**MyCrawler**

**Cahier des charges**

Révisions

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Auteur** | **Commentaires** |
| 1.0 | 25/03/2009 | A. Geoffrey | - Création du modèle.  - Rédaction de l’introduction.  - Début de rédaction « Cadre de travail » |
|  |  |  |  |

Sommaire

[2. Cadre de travail 5](#_Toc225747212)

[1.1. Equipe et répartition des tâches 5](#_Toc225747213)

[1.2. Gestion des versions 5](#_Toc225747214)

[1.3. Phases de conception 5](#_Toc225747215)

Introduction

Ce cahier des charges a pour objectif d’orienter le projet *MyCrawler* et de proposer un démonstrateur évolutif capable de prendre en compte des développements futurs. Un chapitre entier sera donc consacré aux évolutions possibles du démonstrateur. De plus, ce document doit servir de référence tout au long du développement du projet.

Le projet *MyCrawler* a pour objectif de proposer un démonstrateur capable de crawler le Web à la recherche d’informations pertinentes en rapport avec une communauté donnée. Il existe quantité de crawler sur le Web capable de réaliser ce travail mais de façon plus ou moins automatique avec des contraintes importantes (temps d’exécution important du crawler avant d’obtenir une quantité conséquente de données exploitable, ne proposer que de crawler des fichiers spécifiques, …). Outre le crawling, *MyCrawler* se devra de fournir une solution capable de crawler rapidement une communauté à l’aide d’un réseau d’ordinateurs. Le fait de distribuer les ressources de crawl entre plusieurs machines impose une démarche rigoureuse de communication entre les différentes entités. Cette fonctionnalité apporte une innovation importante par rapport à tous les crawler existants sur Internet. A la manière d’eMule, des réseaux Torrent, MyCrawler a pour ambition de proposer un réseau conséquent d’échange d’information à l’échelle du peer-to-peer offrant un nouveau moyen de surfer sur Internet. Ainsi, crawler le Web et fournir un moteur de recherche dynamique et pertinent seront désormais des fonctionnalités envisageables et accessibles.

Les efforts de développement seront principalement concentrés à la réalisation d’un prototype capable de crawler une quantité importante de liens hypertexte sans tenir compte, dans un premier temps, du domaine d’étude. En effet, déterminer la pertinence d’une page par rapport à la communauté visée nécessite de développer des outils de text-mining permettant d’écarter ou non une page de l’agrégat construit par le crawler. Une telle fonctionnalité demande un temps de développement conséquent. C’est pourquoi ce développement n’entre pas dans le cadre de ce projet. *MyCrawler* se devra de fournir un Framework évolutif capable de faire communiquer plusieurs ordinateurs afin d’obtenir une visualisation quasi-temps réel d’un graphe conséquent de sites web. L’accent est dont principalement mis sur le partage d’une puissance de calcul répartie de crawl.

# Cadre de travail

Ce projet entre dans le cadre d’une UV PR (Projet) à l’Université de Technologie de Compiègne durant la période de printemps 2009 soit de février à juin 2009 (4 mois). Ce projet est suivi par M. Franck Ghitalla et est sanctionnée par deux soutenances.

## Equipe et répartition des tâches

**PORTEJOIE Julien (GI04 - ICSI)** chargé de concevoir les interfaces utilisateurs et de proposer un outil de visualisation de graphe.

**ANNEHEIM Geoffrey (GI04 – STRIE)** chargé de développer les composants de crawl et de communication entre les différentes machines du réseau crawler distribué.

## Gestion des versions

La gestion des versions se fait à l’aide de Google code (en anglais).

## Phases de conception

Etude sur 2 mois.

Développement sur 2 mois.