ONYCOM

Ankus Crawler ver 1.0 User's Manual

Update 20170816

ankus

ONYCOM

Library

- httpComponent ver 4.5 (httpclient, httpcore)
- Jsoup ver 1.10.3

Class Detail

- CrawlerDriver.class

  메인 실행 class, 각 사이트에 맞는 다양한 설정 및 수집절차 등을 수집 대상 사이트에 맞도록 정의하여 사용

- PageArgument.class

  스크린트 파일 변수 클래스

- ParseAddr.class

  http 주소 분석 클래스, 주소를 분석하여 주소냐의 변수값을 분리

- ParseScript.class

  스크립트 파일 분석 클래스, 스크립트 파일을 분석하여 추출하고자 하는 정보의 내용 및 주소등을 변수에 저장

- Crawler.class

  입력받은 주소의 페지이를 수집하는 클래스, http 데이터 전송방식에 맞게 post, get 타입 사용

- ParseHTML.class

  스크립트 내용에 맞게 html 페이지에서 원하는 정보를 추출

- ElementDEF.class

  스크립트 파일을 위한 Element 정의 클래스

- ScriptDEF.class

  스크립트 파일 정보 저장 클래스

- GithubParser.class

  Github 서브 페이지 추출을 위한 주소정보 추출을 위한 파서클래스 ???? 말이 좀 이상

Script

Script use method

- Id:0

  스크립트로 추출할 정보 아이디

- addr:https://github.com/Netflix?page=1

  id에 해당하는 수집 페이지 주소

- info-1:TAG,poll-include-fragment, link,html

  페이지에서 추출하고자하는 정보내용

  Info-N 형식으로 다양한 내용 작성가능

  TAG: 추출하려는 정보가 위치에 정의된 타입(TAG,CLASS)

  poll-include-fragment: TAG 나 CLASS의 변수명

  link: 해당 내용에 정의된 정보의 이름

  html: 해당 내용에서 가져올 정보의 타입

- Github 프로젝트별 수집기 스크립터 예시

  스크립트 형태로 탐색과 추출내용을 정의

  id:0

  addrLhttps://github.com/Netflix?page=1

  # 프로젝트 1번 페이지 주소

  info-1:TAG,poll-include-fragment,link,html

  #html 페이지내에 서브 프로젝트 링크가 저장되는 곳의 정보

  # 아래의 태그에서 세부링크주소를 가져오는 부분을 정의

```
<div class="col-3 float-right text-right">
  <poll-include-fragment
    src="/Netflix/lemur/graphs/participation?h
  </poll-include-fragment>
</div>
```

id:1

info-1:CLASS,repository-meta,Description,text

#서브프로젝트 페이지에서 프로젝트 설명정보가 위치한 곳의 정의

#아래의 태그내 class 부분의 정보가 저장된 내용을 수집

```
<div class="repository-meta mb-0 mb-3 js-repo-meta-edit js-details-
  <div class="repository-meta-content col-11 mb-1">
        <span class="col-11 text-gray-dark mr-2" itemprop="about">
          A distributed in-memory data store for the cloud
        </span>
  </div>
```

info-2:CLASS,text-emphasized,info,text

#서브프로젝트 페이지에서 프로젝트 정보가 위차한 곳의 정의 1(commit 등)

#아래의 태그내 class 부분의 정보가 저장된 내용을 수집

```
H6V6h5v2H8v5zM7 1C4.81 1 2.87 2.02 2.02 1.59 3.
.34.03-.67.09-1H.08C.03 7.33 0 7.66 0 8c0
    <span class="num text-emphasized">
      1,301
    </span>
    commits
</a>
```

info-3:CLASS,social-count,subInfo,text

#서브프로젝트 페이지에서 프로젝트 정보가 위치한 곳의 정의 2(watch 등)

#메인함수에서 0번 스크립터로 수집할 링크 정보를 수집 후, 세부 페이지에서 1번 스크립트에 해당하는 정보를 수집 가능하도록 플로우 구성

- 실행 화면 예시

```
Randolui-MacBook-Pro:ac randol$ ls
ac.jar          script.prj
Randolui-MacBook-Pro:ac randol$ java -jar ac.jar
log4j:WARN No appenders could be found for logger (org.apache.http.client.protocol.RequestAddCookies).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Netflix/chaosmonkey/
Description     Chaos Monkey is a resiliency tool that helps applications tolerate random instance failures.
Commits:        90
Branches:       8
Releases:       3
Contributors:   9
Watch:  239
Star:   2,561
Fork:   150
Netflix/sstable-adaptor/
Description     No description, website, or topics provided.
Commits:        18
Branches:       2
Releases:       0
Contributors:   0
Watch:  146
Star:   1
Fork:   1
Netflix/lemur/
Description     Repository for the Lemur Certificate Manager
Commits:        927
Branches:       52
Releases:       11
Contributors:   50
Watch:  233
Star:   640
Fork:   108
```

Github

http://github.com/onycom-ankus/ankus-crawler

Etc (Plan)

1. Web Crawler

조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램이다. 웹크롤러에 대한 다른 용어로는 앤트(ants), 자동 인덱서(automatic indexers), 봇(bots), 웜(worms), 웹 스파이더(web spider), 웹 로봇(web robot) 등이 있다.

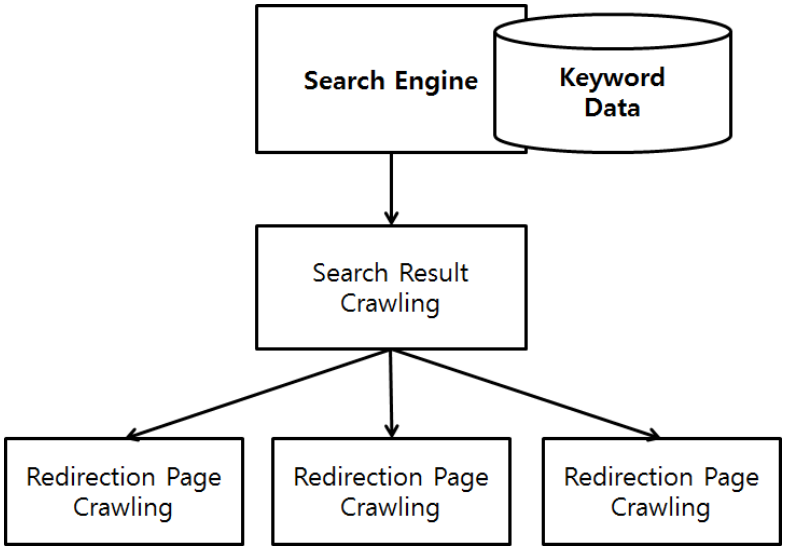
웹크롤러가 하는 작업을 웹 크롤링(web crawling) 혹은 스파이더링(spidering)이라 부른다. 검색 엔진과 같은 여러 사이트에서는 데이터의 최신 상태 유지를 위해 웹크롤링한다. 웹크롤러는 대체로 방문한 사이트의 모든 페이지의 복사본을 생성하는 데 사용되며, 검색 엔진은 이렇게 생성된 페이지를 보다 빠른 검색을 위해 인덱싱한다. 또한 크롤러는 링크 체크나 HTML 코드 검증과 같은 웹 사이트의 자동 유지 관리 작업을 위해 사용되기도 하며, 자동 이메일 수집과 같은 웹 페이지의 특정 형태의 정보를 수집하는 데도 사용된다.

웹크롤러는 봇이나 소프트웨어 에이전트의 한 형태이다. 웹크롤러는 대개 시드(seeds)라고 불리는 URL 리스트에서부터 시작하는데, 페이지의 모든 하이퍼링크를 인식하여 URL 리스트를 갱신한다. 갱신된 URL 리스트는 재귀적으로 다시 방문한다.

2. ankus crawler 웹문서 수집

2.1 구글, 네이버, 다음 검색엔진 활용 특정 키워드 관련 웹문서 수집(2017년 9월 적용예정)

대표적인 검색엔진인 구글(미개발), 네이버, 다음의 검색기능을 활용하여 사용자가 미리 정의한 특정 키워드에 대한 검색결과로 제공되는 웹문서를 수집한다.



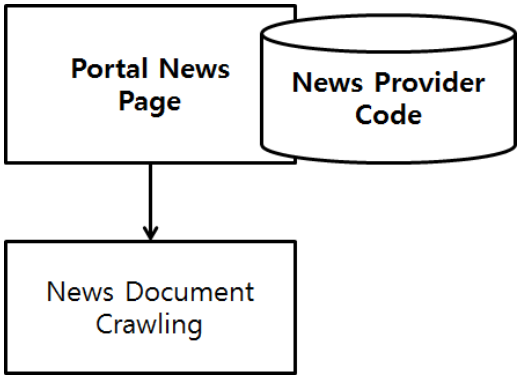
이와 같은 기능을 위해 (1)사용자가 미리 정의한 특정 키워드목록을 순차적으로 검색엔진에 입력한 후 검색 결과에 해당하는 웹페이지를 수집/분석한다. 이 때에는 검색엔진에서 사용하는 검색 날짜, 검색기준, 문서 분류 등과 같은 변수를 미리 정의하여 활용한다. (2)사용자가 해당키워드에 대해 최종적으로 수집을 원하는 웹문서에 대한 웹주소를 추출한다. 이 때의 웹문서의 종류는 웹사이트의 문서, 블로그 문서, 카페등과 같은 커뮤니티에 해당하는 문서 등이 있을 수 있다. 상세페이지에 대한 웹주소를 추출할 때에는 사용자가 원하는 수집대상에 맞는 규칙을 정의하고, 규칙에 맞는 상세 웹주소만을 추출하여 광고, 검색엔진 제공 연관사이트, 배너등과 같은 불필요한 웹주소 링크등이 수집되는 것을 방지한다.

수집대상이 되는 상세페이지의 웹주소 수집 완료 후, 해당하는 (3)웹주소의 웹문서를 순차적으로 수집한다. (4)수집되는 웹문서는 HTML파일 형태로 HDFS에 저장되고, (5)상세 웹문서에 대한 정의된 규칙에 따라 웹문서에서 문서제목, 문서내용, 댓글(reply) 등과 같은 웹 콘텐츠를 추출 및 DB화 한다.

2.2 국내 언론사 뉴스 수집(2017년 9월 적용예정)

웹환경에서 뉴스문서를 생성하는 국내 언론사는 약 100여개에 달한다. 이들 언론사의 웹페이지를

전부 모니터링하며 뉴스문서를 수집하기 위해, 네이버와 같은 대형 포털 사이트의 뉴스 페이지를 활용한다. 포털 사이트에서는 언론사로부터 뉴스문서를 실시간으로 제공받고, 동일한 형식으로 사용자에게 제공하기 때문에 웹문서 수집기의 유지, 보수, 관리가 용이하다.



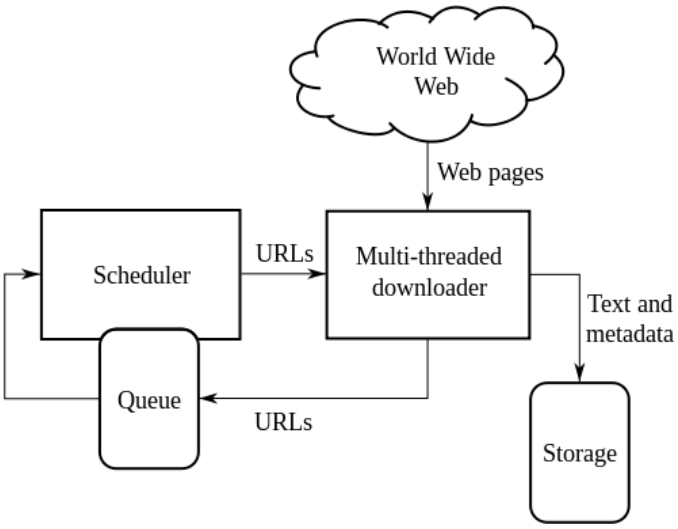
뉴스 수집 기능을 위해, (1)포털 사이트에서 사용하는 각각의 언론사 코드를 확보한 후 관리에 사용한다. 이후 사용자가 원하는 언론사의 코드만을 이용하여 뉴스 문서 수집에 활용한다. (2)사용자가 수집을 원하는 날짜 등과 같은 수집 규칙을 이용하여, 뉴스문서를 수집하고 (3)수집되는 뉴스문서는 HTML파일 형태로 HDFS에 저장한다. (4)상세 뉴스문서에 대한 정의된 규칙에 따라 뉴스제목, 뉴스 내용, 댓글(reply) 등과 같은 뉴스 콘텐츠를 추출 및 DB화 한다.

2.3 지정 웹사이트 게시판 모니터링 및 수집

사용자가 모니터링을 원하는 특정 사이트를 지정하게 되면, 해당하는 사이트의 웹문서를 수집한다. 메인화면에서부터, 홈페이지 전체 내용을 세부링크를 순차적/재귀적으로 탐색하는 방법을 사용하여 수집한다. 수집대상이 되는 웹 주소를 리스트화 하고, 중복되지 않는 웹 주소가 더 이상 존재하지 않을 때 수집을 종료한다. 초기 전체 웹사이트를 수집하고 난 후 주기적으로 갱신되는 웹사이트의 내용을 수집한다. 수집되는 웹사이트의 문서는 HTML파일 형태로 HDFS에 저장한다.

2.4 자동 수집 로봇(2017년 10월 적용예정)

불특정 다수의 홈페이지를 문서 및 특성 구분없이 HTML파일에 존재하는 세부링크를 재귀적으로 탐색하는 방식을 이용하여 수집한다. 수집되는 웹사이트의 문서는 HTML파일 형태로 HDFS에 저장한다. 자동 수집은 다양한 형태를 가진 웹 문서가 수집되기 때문에 제목외의 콘텐츠 내용을 별도로 추출하지 않는다.



3. ankus crawler 웹문서 정제

3.1 키워드 관리(2017년 9월 적용예정)

수집 기준 및 웹문서 정제를 위한 키워드를 관리한다. 저장된 키워드는 검색대상, 검색제외대상, 규칙포함대상, 규칙제외대상 등으로 분류된다. 검색대상 키워드는 검색엔진을 이용하여 웹문서 수집 시 사용하는 검색키워드이다. 검색제외대상 키워드는 검색엔진을 이용하여 수집되는 웹문서 중 수집을 하지 않을 문서를 분류하기 위한 키워드이다. 규칙제외대상키워드는 능동적인 웹 문서 수집시 웹 주소에 포함되지 말아야할 키워드이다.

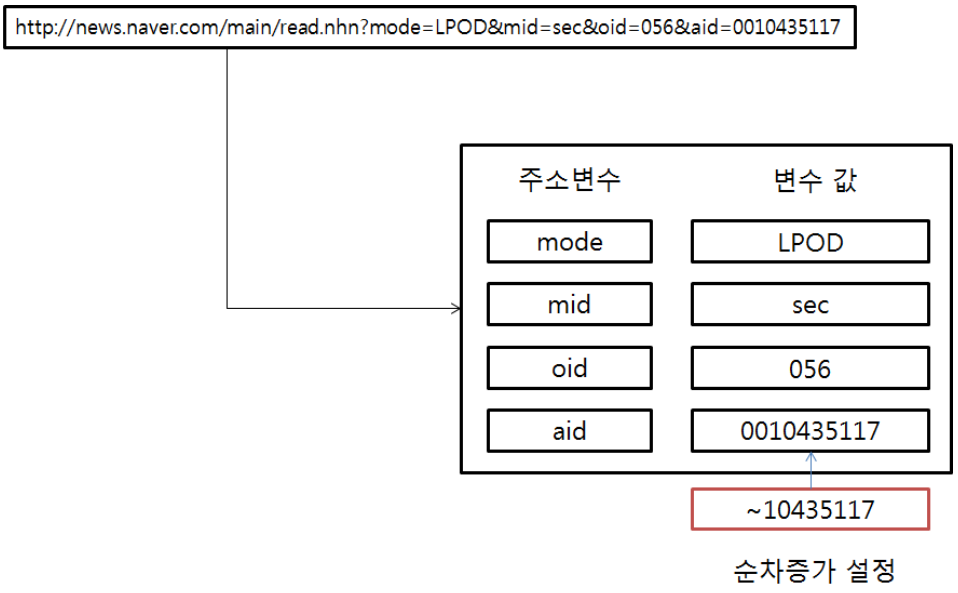
3.2 형태소 분석(2017년 9월 적용예정)

웹 문서 정제를 위하여 apache 라이센스로 공개된 오픈 소스 한글/영문 형태소분석를 각각 사용한다. 사용자는 수집대상이되는 웹문서에 맞게 언어를 선택할 수 있으며, 대상이되는 언어를 이용하여 형태소분석을 수행한다. 형태소분석이 되는 대상은 제목, 내용, 댓글등이며, 형태소분석을 위해 기본 사전 및 사용자가 입력한 키워드가 추가된다. 형태소분석시 사용되는 사용자 사전은 키워드 관리기능과 사용자 사전 관리 기능(품사₩t단어 형태로 저장)을 통해 관리한다.

4. ankus crawler 웹문서 상세기능

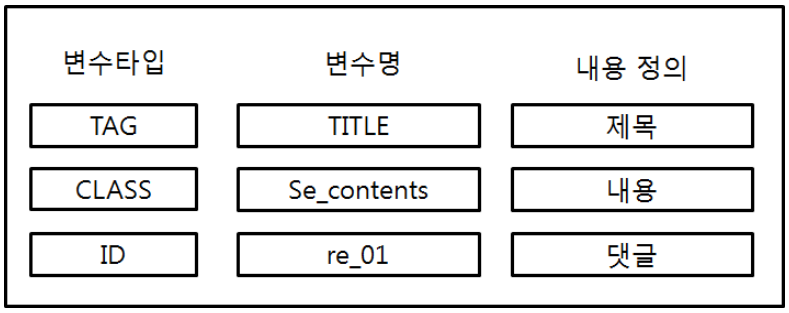
4.1 수집 규칙

사용자가 웹 주소의 예시를 입력하면 해당 사이트가 사용하는 웹 주소 변수를 분석하고, 이 중 콘텐츠를 식별할 수 있는 id, 문서번호 등에 임의의 규칙을 부여할 수 있도록 한다. 예를 들어 순차적으로 증가하는 문서번호의 경우 해당 변수명을 선택하고 순차증가 규칙을 적용하면, 해당 값을 순차적으로 증가시키면서 웹문서를 수집할 수 있다.



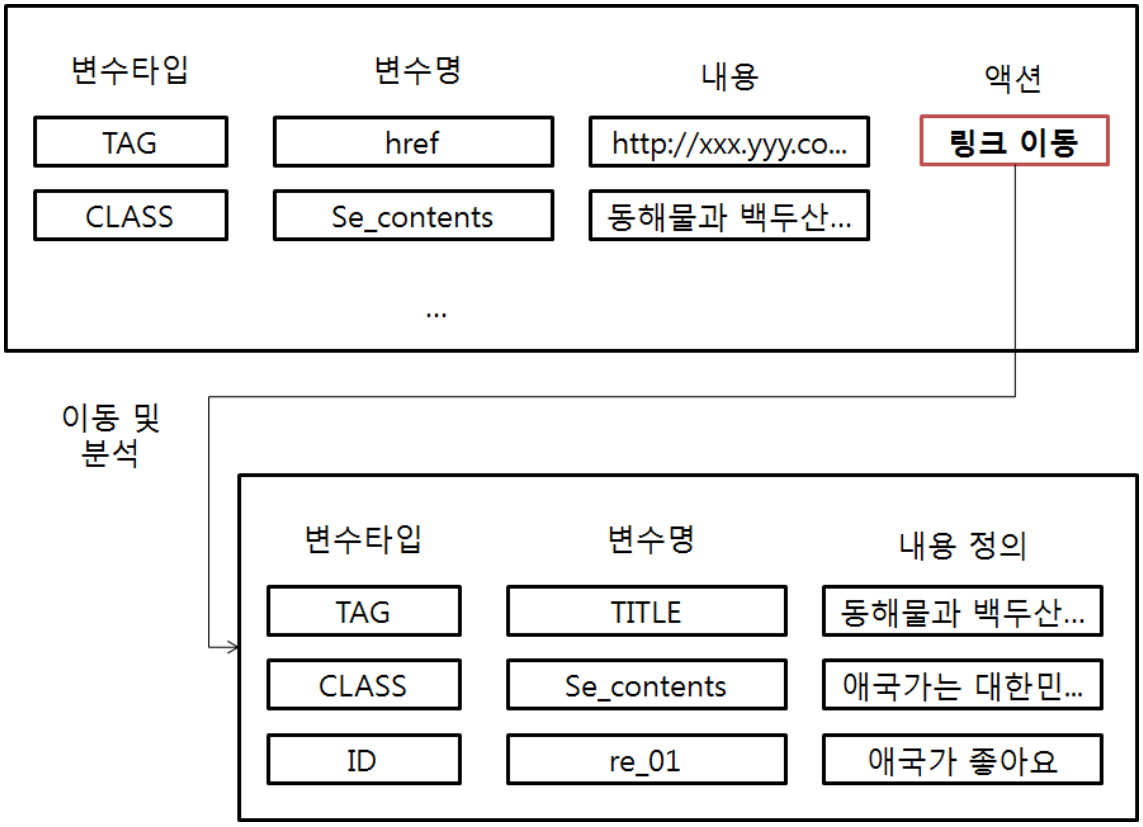
4.2 웹문서 규칙

웹문서는 HTML태그를 기준으로 작성되어 있으며, 이러한 형식의 문서는 jsoup 라이브러리를 통해 태그 단위, 클래스 단위, 아이디 단위등으로 구분 지을 수 있다. 사용자가 특정 웹 문서의 형태에서 제목, 내용 및 댓글과 같은 콘텐츠를 추출하여 DB화하기 위해서는 웹 문서 문석을 통해 추출할 내용을 정의할 수 있는 규칙을 정의해야한다. 이를 위해 aC에서는 태그, 클래스, 아이디등의 단위와 추출된 내용에 대한 정의를 사용자가 손쉽게 할 수 있도록 스크립트(GUI)형태로 제공한다. 사용자는 추출할 내용이 있는 곳의 태그등을 정의하고 해당하는 내용에 대한 분류를 정의함으로써 웹 문서에서 자신이 원하는 내용을 추출 할 수 있다.



4.3 수집 설정 도우미

웹문서 수집을 위해서는 주소방식과 웹 문서를 구성하고 있는 다양한 tag에 대한 충분한 지식을 가지고 있어야 한다. aC에서는 이러한 지식을 갖추지 못한 일반 사용자도 효과적으로 웹 문서를 수집할 수 있도록 수집 설정 도우미 기능을 제공한다. 수집 설정 도우미는 각 웹 문서에 존재하는 링크나, 텍스트 내용을 추출할 수 있는 변수 타입, 변수명별로 사용자에게 해당하는 내용을 제공한다. 사용자는 제공하는 내용을 선택하고, 정의함으로써 수집 단계별 추출 규칙 등을 설정할 수 있다.


변수타입
변수명
내용
액션
TAG
href
http://xxx.yyy.co...
링크 이동
CLASS
Se_contents
동해물과 백두산...
...
이동 및
분석
변수타입
변수명
내용 정의
TAG
TITLE
동해물과 백두산...
CLASS
Se_contents
애국가는 대한민...
ID
re_01
애국가 좋아요