논문 2006-43CI-6-7

주제 중심 수집기를 이용한 RSS 채널 추천 시스템

(RSS Channel Recommendation System using Focused Crawler)

이 영 석*, 조 정 원**, 김 준 일***, 최 병 욱****

(Youngseok Lee, Jungwon Cho, Jun-Il Kim, and Byung-Uk Choi)

요 약

최근 빠른 주기로 많은 양의 새로운 정보가 생성되기 때문에, 개인별 관심 분야의 전문화와 블로그의 보급을 위해 RSS라는 신디케이션 기술이 제공되고 있다. 사용자는 RSS 수집기에 RSS 채널의 주소를 등록함으로써, 새롭게 갱신된 콘텐츠를 자동으로 전달받을 수 있어서 신규 정보를 찾기 위해 사이트에 지속적으로 접근하지 않아도 된다. 본 논문에서는 사용자가 웹상에 존재하는 RSS 문서를 효과적으로 이용할 수 있도록 RSS 채널의 주소를 수집하는 주제 중심의 수집기와 사용자 질의에 따른 RSS 채널의 순위 부여 방안을 제안한다 제안된 RSS 수집기를 이용하면 사용자는 원하는 RSS 채널 주소를 효과적으로 검색할 수 있어서 자료 검색의 효율성을 증진시킬 수 있다.

Abstract

Recently, the internet has seen tremendous growth with plenty of enriched information due to an increasing number of specialized personal interests and popularizations of private cyber space called, blog. Many of today's blog provide internet users, RSS, which is also known as the syndication technology. It enables blog users to receive update automatically by registering their RSS channel address with RSS aggregator. In other words, it keeps internet users wasting their time checking back the web site for update. This paper propose the ways to manage RSS Channel Searching Crawler and collected RSS Channels for internet users to search for a specific RSS channel of their want without any obstacles. At the same time, This paper proposes RSS channel ranking based on user popularity. So, we focus on an idea of adding index to information and web update for users to receive appropriate information according to user property.

Keywords: RSS(Really Simple Syndication), 주제 중심 수집기(Focused Crawler), RSS 채널, 채널 추천(Channel Recommendation), 채널 순위 부여(Channel ranking)

I. 서 론

RSS는 최근 추가되거나 변경된 페이지들에 대한 요약 정보를 제공하는 기술로서, 사용자가 신규 정보를 찾기 위해 반복적으로 사이트에 접근하는 과정을 줄여줄 수 있다.

2004년 PEW INTERNET 조사에 따르면 미국의 1억 2천만 성인 인터넷 사용자 중 7%가 자신의 블로그를 만든 경험이 있고(800만 명), 27%가 블로그를 주기적으로 구독하고 있으며, 5%가 RSS 뉴스 수집기를 사용하고 있다[1]. 사용자는 RSS 수집기에 해당 RSS 채널의 주소만 기억해두면 나중에는 이 파일에 요약된 내용만 보고 중간 과정 없이 직접 해당 페이지로 접근할 수 있게 되므로 방문자들의 수고를 줄일 수 있다. 하지만, 현재의 시스템은 사용자가 직접 블로그를 돌아다니며 RSS 채널의 주소를 찾아 일일이 등록하는 작업을 통해 이루어지고 있다. 또한 아직까지 일반 사용자들에게는 RSS 문서 구분과 링크위치 파악에 있어 어려운 점이 존재한다. 대부분의 블로그 서비스를 제공하는 포털 사이트는 블로그 검색 도구를 제공한다. 그러나 자체 제

* 학생회원, **** 평생회원,
한양대학교 전자통신컴퓨터공학과
(Dept. of Electronics & Computer Engineering, Hanyang University)
** 평생회원, 제주대학교 컴퓨터교육과
(Dept. of Computer Education, Cheju National University)
*** 정회원, 위세아이텍 연구소
(Research Institute, WISE iTech Co., Ltd.)
접수일자: 2006년5월19일, 수정완료일: 2006년10월30일

공 검색 도구는 해당 포탈의 블로그로만 검색이 제한되며 블로그 제목, 소개글 정도를 기준으로 블로그를 검색해 낸다. 또한 글이 올라오지 않는 비어있는 블로그도 다수 검색되므로 효율적이지 못하다. 따라서 사용자는 스스로 블로그의 여러 글을 살펴보며 관심 블로그 여부를 파악하는 수고를 해야 한다.

본 논문에서는 사용자의 편의성을 도모하기 위해 블로그 서비스를 제공하는 여러 포털 사이트로부터 RSS 채널을 자동으로 수집하는 RSS 채널 탐색 수집기와 RSS를 이용하여 블로그의 정보 데이터베이스를 구성하는 시스템을 제안한다. 제안된 시스템을 통해 사용자는 다양한 질의를 통해 블로그의 RSS 채널을 직접 검색함으로써 RSS 수집기 사용 시 등록할만한 채널 주소를 찾는데 도움을 얻을 수 있다.

Ⅱ. 관련 연구

1. RSS

RSS는 Really Simple Syndication, 혹은 Rich Site Summary의 줄임말이며, RDF(Resource Description Framework) 기반의 콘텐츠 배급 프로토콜이다[2]. 이는 웹 정보 제공자 측에서 새로운 정보의 갱신 여부를 알려주는 용도로 쓰인다. 현재 RSS는 표 1과 같이 7가지 버전이 존재한다[3].

기존에도 RSS 수집기에 들어오는 신규 정보를 개인의 선호에 맞게 분류하려는 연구가 있었다[4]. 사용자는 RSS 수집기에 자신이 원하는 단어를 넣으면, 시스템은 RSS 문서 내의 연관 단어를 찾아 분류해 내는 방식이다. 하지만 이는 사용자가 입력한 단어와 연관된 단어

표 1. RSS 버전별 특징
Table 1. Specifications of each RSS versions.

버전	오너	설명	진행
0.90	넷스케이프	초기 버전	1.0에 의해 중단
0.91	유저랜드	0.90 간략화	2.0에 의해 중단되었지만, 많이 쓰임
0.92, 0.93, 0.94	유저랜드	0.91 확장	2.0에 의해 중단
1.0	RSS개발 그룹	RDF 기반	안정화코어, 모듈개발 진행 중
2.0	유저랜드	0.91 확장	안정화코어, 모듈개발 진행 중

를 찾기 위해, 별도의 계층적 지식 구조를 필요로 하는 한계점을 지닌다.

2. 웹 수집기

인터넷 사용자들은 원하는 정보를 찾기 위해 검색 사이트를 이용한다. 이러한 검색 사이트는 내부 데이터베이스를 가지고 있으며, 이러한 데이터베이스 구축을 위해 웹 수집기가 동작한다. 웹 수집기란 웹 문서 내부에 포함 된 하이퍼링크를 따라서 이동하면서 웹상의 정보를 수집하는 일종의 소프트웨어이다. 웹 수집기의 구현에서 고려되는 사항은 DNS 서버에 대한 병목 현상 최소화 웹 사이트에 대한 부하 감소, URL의 중복 처리, 문서 내용의 중복처리, 문서 탐색 방법 등이 있다[5].

웹 수집기는 기본적으로 깊이우선 탐색(Depth First Searching), 너비우선 탐색(Breath First Searching) 방식으로 웹의 모든 문서를 탐색한다. 만약 수집기가 탐색하고자 하는 대상이 웹의 전체가 아닌, CiteSeer와 같이 특정 관심 분야에 대한 웹 문서만 검색을 하는 경우, 문서에 나와 있는 모든 링크를 대상으로 움직일 필요는 없다. 이 경우 특정 주제와 문서와의 연관성에 따라 탐색할 링크의 수를 줄여나가는 수집기를 주제 중심 수집기라고 한다[6]. 주제 중심 수집기의 기본 알고리즘은 웹 문서를 단어의 집합으로 보고 페이지와 주어진 특정 주제와의 연관성을 코사인 유사도 측정으로 계산해 내는 방식으로, 이러한 방식의 우수성은 이전 연구에서 밝혀져 있다[7]. 이외에도 HTML 페이지의 트리구조나 문서 객체 모델(Document Object Model)을 이용하여 링크의 구문을 파악하여 이동하는 수집기 등이 있다[8].

본 논문에서 구현한 RSS 채널 탐색 수집기는 주제 중심 수집기의 너비우선 탐색 방식을 사용하였다. 또한 탐색 도메인을 RSS 채널 링크가 주로 존재하는 블로그 관련 포털 사이트로 국한시킴으로써 채널 탐색 효과를 높이도록 하였다.

3. 블로그 정보 검색

1인 미디어인 블로그의 증가는 활용할 수 있는 정보의 양을 대폭 늘었다. 이러한 블로그들의 많은 정보를 효율적으로 사용자들에게 제시하기 위해 각종 메타 블로그 사이트들이 존재한다.

국내의 대표적 메타 블로그 사이트로는 블로그코리아[9], 올블로그[10], 블로진[11] 등이 존재한다. 하지만 아직까지 이러한 블로그 사이트들은 가입자의 직접 등록에 의해서만 블로그 정보가 수집되며, 가입 시 블로그

주소와 함께 RSS 채널 주소까지 명시적으로 적어주어야만 정보를 모아올 수 있다. 또한 현재의 메타 블로그 사이트의 검색은 블로그 제목, 소개글로만 국한되어 검색된다. 본 논문에서 제안하는 시스템은 RSS 채널 정보 수집이 사용자의 상호작용없이 RSS 채널 탐색 수집기를 통해 자동화 되며 RSS 문서를 통해 글의 작성자, 작성시간 등을 추출하여 저장함으로써 사용자가 블로그 검색 시 다양한 질의를 가능하도록 하였다.

III. 실 험

1. 시스템 구성도

시스템은 그림 1과 같이 RSS 채널을 획득하기 위한 RSS 채널 탐색 수집기 부분과 탐색된 RSS 채널을 분석하여 데이터베이스에 저장하는 모듈로 구성된다. 획득된 RSS 채널은 이후 사용자의 질의에 따라 RSS 채널을 추천하게 된다.

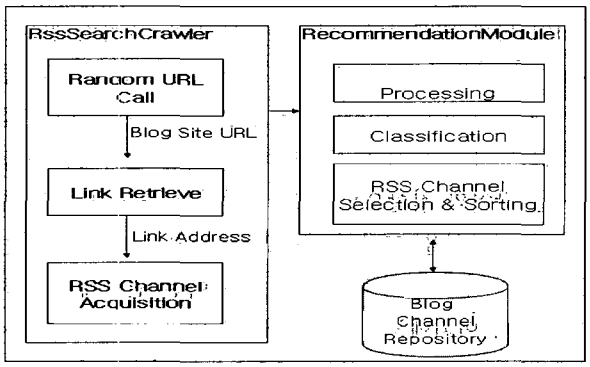


그림 1. 시스템 구성도
Fig. 1. Schematic diagram.

2. 주제 중심 RSS 수집기

주제 중심 RSS 수집기는 블로그에 존재하는 RSS 채널 주소를 찾아서 수집해 온다. 주제 중심 수집기 구현을 위해 고려된 사항은 다음과 같다. 첫째 블로그는 하나의 정형화된 구조를 띄지 않으며, 또 내용을 토대로 분류되는 웹페이지가 아니다. 따라서 웹을 대상으로 블로그 여부를 판단하는 것은 불가능에 가깝다. 둘째, RSS 채널 주소 링크는 보통 블로그 메인 페이지에 존재한다. 따라서 블로그에서 RSS 채널 주소를 찾기 위한 링크 탐색 범위는 블로그의 메인 페이지에 존재하는 링크 개수로 제한되어야 한다.

언급된 두 가지 문제점을 해결하고 RSS 채널 수집 효과를 높이기 위해 본 논문에서 제시한 주제 중심 수집기는 블로그 서비스를 제공하는 포털 사이트의 블로그 글의 접근 모듈을 사용한다. 해당 모듈의 호출을 통해 수집기는 지속적으로 블로그 주소를 획득하며, 블로그 메인 페이지에 존재하는 링크의 수로 탐색 큐의 크기를 제한한 후 너비우선 방식으로 조사한다. 단, 블로그 메인 페이지가 프레임 형태로 되어 있을 경우에는 추가 작업이 필요하다. 프레임에는 실제 링크 정보가 포함되어 있지 않으므로, 실제로 링크가 존재하는 웹페이지를 찾기 위해선 프레임을 구성하고 있는 페이지로의 접근이 필요하다. RSS 채널을 획득하기 위한 과정은 그림 2와 같다.

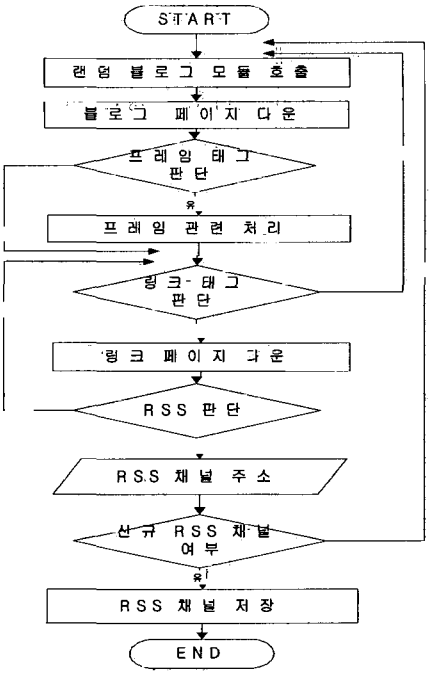


그림 2. RSS 채널 탐색 흐름도
Fig. 2. Flowchart for retrieval of RSS channel.

표 2. RSS 탐색을 위해 사용된 정규 표현식
Table 2. Regular expression for retrieval of RSS.

패턴 종류	정규 표현식
프레임 태그	<frame\\s+(.*?)\\s+src\\s*=\\s*\"?(.*?)[\"|>]
링크 태그	<a\\s+href\\s*=\\s*\"?(.*?)[\"|>]
RSS 문서 태그	^<channel

이때 수집기는 각 블로그 사이트 메인 페이지에 존재하는 태그들을 조사하기 위해 표 2와 같은 정규표현식(Regular Expression)[12]을 이용한다.

해당 블로그가 프레임으로 되어 있을 경우, 각 프레임에 해당하는 주소를 획득하여 실제 블로그 메인 페이지 주소를 재설정한다. 블로그의 메인 페이지에 존재하는 링크들을 조사하기 위해 <a href> 형식의 태그에서

링크주소를 획득한다. 수집기는 획득된 링크주소들에 접근하여 RSS 채널이 가져야 할 태그 여부를 확인 후 RSS 채널이라 판단되면, 기존에 이미 탐색된 채널인지의 비교를 통해, 신규 채널일 경우 해당 RSS 채널 주소를 저장한다.

3. RSS 채널의 관리 방안

습득된 RSS 채널은, 추천 후보 채널과 비추천 채널로 구분되어 관리된다. 여기서 비추천이란 그림 3과 같이 현재 블로그에 올라온 글이 없는 경우를 말한다. 이는 RSS 문서에 포함된 Item 여부로 확인할 수 있다.

비추천 채널이 아닌 나머지 채널은 추천 후보 채널로 배정되며, 이에 해당되는 채널들은 실제 채널 주소에 존재하는 RSS 문서로부터 정보를 추출하여 데이터베이스에 저장한다. RSS 문서로부터 추출 가능한 정보는 그림 4와 같다. 추천 후보 채널은, 이러한 RSS 문서를 통해 주기적으로 관리된다.

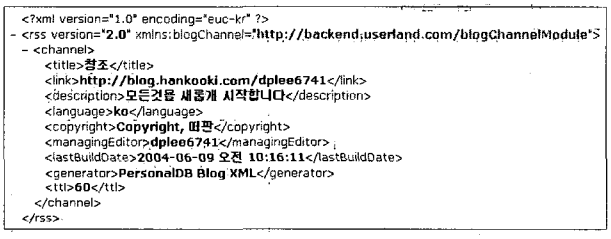


그림 3. 신규 정보를 포함하지 않은 RSS 문서
Fig. 3. RSS document (Not including new data).

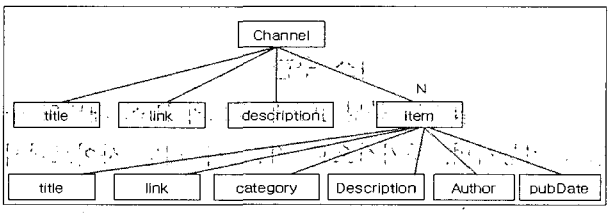


그림 4. RSS 문서의 구조
Fig. 4. The structure of RSS document.

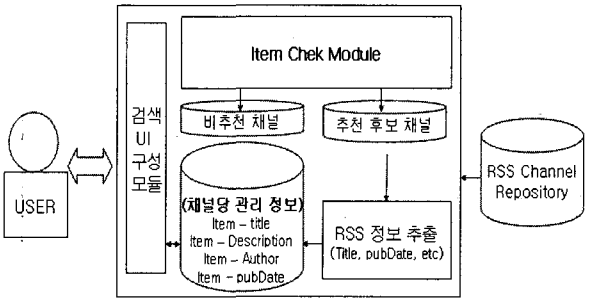


그림 5. RSS 채널 관리 구성도
Fig. 5. Schematic diagram for RSS channel Management.

RSS 채널 관리의 구성은 그림 5와 같다. 한 채널에 대한 정보는 RSS 문서 단위로 저장하는 것이 아닌 채널의 하위 요소인 아이템 단위로 저장한다. 이는 갱신되는 RSS 문서에 포함된 내용이 이전 RSS 문서에 비해 전부 새로운 아이템으로 구성된 것이 아니기 때문에 저장 공간의 효율성을 위해서이다.

아이템 하위의 주제와 설명 정보는 신규 게시물의 제목과 내용부분에 해당된다. 포탈별로 이 부분의 내용에 HTML 태그가 포함되는 경우가 있으므로, HTML 태그가 존재하는 경우 정규식을 이용하여 제거한 후 데이터베이스에 보관한다. 카테고리 정보는 블로거가 자신의 블로그에 설정한 카테고리 정보이다. 이는 비슷한 정보에 대해서도 블로거 별로 카테고리 이름을 달리 설정할 수 있으므로 일관성 있는 정보는 아니다. '다음'과 같은 포탈의 경우 이 정보를 RSS에 포함시키지 않고 있다. 따라서 이 정보는 데이터베이스에 보관하지 않는다. 저자 정보는 글의 작성자에 대한 정보이므로, 이를 데이터베이스에 저장함으로써 같은 작성자의 다른 게시물도 함께 검색이 가능하도록 지원한다. 그림 5의 pubDate 태그는 신규 정보가 올라온 시간을 나타내며, 사용자가 작성 시간에 따른 정보 검색이 가능하도록 한다.

4. 사용자 질의에 따른 RSS 채널 순위 방안

RSS의 기본 목적은, RSS를 통한 콘텐츠의 자동 전달이라고 할 수 있다. 이를 위해서는 사용자가 원하는 정보가 꾸준히 올라오는 순서대로 RSS 채널의 순위를 결정하여 제시해줄 필요성을 가짐에도 불구하고, 정보 전달을 목적으로 정보 갱신을 고려하여 블로그의 순위를 부여하는 검색 도구는 소개된 바 없다.

현재 사용자들이 RSS 채널을 이용하기 위해 차선책으로 사용하는 블로그 검색 도구는 블로그의 제목과 소개글 대상의 검색이 주로 이루어지므로 정보 갱신 여부와는 무관한 순위가 제시된다.

또한 기존에 제안된 순위 부여 방식을 사용하는 것에도 무리가 있다. 기존의 유사도에 따른 순위 부여[13]는 사이트 단위가 아닌 문서 단위의 유사도 계산에 적합하다. 예를 들어, 사용자가 질의한 단어를 다수 개 포함한 글이 1개 올라온 RSS 채널과 사용자가 질의한 단어를 한 개씩 포함한 다수개의 글이 올라온 RSS 채널의 경우를 생각해보면, 단순한 유사도 계산으로 순위를 매기면 전자가 우선순위를 높게 부여될 수도 있게 된다.

또한 하이퍼링크 정보를 이용한 HITS[14]나 PageRank[15]와 같은 순위 부여 방식은, 이 논문에서

RSS 채널의 도메인으로 설정한 블로그에는 적합하지 않다. 그 이유는 블로그에 존재하는 타 블로그와의 하이퍼링크 정보는 유동적으로 변하며, 링크의 수를 제한해 놓은 경우가 대부분이기 때문이다.

본 논문에서는 사용자에게 우선적으로 추천되어져야 할 RSS 채널은 "자료의 갱신주기가 짧고, 사용자가 관심 있어 하는 글이 많이, 그리고 꾸준히 올라올만한 채널"이라 정의하고 이를 위해 사용자 질의 단어를 포함하는 게시물의 수, 자료의 갱신률, 자료의 갱신주기 등을 고려한 RSS 채널 순위 부여 방법을 제안한다. 논문에서 제시된 방법을 적용하기위해서는 샘플 기간과 샘플 주기 설정을 필요로 한다.

RSS 채널의 순위를 정하기 위한 첫 번째 기준은 식 (1)과 같다.

$$\text{point} = \sum_{i=1}^{i=m} (k/n) \times w_i \quad (1)$$

point: 질의 단어의 위치와 게시물 수에 따른 우선순위 값
k: 질의 단어 수
n: 질의 단어 중 해당 게시물과 매칭된 수
m: 샘플 기간 내 질의 단어를 포함한 게시물의 수
w_i: 질의 단어 위치에 따른 가중치 (제목: w_i=1, 본문:w_i=0.5)

첫 번째 기준은 샘플 기간 내에서 존재하는 질의한 단어를 포함한 게시물의 양에 기반 하여 순위를 매기게 된다. 이는 기존의 검색 방법에 본문과 제목의 중요도 차이를 설정하여 사용자의 질의에 적합한 문서를 추출할 수 있도록 하였다. 첫 번째 기준에 의해 추출된 게시물 중에서 우선순위는 식 (2)와 같은 두 번째 기준을 적용하여 정한다. 두 번째 기준은 사용자의 정보 갱신의 일관성 여부를 파악하기 위한 것이다.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (k_i - m)^2 \quad (2)$$

n : (샘플링 기간) / (샘플링 주기)
k_i : i번째 주기의 질의단어 포함 게시물 개수
m : 샘플링 기간 내 평균 게시물 개수

예를 들어, 샘플 기간 2달, 샘플 주기 1주로 설정하였고, 블로그 A와 블로그 B의 샘플 주기 당 질의 단어 포함 게시물의 수가 블로그 A가 1,2,1,2,1,2,1,2 이고 블로그 B가 0,4,0,5,0,2,0,1 으로 가정한다면, 샘플 기간 내 질의 단어 포함 게시물의 수는 같지만 블로그 A가 블로그 B보다 더 작은 분산 값을 가지므로 더 높은 순위를 가지게 된다.

만약 첫 번째, 두 번째 알고리즘의 연산결과 동일한 순위가 있을 경우, 식 (3)과 같이 샘플 기간 내 질의 단어 포함 게시물의 갱신 주기를 파악하여 갱신 주기가 짧은 것이 더 높은 우선순위를 가지도록 한다. 제안된 시스템에서는 샘플 기간을 1년 이내로 제한하며, 샘플 주기와 자료 갱신 주기는 모두 하루 단위로 처리된다.

$$p = \frac{1}{m-1} \sum_{j=1}^{m-1} (d_{j+1} - d_j) \quad (3)$$

m = 샘플링 기간 내 질의어 포함 게시물 개수
d_j : j번째 게시물의 갱신일 (일단위 계산)

Ⅳ. 시스템 구현

1. 시스템 구현 환경

RSS 채널 탐색 수집기와 검색 데이터베이스를 구현하기 위한 실험 환경은 표 3과 같다.

표 3. 실험 환경
Table 3. Implementation environment.

구 분	RSS 탐색 수집기	검색 데이터베이스
H/W	Intel Pentium Dothon CPU 1.6 GHz 768 MB RAM	Intel Celeron CPU 1.6 GHz 512 MB RAM
S/W	Windows XP pro. JDK 1.5 Informa library	Windows 2000 server MS SQL-2000 server

2. 사용자 인터페이스 구현

구현된 수집기는 분당 100개 정도의 RSS 채널의 수집이 가능하며, 주기적으로 접근 시 해당 사이트에서

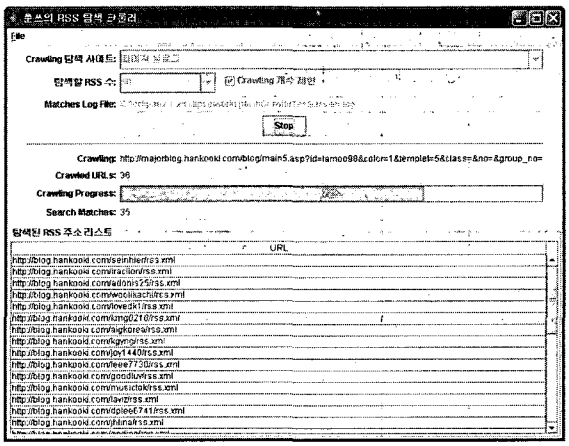


그림 6. RSS 탐색 수집기
Fig. 6. Retrieving crawler of RSS document.

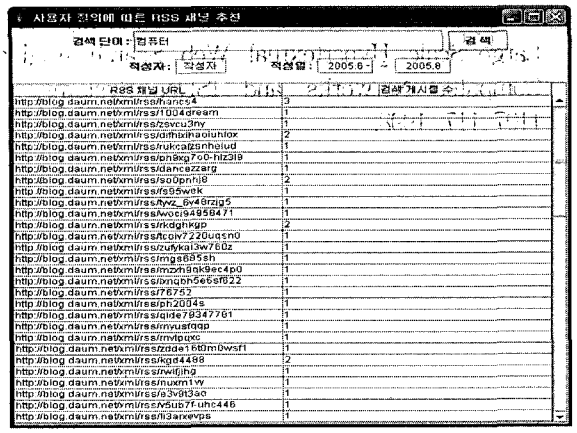


그림 7. 사용자 질의응답 화면

Fig. 7. User Interface for user interactions.

그림 8. 제안 방법을 적용한 검색 화면

Fig. 8. retrieval result of the proposed ranking method.

접근을 막을 수 있으므로, 50개 당 서로 다른 블로그 사이트에 접근한다.

그림 6은 메이저 블로그 포털 사이트의 RSS 채널을 탐색해 나가는 모습을 보여준다. 구현된 RSS 탐색 수집기는 다음, 엠파스, 하나포스, 코리아닷컴, 메이저 등 총 6개의 블로그 사이트의 RSS 채널 검색을 지원한다.

사용자 질의에 대한 구현 화면은 그림 7과 같다. 키워드, 작성일, 작성자로 검색 가능하다.

그림 8은 제안한 RSS 채널 순위 부여 알고리즘을 적용하여 구현된 인터페이스이다. 이곳에서 사용자는 샘플링 주기와 기간을 설정하여 자신에게 알맞은 RSS 채널 검색이 가능하도록 한다.

V. 시스템 평가

본 논문에서 제시한 사용자 질의에 대한 RSS 채널 추천을 사용한 경우와 국내의 블로그 서비스를 제공하는 포털 사이트 중 엠파스에 존재하는 블로그 검색 도구와 비교 실험을 해본 결과 그림 9와 같은 결과를 얻을 수 있었다. 엠파스에서 제공되는 블로그 검색 도구는 블로그 제목(소개글 포함) 대상 검색과 블로그 카테고리 검색, 블로그 신규 게시물 검색을 제공한다. 다음, 파란, 하나포스, 코리아닷컴, 메이저 등도 동일한 검색 도구를 지원한다.

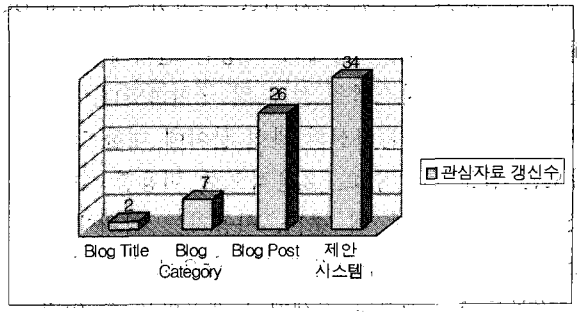


그림 9. RSS 채널 추천의 성능 비교

Fig. 9. Performance evaluation of the RSS Channel recommendation.

2005년 10월 20일 기준으로 검색하였으며, 각 검색 도구의 상위 랭크된 10개 블로그의 RSS 주소를 취하여 향후 15일간 사용자가 질의한 검색어를 포함한 신규 자료의 갱신 정도를 비교하였다. 검색어는 '윈도우' 이며, 영문 'Windows'도 함께 포함하여 검색한다. 제안된 시스템의 RSS 채널 순위 부여 방법을 이용하기 위해서는 샘플링 주기와 기간 설정이 필요한데, 여기서는 샘플 주기는 14일로 선택하고 2005년 7월 29일부터 2005년 10월 20일 까지(84일)를 샘플링 기간으로 설정하였다.

그림 9의 결과에서 알 수 있듯이, 실제 포털 사이트에서 제공하는 검색 결과를 사용한 경우보다 제안된 시스템의 검색 결과를 사용한 경우의 자료 갱신률이 더 뛰어난 것으로 나왔다. 이는 기존 블로그 검색엔진을 통해서는 자체 운영하는 블로그의 사이트 제목과 소개글에 바탕을 둔 결과가 나오기 때문에 앞으로의 갱신성에 대한 고려는 없었다. 하지만 제안된 시스템은 사용자 질의어에 기반을 둔 자료의 수, 갱신 주기, 갱신률의 분산도 등을 모두 고려하였기 때문에 기존 결과에 비해 우수한 성능을 보일 수 있었다.

VI. 결 론

블로그의 증가로 사용할 수 있는 RSS 채널은 많아졌지만, 사용자가 원하는 RSS 채널을 이용하기 위해서는 각 블로그를 직접 돌아다니며 블로그의 내용을 판단하고 RSS 채널 주소를 찾아내야 하는 문제점이 있다.

본 논문에서는 주제 중심 RSS 수집기를 이용한 RSS 채널 자동 탐색과, 사용자의 다양한 질의에 대하여 채널을 추천하는 시스템을 설계 및 구현하였다. 제안하는 시스템은 사용자가 자신의 관심 분야에 대한 정보가 올라오는 블로그를 쉽고, 효율적으로 이용할 수 있도록 RSS 채널을 검색 해 준다. 사용자는 상위에 검색된 RSS 채널을 RSS 수집기에 등록함으로써 자신이 원하는 정보를 효율적으로 검색하고 제공받을 수 있다.

참 고 문 헌

[1] PEW INTERNET & AMERICAN LIFE VPROJECT, http://www.pewinternet.org, 2004.

[2] World Wide Web Consortium, http://www.w3c.org, 2005.

[3] RSS Technology Reports, http://www.oasis-open.org/cover/rss.html, 2005.

[4] Weihong Huang, "Enabling Context-Aware Agents to Understand Semantic Resources on the WWW and The Semantic Web", Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 138-144. 2004.

[5] 김성진, "웹 로봇 구현 및 한국 웹 통계 보고", 한국 정보처리 학회, 10권 C편, 제 4호, 2003.

[6] Soumen Chakrabarti, "Focused crawling: a new approach to topic-specific web resource discovery", In Proc. of 8th International World Wide Web Conference, 1999.

[7] F. Menczer, "Evaluating topic-driven Web crawlers", Proc. 24th annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 241-149. 2001.

[8] Gautam Pant, "Topical Crawling for Business Intelligence", Proc. of ECDL 2003, pp. 233-244, 2003.

[9] BlogKorea, http://www.blogkorea.org, 2005.

[10] AllBlog, http://www.allblog.net, 2005.

[11] BLOZINE, http://www.blozine.com, 2005.

[12] Haruo Hosoya, "Regular expression pattern matching for XML", Proc. of the 28th ACM SIGPLAN-SIGACT symposium on Principles of programming lanuage, pp.67-80. 2001.

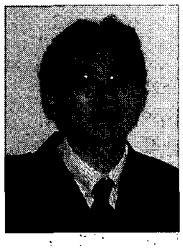
[13] B. Yuwono, "Search and ranking algorithms for locating resources on World Wide Web", Proc. of the Int. Conf. on Data Engineering, pp. 164-171, 1996.

[14] Jon Kleiboemer, "Authoritative sources in a hyperlinked environment", Proc. of the 9th ACM-SIAM symposium on Discrete Algorithms, pp. 668-677, 1998.

[15] Brin, S., & Page, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, pp. 1107-117. 1998.

저 자 소 개

이 영 석(학생회원)
1998년 서울교육대학교 초등교육과 학사 졸업.
2001년 서울교육대학교 교육대학원 컴퓨터교육과 석사 졸업.
2003년~현재 한양대학교 전자통신컴퓨터공학과 박사과정.
<주관심분야 : 모바일 학습, 지능형 교육 시스템, 멀티미디어 콘텐츠 처리, 온톨로지>



김 준 일(정회원)
2003년 동국대학교 전자컴퓨터공학과 학사 졸업.
2006년 한양대학교 정보통신공학과 석사 졸업.
2006년~현재 (주)위세아이텍 연구소 연구원
<주관심분야 : 웹서비스, 온톨로지, 웹 기반 시스템>

조 정 원(평생회원)-교신저자
1996년 인천대학교 정보통신공학과 학사 졸업.
1998년 한양대학교 전자통신공학과 석사 졸업.
2004년 한양대학교 전자통신전파공학과 박사 졸업.
2004년~현재 제주대학교 컴퓨터교육과 조교수.
<주관심분야 : 정보교육, 유비쿼터스 학습, 프로젝트 관리 및 평가, 멀티미디어 정보검색>

최 병 욱(평생회원)
1973년 한양대학교 전자공학과 학사 졸업.
1978년 일본 경응의숙대학(KEIO) 전기공학과 석사 졸업.
1981년 일본 경응의숙대학(KEIO) 전기공학과 박사 졸업.
1981년~현재 한양대학교 정보통신대학 정보통신학부 교수
<주관심분야 : 영상처리, 멀티미디어 공학, 유비쿼터스 학습>