전문가 검색 엔진에서 데이터 마이닝을 이용한 개념 관계 추출

이권국[O]*, 신일수*,이상준**, 김기태*
*중앙대학교 컴퓨터공학과
**제주 국립 대학교 통신,컴퓨터 공학부

Extraction of conceptual relation using Data Mining in expert search engine

Kwon-Kook Lee[O]*, Il-Su Shin*, Sang-Jun Lee**, Ki-Tae Kim*
*Dept. of Computer Science & Engineering, Chung-Ang University
**Faculty of Telecommunication & Computer Engineering, Che-ju National University

요 약

전문가 검색 엔진은 전문가 시스템과 같은 목적에서 특정 전문 분야에 대한 특별한 정보를 모아 특정 정보를 검색하기 위한 엔진이다. 인터넷의 홈페이지는 서로를 연결하는데 하이퍼링크(hyperlink)를 사용하는데 이런 하이퍼링크(hyperlink)에 있는 정보를 이용하여 홈페이지와 홈페이지 사이의 연결관계를 모은 결과를 전문가 검색 엔진에서 모은 키워드와 웹 사이트를 이용하여 각 키워드 간의 관련성을 데이터 마이닝 기법을 사용해서 각 키워드나 웹 페이지간의 상관관계에 대한 개념을 추출한다. 본 논문에서는 이런 홈페이지 간의 추출을 이용한 시스템 설계와 구현 결과를 보여준다.

1. 서 론

인터넷은 정보의 바다라고 할 만큼 많은 정보들로 산재해 있다. 또한 최근에 인터넷에 대한 관심이 폭발적으로 증가하고 있고 그에 맞춰서 수많은 사이트나 웹 페이지들이 새롭게 생겨나고 있다. 이런 수많은 정보들을 효과적으로 이용하고자 Yahoo, Lycos, Altavista같은 여러 검색 사이트 들도 등장했다. 이런 검색 사이트들은 수동으로 하는 디렉토리 구조나 단순 단어 검색으로 사용자의 질의에 대한 결과를 찾기 때문에 정확성, 특히 질의와 관련성이 많은 결과를 찾기가 힘들었다. 이런 문제점을 해결하기 위해 본 논문에서는 데이터 마이닝 기법을 사용하여 웹 페이지나 키워드 간의 관련성을 계산하여 사용자의 질의에 보다 정확하고 관련된 정보를 제공하는 개념 관계를 추출하려 한다. 2장에서는 본 논문의 기반연구로 기존 검색 사이트의 질의 처리 방법과 데이터 마이닝 기법중 하나인 Market Basket Analysis에 대해 알아보고 3장에서는 본 논문에서 제안하는 전문가 검색 엔진 시스템이 어떻게 되어있는지 구성을 살펴본다. 마지막으로 4장에서는 결론과 향후 과제를 논의한다.

2. 기반연구

2.1 기존 검색 사이트의 질의 처리 방법

기존의 검색 사이트들은 크게 2가지 방법으로 질의를 처리한다. 1세대 검색 사이트중 대표적인 Yahoo에서는 디렉토리 구조로 인터넷에 있는 웹사이트를 사람의 손으로 직접 나누어 각각의 사이트를 정리 보관한다. 사용자의 질의가 들어왔을 때 Yahoo에서는 그 질의어를 받아서 자신이 저장하고 있는 웹 디렉토리 내에서 찾아서 있을 경우 그 디렉토리와 웹 페이지를 같이 보여준다. 이 방법은 디렉토리로 나누어져 있어서 사용자가 질의한 것과 관련된 것들이 같은 디렉토리 안에 있을 경우가 많기 때문에 관련된 자료를 모으는데는 편리하나 인터넷에 있는 웹사이트들을 사람의 손으로 직접 디렉토리 구조로 만들기 때문에 새로운 웹사이트들의 반영이 늦고 많은 웹사이트들을 모을 수 없다는 단점으로 정확성이 떨어지게 된다.

2세대로 불리는 검색 사이트 중 대표적인 Altavista는 사용자가 질의어를 넣으면 제목이나 본문에 같은 단어를 찾아서 그 단어가 있을 경우 결과로 보여주게 된다. Yahoo같은 수동으로 하는 검색이 아니기 때문에 비교적 가지고 있는 웹사이트가 많고 기계적으로 질의에 따른 결과를 내므로 검색 속도도 빠르고 그 결과도 많다. 하지만 단순히 질의한 단어와 같은 단어가 있다는 것으로 맞는 결과로 계산하고 그 결과를 내기 때문에 그 자료가 정말 찾고 싶은 자료인지 정확성이 불투명하고 자신이 원하는 결과가 맞다고 하더라도 그 자료와 관련된 다른 사이트를 찾는 것은 거의 불가능하다.

2.2 Market Basket Analysis

Market Basket Analysis(이하 MBA)는 데이터 마이닝 기법의 한 방법이다. 데이터 마이닝이란 수집한 데이터(data)를 가지고 그 데이터(data)안에서 유용한 패턴이나 규칙을 찾아서 분석하는 것이다.

MBA는 가게를 찾은 손님들이 누가 어떤 물건을 왜 사는가를 분석한다. 예를 들어 한 손님이 오렌지 쥬스, 소다수, 표백제를 샀을 때 오렌지 쥬스와 소다수 혹은 표백제와 무슨 관계가 있는지는 분석, 연구한다.

3. 웹(Web)의 정보 관계 추출

[그림 1]는 웹 페이지의 링크에 저장된 정보간의 관계를 찾는 과정을 보여주고 있다. Seed라고 불리는 시작 웹 페이지를 중심으로 하여 넓이 탐색 기법을 사용한다. 링크하는 웹 페이지를 설명하는 앵커

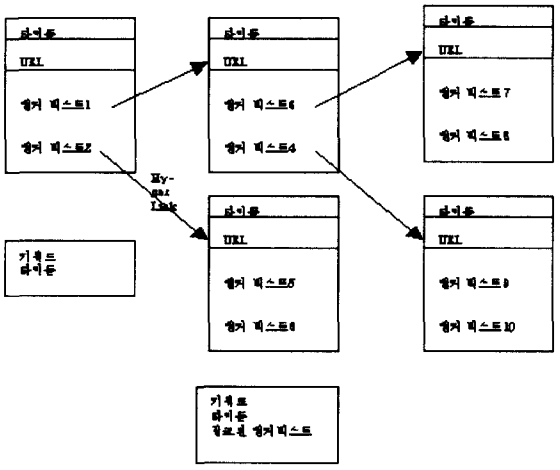


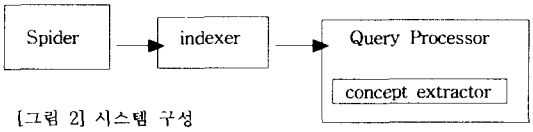
그림 1 정보 관계 추출 과정

텍스트는 대부분 링크하는 웹 페이지를 요약하여 관련 있는 단어를 사용하는 것이 대부분이므로 앵커 텍스트와 링크된 웹 페이지가 관련성이 있다. 그리고 링크를 할 때는 앵커 텍스트를 가지고 있는 웹 페이지와 앵커 텍스트가 가리키고 있는 웹 페이지도 관련이 있기 때문에 앵커 텍스트를 키워드로 하여 서로 관련됨을 나타낼 수 있다.

4. 전문가 검색 엔진 시스템 구성

전문가 검색 엔진은 전문가 시스템과 같은 목적에서 특정 전문 분야에 대한 특별한 정보를 모아 특정 정보를 검색하기 위한 엔진이다.

전문가 검색 엔진 시스템은 [그림 2]과 같다.



[그림 2] 시스템 구성

전체 시스템은 크게 Off-line Batch Job과 On-line Processing 부분으로 구분한다. Off-line Batch Job 부분은 웹문서를 모아서 인덱스를 구축하는 부분으로 최종적으로 인덱스 파일 시스템의 구축을 목표로 한다. On-line Processing 부분은 사용자의 질의를 받아 인덱스 파일 시스템을 이용하여 개념 그래프의 결과를 보여주는 기능을 담당한다.

전체적인 동작은 Off-line으로 웹문서를 수집하여 데이터베이스에 저장한 후, 색인기를 통하여 인덱스 파일을 만든다. 검색 시스템이 웹 서버를 통하여 질의를 받으면, 질의처리기가 인덱스 엔진을 통하여 인덱스 파일을 접근하여 개념그래프를 작성하고, 이를 사용자에게 검색 결과로 제공한다.

웹문서 수집기

수집기(spider)의 역할은 웹문서를 가져와서 데이터베이스에 저장한다. 저장되는 정보로는 웹문서의 URL, Title, 샘플 텍스트(255자), 하이퍼링크 연결 관계를 저장한다.

인덱서부

웹문서와 하이퍼링크 정보로 구성되어진 데이터베이스로부터 실제 검색서비스를 위하여 인덱스 파일 시스템을 구축하는 프로그램이다.

질의처리부

질의어를 이용하여 인덱서를 찾고, 찾아진 웹문서의 핵심어와 하이퍼링크의 정보를 이용하여 개념을 생성한다. 개념을 생성할 때는 데이터 마이닝 기법중 하나인 Market Basket Analysis의 요소중에서 Support, Confidence, Improvement를 사용해서 생성하는데 3가지 각각의 식은 다음과 같다.

식 1) $$support = \frac{Number\ of\ Somethings}{Number\ of\ Total}$$

식 2) $$confidence = \frac{Number\ of\ something\ B}{Number\ of\ Condition}$$

식 3) $$improvement = \frac{p(condition\ and\ result)}{p(condition)p(result)}$$

support는 식 1)에서와 같이 한 집합의 전체에서 하나 혹은 여러 개가 전체의 몇 %를 차지하고 있는가를 나타낸다. confidence는 식 2)에서와 같이 어떤 조건의 총 개수를 분모로 두고 어떠한 것이 조건의 얼마나 많은 비율을 차지고 하고 있는가를 나타낸다. improvement는 조건과 결과를 보고 계산된 확률이 실제 상황에서 무작위로 선택될 때와 비교하여 얼마나 정확한가를 보여준다.

사용자의 질의가 들어오면 그 질의어는 다음의 과정을 거치게 된다.

사용자가 질의를 할 경우 Query processor에 있는 concept extractor에서 사용자의 질의에 대한 관련어들의 support와 confidence, improvement를 각각 구한 후 [그림 2]에서와 같은 결과를 보여주고 있다. 중심에 있는 원은 사용자가 질의한 단어를 나타내고, 주변에 원형을 이루고 있는 원은 데이터 마이닝 기법으로 찾은 사용자의 질의 단어에 상관관계가 있는 단어들이다. 주변의 원들은 3시 방향부터 improvement가 높은 순서대로 시계 방향으로 위치

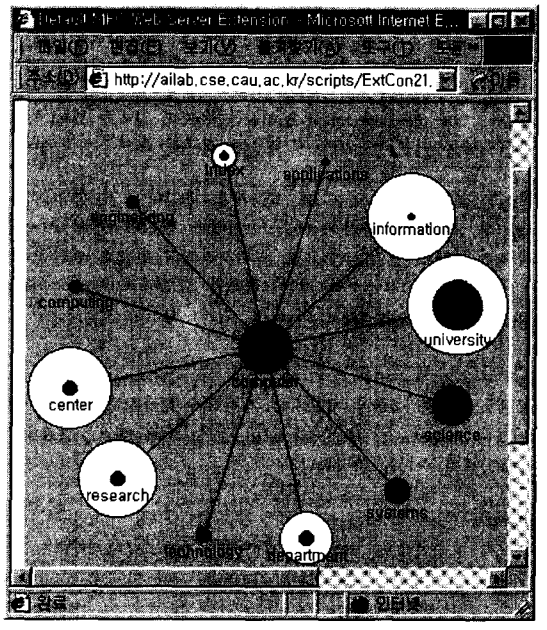


그림 3 데이터 마이닝 기법을 이용한 질의 검색 결과

하고 있다.

각각의 원에서 파란 색 원은 사용자의 질의어에 대한 관련어의 confidence를 나타내고 있다. 흰색 원은 관련어의 support를 나타내고 있다. 질의어와 관련어를 나타내는 원들을 연결하는 선은 두 단어가 서로 관련되어 있음을 나타낸다.

[그림 3]에 있는 각각의 원들을 더블 클릭했을 경우 원에 나타나고 있는 관련어에 관한 웹 사이트 주소 리스트 화면[그림 3]이 뜬다.

[그림 4]은 질의어는 Computer이고 관련어를 science를 선택했을 때 나오는 관련된 웹 페이지 리스트이다. 하이퍼 링크로 연결되어 있는 부분을 클릭하게 되면 해당하는 웹페이지를 찾아 갈 수 있다.

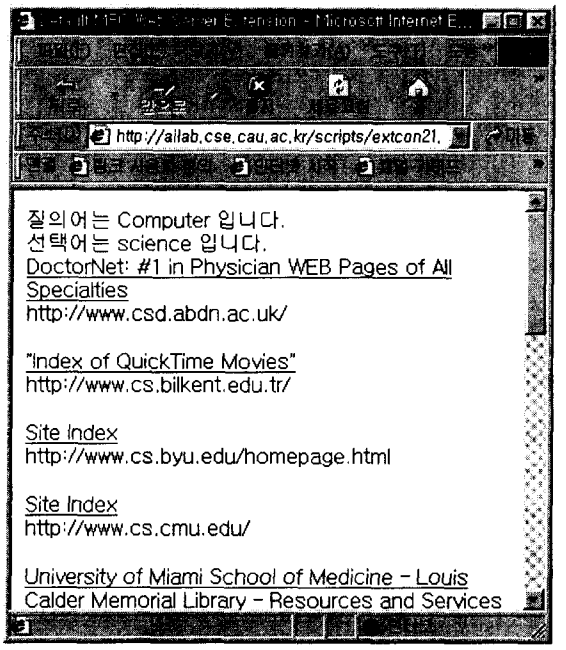


그림 4 결과 웹 페이지 리스트

4. 결론 및 향후 과제

본 논문에서는 사용자가 질의 할 때 보다 정확하고 관련된 자료를 쉽게 찾을 수 있는 방법으로 데이터 마이닝 기법을 사용한 개념 관계 추출 기법을 설계해 보았다. 기존의 수동적인 디렉토리 시스템이나 단순 단어별 검색 사이트와는 달리 각각의 단어로 검색을 하되 그 단어를 키워드로 하여 각 키워드의 관련성을 계산하여 보다 관련성있고 정확한 정보를 사용자에게 제공한다. 향후 과제로는 하이퍼링크(hyperlink) 정보를 사용하여 키워드 관계를 보다 명확하게 분류하고 특정 분야의 관계에 맞게 개념 그래프를 명확하게 할 필요가 있다.

참고문헌

[1] 김홍주, Robot agents and Search Engine, "http/::solgeo.dongguk.ac.kr/~k2/html/TS3-4.html"

[2] 조민재, 웹의 개념 지식을 이용한 자동 시소러스 생성법의 설계 및 구현. 1999.12.

[3] 최준영, 인터넷상의 하이퍼링크를 이용한 개념 그래프 기반 검색 시스템. 1998.12.

[4] Clare Cardie. Empirical Method in Information Extraction. AAAI Magazine. Winter. 1997

[5] Dayne Fritag. Information Extracting from HTML : Application of a General Machine Learning Approach. Information Extraction. AAAI. 1998

[6] Mark Craven, etc. Learning to Extract Symbolic Knowledge from the World Wide Web. Information Extraction. AAAI. 1998

[7] Quilan J.R. Learning logical definitions from relations. Machine Learning 5(3):239-266. 1990

[8] The Web Robots Pages,"http/::info.webcrawler.com/mak/projects/robots/robots.html"