

대용량 콘텐츠를 위한 역 RSS 웹 크롤러 설계

정회원 홍 석 주*, 박 영 배**

Design for WEB Crawler of Reverse RSS for a Large Quantity Contents Search

Seok-Joo Hong*, Young-Bae Park** *Regular Members*

요 약

본 논문은 역(Reverse) RSS(Really Simple Syndication) 기반의 지능형 정보검색시스템 설계에 관한 것으로, 기존의 방식과 같이 사용자가 RSS 주소를 입력하여 제한된 RSS 정보를 받아보는 방식이 아니라, 사용자는 단순히 자신이 원하는 정보를 입력만 하면, 자동화된 RSS 주소수집서버가 수집한 수많은 RSS 주소들로부터 실시간으로 수집하는 RSS 규격 문서들 중 사용자가 원하는 규격 문서에 대한 RSS 정보만을 제공해줌으로써, 수많은 정보를 찾아 그 중 원하는 정보만 추려서 제공해주는 역 RSS 구독(Reverse RSS Subscribe) 방식을 설계하는데 있다. 제안된 역 RSS 기반 지능형 정보검색 시스템을 통하여 양질의 정보를 찾아서 헤매는 시간을 획기적으로 줄일 수 있고 개인 비서를 두게 되는 효과를 얻을 수 있다.

Key Word : RSS(Really Simple Syndication), Reverse RSS, RSS Reader, 크롤링 (Crawling)

ABSTRACT

This study matters the design of an intelligent information search system that is based on the Reverse RSS(Really Simple Syndication). Apart from to the previous method, where the user inputs the RSS address that one intends and obtains limited RSS information, the user just types in the information that one appoints to acquire the RSS information of standard documents that the user is interested among several RSS addresses by a Reverse RSS(Really Simple Syndication) method, which is drawn by the automated RSS address collection server in realtime. Through the proposed Reverse RSS(Really Simple Syndication) based intelligent information search system, time can be significantly saved along with obtaining information with good quality, furthermore, it has the effects of having a personal secretary.

I. 서 론

최근 초고속 통신망의 발달과 함께 인터넷 이용의 급속한 확산으로 빠르게 정보를 얻고 제공할 수 있게 되었다. 이전처럼 정보를 찾거나 얻기 위해 많은 시간을 투자하여 도서관이나 관련 서적 등을 찾아 헤매지 않아도 되고, 언제 어디서든지 인터넷과 연결된 컴퓨터 앞에서 빠르고 쉽게 정보를 찾을 수 있게 되었다. 이전의 노하우(Know-How)가 중요한 시대

보다ノウ웨어(Know-Where)를 더 필요로 하는 시대가 되었다.

예전에는 일방적인 정보를 찾아 받아들이기만 했던 사용자들이 최근 들어 개인 정보와 개인 내부에 대해 관심을 가지며 표현하며 외부에 표출하기를 원하게 되었다. 이러한 시점에 가장 알맞은 도구인 블로그(blog)가 등장하고 사용이 늘어나게 되었다. 자신만의 블로그 뿐만 아니라 관심있는 주제의 다른 사람의 블로그를 방문하면서 유용한 블로그나 관심 있

* 경희사이버대학교 정보통신학과 겸임교수, ** 명지대학교 컴퓨터공학과 교수
논문번호 : 09007-0130, 투고일자 : 2009년 1월 30일

는 블로그는 계속 구독하려는 필요성을 느끼게 되었다. 이러한 이유로 인하여 블로그는 좀 더 쉽게 콘텐츠를 배급하고 흩어져 있는 많은 자료들을 수집하기 위해 XML기반의 RSS 서비스를 사용하고 있다^{11,12}.

기존의 RSS 리더기(Reader)는 사용자가 RSS 주소를 직접 입력하면 해당 RSS 주소를 주기적으로 방문하여 새로운 정보가 올라온 경우 사용자에게 알려주었다. 이러한 RSS 리더기를 사용한 방식은 사용자가 해당 RSS 주소를 직접 알아야 정보를 받아 볼 수 있다는 사용 편리성에 있어서 치명적인 단점과 각 사용자들이 개별적으로 알고 있는 RSS 주소의 수가 많지 않다는 단점이 존재하고, 이러한 방식으로 인해서 커다란 활용 여지가 있음에도 불구하고 RSS 리더기는 사용자층을 많이 확보하지 못하였다¹².

또한, 기존의 RSS 리더기가 가지는 두 번째 문제점을 해결하기 위한 것 중 하나가 메타 블로그(Meta Blog)로서 이는 여러 사람이 수동으로 입력한 RSS 주소를 공유해서 다양한 RSS 주소로부터 콘텐츠를 가져와서 사용자에게 보여 주거나 검색할 수 있도록 한다. 이렇게 RSS 주소가 좀더 많아지기는 하였지만 여전히 사용자가 수동으로 입력한 극소수의 RSS 주소에 의존하고 있다. 또한, 사용자는 다양한 사람들이 입력한 RSS 주소로부터 정보를 받아 보면서 그 중에 자신이 원하는 정보를 선택해야 하는 수고를 해야 한다¹².

본 논문은 이러한 문제점을 해결하기 위하여 제한한 것으로서, 본 논문의 목적은 RSS 규격 문서를 기존에 사용자가 직접 RSS 주소를 미리 알고 있으면서 입력도 해야 하는 불편함을 역 RSS 구독 방식을 사용하여 편리성과 유용성을 증가시킨 역 RSS 기반 지능형 정보 검색 및 모니터링 시스템을 제공하는데 있다.

본 논문의 또 다른 목적은 RSS 리더기 부분에서는 기존의 방식과 같이 사용자가 RSS 주소를 입력하여 제한된 정보를 받아 보는 방식이 아니라 사용자는 단순히 자신이 원하는 정보를 입력만 하면, 자동화된 RSS 주소수집서버가 수집한 수많은 RSS 주소들로부터 실시간으로 수집하는 RSS 정보들 중에서 사용자가 원하는 정보에 대한 역 RSS 문서 정보를 제공하여 RSS의 사용 용이성 한계와 제공되는 정보 범위의 한계를 극복할 수 있도록 한 역 RSS 기반 지능형 정보 검색 시스템을 제공하는데 있다.

이 논문의 구성은 다음과 같다. 2장에서는 본 논문의 이론적 고찰을 위하여 기존 RSS 특징과 문제점을 살펴본다. 3장에서는 기존 RSS를 이용한 정보

검색시스템의 문제점을 해결하기 위한 역 RSS 기반의 지능형 정보검색시스템의 설계와 기술요소를 설명한다. 4장에서는 결론을 맺는다.

II. 관련 연구

2.1 RSS 2.0 소개

본 논문에서는 역 RSS 지능형 정보검색시스템의 설계를 위해 RSS 2.0에 대해 알아보기로 한다. RSS는 일종의 XML 언어이고, 웹 콘텐츠와 메타데이터를 신디케이트 하는데 사용된다. RSS 0.91은 여러 버전들 중 가장 일반적으로 사용되고 있다. 새로운 RSS 피드의 경우, 2.0 버전을 사용하는 것이 더 좋은데, 그 이유는 현재 스펙은 0.91과 백워드 호환이 되기 때문이다.

RSS 파일은 <channel> 엘리먼트와 이것의 하위 엘리먼트로 구성된다. <channel>에는 <title>, <link>, <description>같은, <channel>에 대한 메타데이터를 나타내는 엘리먼트들이 포함된다. 또한 아이템의 형태로 채널 콘텐츠 자체도 추가된다. 아이템들은 채널을 구성하고, 자주 변하는 콘텐츠를 포함하고 있다^{12,16}.

2.2 RSS 활용 분야

RSS는 사이트에 있는 콘텐츠의 배급(syndication)과 수집(aggregation), 그리고 개인 사용자의 콘텐츠 사용 편리성으로 볼 수 있다. 먼저 사이트에서 RSS 활용은 A사이트가 B사이트의 콘텐츠를 배급하기로 했다면 A사이트는 B사이트의 RSS를 파악하고 읽어와서 갱신 주기 등을 결정하여 콘텐츠가 표시될 A의 웹 페이지를 구성한다. 그런 다음, A사이트의 RSS 수집/발행 시스템이 작동하면 해당 시스템은 자동적으로 B사이트의 콘텐츠의 제목, 링크, 내용 등

표 1. RSS 버전

버전	명칭	책임자	특징
RSS 0.9	RDF Site Summary	Netscape	RDF 기준
RSS 0.91	RSS	UserLand S/W	웹 브라우저 제조사에 의한 독자적 확장
RSS 1.0	RDF Site Summary	RSS-DEV 그룹	표준화된 모듈에 의한 확장성
RSS 2.0	Really Simple Summary	Dave Winer	0.9X와의 호환성. 팟캐스팅
Atom	-	IETF	새롭게 만들어진 표준화 사양

을 결정된 주기에 따라 갱신하여 A사이트에서 보여 주게 된다.

또한, 개인의 RSS 활용으로는 이전에 관심 있는 특정 정보를 가진 사이트들을 일일이 방문하여 최신 콘텐츠를 확인하고 언제 업데이트가 되는지 알지 못해 수시로 방문해야 하는 번거로움을 대신하여 RSS 리더기 프로그램을 통해 관심 있는 사이트의 RSS를 등록하고, 갱신 주기 등을 저장해 놓으면 RSS리더 프로그램을 통해 업데이트된 최신의 콘텐츠를 편리하고 빠르게 파악하고 얻을 수 있다.

2.3 RSS 리더기

RSS 리더기(reader)는 등록된 RSS 포맷으로 만들어진 문서를 읽어서 콘텐츠를 쉽게 구독(aggregation)할 수 있게 해주는 프로그램이나 서비스를 말한다. 이러한 RSS 리더기는 RSS를 이용하기 위해서는 필수적인 프로그램이라고 볼 수 있다. 리더기는 등록된 채널(channel)을 자동으로 업데이트해서 보여주며, 등록된 시간에 자동으로 업데이트하여 주기적으로 새로운 콘텐츠를 얻을 수 있다.^[11]

RSS 리더기는 운영 및 활용 형태에 따라 크게 세 가지 종류가 있다. 먼저 가장 많이 활용하고 있는 프로그램을 설치하여 일반 애플리케이션으로 사용하는 방법이 있다. 둘째로는 웹 사이트에서 제공되는 리더 서비스를 사용하는 방법이 있으며, 셋째로 기존에 많이 활용되고 있는 독립 애플리케이션인 아웃룩(outlook)이나 인터넷 브라우저에 add-in이나 plug-in 방식이 포함되어 있는 복합형이 있다. 그리고 최근에는 구글 개인화 페이지, 마이크로소프트의 라이브 닷컴, Netvibes, 위자드닷컴, Pageflakes, Protopage 같은 개인화 포털도 RSS 리더의 역할을 하고 있다. 표 2는 RSS 리더기의 종류에 따라 제공되는 RSS 리더기 들을 보여준다.

이와 같이 RSS 역할이 점점 커지고 있지만, 여전히 방대한 정보의 제공 및 정보의 중복으로 인해 RSS의 필요성을 느끼지 못하고 있으며, 각 회사의 서비스 개발비용에 비해 소수의 유저만이 사용하고, 대다수의 사람이 RSS에 대해 인지하지 못하고 있는

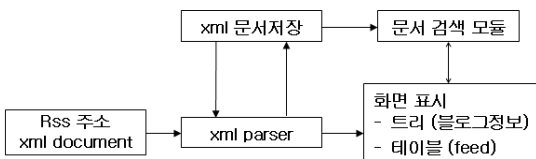


그림 1. RSS 리더기의 전체 구조

표 2. RSS 리더기의 종류

종 류	서비스(리더)이름
독립 애플리케이션	연모, Sharp reader
웹기반 리더기	한rss, Rojo, NewsGator, 구글 RSS 리더
복합형 리더기	Onfolio, Attensa

실정이다. 또한, 대다수의 RSS 유저는 블로그 서비스만을 주로 이용하고 있으며, 각기 다른 RSS 프로그램 사용하고 리더기에 쌓아만 놓는 구독자들 때문에 낭비되는 트래픽이 많이 일어나는 실정이다.

Ⅲ. 역 RSS 기반의 지능형 정보검색 시스템

이 논문에서 제시하는 대용량 콘텐츠 검색을 위한 역 RSS 기반의 정보검색시스템 플랫폼은 그림 2와 같으며, 인터넷 상의 방대한 웹 콘텐츠를 기존의 방식과 같이 사용자가 RSS 주소를 입력하여 제한된 정보를 받아 보는 방식이 아니라 사용자는 단순히 자신이 원하는 정보를 입력만 하면, 자동화된 RSS 주소수집서버가 수집한 수많은 RSS 주소들로부터 실시간으로 수집하는 RSS 정보들 중에서 사용자가 원하는 정보에 대한 역 RSS 문서 정보를 제공하여 RSS의 사용 용이성 한계와 제공되는 정보 범위의 한계를 극복할 수 있도록 한 역 RSS 기반 지능형 정보 검색 시스템의 구조이다.

3.1 시스템 구성 요소

그림 2의 제안시스템은 크게 RSS 주소수집 서버, 데이터베이스 서버, RSS 데이터수집서버, RSS 데이터저장소, RSS 데이터분석 서버, 사용자 RSS 문서 제공 서버, 웹 서버 및 사용자 단말(웹 브라우저) 등을 포함하여 이루어진다.

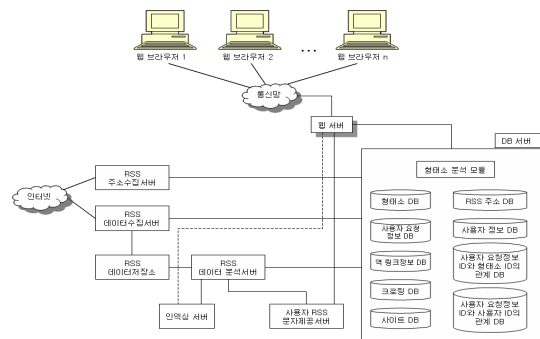


그림 2. 대용량 콘텐츠 검색을 위한 역 RSS 기반의 정보검색시스템 플랫폼

3.1.1 RSS 주소수집 서버

RSS 주소수집서버는 인터넷에 연결되어 통상의 자동적인 확장방식으로 수많은 RSS 주소들을 수집하여 DB 서버에 저장될 수 있도록 전송하는 기능을 수행한다. 이때, 상기 자동적인 확장방식은 대표적인 인터넷 자원(예컨대, RSS 또는 ATOM 등) 주소 표현 형태를 이용하여 인터넷 상에 있는 웹(예컨대, IPv4에서는 일반 웹, IPv6에서는 전자제품을 포함한 웹 등) 문서(HTML 파일)에서 RSS 주소를 자동적으로 추출하고, 해당 웹 문서에 있는 링크(link)에서도 같은 방식으로 RSS주소를 추출하는 방식이다.

즉, 미리 설정된 주요 포털이나 블로그 웹 문서를 시작으로 해서 점차적으로 해당 웹 문서들의 외부로 향하는 링크를 따라 방문하면서 RSS 주소를 자동 추출하거나, RSS 주소를 추출할 웹 문서를 주요 메타 사이트들이 제공해주는 최신 RSS 파일을 주기적으로 방문하면서 이에 들어 있는 링크 주소를 방문하여 RSS 주소를 추출하는 방식이다.

3.1.2 DB 서버

DB 서버는 RSS 주소수집 서버로부터 수집된 수많은 RSS 주소들, 각 주소간의 링크 관계, 각 사용자 정보들 및 각 사용자가 요청한 정보별로 형태소들을 데이터베이스화하여 저장 및 관리하는 기능을 수행한다.

이러한 DB 서버는 형태소 분석모듈, 형태소 DB, RSS 주소 DB, 사용자 정보 DB, 사용자 요청정보 DB, 사용자 요청정보 ID와 형태소 ID의 관계 DB, 및 사용자 요청정보 ID와 사용자 ID의 관계 DB 등을 포함하여 구성되어 있다.

(1) 형태소 분석 모듈

형태소 분석 모듈은 웹 서버를 통해 후술하는 적어도 하나의 사용자 단말로부터 전송된 사용자 요청 정보들을 제공받아 각 사용자 요청정보별로 형태소를 분석하는 기능을 수행한다.

(2) 형태소 DB

형태소 DB는 형태소 분석 모듈로부터 형태소 분석된 각 사용자 요청정보(또는 역 RSS 키워드)에 대한 형태소들을 고유한 형태소 ID와 실제 형태소명으로 테이블화하여 저장 및 관리한다.

(3) RSS 주소 DB

RSS 주소 DB는 RSS 주소수집 서버로부터 수집된 RSS 주소들을 제공받아 국가별(예컨대, 영어권

등)로 테이블화하여 저장 및 관리한다.

(4) 사용자 정보 DB

사용자 정보 DB는 웹 서버를 통해 적어도 하나의 사용자 단말로부터 전송된 사용자 정보들(예컨대, 사용자 ID, 사용자명(User Name) 및 패스워드(Password) 정보 등)을 테이블화하여 저장 및 관리한다. 이때, 상기 사용자 정보들은 해당 서비스의 회원 가입 시 웹 서버에 접속된 각 사용자 단말로부터 제공되며, 사용자 ID와 패스워드는 사용자 로그인 시에 사용된다.

(5) 사용자 요청정보 DB

사용자 요청정보 DB는 웹 서버를 통해 각 사용자 단말로부터 전송된 사용자 요청정보들을 고유한 사용자 요청정보 ID와 실제 사용자 요청정보명(형태소 분석되기 전 상태임)으로 테이블화하여 저장 및 관리한다.

(6) 사용자 요청정보 ID와 형태소 ID의 관계 DB

사용자 요청정보 ID와 형태소 ID의 관계 DB는 사용자 요청정보 DB에 저장된 각 사용자 요청정보 ID와 형태소 DB에 저장된 형태소 ID 간의 연결 관계를 테이블화하여 저장 및 관리한다. 표 6에서 예

표 3. 형태소 DB의 예

형태소 ID	형태소명
82	스크랩
83	로그
84	포털

표 4. RSS 주소 DB의 예

RSS 주소(영어권)	최종 방문날짜	갱신주기
http://stevewynn.net/rss/wynnweb_our_review.rs	2009-1-2	0
http://feeds.feedburner.com/TheCavamansWine	2008-1-2	0
http://mensa-barbie.blogspot.com/rss.xml	2008-1-2	0

표 5. 사용자 요청정보의 예

사용자 요청정보 ID	사용자 요청 정보 명
71	오픈 마루
72	인턴
106	벤처 투자

컨대, 1번 사용자 요청정보 ID는 41번 및 42번 형태소를 포함하고 있다는 의미이다.

(7) 사용자 요청정보 ID와 사용자 ID의 관계 DB
 사용자 요청정보 ID와 사용자 ID의 관계 DB는 사용자 요청정보 DB에 저장된 각 사용자 요청정보 ID와 사용자 정보 DB에 저장된 사용자 ID 간의 연결 관계를 테이블화하여 저장 및 관리한다. 이와 같은 사용자 요청정보 ID와 사용자 ID의 관계 DB를 통해 어떤 사용자가 어떤 정보를 요청하였는지를 용이하게 알 수 있다.

(8) 역 링크정보 DB
 역 링크정보 DB는 웹 문서 랭킹(Ranking)을 하기 위하여 특정한 사이트에 대한 역 링크들을 테이블화하여 저장 및 관리한다. 표 8에서 예컨대, “http://apache.org”에서 “http://w3.org”에 링크를 가지고 있다는 표시이다. 이러한 역 링크정보 DB는 RSS 주소수집서버와 연동되어 RSS 주소들의 수집 시 용이하게 활용될 수 있다.

(9) 크롤링 DB
 크롤링 DB는 RSS 주소와 사이트(SITE)의 관계를 테이블화하여 저장 및 관리한다. 이러한 크롤링 DB는 RSS 주소수집 서버와 연동되어 RSS 주소들을 수집하기 위한 메타 정보 및 랭킹 정보에 활용될 수 있다.

(10) 사이트 DB
 사이트 DB는 각 사이트(SITE)에 대한 각종 정보들(예컨대, 마지막 방문날짜, 랭킹 점수 또는 외부 링크의 수 정보 등)을 테이블화하여 저장 및 관리한다. 여기서, 상기 외부 링크의 수는 추후 웹 문서 랭킹을 할 때에 사용되며, 상기 랭킹 정보는 주기적으로 DB 서버가 상기 역 링크정보 DB에 저장된 정보와 외부 링크의 수 등의 정보를 사용하여 계산되어지는 결과이다.

3.2 역 RSS기반의 지능형 정보검색 방법

그림 3은 역 RSS기반의 지능형 정보검색 방법을 설명하기 위한 “<item>”의 소스를 구체적으로 설명하기 위한 도면으로서, 상기 “<item>”내에는 예컨대, 제목(title), 링크(link), 요약설명(description), 카테고리(category) 및 등록날짜(pubdate) 정보들을 포함하고 있으며, RSS 데이터수집 서버는 상기 링크(link) 정보에

표 6. 사용자 요청정보 ID와 형태소 ID의 관계 DB의 예

사용자 요청정보 ID	형태소 ID
1	41
1	42
2	43
3	44
4	45

표 7. 사용자 요청정보 ID와 사용자 ID의 관계 DB의 예

사용자 ID	사용자 요청정보 ID
sjhong	6
sjhong	7
sjhong	8
sjhong	9
sjhong	15

표 8. 역 링크정보 DB의 예

사이트(Site)	역 링크(Back Link)
http://w3.org	http://-project.org
http://w3.org	http://alphaquam.com
http://w3.org	http://apache.org

표 9. 크롤링 DB의 예

URL	RSS 주소
http://www.allblog.net	http://www.allblog.netRSS/AllPosts.xml
http://kkninja.egloos.com	http://kkninja.egloos.com/index.xml
http://ageagain.egloos.com	http://ageagain.egloos.com/index.xml

표 10. 사이트 DB의 예

URL	최종 방문날짜	랭크(RANK)	외부 링크수
http://nineteen.egloos.com	2009-1-2	2.01627457857	119
http://plluto.egloos.com	2009-1-2	1.46081874982	71
http://psyke.egloos.com	2009-1-2	0.265903513474	9
http://romuska.egloos.com	2009-1-2	0.912660412025	51

```

<item>
  <title><![CDATA[개발자들의 아카데미 상]]>
  </title>
  <link>http://agile.egloos.com/3238987</link>
  <guid>http://agile.egloos.com/3238987</guid>
  <description>
    <![CDATA[
IT업계에도 노벨상과 아카데미상이 있습니다.
튜링상을 노벨상에 비유할 수 있고, 톨트상을
아카데미상에 &nbsp;비유할 수 있습니다.
오늘은 개발자들에게 아카데미상을 보는 기대감과
즐거움을 주는 톨트상에 대해 이야기를 해보겠습니다.
톨트상은 1990년부터 .....
<a href="http://agile.egloos.com/3238987">
<font style="font-size:11px;">글 전체보기
</font></a>]]>
  </description>
  <category>미분류</category>
  <pubDate>Mon, 23 Nov 2008 10:34:50
GMT</pubDate>
</item>
    
```

그림 3. RSS 파일의 한 <Item>의 예

존재하는 “http://agile.egloos.com/3238987” URL 주소를 방문하여 해당 원본 웹 문서 데이터를 수집한다.

RSS 데이터수집 서버는 RSS 데이터저장소에 미리 저장된 RSS 파일 목록과 소스 링크 정보를 통해 새로 웹에서 받아온 RSS 파일을 비교하여 새로 업데이트(Update)된 “<item>~~~<item>”에 대해서만 해당 링크(link)를 방문하여 원본 웹 문서 데이터를 수집하게 된다. 그리고 RSS 데이터저장소는 RSS 데이터수집 서버로부터 수집된 각 RSS 파일에 대한 웹 문서(HTML 파일) 데이터와 함께 해당 RSS 정보들(예컨대, 제목(title), 링크(link), 요약설명(description), 카테고리(category) 및 등록날짜(publication date) 정보 등)를 제공받아 이를 저장 및 관리하는 기능을 수행한다.

또한, RSS 데이터저장소에 있는 최신 웹 문서 데이터들은 통상의 스케줄러(scheduler)에 의해서 RSS 데이터분석 서버로 할당됨이 바람직하다. 그리고 RSS 데이터분석 서버는 RSS 데이터저장소에 미리 저장된 각 RSS 파일에 대한 웹 문서 데이터를 제공받아 형태소 분석모듈을 통해 형태소 분석하고, 상기 형태소 분석된 데이터들과 DB 서버에 미리 저장된 각 사용자 요청정보별 형태소들을 비교 분석하여, DB 서버에 미리 저장된 각 사용자 요청정보와 동일한 범주에 포함되는지의 여부를 판단하는 기능을 수행한다.

이를 구체적으로 설명하면, RSS 데이터분석 서버는 DB 서버의 형태소 DB에 저장되어 있는 형태소명들과 현재 분석중인 웹 문서 데이터를 형태소 분석한 형태소명들을 비교하여, 현재 웹 문서 데이터가 포함하고 있는 형태소 ID를 추출한다. 즉, DB 서버의 사용자 요청정보 ID와 형태소 ID의 관계 DB를 이용하여 현재 웹 문서 데이터가 포함하고 있는 사용자 요청정보 ID를 획득 결국, 현재 웹 문서 데이터가 포함하고 있는 사용자 요청정보 ID들을 획득하게 되고, 이 사용자 요청정보 ID들은 각 서버에서 해당 사용자 요청정보에 대한 RSS 주소(예컨대, http://141.223.~~/rss/1/1.xml)를 가리키게 된다. 따라서, 해당 RSS 파일을 읽어 들인 후 현재 웹 문서 데이터의 RSS 정보들(예컨대, 제목(title), 링크(link), 요약설명(description), 카테고리(category) 및 등록날짜(publication date) 정보 등)를 추가한 후 저장한다. 이때, 추가되는 순서는 날짜 순서대로 저장됨이 바람직하다.

예를 들면, 수집된 웹 문서 데이터의 내용은 “나는 즐겁게 밥을 먹었다”이고, 형태소 분석된 내용은 “나”, “는”, “즐겁”, “게”, “밥”, “을”, “먹”, “었” 및 “다”이며, DB 서버의 형태소 DB 및 사용자 요청정보 ID와 형태소 ID의 관계 DB는 각각 표 11 및 표 12와 같다고 가정한다.

이때, 상기 형태소 분석된 내용은 형태소 DB 중에서 “나”, “즐겁” 및 “밥”이라는 3개의 형태소를

표 11. 형태소 DB

형태소 ID	형태소명
1	나
2	바나나
3	즐겁
4	논문
5	밥

표 12. 사용자 요청정보 ID와 형태소 ID의 관계 DB

사용자 요청정보 ID	형태소 ID
1	4
2	1
2	5
3	2
4	3
4	2

포함하고 있다는 것이 분석되어진다. 이들은 각각 1번, 3번 및 5번 형태소 ID를 가지고 있다.

상기 1번, 3번 및 5번은 결국 현재 웹 문서 데이터가 형태소 DB에 저장되어 있는 형태소 목록 중에서 포함하고 있는 형태소로서 이것을 이용해서 사용자 요청정보 ID와 형태소 ID의 관계 DB를 살펴보면, 2번의 사용자 요청정보 ID가 1번 및 5번의 형태소 ID로 이루어져 있기 때문에, 이 2번의 사용자 요청정보의 범주에 현재 웹 문서 데이터가 포함된다.

한편, 4번의 사용자 요청정보 ID는 3번 및 2번의 형태소 ID로 이루어져 있기 때문에, 단순히 3번의 형태소 ID만 가지고 있는 현재 웹 문서 데이터는 해당 사용자 요청정보의 범주에 속하지 않게 된다. 이렇게 추출된 사용자 요청정보 ID들을 이루는 형태소들이 현재 웹 문서 데이터 내에서 일정 기준 이상 분산되어 있는 경우에 해당 사용자 요청정보 ID는 제외시킨다.

이로써 현재 웹 문서 데이터의 RSS 정보들(예컨대, 제목(title), 링크(link), 요약설명 (description), 카테고리 (category) 및 등록날짜(publication date) 정보 등)은 사용자 RSS 문서제공서버(600)에 전달되어서 2번 사용자 요청정보 RSS에 추가되어진다.

한편, 상기 형태소 분석된 내용(“나”, “는”, “즐겁”, “게”, “밥”, “을”, “떡”, “었” 및 “다”)은 그대로 인덱싱 서버에 전송되어 추후 사용자가 웹 서버를 통해 용이하게 검색할 수 있도록 한다. 그리고 사용자 RSS 문서제공 서버는 RSS 데이터분석 서버에 의해 현재 형태소 분석된 웹 문서 데이터가 각 사용자 요청정보와 동일한 범주에 포함될 경우, 해당 웹 문서 데이터에 대한 RSS 파일의 RSS 정보들을 제공받아 해당 사용자 요청정보에 대한 맞춤형 역 RSS 문서 정보를 생성하여 저장 및 관리하는 기능을 수행한다. 여기서, 역 RSS 문서 정보에는 RSS를 통해서 제공되는 다양한 텍스트, 이미지, 사운드, 동영상 또는 지식 정보 등을 포괄하여 제공하는 정보이다.

웹 서버는 인터넷을 통해 각 사용자 단말로부터 전송된 사용자 요청정보들을 제공받아 DB 서버에 전달하고, 사용자 RSS 문서제공서버와 연동되어 해당 사용자 요청정보에 대한 맞춤형 역 RSS 문서 정보를 해당 사용자 단말의 화면에 디스플레이(Display)해주는 기능을 수행한다. 또한, 웹 서버는 각 사용자 단말로부터 전송된 사용자 ID와 패스워드 정보를 제공받아 DB 서버에 미리 저장된 사용자 정보들과 비교하여 회원 여부를 판별한다.

그리고 사용자 단말은 예컨대, 네트워크(Network) 또는 인터넷(Internet) 등과 같은 유선 또는 무선 통

신망을 통해 웹 서버에 접속되며, 통상적인 웹 브라우저(Web Browser)를 통해 웹 서버에서 제공하는 각종 서비스를 제공받을 수 있게 된다.

IV. 결 론

본 논문에서는 사용자가 RSS 주소를 입력하여 제한된 정보를 받아 보는 방식이 아니라 사용자는 단순히 자신이 원하는 정보를 입력만 하면, 자동화된 RSS 주소수집서버가 수집한 수많은 RSS 주소들로부터 실시간으로 수집하는 RSS 정보들 중에서 사용자가 원하는 정보에 대한 역 RSS 문서 정보를 제공하여 RSS의 사용 용이성 한계와 제공되는 정보 범위의 한계를 극복할 수 있도록 한 역 RSS 기반 지능형 정보 검색 시스템의 플랫폼을 제안하였다.

본 논문의 기대효과는 사용자가 단순히 자신이 원하는 정보를 입력만 하면, 자동 수집된 수많은 RSS 주소들로부터 실시간으로 수집하는 RSS 규격 문서들 중에서 사용자가 원하는 RSS 규격 문서에 대한 RSS 정보만을 제공해줌으로써, 사용자는 수많은 정보를 찾아서 그 중 원하는 정보만 추려서 제공해주는 개인 비서를 두게 되는 효과를 얻게 되어서 양질의 정보를 찾아 헤매는 시간을 획기적으로 줄일 수 있다.

향후의 과제로는 대용량 콘텐츠의 자동 수집을 위한 지능형 정보검색시스템을 역 RSS 기술과 웹 크롤러 기술을 적용하여 구현하는데 있다.

참 고 문 헌

- [1] 장은영, “개별화된 웹 미디어를 이용한 학습 커뮤니티 환경 설계 및 구현”, 한국교원대학교 석사학위논문, 2005.
- [2] 강성후, “XML을 활용한 rss 리더기의 설계 및 구현”. 부산외국어대학교 석사학위논문, 2005.
- [3] 석정화, “XML기반의 RSS를 이용한 협업을 위한 커뮤니케이션 시스템 구현”. 홍익대학교 석사학위 논문, 2004.
- [4] 김종태, “나는 블로그가 좋다”, 이비컴, pp 50-250, 2004.
- [5] 김법목, “동기적 상호작용 증진을 위한 블로그 기반 협동학습 시스템의 설계 및 구현”. 한국교원대학교 석사학위논문, 2005.
- [6] <http://web.resource.org/rss/1.0/spec>
- [7] 전중홍, “컨텐츠 신디케이션 표준화 동향” - RSS, OPML, ATOM/ <http://www.w3c.or.kr/~hollobit/>

data/paper/ TTA-RSS2.htm,

[8] 정희경, “알기쉽게 해설한 XML”, 이한출판 사, pp111-400, 2005.

[9] <http://www.ibm.com/developerworks/web/library/w-rss.html>

[10] 김윤수 <http://yesarang.tistory.com/8>

[11] developerWorks 제공 RSS 피드 <http://www-106.ibm.com/developerworks/rss/>

[12] 권이남, 김재수, 신동구, 전성진, 정택영, 박병희, “RSS기반 과학기술 정보 배급 표준시스템 설계에 관한 연구”, 한국정보처리학회 2005년 추계 학술대회에서 발표한 발표 논문지, pp545-548, 2005.

[13] 신나희, “한국형 블로그(Blog)에 관한 연구-포털 서비스를 중심으로”, 이화여자대학교 석사학위 논문, 2003.

[14] 구중억, “연구장비정보의 RSS 기반 SDI 시스템 설계 및 구현”, 충남대 대학원 석사학위 논문, 2006.

[15] xfiniti Korea, <http://www.xfiniti.com>

[16] <http://www.mnot.net/rss/tutorials> 10(2): 127~160, 2001

[17] Ben Hammersley, *Developing Feeds with RSS and ATOM*, O'Reilly Media, 2005

[18] The RSS 2.0 specification, <http://blogs.law.harvard.edu/tech/rss>

[19] David Chmielewski, Gongzhu Hu, “A Distributed

Platform for Archiving and Retrieving RSS Feeds”, ACIS-ICIS 2005, pp. 215-220

[20] Sean Lyndersay, “Windows and RSS: beyond blogging”, SIGMOD 2006.

홍 석 주 (Hong, Seok Joo)

정회원



1988년 2월 명지대학교 전자계산학과 졸업

1998년 8월 명지대학교 컴퓨터공학과 석사

2001년 8월 명지대학교 컴퓨터공학과 박사수료, 방송정보기술사, ISO 국제심사원, 경희

사이버대학교 정보통신학과 겸임교수, (주)유오시스템즈 대표이사

<관심분야> web computing, ITS, multimedia Database

박 영 배 (Park, Young Bae)

정회원



1993년 2월 서울대학교 대학원 컴퓨터공학과 (공학박사)

1990년~1992년 명지대학교 전산소장

1997년~2001년 명지대학교 산업대학원장, 명지대학교 컴퓨터공학과 교수

<관심분야> Mobile DB, Spatial DB, 한국어정보처리, Large Fingerprint DB, Web computing