**종합 설계 프로젝트**

**Posting Notifier**

**11조**

2014313795 김성섭

2014310312 김하림

2014311062 이진욱

1. **소개**

‘Posting Notifier’는 사용자가 특정 사이트로부터 갱신되는 정보를 계속적으로 알림으로 받고 싶을 때 이용할 수 있는 프로그램입니다.

이 프로그램은 3가지 단계로 구성되어 있습니다.

1. (로컬)웹 서버 (사용자 이용) – 사용자가 정보를 입력해 카카오 메시지를 받을 계정과 크롤링할 웹 사이트의 주소를 얻습니다.
2. 웹 크롤러, 파서 - 특정 웹 사이트를 주기적으로 크롤링하고 내용을 비교해 업데이트 된 정보를 추출합니다.
3. 카카오 메시지 연동 – 크롤링 과정에서 업데이트 된 정보가 있다면 해당 웹 사이트의 링크를 웹 서버에서 얻은 계정으로 카카오톡 메시지로 보냅니다.
4. **관련 연구, 프로그램**

* 연구

고성능 웹크롤러의 설계 및 구현 [1]

인터넷의 성장에 따라 웹 문서 양이 거대해지면서 기존의 웹 크롤러 기술에 대해 수집 속도를 최대화할 수 있는 동적 스케줄링 기법과 함께 시스템과 네트웍 과부하를 막기 위한 동적 부하 분산 기법을 제안함

빅데이터 분석 기반의 정보 검색을 위한 웹 크롤러 서비스 구현 [2]

대학생 및 취업준비생의 대회활동, 공모전, 장학금에 대한 효율적인 정보 수집을 위한 웹 크롤러 서비스 방식을 제안함. 빠른 속도를 유지함과 동시에 중복 정보를 피하기 위해 Jsoup 트리 분석과 Json 형식의 데이터 전송 방식을 이용하였다.

웹크롤러의 수집주기 최적화 [3]

데이터가 폭발적으로 증가하는 시대에 맞춰 최신의 정보를 적은 오버헤드로 얻는 것을 중접으로 두었고, 과거 히스토리로부터 웹사이트 변경이 집중되는 시간을 파악해 웹 수집 주기를 결정하는데 반영하는 것을 주요 골자로 두었다. 해당 논문에서는 JAVA를 이용해 크롤러를 제작하였다.

* 관련 프로그램

상용 웹 크롤러 – Konan Web Crawler

자바스크립트를 이용한 웹 크롤러로, 지능형 수집 옵션, 분류 옵션, 규칙 설정, 통계 분석 기능 등이 있다. 현재 안전행정부에서 사용하는 웹 크롤러이다.

오픈 소스 웹 크롤러

1. Scrapy (Python)

사용하기 쉽고, 문서화가 잘 되어있다. JSON, XML, CSV 포맷 등을 지원한다. 분산환경, 동적확장성, 지속성이 부족하고 큰 데이터를 추출하기 어렵다.

1. Heritrix (Java)

문서화가 잘 되어있고, 설정이 쉽다. 성능 또한 좋고 안정적이며 분산 크롤링이 지원된다. 그러나 연속 크롤링이 지원되지 않고, 수동으로 서버 구조를 설정하고 숫자를 할당해야 한다.

1. Apache Nutch (Java)

하둡 기반의 분산 크롤링을 지원한다. 그러나 문서화가 매우 잘 안 되어있으며 예제도 부족하고, 속도도 느리다.

**3. 구현 계획**

웹 서버는 Python3 로 사용할 수 있는 Django를 이용해 제작합니다.

웹 크롤러는 Python3로 사용할 수 있는 Selenium을 이용해 제작합니다.

카카오 메시지 연동은 카카오 developers에서 사용할 수 있는 Rest API를 사용합니다.

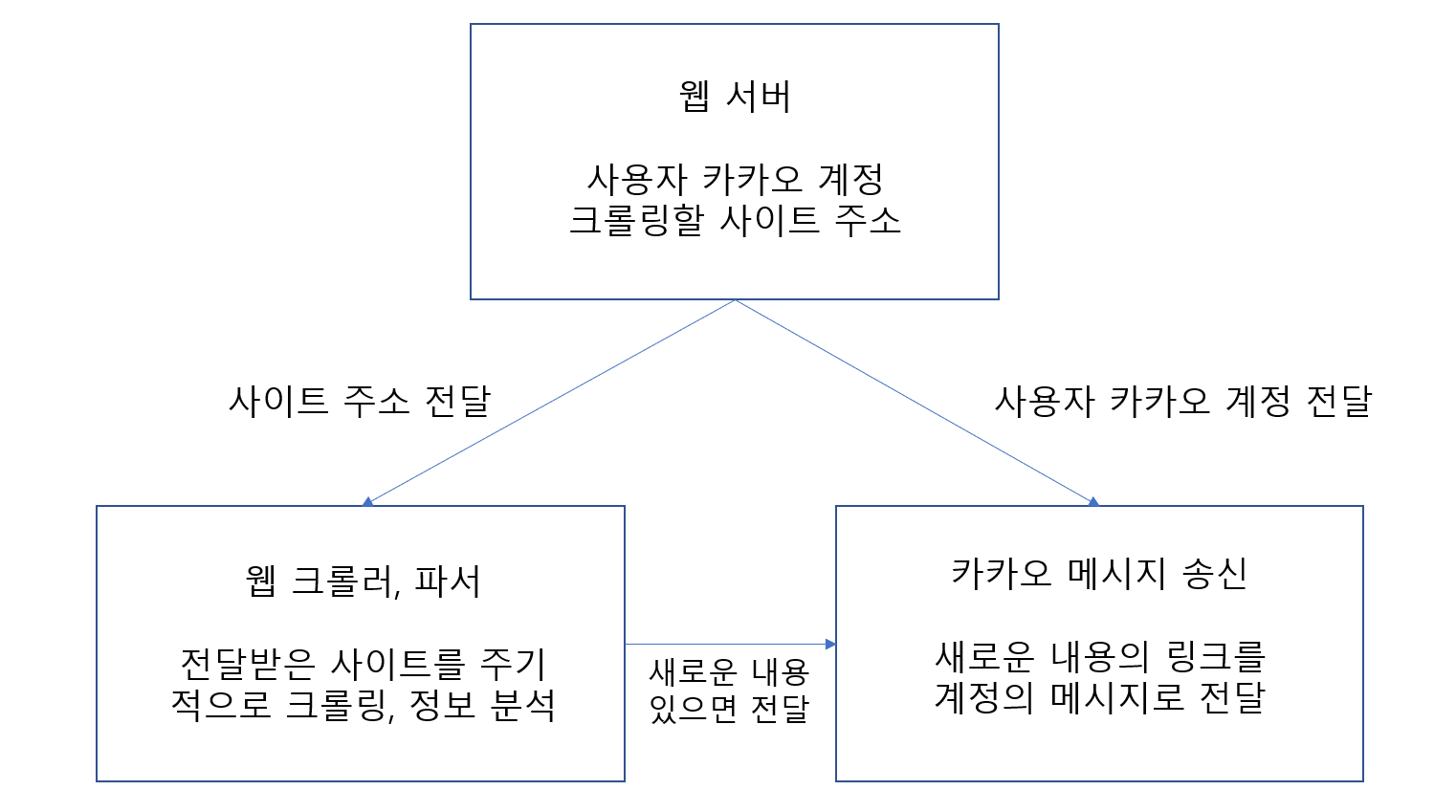


그림 : Posting Notifier의 구조도

부가적인 기능

1. 여러 사이트에 대한 크롤링 가능
2. 키워드 설정으로 필터링 가능 – 기본적으론 업데이트가 있으면 바로 알림이 가지만 키워드를 설정하면 추가된 컨텐츠에 해당하는 내용이 없을 시 필터링이 가능합니다.

**참고 문헌**

[1] 김희철, 채수환, “고성능 웹크롤러의 설계 및 구현”, 디지털콘텐츠학회논문지, 2003년

[2] 김희숙, 한나, 임숙자, “빅데이터 분석 기반의 정보 검색을 위한 웹 크롤러 서비스 구현”, 디지털콘텐츠학회논문지, 2017년

[3] 조완섭, 이정은, 최치환, “웹크롤러의 수집주기 최적화”, 한국콘텐츠학회논문지, 2013년