N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de Master en Informatique

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Système d'information avancée

Par :

Roaissat Messaouda

SUJET

MISE AU POINT D'UN ROBOT D'INDEXAGE WEB

Jury :

Mr.	**Université de M'sila**	**Président**
Mr. ALLAOUA HEMMAK	**Université de M'sila**	**Rapporteur**
Mr.	**Université de M'sila**	**Examinateur**
Mr.	**Université de M'sila**	**Examinateur**
Mr.	**Université de M'sila**	**Examinateur**

Promotion : 2011 /2012

Table des Matières

Table des Matières

Table des Matières

Table des Matières

Introduction

La Recherche d'Information (RI) est un domaine qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur. L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur avec une représentation du contenu des documents au moyen d'une fonction de correspondance.

L'architecture des outils de RI sur le web est généralement caractérisée par l'utilisation d'un index inversé et d'un ensemble de machines fonctionnant en parallèle, de même que Google. La pertinence des réponses est liée à un système de tri de pertinence construit sur la notion de lien existant entre les pages. Ce principe de recherche et d'évaluation est qualifier aujourd'hui de classique, et les approches en RI se sont orientées vers une nouvelle génération de systèmes de recherche basés sur l'accès contextuel et sémantique à l'information.

Notre monde tend à s'informatiser de plus en plus notamment par l'entremise d'Internet. En effet, ce dernier s'est inséminé dans notre quotidien et dans notre vie professionnelle. Dès que l'on se pose une question surfer sur la toile nous permet d'obtenir de promptes réponses bien qu'il existe des milliards de sites. Grâce à des outils rapides et simples d'utilisation que sont les moteurs de recherche, nous sommes à même de trouver ce que nous cherchons au travers d'un petit champ de texte, comme l'itinéraire pour se rendre à notre lieu de vacances ou à un rendez-vous d'affaires, trouver un tutoriel. Internet nous offre une multitude de possibilités et les moteurs de recherche nous permettent d'y accéder plus facilement. Mais que se cache-t-il derrière cette interface sobre, cet outil devenu indispensable? Quels sont leurs principes ? Comment sont indexées toutes ces pages Internet ?

Certains moteurs comme Google préfèrent ou complètent l'indexation des nouvelles pages par des algorithmes mathématiques plus complexes, mais moins coûteuse en main d'œuvre et plus courte en durée. Pour cela, ces moteurs utilisent des programmes automatisés appelés Spiders, Crawlers, Bots ou encore Robots (l'équivalent français, peu utilisé, est « robot d'indexation »). Ces spiders parcourent sans interruption les pages déjà indexées, naviguant de lien en lien à la recherche de nouveaux liens et en recense les pages. Ensuite des logiciels tel que ICE (Intelligence Concept Extraction) permettent d'établir les rapports entre les termes que les spiders ou crawlers ont trouvé déterminants dans ces pages, les mots clés et les autres paramètres.

Le présent travail montre comment construire un robot d'indexation. Un robot d'indexation est un programme automatisé qui recherche nouveaux documents Web et puis index de leurs adresses en contenu de l'information dans une base de données, Spiders Web sont généralement considéré comme un type de robot, ou un robot Internet. La principale différence entre une araignée et un bot est simple: l'araignée est capable déplacer vers de nouvelles pages son demande programmée . Spiders s'est avérée très utile pour l'un des sites de services publics premiers à apparaître sur le Web le moteur de recherche.

Introduction

Organisation du mémoire

Le travail présenté dans ce mémoire est divisé en quatre chapitres :

- ✓ Le chapitre 1 : La recherche d'information sur INTERNET

  Ce chapitre permet une compréhension claire et précise les concepts de recherche d'information, des systèmes de recherche d'information , ceux des outils de recherche sur le web et le moteur de recherche.

- ✓ Le chapitre 2 :Les Robots d'indexation

  L'objectif de ce chapitre a compréhension claire et précise du concept d'un robot d'indexation et également les notions apparentées aux robots , le comportement et les différentes stratégies crawling.

- ✓ Le chapitre 3 : Problématique

  Ce chapitre a pour objectif de Comment sélectionner le meilleur de stratégie crawling.

- ✓ Le chapitre 4 : Etude Conceptuelle

  ce chapitre entoure l'étude conceptuelle de notre système utilise UML.

- ✓ Le chapitre 5 : Réalisation et Mise en œuvre

  Mise en œuvre les concepts précédentes dans chapitre 4

En fin, nous concluons ce mémoire par une conclusion générale et la présentation de quelques perspectives.

Références

Bibliographie :

[Abdelkrim Bouramoul ,1] : MR. ABDELKRIM BOURAMOUL, Recherche d'information contextuelle et sémantique sur le web. Thèse Docteur en Sciences : Spécialité : INFORMATIQUE. Alegria : Université MENTOURI de Constantine, le 25/09/2011

[Fabien Picarougne, 7] : Fabien Picarougne , Recherche d'information sur Internet par algorithmes évolutionnaires, thèse pour obtenir le grade de Docteur , l'université de tours ,le 19/11/2004.

[Fred colantonio ,10] : Fred colantonio, Référencement, e-marketing et visibilité web : 30 pratiques pour décideurs et webmasters

Webographie

[Aurélien Bardon ,2] : Aurélien Bardon, L'histoire des moteurs de recherche
Disponible sur :
http ://oseox.fr/blog/index.php/594-histoire-moteurs

(Consulté le 01/2012)

[Mickaël MARCHAL ,3] : Mickaël MARCHAL Nadia TEA ; Les moteurs de recherche. Comment indexent-ils l'information, et comment la restituent-ils ? Promo 2007
Disponible sur : www.lesitedemika.org/ressources/**moteurs recherche**.pdf
(Consulté le 11.2011).

[Jean-Christophe Féraud ,4] : Jean-Christophe Féraud, Petite histoire des moteurs de recherche
Disponible sur :
http://www.01net.com/editorial/186972/petite-histoire-des-moteurs-de-recherche

Références

[wikipedia ,5] :

- ✓ Moteur de recherche

Disponible sur :http://fr.wikipedia.org/wiki/Moteur_de_recherche

- ✓ web crawler
  Disponible sur : http://en.wikipedia.org/wiki/Web_crawler

- ✓ URL normalization
  Disponible sur : http://en.wikipedia.org/wiki/URL_normalization

[Algorithmes, moteur et technique d'indexation ,6] : Algorithmes, moteur et technique d'indexation
Disponible sur :
http://www.webmaster-hub.com/publication/-Algorithmes-moteurs-et-techniques-.html

[web crawling ,8] :

- ✓ web crawling
  Disponible sur : grupoweb.upf.es/WRG/course/slides/crawling.pdf

- ✓ carlos castilo , web crawler

Disponible sur :
http://www.slideshare.net/ChaToX/web-crawling-2006

[annuaire-info ,9]:

- ✓ Fiche robot.txt

Disponible :

http://www.annuaire-info.com/robots-txt/fichier

- ✓ Balise META « robots »

Disponible sur :

http://www.annuaire-info.com/robots-txt/meta-robots/

Références

- ✓ Protocole d'exclusion des robots :
  Fichier .htaccess/X-robots-tag

  Disponible sur :
  http://www.annuaire-info.com/robots-txt/

[High Performance Crawling System ,11] : High Performance Crawling System
Disponible sur :

https://wing.comp.nus.edu.sg/downloads/.../100/100.pdf

[Spiders ,12] : Spiders, crawlers, harvesters, bots

Disponible sur:
clgiles.ist.psu.edu/IST441/materials/.../crawlers.pp

[Breadth-First ,13]: Breadth-First

Disponible sur :http://combine.it.lth.se/CrawlSim/report/node18.html

الملخص:

قد تكون مهمة العاملين في الفهرسة وتصنيف شبكة ويب العالمية بأكملها على نطاق واسع صعبة ومعقدة. هذا العمل هو دائما أو تقريبا مخصص لعناكب البحث . زاحف الشبكة هو برنامج الكمبيوتر الذي يتصفح الشبكة العالمية بطريقة منهجية و منظمة. إن الزحف على شبكة الإنترنت هو وسيلة هامة لجمع بيانات، ومواكبة تزايد حجم شبكة الإنترنت، حيث بإستمرار هناك عدد كبير من صفحات الويب التي تضاف كل يوم، والمعلومات في تغير مستمر. هذه المذكرة هي لمحة عامة عن مختلف أنواع عناكب البحث والسياسات مثل، اختيار، إعادة الزيارة، عملية الموازاة، المداراة التي ينطوي عليها ذلك.

كلمات-المفتاح : عناكب البحث , محرك البحث, الشبكة العنكبوتية, سلوك , سياسات الزاحف .

RESUME:

La tache des agents d'indexation et de classification de documents web est devenue absolue laborieuse et complexe. C'est un travail qui est presque toujours réservé pour les robots d'indexation.

Un robot d'indexation est un programme informatique qui navigue sur le web en une approche méthodique, de manière automatisée ou d'une manière ordonnée. Le crawling du Web est une opération très importante pour la collecte des données sures, et de tenir avec l'Internet en pleine expansion. Un grand nombre de pages Web sont continuellement ajoutées chaque jour, de l'information est en évolution croissante. Ce mémoire est un aperçu des différents types de robots d'indexation et les politiques, comme la sélection, la re-visite, la politesse, la parallélisation qui en sont responsables.

MOT-CLES: Robot d'indexation, le comportement, les politiques, Moteurs de recherche, World Wide Web, Algorithme de parcours en largeur

ABSTRACT:

It would be entirely too large a job for human workers to index and categorize the entire World Wide Web. This is a job that is almost always reserved for Web spiders. A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Web crawling is an important method for collecting data on, and keeping up with, the rapidly expanding Internet. A vast number of web pages are continually being added every day, and information is constantly changing. This thesis is an overview of various types of Web Crawlers and the policies like selection, re-visit, politeness, parallelization involved in it

KEY-WORDs: Web Crawler, Behavior, Policies, Search Engine, World Wide Web.