Filtrage de l'Information

*O. Nouali, S. Benmezùine, **M. Djaid,** S. **Sidi-Boumediène** ,*
Laboratoire Intelligence Artificielle, CE.R.I.S. T
Rue des 3 frères Aïssou, Ben Aknoun, Alger, Algérie
Fax : 213 (2) 91.21.26 - Tél. : 213 (2) 91.18.21
E-mail : (benmcziane, djaid, nouali)(àtassili.cerisl.dz

1-Introduction:

r~ #A **r** *-4* • vec l'avènement de l'Internet, la très grande masse d'informations devenue disponible nécessite aujourd'hui de consacrer une partie considérable de notre temps à l'extraction de l'information pertinente.

L'utilisateur équipé seulement d'outils de recherche d'information sur les réseaux, ne peut pas faire face au flux d'information générée.

Au lieu de laisser l'utilisateur dépenser son temps à chercher l'information dont il a besoin, la tendance actuelle est de concevoir des mécanismes qui permettent de lui faciliter la tâche en lui faisant parvenir continuellement l'information qui l'intéresse, c'est ce qu'on appelle les services de Dissémination Sélective de l'Information [1].

Le filtrage de l'information est un nom donné à une variété de processus dont le but est de faire parvenir justement, à partir de larges volumes d'informations générées dynamiquement, les informations aux personnes qui en ont besoin. Le rôle du filtrage de l'information est donc d'augmenter la quantité d'informations pertinentes collectées à partir de différentes sources.
Les domaines d'application du filtrage de l'information sont assez variés, et d'une grande importance économique, parmi eux: Mailing list, Usenet News, E-mail, Web[2], etc..

2. Filtrage d'information et recherche d'information

La recherche d'information est étroitement liée au filtrage de l'information dans le sens où ils ont le même but qui est de retrouver l'information pertinente pour un certain utilisateur. La distinction entre les deux processus n'est souvent pas claire. Ainsi, si **l'on** considère la recherche d'information, d'un point de vue très large, comme étant un processus de sélection de l'information, alors le filtrage de l'information est simplement

un cas particulier de la recherche d'information où l'information arrive d'une manière dynamique. D'un autre point de vue, si l'on considère que la recherche d'information est un processus assurant la sélection d'une information relativement statique en réponse à des requêtes relativement dynamiques, alors le filtrage de l'information est mieux vu comme étant le problème dual de la recherche d'information [5].

Les principales différences sont:
la recherche d'information assure la collection et l'organisation des documents, le filtrage de l'information assure la distribution des documents aux personnes qui en ont besoin.

un système de recherche d'information établit une sélection de documents à partir d'une bases de données statique, le filtrage est une sélection et/ou souvent une élimination d'information à partir d'une source d'information dynamique.

un système de recherche d'information est utilisé par une seule personne à un moment donné (une requête à la fois). Par contre un système de filtrage est un processus itératif (multiples parcours) pouvant être utilisé par une ou plusieurs personnes avec des intérêts à long terme.

le filtrage de l'information est généralement appliqué à des flux arrivants de nouvelles données tandis que dans la recherche d'information, les modifications dans la base de données ne sont pas fréquentes et la recherche n'est pas limitée aux nouveaux documents.

le filtrage de l'information entraîne le processus de déplacement *(ftemoving)* de l'information du flux de données alors que la recherche d'information implique le processus de collecte *(Finding)* de l'information dans la base de données.

Le filtrage est basé sur des descriptions des préférences d'un individu ou d'un groupe d'individus appelées profils: Ils représentent leurs intérêts à long terme. Contrairement dans la recherche d'information les requêtes reflètent des intérêts à court terme.

3. Méthodes de Filtrage

Le filtrage de l'information constitue un domaine proche de la recherche d'information, et les méthodes employées sont similaires: c'est seulement l'approche ou la vision qui diffère. En effet, le processus de filtrage se trouve être le problème dual de la recherche d'information du fait que dans le cas de la recherche d'information, une base de documents indexée par les requêtes des utilisateurs et dans le cas du filtrage de l'information, des documents indexant une base de profils(figurel).

Requête Utilisateur

Nouveau Document

{m_{R1},m_{R2},.. m_{Ri}, .. $m_{K,}$}

{m,,,, **m_{D2},.. mu;,.. m_{Un}}**

Liste des **Liste des mots Clés**

m, **m,**

m_2 **m_2**

m. *m,*

niL **m_k**

Base de données des Documents

{m_Dn mu 12, •• muni, • muinl}
{m_D2i m_{1J2}2, •• ni|$_{)2}$i2, • m_{D2n}2}

i^{m} ni i $^{m1)j2,}$ • niujij,.• $m_{l)jnj}$}

{^{m}Dml> ^{ta}Dm2,• $^{ml>niim,}$ •• ^{m}Dmnm}

Figure 2. a: Indexation de la base des Documents dans le processus de Recherche d'Information

Base de données des Documents

{lïipn mpi2, • Blpmi •• m,.,$_{nl}$}
{mp2i mp22, • mp$_{2i}$2,.• m,.2n2}

{m,.; , m,.$_{i2}$, •• m$_{Pjr}$„ . mi-inj}

{m,.$_{m}$i, rapm2.• "•l'mûm -• ^Mninm J

Figure 2. h: Indexation de la base des Profils dans le processus de Filtrage de l'Information

***Figurel** : Indexation dans les processus de Recherche et de Filtrage d'Information*

L'une des toutes premières formes de filtrage de l'information se trouve être la dissémination sélective et automatique de l'information: courrier électronique, conférences électroniques, distribution d'articles, etc..

Un bon outil de filtrage de l'information doit répondre à deux questions fondamentales:

(i) quelles sont les méthodes les plus efficaces pour la correspondance (*matching)* des intérêts des utilisateurs avec l'information disponible ?
(ii) comment devrait-on décrire les intérêts d'un utilisateur ?

Dans ce qui suit, nous allons présenter les principales techniques employées dans le domaine du filtrage, ainsi que les diverses manières de représenter les intérêts d'un utilisateur.

3.1. **Les techniques traditionnelles**

Elles se basent sur l'occurrence d'un ensemble de mots clés pour identifier ou reconnaître les documents pertinents.

Filtrage Full-text: C'est une méthode directe qui consiste, en se basant sur le parcours du texte, à sélectionner tous les documents contenant une certaine chaîne de caractères(mots clés, expression booléenne,...).

Filtrage basé sur l'indexation: Chaque document est représenté par une liste de mots clés décrivant le contenu du document. Les mots clés sont stockés dans un fichier indexe et pour chaque mot clé est établie une liste de pointeurs désignant les documents qui lui sont relatifs (Figure 1).

Filtrage Booléen: L'utilisateur exprime ses profils par des mots qui doivent exister ou ne doivent pas exister dans le document à recevoir. Le modèle n'autorise que la conjonction et la négation des mots. Cependant, l'utilisateur peut simuler la disjonction en plusieurs profils.

Filtrage Vectoriel: La requête et le document sont identifiés par un ensemble de termes ou mots clés. Si il y a m termes pour identifier un document D, alors le document est représenté conceptuellement par un espace vectoriel de k dimensions:

D = {(T1,W1), (T2,W2, ... , (Tk,Wk) et Wk différent de 0} (Ti: terme, Wi: poids).

La requête est représentée de la même façon. Pour un couple (Document, Requête), un traitement de mesure de similarité est effectué pour déterminer si le document est intéressent ou non.

3.2. Les techniques utilisant l'information sémantique :

Traitement du langage naturel: Les techniques du traitement du langage naturel

cherchent à améliorer les performances des systèmes de filtrage en unifiant la sémantique des documents et les profils des utilisateurs. Par exemple, l'utilisation de phrases au lieu de termes pour représenter les profils et les documents (pour l'indexation) s'avère meilleure car les phrases apportent un contenu sémantique plus fort [6].

Latent Semantic Indexing (LSI): C'est une extension de la méthode vectorielle standard. Elle utilise l'information sémantique cachée des mots (Latent Semantic). La méthode nécessite une étude de tout le texte pour en extraire des relations utiles entre les termes et les documents. Des techniques statistiques sont utilisées pour calculer et simuler ces associations. Le principe de la méthode consiste à construire une matrice (termes-documents), ensuite est réduite en lui appliquant la méthode de décomposition SVD (approximation par combinaisons linéaires). Cette méthode LSI a prouve une meilleur performance (7j. En effet, elle permet de sélectionner des documents même s'ils n'ont pas de mots communs avec les profils.

3.3. Modélisation du profil de l'utilisateur

Décrire les intérêts d'un utilisateur est une tâche difficile.
Pour savoir comment configurer les modèles des utilisateurs pour un système de filtrage de l'information, une étude d'observation peut être entreprise pour noter comment font les lecteurs pour décider des messages intéressants [9].

Profil de mots clés: Généralement, un utilisateur fournit un ensemble de mots clés. Néanmoins, cette technique est ambiguë du fait:

- qu'un mot peut avoir plus d'un sens

- et qu'un concept peut être décrit par plusieurs mots.

C'est pourquoi, plusieurs autres sources d'informations peuvent être utilisées telles que:

- dans quelle organisation travaille l'utilisateur ?
- quels articles a-t-il déjà lu dans le passé ?
- quels articles a-t-il commandé ?

Profil de documents: L'idée de filtrage dans les techniques utilisant ce type particulier de profils, est généralement de créer un espace de documents jugés intéressants par un utilisateur. Et chaque nouveau document se trouvant être proche aux documents dans cet espace, est alors considéré comme pertinent. De ce fait, le profil de document fournit une représentation simple et très efficace des intérêts d'un utilisateur. De plus, l'indication d'un petit nombre de documents pertinents est aussi efficace (si ce n'est plus) mais beaucoup plus simple que la génération d'une longue liste de mots et/ou d'expressions

pour décrire, souvent difficilement, les intérêts d'une personne.

Filtrage en collaboration: les utilisateurs insèrent leurs remarques à l'intérieur du document (par exemple, dire qu'un document est intéressant ou pas) et ces remarques dites annotations sont accessibles par les filtres d'autres utilisateurs [12]. La modification des profils se basent sur les annotations des autres utilisateurs.

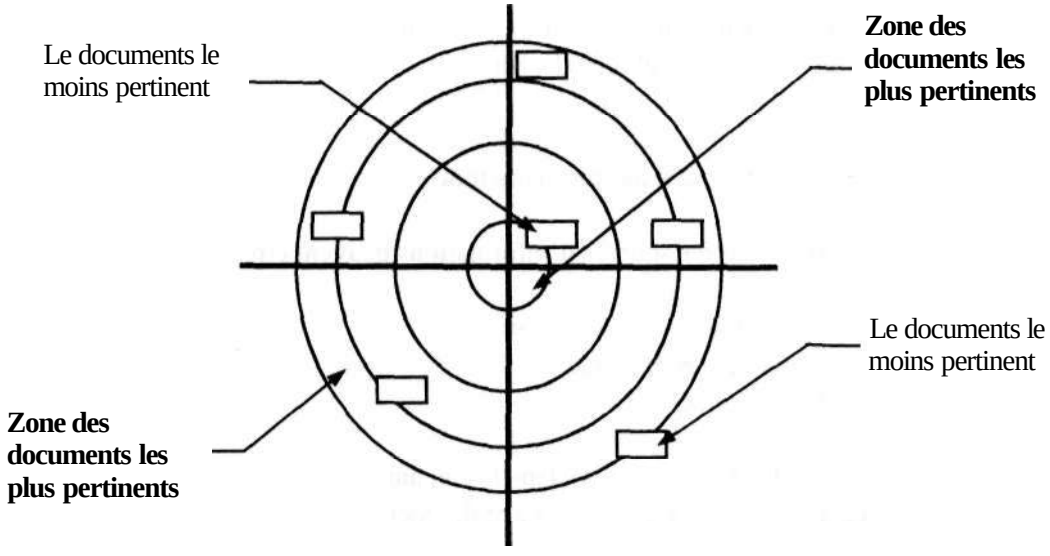
4. Quelques systèmes de filtrage existants:

4.1. INFOSCAN

C'est un outil qui permet de filtrer l'information suivant les intérêts spécifiques des utilisateurs. Il sert à filtrer, à repérer ou à classer n'importe quels documents de format **texte.**

L'utilisateur doit décrire ses intérêts au système en tapant des mots clés dans des filtres (profils). Les filtres sont des descriptions de sujets qui intéressent l'utilisateur. Chaque mot clé d'un filtre est accompagné d'une pondération et d'une portée. La pondération pour l'aspect priorité du filtre et la portée permet de chercher les mots clés dans les documents.

Le système cherche les mots clés dans les documents et affiche les résultats sur un écran de radar qui permet à l'utilisateur de voir, d'un simple coup d'oeil, les documents les plu^

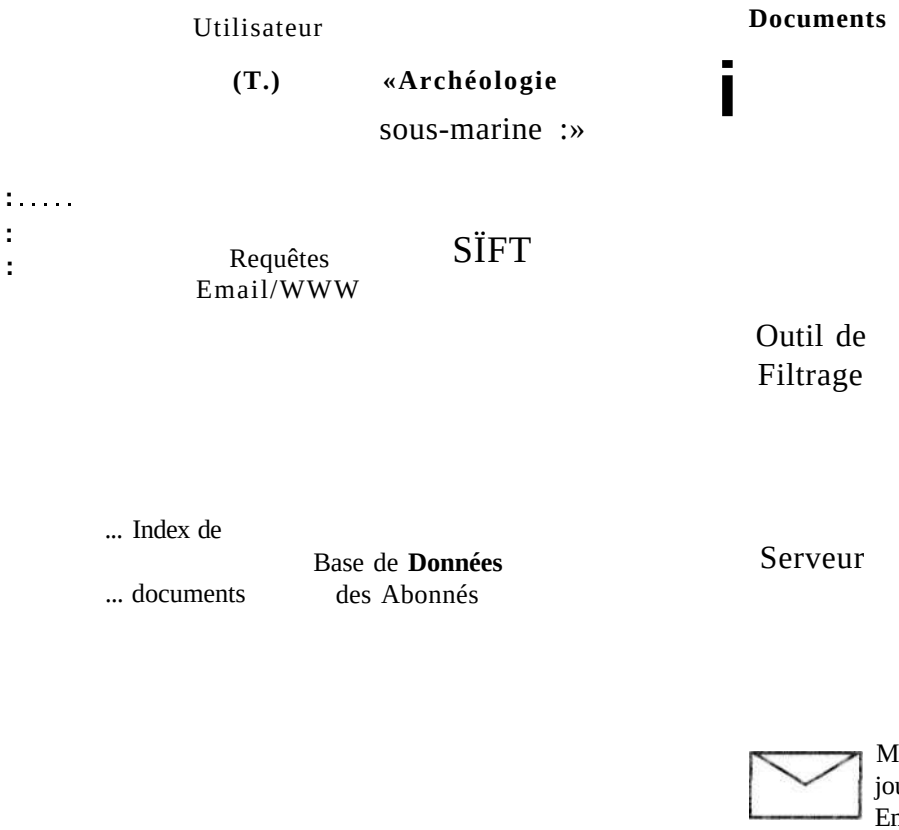


pertinents sans même lire un seul mot (figure 2).

***Figure** 2: Présentation des documents par le système INFOSCAN*

4.2. SIFT (Stanford Information Filtering Tool)

C'est un outil de dissémination de l'information qui permet de sélectionner, à partir de



larges volumes d'informations, les informations pertinentes et de les envoyer aux personnes qui en ont besoin (figure 3).

Figure 3: Architecture et fonctionnement de S/FT

L'utilisateur intéressé par un tel service s'inscrit en soumettant les profils qui décrivent ses intcrcts. Ensuite, il reçoit passivement les nouvelles et les informations filtrées qui répondent à ses besoins.

Caractéristiques

1) Supporte le filtrage **full-text** en utilisant les modèles classiques de recherche

d'information.

2) Capable de traiter un large volume d'information avec un grand nombre de profiles en utilisant les nouvelles techniques d'indexation.

3) S'exécute sous UNIX.
Les capacités du système ont été testé sur différents serveurs: exemple Usenet.

5. Conclusion

Dans cet article, nous avons présenté la définition, le rôle et les différentes méthodes du processus de filtrage de l'information. Du fait que ce dernier soit étroitement lié à la recherche d'information, les méthodes de filtrage actuelles sont basées d'une façon directe ou indirecte sur les techniques et des méthodes traditionnelles de recherche d'information.

Néanmoins, ce domaine reste ouvert vers d'autres tendances. Ainsi, certains chercheurs utilisent les techniques traditionnelles en essayant toujours de les améliorer en proposant de nouvelles approches qui tentent de capter le plus d'information sémantique.

De ce fait et vu la diversification des tendances classiques et actuelles, il n'y a pas encore de conclusion concrète. Indexer des phrases ou expressions au lieu de mots clés apporte des améliorations certaines à l'efficacité du filtrage au frais d'un pré-traitement élaboré (analyse partielle ou totale et analyse syntaxique de la phrase).
Par analogie, la méthode LSI par exemple, nécessite
(1) la disponibilité d'un corpus pour construire la matrice termes-profils et
(2) un temps d'exécution important pour la méthode SVD pour donner un résultat assez satisfaisant.

De ce fait les méthodes les plus récentes (traitement du langage naturel, **LSI** et réseaux neuronaux) semblent prometteuses.

» ^ — Références Bibliographiques

[I] Tak W. Tan, Hector (iarcia-Molina - {tyan,hcclor}(W cs.stanford.edu
SIFT, A Tool for Wide-Area Information Dissémination
In Proceedings oflhe 1995 USHNIX Technical Conférence, pp. 177-186, 1995

[2| Steve Gant - gants(« ils.unc.edu
A Samplc Information F Hier for the Web
10 Avril 1995

[3] Belkin Nicholas J, Croit W Bruce
Information Filtering and Information Retrieval : Two s ides of the same coin ?
Communication of the ACM, volume 35, N° 12, pp. 29-38, Décembre 1992

[4j Juha Takkinen - juhta(a ida.liu.se
Introduction to Course
Information Retrieval and Information Filtering (IRIF), Spring 1996
http://www.ida.liu.se/labs/iislab/courses/IRIF_introduktion.html

(5) *Information Filtering Defined*
Douglas Oard
12 Décembre 1995

[6] Croft

[7] Foltz, P. W.
Using Latent Semanlic Indexing for information filtering
In Proceedings of the ACM Conférence on Office Information Systems
ACM/SIGOIS, New York, Avril 1990, pp. 40-47.

(8) Roland Hjerppe - rfaj@ida.liu.se, Juha Takkinen - juhta(<7 ida.liu.se
Personalized Information Filtering and ML
Information Retrieval and Information Filtering (IRIF), Spring 1996
http://www.ida.liu.se/labs/iislab/courses/IRIF_ml.html

[9] Stadnyk Irène, Kass Robert
Modeling Décision Making of USENET News Readers
Technology Représentation, CFAR, pp. 91-103, 1991

[10] Stadnyk Irène, Kass Robert
Modeling User 's Interests in Information Filters
Communication of the ACM, volume 35, N° 12, pp. 49-50, Décembre 1992

[II] Frederik Kilander - fk(« dsv.su.se
Message Classification and Filtering
Sweden, 4 Janvier 1995