de CREATION_DE) permet de construire, dans le cadre de la grammaire applicative et cognitive [desclés 1990], un schéma sémantico-cognitif.

## IV/2/2 LES VERBES INDICATEURS D'UNE CAUSALITE PRECISANT L'EFFET PRODUIT

La langue exprime l'effet d'une causalité directement par l'expression d'une situation ou à l'aide d'une construction sémantique (basée sur la description d'une situation). Nous appliquons sur la situation un opérateur de « *modalité d'action* » ou un « *opérateur d'influence* ».

***a)Opérateur de modalité d'action :*** appliquée à une situation, il ne la modifie pas mais situe une partie de son déroulement. Exemple *entamer, tenter de* ou encore *finir de*.

Il existe quatre valeurs sémantiques associées à des relations causales (causalités) ce sont des relations */tenter_de/*, */commencer_à/*, */continuer_à/* et */achever_de/* .

***b/Opérateur d'influence :*** appliqué à une situation, il constitue une situation nouvelle.
Notion de *création* */pousser-à/*, */créer/*, */entretenir/*, */arrêter_d'entretenir/*.
Notion d'*annihilation* /s'opposer-à/, /annihiler/, /bloquer/, /débloquer/.
Notion de *possibilité de réalisation* /laisser-faire/, /empêcher/ et également /faciliter/ et /gêner/ où les causes n'ont pas la capacité de créer ou d'inhiber une situation mais qui peuvent la modifier.

## IV/3/3 LE SYSTEME COATIS

### *a) Présentation du modèle :*

Le système COATIS analyse des textes d'un domaine technique quelconque pour élaborer, à partir des connaissances linguistiques sur l'expression de la notion de causalité en français, une structuration des connaissances causales repérées dans les documents.

Les connaissances obtenues par ce traitement et le modèle linguistique sous-jacent, permettent d'élaborer de nouvelles requêtes de filtrage des textes, selon l'activité du consultant : diagnostiques, résolution de problèmes, exploration de nouveaux documents.

Le système COATIS filtre dans un premier temps le document d'origine et propose un ensemble de relations causales, entre les situations exprimées dans le texte.

L'analyse de ces résultats constitue un nouveau point de départ pour le filtrage du texte, car la sémantique révélée par les causalités identifiées dans le document permet de construire certaines inférences qui permettent :

- De proposer des informations nouvelles( non exprimées dans le document).
- De construire un filtrage motivé, plus restreint que le filtrage exhaustif effectué par le système COATIS.

***b/ exploitation des résultats***

L'information causale recueillie par COATIS sert dans :

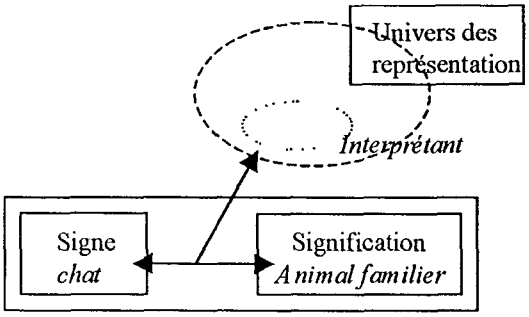
- le processus de modélisation d'un domaine [Garcia et al 1999].
- la structuration d'une terminologie [Garcia 1998].
- le filtrage automatique des textes [Garcia 1998].
- l'indexation sémantique de documents [Garcia 1998], [Gros et al. 1997] et [Assadi 1998].

## IV/3 LA DEFINITION

### IV/3/1 PRESENTATION DE LA DEFINITION

La définition [Cartier 1998] est une opération complexe, que nous pouvons, en première approximation, comprendre comme effectuant une identification entre une expression et une autre expression, cependant, la définition peut être modélisée en se basant sur certains concepts tels que :

- Le signe/ signification et concepts,




Nous attachons au signe "*chat*" une signification qui peut être "*animal familier*" par exemple.

L'ensemble signe plus signification est appelée concept.

L'utilisation du signe peut se faire en *usage* ou en *mention,* c'est ce que nous appelons le processus *d'autonymie*.

En usage, nous désignons des instances de la signification, alors qu'en mention, le concept se désigne lui-même, il est sa propre référence.

Cependant, la tradition y voie dans la définition des processus plus primaires tel que : attribution, catégorisation, spécification. La définition comporte donc plusieurs relations genre_espèces et un ou plusieurs attributs, c'est ce que nous désignons par *l'identification.*

## IV/3/2 LE FILTRAGE A L'AIDE DE LA DEFINITION

Pour son aspect bien défini et claire, la définition a été utilisée par plusieurs systèmes de filtrage tels que LEXTER [Bourigault 1994], SAFIR [Berri et all 1996].

Rebeyrolle et Pery-woodley [Rebeyrolle AND pery-woodley 1998] ont essayé de présenter une perspective d'identification et de description d'objets participant à l'organisation d'un texte, c'est à dire l'architecture textuelle. Pour ce fait, ils se sont appuyés sur la définition dans les textes scientifiques et techniques, dans une perspective de filtrage automatique dans des bases de données textuelles.

Leur analyse se fond sur un modèle de la structure des textes écrit dans lequel la signalisation des segments (objets textuels) et de leurs relations, par le biais de *la mise en forme matérielle*, est un aspect central de la mise en texte. En particulier, ce modèle met l'accent sur le texte en tant qu'objet visuel.

Dans leur modèle de représentation de l'architecture textuelle, ils se sont appuyés sur les travaux de Virbel et son équipe ([Virbel 85], [Virbel 89], [Pascual 91]) ils ont proposé une définition étendue de la notion de la *mise en forme matérielle* du texte, qui met en relation les moyens visuels (typographique et dis positionnels) et les moyens discursifs (marqueurs lexicaux et syntaxiques) pour signaler l'organisation du texte.

| Définitions | A est |
|---|---|
| A : _____ | B peut être défini |
| B : _____ | comme |
| C : _____ | on appelle C |

fig.6 Mise en forme matérielle *VS* formulation discursive

# V/4 FILTRAGE AUTOMATIQUE DE PHRASES TEMPORELLES D'UN TEXTE

## V/4/1 INTRODUCTION

Dans cette section, nous allons nous intéresser au filtrage selon les types des textes. Pour ce fait nous allons prendre l'exemple du filtrage des phrases temporelles [Faiz 1998].

Les phrases qui comportent des dates ou des délais, qui posent des imputations et des déclenchements d'ordre sont relativement importantes en droit. Dans les textes du type réglementations sociales, la date, la durée et le rapport temporel entre différents événements sont très importants.

Le principe de la méthode consiste à repérer les unités temporelles pertinentes (les indicateurs et les opérateurs temporel dans le texte analysé), ses opérateurs et indicateurs sont stockés dans une base de données qui constitue une base de connaissances pour le modèle.

## V/4/2 DESCRIPTIONS DU FONCTIONNEMENT DU SYSTEME

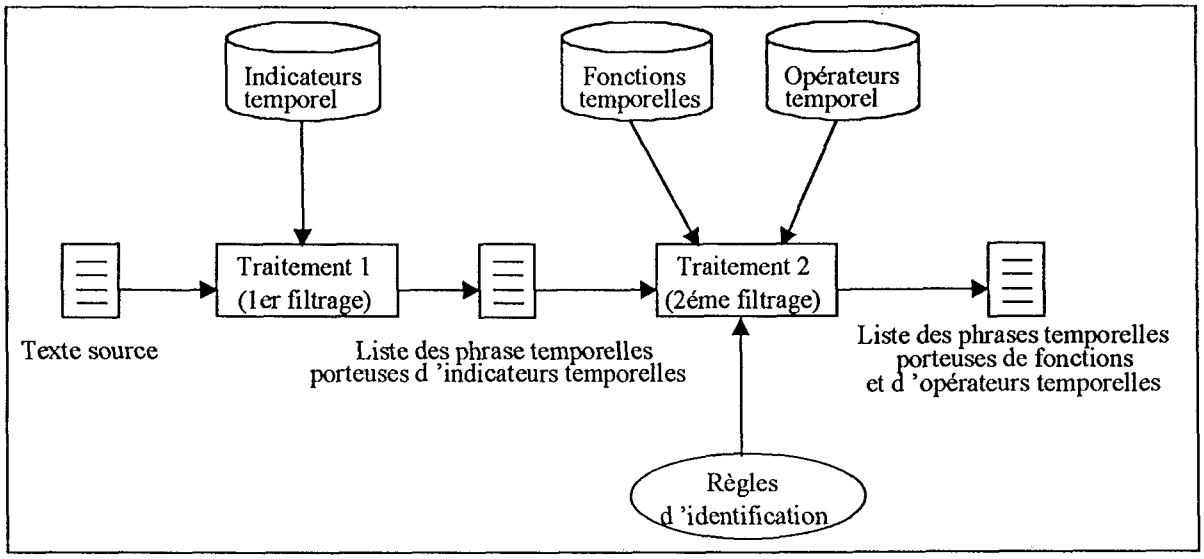



Fig7. Architecture générale du système

Dans ce système, nous identifions deux étapes principales :

1) L'extraction des phrases contenant les indicateurs temporels : jours, mois, années, dates, durées...

2) L'extraction des phrases contenant les opérateurs temporels : le système prend en entrée les résultats de la phase précédente pour identifier les phrases pertinentes.

le système utilise deux bases de connaissances :

- une base de connaissances contenant des opérateurs temporels : du, avant le, à compter du...
- une base de connaissances des fonctions temporelles : premier jour de, troisième jour du...

Exemple:

extrait du texte source

Les conditions générales d'ouverture du droit et de versement des prestations familiales, définies au chapitre 511_A, doivent être remplies.
L'allocation pour jeunes enfants revêt deux formes :
L'allocation pour jeunes enfants sans conditions de ressources est versé du premier jour du mois civil suivant le troisième mois de grossesse, au troisième mois de vie de l'enfant : Elle est versée autant de fois qu'il y d'enfants nés ou à naître.

L'allocation pour jeune enfant avec condition de ressources est versée à compter du quatrième mois de vie : une seule A.P.J.E. est servie quel que soit le nombre d'enfants de moins de trois ans. Une famille qui a déjà un enfant de trois ans et perçoit à ce titre une (A.P.J.E.) ne peut ouvrir droit à une nouvelle allocation pour un enfant de rang suivant.

"Manuel Pratique des questions de personnel" de Électricité de France, 1989.

Les phrases retenues après le premier filtrage :

P1: L'allocation pour jeunes enfants sans conditions de ressources est versé du premier **jour** du **mois** civil suivant le troisième **mois** de grossesse, au troisième mois de vie de l'enfant

P2: L'allocation pour jeune enfant avec condition de ressources est versée à compter du quatrième **mois** de vie

P3: une seule A.P.J.E. est servie quel que soit le nombre d'enfants de moins de trois **ans**

P4: Une famille qui a déjà un enfant de trois **ans** et perçoit à ce titre une (A.P.J.E.) ne peut ouvrir droit à une nouvelle allocation pour un enfant de rang suivant.

Les phrases retenues après le premier filtrage :

P1: L'allocation pour jeunes enfants sans conditions de ressources est versé **du premier jour du mois civil suivant le troisième mois de** grossesse, au troisième mois de vie de l'enfant

P2: L'allocation pour jeune enfant avec condition de ressources est versée **à compter du quatrième mois de** vie

# DEUXIEME PARTIE
# LE FILTRAGE SUR LE WEB

# I/ INTRODUCTION

Le développement récent et rapide d'Internet a provoqué l'écoulement de grand flots d'informations sur ce réseau des réseaux. Par conséquent, les utilisateurs se trouvent perdu dans un océan d'informations qu'il ne peuvent pas exploiter [Loeb 1992].

Les outils de recherche d'informations classique n'ont pas résolu le problème. Malgré qu'il délivrent à l'utilisateur une petite partie de la masse globale des informations sur le réseaux, mais cette partie reste toujours inexploitable, à cause de son grand volume qui peut atteindre des milliers de documents.

Le recours à des nouvelles techniques s'avère inévitable, c'est pour cela que les chercheurs se sont penché sur la recherche de nouveaux médiateurs, entre la source d'information et l'utilisateur, a savoir les filtres des informations.

Les filtres, qui sont logiquement positionnés comme une "troisième partie" dans la communication, entre la source et l'utilisateur, doivent posséder les connaissances et les méthodes pour examiner l'information à la source et envoyer l'information qu'il juge pertinente à l'utilisateur.

Le filtre, peut être placé du côté de l'utilisateur ou bien de coté de la source de données, dans le premier cas, qui est le plus utilisé, le filtre assiste l'utilisateur dans sa recherche d'informations pertinentes, dans le deuxième cas, le filtre est utilisé pour cibler les utilisateurs qui ont besoin d'un certain type d'informations[Loeb 1992].

## II/1/ DESCRIPTION GENERALE DUN SYSTEME DE FILTRAGE SUR LE WEB

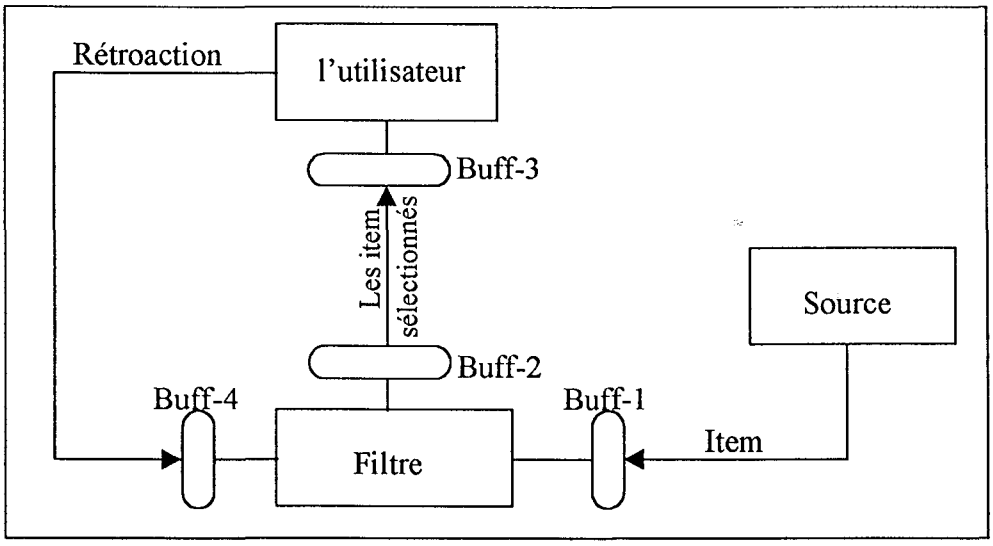



Fig.8 description générale d'un système de filtrage sur le web

La figure ci dessus, montre un système de filtrage des informations composé de sept unités [Loeb 1992], la source d'informations, le filtre des informations, l'utilisateur et quatre buffeurs.

Dans ce système, la source présente quelques descripteurs de l'information au filtre. Le filtre l'achemine à l'utilisateur, un sous-ensemble des items sélectionnés, en se basant sur des connaissances antérieures sur les besoins de l'utilisateur, enregistrés dans le profil.

Le profil est une sorte de base de données de références représentant les besoins en informations ou des références à l'utilisateur. L'utilisateur peut intervenir directement sur le filtre (la rétroaction), en exprimant ses opinons sur les documents présentés par le filtre.

Il y quatre buffeurs entre la source d'information, le filtre d'information et l'utilisateur. Ces buffeurs sont recommandés dans l'objectif de sauvegarder différents types d'informations ou les différents scénarios d'usage. Par exemple le buffeur 4 enregistre les profils des utilisateurs.

## I/2 LES ELEMENTS DU MODELE DE FILTRAGE

Loeb [Loeb 1992] a identifié 11 dimensions identifiant le paysage de filtrage. Ces dimensions peuvent être classifiées dans 4 catégories.

### *1) Disposition de l'utilisateur*

Type de l'utilisateur - proactive, casual

Tous les utilisateurs d'un système de filtrage des informations n'ont pas les mêmes besoins et espérances, ainsi ils peuvent être classés selon la nature de leurs besoins en informations.

Protection - profil protégé, historique d'utilisation protégé, information protégée.

L'habileté à personnaliser la livraison d'information dépend de la disponibilité des informations sur les utilisateurs et leurs besoins.

### *2)L'échelle du temps*

La durée de vie des informations- Minute (stock market), jour (new event, mail), décennie (technology reports), siècle (entertainment).

La valeur de l'information par rapport aux besoins des utilisateurs varie aux fil du temps.

Le modèle de stockage disponible - information sauvegardée, la vie de l'information :

Pour évaluer la valeur de l'information, le filtre se base sur :

- Les types d'informations élémentaires fournies par la source.
- La fréquence d'arrivée des informations.
- La disponibilité des informations (la rareté).

Le modèle de livraison du filtre - continue, synchrone, asynchrone :

Le filtre peut livrer les informations d'une manière continue (au fur et à mesure de leurs arrivées), d'une manière synchrone ou asynchrone selon la demande de l'utilisateur.

Le modèle d'utilisation de l'usager - continue, régulier, irrégulier, une seule fois
La durée de la session et la fréquence varie d'un usager à l'autre.

Le mécanisme de rétroaction de l'utilisateur - en temps réel ou en différé.
La réaction de l'utilisateur vis à vis des documents livrés peut être prise en compte dans la session courante ou sauvegardée pour des utilisations futures.

***3)La livraison de l'information***

Les caractéristiques de l'information - composition et taille.
La forme de l'information (texte, son, image...) a un effet sur la performance du filtre.

La manière de transport de l'information - broadcast, narrowcast, switched :
L'acheminement de l'information à l'utilisateur peut être fait de plusieurs manières.

Le matériel de l'utilisateur - télévision, PC :
L'intelligence, la performance et la capacité de stockage du terminal utilisé par l'usager influencent la capacité du filtre.

***4) le contenue de l'information***

L'attribut contenu dans l'information - descripteur, indexation :
Le filtre doit examiner les items des informations ou des descriptions avant de décider s'il doit la livrer à l'utilisateur ou non.

Comme pour le filtrage linguistique, différentes techniques sont utilisées pour le filtrage sur le web tels que les mots clés, les profils utilisateur, les méthodes de l'intelligence artificielle, des méthodes probabiliste, etc.

# II/ LES MOTS CLES

C'est une méthode très simple de détermination des intérêts des usagers et les liens entre les différents mots clés.

Si un besoin particulier d'un utilisateur est décrit par un ensemble de mots clés alors les documents contenant ces mots clés sont jugés pertinents.

## II/1 PROBLEMES AVEC LES MOTS CLES.

Selon Foltz et Dumais [Foltz and Dumais 1992], les méthodes de filtrage basé sur les mots clés peuvent conduire à l'extraction de certains documents non pertinents pour l'utilisateur, car les mots clés ne décrivent pas correctement le contexte. Un mot clé a un seul

sens, mais plusieurs concepts peuvent être décrits avec ce même mot clé, le concept a son tour peut être décrit avec plusieurs mots clés.

## II/2 UTILISATION DES MOTS CLE

Allen[Allen 1990] a conduit une série d'expériences pour explorer des modèles utilisateurs qui prédisent les préférences concernant les nouveaux articles. Dans sa méthode, il s'est basé sur les anciens articles lus et approuvés par l'usager. Il essaye d'établir une relation entre les noms, les mots clés des articles lus et ceux des nouveaux articles.

Selon Foltz et Dumais [Foltz and Dumais 1992], ce modèle a eu un succès avec les articles d'ordre général mais pas avec les articles scientifiques.

"The Information lens System" [Mackay et all 1989] [Malone 1987] permet aux usagers de créer des règles pour filtrer les messages en se basant sur les mots clés. (Ceci est appliqué surtout à la messagerie électronique car les mails sont bien structurés (expéditeur, sujet..).

## II/3 INCONVENIENTS

*Problème de vocabulaire* [Furnas 1987] : Les termes utilisés dans les documents sont variables et ne sont pas les mêmes que ceux utilisés par l'usager. D'ailleurs Furnas, Landouer, Gomez et Dumais [Furnas et all 1986] ont montré qu'il y une très faible probabilité (de 0.1 à 0.2) que deux personnes utilisent le même mot clé pour décrire le même objet.

*Problème conceptuel* [Gaines and Shaw 1989] : Le concept utilisé pour la représentation des informations peut être différent du concept recherché par l'utilisateur.

Par conséquent, les modèles basés sur les mots clés ne sont pas suffisants, les informations contextuelles et sémantiques doivent être impliquées, ce qui a conduit certains chercheurs à la modélisation des usagers.

# III/ FILTRAGE SELON LE PROFIL UTILISATEUR

## III/1 INTRODUCTION

La plupart des filtres d'informations sont limités à des contextes spécifiques, des types d'utilisateurs spécifiques, et des sources d'information spécifiques. Ceux-ci proviennent du fait que, d'une part, les besoins des utilisateurs sont souvent très ambigus, ce qui rend difficile la modélisation des utilisateurs[Allen 1990], d'autre part, les sources d'informations ne donnent pas suffisamment de détail à propos de leurs contenus [Foltz and Dumais 1992].

Loeb [Loeb 1992], Belkin et Croft [Belkin and Croft 1992], considèrent que le filtrage des informations dépendant du domaine d'application dans lequel il opère et du contexte dans lequel il est utilisé.

Les filtres d'informations doivent présenter à chaque utilisateur les informations qui correspondent à ses besoins. Mais il existe des utilisateurs occasionnels, qui n'ont pas un

besoin précis et qui cherchent des informations intéressantes, donc leur besoin n'est pas le même aux fil du temps. Comment faut-il concevoir les filtres pour ces utilisateurs ?

## III/2 LE MODELE UTILISATEURS

Un modèle utilisateur selon Kass et Finin [Kass and Finin 1989]

" *a system knowledge source containing explicit assumption on all aspects of the user that may be relevant for the behavior of the system*"

Tout modèle utilisateur, incorporé dans n'importe quel système, doit avoir trois composantes de fonctionnements principales

1- Une composante de maintenance et de représentation pour la gestion des connaissances sur l'utilisateur.
2- Une composante pour ajouter des nouvelles connaissances au modèle utilisateur.
3- Une composante qui facilite l'accès pour aider et répondre aux besoins du système dans lequel en exécute le modèle utilisateur.

## III/3 PROBLEME DE DEVELOPPEMENT DE MODELE UTILISATEURS

Plusieurs facteurs doivent être utilisés pour déterminer l'intérêt d'un utilisateur. Généralement, les usagers utilisent un ensemble de mots pour illustrer leurs besoins, mais ceci n'évite que d'autres informations puissent intervenir dans le choix des documents telles que :

- Les articles que l'utilisateur a lus dans le passé.
- L'organisme dans lequel il travaille.
- Les livres qu'il a commandés.
- La familiarité avec les articles.
- L'importance ou l'urgence.

En plus, il peut y avoir une interaction entre certains de ces facteurs et les applications.

Les difficultés de modélisation des utilisateurs, recensées par Kass et Stadnyk [Kass and Stadnyk 1992] sont les suivantes :

- L'utilisateur ne peut pas exprimer ses besoins de manière précise.
- Les besoins différents d'un utilisateur à l'autre.
- Les besoins des utilisateurs sont en continuel changement.
- L'intérêt de l'utilisateur peut être relié :
  * À la qualité et la complexité des informations.
  * Aux domaines d'intérêts.
  * Aux buts des informations.
  * Aux types des informations.
  * Aux caractéristiques des informations.

Kass et Stadnyk [Kass and Stadnyk 1992] ont recensé certaines régularités chez tous les lecteurs pour décrire leurs besoins, et ils les ont classés en 5 catégories.

| Types de catégorie | Définitions | Exemples de catégorie |
|---|---|---|
| Domaines de concept | La catégorie sémantique, utilisées pour décrire le sujet du message et | Mac Iifx |

| | | |
|---|---|---|
| | si le message va être lu ou pas. | |
| Buts | Dépend des intérêts de l'utilisateur | Besoin de savoir, capable d'aider |
| Type de messages | Décrire les majeures parties d'une classe | Vente, discussion, solution au problème. |
| Caractéristiques du message | Informations contextuelles sur le message | La longueur du message. |
| Relations | Généralement des buts reliés au concept du domaine | Besoin de vendre un moniteur, j'ai une documentation sur les HFS. |

Tab.1 Catégories de description

Ils ont trouvé aussi, des régularités dans la manière de lecture des réponses de la part des usagers.

Ces régularités sont formulées dans des règles de corrélations et des conjonctions des catégories de descriptions.

| **Format des règles** |
|---|
| [about(*topic,msg*)\|*goal*(*topic*)\|*msg-type*(*msg*)\|*msg-characteristic*([*msg*\|*topic*])] -> [read(*msg*)\|not read(*msg*)]. |
| **Exemple des règles** |
| Help(*msg*) ->not read (*msg*) |
| About(SE, *msg*) ->not read(*msg*) |
| About(printer, *msg*)^ not have (printer) -> not read(*msg*) |
| About(printer, *msg*)^want(printer)^for-sale(*msg*) -> read(*msg*) |
| About(LC, *msg*)^not know(LC)^short(*msg*) -> read(*msg*) |

Tab.2 les règles de décodage des intérêts de l'utilisateur

Ces règles peuvent être utilisées comme des filtres. En effet, chaque règle va former la base d'un stéréotype, qui va être utilisé comme un filtre initial pour les nouveaux utilisateurs.

# IV/ LES STEREOTYPES

## IV/1 INTRODUCTION

Un stéréotype représente une collection d'attribut commun à un ensemble de personnes. Les stéréotypes sont très utilisés dans les systèmes d'information, qui ont besoin d'une modélisation des usagers, pour améliorer les interactions entre le système et les utilisateurs.

Les systèmes qui utilisent le plus les stéréotypes sont les systèmes experts [Rich 1979][Rich 1983] et les systèmes d'enseignements [Chin 1989].

Bracha, Peretz et Uri [Bracha, Peretz and Uri 1997] ont présenté l'utilisation des stéréotypes comme moyen d'amélioration de l'efficacité des systèmes de filtrage d'informations, qui se base sur les profils des utilisateurs.

## IV/2 LA MODELISATION DES UTILISATEURS PAR LES STEREOTYPES

Les stéréotypes sont utilisés lorsque nous disposons d'informations incomplètes sur l'utilisateur ou lorsque la source d'information n'est pas fiable. Au cours de leurs exécutions,

les systèmes utilisant les stéréotypes, recensent des informations supplémentaires ou vérifient des informations déjà existantes sur l'utilisateur

## IV/3 INTEGRATION DES STEREOTYPES DANS LE MODELE DE FILTRAGE D'INFORMATION

L'acquisition des connaissances sur l'utilisateur peut être faite par plusieurs méthodes comme :

- Les questionnaires [Kass and Stadnyk 1992].
- Les réponses des utilisateurs aux filtrages précédents [Foltz & Dumais, 1992].
- En mémorisant le comportement des usagers vis à vis des filtrages précédents [Morita and Shinoda 1994].

## VI/3/1 DESCRIPTION DU FONCTIONNEMENT DU MODELE

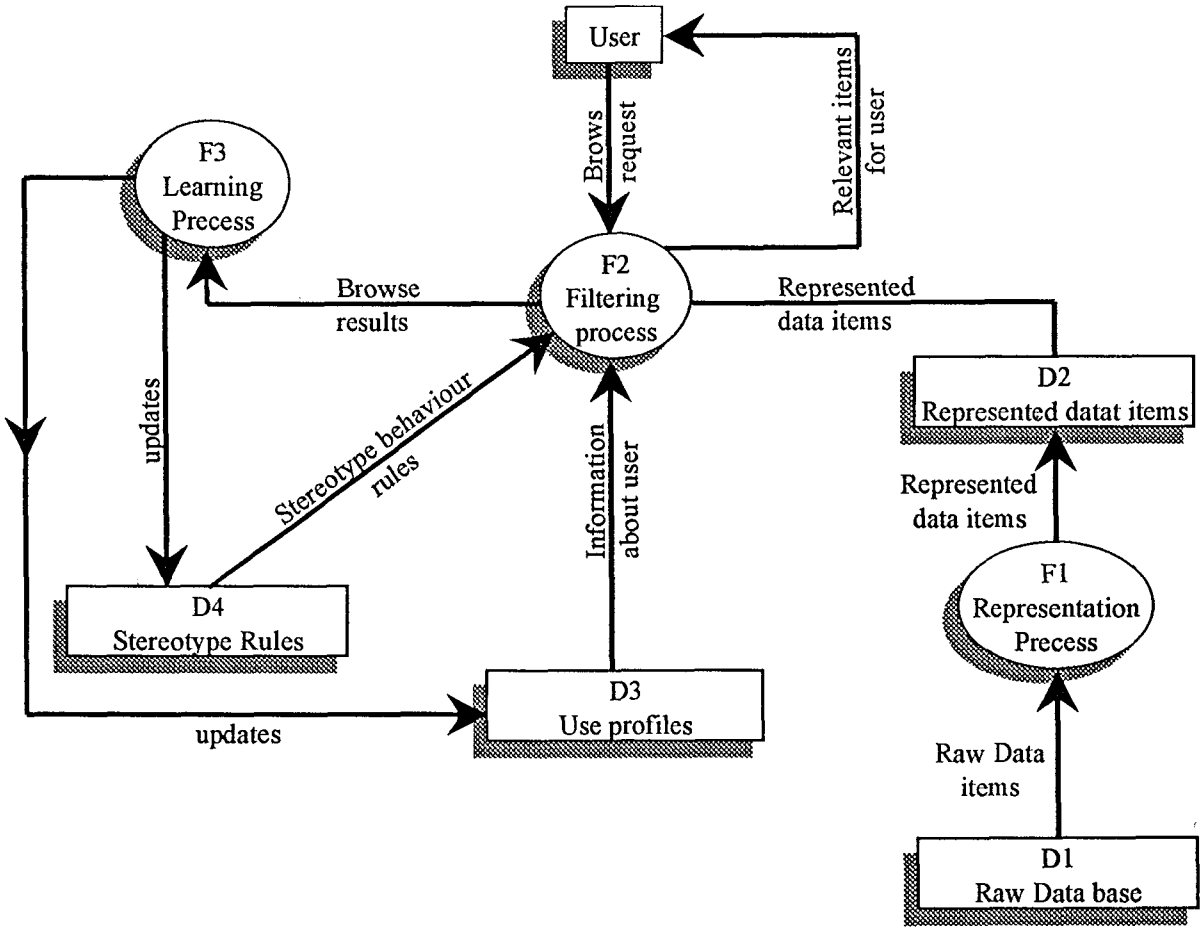



Fig. Description fonctionnelle du modèle
[Shapira, Shoval and Hanani 1995]

D2, "Represented Data Item", est une base de données qui contient des vecteurs de poids des différents termes, chaque vecteur représente un élément de donnée dans la base de

données (D1) "Raw Data Itens". Nous supposons l'existence d'un processus de représentation (F1), qui transforme les "raw data item" de la base D1 en des "represented data item" de la base D2.

D3 la base de données "user profiles", elle contient des données sur quelques utilisateurs pour le calcule des éléments pertinents. Chaque utilisateur est représenté par des vecteurs de poids des termes (les valeurs initiales du vecteur sont établies à l'aide d'un questionnaire adresser aux utilisateurs).

D4 la base de données "Stereotype Rules", elle contient des informations (représentés sous forme de règles) sur des stéréotypes pré-établis sur le comportement et les habitudes des utilisateurs. Chaque utilisateur est relier à un ou plusieurs stéréotypes.

(f2) "filtering process", c'est le processus central du modèle, il doit chercher les informations les plus pertinentes pour l'utilisateur, en fonction de sont profils est les stéréotypes aux quels il est associer.

(f3) "Learning process", ce processus est activé après le filtrage, il met à jours le profil personnel de l'utilisateur et les règles de comportement des stéréotypes.

# V/ LE MODELE DE FILTRAGE PROBABILISTE

## V/1 DEFINITION

Belkin et Croft [Belkin and Croft 1992] définissent Le filtrage probabiliste ainsi :
*"Étant donnée un objet quelconque, d'un ensemble d'objets en entrées, et un ensemble de profils, nous devons sélectionner les meilleurs couples objet-profils"*

## V/2 PRESENTATION DU MODELE

Pour tout objet $O_j$, du flux des données entrant, nous calculons la probabilité $P_i$ (i=1..N) associer à chaque profil utilisateur. En nous basant sur ce calcul, nous "filtrons" les objets (en l'éliminant pour un profil donné ou en le sélectionnant pour d'autres profils, selon les applications).

Ce modèle s'inspire des *modèles des réseaux d'inférences baysien* [Pearl 1988 ], ces réseaux représente un graphe de dépendance acyclique. Dans lequel les nœuds représentent des variables propositionnelles et les arcs représentent les relations de dépendances entre les propositions.

Si un nœud *p* cause ou implique la préposition représentée par le nœud *q*, les deux nœuds sont reliés par un arc. Le nœud *q* contient une matrice de lien qui spécifie la probabilité $P(p/q)$ pour toutes les valeurs possibles de ces deux variables.

Avec le modèle du réseau d'inférence, nous calculons la probabilité qu'un profil donné (ayant un besoin en information) est vrai sachant l'objet en entrée.

Ce qui veut dire, qu'un objet $O_j$ est pertinent pour un profil $P_i$ si $P(P_i \text{ est vrai}/ O_j) > P(P_i \text{ est faux}/ O_j)$. Ainsi le filtrage est défini comme étant le processus de détermination du profil

qui a la plus grande probabilité d'être satisfait par un objet donné. Les objets avec une faible probabilité sont éliminés.

## V/3 LES DIFFICULTES DE LA MISE EN PLACE

Belkin et Croft [Belkin and Croft 1992] soulignent que l'implémentation d'un système de filtrage utilisant cette approche, évoquent deux difficultés conceptuelles et certains problèmes d'efficacité.

- une première difficulté conceptuelle, reliée à l'indexation ou la représentation du contenu des objets, il faut donc utiliser des mots simples et des phrases des bases dans l'indexation.

- une deuxième difficulté conceptuelle provenant du fait que les objets ne sont pas statiques, comme dans une base de données, mais arrivent sous forme de flots de données très importants.

- un problème d'efficacité, dû au fait que chaque objet doit être comparé à un ou plusieurs profils, voir même des centaines.

# VI/ UTILISATION DES METHODES DE L'INTELLIGENCE ARTIFICIELLE

## VI/1 LES AGENTS INTELLIGENTS

Baclace [Baclace 1992] a mis en œuvre un modèle de construction et de raffinement automatique du profil de l'utilisateur.

### VI/1/1 DESCRIPTION DU FONCTIONNEMENT DU SYSTEME

Pour créer ce système, Baclace [Baclace 1991] s'est inspiré du *Agonic Open System* [Miller et All 1988], qui est une technique d'optimisation basée sur un modèle économique, dans lequel des agents sont en compétition à travers un système "de vente et d'achat".

Le système commence par l'identification des propriétés du document (exp. auteur "descle" mot clé "schème"), en suite, il active les agents qui sont sensibles aux propriétés trouvées.

Un algorithme d'apprentissage associe l'évaluation de l'utilisateur avec certaines propriétés du document, telles que, l'auteur, le sujet, les mots clés, l'organisation ( en se basant sur des anciennes évaluations de l'usager), ensuite, il donne des priorités aux documents.

Toute propriété du document non trouvée provoque la création d'un nouvel agent (les agents sont stockés dans une base de données) pour les utilisations futures. Ainsi le système se réamorce lui-même et donne des priorités aux informations sans aucune expérience.

Chaque agent est activé par une unique propriété ou une conjonction de propriétés (exp. mot clé "Orienté" et mot clé "Objet. Chaque agent possède une provision d'argent et un numéro de classement dans l'intervalle des réels [-1, 1] (il existe une fonction de passage du

classement discret fait par l'utilisateur vers cet ensemble). A chaque fois qu'un agent est activé, il paye un coût fixe de transaction.

Après avoir scanner les propriétés du document et l'activation de certains, nous additionnons le classement estimé auparavant de tous ces agents, pour former la priorité du document, et ainsi nous établissons une liste de priorité.

Après l'évaluation de l'utilisateur, les agents qui sont loin de la moyenne de l'évaluation de l'utilisateur sont sanctionnés. Le montant de la sanction est redistribué entre les agents qui ont une note supérieure à la moyenne. Ensuite, cela les agents activés ajustent leurs classements.

Cette compétition utilise le facteur d'encombrement (crowding-factor) qui limite le nombre d'agent actif a un maximum.

## VI/2 METHODE D'APPRENTISSAGE AUTOMATIQUE : LE SYSTEME INFOS

### VI/2/1 INTRODUCTION

Le système INFOS( Intelligent News Filtering Organisationnal System) [Mock and Vemuri 1997] est un travail qui s'inscrit dans une perspective d'amélioration de la recherche et du filtrage sur le web. INFOS organise automatiquement les données selon leurs degrés de pertinence en se basant sur les intérêts de l'utilisateur.

Le système apprend automatiquement la méthode de classement de l'utilisateur en se basant sur les propriétés des articles déjà lus, par cet utilisateur, et sur certaines autres propriétés dégagées à partir du comportement d'autres usages (ayant le même profil que cet utilisateur).

### VI/2/2 DESCRIPTIONS DU FONCTIONNEMENT DU SYSTEME

Le filtrage est réalisé à l'aide d'une technique hybride qui combine :

- des éléments de la recherche "HILL CLIMBING" basés sur les mots clés,
- une conception de la représentation des connaissances à travers le WordNet
- une analyse partielle à travers le modèle d'index.

INFOS combine les avantages de chacune de ces méthodes tout en gardant sa robustesse. L'objectif du système, est de déterminer si un article donné intéresse l'utilisateur ou non.

À partir des perspectives de l'utilisateur, INFOS construit un profil correspondant à ce dernier.

Après que l'utilisateur aurait finit de lire un article, il lui donne une notation. Mock et Vemuri utilisent des notations de type (Accepté, Rejeté, Non évalué), afin que le modèle soit le plus simple possible [Mock, 1996].

Pour classifier les articles au départ, nous utilisons une méthode simple qui se base sur les mots clé "Global Hill Climbing" (GHC).

Cette méthode a un taux d'erreur très faible, mais une proportion assez importante de documents non classés.

Ainsi, le GHC est utilisé comme une méthode simple et rapide pour un premier classement. Ensuite le GHC est combiné avec un module appelé Case_Base_Reasoning (CBR), dans lequel est incorporons un WorldNet Knewledge, le CBR est plus lent et plus complexe que le GHC mais il fournit une bonne classification.

Ainsi, si le GHC retourne des documents non classé alors le CBR est invoqué

# CONCLUSION

# CONCLUSION

Dans cette note de synthèse, nous avons essayé de présenter différents modèles et méthodes de filtrage des informations, élaborer jusqu'à nos jours. Lors de cette l'exploration de la littérature, concernant notre sujet, nous avons remarqué que la plupart des auteurs ont tendance à considérer deux grandes classes filtrage des informations en deux grandes classes, à savoir de *filtrage linguistique* et le *filtrage sur le web*.

Cependant, nous ne pouvons pas parler de deux classes disjointes. En effet, un bon système de filtrage des documents électroniques se base, d'une part, sur une bonne modélisation des utilisateurs et des outils de recherche sur le web, d'autre part, sur des techniques linguistiques afin de raffiner la qualité des réponses retourner par le système de filtrage des informations.

# BIBLIOGRAPHIE

**AHO, A. V., and ULLMAN, J. D.** (1972). The theory op parsing, translation, and compiling (Vol. I, parsing). Englewood Cliffs, NJ: Prentice-Hall.

**ALLEN, R.**(1990). *User models: Theory, method and practice*. Int. J. Man-Machine Stud. 32(1990), pp 511-543.

**ASHWIN RAM** (1992). *Natural language understands for information filtering systems.* In Communication of the ACM Vol.35 No.12, pp 80-81, December 1992.

**ASSADI H.,** 1998. *Construction d'anthologie à partir de textes techniques.* Application aux systèmes documentaires. Thèse de doctorat, Université Pierre et Marie Curie (ParisIV), octobre 1998.

**BACLACE PAUL E.** (1992). *Competitive agent for information filtering.* In communication of the ACM Vol.35 No.12, pp 39-48 ,December 1992.

**BACLACE P.** (1991). *Personal information intake filtering.* In Proceeding of Bellcore Workshop on High-performance Information Filtering (Morristown, N.J., Nov. 1991).

**BELKIN NICHOLAS J. AND CROFT W.BRUCE** (1992), *information filtering and information retrieval : two sides of the same coin ?.* In communication of the ACM Vol.35 No.12, pp 29-38, December 1992.

**BERRI J., CARTER E., DESCLES J-P., JACKIEWICZ A., MINEL J-L.**, 1996. *SAFIR, système automatique de filtrage de textes.* In Actes de la conférence traitement automatique du langage naturel (TALN'96), Marseille.

**BOURIGAULT D.**, (1993). *Analyse syntaxique locale pour le repérage de termes complexes dans un texte.* In Traitement automatique des langues. Revue éditée par l'association pour de traitement automatique des langues (ATALA), 34(2), Paris.

**BOURIGAULT D.** (1994*). LEXTER, un logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de textes,* thèse EHESS (29 JUIN 1994).

**BRACHA SHAPIRA ,PERETZ SHOVAL AND URI HANANI** (1997), *stereotypes in information filtering system.* In Information Processing & Management Vol.33, No.3, pp 273-287, 1997.

**CARTIER EMMANUEL** (1998), *Analyse automatique des textes: l'exemple des informations définitoires,* In RIFRA'98 Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique, pp 6-18, 1998.

**CHANDRASEKAR R. AND SRINIVAS B.** (1998), *GLEAN : using syntactic information in document filtering.* In Information Processing & Management Vol.34, No.5, pp 623-640, 1998.

**CHIN D.N.** (1989). *KNOME: modelling what the user knows in UC.* In A. Kobsa, & W. Wahster (Eds), User Modes in Dialog Systems (Ch.3, pp. 74-107). Berlin: Springer-Verlag.

**CURT STEVENS** (1992), Automating the creation of information filters. In communication of the ACM Vol.35 No.12, pp 48, December 1992.

**DEERWESTER S., DUMAIS S.T., FURNAS G.W., LANDAUER T.K AND HARSHMAN R.,** (1990). *Indexing by Latent Semantic Analysis.* J. Am. Soc. Inf. Sci. 41, 6(1990), pp 391-407.

**DESCLE J-P**, 1990, *Langage applicatif, langage naturelle et cognition.* Hermés. Paris.

**FAIZ RIM** (1998), *Filtrage automatique de phrases temporelles d'un texte.* In RIFRA'98 Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique, pp 55-63, 1998.

**FOLTZ PETER W. AND DUMAIS SUSAN T.** (1992), *Personalised information delivery : an analysis in information filtering methods.* In communication of the ACM Vol.35 No.12, pp 51-60, December 1992.

**FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M. AND DUMAIS, S.T** (1987). *The vocabulary problem in human-system.* communication. Communication of the ACM Vol.30 No. 11 pp 964-971, 1987.

**FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M. AND DUMAIS, S.T**(1983). *Statistical semantics: Analysis of the potential performance of keyword information system.* Bell Syst. Tech. J 62,6 (1983), 1753-1806.

**GAINES, B.R. AND SHAW, M.L.G** (1989*). comparing the conceptual systems of expert.* In Eleventh International Conference on Artificial Intelligence (1989) pp. 633-638.

**GARCIA D., AUSSENAC-GILLES N. ET COURCELLE A.**, 1999. *Exploitation, pour la modélisation, des connaissances causales détectées par COATIS dans des textes.* In Ingénierie des connaissances. Eds G. Kassel, J. Charlet et M. Zacklad. Édition Eyrolles, Paris. A paraître début 1999.

**GARCIA D.,** 1998. *Analyse automatique des textes pour l'organisation causale des actions, réalisation du système informatique COATIS.* Thèse de doctorat, Université paris-sorbonne (Paris IV), 27 Mais.

**GARCIA D.** (1998), *Exploitation pour l'élaboration de requêtes de filtrage de texte, des connaissances causales détecte par COATIS,* In RIFRA'98 Rencontre internationale sur l'extraction le Filtrage et le Résumé Automatique, pp 44-54, 1998.

**GOLDBDERG DAVID, NICHOLS DAVID, OKI BRIAN M., AND DOUGLAS TERRY** (1992), using collaborative filtering to weave an information tapestry. In communication of the ACM Vol.35 No.12, pp 61-81, December 1992.

**GROS C., ASSADI H., AUSSENAC-GILLES N. ET COURCELLE A.,** 1996. *Task models for Technical Documentation Accessing.* In proceedings of the 9th European Knowledge Acquisition Workshop (EKAW'96), Position paper. Nottingham (UK).

**HYUN-KYU KANG AND KEY-SUN CHOI** (1997), *two-level document ranking using mutual information in natural language information retrieval.* In Information Processing & Management Vol.33, No.3, pp 289-306, 1997.

**JACKIEWICZ A.**, (1998). *L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploitation contextuelle.* Thèse de doctorat, Université de Paris-Sorbonne (Paris IV).

**KASS ROBERT AND STADNYK IREN** (1992), Modelling user' interest in information filtering. In communication of the ACM Vol.35 No.12, pp 49-50, December 1992.

**KASS R., AND FININ T.** (1989). *The role of user models in cooperative interactive systems.* International Journal of Intelligent Systems, 4, pp. 81-112.

**KUIKKA EILA AND SALMINEN AIRI** (1997), *Two-dimensional filter for structured text.* In Information Processing & Management Vol.33, No.1, pp 37-54, 1997.

**LOEB SHOSHANA** (1992). *Architecting personalised delivery of multimedia information.* In communication of the ACM Vol.35 No.12, pp 39-48 ,December 1992.

**MACKAY, W.E., MALONE, T.W., CROWSTON, K., RAO, R., ROSENBLITT, D., AND CARD, S.K.**(1989*). How be experienced information lens user use rules?.* In proceeding of the ACM CHI'91 Conference on Human Factors in Computing Systems (Austin, Tex. Apr. 30-May 4). ACM/SIGCHI, New York, 1989, pp.211-216.

**MALONE, T.W., GRANT, K.R., LAI, K.Y., RAO, R. AND ROSENBLITT, D.R.**(1987). *Semistructured message are surprisingly useful for computer-supported coordination.* ACM Trans. Off. Inf. Sust. 5,2(1987), pp 115-131.

**MILLER, M.S AND DREXLER, E. MARKETS and computation** (1988). *Agoric open Systems.* In the Ecology of computation, B.A. Huberman, Ed., Elsevier, 1988, pp. 133-176.

**MOCK, K., AND VEMURI, V.** (1994). *Adaptive user interface for intelligent information filtering.* Proceedings of the third Golden West International Conference on Intelligent systems, Las Vegas, Nevada (pp. 506-517). London: Aslib.

**MOCK K.** (1996). *Hybrid hill-climbing and knowledge-base techniques for intelligent news filtering.* Thirteenth National Conference on Artificial intelligence, Portland, Oregon. Menlo Park, CA: AAAI Press; Cambridge, MA: MIT Press.

**MOCK KENRICK J.. AND VEMURI V.RAO** (1997). *information filtering via hill climbing, wordnet, and index patterns.* In Information Processing & Management Vol.33, No.5, pp 633-644, 1997.

**MORITA MASAHIRO AND SHINODA YOICHI** (1994), *Information Filtering on user behaviour analysis and best match text retrieval.* In SIGIR '94 : proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval pp 272-281.

**PASCUAL E.,** (1991). *Représentation de l'architecture textuelle et génération de textes.* Thèse de doctorat en informatique, Toulouse, 1991.

**PEARL J**(1988*). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

**PERSIN MICHEL** (1994). *Document filtering for fast ranking.* In SIGIR '94 : proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval pp 339-348.

**REBEYROLLE JOSETTE AND PERY-WOODLY MARIE-PAULE** (1998),*Repérage d'objet textuels fonctionnels pour le filtrage d'information: le cas de la définition.* In RIFRA'98 Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique, pp 19-30, 1998.

**RICH E.** (1979). *User modelling via stereotypes.* Cognitive Science, 3, 329-354.

**RICH E.** (1983). *User are individuals: individualising use models.* International journal of Man-Machine Studies, 18, pp. 199-214.

**SHAPIRA B., SHOVAL P. AND HANANI, U** (1995). *Hypertext browsing: a new model based on hypergraph dynamic construction using data analysis methods.* Proceedings of NGITS'95, the Second International Workshop on Next Generation information Technologies and systems. Naharia, Israel.

**VACHIER C., MAYER F., GRATIN C. AND TALBOT H.** *filtrage par décomposition morphologique : application a l'extraction de structures rectilignes.* In 9éme congres de reconnaissances des formes et intelligence artificielle.

**VIRBEL J.** (1989). *The contribution of linguistic knowledge to the interpretation of text structures.* In J. André, V? Q uint, R. K. Furuta (Ed.), structured documents, Cambridge University Press, Cambridge, 161-181, 1989.

**VIRBEL J.**(1985). *Langage et méta-language dans le texte du point de vue de l'édition en informatique textuelle.* Cahier de grammaire, 10, pp1-72, 1985.