지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템 연구

이성열ᵘ 백혜정 박영택
숭실대학교 컴퓨터학과
{trident, hjbaek}@multi.soongsil.ac.kr park@computing.soongsil.ac.kr

A Study of Knowledge Based Agent System for Web News-Document Retrieval

Sung-Yull Leeᵘ Hea-Jung Baek Young-Tack Park
Dept. of Computer Science, Soongsil University

요 약

현재 인터넷상의 정보와 문서의 양은 상상을 초월하는 증가추이를 나타내고 있다. 이와 더불어 표현하려는 목적에 따라 체계적으로 정리되고 정형화된 문서들 또한 증가하고 있다. 이러한 문서들 중에는 각 인터넷 신문사나 웹진과 같은 문서들이 포함되는데, 이러한 문서들은 각각의 내용구성과 표현 형식에 있어서 비슷한 구성을 지니고 있다. 본 논문에서는 이러한 체계적이고 정형화된 웹 뉴스 문서검색을 위하여 '지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템'을 제안한다. 사용자는 시스템에서 제공하는 지식을 기반으로 검색하고자 하는 대상을 에이전트 시스템에게 요청하게 되고 지식기반을 이용한 에이전트 시스템은 보다 정확한 정보를 사용자에게 제공하게 된다.

1. 서 론

현재 인터넷상에서의 문서증가 속도는 검색시스템의 한계를 뛰어넘은지 오래되었다. 새로운 문서를 검색시스템이 발견하기도 전에 또다른 문서가 생성되고 있는 것이다. 또한 사용자들은 원하는 정보를 검색하기 위하여 일반적인 검색시스템을 사용하지만 자신이 원하는 정보를 얻기란 쉽지 않다. 이러한 이유로 인하여 현재는 특정 분야, 특정 정보들에 대해서만 검색을 해주는 검색시스템들이 등장하게 되었고 일반적인 검색시스템 보다는 사용자 만족도를 높여주고 있다. 이렇게 특정 분야에 대한 검색시스템이 사용자 만족도가 높은 이유는 인터넷상의 문서의 양이 증가함에 따라 특정 분야에 대해서 보다 체계적이고 정형화된 문서들 또한 증가하고 있고 이에대한 검색만을 전문적으로 하는 검색시스템은 그 결과가 일반적인 검색 시스템에 비하여 만족스럽게 나타나기 때문이다.

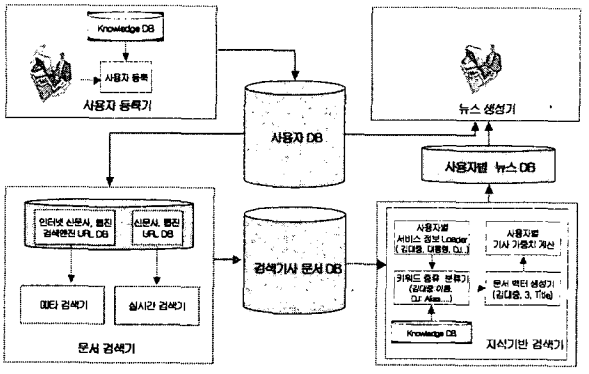
본 논문에서는 이러한 일반 사용자들의 일반적인 정보검색시 발생하는 문제점을 극복하고 특정한 분야에 대한 집중적인 문서검색을 위하여 '지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템'을 제안한다.

제안하는 시스템은 인터넷신문이나 웹진등에서 사용자들에게 뉴스를 전달하기 위한 목적으로 생성하는 뉴스문서를 특정 대상으로 선정하였고 이에 대한 검색을 실시하고 결과를 사용자에게 제공한다. 또한 단순한 검색시스템을 탈피하여 사용자가 원하는 지식을 기반으로 검색결과의 정확성을 향상시키고 검색 결과를 볼 수 있도록 지식기반 검색 에이전트 기능을 수행한다. 본 논문의 구성은 다음과 같다. 2장에서는 본 시스템의 개략적인 구조를 설명하고 3장에서는 지식기반 검색방식을 설명한다. 4장에서는 시스템구현에 대하여 간략히 설명하고 5장에서는 결론 및 향후 연구에 대하여 논하겠다.

2. 시스템 구성

본 논문에서 제안하는 '지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템'은 크게 4부분으로 구성된다. 먼저 사용자가 시스템이 제공하는 지식을 바탕으로 자신이 원하는 정보를 선택하는 사용자 등록기, 입력된 정보를 키워드(keyword)로 하여 인터넷 신문과 웹진등에서 문서를 검색하는 문서검색기, 검색된 문서들을 지식기반 방식을 이용하여 제 분류하고 문서 가중치를 계산하는 지식기반 검색기, 사용자에게 검색된 정보를 제공하는 뉴스생성기로 구성된다.

다음 [그림 1]은 본 시스템의 구성을 간략화 시킨 시스템 개략도이다.



[그림 1] 지식기반 방식을 이용한 웹 뉴스검색 시스템 개략도

• 사용자 등록기 : 사용자는 원하는 정보를 키워드로 입력한 후 시스템에서 해당하는 키워드에 대하여 제공하는 지식을 기반으로 상세 지식을 작성하여 나간다. 이때 입력되는 사용자 지식은 지식기반 검색기에서 해당문서를 분류하는데 있어서 시스템에서 제공하는 지식과 함께 영향을 미치게 된다.

• 문서 검색기 : 사용자가 입력한 키워드를 기반으로 문서를 검색한다. 이때 문서를 검색하는 방법으로 메타(meta)검색과 실시간(real-time) 검색기를 사용한다. 메타검색기는 각 인터넷 신문사와 웹진등의 검색엔진을 이용하는 메타검색을 실시하여 해당하는 키워드의 문서를 수집한다. 실시간 검색기는 사용자의 키워드에 대한 문서수집이 아닌 시스템에 등록되어 있는 인터넷 신문사와 웹진등의 페이지에 대하여 하이퍼링크(hyper-link)를 기준으로 해당하는 하위 페이지들에 대하여 페이지 수집을 한다. 이러한 무작위적 문서수집 방법은 메타검색으로 찾지 못하는 문서들을 수집하기 위해서이다. 이때 두가지 방법모두 수집되는 문서의 URL만을 수집한다. 이러한 이유는 문서의 수가 증가함에 따라서 부가되는 문서 저장용량의 부담을 줄이기 위해서 이다.

• 지식기반 검색기 : 문서검색기에서 수집된 문서는 사용자 키워드를 반영하여 수집된 문서도 있지만 그렇지 않은 실시간 검색기에서 수집된 문서와 같이 혼재하게 된다. 지식기반 검색기에서는 이러한 문서들에 대하여 지식기반 방식을 이용한 가중치 계산과 키워드를 이용한 분류를 통하여 각 사용자가 원하는 문서에 대한 필터링(filtering) 작업과 문서랭킹(document ranking)작업, 분류작업을 수행하게 된다. 지식기반 방식을 이용한 가중치 계산에 대해서는 3장에서 자세히 설명하겠다.

• 뉴스 생성기 : 사용자는 신청한 뉴스 정보를 검색하기 위하여 뉴스 생성기를 이용하게 된다. 뉴스 생성기는 사용자의 뉴스정보를 종류별로 분류하여 보여주게 된다. 사용되는 뉴스의 종류는 최신뉴스(중요), 지난뉴스(중요), 최신뉴스(보통), 지난뉴스(보통)으로 분류된다. 각 분류의 방법은 뉴스의 가중치가 우선이 되어 '중요', '보통'으로 먼저 분류되며 뉴스의 일자에 따라서 '최신', '지난'으로 분류된다.

3. 지식기반 방식의 문서검색

특정한 단어에 대하여 문서 검색을 수행할 때 단일 키워드만을 이용한 검색보다는 연관되는 키워드들을 이용한 검색 결과가 더욱 만족스러운 결과를 보일 수 있다. 이러한 문서 검색의 효율은 뉴스문서라는 특정 범위 안에서의 문서검색에 있어서 뚜렷이 반영된다. 예를들어 어떠한 사용자가 '박찬호(야구선수)'라는 인물에 대하여 뉴스검색을 신청하였다. 이때 검색되는 문서들은 '박찬호'라는 키워드가 들어가 있는 문서들을 집중적으로 검색하여 올 것이다. 하지만 '박찬호'라는 키워드가 들어가 있다고 하여서 모두 '박찬호'에 대한 뉴스 문서는 아닐 것이다. 다른 뉴스내용의 보도를 위하여 '박찬호'라는 이름이 언급되어 있을수도 있고, 뉴스의 주요 내용이 '박찬호'가 아닌 제 3자의 내용일 수도 있는 것이다.

이러한 검색결과의 부정확성을 감소시키기 위하여 별명, 직업, 소속, 영, 한문이름 같은 '박찬호'의 지식을 이용하여 지식기반 방식의 문서검색을 실시하게 된다. 본 시스템에서는 문서검색을 위한 키워드의 종류를 크게 4가지로 분류하고 그에 따른 지식의 종류를 정의하였다. 다음 [표 1]은 키워드의 종류와 그에따른 지식의 정의이다.

키워드분류	개인	기관, 회사	학교
지식정의	키워드(이름)	키워드(기관명, 회사명)	키워드(이름)
	별명	약어	약어
	직업	약어	단과대
	소속	부서	학과
	영문이름	업종	지역
	한문이름	영문이름	한문이름
	세부키워드	한문이름	영문이름
		세부키워드	세부키워드

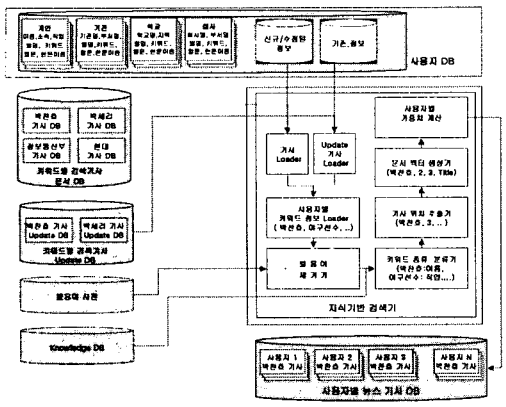
[표 1] 키워드 종류와 지식의 정의

다음은 예를 들어서 본 시스템에서 사용한 '박찬호'에 대한 지식을 살펴보겠다.

-키워드 분류 : 개인
-지 식 : 키워드(박찬호), 별명(코리안 특급, 볼 파크), 직업(야구선수, 스포츠인), 소속(미국 LA다저스), 한문이름(朴贊浩), 세부키워드(메이저리그, 야구, 투수)

위와 같은 '박찬호'에 대한 확장지식은 문서분류 작업에

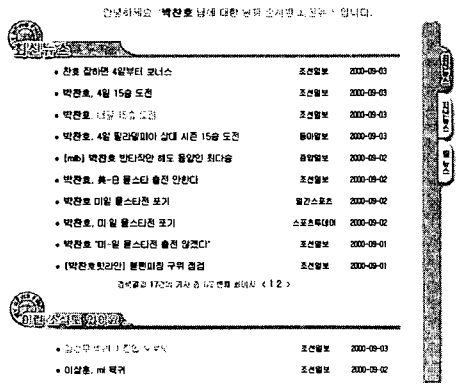
있어서 문서에 대한 가중치 계산에 큰 영향을 미치게 된다. 즉, '박찬호'라는 키워드가 들어간 문서라 할지라도 '박찬호'지식에 포함되어 있는 단어들이 어느 정도 문서에 포함되었는지 또한 '박찬호'라는 키워드가 나타난 위치와 어느정도 밀접해 있는지 등에 따라서 문서의 가중치 값이 틀리게 나타난다. 다음 [그림 2]는 지식기반 검색기의 상세 구조도 이다.



[그림 2] 지식기반 검색기 상세구조

4. 시스템 구현

본 '지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템'은 팬티엄III 듀얼550Mhz 환경에서 서버 프로그램은 JAVA 1.2로 구현되었으며 웹 서비스 프로그램은 ASP(active server page)를 기반으로 구현 되었다. DB는 오라클(oracle)을 사용하였다. 서버 프로그램은 뉴스의 수집과 검색, 지식기반 문서검색 작업을 수행하며 웹 프로그램은 사용자의 검색신청과 검색 결과를 표시하는데 사용 되었다. 다음 [그림 3]은 본 시스템의 사용자 뉴스 정보 검색의 결과 화면이다.



[그림 3] 사용자 뉴스 정보 검색 결과 화면

5. 결론 및 향후 연구

본 '지식기반 방식을 이용한 웹 뉴스문서 검색 에이전트 시스템'은 기존 키워드 위주의 문서검색 시스템의 정확도를 향상시키고 사용자가 원하는 문서를 정확히 검색하기 위하여 문서검색 범위를 인터넷 뉴스문서로 제한하는 대신에 지식기반 방식을 이용하였다. 이는 키워드에 대한 확장 지식을 시스템이 미리 제공하여 줌으로써 해당 키워드의 뉴스분야에 대한 검색에서는 만족할만한 결과를 얻을 수 있었다. 이후 지식기반 방식을 이용하여 뉴스 분야의 문서만이 아니라 더욱 확장된 범위를 가지는 지식기반 검색 에이전트 시스템으로의 확장성에 대한 연구가 지속 되어야 할 것이다.

6. 참고 문헌

[1] Goldszmidt, M., and Sahami, M. 1998. *A Probabilistic Approach to Full-Text Document Clustering*. Technical Report ITAD-433-MS-98-044, SRI International.

[2] Oren Zamir, Oren Etzioni, Omid Madani and Richard M. Karp. *Fast and Intuitive Clustering of Web Documents*. KDD'97.

[3] Mladenic, D., (1999) *Text-learning and related intelligent agents* (Revised version in IEEE EXPERT, Special Issue on Applications of Intelligent Information Retrieval), July-August 1999.

[4] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery. *Learning to Construct Knowledge Bases from the World Wide Web*. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/

[5] Marko Balabanovic and Yoav Shoham, *Learning Information Retrieval Agents: Experiments with Automated Web Browsing*, AAAI Spring Symposium on Information Gathering, Stanford, CA, March 1995.