Problématique de la recherche textuelle et Intranet Qualité

Travail de diplôme réalisé en vue de l'obtention du diplôme HES

Par :

Xavier VITALI

Conseiller au travail de diplôme :

Peter DAEHNE, Professeur HES

Mandant :

Xavier BURDET, Professeur HES et Responsable Qualité

Genève, le 24.11.2006

Haute École de Gestion de Genève (HEG-GE)

Filière Informatique de Gestion

Déclaration

Ce travail de diplôme est réalisé dans le cadre de l'examen final de la Haute école Ecole de gestion de Genève, en vue de l'obtention du titre de bachelor en informatique de gestion. L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de diplôme, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de diplôme, du juré et de la HEG.

« J'atteste avoir réalisé seul(e) le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 24.11.2006

Xavier VITALI

« Internet : on ne sait pas ce qu'on y cherche
mais on trouve tout ce qu'on ne cherche pas. »

Anne Roumanoff, humoriste française.

Remerciements

Je tiens à remercier tout particulièrement les personnes m'ayant permis de réaliser ce travail de diplôme pour leur soutien, leurs conseils, leur patience et leur implication.

M. BURDET Xavier, qui m'a donné l'occasion de travailler dans un domaine nouveau et actuel me permettant de mettre à profit les diverses compétences acquises durant ma formation.

M. DAEHNE Peter m'ayant suivi tout au long de ce mémoire, pour sa patience et sa disponibilité.

M. LECLERC Olivier pour ses nombreuses explications à propos du fonctionnement des produits Google ainsi que l'ASP et le XML.

Mme WEINMANN Danielle pour ses éclaircissements à propos du produit MEGA.

M. BERTHERAT-PACCARD Bruno qui m'a fait connaître le produit MondoSearch et conseillé tout au long de la configuration lors de ma phase de test personnel. Les nombreux documents qu'il m'a fourni m'ont éclairci sur le système d'indexation utilisé par MondoSearch et permis de situer cet outil par rapport aux autres produits existants sur le marché.

La société MondoSoft qui m'a mise à disposition une version du logiciel MondoSearch et accordée un délai de 2 semaines supplémentaires pour l'évaluation de la version d'essai.

Sommaire

Ce travail de diplôme est le résultat d'une étude réalisée en vue de l'obtention du titre de Bachelor en Informatique de Gestion à la Haute Ecole de Gestion de Genève. Pour réaliser ce mémoire, je me suis tourné vers M BURDET Xavier, Responsable Qualité à la HEG qui m'a proposé un sujet et M DAEHNE Peter, Professeur HES pour suivre l'évolution de mon travail.

L'objectif à atteindre fut défini dès la fin du mois de septembre et concernait une étude de faisabilité d'un système de recherche à l'intérieur du site Intranet de l'école.

En effet, M BURDET désirait un outil intégré au site lui permettant, ainsi qu'aux autres collaborateurs de la HEG, d'effectuer des recherches par mots clé sur le site Intranet.

C'est la raison pour laquelle nous avons donc axé notre travail de diplôme sur le thème de la recherche textuelle de l'information et le principe d'indexation avec comme objectif de réaliser un prototype fonctionnel sur le site Intranet. Pour cela, nous avons étudié différents outils du marché et choisi l'un d'entre eux pour la réalisation du prototype.

Nous avons opté pour une méthodologie en entonnoir afin de partir sur les principes généraux de la recherche d'informations et le fonctionnement des moteurs de recherche. Puis, nous présentons et sélectionnons des outils candidats en les confrontant grâce à des critères prédéfinis avec le mandant pour enfin choisir l'un d'entre eux. Enfin, je réalise le prototype en détaillant chaque étape de la démarche pour obtenir un système de recherche viable et fonctionnel à l'intérieur du site Intranet Qualité de la HEG.

Nous avons accepté ce sujet car il permettait de mettre en œuvre une démarche de recherche d'information complexe sur un thème qui nous parait aujourd'hui commun : la recherche d'information à travers Internet. De plus, la réalisation d'un prototype fonctionnel prouvant la faisabilité du mandat nous semblait importante. Nous avons donc eu la chance de manipuler des environnements qui nous étaient inconnus jusqu'à présent tels que les stations de travail virtuelles ou MondoSearch.

Ce travail de diplôme nous permet aussi de mettre en œuvre les compétences tout au long de notre cursus au sein de la HEG dans des domaines variés : communication, réseau, programmation, technologie du Web.

Table des matières

Liste des Tableaux

Liste des Figures

Introduction

L'intranet Qualité de la Haute Ecole de Gestion permet aux différents collaborateurs d'accéder aux formulaires, directives et procédures de l'école. En outre, l'information est à constamment à jour, elle se trouve à une seule place et le stockage papier est ainsi supprimé. Les principaux services à disposition des utilisateurs sont la disponibilité et l'échange de documents, le partage des données de l'école et la gestion de la circulation des documents et du travail associé.

Actuellement, l'intranet Qualité de la HEG est un site généré automatiquement par le système MEGA qui permet d'effectuer des recherches par procédure, par processus, par acteur ou par nom de document.

Afin de se repérer dans l'Intranet, il est indispensable de faire le lien entre le site et son contenu afin d'assister l'utilisateur dans sa recherche d'information. C'est ce à quoi s'emploie le moteur de recherche. Dès lors, il apparaît utile pour les utilisateurs de retrouver des documents à partir de mots clé afin d'éviter les pertes de temps liées à la recherche d'information, d'où un net gain en terme de productivité.

D'après le rapport d'activité d'un mandat de la HEG datant de Décembre 2005, il apparaît que depuis la création de l'Intranet Qualité (automne 2004), 57% des collaborateurs interrogés (21 personnes) utilisent rarement ou jamais le site Intranet mis en place. *« Pour toutes ces personnes, la recherche des informations utiles est difficile : la logique n'est pas la leur, il faut trouver une autre structure … ajouter un moteur de recherche»*[1]. En effet, la plupart des personnes interrogées voudraient avoir un accès direct aux objets sans avoir à passer par toute l'arborescence actuelle.

Pour cela, nous allons dans un premier temps définir la problématique de la recherche textuelle en définissant le fonctionnement des moteurs de recherche. Nous analyserons et comparerons ensuite divers outils existant sur le marché. Puis, nous mettrons en place une solution choisie permettant de répondre aux attentes des utilisateurs. Enfin, nous établirons une conclusion nous permettant de mettre en exergue l'utilité d'un système évolué de recherche d'informations au sein des PME

[1] Annexe – rapportMandat.pdf

1. Problématique de la recherche textuelle

La recherche d'information comporte deux étapes essentielles. La première étape est l'indexation préalable des documents et des pages contenants les informations ; celle-ci s'effectue en arrière plan, sans aucune intervention de l'utilisateur. La seconde est la recherche proprement dite, déclenchée par l'utilisateur qui interroge le système.

1.1 Fonctionnement des moteurs de recherche

1.1.1 Principe de fonctionnement

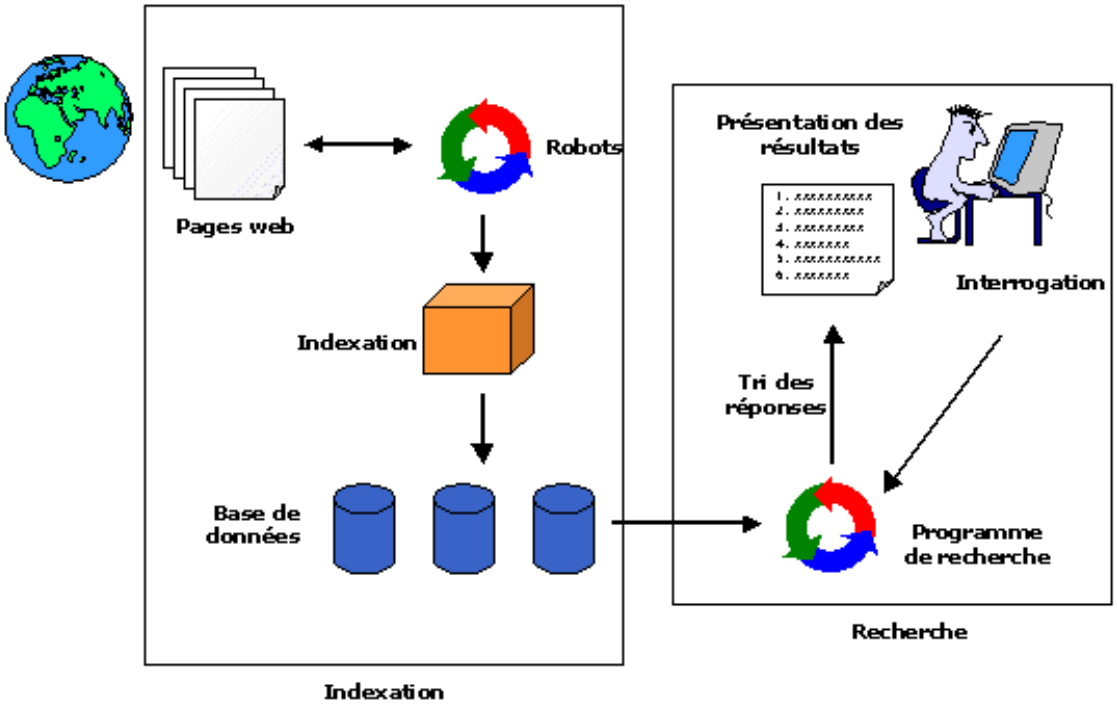


FIG.1 - Fonctionnement d'un moteur de recherche

Un moteur, c'est avant tout un robot (appelé aussi 'crawler', 'worm' ou 'spider') qui récupère chaque jour des millions de pages Web pour en extraire les mots à l'exception des termes vides (mots courant) afin de les stocker dans une base de données qui servira d'index.

Un moteur de recherche, c'est également un logiciel (appelé 'search engine') qui, à partir d'une interrogation formulée par un internaute, va parcourir cet index afin d'identifier les termes recherchés par l'utilisateur et renvoyer une page de résultats selon des critères de pertinence bien définis.

La pertinence des résultats va dépendre de deux facteurs : la syntaxe des requêtes (c'est-à-dire la manière dont l'utilisateur va formuler sa question pour le programme de recherche) et l'ordre d'affichage des résultats. On verra par la suite les différentes manières d'interroger un moteur de recherche, la syntaxe bien particulière qui s'y rapporte ainsi que la manière dont sont triés les résultats.

L'efficacité d'un moteur de recherche peut également être mesurée au travers des informations disponibles sur les résultats (titre de la page, résumé, date de création ou de dernière modification …).

La base de données d'index tient donc un rôle primordial dans la recherche d'information. Or, les documents électroniques apparaissent et disparaissent tous les jours ; il faut donc constamment réactualiser cette base de données. De plus, aucun moteur de recherche ne peut parcourir la totalité des pages disponibles sur le Web en une journée ! (Ce processus prendrait des semaines). Chaque moteur va donc utiliser une stratégie d'indexation propre, certains allant jusqu'à calculer la fréquence de mise à jour de sites.

En résumé, le fonctionnement d'un moteur de recherche se décompose en trois étapes distinctes :

- L'exploration du Web par un robot d'indexation[1] en parcourant récursivement le graphe formé par les hyperliens.
- L'indexation des fichiers récupérés en stockant les mots dans une base de données qui fonctionne comme un dictionnaire inverse.
- L'interrogation de la base de données et le renvoi des résultats en réponse à une requête utilisateur utilisant un des mots clés de cet index.

[1] Glossaire – [R]

1.1.2 Principe de l'indexation en texte intégral

Afin de répondre aux requêtes des utilisateurs qui sont à la recherche d'informations, les moteurs de recherche indexent au préalable (et continuellement) les pages Web. Chaque moteur possède des logiciels appelés robot d'indexation (spiders ou crawlers) qui parcourent de manière continuelle les pages Web.

Les techniques utilisées pour réaliser une indexation en texte intégral d'un document sont similaires à tous les moteurs de recherche. Il existe néanmoins des problèmes liés à la langue naturelle : erreurs de syntaxe liées aux fautes de frappe, emploi d'abréviations ou d'acronymes[1], de synonymes[2], de la polysémie[3], de l'antonymie[4] ou encore de l'anaphore[5].

À ces problèmes liés à la langue naturelle, il faut rajouter la recherche sur un groupe nominal. Par exemple, une recherche sur l'expression « avion à réaction » effectuera les traitements de recherche sur les termes « avion » et « réaction » augmentant ainsi la quantité de résultats de la recherche en réduisant la précision de celle-ci.

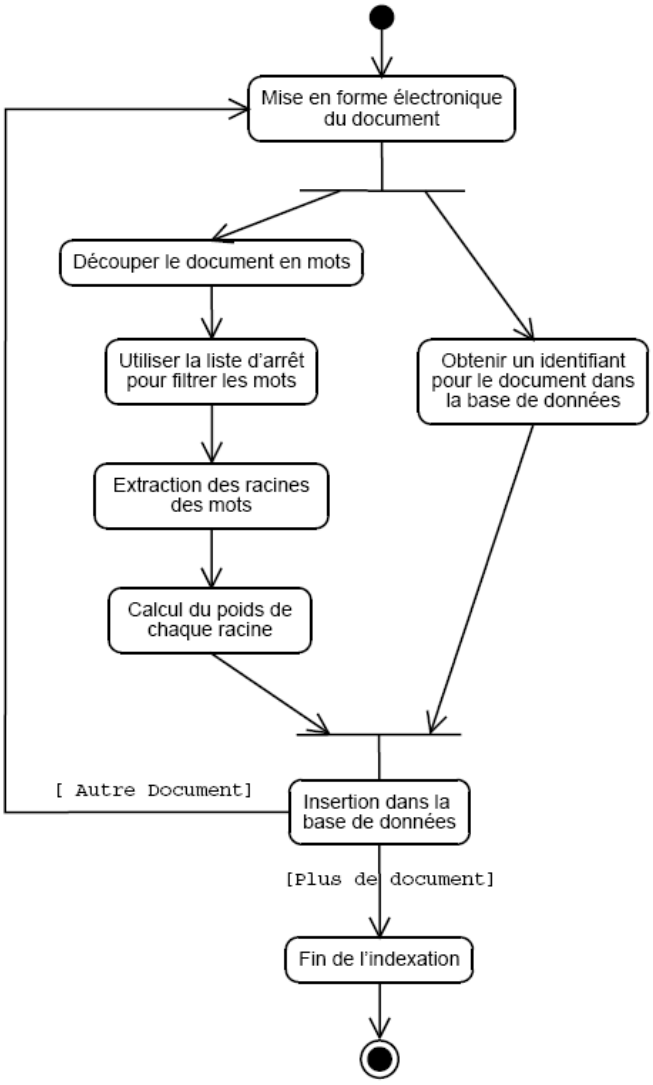


FIG.2 - Modèle général d'une indexation en texte intégral

[1] Glossaire – [A]
[2] Glossaire – [S]
[3] Glossaire – [P]
[4] Glossaire – [A]
[5] Glossaire – [A]

Description des différentes étapes mentionnées dans la FIG.2

- Mise en forme des documents électroniques : cette opération est réalisée en interne à la HEG et fait partie des taches administratives réalisées par le secrétariat ou par le service Qualité de la HEG.

- Découper le document en mot : tous les moteurs cités plus bas prennent en compte la plupart des formats usuels de documents (MS Office, PDF, etc.). Il s'agit ici d'analyser les phrases et de les segmenter en plusieurs mots à partir d'un délimiteur choisi (dans notre cas, l'espace).

- Utiliser la liste d'arrêt pour filtrer les mots : cette phase consiste à supprimer l'ensemble des termes les moins pertinents et est appelée « stoplist ». Cette stoplist contient généralement les articles, les déterminants, les adverbes et les termes trop génériques. Par exemple, on intègrerait le terme « avion » dans une liste de documentation traitant de l'histoire de l'aviation.

- Extraction des racines des mots : cette opération sert à supprimer un maximum de formes fléchies[1] d'un mot afin d'améliorer la pertinence du résultat de la requête. De plus, on réduit ainsi le nombre de termes utilisés pour l'indexation, améliorant de ce fait le temps de réponse et réduisant l'espace de stockage nécessaire à l'indexation.

- Calcul du poids de chaque racine : un poids est attribué à chaque terme employé dans un site en fonction de l'endroit où il a été trouvé. Par exemple, le mot aura plus d'importance s'il fait partie du titre que s'il est situé dans le corps même du texte.

Malgré ces opérations, on peut dégager des conséquences sur une mauvaise indexation des documents :

- Le **bruit** : il s'agit d'un nombre trop important de réponses qui n'ont aucun rapport avec le terme recherché par rapport aux réponses attendues.

- Le **silence** : cela concerne l'absence d'un document pertinent lié à la recherche mais qui n'apparait pas dans la liste des résultats renvoyés par le moteur de recherche.

- La **pertinence** (ou taux de rappel) : elle traduit le caractère plus ou moins exhaustif de la recherche. Elle se calcule de la manière suivante :

Nb_de_docs_pertinents_obtenus / Nb_de_docs_pertinents_de_la_base

[1] Glossaire – [F]

- La **précision** (ou taux de précision) : elle indique dans quel mesure les résultats trouvés sont pertinents par rapport à la recherche effectuée. Elle se calcule de la manière suivante :

Nb_de_docs_pertinents_obtenus / Nb_de_docs_obtenus

Bien sur, l'objectif à atteindre serait une pertinence et une précision qui tendraient toutes les deux vers 100%. Cependant, les moteurs de recherche actuels dépassent rarement les 50%.

1.1.3 Recherche d'information

Lors de la recherche d'information, deux problèmes majeurs peuvent intervenir :

- Aucune réponse : il faut alors élargir le champ de la recherche avec d'autres mots clés en rapport avec le sujet ou bien en enlever certains.

- Trop de réponses : dans ce cas précis, il faut mettre des restrictions au niveau de la recherche en utilisant par exemple les opérateurs booléens ou de proximité, opérateurs que tous les moteurs ne supportent pas. De plus, il existe des fonctions spéciales propres aux moteurs de recherche qui permettent de restreindre encore plus le nombre de réponses renvoyées.

Opérateurs	Exemple	Fonctionnement
	Arbre fruiter	Cherche Arbre ET fruitier
« »	« George Boole »	Cherche l'expression complète sachant que les mots se suivent dans le document.
-	Étude –universitaire	Cherche Étude sans le mot universitaire
+	Étude + universitaire	Cherche Étude ET universitaire sachant que pour certains moteurs de recherche, le ET est implicite (voir exemple ci-dessus).
*	Inter*	La troncature remplace une ou plusieurs lettres d'un mot. Ici, la recherche se fera sur tous les mots commençant par 'inter'.
NEAR / ADJ	Recherche NEAR information NEAR Internet	Les réponses contiendront obligatoirement les mots recherchés mais qui, dans un texte, peuvent être séparés par plusieurs mots.

Tableau 1 : Les opérateurs de recherche d'informations sur le Web

Il est à noter que certains opérateurs ne sont plus gérés par les moteurs de recherche, notamment la troncature et le NEAR-ADJ que l'on ne retrouve plus que sur Altavista.

1.1.4 Avec Google … et le Page Rank

La recherche d'information de la plupart des moteurs de recherche s'effectue de manière similaire en respectant la formule de Salton[1]. Là où les autres moteurs de recherche présentent leurs résultats de recherche d'une manière qui peut sembler anarchique, Google renvoie des résultats trié par pertinence : ceci est le résultat de l'algorithme PageRank[2].

Google attribue une note à chaque document qui détermine directement sa pertinence dans les résultats. Le facteur principal du PR (PageRank) est le nombre de liens qui pointent vers un document. En effet, plus un document est lié par d'autres, plus il doit être pertinent.

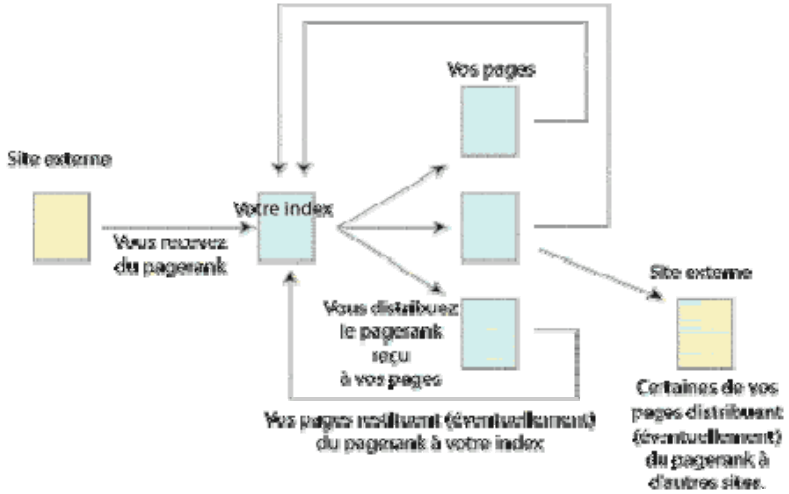


FIG.3 – Schéma explicatif du PageRank

Considérons une page A recevant des liens par les pages T1 à Tn. Le paramètre d est un simple facteur d'amortissement pouvant être ajusté entre 0 et 1 (habituellement, on donne d = 0,85). Enfin, C(A) est défini comme le nombre de liens sortants de la page A. Le PageRank de la page A est donc :

$$PR(A) = (1-d) + d(PR(T1) / C(T1) + \ldots + PR(Tn) / C(Tn))$$

En pratique, on pourrait se demander comment ce processus itératif peut faire bon ménage avec la modification permanente de l'index, (influence sur les autres documents, cohérence de calcul dans tous les Datacenter). Cela s'effectue en réalisant le calcul du PR et la synchronisation des serveurs environ une fois par mois : la Google Dance.

---

[1] Glossaire – [S]

[2] Glossaire – [P]

Analyse de la formule :

PR(A)	Le PageRank de la page A
PR(Tn)	Le PageRank de la page Tn
C(Tn)	Le nombre de lien sortant de la page Tn
d	Tous les liens sont additionnés mais pour en limiter l'importance, le total est multiplié par un coefficient d'amortissement (0,85)
1-d	Permet de garantir que la moyenne des PageRank de l'ensemble des pages Web sera égale à 1

Exemple :

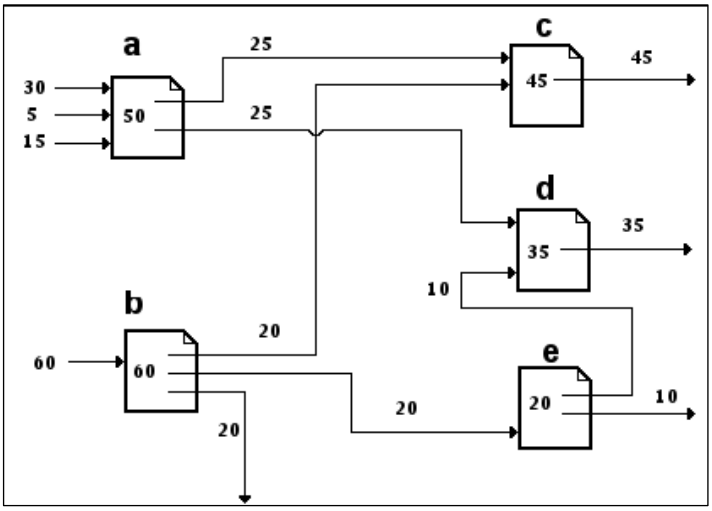


FIG.4 – Calcul simplifié du PageRank de plusieurs pages interconnectées

Dans la figure ci-dessus, les pages 'a' et 'b' ont 3 et 1 liens entrants, de poids respectivement 30, 5 et 15, et 60. La page 'a' a donc un PageRank de 50 et la b un PR de 60. Les 2 liens sortant de 'a' ont un poids de 50/2=25 que se partagent les pages 'c' et 'd'. La page 'b' a trois liens sortant équivalent à 60/3=20 que se partagent les pages 'c', 'e' et une autre page quelconque.

La page 'c' a donc un PR de [(PR(a)/2) + (PR(b)/3)] = 25 + 20 = 45. La méthode de calcul est la même pour les pages 'd' et 'e'.

1.2 Les modes de recherche

1.2.1 Recherche par navigation arborescente

La page d'accueil de l'intranet Qualité permet d'accéder à 3 arbres :

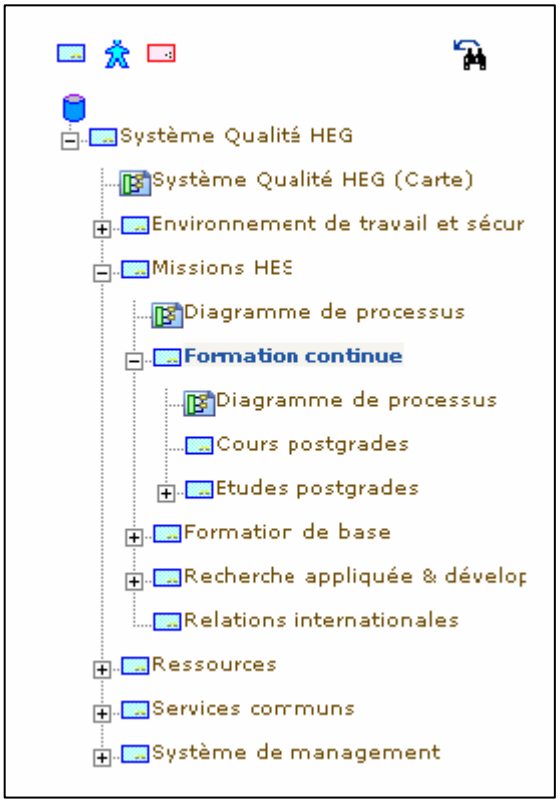
- L'arbre des processus.
- L'arbre des procédures.
- L'arbre des acteurs.

La recherche s'effectue par navigation dans des menus successifs.

L'information est structurée, organisée en plan de classement. Les informations secondaires se situent le plus bas dans la hiérarchie de l'arbre.

La démarche est systématique : on va toujours du général au particulier.

Exemple de recherche par navigation arborescente : classifications documentaires (CDU, annuaires thématiques du Web, page d'accueil d'un site Web). On retrouve aussi toutes sortes de représentations de l'architecture d'un arbre : répertoire, sous répertoire, etc. … ou même le fil d'Ariane dans les sites de E-Commerce qui utilisent ce principe.



1.2.2 Recherche par navigation hypertextuelle

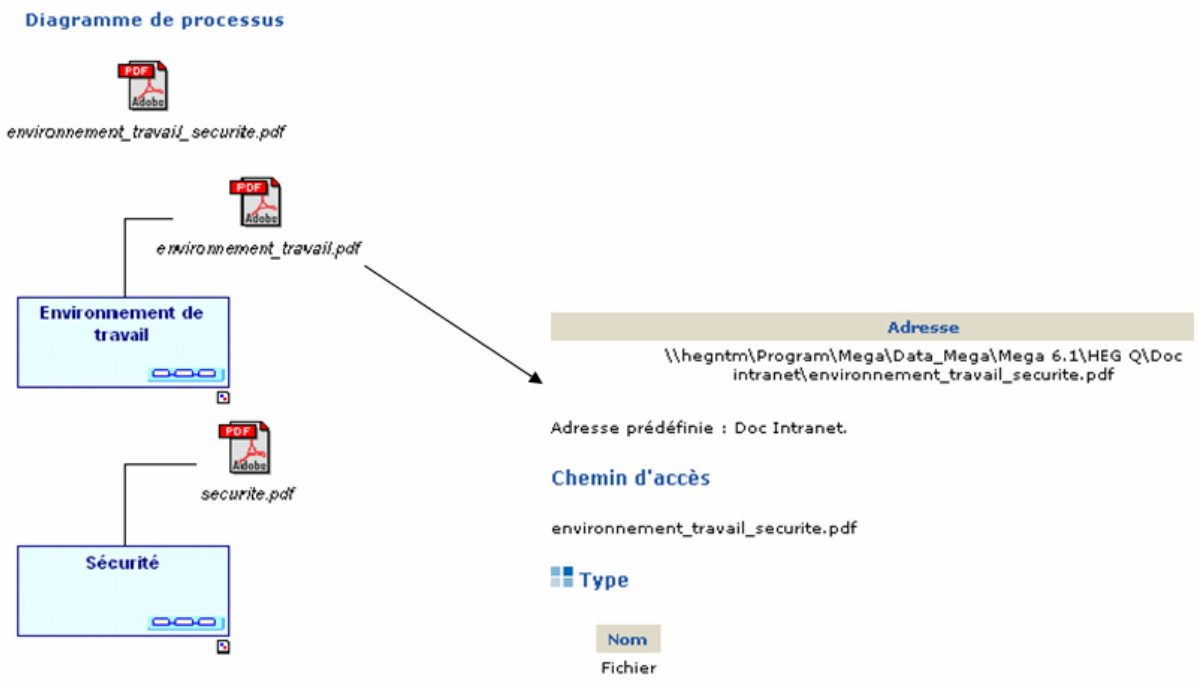


FIG.5 – Schéma de navigation hypertextuelle

Par définition, c'est l'exemple typique de navigation présente sur les sites Web. On retrouve aussi ce type de navigation dans les hypertextes sur CD-ROM. On peut comparer celle-ci à une navigation dans un réseau de nœuds et de liens créés par association entre des mots et des documents.

Cela permet, comme le montre la figure ci-dessus, d'accéder à de l'information brute (fichiers PDF, documents MS Office …) par simple clic ou bien d'accéder à d'autres pages de l'Intranet Qualité.

2. Les outils existants

2.1 Google Search Appliance (État de Genève)

2.1.1 Principe

L'État de Genève a acquis en 2005 un serveur Google Search Appliance pour la somme de 80'000.- CHF et est depuis client de Google. Le serveur Appliance fonctionne donc en interne à l'État.

Le Google Search Appliance se compose d'un Hardware et d'un logiciel conçus pour mettre à la disposition des entreprises la puissance du service de recherche de Google.

Le GSA permet aux utilisateurs finaux de rechercher au sein de leur Intranet de la même manière que sur le Web traditionnel. Ainsi, on retrouve les mêmes fonctionnalités de recherche qui sont :

- le classement des résultats.
- le résumé dynamique des pages.
- le regroupement des résultats.
- le correcteur orthographique automatique.
- les pages en cache.
- le tri par date, etc.

L'utilisateur n'est donc pas plongé dans un univers qui lui est totalement inconnu et cela réduit considérablement le temps de prise en main.

L'administration du GSA se fait via l'interface Web. Il peut donc être déployé rapidement et être géré par une seule personne n'importe où dans le monde et en plusieurs langues différentes.
L'administrateur peut en outre :

- segmenter les index pour fournir des résultats différents selon les utilisateurs (collections).
- appliquer des filtres sur les langues, les types de fichiers, les balises Meta.
- définir des synonymes (recherche de l'utilisateur sur « Natel », le GSA propose aussi de rechercher pour l'expression « téléphone portable ».
- d'établir des correspondances entre des URL et une liste de mots-clés.
- de configurer la présentation des pages grâce à des feuilles XSLT.

2.1.2 Caractéristiques

Il existe trois modèles du Google Search Appliance chacun ayant une licence d'utilisation différente, notamment au niveau de la quantité de fichiers indexables. Par exemple, le premier modèle GB-1001 couvre 500'000 documents (30'000€ - 50'000€) et traite 300 requêtes/min. Les modèles GB-5005 et GB-8008 explorent respectivement jusqu'à 5 millions et 15 millions de documents.

Les modèles GSA permet de détecter automatiquement un cinquantaine de langues. Les fichiers indexés doivent respecter une taille maximum : 2,5Mo pour les fichiers HTML et 30Mo pour les 220 autres formats supportés (MS Office, PDF …).

2.1.3 Mise en place[1]

D'après le schéma du réseau global genevois et l'interview de M. INEICHEN, les modifications à apporter au système pour indexer les fichiers de l'Intranet Qualité sont minimes et rapidement réalisables.

Pour cela, une phase de préparation est nécessaire. Celle-ci consiste à produire une page d'accueil HTML[2] contenant les liens hypertextes vers les différents Intranet Qualité des écoles faisant parties des HES-SO//Genève[3]. En effet, tous les fichiers de HEGNTM (noté E) doivent être atteignables depuis la racine et cela entraine une arborescence du type :

- homepageIntranet.htm
    - lien vers l'Intranet Qualité de la HEG
    - lien vers l'Intranet Qualité de la HETS
    - etc. …

Il faut donc modifier le firewall de la HEG (noté B) pour HEGNTM et définir comme point d'entrée sur le GSA la page d'accueil citée précédemment de la manière suivante : http:// hegntmhegntm.ceti.etat-ge.ch/homepageIntranet.htm[4].

---

[1] Annexe – [A2.1]

[2] Annexe – [A3.1]

[3] Glossaire – [H]

[4] Glossaire – [F] (FQDN)

La deuxième phase est la configuration du Google Search Appliance située dans la DMZ[1] de l'État de Genève. En effet, celui-ci doit être configuré de manière à pouvoir explorer les fichiers contenus sur le serveur HEGNTM. Ce procédé s'effectue en saisissant l'URL de la page d'accueil HTML comme le montre la figure ci-dessous.

FIG.6 – Configuration du Google Search Appliance

Une autre phase de configuration du GSA consiste à indiquer le chemin que peut suivre le robot d'indexation à partir du point d'entrée. Dans notre cas, nous mettrions par exemple : http://hegntm.ceti.etat-ge.ch/homepageIntranet.htm

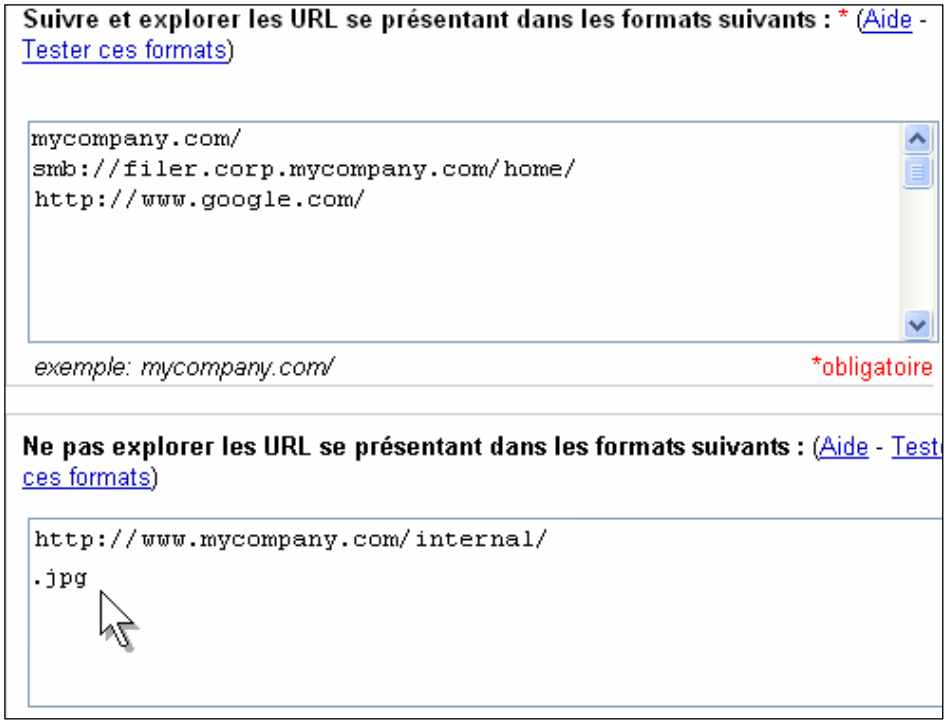


FIG.7 – URLs d'exploration du Google Search Appliance

[1] Glossaire – [D]

Il existe bien la possibilité d'interdire certains fichiers contenus dans le serveur HEGNTM afin d'empêcher certains utilisateurs d'y avoir accès ou bien parce qu'ils représentent un nombre trop important de fichiers pouvant augmenter le bruit[1] dans la page de résultats. Dans la figure FIG.7 ci-dessus, on s'aperçoit que les fichiers contenus dans le répertoire nommé « internal » ainsi que tous les fichiers « .jpg » seront exclus de la page de résultats.

De même, l'administrateur a la possibilité de paramétrer l'accès au contenu en attribuant des droits (collections) à certains utilisateurs pour certains répertoires.

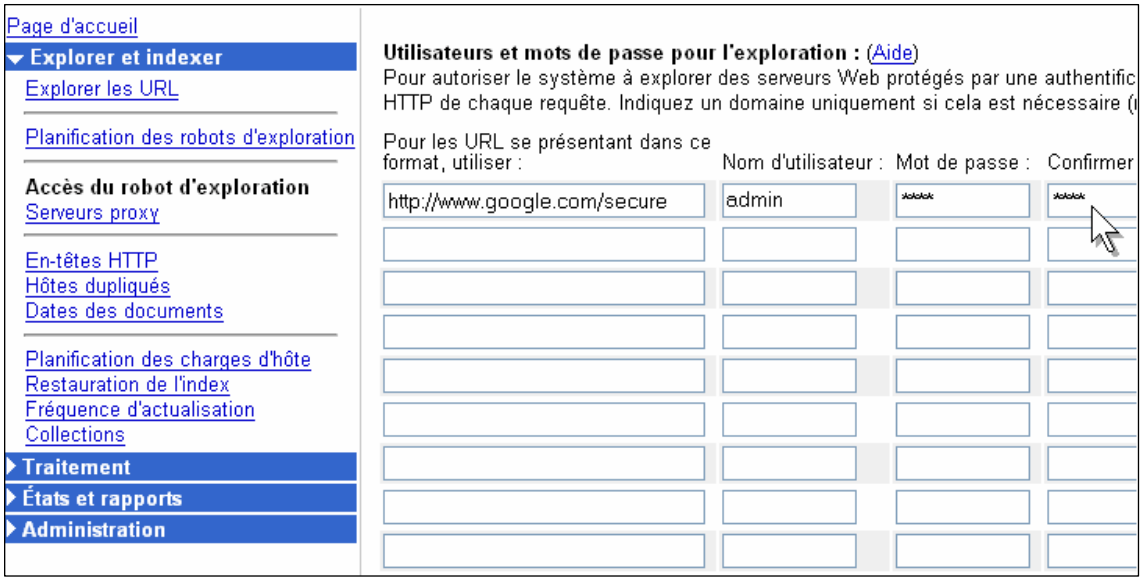


FIG.8 – Limitation de l'exploration pour les utilisateurs

2.1.4 Inconvénients

Il convient de rappeler que le Google Search Appliance est situé au niveau de la DMZ de l'État et est donc la propriété de celui-ci. Ainsi, il y aurait une certaine dépendance matérielle des HES-SO//Genève avec l'État ce qui entrainerait l'indisponibilité des divers Intranets en cas de panne du GSA. De plus, l'État de Genève serait en droit de demander une participation financière pour la « location » du GSA. Rappelons que le GSA de l'État de Genève peut indexer un million de fichiers et que le nombre de fichiers des Intranet des HES-SO//Genève est estimé à 9000 fichiers, soit 0,009% de la capacité du GSA.

[1] Glossaire – [B]

Le GSA à une durée de vie limitée à 2 ans. Au-delà, le support de Google n'est plus assuré. Cela peut devenir un réel danger pour la gestion des nouveaux formats de fichiers qui ne seront donc plus pris en compte sans financer une mise à jour (matérielle ou logicielle).

Il sera à notre avis difficile de faire indexer les fichiers des Intranets des HES-SO//Genève des Cantons voisins (Vaud, Neuchâtel, …) par l'État de Genève sans soulever des problèmes juridiques, administratifs, financiers et techniques (problème de l'interconnexion des réseaux).

2.2 Google Mini

2.2.1 Principe

Il pourrait aussi être intéressant de disposer d'un serveur propre afin de stocker l'Intranet Qualité de la HEG mais aussi l'ensemble des Intranets Qualité des autres Haute Écoles Spécialisés de l'état.

Le principe du Google Mini est le même que celui du Google Search Appliance dans la majorité des cas. Les différences notables se retrouvent au niveau des caractéristiques physiques et techniques.

Tout comme pour son homologue Appliance, le Google Mini permet en outre d'analyser le comportement des utilisateurs en matière de recherche et d'établir de rapports détaillés sur :

- Le nombre total de recherches et de requêtes uniques
- Le nombre de recherches effectuées un jour donné
- Le nombre moyen de recherches à différentes heures de la journée
- Les 100 mots clés et requêtes les plus employés

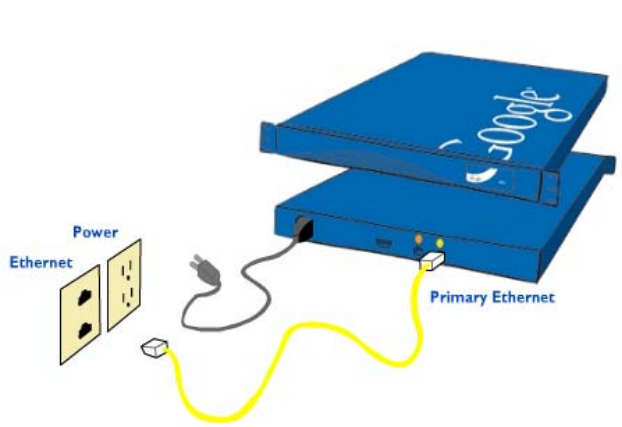
Nous avons vu plus haut qu'un des problèmes liés à la recherche d'information concerne l'orthographe et la subjectivité de l'utilisateur. Le problème peut-être résolu grâce aux suggestions de variantes orthographiques.

2.2.2 Caractéristiques

Les nouvelles versions de Google Mini permettent d'indexer de 50'000 à 300'000 documents pour un prix de licence variant de 1'995€ à 8'995€. Il convient de préciser qu'un document équivaut à une adresse URL[1]. Le nombre de formats traités par le Mini est le même que celui du GSA, en effet, les protocoles de reconnaissance de type de fichiers utilisés sont identiques.

Dans les entrailles du Mini, on retrouve 2 Pentium III-S à 1,26 GHz, un disque dur de 120 Go et 2GB de SDRAM (4x512 MB).

2.2.3 Mise en place



Tout comme le Google Search Appliance, le Mini est installable et configurable en quelques heures. Il ne suffit que d'une prise de courant et d'une connexion Ethernet[2] pour le faire fonctionner.

Google Mini peut être facilement installé dans un rack de centre de données standard. Une fois le système mis en place dans le centre de données, on peut le connecter au réseau à l'aide d'un câble Ethernet, définir les paramètres de réseau standard (adresse IP, par exemple), puis configurer et gérer le système à l'aide d'une interface Web similaire à celle du Google Search Appliance.

Le système Mini accède aux documents du réseau interne. Il est par conséquent inutile de le connecter directement à un serveur. En outre, le nombre de serveurs qu'il peut explorer n'est pas limité.

Les démarches de l'administrateur sont les mêmes que sur le GSA : connexion, paramétrage des adresses URL que le Google Mini doit parcourir et éventuellement restriction des URL à parcourir et création de répertoires protégés pour un type d'utilisateur.

---

[1] Glossaire – [U]

[2] Glossaire – [E]

Il convient d'ajouter que la durée nécessaire à l'exploration initiale dépend de plusieurs facteurs :

- Le nombre et la taille des fichiers à indexer.
- La vitesse des serveurs à parcourir.

Cependant, les Google Mini et Appliance sont capables d'explorer la totalité des fichiers en quelques heures.

Enfin, il existe deux systèmes de mise à jour des index : le mode d'exploration continu et le mode d'exploration planifié. Les robots d'indexation Google sont capables d'analyser les sections d'un site qui évoluent souvent et les explorent donc plus fréquemment. Il est aussi possible pour l'administrateur de planifier une exploration totale des fichiers (par exemple le soir afin d'avoir une version à jour des index le lendemain matin).

2.2.4 Inconvénients

L'achat d'une licence du Google Mini ne correspond pas à la politique de centralisation de l'informatique voulue par l'État de Genève dans un souci de pérennité et d'économies. Il convient de rappeler que les HES-SO//Genève sont totalement indépendantes au niveau du choix des outils pédagogiques.

De plus, l'ensemble des HES-SO//Genève partagent les frais de chaque nouvel achat de serveur afin de diviser le coût (par 7). Ainsi, le serveur HEGNTM va être remplacé courant Octobre et l'achat d'un Google Mini n'est donc pas d'actualité avant 3 ou 4 ans au vu du budget déjà alloué.

Au niveau de l'assistance, Google Mini n'est « assuré » que pour une durée d'un an (au lieu de 2 pour le Google Search Appliance). L'achat d'une deuxième année d'assistance est possible pour la somme de 1'043,37€ (soit 1'657 CHF).

Le personnel de la HEG doit être formé et la HES-SO//Genève doit prévoir un budget d'exploitation et de mise à jour de la base d'index.

2.3 MondoSearch

Fondé en 1998, Mondosoft est une société qui fournit des logiciels d'indexation, de recherche et d'étude comportementale pour les sites Internet et Intranet.

La version d'évaluation téléchargée de MondoSearch permet d'indexer jusqu'à 50'000 documents en utilisant le serveur IIS de Microsoft. La prise en main de base pour de premiers résultats concluants a été effectuée en une heure seulement sans connaissance particulière de l'environnement de travail. Par la suite, de nombreux paramétrages sont réalisables afin de personnaliser l'ensemble du module de recherche.

2.3.1 Principe[1]

MondoSearch est un outil de recherche avancé et multi-langues. Il utilise pour cela un langage formel propriétaire nommé MQL (Mondosoft Query Langage). Dans le prototype mis en place, MondoSearch se sert de l'IIS[2] de Microsoft pour indexer les sites Internet. Lors de l'indexation (crawling), une base de données est créée afin de constituer l'index des mots provenant de l'ensemble des fichiers du site Web.

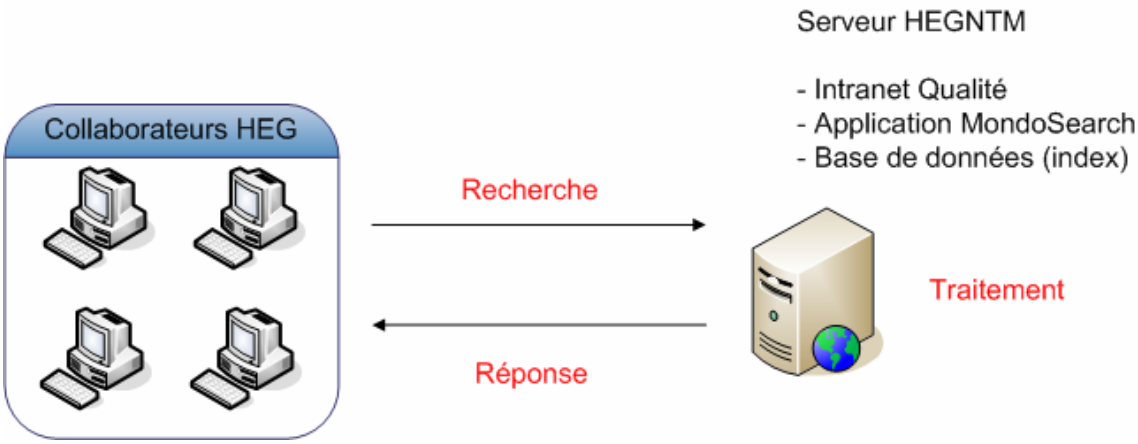


Figure 9 – Principe de Fonctionnement de MondoSearch

[1] Annexe – [A5.1]

[2] Glossaire – [I]

2.3.2 Caractéristiques

Construit sur .NET, MondoSearch est pourvu de modèles de recherche pour.NET, d'un connecteur pour Microsoft CMS[1] (gestion de contenu web) et d'un kit pour une exploitation sous la forme de service web. MondoSearch permet l'indexation en texte intégral de fichiers Office, PDF, HTML et Flash. L'interface du logiciel est en anglais.

La fiche technique de l'éditeur annonce un minimum de 200'000 pages indexables. En ce qui concerne le serveur, celui-ci doit être composé au minimum d'un processeur 4Ghz Intel Pentium, de 1 GB RAM et disposer de 15 GB de stockage disque libre. D'un point de vue logiciel, Windows Serveur 2000/2003 ou Windows XP avec IIS 5.1

La version d'évaluation disponible sur le site officiel[2] est utilisable durant 30 jours après l'installation et permet l'indexation de 50'000 fichiers. Pour utiliser le logiciel, une clé de licence et un code PIN sont nécessaires.

2.3.3 Inconvénients

L'inconvénient majeur de MondoSearch est son prix. Après avoir demandé un devis auprès de l'éditeur, il s'avère que cette solution est disponible à 9'095€ par processeur (soit 14'500 CHF) incluant le support technique et les mises à jour de la première année. Au-delà, le coût du support et des mises à jour sont calculées à partir du prix total de la licence (20%), soit 1'515€/an (2'400 CHF).

En comparaison, le Google Mini est disponible à 2'900€ pour une indexation de 100'000 pages.

Il convient d'ajouter que MondoSoft propose à leurs clients l'option Behavior Tracking (à partir de 10'000€, soit 16'000 CHF) qui permet d'analyser le comportement des utilisateurs pour optimiser le paramétrage du moteur. Ce module restitue tout d'abord des statistiques classiques, telles que les mots-clés les plus utilisés, le nombre de clics, ou les recherches qui n'ont donné aucun résultat et celles qui n'ont engendré aucun clic. Cette option est relativement chère et difficile à configurer.

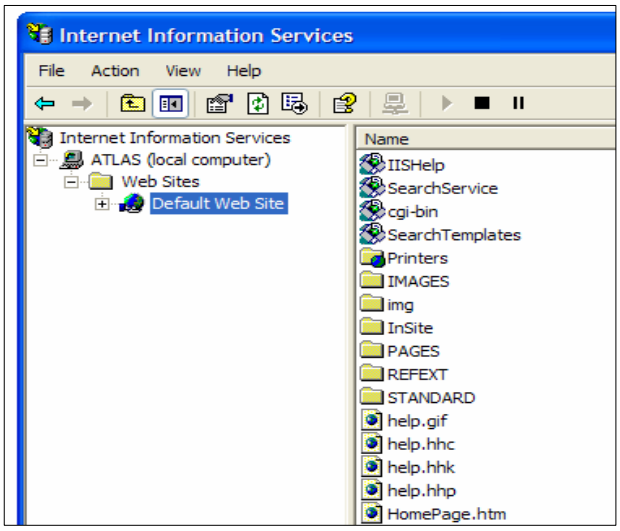
---

[1] Glossaire – [C]

[2] www.mondosoft.com

2.3.4 Fonctionnement

Dans un premier temps, il faut s'assurer de la présence de l'IIS[1] sur le système d'exploitation. Le site Intranet est ajouté au dossier Web Sites de l'IIS et démarré. Il faut bien sûr s'assurer d'avoir défini la page d'accueil du site dans les préférences (click droit sur Default Web Site). On peut observer sur l'image ci-contre le dossier de l'Intranet Qualité de la HEG ainsi que la page d'accueil HomePage.htm.



La version d'évaluation (30 jours) de MondoSearch est téléchargée depuis le site officiel à la suite de quoi MondoSearch envoie un mail contenant la clé de licence. Ce mail contient aussi un lien vers un Code PIN nécessaire lors de la phase d'installation. Lors de l'installation, il suffit de nommer le DNS de la machine (ici, ATLAS) et de renseigner les informations fournies par MondoSoft.

Une fois installée, on peut lancer le moteur de recherche MondoSearch en tapant l'URL suivante : http://atlas/Insite/

L'administrateur arrive sur une page de configuration qui lui permet de modifier les options d'indexation, de la page de recherche et de résultats. Ensuite, il peut lancer l'indexation du site (Crawling) afin de créer la base d'index. Ce processus peut prendre de plusieurs minutes à plusieurs heures suivant la taille du site.

Une fois la base de données publiée, l'administrateur peut tester le moteur de recherche et configurer à souhait les paramètres d'utilisation et d'interface. Enfin, la page de recherche peut être mise à la disposition des utilisateurs qui emploieront le moteur de recherche sur la part serveur dédiée.

[1] Glossaire – [I]

2.4 Solution Open Source

2.4.1 Principe

Pour la mise en œuvre d'une solution Open Source, nous avons choisi le moteur MnoGoSearch, principalement en raison de la popularité intéressante dont il jouit sur la toile. Tout comme les logiciels Google et MondoSearch, il permet d'indexer les documents d'une site.

MnoGoSearch est constitué de deux parties :

- L'indexeur qui passe en revue récursivement les sites Web ou les fichiers locaux et enregistre les méta-données ainsi recueillies dans une base de données MySQL pour optimiser les recherches ultérieures.
- Le moteur de recherche est accessible par une interface Web.

Il prend lui aussi en compte les principaux formats de fichiers utilisés sur le Web (HTML, PDF, MS Office) et, grâce à des modules spécifiques, permet d'utiliser son moteur de recherche pour des sites Internet et Intranet.

MnoGoSearch peut traiter des requêtes de recherche par mots clé, par date ou type de fichier et gère le multilinguisme comme ses concurrent payants. Il est capable d'indexer des millions de documents si le serveur est suffisamment puissant pour supporter une telle charge.

Il convient de noter que 2 versions de MnoGoSearch existent :

- MnoGoSearch Lite : cette version gratuite (Linux) pour environnement de travail réduit permet d'indexer entre 1'000 et 3'000 documents.
- MnoGoSearch Pro : permet en outre l'administration à distance et le travail avec les environnements MS Access et MySQL.

2.4.2 Caractéristiques

MnoGoSearch est un moteur d'indexation full-text et de recherche Open Source écrit en C. Cet outil est utilisable sous la plupart des environnements de travail (Windows, Mac OS, GNU/Linux).

La licence publique générale GNU est gratuite pour Linux et les OS de type Unix mais la version Windows est payante.

- MnoGoSearch Lite pour Windows : 99$ (soit 125 CHF)
- MnoGoSearch pour Windows – Edition Standard : 995$ (soit 1'250 CHF)
- MnoGoSearch pour Windows – Edition MSSQL : 1’750$ (soit 2'200 CHF)
- MnoGoSearch pour Windows – Edition Oracle : 995$ (soit 25'000 CHF)

Le support s’effectue par mail uniquement et coûte 2’000 $/an, soit 2'500 CHF/an. Les mails reçus par les développeurs de MnoGoSearch sont simplement considérés comme prioritaires et théoriquement traités en un jour ouvrable. De nombreux services sont disponibles pour la personnalisation du système de recherche (installation, configuration) si le type de support le plus onéreux a été choisi.

Il existe toute une série de fonctions PHP disponibles sur Internet et utilisables avec MnoGoSearch sous MySQL, naturellement supporté par PHP. Pour utiliser ces fonctions, il est nécessaire d'installer les dernières versions du moteur d’indexation et de recherche. C’est en effet une API complète pour PHP (fonctions commençant par « udm_ » qui permet de s’interfacer avec moteur de MnoGoSearch.

2.4.3 Fonctionnement

L’installation de MnoGoSearch s’effectue en lançant l’exécutable fourni après téléchargement depuis le site officiel. La mise en place de cet outil est lourde : il faut installer la base de données SQL[1], installer les drivers ODBC[2], créer les tables (pour la version Lite). La création de la base de données pour le stockage des fichiers indexés est automatique dans la version pro du logiciel.

Une fois le produit installé, il faut configurer les paramètres de recherche, d’indexation (pour l’ensemble des types de fichiers) et de pertinence (afin d’optimiser les temps de traitement) et créer les interfaces PHP/mySQL de recherche et d’affichage des résultats.

---

[1] Glossaire – [S]

[2] Glossaire – [O]

2.4.4 Inconvénients

Comme dit plus haut, la version gratuite ne concerne que les utilisateurs du système d'exploitation Linux. Aussi, toutes les versions sous Windows sont payantes : de 125 CHF (1'000-3'000 documents) à 25'000 CHF pour plusieurs millions de documents et un support complet.

Il est indispensable de télécharger un module supplémentaire pour traiter les documents PDF et de paramétrer manuellement les parsers[1] utilisés, tâche pas forcément aisée pour un administrateur non initié à la programmation.

La version Lite ne comprend pas de support SQL si l'on rencontre un quelconque problème de développement et il n'existe qu'une communication par mail (ou forum) entre les utilisateurs et le service développement de MnoGoSearch.

Enfin, nous estimons qu'il sera nécessaire de former une équipe apte à utiliser ce genre de produit libre, tant sa configuration et sa maintenance apparaissent ardues.

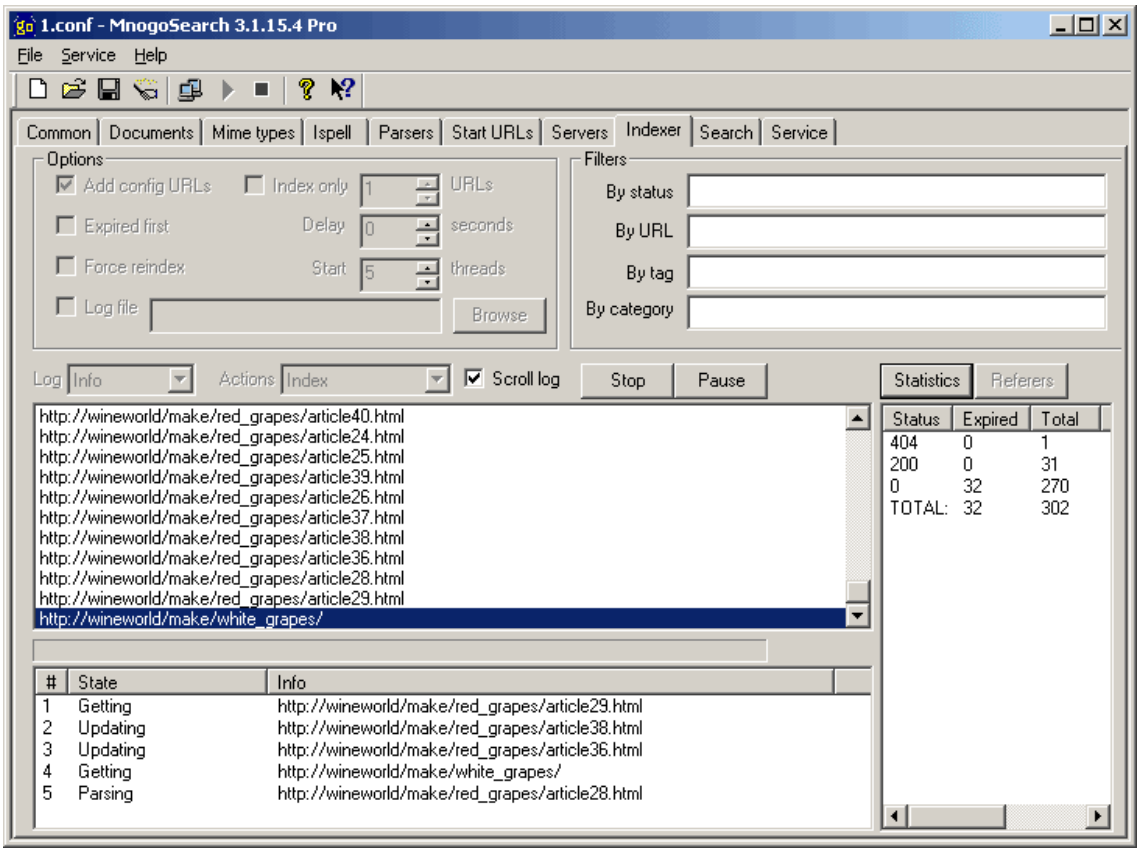


Figure 10 – Écran de configuration de MnoGoSearch

[1] Glossaire – [P]

3. Comparaison des solutions

Les outils candidats ont été sélectionnés selon leur popularité et parmi un choix large de mise en place (nouveau serveur, logiciel à mettre en place sur le serveur, modification interne au niveau du site). Ainsi, on peut couvrir les différentes possibilités d'insertion de module de recherche.

3.1 Tableau

CRITERES	Poids (/5)	GSA	Mini	Solution Open Source	MondoSearch
Coût financier	**2**	2	1.5	2	1
Installation, Maintenance	**3**	2	1.5	0.5	1
Rapidité indexation	**2**	2	2	1	2
Contrainte serveur	**4**	4	4	4	4
Nb docs indexables	**2**	2	2	1	2
Interface, ergonomie	**3**	3	3	1	2.5
Paramétrage de la page des résultats	**3**	2.5	2.5	2	3
Présentation des résultats	**4**	3.5	3.5	2.5	3
Rapidité de la recherche	**5**	5	5	3	5
Distinction maj/min, accents	**1**	1	1	0.5	1
Recherche booléenne et de proximité	**2**	0.5	0.5	0	1.5
Classement des résultats	**5**	4.5	4.5	2.5	4.5
Recherche avancée (filtrage, outils lang.)	**2**	2	2	1	2
TOTAL	**130**	**118**	**115.5**	**74.5**	**113**

Tableau 2 – Tableau de comparaison des outils

Le tableau de comparaison ci-dessus permet une analyse discriminante des outils existants sélectionnés. Le choix des critères a été élaboré en nous positionnant en tant qu'administrateur, puis utilisateur final. Par la suite, une interview de M. BURDET (administrateur/utilisateur de l'Intranet Qualité de la HEG) nous a permis d'affiner les poids de chaque critère.

La dotation totale en points de chaque logiciel doit permettre de choisir l'outil le plus adapté à la problématique. Cependant, le choix de la solution reste à la charge de l'équipe d'administration de la HEG.

3.2 Critères et interprétation

3.2.1 Définition des critères :

- Coût financier : L'ensemble des HES-SO//Genève est constitué de 7 Écoles partageant les coûts d'achat de matériels leur permettant de mettre en place des outils de communication comme l'Intranet Qualité. Aussi, le critère financier apparaît peu important car toujours divisé par 7. Dans le tableau ci-dessus, plus la note des outils testés est haute, moins le coût est important.

- Installation, maintenance : La simplicité des procédures d'installation est l'un des facteurs de réussite des logiciels et les administrateurs sont le plus souvent guidés par leur automatisation. En revanche, le paramétrage et la configuration des outils varient en fonction de leur complexité et du degré de granularité désiré par l'administrateur. Aussi, il faut s'assurer que l'outil choisi ne nécessite pas une prise en main laborieuse. Au niveau technique, il faut aussi veiller à ce que la maintenance fournisseur soit suivie (patch en cas de bug, par exemple). On peut se demander aussi comment réagira le système en cas de surcharge dû à une utilisation simultanée de plusieurs collaborateurs. De même, le protocole utilisé lors de l'indexation vient s'ajouter aux questions techniques. Enfin, le paramétrage des ports doit être possible en cas d'évolution de l'utilisation du serveur Web (passer du port 80 actuel au port 8003, par exemple.

- Rapidité indexation : la rapidité d'indexation apparaît peu importante car le processus d'indexation des fichiers (mise à jour de l'index) peut-être réalisé la nuit en dehors de toute utilisation. En fonction des outils, l'indexation peut-être configurée (date et heure de début, sélection des dossiers à traiter). D'autres

caractéristiques viennent modifier la durée de l'indexation : mémoire vive, processeur, nombre et taille des fichiers.

- Contrainte serveur : l'achat d'un nouveau serveur n'est pas à l'ordre du jour car les HES-SO//Genève viennent d'acquérir début Octobre un nouveau serveur de fichiers remplaçant HEGNTM. Toutefois, les différentes solutions proposées ne nécessitent pas l'achat d'un nouveau serveur (excepté l'achat du robot d'indexation pour les outils Google). Cependant, d'autres outils peuvent nécessiter l'achat d'un serveur spécifiques d'où la forte pondération.

- Nb docs indexables : Le nombre de fichiers indexables n'est pas un critère déterminant car la quasi-totalité des solutions que l'on trouve dans le commerce peuvent indexer de plusieurs dizaines de milliers à plusieurs millions de fichiers. Rappelons que l'ensemble des fichiers des Intranet Qualité de la HES-SO//Genève ne représentent que 9'000 fichiers.

- Interface, ergonomie : D'après le rapport d'activité du mandat de la HEG[1] de Décembre 2005, les premières raisons de la non-utilisation de l'Intranet Qualité se situent au niveau du manque de clarté, de structure, de l'aspect non-intuitif du site et du manque d'un moteur de recherche. Pour toutes ces personnes, la recherche d'informations utiles est difficile. Aussi, l'ajout d'un module de recherche se doit d'être convivial et simple d'utilisation afin de renouer les liens entre les collaborateurs et l'Intranet Qualité. L'ensemble des utilisateurs doivent pouvoir prendre en main de manière immédiate le système de recherche et s'approprier les différentes options leur permettant d'optimiser leur processus de recherche.

- Paramétrage des résultats : Parmi les solutions proposées, la plupart d'entre elles disposent d'un paramétrage avancé de la page de résultat en donnant aux utilisateurs diverses informations sur les différents résultats de la recherche : nombre de documents trouvés, date et type du fichier, résumé de la page ou affichage de synonymes pour les termes recherchés. Ce critère est bien pondéré en raison de l'interaction directe des collaborateurs avec le système de recherche.

- Présentation des résultats : La clarté et la lisibilité de la page de résultats sont essentielles pour l'utilisateur qui pourra accéder de manière directe aux

[1] Annexe – [A4.1]

documents qu'il souhaite. Aussi, chaque occurrence de résultat devra se distinguer nettement avec la possibilité supplémentaire de regrouper les résultats par catégories (ce qui est réalisable suivant l'outil choisi). Ce critère est déterminant pour l'utilisation régulière et conviviale de l'Intranet Qualité.

- Rapidité de la recherche : L'internaute se décourage vite ! En effet, la lenteur de chargement de la page de résultats peut avoir un effet négatif sur l'intérêt que porte l'utilisateur à celle-ci. Le but du moteur de recherche est de fournir un maximum de résultats le plus rapidement possible avec un minimum de bruit documentaire. En général, l'internaute ne patiente que rarement plus de quelques secondes pour le téléchargement d'une page Web.

- Distinction maj/min, accents : La distinction des majuscules/minuscules et des accents varie en fonction des moteurs de recherche. Si Google et MondoSearch parviennent à reconnaitre les mots exacts saisis sans accent et/ou en majuscule, la solution Open Source devra contenir des procédures d'identification des termes afin d'effectuer la recherche avec le terme réel. Cependant, les utilisateurs ont l'habitude de saisir les mots avec une orthographe correcte grâce à leur familiarité avec le traitement de texte. Ce critère n'est donc que légèrement coté.

- Recherche booléenne (logique) et de proximité : Certains opérateurs (OR, AND, NEAR, ADJ …) ainsi que les troncatures peuvent être utilisés pour la recherche de documents. Certains internautes affectionnent particulièrement l'utilisation de ces opérateurs afin de restreindre au maximum les résultats obtenus par leur requête. Rappelons que Google utilise « AND » par défaut et que MondoSearch permet de paramétrer l'utilisation par défaut soit du « OR » soit du « AND ».

- Classement des résultats : La pertinence et le tri des résultats est l'un des critères les plus importants pour la plupart des internautes. En effet, rares sont les utilisateurs qui navigueront pendant plus de 20 secondes afin de trouver le document souhaité. En revanche, des documents triés selon leur pertinence l'aideront à atteindre son objectif et le système de recherche remplira son rôle efficacement. Si le système de recherche ne parvient pas à répondre aux attentes premières des utilisateurs, alors celui-ci devient inutile et la frustration de l'utilisateur aura des effets négatifs sur l'Intranet Qualité de la HEG.

- Recherche avancée : certains outils proposent des fonctionnalités de recherche avancées tel que la sélection du type de fichier, la langue du document (multi-langues), la recherche par catégorie si cela a été préalablement configuré ou la possibilité de rechercher un ou plusieurs mots d'une expression. Bien que ces options puissent paraître pratiques pour l'utilisateur, ce dernier ne les utilise que rarement car il estime la plupart du temps que l'expression de la recherche est amplement suffisante. N'oublions pas que l'intérêt premier de l'utilisateur est le résultat de la recherche et non son mode de fonctionnement.

3.2.2 Interprétation des résultats :

D'après le tableau d'analyse discriminante, nous obtenons les résultats suivants :

- Google Search Appliance : 118
- Google Mini : 115,5
- Solution Open Source : 74,5
- MondoSearch : 113

Il existe une très forte disparité entre les solutions. Certains critères pénalisent la solution Open source qui se retrouve dès lors vite dépassée par les attentes premières des utilisateurs dans une structure comme la HEG. De plus, la mise en place d'une telle solution n'est pas évidente et l'on ne dispose pas toujours des ressources humaines nécessaires à l'installation de tels produits.

MondoSearch et les solutions Google obtiennent un score relativement proche et l'on ne peut dégager de « vainqueur ». Désormais, le choix de l'outil dépendra de facteurs politico-administratifs (Google Search Appliance), organisationnels (Google Mini) et financiers (MondoSearch).

Les problèmes politiques et organisationnels sont plus difficiles à gérer que les problèmes financiers ou techniques. En effet, le financement d'un projet peut s'effectuer sans moindre mal car les coûts sont pris en charge par les 7 écoles de la HES-SO//Genève. Cependant, la solution doit faire l'unanimité et là se pose un problème de nature nouvelle. Qui peut/doit décider de la solution adéquate pour les HES-SO ?

4. Mise en place d'une solution

4.1.1 Architecture de l'Intranet Qualité

Le site Web actuel de l'Intranet Qualité a été réalisé par le Laboratoire Qualité de la HEG grâce au logiciel MEGA, une plateforme de modélisation intégrée comprenant tous les outils de modélisation graphique pour l'analyse des processus et l'architecture d'entreprise.

Le site Intranet est composée de pages générées automatiquement à partir de templates et de pages HTML personnalisées (et donc modifiables pour la mise en forme du module de recherche). La navigation s'effectue via des Frames.

Au niveau de l'architecture du site Intranet, on retrouve une page d'accueil « homepage.htm » et 4 dossiers qui contiennent les pages générées, les images, les fichiers PDF et MS Office ainsi que les pages personnalisées[1]. Toutes les pages personnalisées (non générées par MEGA) ont un nom de fichier commençant par « ws_mp_ » ce qui permet au webmaster de les reconnaitre plus aisément. Les pages générées sont définies à partir d'une clé primaire car chaque page concerne un unique objet. Les noms de fichier sont des nombres hexadécimaux générés par le système, par exemple « 2BCD6436406D0335.htm ».

L'ensemble de la navigation s'effectue par enroulement et déroulement de menus en JavaScript avec l'affichage du contenu au centre de la page, dans une Frame nommée « Content ».

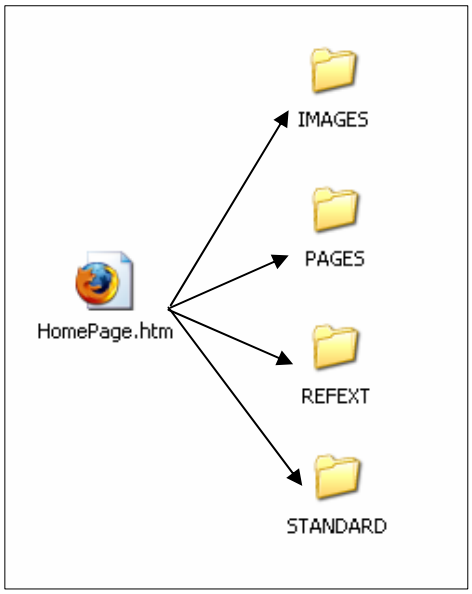


Figure 11 – Hiérarchie des répertoires du site Intranet

[1] Annexe – [A4.1]

4.1.2 Démarche de la mise en place

4.1.2.1 Téléchargement et Contact avec Mondosoft

L'étude Gartner[1] du 6 Octobre 2006, montre que le produit de l'éditeur MondoSoft est un choix pertinent. En effet, MondoSearch est parfaitement compatible avec l'environnement Microsoft et son serveur IIS. De plus, le moteur d'indexation et de recherche est disponible pour un coût relativement bas (15'000 CHF d'après devis pour la HEG).

Ne pouvant pas réaliser de tests sur les produits Google (pas de location disponible pour le Google Mini et pas d'autorisation pour le Google Search Appliance de l'État de Genève) et ne désirant pas créer un prototype à partir d'un système trop complexe que proposait MnoGoSearch, je me suis naturellement tourné vers MondoSearch.

La version d'évaluation de MondoSearch est disponible à l'adresse suivante : http://www.mondosearch.com/Download.aspx. Il est nécessaire de remplir un formulaire de renseignement que le service Ventes de MondoSearch valide. Après quoi, nous recevons un mail nous fournissant une clé de licence ainsi qu'un lien vers un deuxième formulaire à remplir pour obtenir un code PIN utile lors de l'installation du logiciel. Ce formulaire nous demande en effet le nom de la machine sur lequel sera installée la version d'évaluation et nous obtenons ainsi un code PIN unique en fonction de la machine.

Malgré le téléchargement d'une version d'évaluation, j'ai demandé un devis afin d'évaluer le coût de la solution auprès de MondoSoft. J'ai donc rempli un devis et pris contact avec le service de Ventes basé à Copenhague afin de spécifier les caractéristiques de l'environnement de la HEG (nombre de CPU pour le serveur, nombre de fichiers à indexer).

4.1.2.2 Installation de MondoSearch à la HEG

Il est nécessaire de rappeler qu'une première phase de tests à été réalisé sur notre ordinateur personnel avec le serveur Web de Microsoft et que ceux-ci ont été concluants. C'est grâce à la facilité d'installation et de paramétrage que cet outil a été choisi pour créer un prototype d'indexation de site. A noter que l'installation, le paramétrage de base (langue, prise en charge des frames, …) ne prend qu'une dizaine

[1] Bibliographie

de minutes et que l'on obtient des résultats satisfaisants. Un paramétrage plus fin permet de réduire le bruit et d'augmenter la pertinence des résultats.

Pour utiliser MondoSearch, nous allons utiliser un environnement virtuel appelé WMWare, concurrent direct de Virtual PC (Microsoft) beaucoup plus lent. Une fois WMWare installé, on lance l'exécutable « Vitali.wmx » pour travailler dans l'environnement virtuel et copier dans le répertoire « C:\Inetpub\wwwroot\ » le dossier contenant le site Intranet. Ce processus intègre sous IIS le site Web dans un dossier « Intranet ». J'ai utilisé un environnement virtuel car il ne m'était pas possible d'utiliser HEGNTM, le serveur de production.

Il faut tout d'abord définir une page d'accueil par défaut pour le site en question dans les propriétés du dossier « Intranet » en saisissant le nom « homepage.htm ».

À noter par ailleurs qu'il faut autoriser le port 80 au niveau du firewall pour accéder à la machine virtuelle WMWare pour pouvoir travailler en local sur le site Intranet copié dans l'environnement virtuel.

MondoSearch peut ensuite être installé. La procédure requiert la clé de licence (obtenue par mail) ainsi qu'un code PIN récupéré sur Internet en donnant le nom donné à la machine virtuelle, ici « heg-mondo ». Une fois installé, MondoSearch se lance en saisissant l'URL suivante : http://heg-mondo/insite.

4.1.2.3 Paramétrage de l'application

L'application étant installée, il faut ensuite configurer les paramètres d'indexation, de recherche et de résultats. Dans un premier temps, il faut configurer le robot d'indexation (crawler) : indexation des mots visibles sur la page et des méta informations, type des fichiers à indexer, catégorisation des fichiers indexés, le nom de la machine (http://heg-mondo), prise en compte des frames du site Intranet, définition de la page de départ et des dossiers constituant la hiérarchie du site …

L'interface utilisateur est aussi entièrement paramétrable au niveau des fonctionnalités avancées (langue, type de fichiers à rechercher, proposition de termes proches) et du design. Il est même possible d'utiliser des gabarits existants pour la page de recherche et de résultat : on pourrait mettre en place un modèle de page de résultat basé sur le design du moteur de recherche Google™.

La manière de rechercher des informations peut aussi être configurée. Ainsi, lorsque l'utilisateur saisit plusieurs mots, il a la possibilité de rechercher explicitement

l'ensemble des termes ou au moins l'un des termes. La page de résultats indiquera, en fonction de ses choix, les mots trouvés dans chaque document retourné. A noter par ailleurs que la gestion de la troncature est prise en compte.

Au niveau de la page des résultats, l'administrateur peut paramétrer les informations à afficher pour chaque occurrence : type de fichier, langue, date de création ou de dernière modification, nombre de documents trouvés, suggestions de recherche …

4.1.2.4 Phase de test

Une fois l'application paramétrée, il faut réaliser l'indexation du site Intranet. Lors du crawling (indexation des fichiers), il est possible de voir le log qui affiche en temps réel les processus en cours d'exécution. Par exemple, pour traiter un fichier PDF, il y a plusieurs processus : ouverture d'une ligne dans la base de données, conversion temporaire du fichier PDF en fichier texte, indexation des mots dans la base de données et fin de tâche pour le fichier.

Le processus d'indexation (grabbing process) est divisé en 4 phases distinctes :

- Initialisation : chargement des paramètres et préparation de la Grab Database.
- Indexation des hôtes : chargement des points de départs (starting points) définis, indexation de tous les mots rencontrés lors du parcours de l'ensemble des liens.
- Fusion des données : après l'indexation, MondoSearch catégorise les résultats, supprime les pages dupliquées et optimise la base de données.
- Publication de l'interface utilisateur (page de recherche) pour la phase de test.

Il est important de préciser qu'il existe deux types de bases de données utilisées par MondoSearch :

- Searchable Database : cette base de données est employée par les utilisateurs lors de la mise en production du système de recherche.
- Grab Database : cette base de données est mise à jour lors d'une nouvelle indexation de sites Web et utilisée par l'administrateur pour tester les nouvelles configurations ou pour l'indexation de nouvelles parties du site. Pour que les utilisateurs bénéficient de la nouvelle base de données, il faut faire la mise à jour de la base de données de production par simple copie de la Grab BDD vers la Searchable BDD.

Après l'indexation du site, on peut publier la base de données afin de tester le système de recherche.

- Page de recherche :

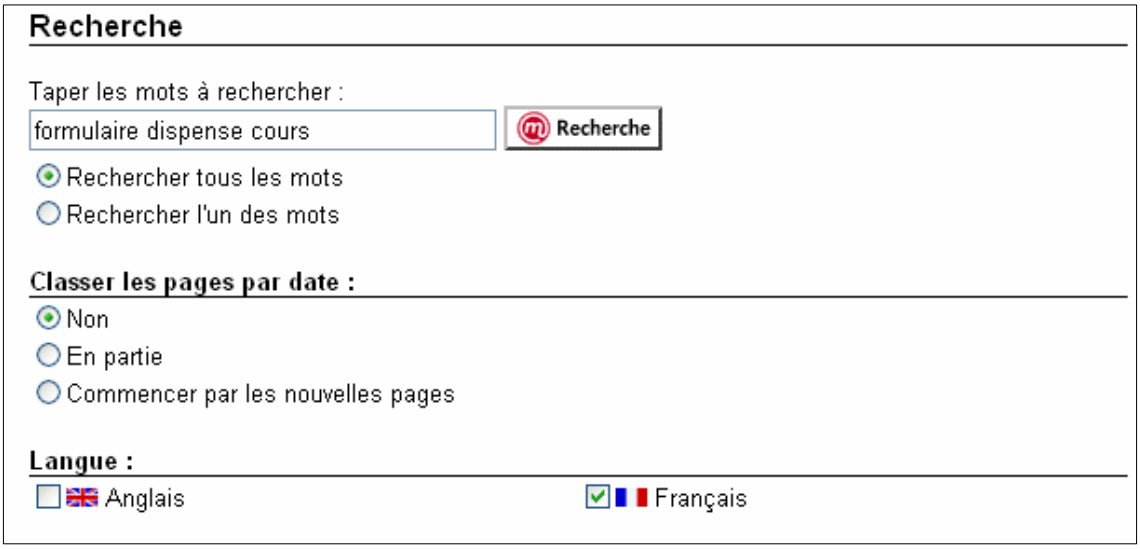


Figure 12 – Page de recherche sous MondoSearch

Comme on peut le voir sur la figure ci-dessus, l'utilisateur dispose d'un formulaire comprenant une zone de recherche, qui lui offre la possibilité de rechercher l'ensemble ou l'un des termes saisis et qui lui permet de choisir la méthode de classement des résultats. Enfin, il lui est possible d'effectuer une recherche en plusieurs langues si le document recherché n'est pas disponible en français grâce à la reconnaissance des termes (dictionnaire MondoSearch).

- Page de résultats :

Figure 13 – Page de résultat sous MondoSearch

L'administrateur paramètre lui-même les éléments à afficher sur la page de résultats. Ici, on retrouve les informations de la recherche en haut de page (termes recherchés, catégorisation, nombre de documents trouvés et nombre de termes trouvés sur chaque documents).

Sur la figure ci-dessus, nous recherchions le formulaire traitant de la procédure de dispense de cours pour un étudiant. Tout naturellement, j'ai saisi les 3 mots clés suivants : *formulaire – dispense – cours*. Le premier résultat de la page de recherche correspond exactement au document recherché, à savoir le fichier PDF contenant le formulaire de dispense d'unité de cours / module.

4.1.2.5 Mise en place de MondoSearch dans l'Intranet Qualité de la HEG

Le système de recherche fonctionnant à travers MondoSearch, il nous faut donc l'intégrer dans le site Intranet Qualité de la HEG. Pour cela, le guide du développeur fourni avec l'application nous servira de support. En effet, ce dernier contient un exemple de script à intégrer et à paramétrer de sorte que l'utilisateur final puisse utiliser le moteur de recherche MondoSearch.

Dans un premier, il a fallu mettre en place le formulaire de recherche qu'utilisera l'utilisateur final. Il s'agit ici de simples balises HTML contenant la barre de recherche et le bouton d'envoie de la requête que l'on place dans la page « WS_MP_PERMANENT_FR.HTM » (frame Permanent)

```
<FORM ACTION="search.asp" METHOD="GET" target="Content">
        <INPUT TYPE="TEXT" NAME="QUERY" SIZE="30">
        <INPUT TYPE="SUBMIT" VALUE="Rechercher">
</FORM>
```

Un script ASP est nécessaire afin de récupérer le paramètre constitué des mots clé saisis par l'utilisateur. Ce fichier est nommé search.asp[1] et est disponible en annexe de même que les explications le concernant.

La tâche du script « search.asp » est de récupérer la requête formulée par l'utilisateur et de générer la page de résultats qui sera contenue dans la frame nommée « Content ». Une arborescence du site est disponible en annexe[2].

Afin d'obtenir un bon rendu visuel, j'ai ajouté dans le fichier CSS de l'Intranet Qualité une partie me permettant de spécifier la forme du texte :

```
h1.mondoCategory {
        FONT-SIZE: 1.2em;
        FONT-WEIGHT: normal;
        COLOR: gray;
}
```

Dans le script search.asp, il suffit de déclarer l'utilisation de la feuille de style :

```
<LINK REL="STYLESHEET" TYPE="text/css" HREF="ws_intranet.css">
```

Puis dans la section du code souhaitée, ici l'affichage des catégories, on utilise la balise suivante :

```
<h1 class='mondoCategory'> texte  </h1>
```

---

[1] Annexe – [A4.3]

[2] Annexe – [A4.1] [A4.2]

Problèmes survenus :

- Un des points problématiques concernait les réponses fournies par le moteur de recherche de MondoSearch. En effet, celui-ci renvoyait systématiquement une quantité d'information trop importante pour l'utilisateur final bien que les résultats soient pertinents. On trouvait par exemple l'ensemble des pages HTM qui contenait les mots clés saisis, alors que l'internaute désire seulement les documents de type PDF et MS Office. Pour cela, deux solutions se présentent : soit de n'indexer que le dossier REFEXT contenant les fichiers PDF et MS Office, soit d'indiquer que l'on peut suivre mais ne pas indexer l'ensemble des fichiers du site.

- Un problème majeur est intervenu lors du déploiement du système de recherche à l'intérieur du site Intranet. En effet, le script contenu dans le manuel du développeur de MondoSearch contenait une erreur de syntaxe renvoyant systématiquement une page d'erreur au lieu de la page de résultat. Après modification du script (simple suppression d'un caractère), il s'avérait que l'outil fonctionnait parfaitement.

- Enfin, l'utilisateur peut s'apercevoir que les titres des documents présents dans la page de résultats peuvent ne pas sembler pertinents. Il convient alors de préciser que certaines informations présentées dans la page de résultats sont issues des méta-informations. Par exemple, le titre d'un document PDF ne sera par forcément celui que l'on peut voir dans l'explorateur de fichiers mais celui qui aura été attribué aux propriétés du document par son créateur.

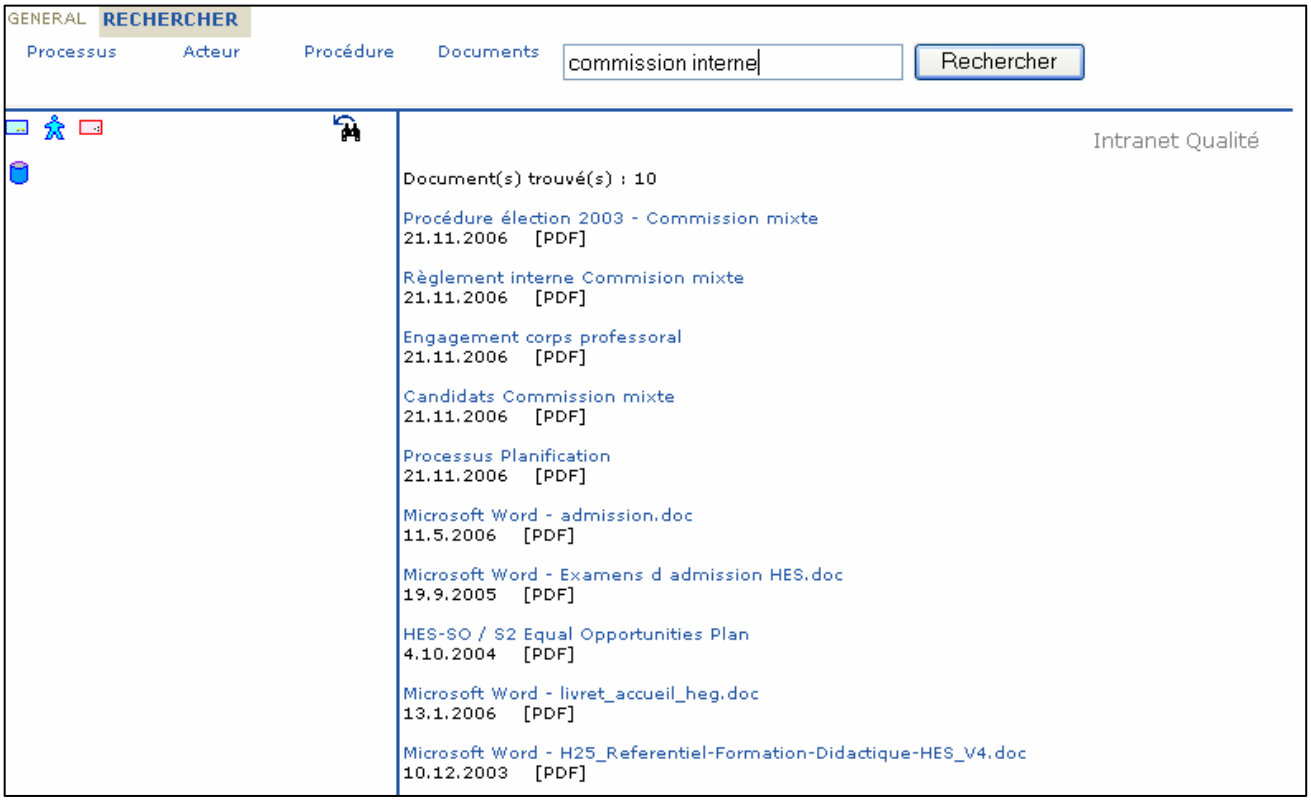


Figure 14 – Page de résultat dans l'Intranet Qualité

La page de résultat que l'on obtient contient les informations suivantes :

- Nombre de documents trouvés
- Titre des documents ainsi que le lien pour y accéder
- Date de création ou dernière modification du fichier
- Type du fichier

Avec les informations présentées et les différents filtres appliqués, l'utilisateur obtient donc bien des résultats cohérents avec un bruit minimal.

5. Conclusion

Les intranets ont fait une percée remarquable ces dernières années. Il existe cependant une fracture entre les utilisateurs et les fournisseurs de contenu, comme le montre le rapport d'activité lié au mandat de la HEG (document PDF disponible sur le CD). L'une des premières fonctions proposées par les sites Intranets est la mise à disposition de l'information. Encore faut-il trouver celle-ci ! C'est d'ailleurs ce qui faisait défaut au site Intranet Qualité de la HEG; l'information n'étant disponible que par navigation hypertextuelle et arborescente, l'utilisateur devait avoir conscience de l'architecture du système pour y naviguer. Un outil de recherche fiable et simple d'utilisation s'imposait donc afin de ramener les utilisateurs à utiliser cet espace de communication. En effet, le plus grand danger qui menace les utilisateurs lors d'une recherche sur le Web ou l'intranet n'est pas de ne rien trouver, mais au contraire d'obtenir un trop grand nombre de résultats qu'il est impossible d'exploiter.

Après la réalisation du prototype avec MondoSearch, nous avons pu mettre en évidence une des réponses plausibles à la problématique. En effet, il est possible d'indexer l'ensemble des Intranets Qualité des HES-SO de Genève afin d'introduire un système de recherche par mots clé. Pour cela, l'ensemble des fichiers des Intranets Qualité doivent être stockés sur le serveur HEGNTM, parcourus par le moteur d'indexation afin de constituer une base de données exploitable lors de la recherche d'informations. Il est tout à fait possible avec les outils Google ou MondoSoft de mettre en place une politique de catégorisation des documents indexés afin de fournir aux utilisateurs des résultats selon un domaine prédéfini par exemple, par école ou par service (qualité, ressources humaines). Ce dernier point impliquerait cependant une réorganisation générale de la hiérarchie des dossiers sur le serveur de fichiers.

Nous pouvons observer, comme cela a été mis en évidence par le prototype que l'affichage est entièrement personnalisable tant au niveau graphique que pour le contenu. En effet, la récupération du flux XML nous permet de faire apparaître nombre de paramètres propres aux résultats. Pour des raisons de lisibilité évidente, nous n'avons fait afficher que les renseignements les plus pertinents et aptes à aider l'utilisateur (titre, URL, type de fichier …) mais beaucoup d'autres sont disponibles : image, logo du type de fichier, etc.

Pour l'utilisateur, un résultat peut être considéré comme exploitable lorsqu'il se retrouve devant un choix restreint de réponses, chacune d'entre elles ayant un taux de

pertinence élevé. De plus, la description des résultats obtenus doit lui permettre d'identifier rapidement le document recherché parmi un ensemble pourtant cohérent.

Nous avons montré que la pertinence des résultats et leur taux de précision dépendent fortement des méta informations associées aux documents indexés par le logiciel de recherche. Il est donc essentiel que ces méta informations soient correctement définies pour obtenir des résultats effectivement exploitables. Pour cette raison, il est nécessaire d'effectuer un travail en amont, d'une part lors de la création de documents, d'autre part pour les documents existants, de sorte que le titre du document, les mots clés et la description soient renseignés avec des informations cohérentes avec le contenu.

Beaucoup de PME misent sur les systèmes d'information comme atout stratégique. Le taux d'utilisation de l'espace informatif que représentent les Intranets ne cesse de croître, notamment par l'intégration de modules de recherche aux interfaces connues et familières, comme GSA ou Google Mini, et jouissant d'une bonne réputation. L'implantation d'un tel système de recherche apparait donc comme essentielle pour ces entreprises et leur apporte de nombreux avantages tant au niveau fonctionnel que financier. En effet, on remarque que l'installation des outils nécessaires est relativement rapide et que le coût global de déploiement est inférieur au coût de développement d'un projet interne. Le développement d'un module de recherche aurait pris beaucoup plus de temps et le résultat en termes de fonctionnalité, rapidité et fiabilité aurait été de moindre qualité.

On peut se demander quel est le degré de confiance des entreprises pour les solutions gratuites. Si ces dernières sont souvent retenues en tant que solutions envisageables, il est parfois difficile de les implanter, tant l'installation et la configuration représentent un travail conséquent pour des résultats parfois irréguliers. Le manque de convivialité et de contrôle des systèmes de recherche représentent souvent un frein pour les entreprises qui désirent avant tout un module stable, rapide de déploiement et efficace.

Enfin, si la recherche d'information tient une place prépondérante au sein des entreprises, il n'en demeure pas moins qu'une étude sérieuse doit être menée avant le choix d'une solution. On ne remet plus en cause l'utilité d'un Intranet ni ses multiples avantages; on cherche aujourd'hui à optimiser son utilisation, notamment en concentrant la totalité de l'information pour que cet outil de travail soit le plus performant possible. L'Intranet est un outil de communication et de travail collaboratif, son utilisation régulière est donc directement liée tant à la mise à jour des informations qu'il contient qu'aux outils de recherche d'information offerts.

Bibliographie

GIROUX, Alain, MADINIER Hélène, GRAS, Elphège. *Mandat HEG Phase 1 – Rapport d'activité*. HES-SO//Genève. Genève : Décembre 2005. 31 pages.

Article Moteur de recherche, *Site Wikipédia* [en ligne]. http://fr.wikipedia.org/wiki/Moteur_de_recherche (consulté le 1.10.2006).

Quelles sont les conséquences d'une mauvaise indexation ?, *Site Gipro, Glossaire des termes de la Gestion de l'Information* [en ligne]. http://fr.wikipedia.org/wiki/Moteur_de_recherche (consulté le 7.10.2006).

Un moteur de recherche sur la toile, *MEDIALOG N°43 – Mai 2002* [en ligne]. http://www.ac-creteil.fr/Medialog/ARCHIVE43/moteurs_de_recherche43.pdf (consulté le 1.10.2006).

Comment fonctionne un moteur de recherche ?, *Site de création de site Web PMESoft* [en ligne]. http://www.pmesoft.be/LibFR/Moteur.php (consulté le 2.10.2006).

Moteurs de recherche, *Rapport de mini-projet ENSICAEN - 2ème semestre 2005-2006* [en ligne]. http://www.ecole.ensicaen.fr/~lrachi/Projet1A/Projet1A.pdf (consulté le 2.10.2006).

Problématique générale de la recherche d'information, Support de cours élaboré par SERRES Alexandre, *Maître de conférences en Sciences de l'Information et de la Communication, 8 novembre 2004* [en ligne]. http://www.uhb.fr/urfist/Supports/RechInfoInit/RechInfo3Problematique.html (consulté le 28.09.2006).

Comment fonctionne Google ?, Article issu d'un blog personnel - Février 2004 [en ligne]. http://padawan.info/fr/web/comment_fonctionne_google.html (consulté le 18.10.2006).

Support technique de MnoGoSearch, Site non officiel [en ligne]. http://mnogosearch.free.fr/support.html (consulté le 4.11.2006).

Magic Quadrant for Information Access Technology, 2006, Document PDF. [en ligne] http://mediaproducts.http://mediaproducts.gartner.com/reprints/fast/143690.html (consulté le 13.11.2006).

A1.1 - Glossaire

A

- **API Google** : Technologie permettant aux développeurs d'utiliser les services de recherche Google. L'API Google est un ensemble d'outils mis à disposition de tous qui permet d'interroger à distance les serveurs du moteur de recherches Google.

- **Acronyme** : Mot formé de la première ou de quelques-unes des premières lettres de plusieurs mots.

- **Anaphore** : Répétition d'un même mot au début de plusieurs membres d'une phrase afin de renforcer l'idée exprimée ou d'opérer une symétrie.

- **Antonymie** : Désigne des termes de sens opposés.

B

- **Bruit** : Il s'agit d'un nombre trop important de réponses qui n'ont aucun rapport avec le terme recherché par rapport aux réponses attendues.

C

- **CDU** : Classification Décimale Universelle ; système de classification de bibliothèque développé par Paul Otlet et Henri La Fontaine à partir de la classification décimale de Dewey (CDD), et avec l'autorisation de Melvil Dewey.

- **CMS** : Les "Content Management System (CMS)" permettent de réduire le temps de programmation d'un site (structure, design, fonctions). En général ils comportent une interface d'administration très développée qui permet une gestion aisée, rapide et rigoureuse du contenu. Des fonctions supplémentaires peuvent se greffer avec aisance via des "modules" ou autre "add-on".

D

- **DMZ** : Sous-réseau délimité par un firewall.

E

- **Ethernet** : Protocole de réseau informatique à commutation de paquet permettant à des ordinateurs de communiquer sur un réseau local.

F

- **Formes fléchies** : formes plurielle, gérondive, possessive d'un mot.

- **FQDN**: Fully Qualified Domain Name. Représente un nom de domaine non ambigu pour spécifier la position absolue d'un nœud dans un arbre DNS (Domain Name Server). Il s'agit donc de l'adresse complète d'une machine incluant son nom d'ôte et son nom de domaine.

G

- **GSA** : Google Search Appliance.

H

- **HES-SO**//Genève : Haute Ecole Spécialisé de Suisse Occidentale (EIG, EIL, HEAD, HEG, HEdS, HETS, HEM).

I

- **Indexation** : opération consistant à repérer dans un texte certains mots ou expressions particulièrement significatifs (appelés *termes*) dans un contexte donné, et à créer un lien entre ces termes et le texte original.

- **Index** : en informatique, l'index est une fonctionnalité permettant d'effectuer un accès rapide aux enregistrements d'une table.

- **Intranet** : Un intranet est un réseau informatique utilisé à l'intérieur d'une entreprise ou de toute autre entité organisationnelle utilisant les techniques de communication d'Internet (IP).

- **IIS** : Internet Information Services. Serveur http et/ou FTP créé par Microsoft pour ses systèmes d'exploitation Windows.

M

- **Moteur de recherche** : machine spécifique (matérielle et logicielle) chargée d'indexer des pages web afin de permettre une recherche à l'aide de mots-clés dans un formulaire de recherche.

- **Métamoteur** : outil de recherche qui interroge en parallèle plusieurs moteurs et annuaires de recherche, rapatrie leurs réponses et les organise, selon des méthodes de classement spécifiques, pour fournir aux utilisateurs une présentation structurée des résultats.

- **Mot clé** : Mots ou ensemble de mots qu'on inscrit dans un outil de recherche pour signaler l'objet recherché.

O

- **ODBC** : signifie *Open DataBase Connectivity*. Il s'agit d'un format défini par Microsoft permettant la communication entre des clients bases de données fonctionnant sous Windows et les SGBD du marché.

P

- **PME** : Petite et Moyenne Entreprise.

- **Polysémie** : Caractérise les différentes significations d'un même terme.

- **PageRank** : Appelé aussi PR est l'indice de popularité d'une page Web utilisé par le moteur de recherche Google.

- **Pertinence** (ou taux de rappel) : il traduit le caractère plus ou moins exhaustif de la recherche.

- **Précision** (ou taux de précision) : il indique dans quel mesure les résultats trouvés sont pertinents par rapport à la recherche effectuée.

- **Parser** : appelé aussi analyseur syntaxique, est un programme informatique permettant d'exhiber la structure d'un texte.

R

- **Recherche documentaire** : Action, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents (Vocabulaire de la documentation, AFNOR, 1987)

- **Recherche documentaire informatisée** : ou RDI. Recherche documentaire utilisant un logiciel documentaire, sur ordinateur ; implique l'élaboration d'une démarche, d'une stratégie de recherche (définition des mots-clés, des clés d'accès, l'élaboration d'une équation de recherche). La RDI est plutôt assimilée à l'interrogation des banques de données.

- **Recherche de l'information** : Action, méthodes et procédures ayant pour objet d'extraire un ensemble de documents les informations voulues (d'après AFNOR, 1979).

- **Robot d'indexation** : ou crawler, est un logiciel qui explore automatiquement le Web et collecte les ressources indexées par un moteur de recherche.

S

- **Système d'information** : Un système d'information (noté SI) représente l'ensemble des éléments participant à la gestion, au stockage, au traitement, au transport et à la diffusion de l'information au sein d'une organisation.

- **Salton** (Formule de) : Cette formule tient compte de l'occurrence d'un mot mais également de la taille du document.

- **Silence** : cela concerne l'absence d'un document pertinent lié à la recherche mais qui n'apparait pas dans la liste des résultats renvoyés par le moteur de recherche.

- **SQL** : Structured Query Language (ou langage structuré de requêtes) est un pseudo-langage informatique standard et normalisé permettant d'interroger et manipuler une base de données relationnelle.

T

- **Thésaurus** : sorte de dictionnaire hiérarchisé ; un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. Il ne fournit qu'accessoirement des définitions, les relations des termes et leur choix l'emportant sur les significations.

U

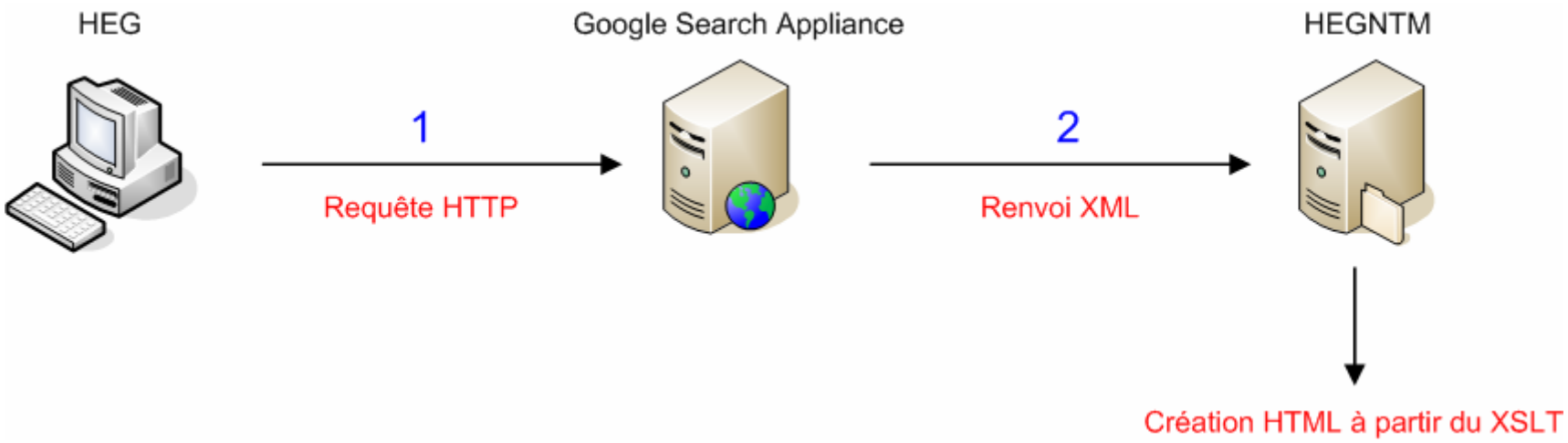
- URL : Uniform Ressource Locator (littéralement « repère uniforme de ressource »). Les URL ont été conçues pour le Web afin d'identifier les pages et sites Web, elles sont donc aussi appelés adresses Web.

X

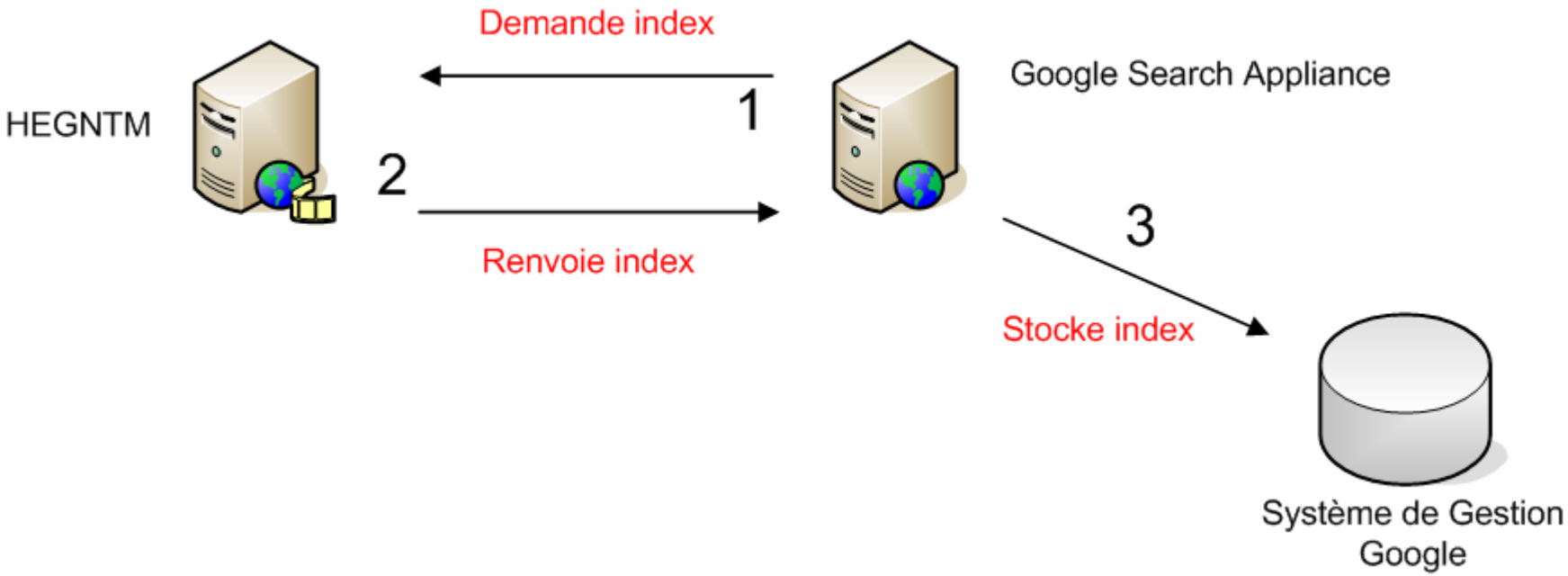
- **XSLT** : Extensible Style Language Transformations est un langage destiné à transformer un fichier XML en un fichier XML ou HTML. Mais ce pourra être tout aussi bien un fichier d'un autre format : par exemple du texte pur, ou du RTF.

A2.1 - Schéma réseau global

A2.2 – Solution avec GSA



A2.3 – Principe de l'indexation avec GSA

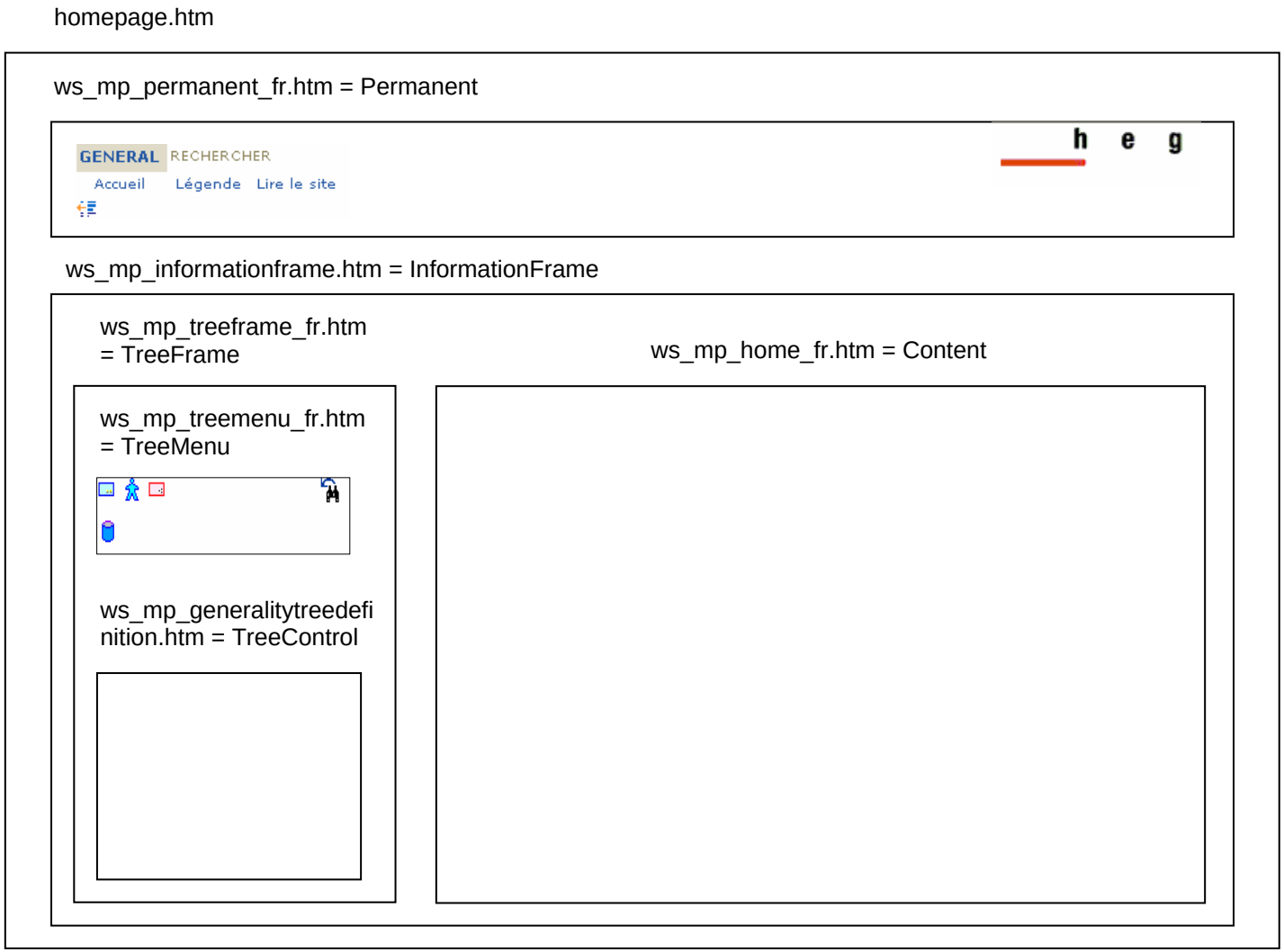


A3.1 – Exemple de page d'accueil

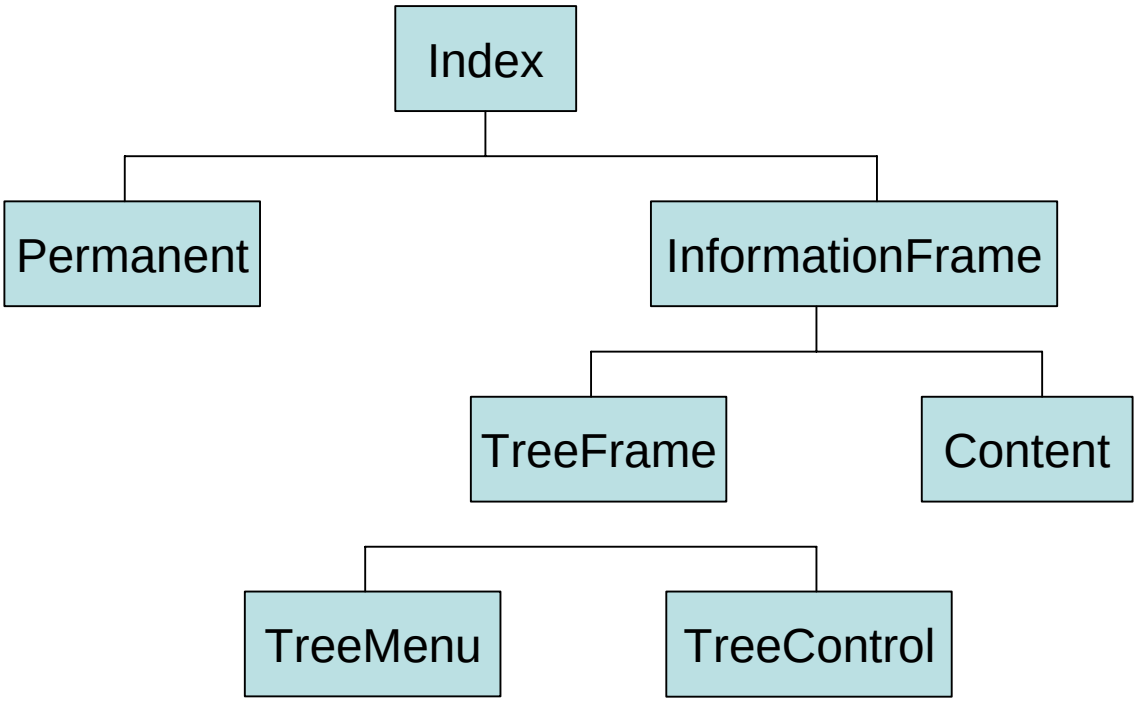
Cette page est issue du site de la HES-SO//Genève dont j'ai modifié le cadre intérieur afin de réaliser une page d'accueil permettant aux collaborateurs d'accéder aux différents Intranet Qualité des écoles de la HES-SO//Genève. De plus, cette page est nécessaire car son URL constitue le point d'entrée des robots d'indexation Google Search Appliance et Google Mini.

Le code source de cette page est disponible sur le CD rendu et la page respecte le style CSS réalisé par le LTI (Laboratoire des Technologies de l'Information).

A4.1 – Structure du site Intranet



A4.2 – Organisation des Frames – Intranet Qualité



A4.2 – MondoSearch Intranet – Search.asp

```
<LINK REL="STYLESHEET" TYPE="text/css" HREF="ws_intranet.css">
	<%
		on error resume next

		Set xmldoc = Server.CreateObject("MSXML2.DOMDocument")
		xmldoc.async = False
		xmldoc.setProperty "ServerHTTPRequest", true
		xmldoc.Load "http://heg-mondo/cgi-bin/MsmFind.exe?resmask=MsmResXML.msk&QUERY=" &
		Server.HTMLEncode(Request.QueryString("Query"))
		set rp = xmldoc.selectsinglenode("resultpage")
		set res = rp.selectsinglenode("results")
		set sum = res.selectsinglenode("summary")

		for each c in res.selectnodes("category")
		  response.write "<h1 class='mondoCategory'>"&c.selectsinglenode("title").text&"</h1>"&chr(13)&chr(10)

		  'AFFICHAGE DU NOMBRE DE DOCUMENTS TROUVES
		  response.write "Document(s) trouvé(s) : "&c.selectsinglenode("pagecount").text&"<BR /><BR />"

		  'AFFICHAGE DU LIEN + URL
		  for each r in c.selectnodes("page")
			response.write "<A HREF="""
			response.write r.selectsinglenode("url_to_go").text&"""
			response.write " title="""&r.selectsinglenode("description").text&"""
			response.write ">"
			response.write r.selectsinglenode("title").text
			response.write "</a><BR />"&chr(13)&chr(10)
```

```
        ' AFFICHAGE DE LA DATE
        response.write r.selectsinglenode("date").text
        response.write "    "

        ' AFFICHAGE DU TYPE DE FICHIER
        response.write r.selectsinglenode("mime_text").text

        ' AFFICHAGE DE LA DESCRIPTION
        response.write "<BR /><BR />"
  next
next



Set xmldoc = Nothing

if err.number > 0 then
  'response.write(err.number&" - " & err.Description&"<br>")
  'response.write("ligne : "&err.Line&" - colonne" & err.Column&"<br>")
  'response.write("Source : "&err.Source&"<br>")
  'response.write("File : "&err.File&"<br>")
end if
%>
```

Le code de débogage nous a permis de mettre en exergue les différentes erreurs survenues lors de la phase de déploiement du système de recherche en faisant afficher le numéro et la description des erreurs ainsi que d'autres informations pouvant être utiles.

Commentaire du script :

- xmldoc.async = false : cet attribut détermine si le téléchargement du document XML s'effectue de manière asynchrone ou pas. Un téléchargement asynchrone concerne un fichier commençant à s'enregistrer à l'emplacement souhaité sans bloquer le script. Lorsque la fonction aura terminée son traitement, le fichier ne sera pas encore disponible.
- xmldoc.setProperty "ServerHTTPRequest", true : la méthode setProperty affecte un attribut de l'objet auquel elle est affectée. C'est un accesseur en écriture.
- Server.HTMLEncode : Cette méthode applique les règles d'encodage et de syntaxe du HTML à la chaîne de caractères passée en paramètre. L'intérêt se situe au niveau des caractères réservés du HTML ; cela permet entre autre d'écrire des balises ne devant pas être interprétées.
- Selectsinglenode : La méthode SelectSingleNode permet de récupérer la valeur d'un nœud XML (ou tag XML).
- SelectNodes : Cette méthode sélectionne une liste de nœuds correspondant à l'expression passée en paramètre.
- xmldoc.Load "Param" : Cette méthode charge un document XML depuis la source spécifiée par le paramètre.
- Heg-mondo : Nom de l'hôte pris en compte par MondoSearch.
- MsmFind.exe : Fichier qui exécute la recherche (moteur de recherche).
- MsmResXML.msk : Le fichier de sortie XML par un masque de résultat nommé MsmResXML.msk. Après la recherche, on obtient donc un fichier XML traité plutôt que du HTML.
- QUERY : Il s'agit ici des mots clé saisis par l'utilisateur lors de sa recherche.

A5.1 – Fonctionnement de MondoSearch