키워드 관련도를 이용한 뉴스기사의 연관검색 기법

김지혜○
한성대학교
컴퓨터공학과
id_mana@hanmail.net

장재영
한성대학교
컴퓨터공학과
jychang@hansung.ac.kr

윤홍준
서울시립대학교
전자전기컴퓨터공학부
flow@uos.ac.kr

김한준
서울시립대학교
전자전기컴퓨터공학부
khj@uos.ac.kr

A Relationship Search in News Articles Using a Keyword Association Frequency

Ji-hye Kim○
Department of Computer Engineering,
Hansung University, Seoul, Korea

Jae-Young Jang
Department of Computer Engineering,
Hansung University, Seoul, Korea

Han-joon Kim
Department of Electrical and Computer Engineering,
University of Seoul, Korea

Hongjune Yune○
Department of Electrical and Computer Engineering,
University of Seoul, Korea

요 약

현재 많은 포털 사이트에서는 인기가 있거나 중요도가 높은 키워드에 대해 정보를 제공해주는 태그 클라우드나 연관 검색어 등의 기능이 제공되고 있다. 하지만 대부분의 뉴스기사 페이지들은 날짜와 분야별로 기사들이 나열되어 있으며 사용자는 카테고리별로 나누어진 기사를 읽을 수만 있을 뿐 그 기사와 연관된 다른 기사의 정보에 대해서 한눈에 알아 볼 수 있는 방법은 미흡한 실정이다. 또한 연관 검색어 서비스도 사용자가 검색한 입력 내용을 기반으로 연관성 정도를 분석하여 객관성을 보장하지 못하고 있다. 본 논문에서는 기존의 태그 클라우드 방식에서 좀 더 나아가 축적된 뉴스 기사로 부터 검색 키워드와 밀접히 연관된 키워드를 추출하여 제공하는 기사 검색 시스템을 소개한다. 이 시스템은 사용자가 기사 검색을 하였을 때, 키워드와 가장 밀접한 기사를 검색해 주는 것뿐만 아니라 검색어와 관련된 연관 키워드들을 보여주고 연관된 키워드간의 관계성을 보여줌으로써 뉴스 기사들 속에 숨겨진 연관정보의 탐색을 가능하게 한다.

1. 서 론

기존의 포털 사이트에 제공되는 블로그(blog)나 검색(search)으로 대표되는 인터넷 환경을 웹 1.0으로 본다면 개방적인 웹 환경을 기반으로 네티즌들의 정보공유와 참여가 가능해진 현재 인터넷 환경을 웹 2.0이라고 부르고 있다. 웹 2.0 시대가 도래하면서 사용자들의 편의성을 추구하는 다양한 서비스 기술들이 등장하고 있다. 대표적으로 사이트에 새롭게 올라 온 글을 사용자가 원하는 정보만 볼 수 있도록 서비스화 되어있는 RSS(Really Simple Syndication)[1]나 사용자가 마음대로 지정해 놓은 단어가 중요한 정보로 검색 될 수 있는 태그(tag) 등을 예로 들 수 있다. 이러한 기술의 발달로 인해 몇몇 포털 사이트의 뉴스 기사페이지는 인기가 있거나 중요도가 높은 내용들을 사용자가 질문하기 전에 미리 보여주고 정보의 동향을 제공해주는 태그 클라우드(tag cloud) 기능을 제공하고 있다[2]. 하지만 아직도 대부분의 뉴스 기사 페이지들은 날짜와 분야별로 기사들이 나열되어 있으며 사용자는 카테고리별로 나누어진 기사를 읽을 수만 있을 뿐 그 기사와 연관된 다른 기사의 정보에 대해서 한눈에 알아 볼 수 있는 방법은 미흡한 실정이다.

본 논문에서는 기존의 태그 클라우드 방식에서 좀 더 나아가 축적된 뉴스 기사로 부터 검색 키워드와 밀접히 연관된 키워드를 추출하여 제공하는 검색 시스템을 소개한다. 본 시스템의 핵심 요소인 연관검색 방식은 사용자가 기사 검색을 하였을 때, 키워드와 가장 밀접한 기사를 검색해 줄 뿐만 아니라, 키워드와 연관된 키워드들과 연관 정도를 보여준다. 이는 사용자가 검색한 기사의 내용을 파악할 뿐만 아니라 연관된 키워드에 대한 정보를 얻게 된다. 나아가 연관 기사들끼리의 관계성을 확인함으로서 자기가 원하는 정보이외에 관심이 없었던 기사의 연관까지 확인하여 정보 습득에 대한 시야를 더욱 넓혀 줄 수 있다. 이 시스템은 또한 최신 기사와 검색 내용의 월별 추이 그래프를 제공함으로써 검색 키워드의 상세 정보를 제공하는 기능을 한다.

기존의 일부 포털 사이트에서도 본 시스템의 기능과 유사한 연관 검색 기능을 제공하고 있다. 그러나 포털 사이트에서 제공되는 연관검색 기능은 사용자들이 입력한 키워드 간에 유사성이나 동일한 사용자에 의해 연속적으로 입력된 키워드들을 분석함으로써 키워드들 간의 연관성을 부여하고 있다. 반면에 본 논문에서 소개하는 검색 시스템에서는 텍스트 형식의 뉴스 기사들을 분석하여 기사 내에 키워드들의 연관성을 분석하여 제공함으로써 객관성을 보장하고 있다.

본 논문에서 개발한 검색 시스템은 포털사이트에 축적된 뉴스 정보를 웹 크롤링(web crawling) 방식으로 수집한 후, 다양한 검색 기술들을 활용하여 키워드간의 연관성을 분석하였다. 키워드의 연관 정도를 분석하기 위해 본 논문에서는 통계에 기반한 키워드 관련도(keyword association)를 정의하였다. 키워드 관련도란 하나의 문서 혹은 문서 집합에 대해서 주어진 두 개의 키워드가 어느 정도의 연관성을 갖는가를 판단하기 위해 본 논문이 제안한 방법으로써, 기존 문서 검색의 대표적인 방법인 TF-IDF 방식[3][4]을 변형하여 고안하였다.

분석된 연관정보는 관계도(relation chart)를 통하여

서비스되는데, 이 기능은 기존에 알고 있던 관계 외에 새로운 연관관계를 확인함으로써 의외의 연관성을 발견할 수 있도록 도와준다. 추가적으로 연관 기사를 통해 키워드 간의 어떤 이유로 연관이 있는지 관련 원인도 분석이 가능하다. 그리고 월별 추이 그래프와 기사를 통해 기간별 키워드의 중요도와 인용 정도에 대한 분석 기능도 제공된다. 본 시스템은 현재까지 연예인, 정치인, 스포츠인 등 사회적으로 알려짐 유명 인사를 대상으로 구축하였으며, 향후 그 대상을 인물 이외의 사회적 이슈가 되는 키워드들로 확대해나갈 계획이다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템의 구성과 주요 모듈에 대해 설명하고, 3장에서는 연관검색 기법을 소개한다. 4장에서는 개발결과와 그에 대한 평가를 하고, 마지막으로 5장에서는 결론을 맺는다.

2. 시스템 구성

본 논문에서 개발한 시스템의 전체적인 구성도는 그림 1과 같다. 이 시스템은 크게 관리자와 사용자 부분으로 나누어진다. 우선 관리자 측면에서는 임의의 시간이나 주기적으로 업데이트 모듈을 이용하여 최신 뉴스기사를 수집한다. 업데이트 모듈은 인터넷에서 뉴스 기사를 수집하여 키워드를 추출하고 키워드들 간의 연관정보를 분석하여 그 결과를 데이터베이스에 저장한다. 이러한 일련의 작업을 수행하기 위해서는 웹 크롤링과 데이터 마이닝 그리고 형태소 분석기 등의 기술이 요구된다. 다음으로 사용자 측면에서는 검색하고자하는 키워드가 입력되면 데이터베이스에 저장된 키워드에 해당하는 뉴스기사들과 연관 키워드 정보 등 관련 자료를 추출하여 관계도를 비롯한 여러 가지 방식을 통해 사용자에게 제공된다.

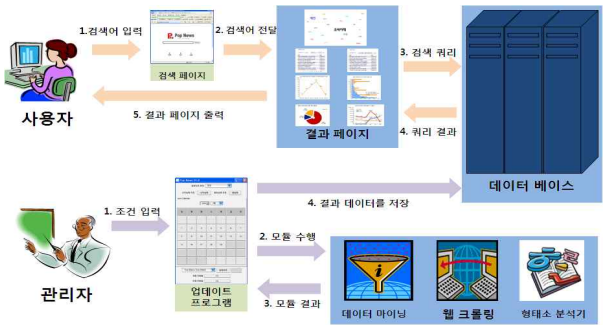


그림 1. 시스템 구성도

그림 2는 관리자 모듈의 상세한 구조를 보여준다. 첫 단계로 업데이트 프로그램에서 임의의 기간 혹은 최신의 기사를 수집하는 명령을 내리면 웹 크롤링 기법을 이용하여 지정된 URL을 통해 기사를 수집하게 된다. 수집된 기사는 HTML형식으로 구성되어 있으므로 태그 등을 삭제하여 기사 제목과 본문을 텍스트 형식으로 변환하고 데이터베이스에 저장한다. 다음 단계로 수집된 각 기사에 대해서 형태소 분석을 거쳐 주요 키워드를 추출하게 한다. 추출된 키워드는 데이터 마이닝과 정보검색 기법을 통해 기사 내에서 키워드의 중요도나 키워드들 간의 연관성 정보를 분석하여 그 결과를 다시 데이터베이스에 저장하게 된다.

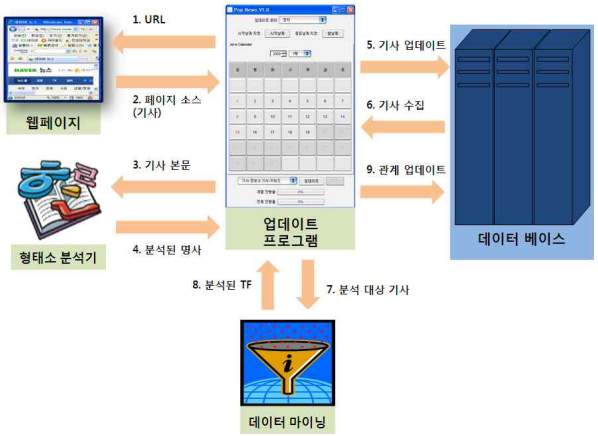


그림 2. 관리자 서브시스템 구조

이와 같은 연관성 분석 작업을 통해 사용자에게 제공되는 구체적인 기능은 표 1과 같다. 이 중에서 본 논문에서 가장 중점을 두고 개발한 것은 키워드 간의 연관성 분석이다. 분석 결과는 연관성을 표현하는 관계도와 연관 분석 그래프를 통해 보여준다. 관계도란 입력된 키워드와 다른 키워드들 간의 연관성 정도가 나타내는 차트를 의미하며, 분석 그래프는 연관성을 가진 키워드들의 연관정도와 해당 키워드들의 중요도를 나타내게 된다. 또한 키워드의 기간별 추이 그래프는 과거로부터 현재까지 해당 키워드들이 뉴스에서 어느 정도 중요성을 갖고 다뤄졌으며, 그 중요도가 기간에 따라 어떻게 변화하였는가를 나타낸다. 마지막으로 키워드와 가장 관련성이 높은 기사의 원문을 직접 확인할 수 있으며, 다른 키워드들과의 연관성에 가장 영향력을 미친 기사의 원문도 확인할 수 있다.

3. 연관성 분석 기법

본 장에서는 두 키워드간의 관련성의 정도를 정량화하기 위한 방안을 기술한다. 자연어처리(natural language processing)의 구문 분석 및 의미 분석 과정을 통해 명확하게 고유명사간의 관련성을 측정하는 방법이 있을 수 있으나, 자연어처리의 정확도 문제와 시간복잡도 문제를 감안하여 본 논문에서는 통계기반 방안을 제안한다.

우선, 제안 기법의 아이디어를 도출하기 위해 정보검색에서 흔히 사용하는 TF-IDF 가중치 모델[3][4]을 분석해보자. 기본적으로 TF-IDF 가중치는 한 문서 내에서 한 단어의 가중치를 결정할 수 있는 모델로서, 이는 TF(term frequency)와 IDF(inverse document frequency)값을 곱한 것이다. 여기서 TF값은 한 문서 내에서의 특정 단어가 출현하는 횟수 (또는 이를 정규화한 값)를 의미하며 이는 한 문서 내에서의 그 단어의 중요도를 반영한다. 그리고 IDF값은 특정 단어가 출현하는 문서의 수(DF, document frequency)의 역수(또는 이를 정규화한 값)를 의미하며, 이는 전체문서집합에서 단어

의 중요도를 반영한다. DF값이 큰 단어는 많은 문서에서 는 것은 그 관련성의 정도가 크다고 판단하는 것이 합리

표 1. 시스템 기능

시스템 주요기능	
키워드와 연관 키워드 사이의 관계도	뉴스기사 키워드와 관련된 기사 중 연관성이 높은 연관 키워드와의 관계를 보여준다. 또한 연관 키워드 사이의 관계성도 보여준다.
뉴스기사 키워드와 연관 키워드 사이의 분석 그래프	뉴스기사 키워드와 연관성을 가진 기사들 사이의 기사 수를 그래프로 보여준다.
뉴스기사 키워드에 대한 기간별 추이 그래프	뉴스기사 키워드에 대해 현재를 기준으로 일정기간동안의 기사에서 언급되는 빈도수의 시간대별 추이를 보여준다.
뉴스기사 키워드의 기사 및 연관 키워드의 관련 기사	뉴스기사 키워드에 관련된 기사를 정확도 및 기간별로 보여준다. 기사의 제목을 선택하면 기사의 원문내용을 확인할 수 있다.

공통적으로 사용되는 단어이기 때문에 TF값이 크다 할지라도 그 중요도를 낮춰져야 한다. IDF값은 TF값의 오류를 보정한다는 측면에서 중요한 작용을 하게 된다.

본 논문에서는 두 키워드간의 관련도를 측정하기 위해서 TF-IDF 가중치 모델과 유사한 접근방식을 취한다.

첫째, TF 측면에서 문서 내에서 두 키워드간의 관련도(Association Frequency, AF)를 정의한다. AF는 TF와 유사하게 한 문서 내부에 존재하는 두 개체간의 횟수를 정규화한 값이다. 이는 기본적으로 두 개체간의 관련성은 한 문장 내에서 두 개체가 공존하는 횟수로부터 도출될 수 있다.

다음은 문서 d_x내의 키워드 e_i와 e_j간의 관련도 $AF_{d_x}(e_i,e_j)$를 정의한 것이다.

$$AF_{d_x}(e_i,e_j)=\sum_{s\in d_x} SentenceAssoc_{d_x,s_p}(e_i,e_j) \qquad (1)$$

$SentenceAssoc_{d_x,s_p}(e_i,e_j)$는 문서 d_x의 각 문장 s_p에 존재하는 키워드 e_i와 e_j의 관련도를 측정한 값으로서, 이는 한 문장에 존재하는 키워드들 간의 모든 조합 수의 역수로 정의한다. 즉, 한 문장에서 다른 키워드들이 포함된 경우에 e_i와 e_j간의 관련도가 약화됨을 반영한 것이다. 예를 들어, 그림 3의 첫 문장에 존재하는 키워드의 수는 3개이므로, $SentenceAssoc_{d_x,s_p}(e_i,e_j)$값은 $1/_3C_2$ = 1/3이 된다. 이러한 SentenceAssoc값을 모두 더하여 한 문서 내에서의 Association Frequency (AF) 값으로 측정한다. 한 문장에서 e_i 또는 e_j만이 출현하는 경우에는 합산하지 않는다.

둘째, IDF측면에서, 전체 문서 집합 속에서 두 키워드간의 관련성을 정의해보자. TF-IDF 가중치 모델에서 IDF 인자는 전체문서집합에 퍼져 있는 단어에 대한 가중치를 줄이기 위한 용도로 사용한다. 이에 반해서, 두 키워드간의 관계에 대한 정보가 전체 문서집합에 퍼져 있적이다. 그래서 본 연구에서는 DF 개념을 그대로 반영하여 키워드간 관련도를 계산한다.

결론적으로 위 두 가지 사항을 고려하여 모든 문서 내에서 두 키워드 e_i와 e_j의 관련도 $Assoc(e_i,e_j)$는 다음과 같다.

$$Assoc(e_i,e_j)=AF(e_i,e_j)*\{1+\log DF(e_i,e_j)\} \qquad (2)$$

여기서, $AF(e_i,e_j)$는 각 문서에 대해 e_i와 e_j의 관련도를 모두 더한 값이며, $DF(e_i,e_j)$는 e_i와 e_j가 공존하는 문장이 존재하는 문서의 개수를 의미한다. 이 때 $DF(e_i,e_j)$의 값이 커지는 경우에 AF값의 신뢰성을 떨어뜨릴 수 있으므로, DF값에 log를 취하여 DF값의 증감에 따른 영향을 줄인다. DF를 고려한 이유는 e_i와 e_j가 하나의 문서에 집중적으로 나타남으로써 AF값을 증가시키는 것보다는 여러 문서에 걸쳐 동시에 출현하는 것이 관련성이 더 높은 것으로 간주하기 위해 설정한 요소이다.

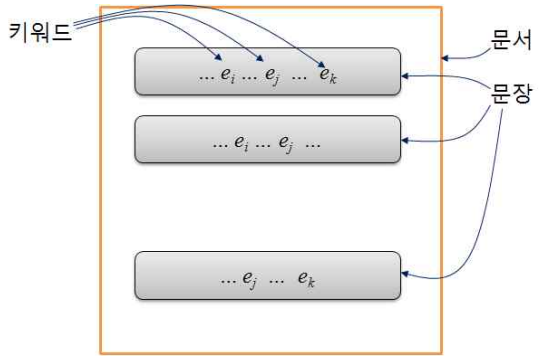


그림 3 문서, 문장, 키워드의 개념도

추가적으로 e_i와 e_j의 관련도 $Assoc(e_i,e_j)$는 다음의 요소들을 고려할 수 있다. 우선 뉴스 기사는 대부분 기사 제목을 포함하고 있다. 따라서 제목에 키워드가 나타

나는 경우는 키워드의 중요도가 당연히 높아짐으로 이에 대한 가중치를 부여할 수 있다. 또한 뉴스 기사 내에서 문서의 앞부분에 나타나는 키워드일수록 해당 기사와 관련성이 크다고 볼 수 있으므로 키워드들이 나타나는 문장이 문서의 어느 위치에 존재하는 가에 대해 가중치를 부과할 수도 있다. 이러한 요소들은 기존의 대부분의 검색 시스템에서도 적용하는 방식으로 본 논문에서는 구체적인 방안에 대해서는 더 이상 논의하지 않는다.

4. 개발 결과

본 논문에서 개발한 시스템은 Java와 JSP를 이용하여 개발되었으며, 데이터베이스는 MySQL 5.0을 이용하였다. 사용자 인터페이스의 그래픽을 보여주기 위해 Flex와 Swiff Chart 3.0을 이용하였다. 또한 한글 형태소 분석은 KLT[5]를 활용하였다. 기사는 네이버를 대상으로 수집하였으며, 현재까지 총 24만 여 건의 기사를 수집하였다. 종류별로는 연예 분야가 17만여 건, 정치 분야 4만여 건 그리고 스포츠 분야 3만여 건 등이다. 서론에서 기술한 바와 같이 본 시스템은 우선적으로 인물에 대한 연관성 분석을 실시하였으며 이를 위해 다음(Daum) 포털 사이트[6]의 인물 데이터베이스를 활용하였다. 여기서 연예인 5,576명, 정치인 1,787명, 스포츠인 6,759명을 대상으로 자체적인 데이터베이스를 구축하였다.

그림 4는 키워드 간의 연관성을 보여주는 관계도 화면이다. 관계도의 중앙에 사용자가 입력한 키워드를 중심으로 최대 20개의 연관 키워드들로 구성되어있다. 연관 키워드의 폰트 크기는 연관 정도를 나타내며 크기가 클수록 그 정도가 높다는 것을 의미한다. 또한 연관 키워드들 간의 연관성도 확인할 수 있는데, 특정 연관 키워드에 마우스를 올리면 그 키워드와 연관이 있는 키워드들만 남고 연관이 없는 키워드들은 사라지게 된다.

그림 4. 관계도 출력 결과

이 그림의 예는 “설경구”라는 연예인 이름으로 검색한 결과로, “설경구”와 가장 연관성이 높은 인물은 “김제동”이며, 그 이외에 “송윤아”, “박지성” 등도 비교적 연관성이 높은 것을 알 수 있다. 이 그림에서 각 키워드들을 클릭하면 해당 키워드들에 대한 관계도 화면으로 바로 전환하게 된다. 예를 들어 “김제동”를 클릭하면 “김제동”에 관한 관계도 화면으로 전환 된다. 그러나 “김제동”에 관한 관계도에서는 “설경구”가 가장 연관이 높게 나타나지 않을 수도 있다. “김제동”과 가장 연관도가 높은 다른 인물이 있을 수 있기 때문이다.

다음으로 그림 5는 검색 키워드와 연관된 키워드들이 어느 정도 연관되었으며, 또한 해당 키워드가 전체 연관 키워드 중에 어느 정도 영향력이 있는가를 나타낸다. 예를 들어 그림 5에서 “설경구”와 가장 연관성 높은 "김제동"의 경우는 30%로 나타났는데, 이 수치는 “설경구”와 연관된 전체 키워드 중에 “김제동”과의 연관성이 30%를 차지한다는 것을 의미한다. 또한 막대그래프의 오른편에 있는 부분은 연관 키워드 중에서 “김제동”의 객관적이 중요도를 상대적으로 보여주는 수치이다. 따라서 이 결과를 볼 때, “김제동”은 설경구와 가장 연관성이 높은 인물인 반면에, “동방신기”는 “설경구“와의 연관성이 크지는 않지만 객관적인 영향력은 매우 높다는 것을 알 수 있다.

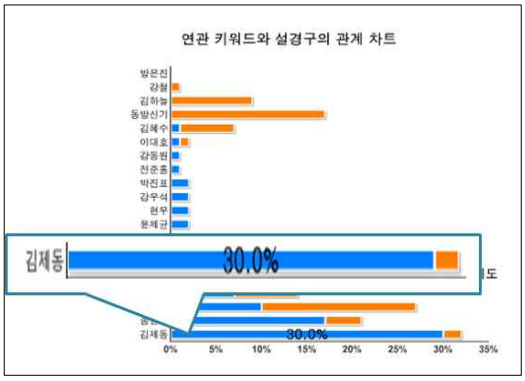


그림 5. 연관성 차트

그림 6은 기간별 검색 키워드의 뉴스 기사 통계를 보여준다. “설경구”의 경우 2009년 5월에 기사 건수가 급격히 증가된 것을 확인할 수 있다. 이 결과로부터 “설경구”는 이 기간에 다른 시기에 비해 뉴스에 자주 인용된다는 사실을 알 수 있다.

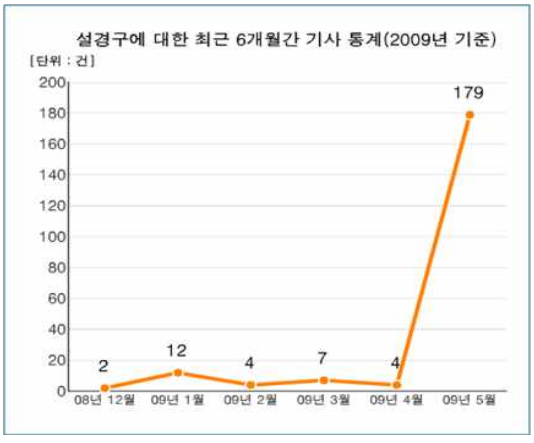
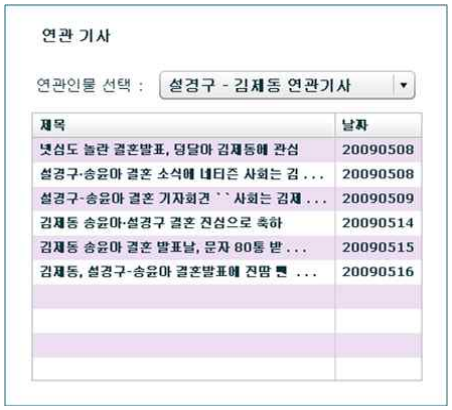


그림 6. 기간별 키워드의 기사 빈도 추이

마지막으로 그림 7의 (a)는 검색 키워드에 해당하는 기사목록을 관련성 정도에 따라 정렬하여 보여준다. 각

기사 제목을 클릭하면 해당 뉴스기사 홈페이지를 직접 확인할 수 있다. (b)는 키워드와 연관 키워드의 관련 기사 목록을 보여주고 있다. 이 리스트도 역시 기사의 연관성 정도에 따라 정렬하여 보여주고 있다.

(a)



(b)

그림 7. 기사원문제공 서비스

5. 결 론

본 논문에서는 뉴스 기사로 부터 검색 키워드와 연관성 정도를 분석하여 정보를 제공하는 검색 시스템을 소개하였다. 연관검색 방식은 사용자가 검색한 기사의 내용을 파악할 뿐만 아니라 연관된 키워드에 대한 정보를 얻음으로써 키워드들 간의 상관관계를 파악하는데 많은 정보를 제공하게 된다. 본 시스템은 현재까지 연예인, 정치인, 스포츠인 등 사회적으로 알려짐 유명 인사를 대상으로 구축하였으며, 향후 그 대상을 인물 이외의 사회적 이슈가 되는 키워드들로 확대하여 [7]과 같은 종합적인 연관분석 시스템으로 확대할 계획이다.

참고문헌

[1] D. Ayers and A. Watt, Beginning Rss And Atom Programming, John Wiley & Sons Inc., 2005
[2] 이강표, 김두남, 김형주, “웹 2.0 환경에서의 태깅기술 동향”, 정보과학회지, 제25권 10호 pp. 36-42, 2007년 10월
[3] http://lucene.apache.org/nutch/
[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques 2nd ed., Morgan Kaudman, 2006
[5] 강승식, 한국어 형태소 분석과 정보 검색, 홍릉과학출판사, 2003
[6] http://www.daum.net
[7] http://www.textmap.com