**Tài liệu mô tả bài tập lớn nhóm 10**

1. Giới thiệu chung

Mục đích của việc xây dựng chương trình Crawler từ các trang tin tức là để đáp ứng nhu cầu tự động tìm kiếm, thu thập một lượng lớn dữ liệu trên Internet, cụ thể ở đây là trong các trang tin tức tiếng Việt, ví dụ như vnexpress.net, dantri.com,vietnamnet.vn,… Chương trình crawler là một phần mềm được viết bằng Java, trên nền Netbeans IDE. Ý tưởng của chương trình là từ một link nguồn cho trước ( ví dụ <http://vnexpress.net/tin-tuc/thoi-su> ), các tác giả sử dụng xpath để trích xuất ra được link các bài báo từ trong html của trang tin, sau đó đọc html của từng trang tin để lấy ra các thông tin cần thiết ( ví dụ: tiêu đề, nội dung, ngày đăng, tác giả… của bài báo). Sau đó thông tin trích xuất được sẽ được lưu trữ dưới dạng các bản ghi trong MongoDB. Chi tiết cụ thể về cấu trúc của chương trình sẽ được mô tả ở trong phần 2.

1. Cấu trúc của chương trình Crawler

Chương trình crawler có 2 nhiệm vụ chính, nhiệm vụ thứ nhất là dùng xpath lấy thông tin từ trang tin tức dựa trên link khởi tạo ban đầu, nhiệm vu thứ hai là sau khi lấy được thông tin, thì lưu trữ chúng dưới dạng các bản ghi vào trong MongoDB

Các mã nguồn mở được sử dụng trong quá trình xây dựng crawler:

* Netbeans IDE : môi trường phát triển tích hợp dùng để phát triển chương trình crawler bằng ngôn ngữ java.
* MongoDB: hệ thống cơ sở dữ liệu mã nguồn mở, được sử dụng để lưu trữ thông tin crawl về.
* HtmlUnit library: thư viện mã nguồn mở của java, hỗ trợ lập trình viên trong việc đọc cũng như xử lý file Html. Trong quá trình lập trình, các tác giả sử dụng thư viện này để lấy được html content của trang web, từ đó sử dụng xpath để lấy được các thông tin cần thiết
* MongoDB Java Driver: thư viện mã nguồn mở giúp cho việc tương tác với mongoDB thông qua java.

Phần mềm crawler được lập trình gồm 4 class:

* MongoDB: class chứa các hàm tương tác với MongoDB qua java, như kết nối với cơ sở dữ liệu, kết nối đến các collection, insert bản ghi vào trong collection, tìm kiếm bản ghi,…
* Article: class đại diện cho một bài báo được crawl về. Trong đối tượng có chứa các thuộc tính như Url ( link của bài báo), HtmlContent( html của bài báo), Title(tiêu đề), pubDate(ngày đăng), Content( nội dung của bài báo)…
* Channel: đại diện cho một link khởi đầu (hay có thể gọi là link nguồn). Trong đó có các thuộc tính : Url ( link nguồn) và các xpath
* Crawler: đây là class main, gồm các hàm xử lý html cũng như đưa dữ liệu vào mongodb. Nôi dung cụ thể của từng hàm:
  + createClient: tạo một client để xử lý Html (được implement từ trong thư viện HtmlUnit)
  + getUrlList: lấy danh sách các url của các bài báo ( sử dụng xpath)
  + parseUrl: xử lý bóc tách nội dung từ html của các bài báo, lưu các nội dung cần lấy(như title, content, pubDate, html…) vào đối tượng Article.
  + getArticle: đưa bài báo dưới dạng bản ghi vào trong MongoDb
  + class SingleCrawler: khởi tạo multithread, để tăng tốc độ crawler.
  + crawlExecute: thực hiện crawl. Hàm có 3 tham số đầu vào là initUrl( link khởi đầu), limit( số lượng bài báo cần crawl), và numOfThread( số lượng thread sử dụng)

1. Hướng dẫn sử dụng: có hai cách để chạy Crawler

Cách 1: Sử dụng IDE

Phần mềm crawler khi chạy yêu cầu sử dụng Netbeans IDE và bật sẵn mongoDB. Để chạy chương trình crawler, ta cần khởi tạo một đối tượng Channel trong hàm main của class Crawler với các xpath tương ứng, sau đó set biến limit ( số lượng bài báo cần crawl) và numOfThread( số lượng thread sử dụng). Sau khi set xong xuôi các thứ ta chạy project để bắt đầu crawl. Chương trình sẽ tự động dựa vào link và các xpath trong Channel để lấy về các bài báo và đẩy vào mongoDB.

Cách 2: Sử dụng Crawler.bat file và tham số môi trường

Mở file Crawler.bat bằng notepad và sửa cấu trúc tùy biến:

Java –jar Crawler.jar [LIMIT] [NUM\_OF\_THREAD]

VD: java –jar Crawler.jar 100 3