

Vers un SYstème de Fouille Automatique et d'eXtraction d'information : SYFAX Cas de l'enseignement

Habib SMEI, Abdelmajid BEN HAMADOU - LARIS SFAX
Mesaac MAKPANGOU - INRIA PARIS

Habib.smei@isetsf.rnu.tn, Abdelmajid.benhamadou@fsegs.rnu.tn, Mesaac.makpangou@inria.fr

Résumé :

Internet est un vaste et complexe environnement contenant des téraoctets de ressources scientifiques, pédagogiques, éducatives, commerciales, ludiques, médicales,...

*L'exploitation de ce volume sans cesse croissant nécessite des outils de recherche, de sélection et d'analyse, localisant et identifiant précisément l'information. En effet, lorsqu'on interroge un outil de recherche, on voit afficher parfois beaucoup de résultats dont une majorité se rapportant peu ou pas au sujet (**bruit**) et parfois très peu de résultats (**silence**).*

Une solution à ces problèmes peut être la spécialisation (se préoccuper d'un domaine précis) dans le stockage et la gestion des documents.

*Dans cet article, nous présentons une étude comparative des différents outils et systèmes de recherche existants, leurs caractéristiques et leurs limites. Nous proposons **SYFAX** un système de stockage (entrepôt de données) et de recherche de documents qui permet de gérer des profils utilisateurs afin de leurs proposer des recommandations et les notifier lorsque des nouveaux documents apparaissent.*

Mots clés : moteurs de recherche ; Web ; recommandations ; profilage ; notification ; indexation ; entrepôt.

Abstract

Internet is a vast and complex environment containing terabytes of scientific, educational, commercial, playful, medical resources...

The management of this volume constantly increasing requires tools of research, selection and analysis, localizing and identifying information precisely. Indeed, when we interrogate a tool of research, we see to display a lot of results sometimes of which a majority relating little or not to the topic (noise) and sometimes very few results (silence).

A solution to these problems can be the specialization (to worry of a precise domain) in the storage and the management of the documentations.

In this paper, we present a survey critical of the different tools and system of research, their features and their limits, we propose SYFAX a system of storage (data warehouse) and documentation research that permits to manage some profiles users in order to their to make some recommendations and to notify them when some new documentations appear.

Keywords: search engine; Web; recommendations; streamlining; data warehouse.

I. INTRODUCTION :

Internet est une gigantesque mine d'informations contenant non seulement des livres et des articles, mais aussi des données scientifiques, des menus, des comptes rendus de conférences, des publicités, des enregistrements audio et vidéo et des transcriptions de conversations interactives. L'anecdotique et l'éphémère sont jetés pèle mèle dans l'important et le durable.

En effet, plus de deux milliards de pages sont accessibles par l'Internet. Ce nombre peut être beaucoup plus important si nous prenons en compte les pages non directement vues par les moteurs de recherche. Une étude publiée dernièrement dans le magazine scientifique *Nature* (<http://www.nature.com>) relève que les moteurs de recherche les plus performants n'indexent pas plus d'un sixième du volume d'informations sur le Web.

Cette masse de documents représente ainsi l'équivalent de 10 milliards de pages A4. L'Internet invisible serait constitué de près de 1 million de Gigaoctets. [Sources : LEAT, UNSA - CNRS.]

L'exploitation de ce volume sans cesse croissant de document, nécessite des outils de recherche, de sélection, d'analyse et de traduction performants, localisant et identifiant précisément l'information. En effet, les outils de recherche actuels rencontrent des difficultés quand à l'accès et la localisation (indexation) des documents requis par les utilisateurs, de plus, ils n'offrent pas des moyens efficaces d'expression précises des requêtes et ne sont pas dotés de moyens efficaces de filtrage permettant de réduire le bruit.

Il va s'en dire que si on se projette dans un domaine limité ou la communauté d'utilisateurs s'intéresse à des sujets communs (enseignement, médecine, commerce,...) on peut trouver un langage commun où des répertoires peuvent stocker, indexer et gérer seulement les documents ayant une relation au sujet traité. C'est une solution pour collecter le maximum de documents se rapportant au sujet traité pour pouvoir par la suite les indexer et fournir des outils de recherche permettant de trouver les documents appropriés à une requête donnée.

Cet article se divise en deux parties :

Dans la première, nous présentons une étude comparative des outils actuels de recherche et d'indexation d'informations. Nous donnons leurs caractéristiques et leurs limites.

Dans la deuxième partie, nous présentons SYFAX, un Système de Fouille Automatique et d'eXtraction d'information appliqué à l'Enseignement. Nous détaillons ses composants et leurs caractéristiques.

II. LES OUTILS DE RECHERCHE ET D'INDEXATION

Plusieurs outils ont été mis en place afin de présenter à l'utilisateur des interfaces faciles à utiliser et permettent la recherche rapide des documents dans le réseau. Les premiers outils qui sont apparus sont les moteurs de recherche, viennent ensuite les annuaires, puis les métamoteurs qui sont des agents de collecte d'informations.

II.1. Les moteurs de recherche

Les premiers robots indexeurs sont apparus en 1994, se sont des logiciels qui visitent les serveurs Web de tous pays et indexent les documents trouvés.

A partir d'une première liste d'URL, le robot indexe chaque page HTML, et progresse vers d'autres documents en suivant les liens hypertextes.

Les méthodes d'indexation des données sont diverses. *Lycos* par exemple indexe le début des pages, d'autres le document complet, certains prennent en compte les données de l'en-tête, etc.

Les efforts actuels portent sur l'amélioration des algorithmes de recherche et de l'affichage des résultats : le calcul d'occurrence des termes, pondéré suivant la place occupée dans les pages indexées, permet d'attribuer un score à chaque document et d'établir un classement par ordre de pertinence. Les résultats actuels restent insuffisants et présentent néanmoins quelques limites que celle de la recherche d'information en texte intégral : l'utilisation du langage naturel et l'absence de traitement linguistique sont à l'origine de bruit et de silence. A noter cependant que quelques serveurs (*Echo et google par exemple*) ont intégré dans leurs dernières versions quelques règles grammaticales.

Un moteur de recherche est composé d'un **robot**, d'une **base de données** et d'un **agent**.

a- Les robots :

Ils sont appelés des "wanderers" (du verbe to wander : vagabonder, errer), des "crawlers" (du verbe to crawl : ramper, se traîner) et aussi des "spiders" (araignées). Ce sont des programmes informatiques qui parcourent le WEB pour référencer les liens qui existent dans les pages. Le robot démarre d'une page de liens et suivra de façon récursive tous les liens qu'il trouvera à partir de cette page initiale.

Ces robots utilisent le protocole http pour repérer les documents chez les serveurs, indexer l'espace pour la recherche par mots - clés, rechercher les liens morts pour la maintenance des sites à jour. C'est la qualité de la démarche du robot lorsqu'il parcourt la toile qui détermine la qualité et la quantité des informations ramenées pour alimenter sa base de données.

b- La base de données

Les données ramenées par les robots sont indexées dans des catalogues qui contiennent des informations descriptives d'un document (adresse, titres, sous-titres, mots des premières lignes des textes, résumés, éventuellement texte intégral...). Ces données sont stockées dans la base de données du moteur avec une adresse qui localise les documents.

La taille de la base de données détermine la couverture de la recherche.

c- L'agent

Il effectue la recherche pour l'utilisateur et propose une liste de réponses classées, dans un certain ordre de pertinence.....

Les moteurs de recherche affichent les adresses des documents qui mentionnent le plus fréquemment le mot-clé recherché et un résumé d'une ou deux lignes du début de document.

II.2. Les catalogues matières ou annuaires.

Par rapport aux moteurs de recherche qui offrent une recherche ouverte sur toutes les pages d'un document, les catalogues indexent d'une manière hiérarchique les documents, en plus se sont des individus qui les sélectionnent et les annotent. Ils constituent une table des matières géante et sont interrogeables par sujet matière.

Bien qu'ils sont suffisamment structurés et organisés, les catalogues ont malgré tout des limites telles que la fréquence des mises à jour et une couverture moins large (**seule une petite partie du réseau est référencé**), en plus du coût de maintenance qu'engendre le maintien de leurs bases de données. L'utilisation de ces annuaires nécessite de connaître un minimum sur le sujet pour choisir la catégorie.

II.3. Les métamoteurs

Ce sont des agents de collecte d'informations dont le rôle est d'optimiser la recherche et de l'enrichir en associant plusieurs moteurs en même temps.

L'objectif étant d'aboutir à une large couverture du Web en diversifiant les méthodes de recherche et en associant les bases d'index des différents moteurs concernés.

A part la couverture, ces outils optimisent le temps de recherche en exécutant en parallèle les différents moteurs de recherche concernés.

L'inconvénient est qu'il est difficile d'utiliser les possibilités de recherche avancée des moteurs (On ne bénéficie pas des particularités de chaque outil), de plus le volume de l'espace disque utilisé pour enregistrer des pages Web trouvés surtout dans le cas d'une recherche offline est très important. Ces métamoteurs sont souvent trop bavards.

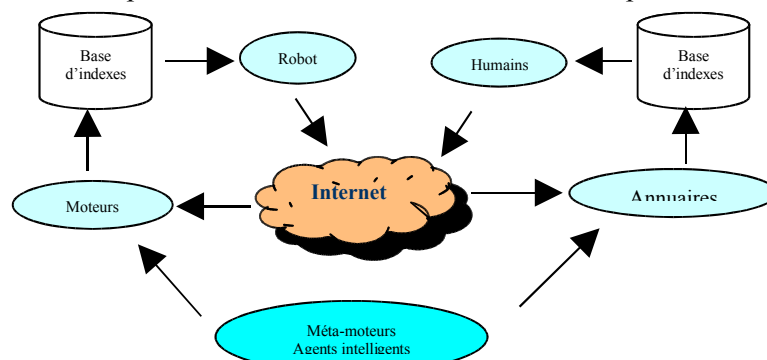


Figure 1. Recherche d'informations

II.4. Autre solution : (Les systèmes de recommandations)

Un système de recommandation est une solution pour partager les informations entre un groupe d'utilisateurs. Son objectif est d'aider les utilisateurs à faire leurs choix dans un domaine où peu d'informations leur sont disponibles, afin de trier et d'évaluer les alternatives possibles. Il est décomposé en trois entités de base : le groupe d'agents producteurs de recommandations, permettant ainsi d'envoyer des propositions et suggestions aux utilisateurs, le module de calcul de recommandations, qui permet de décider qu'une telle ressource doit être envoyée à tel utilisateur et le groupe de consommateurs des recommandations, qui sont les utilisateurs désirant recevoir des recommandations du système.

Les recommandations sont les entrées qui permettent l'évaluation des sites (agrégation et envoi aux destinataires appropriés).

Ces systèmes se basent principalement sur les profils utilisateurs qui sont des structures de données décrivant les centres d'intérêts des utilisateurs dans l'espace des objets à recommander. Une fois une telle structure construite, on peut l'utiliser soit pour filtrer les objets disponibles (on parle alors de filtrage basé sur le contenu), soit pour recommander à l'utilisateur ce qui satisfait d'autres utilisateurs ayant un profil similaire (on parle alors de filtrage collaboratif).

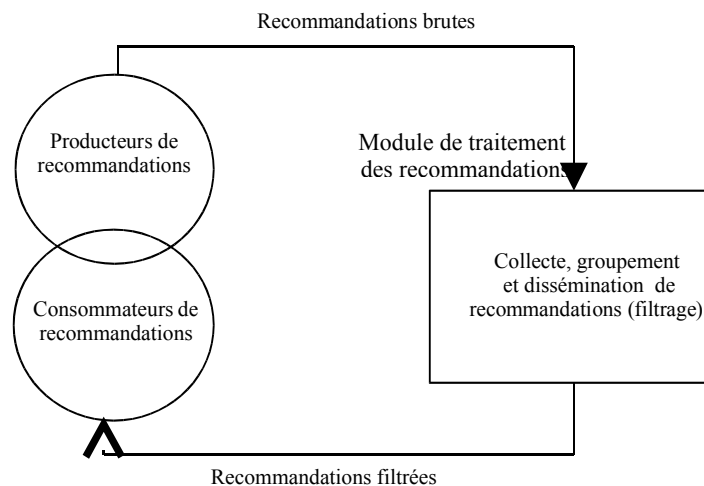


Figure 2. Système de recommandations

III. LE SYSTÈME SYFAX

III.1. Principales motivations :

Comme nous l'avons ci-dessus indiqué, les outils de recherche et d'indexation d'informations ont le mérite d'aider l'utilisateur à mieux trouver et extraire l'information du Web. Cependant, l'examen de ces outils fait ressortir certaines insuffisances telles que la non couverture de tout le contenu du Web, l'absence de pertinence et une structuration non efficace des informations indexées.

Pour pallier ces insuffisances, d'autres solutions sont apparues, telles que les portails qui constituent des portes d'entrée à un assortiment de services en tout genre (annuaire de pages jaunes, météo, bourse, actualités, informations touristiques, shopping, petites annonces, etc.). Ils proposent des pointeurs vers des sujets bien répertoriés et structurés afin d'aider l'utilisateur à trouver les informations recherchées sans qu'il fait recours à des outils de recherche.

Bien qu'ils facilitent l'accès aux informations, ces portails peuvent égarer l'utilisateur dans la multitude de services et sujets proposés. C'est pourquoi, on voit apparaître des portails dits communautaires qui empruntent au portail traditionnel l'aspect généraliste de sa gamme de services proposés, mais centrent le contenu de l'information sur une thématique bien déterminée, comme par exemple le sport, la musique, les voyages, la gastronomie, etc.

Ces portails correspondent plus aux attentes des Internauts, puisqu'ils permettent de réunir des gens qui partagent la même passion et un accès de qualité aux informations pour les membres de la communauté.

Cependant, les utilisateurs de ces portails sont des simples consommateurs d'information et de services. Ces utilisateurs viennent chercher des informations auprès de ces portails communautaires sans pour autant qu'il y aura de réactions de leurs part pouvant contribuer à améliorer les services proposés. En effet, une participation active des utilisateurs s'avère très importante. Elle permet d'enrichir la base de données des portails par l'ajout de nouveaux documents ou aussi, en fournissant un outil d'annotation de documents pour améliorer par la suite la qualité des services proposés. Tous les acteurs de la communauté seront potentiellement actifs. Il ne s'agit pas qu'il y ait un petit nombre de producteurs d'information pour un grand nombre de consommateurs. Les utilisateurs de ces portails peuvent en effet échanger des documents et des idées entre eux, on peut donc intégrer des services permettant d'établir des relations d'échange entre les membres de la communauté.

III.2. Présentation générale :

L'objectif du système SYFAX est de mettre à disposition à une communauté (i.e., universitaire) un portail gérant des documents pertinents se rapportant à leur domaine. Il utilise des outils de recherche et de filtrage qui assurent la livraison de documents pertinents aux utilisateurs. Le filtrage est basé sur un système de recommandation qui collecte les avis des utilisateurs sur un document pour éventuellement le recommander à d'autres utilisateurs. En effet, l'utilisateur donne son avis sur les documents lus et ses réactions peuvent être annotées et consultées par d'autres. On établit ainsi des relations document-document et document-utilisateur.

Le système gère une base de données contenant des informations sur les utilisateurs de la communauté, pour qu'il détienne une vue sur leurs préférences précises, et des informations sur les documents qu'ils consultent. Cela lui permet de recommander des documents aux utilisateurs désirant que le système les informe lorsqu'un nouveau document se rapportant à leurs profils apparaît ou mise à jour.

Quant à l'accès aux informations, SYFAX assure une haute disponibilité des informations contenues, puisqu'il est conçu dans un environnement Intranet. Il contient des entrepôts de données proches des utilisateurs accessibles par des liaisons rapides. Néanmoins, le système peut être utilisé dans un environnement Extranet ou Internet.

Dans sa conception globale, le système est réparti, c'est-à-dire il est constitué d'un ensemble de sites contenant chacun des documents stockés avec leurs descriptions détaillées. Donc pour éviter toute redondance et duplication inutile, le système est doté d'outils spécifiques de contrôle et de maintien de cohérence au sein de l'entrepôt. Actuellement, on s'intéresse à la version centralisée seulement.

Le système SYFAX comporte une interface de gestion de l'entrepôt qui permet d'effectuer les opérations de stockage, de modification et de recherche des documents, une interface de gestion de profils et recommandations qui permet de gérer les profils utilisateurs ainsi que les recommandations des documents aux utilisateurs. Enfin, un module qui gère la coopération entre serveurs de données et qui permet d'assurer la gestion des notifications et la cohérence. (Voir figure 3).

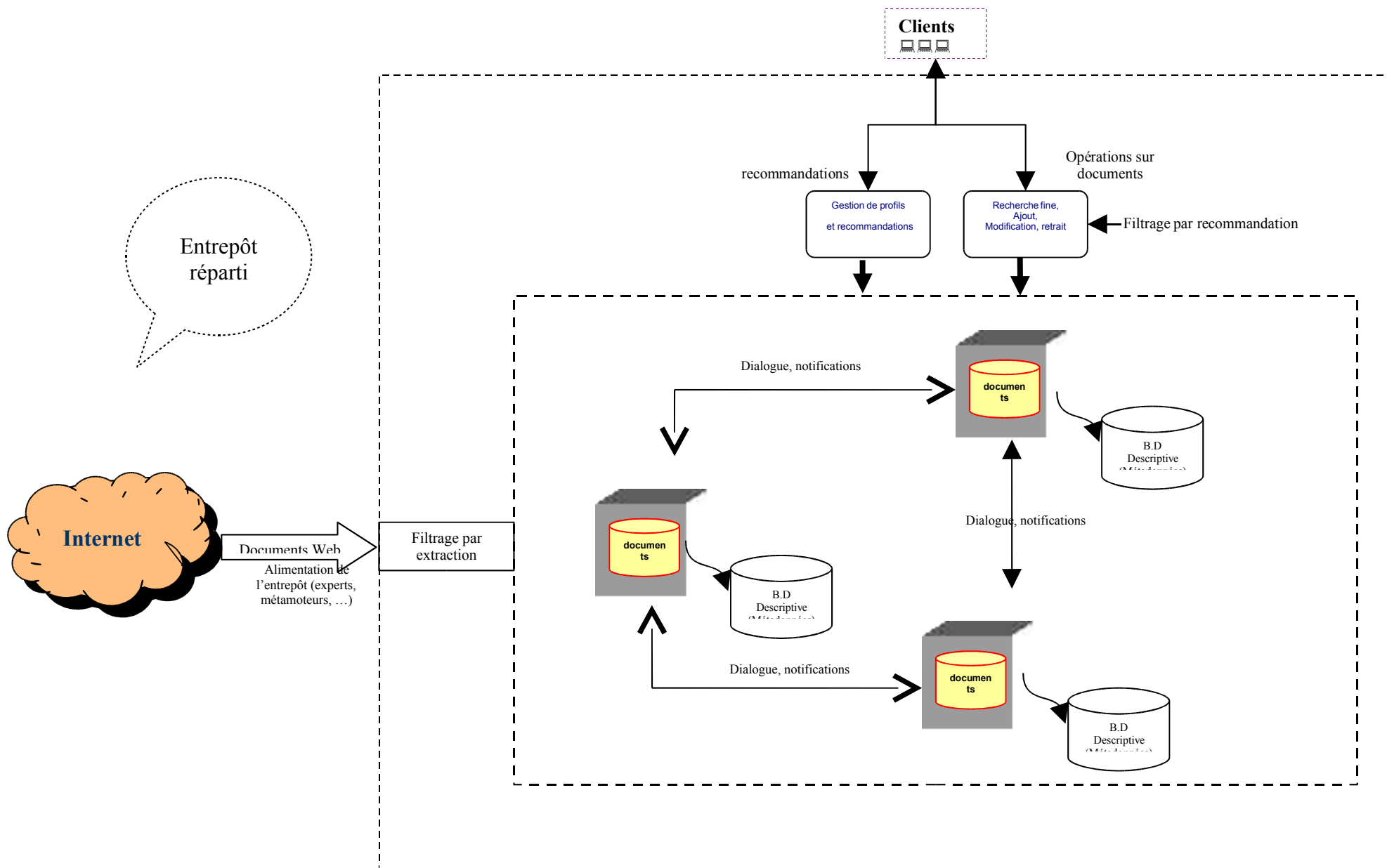


Figure3. Présentation du système SYFAX

III.3. Gestion des l'entrepôt

L'entrepôt est une base de données qui contient des documents se rapportant au domaine de la communauté.

Il est alimenté par deux sources différentes :

- Les auteurs (propriétaires) qui déposent leurs documents (i.e. cours, thèse,...) en spécifiant, à chaque fois, les informations descriptives (ou métadonnées) du document stocké. Il s'agit dans ce cas d'un indexage humain fait par l'auteur du document.
- Le Web, qui est exploité par des experts qui cherchent des nouveaux documents se rapportant aux sujets traités et qui extraient les métadonnées correspondants par des outils d'extraction automatiques ou manuelles.

Parmi les informations descriptives des documents, on peut trouver : le titre du document, le résumé, un ensemble de mots clés, le type du document ainsi que d'autres informations.

Le résumé peut être créé soit manuellement par l'expert ou automatiquement en utilisant des outils de résumé automatique développés par notre équipe [Jawa – Ellouze].

Des opérations sur les documents stockés peuvent être réalisées par les utilisateurs, telles que la recherche, l'ajout et la modification (pour les auteurs des documents), ainsi que l'annotation.

a. La Recherche

Suite à une demande de recherche de documents à partir de mots clés fournis par l'utilisateur, le module de recherche livre une liste de documents répondants aux critères donnés.

La densité est basée sur la fréquence d'occurrence de mots clés dans un document par rapport à la taille du document. Si deux documents contiennent le même nombre d'occurrences, le document le plus petit sera favorisé.

Lors de l'affichage des résultats, et pour une même catégorie de documents, on se base sur :

- Le score (les recommandations des utilisateurs)
- Le nombre de visites à ce site (on détient un compteur de visite pour chaque site, qui s'incrémente chaque fois que le site est visité)

Chaque document recherché est affiché avec son résumé, ainsi l'utilisateur peut accéder au document ou à son résumé.

Le résumé peut être une description textuelle du document recherché, plus éventuellement d'autres informations, une photo, une séquence audio ou vidéo,

b. L'Ajout

L'ajout des documents et des métadonnées se fait par les auteurs qui doivent recevoir en remplissant un formulaire de demande d'hébergement un compte et un mot de passe les identifiants.

Chaque soumission de document serait rigoureusement contrôlée et validée uniquement si elle correspond à un thème traité par le système, empêchant ainsi la présence d'intrus.

c. La Modification

Se sont les informations descriptives qui peuvent être modifiées à travers l'interface par le propriétaire du document qui devra donner son compte et son mot de passe pour pouvoir réaliser l'opération.

Pour le document en question l'utilisateur doit accéder à son système pour supprimer l'ancienne version et mettre la nouvelle. L'interface doit vérifier à la fin de chaque séance les documents qui ont été modifiés (en vérifiant la cohérence de sa base avec le système).

d. Le Retrait

La suppression d'un document se fait par son propriétaire, l'interface doit par conséquent mettre à jour la base de description des documents.

Au moment de la suppression d'un document et tant que tous les autres sites n'ont pas encore cette suppression, le système doit générer un objet qui répondra à toute demande de ce document.

e. Ajouter des mots clés pour un document

Un utilisateur peut proposer un certain nombre (fixer au max) de mots clés pour un document.

Le système n'ajoute effectivement un mot clés à l'ensemble des mots clés du document que si ce mot a été proposé par plusieurs (e.g., 10) clients.

f. Annoter un document

Un utilisateur peut donner son avis sur un document, de ce fait le système propose au client au moment de la lecture d'annoter le document. Il lui propose de 2 types d'annotations :

La première concernant la correspondance du document avec son profil, si oui ce document sera marqué comme convenant au profil de l'utilisateur et des autres ayant les mêmes préférences.

La seconde concernant le point de vue de l'utilisateur sur le document (intéressant / moyen / peu intéressant).

On peut ne donner la possibilité de donner son avis que lorsque le document convient au client.

III.4. Le Filtrage

Dans le système, on rencontre deux types de filtrage, par extraction qui s'applique sur les documents provenant du Web et par recommandation qui s'applique à tous les documents du système.

a. Filtrage par extraction

Ce filtrage est réalisé lorsque des documents sont recherchés du Web, on aura donc à extraire les documents se rapportant au sujet et d'autre part d'extraire les métadonnées décrivant chaque document.

A ce niveau, on peut faire recours à des systèmes de filtrage de résumé automatique et d'extraction de métadonnées.

b. Filtrage par recommandation

Notre système s'oriente vers le modèle dit de filtrage par collaboration (*collaborative filtering*), qui essaye de gagner profit des systèmes de filtrage classique et des systèmes de collaboration.

Le filtrage par collaboration concerne une communauté d'utilisateurs ayant un même intérêt informationnel. Les usagers collaborent entre eux et avec le système. Des humains (les utilisateurs) annotent des documents électroniques, et leurs attribuent un score (le filtrage est ici humain). Une personne qui désire consulter un document peut être informé des recommandations de ses collègues au sujet de ce document.

Lorsqu'un document conviendra au profil d'un groupe il sera envoyé à l'ensemble des individus qui pourront l'annoter ; ces annotations seront comparées entres elles confirmant alors le profil du groupe.

III.5. L'interface de gestion de profils et recommandations

Il assure les opérations suivantes :

- Gestion de profils utilisateurs (création, modification, suppression de profils)
- Recommandations de documents aux utilisateurs

Le système permet de créer des groupes de profils des utilisateurs désirant s'inscrire.

Chaque utilisateur peut recevoir par email, les adresses des documents modifiées ou nouvellement créés, ainsi que la liste des utilisateurs de son profil pour lui permettre de les contacter s'il désire.

Le filtre dispose donc d'informations variées sur un document et peut par exemple proposer à un utilisateur la liste des personnes travaillant sur le même sujet et lui sélectionner les documents qu'elles ont consultés. Si parmi ces documents certains sont pertinents, le filtre pourra trouver l'ensemble des documents ayant les mêmes annotations.

IV. CONCLUSION

Aujourd'hui, personne ne peut nier le besoin de trouver une information pertinente en un minimum de temps dans l'effervescence qui règne aujourd'hui dans le Web. Les portails communautaires tente de répondre à ce besoin. Cependant d'autres services peuvent être ajoutés à ces portails, afin d'améliorer la qualité des informations recherchées.

SYFAX est un portail communautaire qui permet de proposer un système de stockage et d'indexation des documents extraits du Web ou proposés par des auteurs voulant publier leurs documents à la communauté du système.

Il permet en plus, une participation active des membres de la communauté par l'intermédiaire des annotations faites par les utilisateurs et des recommandations envoyés par le système de filtrage aux utilisateurs désirant que le système les notifie lorsqu'une ressource se rapportant à leurs profils apparaisse ou mise à jour.

L'utilisateur du système bénéficiera d'un temps de réponse rapide, une grande disponibilité et un ciblage facile des informations recherchées, puisqu'il accède aux documents qui sont situés dans un environnement Intranet.

Bibliographie

- Trouver l'info sur le Web, Olivier Andrieu, Edition : Eyrolles
- La recherche d'informations, du texte intégral au thésaurus, auteur : Philippe Lefèvre, Edition : Hermes
- Evaluation de systèmes de recherche d'information comportant une fonctionnalité de filtrage par des mesures internes, Thèse de Christine Michel, Université Lumière, Lyon 2, Institut de la communication, Année 1999
- SPECIAL REPORT, Interfaces for Searching the Web, by Marti A. Hearst
<http://www.sciam.com/0397issue/0397hearst.html>
- Etude de l'université de Californie, Berkeley Principe de fonctionnement des méta-moteurs, leurs limites, sélection des plus pertinents d'entre eux. Avril 2000
- Inktomi (1997), <http://www.inktomi.com/press4.html>.
- Recommendations Sharing Forum for Internet Communities
<http://webtools.dyade.fr/pharos/>
- Systèmes de recommandations
http://www.inria.fr/rapportsactivite/RA98/aid/resul_man.html
- E-Commerce Recommendation Applications (2001), J. Ben Schafer, Joseph A. Konstan, and John Riedl, Data Mining and Knowledge Discovery
- Thèse de doctorat en informatique, Un système pour la recherche plein texte et la consultation hypertexte de documents techniques, Quentin Delacroix, Année 1999
- IEEE Communications, Volume 37, Number 1, pp. 116-122, 1999. Searching the Web: General and Scientific Information Access, Steve Lawrence and C. Lee Giles, NEC Research Institute
- Media Metrix (January 1999); <http://www.mediametrix.com/>
- Huberman, B. A. & Adamic, L. A. Evolutionary Dynamics of the *World Wide Web* (1999); <http://www.parc.xerox.com/istl/groups/iea/www/growth.html>
- Dublin Core. The Dublin Core: A Simple Content Description Model for Electronic Resources (1999); <http://purl.oclc.org/dc/>
- Best search engines for finding scientific information in the Web Alexander Lebedev, Moscow State University,
<http://www.chem.msu.su/eng/comparison.html>
- Illustrations pour l'aide à la recherche d'informations sur le Web Brigitte Trousse, INRIA Sophia Antipolis, Action AID