웹 자원 검색을 위한 효율적인 콘텐츠 배급

심현진, 고훈준, 김기태, 조선문, 유원희
인하대학교 전자계산공학과
e-mail:g2022035@inhavision.inha.ac.kr

A Efficient Contents Syndication for the Web Resource Retrieval

Hyun-Jin Sim, Hoon-Joon Kouh, Ki-Tae Kim,
Sun-Moon Jo, Weon-Hee Yoo
Dept. of Computer Science & Engineering, Inha University

요 약

RSS(RDF Site Summary)는 웹 사이트의 콘텐츠를 간단하게 기술하고 RSS 수집기는 사용자에게 이를 적절한 형식으로 보여준다. 그러나 현재의 RSS는 자신의 콘텐츠에 대해서만 기술하고 있어서 웹 자원의 검색 측면에서 볼 때 사용자는 그와 유사한 콘텐츠를 갖는 다른 웹 사이트를 찾는데 또 다른 노력을 들여야 한다. 본 논문에서는 현재의 RSS에 다른 웹 사이트로 연결되는 추가적인 요소를 통하여 웹 자원의 탐색에 RSS를 이용할 것을 제안한다.

1. 서론

시맨틱 웹(Semantic Web)에서 사용하는 온톨로지 언어인 RDF(Resource Description Framework)의 응용된 사례로 현재 사용되고 있는 RSS는 XML을 기반으로 하는 경량의 콘텐츠 배급(Syndication) 포맷이다[1]. RSS는 웹 콘텐츠의 제목과 내용 등을 공유하기 위한 목적으로 사용되고 있다. 즉, 웹 사이트 내의 콘텐츠를 RSS 포맷으로 요약된 메타데이터를 만들고 이 메타데이터를 수집하는 응용프로그램을 이용하면 웹 브라우저를 통해 매번 각각의 웹사이트를 방문하지 않고도 정렬된 최신의 콘텐츠를 볼 수 있다.

현재 개개인의 RSS 파일들을 등록할 수 있는 일종의 검색엔진과 같은 웹 사이트들이 큰 규모로 생겨나고 있고, 이 곳에 자신의 RSS 파일의 경로를 등록하는 것은 기존의 검색엔진에 자신의 웹 사이트를 등록하는 절차보다 훨씬 간편하며 등록되는데 걸리는 시간도 거의 실시간에 가깝다. 따라서 RSS 파일을 만들어 콘텐츠를 배포하려는 사람은 적은 시간에 많은 곳에 자신의 웹 사이트를 알릴 수 있고 방문객을 모을 수 있다.

RSS는 넷스케이프사의 서비스에서 시작하여 오늘에 이르렀다. 이미 널리 알려진 대형 사이트들도 RSS로 최신의 갱신상황을 인터넷 사용자들에게 알려주고 있고 점차 개인에게까지 확대되어 사용되고 있는 실정이다.

W3C의 주도 하에 다음 세대의 웹 서비스로 연구가 진행중인 시맨틱 웹으로 가는 과정에서 현재까지 RSS는 가장 성공한 XML 기반의 응용으로 평가되고 있다. 그러나 현재의 RSS는 자신의 웹 사이트의 콘텐츠에 대해서만 요약한 정보이다. 이를 이용하는 사용자는 그 콘텐츠와 유사하거나 관련성이 깊은 콘텐츠를 가진 다른 웹 사이트를 검색할 가능성이 높지만 RSS는 자신의 웹 사이트에 대한 1차적인 정보만을 제공해 줄뿐이므로 사용자는 관련성이 있는 다른 RSS를 찾기 위해 다시 RSS 파일들의 주소가 등록되어 있는 목록들을 눈으로 일일이 확인해야 한다.

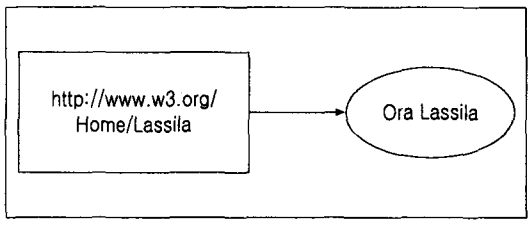
본 논문에서는 기존의 RSS 파일에 콘텐츠 배포자가 자신의 RSS 파일과 유사하거나 더 나은 콘텐

츠를 가진 다른 웹 사이트의 RSS 파일을 연결할 수 있는 링크를 제공하여 RSS 파일의 이용자가 콘텐츠를 수집하는데 드는 시간과 노력을 줄일 수 있도록 제안한다.

2. 관련 연구

메타데이터(Metadata)는 데이터에 관한 데이터이다. 전달하려고 하는 1차 자료에 접근하기 위한 2차 자료로써 도서관의 도서목록이나 지도의 범례 등도 일종의 메타데이터라고 할 수 있다. 메타데이터를 이용해서 웹에 있는 방대한 자원들을 종류와 특성에 따라 간단하게 기술하며 원하는 자원에 쉽게 접근할 수 있고 정확한 검색에 유용하게 쓰이고 있다. 메타데이터의 핵심은 간결성이다. 현재 대표적인 메타데이터로는 더블린 코어(Dublin Core)가 있다. 네트워크에 있는 자원의 기술에 필요한 요소를 정하고 쉽고 정확한 검색을 목적으로 1995년에 합의된 메타데이터이다. 더블린 코어는 15개의 핵심 기술요소로 이루어져 있다. 이들 기술요소들은 자원의 내용, 다른 자원과의 관계, 자원의 속성, 형식, 시간을 나타낸다. 더블린코어를 기초로 다른 분야에서도 확장하여 사용하는 것이 가능하다[2].

RDF는 메타데이터를 처리하기 위한 기초로써 현재 W3C의 표준으로 정해져 있다. 기계가 이해할 수 있는 정보를 교환하는 응용프로그램 간에 상호운용성을 제공하며 자원의 자동적인 처리를 하는데 편리하다[3]. 자원검색에 있어서는 개선된 검색능력을 제공하고 목록분야에서는 콘텐츠나 콘텐츠 간의 관계를 기술할 수 있다. RDF는 자원의 속성과 값을 XML의 방식으로 표현하고 데이터에 의미를 부여한다. 기본적인 데이터 모델은 자원, 속성, 문으로 구성되어 있다[4].



[그림 1] 간단한 RDF 데이터 모델

자원은 속성을 가지며 웹에서 URI(Uniform Resource Identifiers)로 식별되는 모든 객체를 가리킨다. 웹 페이지의 전체나 일부가 될 수도 있고 웹 페이지들의 집합을 나타낼 수도 있다. 속성은 자원의 기술을 위해 사용된 특징, 관계를 말한다. 문은 자원과 특성과 값을 말하며 각각 주부, 술부, 객체라고 한다. 객체는 다른 자원일 수도 있고 문자열을 가리키기도 한다. [그림 1]은 간단한 RDF 데이터 모델이다.

RSS는 웹 사이트가 가지고 있는 콘텐츠의 배급 포맷이며 다목적으로 확장 가능한 메타데이터 기술 방법이다. XML 응용으로써 앞에서 설명한 RDF의 명세를 따르고 있다. RSS는 XML 이름공간을 사용하여 확장 가능하다. 아이템(Items)으로 표현되는 각각의 콘텐츠가 URL로써 표현되고 아이템을 포함하는 채널(Channel)이 RSS를 기술한 포맷이 된다. 각 아이템은 제목, 링크, 콘텐츠에 대한 간단한 기술로 이루어진다. 1999년에 넷스케이프사(Netscape)에서 처음 소개되었고 자사의 웹 사이트에 대해 콘텐츠를 모으는 메커니즘으로 사용되었다. 이후 RSS는 메타데이터에 대한 필요로 수집되고 분류되는 과정을 거치며 성장했다.

```
<XML 선언>
<이름공간 선언>

<채널>
        <RSS의 주소>
        <제목>
        <웹 사이트의 주소>
        <간단한 기술>
        <각 콘텐츠의 주소>
</채널>

<콘텐츠에 상세한 기술>
```

[그림 2] RSS의 개략적인 구조

[그림 2]는 RSS 파일의 개략적인 구조를 보여준다. 채널은 요약정보를 나타내고자 하는 웹 사이트 자체를 나타내고 제목은 배포하려고 하는 콘텐츠를 대표해야하고 하단에는 각 콘텐츠에 대해 주소와 상세한 기술을 덧붙인다. 이렇게 만든 RSS 파일은 자신의 서버 컴퓨터에 저장하고 웹 사이트에서 RSS 파일의 경로를 알려준다. 웹 사이트를 방문한 사람은 그 경로를 자신의 응용프로그램에 저장하게 되면 응용프로그램은 사용자에게 적절한 형식으로 그 웹 사이트의 콘텐츠들을 보여주게 된다.

3. RSS에 새로운 요소 추가

위에서 본 것처럼 현재의 RSS 파일은 방문자들에게 자신의 웹 사이트에 대한 콘텐츠들을 간략한

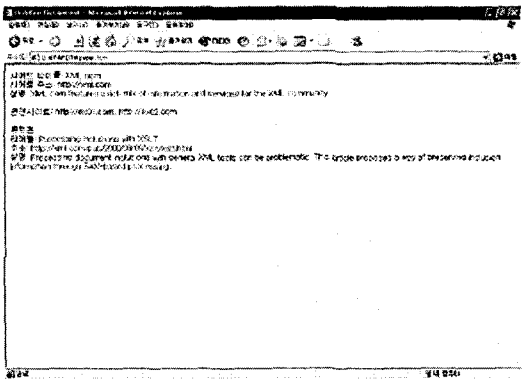
정보로 제공한다. 그러나 본 논문에서는 현재의 용도로만 RSS 파일을 이용하는 것이 아니라 각각의 RSS 파일을 웹에 있는 하나의 리소스로 보고 이것들간의 관계를 링크로써 제공하려고 한다. 더불어 이에 맞는 적절한 RSS 수집기가 제공된다면 RSS 수집기를 사용하여 웹 사이트들의 콘텐츠를 제공받는 이용자들은 자신이 현재 발견한 웹 사이트와 유사하거나 더 나은 콘텐츠를 가진 웹 사이트의 주소도 함께 알게 되어 이용이 가능할 것이다. 이러한 과정들을 구현하기 위하여 본 논문에서는 먼저 기존의 RSS 파일 안에 다른 웹 사이트로 연결되는 링크 요소를 추가하려고 한다. 그리고 추가된 요소까지 함께 파싱하여 사용자에게 적절한 형태로 보여주는 파서를 윈도우즈환경에서 웹 브라우저로 구현하였다.

```
<channel rdf:about="http://example.com/news.rss">
    <title>Example Channel</title>
    <link>http://example.com/</link>
    <description>My example channel</description>
    <items>
      <rdf:Seq>
        <rdf:li resource="http://example.com/2002/09/01/"/>
        <rdf:li resource="http://example.com/2002/09/02/"/>
      </rdf:Seq>
    </items>
    <rsslink>
      <rdf:Seq>
        <rdf:li resource="http://ex01.com/ex01.rss"/>
        <rdf:li resource="http://ex02.com/ex02.rss"/>
      </rdf:Seq>
    </rsslink>
  </channel>
```

[그림 3] RSS 파일의 링크 추가

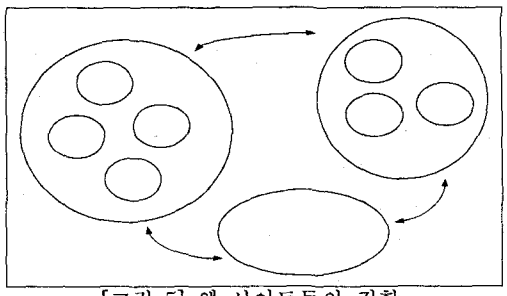
[그림 3]은 기존의 RSS 파일 내부에 비슷한 콘텐츠를 가진 다른 웹 사이트의 RSS 파일의 링크를 추가한 것이다. 이와 같은 형태는 현재의 RSS 명세에 추가적인 요소가 필요한 경우이고 다른 표현으로는 RSS 파일의 상단에 이름공간을 선언하는 부분에서 적절한 요소를 가진 이름공간을 선언하여 요소를 사용하는 방법도 가능할 것이다.

관련성을 가진 콘텐츠가 포함된 다른 웹 사이트를 검색하는 시간을 절약하며 검색과정에서 있을 수 있는 실패나 필요없는 자료를 걸러야하는 노력을 줄일 수 있을 것으로 기대한다. 현재의 웹은 기본적으로 분산환경인 성격을 가지고 있다. 그러한 웹에 수없이 흩어져있는 자원들간에 연관성을 갖도록 연결한다면 사용자들은 원하는 정보를 더 쉽게 찾아낼 수 있을 것이다. 현재 연구가 진행중인 시맨틱 웹에서도 리소스들 간의 관계는 매우 중요하다[5]. 시맨틱 웹이 말하는 자동화된 웹 서비스나 검색은 결국 RDF와 같은 언어에 의해 표현된 의미있는 관계들로부터 출발한다. 이 관계들이 지식 베이스를 만들고 시맨틱 웹은 이것들을 사용하여 이루어지기 때문이다.



[그림 4] 추가된 RSS를 이용한 RSS 수집기

[그림 4]는 앞에서 추가된 요소를 적용하여 RSS 파일을 읽어들인 RSS 수집기의 모습이다. 구현된 RSS 수집기는 웹 브라우저로 보여지며 연관된 콘텐츠를 가진 다른 웹 사이트의 연결까지 보여주고 있다.



[그림 5] 웹 사이트들의 집합

[그림 5]는 이런 연결들로 이루어진 웹 사이트들의 집합이다. 작은 원들이 유사한 콘텐츠를 가진 웹 사이트들이고 큰 원이 그 웹 사이트들의 집합이다. 사용자는 그 집합을 탐색하는 것으로 원하는 정보를 얻어내고 응용 프로그램을 이용하여 RSS 파일을 등록하면 그 웹 사이트들에 대해 계속 최신의 정보를 얻어낼 수 있고 RSS 파일의 제작자에 의해 새로운 연결도 알아낼 수 있다. 결과적으로 원하는 자원이 있는 위치를 검색하는 데에 드는 시간과 노력을 줄일 수 있을 것이다.

4. 결론 및 향후 연구

현재 웹은 수많은 정보들로 이루어져 있어서 이

정보들을 효율적으로 찾아내기 위해서는 도서관의 도서목록과 같은 역할을 하는 메타데이터의 중요성이 점점 부각되고 있다. 또한 메타데이터를 만든 후에는 그 메타데이터들의 상호운용성과 메타데이터들 사이의 관계가 중요성을 가진다. 본 논문에서는 웹 사이트의 요약된 정보를 나타내는 RSS 파일들 사이의 연결을 통해 유사한 의미를 가진 자원들의 연결을 제안하여 개인이 필요로 하는 자원을 RSS 파일과 응용 프로그램으로 쉽게 발견할 수 있도록 하였다. 이렇게 기술한 RSS 파일에 맞추어 응용 프로그램도 함께 개발되어져야 한다.

향후 연구로는 RSS 파일의 내용을 보여주는 응용 프로그램에서 RSS 파일이 가지고 있는 각 아이템에 관한 것이다. 응용 프로그램에서 보여지는 모든 아이템들 중에는 검색하는 사용자에게 불필요한 아이템도 존재할 것이다. 사용자가 한 웹 사이트에서 이용하는 아이템들에 대해 선호도를 평가하여 검색하는 사람에게는 자신의 선호도에 따라 보여주고 콘텐츠를 배급하는 사람에게는 그 아이템에 주력할 수 있도록 하는 방법을 향후 연구로 진행해야 할 것이다.

참고문헌

[1] Gabe Beged-Dov, Dan Brickley, Rael Dornfest, Ian Davis, RDF Site Summary (RSS) 1.0, http://web.resource.org/rss/1.0/spec, 2001

[2] Diane Hillmann, Using Dublin Core, http://dublincore.org/documents/usageguide, 2003

[3] Ora Lassila, Ralph R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222, 1999

[4] Michel Klein, XML, RDF, and Relatives, IEEE INTELLIGENT SYSTEMS, 2001

[5] Sheila A. McIlraith, Tran Cao Son, Hongglei Zeng, Semantic Web Services, IEEE INTELLIGENT SYSTEMS, 2001