韓國IT서비스學會誌
第13卷 第1號
2014年 3月, pp.147-159
http://dx.doi.org/10.9716/KITS.2014.13.1.147

웹 페이지 비교통합 기반의 정보 수집 시스템 설계 및 개발에 대한 연구

장 진 욱*

A Study on Design and Development of Web Information Collection System Based Compare and Merge Method

Jin-Wook Jang*

■ Abstract ■

Recently, the quantity of information that is accessible from the Internet is being dramatically increased. Searching the Web for useful information has therefore become increasingly difficult. Thus, much research has been done on web robots which perform internet information filtering based on user interest.

If a web site which users want to visit is found, its content is searched by following the searching list or Web sites links in order. This search process takes a long time according as the number of page or site increases so that its performance need to be improved.

In order to minimize unnecessary search with web robots, this paper proposes an efficient information collection system based on compare and merge method. In the proposed system, a web robot initially collects information from web sites which users register. From the next visit to the web sites, the web robot compares what it collected with what the web sites have currently. If they are different, the web robot updates what it collected. Only updated web page information is classified according to subject and provided to users so that users can access the updated information quickly.

Keyword : Web Robot, Web Searching, Web Crawler, Information Collection System, Compare and Merge Method.

논문투고일 : 2013년 12월 30일 논문수정완료일 : 2014년 03월 01일 논문게재확정일 : 2014년 3월 04일
* 건국대학교 정보통신대학교 인터넷미디어공학부 산학교수

1. 서 론

인터넷을 통한 정보교환이 활발해 지면서 필요한 정보를 짧은 시간에 손쉽게 수집 및 검색에 대한 필요성이 지속적으로 요구되고 있다.

사용자는 반복적인 주제의 정보수집과 검색을 하는 과정에서 최소한의 리소스로 원하는 정보를 얻고자 한다. 사용자가 웹 검색을 통해 필요한 정보를 찾았다면 그 페이지를 북마크 하고 그 페이지에 새롭게 업데이트되는 내용을 보고 싶어 하게 될 것이다. 그래서 웹에서 사용자가 원하는 정보를 검색하기 위하여 다양한 검색도구가 필요할 것이며 이러한 사용자 요구사항을 충족시키기 위하여 정보검색을 위한 도구로서 다양한 인터넷 검색엔진이 개발되어 사용되고 있다.

본 논문에서는 이런 점을 충족하기 위하여 사용자 맞춤형 관심 사이트를 대상으로 최초 수집한 데이터베이스를 기준으로 비교 통합한다. 그래서 수집시간 단축 및 중복수집 최소화를 목표로 정보수집 시스템 설계 및 개발을 하였다.

2. 관련 연구

2.1 에이전트 시스템 기술

에이전트(agent)란 기존의 작업 처리방식에서 사용자가 원하는 것이 무엇인지를 자동으로 판단하여 사용자의 작업을 대행해 주는 프로그램으로써 에이전트간 정해진 규약에 따라 통신하며 필요한 에이전트를 찾아 네트워크를 항해하는 프로그램으로 알려져 있다[1, 4].

본 논문에서 보조 에이전트(assistant agent)는 사용자의 작업을 돕는 프로그램으로 정의한다. 이는 어떤 문제에 대하여 인간을 대신하여 해결하는 프로세서로 정의하기도 하며 하나의 응용프로그램으로 편리한 컴퓨터 사용 환경을 제공하는 것을 목적으로 하고 있다[3, 4].

기존의 컴퓨터 환경을 제공하는 것을 목적으로 하고 있다. 기존의 편리한 컴퓨터 환경을 제공하는 예로는 자료검색 에이전트, 뉴스 그룹 프로그램 에이전트 등을 들 수 있다. 그러나 현재의 보조 에이전트 시스템의 특징은 보다 능동적으로 사용자의 작업을 대행해 주는 프로세서들로 로봇이라 칭한다. 이러한 로봇들은 사용자 지식을 기반으로 하여 사용자가 요구하는 작업을 대행해 주는 역할을 담당한다.

2.2 WebCQ

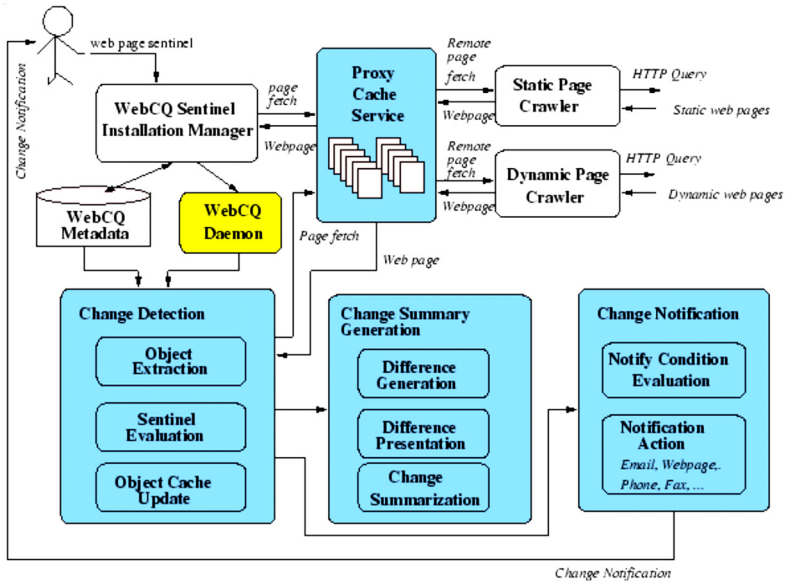
WebCQ[1]는 정적인 페이지나 동적 웹 페이지 변화들에 대한 여러 가지 다양한 변화를 모니터링하고 추가하는 기능을 가지고 있다. WebCQ의 주된 기능은 페이지의 갱신 여부를 파악했을 때 사용자에게 그 내용을 전달하는 기능과 해당페이지의 객체추출 알고리즘을 이용하여 객체의 위치를 파악하고 웹 페이지 내에서 관심 있는 객체를 파악하는 기능이 있으며 마지막으로 갱신 탐지 객체를 통하여 어떤 페이지가 갱신되었는지 파악하는 기능을 가지고 있다[13].

WebCQ는 웹 페이지의 정보 모니터링과 모니터링된 정보를 전달하는 시스템으로 구성되어 있다. 웹 페이지의 변화를 발견하고 검색하기 위하여 4개의 중요한 컴퍼넌트로 구성되어 있다. 첫 번째 웹 페이지의 변화를 탐지하고 검색하는 로봇, 두 번째 시간의 정의로 사용자에게 새로운 정보를 서비스하는 모듈, 세 번째 웹 페이지들의 여러 가지 유형을 지원할 수 있는 페이지 변환 탐지 로봇, 마지막으로 네 번째로 사용자가 관심을 가지고 정보가 변화가 되었다면 추천하는 모듈로 구성되어있다.

WebCQ의 경우 로봇을 이용하여 웹 페이지를 가져온 후 그 페이지의 변환 연부를 파악하게 된다.

---

1) 웹 정보의 변경내용을 모니터링 하여 감지하는 기술 (Detecting and Delivering Information Change on the Web).



[그림 1] WebCQ 시스템 구조도

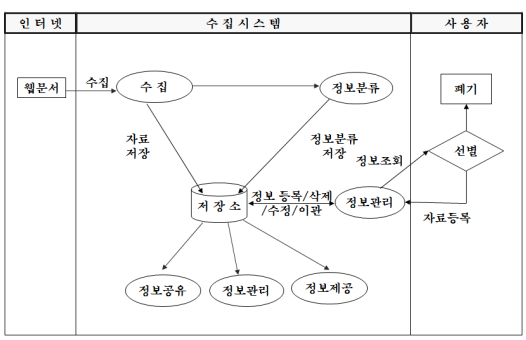
따라서 WebCQ에서 페이지 변환 여부를 파악하는 방법은 첫째, 페이지 내의 링크 수의 변화, 둘째, WebCQ에서는 페이지를 가지고 그 페이지의 수정날짜를 통해 페이지의 변화를 파악한다. 셋째, 웹주소가 나타내는 페이지의 콘텐츠 변화를 통해 페이지의 변화를 파악하는데 콘텐츠의 수정이나 삭제, 추가를 통하여 페이지를 나타내는 파일의 크기의 변화를 통해 그 변화를 알아내는 방법이다.

3. 웹 수집 시스템 연구

3.1 웹 수집 시스템

인터넷을 통한 정보의 폭발로 이제는 정보의 존재유무가 문제가 아니라 어디에 무슨 정보가 있으며 필요한 정보를 선별하는 것이 더 중요한 문제가 되었다. 이를 위하여 검색 서비스가 발전하게 된 것이다. 현재 개발된 많은 검색 서비스는 각각 범위(coverage)면이나 정밀도(precision)면에서 많은 차이를 보이고 있다.

본 논문에서 웹 로봇(web robot)을 이용하는 이유는 최대한 빠른 시간 내에 사용자가 원하는 정보를 얻기 위한 것이다. 기존의 웹 로봇의 경우 사용자가 선호하는 페이지의 주소(URL)를 리스트 형태로 주어지면 순차적으로 해당 사이트를 반복 방문한다. 따라서 서버의 부하와 검색시간이 오래 걸리는 문제점이 있었다. 이러한 문제점을 보완하기 위해서 본 논문에서는 웹 페이지의 변화를 탐지하여 로봇의 불규칙한 검색을 최소화하기 위한 웹 페이지 변환탐지 정보수집 시스템을 구현하였다. 아래의 [그림 2]은 웹 로봇을 이용한 정보수집 시스템의 구성 흐름도이다.



[그림 2] 시스템 구조도

웹 로봇은 사용자의 키워드에 대한 문서수집이 아닌 시스템에 등록되어 있는 인터넷 신문사, 연구소 등의 페이지에 대하여 하이퍼링크(hyper link)를 기준으로 해당하는 하위 페이지들에 대하여 수집을 한다. 인터넷 정보의 특징은 방대할 뿐만 아니라 매시간 정보가 변화하는 속성을 가진다. 최대한 많은 정보를 보유하려는 기존의 검색엔진에서는 특정사용자의 관심영역에 부합된 정보수집에 한계가 있다.

본 논문에서는 사용자가 관심을 가지는 인터넷 사이트와 특정 신문사만을 집중적으로 모니터링 함으로써 정보가 업데이트 되는 즉시 사용자에게 제공하기 위한 엔진을 구축하였다. 인터넷상에 존재하는 방대한 양의 데이터 중에서 사용자가 원하는 정보를 신속하고 정확하게 검색하기 위해서 본 논문에서는 사용자가 관심을 가지는 사이트의 주소를 기준으로 모니터링 사이트 목록을 만들었다. 사용자 모니터링 사이트 목록 사용자가 직접 입력한 정보를 기반으로 신규 정보의 갱신을 알려준다. 페이지의 변화 여부를 파악하기 위해서는 어떠한 형태로든지 페이지의 내용 즉 페이지의 소스 프로그램을 가져와야 한다. 가져온 소스의 웹 페이지 태그를 기반으로 변화 여부를 파악할 수 있기 때문이다. 수집된 웹 페이지는 링크구조에 따라 계층구조 생성기로 전해지는데 이 계층구조 생성기는 페이지 내의 프로그램을 계층 구조의 형태로 만든다. 이렇게 만들어진 페이지의 계층구조는 이후에 페이지 파서(parser)에서 페이지 변화 여부를 파악하는데 사용되어진다.

계층 구조의 형태로 만들어지게 되면 이후에 설명하겠지만 가장 하위 레벨에 있는 것은 해당 문서의 콘텐츠가 위치하게 되는데 이것은 사용자가 등록하는 과정에서 자신이 선호하는 웹 페이지에서 어느 한 콘텐츠만 선택하여 그 변화까지 파악할 수 있다. 이렇게 페이지 갱신 여부가 결정되고 페이지가 변했다면 분석을 통하여 사용자가 원하는 정보인지를 선별하여 수집하게 된다. 만약 페이지의 변화가 없다면 모니터링 웹 로봇은 그 페이지를 생략하고 다음 수집 리스트로 넘어가게 된다.

웹 로봇은 사용자에 의해 등록된 특정 사이트만을 집중적으로 수집한다. 이러한 기능은 빠른 갱신주기를 가지는 신문사나 특정 게시판, 특정 웹 사이트의 정보검색 및 수집에 효과적이며 정보가 업데이트 되는 즉시 분석을 거쳐 사용자에게 정보를 서비스할 수 있도록 구성하였다.

기존 연구가 다사용자를 대상으로 한 서비스 시스템 이었다면 본 연구는 개인화된 맞춤형 시스템에 활용할 수 있다는 점에서 차이점이 있다.

3.2 웹 수집 시스템 구성

정보 수집 시스템에서 미리 등록해 놓은 주소 수집 리스트 정보를 참조하여 해당 사이트 범위만을 수집한다. 이러한 수집 시 가장 주의할 점은 해당되는 사이트의 하부 주소범위 설정이다. 모니터링의 성격상 해당 주소 웹 페이지를 기반으로 관련 링크들을 추출하고 정리한 후 다음 모니터링 대상 페이지를 찾아가게 되는데 이러한 과정에서 다음과 같은 문제점이 나타난다.

첫째, 수집범위를 넘어서는 수행으로 인한 서버 부하가 발생하며, 둘째, 게시판과 같은 반복적인 페이지에 대한 순환적인(recursive) 모니터링 현상과, 셋째, 모니터링 페이지에 대한 반복적인 모니터링으로 인한 시스템 자원낭비가 발생한다.

본 논문에서는 모니터링 웹 로봇에서 나타나는 위와 같은 문제점들을 해결하기 위하여 다음과 같은 방법을 사용 하였다.

첫째, 웹 수집의 범위는 모니터링 되는 페이지의 주소정보를 참고하여 해당 사이트의 깊이(depth)를 규정하고 수행범위를 정형화 시켰다. 둘째, 반복적이고 순환적인(recursive) 수집 현상을 막기 위하여 해당 사이트에 대한 모니터링 로그정보를 저장하며 웹 수집 시 순환적인 수집문제를 해결 하였다. 셋째, 매 모니터링 수행시마다 해당 사이트의 모든 페이지를 수집하는 방식을 탈피하여 수집대상 페이지의 수집 로그정보를 저장함으로써 새로이 업데이

트되거나 추가된 페이지만을 추가적으로 수집하여 기존의 데이터베이스에 추가하는 방식을 택하였다.

3.3 웹 수집 시스템 동작 개요

웹 수집 시스템은 주어진 웹사이트를 대상으로 링크(hyper link)를 운행하면서 웹 서버로부터 가지고온 데이터를 저장하는 기능을 제공한다. 인터넷 정보수집기의 동작 방식은 다음과 같다.

첫째, 사용자는 수집할 웹사이트의 주소를 지정한다. 웹 수집기는 사용자가 지정한 사이트를 수집할 목록에 저장한다. 둘째, 수집기는 수집할 웹사이트 목록에서 한 개를 꺼내어 해당 웹사이트 문서를 웹 서버로부터 받아온다. 셋째, 수집한 문서가 웹 문서인 경우 문서를 파싱해서 링크정보를 추출한다. 셋째, 링크정보를 수집할 웹사이트 목록에 추가한다. 단, 이미 방문한 사이트를 목록에 추가하지는 않는다. 다섯째, 수집 할 사이트 목록이 더 이상 없는 경우에 종료하고 수집할 웹사이트 목록이 계속 있으면 두 번째 단계에서 다시 시작한다.

4. 정보수집 시스템 설계

4.1 웹 문서 분석

대부분의 웹 문서는 관련이 있는 내용의 참조를 위해 링크를 포함하고 있다. 이러한 링크는 연결된 웹 문서들 간에 연관성이 있음을 암시한다. 그러나 기존의 검색 시스템에서는 대부분 이러한 하이퍼링크로 연결되어 있는 웹 문서에 대한 정보를 고려하지 않는다.

본 논문에서는 웹 문서의 하이퍼링크는 관련 있는 웹 문서를 연결하도록 작성되어져 있다는 가정하에 하이퍼링크를 이용한다. 하이퍼링크는 웹 문서의 주소와 텍스트로 이루어져 있다. 보통 앵커 텍스트(anchor text)[2]는 해당 웹 문서에 대한 간략한 정보로 표현되어지고, 웹 문서의 타이틀(title) 태그에는 제목이 기입되어 있다.

웹 문서 분석기는 text와 title 태그를 이용하여 키워드를 추출한다. 일반적인 검색시스템에서는 하나의 웹 문서에 대하여 단어의 빈도나 구문의미를 분석하여 해당 문서에 대한 다중 키워드를 추출하나 여기서는 개념의 구성이 목적이므로 단일 키워드만을 추출한다.

```
document A
  <a href=http://URL B>
   Information Retrieval Papers </a>
  document B
  <title> IR Papers </title>
```

이 방법은 기존의 키워드 검색 시스템들에서 사용되는 키워드 추출 방법보다 단순하다. 그러나 추출된 키워드가 단순하기 때문에 웹 문서에 내포된 여타 키워드에 대한정보가 생략되는 단점을 가진다. 웹 문서를 대표하는 키워드가 추출되면, 본 논문에서는 하이퍼링크를 이용한다. 즉, 웹 문서마다 자신을 대표할 수 있는 태그(tag)를 만든 후, 웹 문서의 연결 관계를 태그와 태그의 연결 관계로 대치한다. 그리고 태그 연결 관계를 이용한다.

즉, 개념의 연결 강도가 더 높아지게 된다. 연결 강도가 높은 개념은 실세계에서도 밀접한 관계가 있을 확률이 높다. 따라서 연결강도에 분기 값을 주어 추출되는 개념을 간략하게 함으로써 인터넷상에 존재하는 많은 관계들 중 대표적인 것들만을 보여줄 수 있다.

개념에 의한 관계를 분석해 보면 첫째, 일반화-특수화 관계가 있다. 예를 들어 키워드 A를 가진 웹 문서에서 키워드 B를 가진 웹 문서로의 하이퍼링크가 있다면, A는 B를 포함한다. 즉 A는 B의 일반화이고, B는 A의 특수화이다. 둘째, 평등 관계가 있다. 예를 들어 키워드 A를 가진 웹 문서와 키워드 B를 가진 웹 문서가 서로 하이퍼링크로 연결되어 있는 관계를 말한다.

2) Anchor의 <A>태그와 </A> 사이에 놓이는 문장을 지정하는 HTML 문장.

지금까지 웹 문서에서의 전처리 과정은 단순히 웹 문서 태그를 찾아내어 이 태그안의 키워드들은 실제 관심 키워드로 추출되지 않도록 제거하는 작업으로 수행되어왔다. 그러나 웹 문서에서 표현 할 수 있는 여러 기법들이 많이 등장함에 따라 단순히 웹 문서 태그를 사용하지 않고도 웹 문서에 기능적은 부분을 추가하고 있는 기술들이 문제가 될 수 있다.

바로 스크립트 언어 기법이 그것인데 이는 단순히 웹 문서 내에 웹 문서 태그 이외의 내용으로 작성하여 웹 문서의 출력 기능에 추가되는 기능이다. 그러므로 이 기법을 사용하고 있는 웹 문서에서 기존 방법의 키워드 추출 작업을 수행하였을 경우에는 스크립트 언어 코드의 내용이 키워드로 추출되는 결과가 발생한다. 또한 웹 문서에서 이 기법을 반복적으로 사용한다면 일정 가중치 이상이 매겨진 비중을 가진 중요 키워드로 판단될 수 있는 상황에까지 이르게 된다.

뿐만 아니라 대부분의 웹 페이지들이 스크립트 언어를 사용하여 동적인 홈페이지를 구축하고 있는 추세인데 스크립트 언어는 사용자의 관심 키워드가 될 수 없으므로 제거되어야 한다. 본 시스템은 가장 대중적으로 널리 이용되는 자바스크립트에 대한 처리 기능을 수행한다. 웹 문서들을 읽을 때 자바스크립트의 키워드를 검색하여 스크립트가 삽입된 부분은 알아내고 이 부분을 제거한 후 다음 처리 단계의 입력으로 사용한다. 자바스크립트는 웹 문서 “<HEAD>” 태그 안에 선언이 되는데 “<script>”와 “</script>” 사이에 스크립트의 내용이 선언된다. 그러므로 입력 스트림을 읽으면서 “<script>” 태그가 발견되면 “</script>” 태그가 나올 때 까지 무시하면 자바스크립트를 제거 할 수 있다.

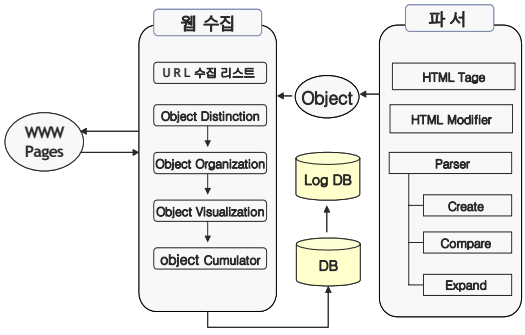
위의 방법을 사용하여 웹 문서내의 기술된 스크립트언어 코드를 제거하는 문제는 해결할 수 있다. 그러나 웹 문서에서 나타나는 또 다른 문제가 있을 수 있는데 이는 웹 문서에 특수문자가 사용될 경우이다. 웹 문서에서는 각종 특수문자를 사용하는 경우 특수문자는 사용자의 관심 키워드가 될 수 없으므로 제거되어야 한다. 특수문자는 “&”로 시작을 하므로 “&”로 시작하는 문자열이 나오면 “;”로 만날때까지 제외시키는 방법을 사용하였다.

4.2 웹 로봇

본 논문에서 가장 중요한 부분을 담당하는 부분이 사용자가 등록한 사이트의 변경 여부를 파악하는 부분이다. 웹 페이지의 갱신 판단 방법에는 흔히 “HTML header”파일을 읽어 수정 날짜와 시간을 비교하여 그 페이지의 변화여부를 파악하거나 또는 소스를 가지고 있는 파일의 크기 즉, 파일의 바이트 수의 변화로 파악 할 수 있었다. 그러나 파일의 크기 변화를 탐지하는 것에 대해서는 문제가 없으나 파일내용에 변화가 있는 경우 그 여부를 파악하지 못하는 경우가 발생할 수 있다.

본 논문에서는 이러한 문제점을 여러 각도로 보완하고 갱신여부 판단을 위하여 다음과 같은 연구를 하였다. 사용자가 등록한 수집대상 주소정보를 기반으로 웹 페이지의 계층구조를 방문함과 동시에 해당 사이트의 태그들을 오브젝트로 구별 및 생성 하고 이후에 다시 방문하였을 때 비교하는 방법을 사용한다.

즉, 헤더(header) 파일이나 단순한 콘텐츠의 삭제나 수정, 삽입으로 인해 파일 크기 변화가 아니라 해당 페이지 내에 존재하는 콘텐츠들의 변화를 직접적으로 비교하여 그 갱신여부를 판별하게 되는 것이다. 다음은 갱신여부 판단을 위한 웹 로봇의 구조도이다.

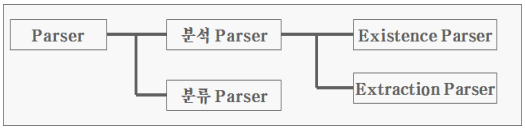


[그림 3] 웹 로봇 구조도

[그림 3]에서 사용자가 등록한 주소 수집 리스트를 페이지를 웹 수집 하여 해당 페이지의 태그를 가져오게 된다. 이렇게 수집되어진 페이지의 소스를 계층구조 생성기를 이용하여 페이지의 계층 구조를 만들게 되고 페이지 변화 분석기를 통하여 이전의 계층 구조와 비교하여 변화된 페이지 여부를 파악하고 페이지가 변화되었다면 변화된 콘텐츠만을 웹 로봇에게 전달하게 되고 이를 최종적으로 해당하는 콘텐츠의 웹 페이지를 검색하여 문서로 저장하게 된다.

웹 로봇에 의해 수집된 웹 페이지는 파서(parser)에 의해 기존 수집된 웹 문서의 태그정보와 새롭게 수집된 웹 페이지의 정보와 비교하여 수집하는 과정에서 다음과 같은 구조로 분석 및 분류가 이루어진다.

그리고 파서는 [그림 4]와 같이 크게 태그분석을 위한 파서와 분류를 위한 파서로 구별되는데 분석 파서는 웹 로봇이 방문한 웹사이트에 해당정보가 존재하는지의 여부를 분석하는 “Existence Parser”와 해당정보가 있는 경우에 그 정보를 추출하는 “Extraction Parser”로 구성된다. 분류 파서는 수집하고자 하는 정보가 속한 웹사이트의 분류와 설정한 분류가 다를 경우 설정한 분류내로 해당 정보를 일치하는 역할을 한다.



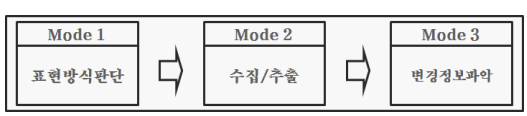
[그림 4] 파서의 종류

페이지의 변화 탐지 방법은 계층구조트리의 하위 레벨 노드의 변화를 통해 알 수 있다. 우선은 페이지 파일 크기의 변화를 통해 알 수 있는 방법이 있다. 이 방법은 페이지의 콘텐츠가 수정, 삭제, 추가 등을 했을 때 파일의 크기 변화를 통해 파악하는 방법이 있다.

그러나 이 방법은 페이지 내의 콘텐츠가 수정, 삭제, 추가되었다 하더라도 파일의 크기 변화가 없을 수 있는 경우가 발생한다. 이러한 문제점으로 인해 이와 병행하여 몇 가지 방법을 제시하고자 한다. 첫째, 페이지 내의 주소변화를 파악하는 것이다. 둘째, 콘텐츠 내의 키워드 개수를 비교하여 변화 여부를 파악한다. 사용자가 검색하고 싶어 하는 키워드를 기준으로 개수를 확인하여 변화여부를 파악한다. 이 방법에는 앞에서 파일의 크기 변화에서 수정을 할 때 파일의 크기 변화가 없을 수 있다는 문제점과 같은 문제를 가지고 있다. 이를 보완하기 위해 콘텐츠의 길이, 명사의 개수, 조사의 개수 및 링크된 웹 페이지의 개수 등을 이용하여 변화 여부를 파악한다. 지금까지 언급한 방법들을 복합적으로 사용하여 페이지의 변화 여부와 콘텐츠의 변화 여부를 파악한다.

웹 로봇으로 웹 페이지를 가져오게 되면 태그정보가 만들어지게 된다. 기존에 만들어졌던 태그정보와 새롭게 생성된 태그정보의 비교를 통해 변화여부를 파악하게 된다. 트리의 마지막 레벨에 위치한 노드에는 콘텐츠가 위치하게 된다. 이 콘텐츠를 각각 하나의 오브젝트 단위로 구분될 수 있게 되고 이 오브젝트로 갱신여부를 파악하게 된다. 아래 [그림 5]에서 웹 로봇으로부터 웹 페이지 정보를 가져와 계층 구조로 이루어진 웹사이트의 갱신된 오브젝트단위로 콘텐츠를 가져오게 되고 웹 페이지의 갱신탐지를 통해 변화된 페이지라면 기존의 계층 구조 트리를 버리고 새로운 계층 구조 트리를 업데이트하게 되는 과정을 나타낸다.

웹 로봇으로 웹 페이지를 가져오게 되면 태그정보가 만들어지게 된다. 기존에 만들어졌던 태그정보와 새롭게 생성된 태그정보의 비교를 통해 변화여부를 파악하게 된다. 트리의 마지막 레벨에 위치한 노드에는 콘텐츠가 위치하게 된다. 이 콘텐츠를 각각 하나의 오브젝트 단위로 구분될 수 있게 되고 이 오브젝트로 갱신 여부를 파악하게 된다.

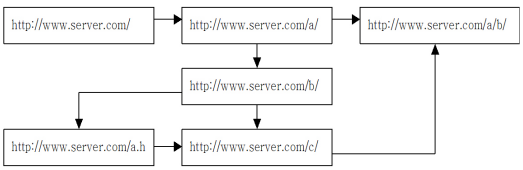


[그림 5] 웹 페이지 갱신 판단 수집 단계

다음과 같은 과정을 통하여 계층 구조 트리를 업데이트하게 된다. 첫째, 웹 로봇이 수집한 특정 사이트를 처음 검색할 경우에 해당 웹사이트의 태그정보 표현방식을 판단하고 파싱에 의해 특정 정보가 있다고 판단되면 이를 데이터베이스에 등록하고 해당정보를 추출 한다. 둘째, 파악한 정보를 바탕으로 빠른 속도로 탐색, 변경된 정보를 갱신 추가, 삭제, 변경된 경우에 차이점을 비교 및 수집한다. 셋째, 변경된 태그정보가 없다고 판단된 페이지를 제외한 모든 페이지를 주기적으로 방문 수집한다.

4.3 웹 수집 범위

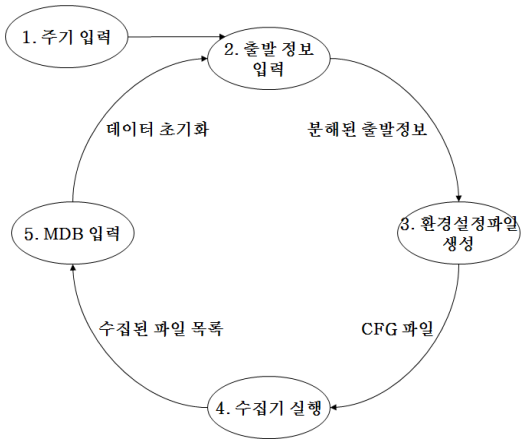
관리자가 웹 로봇을 실행 동작시켜 웹사이트 상에서 자신이 원하는 문서들을 수집하는 경우 무조건 수집기를 실행하여 웹사이트를 수집하는 것은 불필요한 문서들을 수집하게 되는 경우가 많다. 수집대상 주소들이 어떤 링크 구조를 가지고 있는지 파악을 해야 한다. 그리고 수집기를 구동시킬 첫 URL을 제대로 지정해야 사용자가 원하는 웹사이트 문서를 정상적으로 수집 할 수 있다. 다음 의 웹사이트 구조를 살펴보면 다음 [그림 6]과 같이 관리자가 수집기를 구동시킬 첫 주소로 'http://www.sever.com/'을 지정하면 위의 링크 구조상 모든 문서를 수집하게 된다. 이때 수집기를 구동시킬 첫 주소로 'http://www.sever.com/b/'을 지정하게 되면 수집기는 'http://www.sever.com/'과 'http://www.sever.com/a/' 링크를 수집하지 못하게 된다. 이처럼 수집기를 구동시킬 첫 주소를 어떻게 지정하느냐에 따라서 수집되는 문서와 수집되지 않는 문서가 달라진다.



[그림 6] 웹사이트 수집 범위

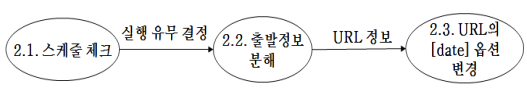
4.4 웹 수집 순서

관리자가 웹 로봇을 실행 동작시키기 위해서는 우선 관심 사이트의 갱신주기가 일일, 주간, 월간인지 기타 어떤 주기로 갱신 및 수정되는지를 판단하여야 한다. 그리고 사이트의 주소단위로 분해하여 단지 그 주소의 첫 페이지만 모니터링 할 것인지 아니면 한 단계 깊은 페이지가지 모니터링 할 것인지를 결정하여 수집기를 실행 하여야 한다. 수집된 주소의 웹 페이지는 데이터베이스에 입력되고 다음 수집을 수행하면서 그 주소의 웹 페이지가 갱신되었는지 비교 수집하게 된다. 다음 [그림 7] 은 웹 수집의 순서를 표현한 그림이다.



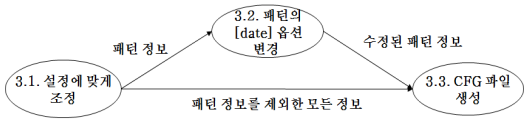
[그림 7] 웹 수집 순서

관심 주소의 출발정보를 입력하는 경우 [그림 8]과 같이 세부적으로 이 사이트의 해당 웹 페이지가 주로 어느 시간대에 갱신된다는 특징이 있다면 특정 시간대의 스케줄을 설정하여 실행유무를 결정한다. 이 경우도 해당 주소의 출발정보를 분해하여 관심 주소의 해당 페이지만 모니터링 대상으로 설정하여야 한다.



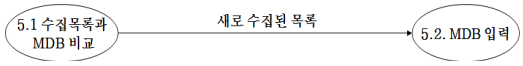
[그림 8] 출발정보 입력

최종 결정된 주소의 디렉터리 구조에 따라 규칙적인 패턴에 적합한 환경 설정을 통하여 [그림 9]와 같이 수집기를 실행한다.



[그림 9] 환경 설정 파일 생성

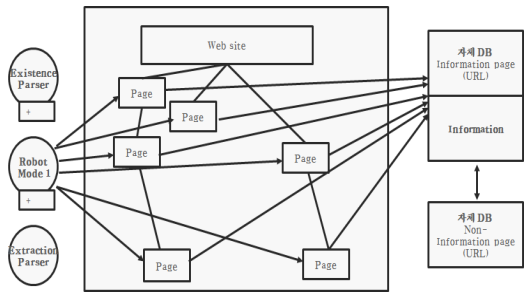
최근 수집된 웹 문서들은 기존에 수집된 문서와 비교되어 주소명이나 웹 문서 태그정보에 따라 갱신되었다고 판단되면 [그림 9]와 같이 데이터베이스에 입력 저장된다.



[그림 10] DB 입력

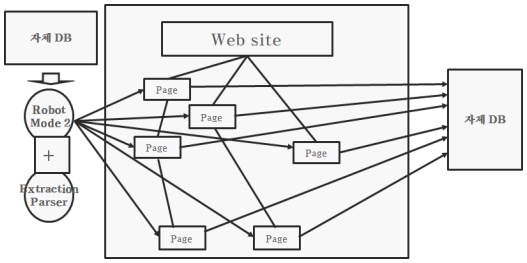
4.5 주기별 검색

웹 로봇에 관심 주소 리스트가 등록되면 'Mode 1'의 웹 로봇이 등록된 웹사이트의 모든 페이지를 방문하고, 'Existence Parser'가 정보의 존재여부를 판단하여 정보가 있는 주소가 없는 주소를 자체데이터베이스에 등록하면서 동시에 'Extraction Parser'가 해당정보를 추출하게 된다[10, 11].



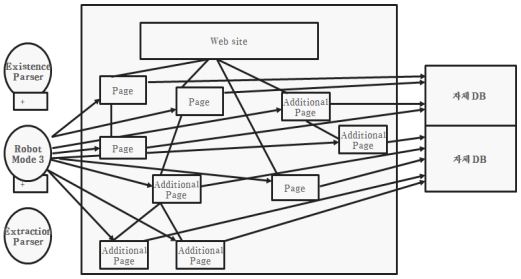
[그림 11] 주기별 검색 Mode 1

Mode 1을 통해 수집된 정보존재 페이지만을 방문하고 'Extraction Paser'에 의해 해당정보를 추출한다, 이 과정을 통하여 빠른 탐색과 정보갱신이 가능하게 된다.



[그림 12] 주기별 검색 Mode 2

Mode 1에서 정보가 없다고 판단된 페이지를 제외한 모든 페이지를 방문하여 추가된 페이지가 있는지 탐색하고 추가된 페이지에 추가된 정보가 있으면 'Extraction parser'에 의해 정보를 추출하게 된다.



[그림 13] 주기별 검색 Mode 3

4.6 웹 로봇 적용 시 특징

일반적으로 검색엔진에서의 소프트웨어 로봇은 통계적 분석, 유지관리, 미러링, 리소스탐사 등의 목적으로 이용되고 있다. 로봇은 네트워크와 서버에 과다한 부하를 줄 수 있기 때문에 로봇 배제 표준이 제안되고 있으며 신중한 설계가 요구되고 있다. 탐색 방법과 탐색주기에 따라 웹 문서수집기의 성능이 달라 질 수 있는데 주어진 주소 집합을 이용해서 너비 우선 탐색을 수행하거나 깊이 우선 탐색을 수행한다, 너비 우선 탐색의 경우 현재 페이지가 링크한 모든 페이지를 탐색하는 방식이며 깊이우선 탐색은 동일 웹 페이지에서 재귀적

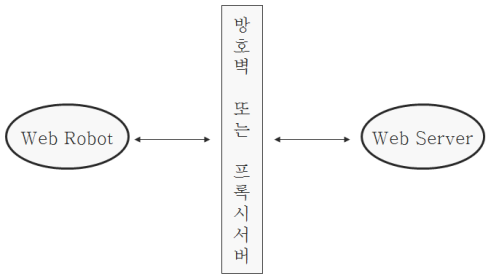
으로 수행하는 방식이다.

탐색 주기는 검색 시스템의 도메인에 따라 하루에 한번 또는 2~3일에 한 번씩 탐색하도록 한다. 주소수집 리스트 관리 시 탐색여부를 확인하는 필드를 두어 주기적으로 자료를 갱신할 수 있도록 하고 웹 로봇은 네트워크 부하를 줄이기 위하여 사용자 트래픽을 고려하여 운영한다[16].

4.7 수집 속도

웹 로봇는 사용자가 웹브라우저로 일일이 웹사이트를 방문하여 해당 문서를 사용자의 하드디스크에 저장하는 것을 자동으로 수행하는 것이다. 따라서 웹 로봇를 통하여 웹사이트의 문서를 수집하는 데 걸리는 시간은 웹 로봇가 동작하는 서버의 트래픽에 따라 달라질 수 있다. 또한 웹 로봇이 동작하는 서버의 네트워크 대역폭이 높다고 하여도 프록시나 방화벽을 거치는 경우에 웹 로봇의 수집 속도가 느려질 수 있다. 또한 프록시가 캐시역할을 담당하는 경우 다음과 같은 문제가 발생 할 수 있다.

즉, 웹 로봇은 실제 웹 서버에 연결해서 문서를 가지고 오는 것이 아니라 프록시에서 페이지를 가져오는 작업이 처리된다. 이러한 경우 캐시에 저장된 문서를 제거하거나 웹 로봇을 캐시서버와 연결되어 있지 않은 머신에서 동작하도록 하여야 한다.

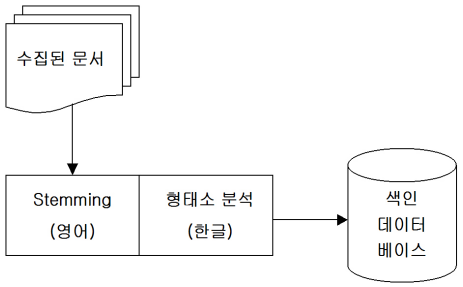


[그림 14] 웹 수집

4.8 색인 생성

색인 생성 에이전트는 웹 로봇에 의해서 수집된 문서들에 대하여 문서에 포함된 단어들을 처리하고 검색이 용이하도록 문서의 색인정보를 생성하는 작업을 수행한다. 색인 생성 에이전트는 문서요약작업과 동시에 이루어지며 문서의 키워드 추출을 기반으로 하여 이루어진다. 또한 색인 데이터베이스는 문서의 주소 및 요약된 문서를 같이 연결시켜준다. 다음 [그림 15]는 색인 생성 에이전트의 구성을 보여준다.

색인 생성 에이전트는 수집된 문서에 대한 요약작업에서 생성되는 키워드들에 대하여 영문의 경우 스테밍작업[3]을 하며 한글의 경우 형태소 분석을 하여 사용자의 질의에 대한 응답으로서 색인검색을 보다 유연하게 해줄 수 있다.



[그림 15] 색인 생성

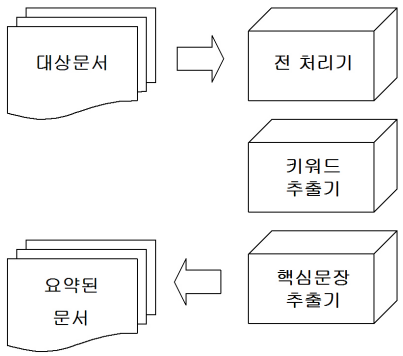
4.9 문서요약

문서요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도 즉, 문서의 길이를 줄이는 작업이다. 문서요약에서는 요약문을 생성하는 방식에 따라 추출과 요약으로 나눌 수 있다. 본 논문에서는 문서의 내용을 요약하기 위한 방식으로 핵심문장 추출방식을 사용하고 있으며 문서에서 핵심내용을 추출하기 위해서 우선 문서의 내용 중에서 키워드를 추출하는 방식을 사용한다. 본 논문에서는 키워드를 좀 더 효과적으로 추출하기 위하여 단어 간의 공백관계를 이용한 방식을 키워드 추출

3) 스테밍(Stemming) : 정보검색에 필요한 요소들인 조사, 특수기호, 존칭, stop word 등 을 제거하는 작업.

방법에 적용한다. 대상문서를 기반으로 전처리에서는 단어 간의 공백관계를 분석하고 키워드와 핵심 문장을 추출한다.

다음 [그림 16]의 전 처리기는 문서를 요약하기에 앞서 문서에 포함된 단어들에 대한 형태소 분석을 하여 문서를 요약하는 기초 작업을 수행하며, 키워드 추출기에서 추출된 키워드들은 색인 생성 에이전트에 의해서 색인 데이터를 생성하는 기준이 된다. 핵심 문장 추출기는 추출된 키워드를 기반으로 하여 문서상에서 가장 중요도가 높은 핵심 문장들을 추출하게 되며 사용자의 검색에 대한 결과로서 제공되어 사용자가 쉽게 문서의 내용을 파악 할 수 있도록 한다.



[그림 16] 문서 요약

5. 구현 및 평가

본 논문에서 구현한 정보수집시스템은 궁극적으로 인터넷상의 게제 기사자료를 수집 하여 사용자에게 제공하는 형태를 취하고 있다, 그러기 위하여 다음과 같은 성능평가 기준에 의해 시스템의 평가를 하려고 한다. 주요개선 요소로는 관심 주소의 수집여부, 동일 게시물의 중복수집 여부, 수집시간으로 구분할 수 있다.

시스템 구현은 Windows 서버환경에서 C 언어를 이용하여 웹 로봇을 개발하고 MS-Access 데이터베이스를 이용하였다.

CNN, Yahoo 뉴스, 조선신보를 대상으로 수집

〈표 1〉 수집 성능 평가

대상	수집여부		중복여부		수집시간(분)	
	전	후	전	후	전	후
CNN	△	○	△	×	8	4
야후 뉴스	△	○	○	×	7	3
조선신보	△	○	○	×	6	2.5

※ △ : 부분수집 및 일부중복, × : 중복 없음,
○ : 정상수집 및 중복.

한 결과 수집여부, 중복여부가 최소화 되었으며 며 수집시간의 경우 평균 50% 이상 향상되었다. 그리고 수집대상 페이지의 이미지나 링크정보에 따라 수집시간의 차이를 보이고 있다. 추가로 주소에 대한 반복적인 모니터링으로 인한 시스템 자원낭비가 일어났고 이미지나 텍스트가 깨어지는 현상이 발생 되었다.

6. 결 론

웹 로봇은 그 성격상 정보검색이나 정보여과 등의 인터넷 정보가공을 위해 적용 될 수 있는 기술이다. 특히, 인터넷의 정보홍수 속에서 원하는 정보를 정확하게 제시간에 획득하기란 쉬운 일이 아니며 따라서 이러한 관점에서 본 논문은 시사, 뉴스 정보 확인에 대하여 기존의 수작업 혹은 범용 검색엔진에 의한 반복적 작업에 효율적으로 활용할 수 있는 시스템이다.

본 논문은 비용 및 시간을 절약 할 수 있었으며 빠른 속도로 증가하는 정보에 대한 대처를 보다 효과적으로 할 수 있게 하였다. 그리고 실시간 갱신 사이트를 대상으로 한 정치, 시사 뉴스 수집에는 수집대상 웹 페이지의 불규칙적인 주소체계 특징으로 인하여 적용의 한계가 있으나 연구소나 국가기관 및 다른 특화된 사이트를 대상으로 적용될 수 있다. 기존의 범용 검색엔진이 할 수 없는 제한된 웹 페이지의 갱신정도를 갱진 전 데이터베이스와 비교 후 통합함으로서 집중적으로 모니터링 할 수 있다는 장점을 가지고 있다.

그리고 본 시스템을 실무적인 관점에서 서버 및

클라이언트 모바일 환경을 접목 한다면 실시간 사용자 맞춤 정보를 공유하고 전파하는데 활용할 수 있다.

참 고 문 헌

[1] 이상렬, “최신 정보검색론”, 2013.

[2] 임해창, 임희석 외 1명 역, “검색엔진 최신정보검색”, 휴먼싸이언스, 2012.

[3] Christopher D. Manning, Prabhakar Raghavan 저 안동언, 김재훈 외 1명 역, “최신 정보검색론”, 2010.

[4] 도용태, “인공지능 개념 및 응용”, 사이텍미디어, 2003.

[5] 김광영, 이원구, 이민호, 윤화묵, 신성호, “웹 자원 아카이빙을 위한 웹 크롤러 연구개발”, 「한국콘텐츠학회논문지」, 제11권, 제9호(2013).

[6] 김광영, 이원구, 이민호, 윤화묵, 신성호, “웹 자원 아카이빙을 위한 웹 크롤러 연구 개발”, 「한국콘텐츠학회논문지」, 제11권, 제9호(2011).

[7] 강한훈, 유성준, 한동일, “전문 분야 정보검색 시스템을 위한 웹 크롤러 래퍼의 설계 및 구현”, Proceedings of KIIS Fall Conference 2010.

[8] 이홍주, 양근우, 김규중, 백승기, 김종우, 허순영, 박성주, “과학기술 연구팀을 위한 지식포탈 아키텍쳐”, 대한산업공학회/한국경영과학회 2002 춘계공동학술대회 한국과학기술원(KAIST) 2002.

[9] Lee, S.-M. and Kim, T.-Y., “A News on Demand Service System based on Robot Agent”, 1998.

[10] Coffman, E. G., Z. Lin, and R. R. Weber, “Optimal robot scheduling for Web search engines”, France, December, 1997.

[11] Spiders, Wanderers, Broker, and Bots “Fah-Chun Cheong, Internet Agent”, New Rider Publishing, 1996.

[12] Baeza-Yates, “Ribeiro-Neto Modern Information Retrieval”, 1995.

[13] Wei Tang, Ling Liu, and Calton Pu, “Web-CQ : Detecting and delivering Information Changes on the Web”, Georgia Institute of Technolgy College of computing, http://www.cc.gatech.edu/projects/disl/WebCQ/, 1995.

[14] Junghoo Cho, and Hector Garcia-Molina, “Efficient Crawing Through URL Ordering”, Lawrence Pages Department of Computer science stanford University, 1995.

[15] Selberg, E. and O. Etzioni, “Multi-services search and comparison using the Meta-Crawler”, 4th Int WWW Conference, December, 1995.

[16] David Butter, Ling Liu, and Calton Pu, “A Fully Automated Object Extraction System for the World Wide Web”, Georgia Institute of Technology College of conputing Atlanta, GA 30332, U.S.A., 1995.

[17] Martijn Korster, “Guidelines for Robot writers”, 1993.

◈ 저 자 소 개 ◈

장 진 욱 (jwjang@konkuk.ac.kr)

건국대학교 신산업융합학과 경영공학박사를 취득하였으며, 국방부 정보사령부 전산장교와 SK 커뮤니케이션즈 PMO 매니저를 거쳐 현재 건국대학교 정보통신대학교 인터넷미디어공학부 산학교수로 재직하고 있다. 주요 관심분야로는 정보검색, 프로젝트 매니지먼트, 소프트웨어 품질, 테스트 프로세스 등이다.