연관 규칙 탐사 기법을 이용한 선박 부품 전문 검색 엔진의 설계 및 구현

하 창 승*, 윤 병 수**, 성 창 규**,김 종 화***, 류 길 수***

Design and Implementation of the Specialized Internet Search Engine for Ship's Parts Using Method of Mining for the Association Rule Discovery

Chang-Sung Ha, Kyung-Youl Jung, Byung-Soo Youn, Chang-Gyu, Sung, Keel-Soo Rhyu

* 동명대학 정보통신계열
** 한국해양대학교 대학원 컴퓨터공학과
*** 한국해양대학교 기계·정보공학부

Abstract: A specialized web search engine is an internet tool for detecting information in finite cyber world. It helps to retrieve necessary information in internet sites quickly.

In this paper, we design and implement a prototype search engine using method of mining for the association rule discovery. It consists of a search engine part and a search robot part. The search engine uses keyword method and is considered as various user oriented interface. The search robot fetches information related to ship parts in world wide web. The experiments show that our search engine(AISE) is superior to other search engines in collecting necessary informations.

Key words : Data mining, Specialized search engine, Association rule

1. 서 론

최근 인터넷 사용이 일반화됨에 따라 일반 사용자나 기업에서 제공하거나 요구하는 정보의 양은 크게 증가하며 특정영역의 정보를 찾아주는 검색엔진의 사용도 함께 늘어나고 있다. 하지만 사용자들은 자신과 관련된 특정 분야(application domain)의 전문 지식 및 정보를 집중적으로 필요로 하는데 비해 범용 검색엔진을 통한 정보의 획득에는 근본적인 문제점을 지니고 있다. 즉 기존의 검색엔진들은 웹 상의 페이지 문서들과 주어진 검색어(key word)를 패턴 비교하는 검색 기법을 사용하기 때문에 정보의 질적 효율이 낮으며 인터넷 상에서 정보를 획득하는 검색 로봇은 단순한 반복 동작을 통해 자료를 수집하기 때문에 네트워크의 트래픽(traffic)을 증가시키고 있다. 또한 주어진 검색어와 연관된 정보가 전문 분야별로 분류되지 않아 관련 없는 정보가 함께 처리됨으로서 응답시간을 크게 저해시키고 있다. 특히 문서 양의 급속한 증가와 문서 내용의 빈번한 변화를 신속히 반영하거나 특정 영역만을 고려하는 기능도

제대로 제공되고 있지 않다[1, 2]. 따라서 정보량의 급속한 증가에 따른 색인 데이터베이스 구축 시간이 커지며, 검색엔진에 대한 효율 및 신뢰도가 감소하는 문제점을 지니고 있다.

인터넷상의 검색엔진은 빠른 검색 속도와 많은 검색 자료의 제공뿐만 아니라 사용자의 검색 목적에 따른 특성화된 정보의 제공이 가능한 전문검색 서비스가 요구된다. 이러한 전문 검색 서비스 기능을 위해 전문화된 영역별로 분류된 정보를 검색할 수 있는 도메인 검색 엔진에 대한 연구가 필요하다[2].

이러한 전문화된 영역별 검색은 선진 외국에서도 하나의 보편화된 인터넷 검색 서비스로서 미국만 해도 1,800여 개의 전문 검색엔진이 사용되고 있는데, 뉴스의 헤드라인만을 검색해 주는 사이트, 연방법과 정부의 웹사이트만을 전문적으로 검색해주는 사이트, 과학 전문 검색 엔진 사이트 등으로 분야별로 특성화되어 가고 있는 추세이다[5].

이에 본 연구에서도 영역별 전문 지식의 한 분야로서 선박에서 사용되는 부품 및 장비와 관련된 전문 지식을 제공하는 지능화된 전문 도메인 검색 엔진을 구현하고자 한다. 본 검색 엔진은 검색방법에 있어 데이터마이닝(data mining)의 추론 기법 중 상품 및 서비스간의 관계를 이용한 연관 규칙 탐사기법을 이용하여 선박 장비와 부품간의 관계를 정의하였으며, 그 구현은 분산 인터넷 환경에 적합한 객체기반 언어인 자바를 사용하여 도메인 전문 검색 엔진을 설계하고 구축하였다.

2. 연관규칙탐사 검색엔진의 이론적 연구

본 연구에서 구현한 전문 검색엔진은 다음과 같은 세 가지 기술을 적용하였다. 즉 인터넷상의 사이트 정보들을 추출하여 색인 데이터베이스로 재구성하는 로봇 에이전트 기술과 색인 데이터베이스를 이용하여 사용자의 검색요구를 처리해 주는 검색 에이전트 기술 그리고 검색된 정보를 사용자의 검색요구에 대한 부합 정도를 측정하는 데이터마이닝 기술이다.

2.1 로봇 에이전트

검색엔진은 검색어를 통한 검색 방법을 제공하기 위해 로봇 에이전트를 이용하여 색인 데이터베이스를 구성한다. 로봇 에이전트는 웹 페이지의 하이퍼텍스트 구조를 추적하여 HTML 문서를 추출하고 다시 그 HTML 문서에서 참조되는 다른 HTML 문서들을 순환적으로 분석하여 필요한 정보를 추출하는 기능을 수행하는 프로그램이다.

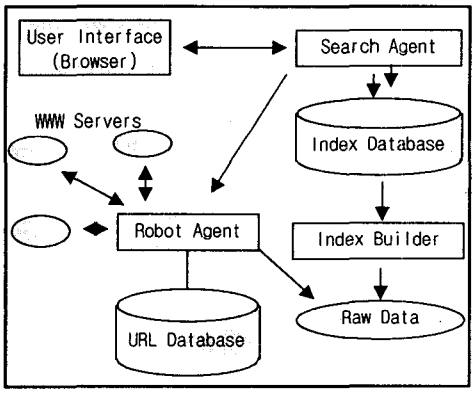

그림 1 로봇 에이전트의 내부 처리 구조

그림 1은 로봇 에이전트의 기능적 위치를 나타낸다. 여기서 로봇 에이전트는 URL 데이터베이스로부터의 사이트 정보를 참조하여 해당 웹 서버들을 탐색하여 원시자료(Raw Data)를 수집한다. 그리고 수집된 자료는 색인 구축기(Index Builder)에 전달되어 색인 데이터베이스 구축에 이용된다. 색인 데이터베이스 구축은 주제어 추출 과정을 통해 이루어지는데 주제어란 조사나 어미를 제외한 순수 명사만을 의미한다. 로봇 에이전트에서의 내용 검색은 형태소 분석

을 이용한 주제어 추출을 통해 수행되며 형태소 분석은 원시자료 작성자의 의도 및 의미에 부합하도록 띄어쓰기 단위를 기준으로 이루어졌다.

2.2 검색 에이전트

검색 에이전트는 크게 프리젠테이션 계층과 트리거 계층으로 구성된다. 프리젠테이션 계층은 사용자의 질의 검색어와 색인 구축기를 통해 구성된 색인 데이터베이스의 주제어를 비교하여 같으면 해당 정보를 HTML 문서 형태로 표현하여 사용자에게 보여준다. 그리고 트리거 계층은 로봇 에이전트와 연동하여 상호구동적(interactive)으로 동작하는 계층이다. 트리거 계층은 사용자의 질의 검색어가 색인 데이터베이스의 주제어로 존재하지 않을 경우 로봇 에이전트에게 URL 데이터베이스의 등록 도메인 이름을 이용하여 구한 검색어와 카테고리를 매개변수로 웹에서 다시 관련된 내용을 검색하여 색인 데이터베이스를 재구성하도록 요구한다.

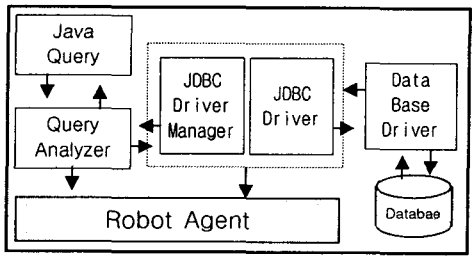

그림 2 검색 에이전트의 처리 구조

그림 2는 검색 에이전트가 색인 데이터베이스 및 로봇 에이전트와 연동되는 과정을 보여 주고 있다. 자바기반의 검색 에이전트는 색인 데이터베이스와 연동하는 과정에서 기본적으로 JDBC를 사용한다[6].

질의 분석기(query analyzer)는 통합형 SQL 데이터 질의어를 처리하는 프레임워크(frame work)로서 질의 및 처리결과를 분석하여 색인 데이터베이스 및 로봇 에이전트로 제어를 스위칭(switching)시키는 기능을 수행하는 제어모듈이다.

2.3 연관 규칙 탐사 기법

연관 규칙 탐사기법이란 유효성 측정을 위한 데이터마이닝의 분석기법으로 항목들의 집합으로 표현된 트랜잭션들에서 각 항목간의 연관성을 반영하는 규칙을 찾아내는 방법이다[3]. 데이터마이닝은 대용량의 데이터 내에 존재하는 데이터간의 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화함으로써 유용한 정보를 추출하는 과정으로 연관 규칙 탐사, 항목 분류, 클러스터링, 요약, 순차 패턴 탐사 기법 등이 있다[4, 8, 9].

연관 규칙 탐사 기법은 동시 혹은 순차적으로 발생한 데이터들을 탐사하여 데이터간의 연관성을 정의한다. 구매정보 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용되는 기법으로 상품이나 서비스의 거래기록 데이터로부터 상품간의 연관성 정도를 측정하여 연관성이 많은 상품들을 그룹화 하여 동시에 구매될 가능성이 큰 상품들을 찾아냄으로서 이를 적용할 수 있다.

연관 규칙 "R: X→Y"는 상품 X가 구매되어진 경우는 상품 Y도 구매된다는 규칙을 표현한다. 이러한 연관 규칙의 장점은 규칙이 조건부→결과로 표현되어 연관성 분석이 쉽고 분석방향이나 목적변수가 없는 경우에도 유용하며 거래내용 데이터를 변환 없이 이용할 수 있으며 분석을 위한 연산이 간단하다[10].

연관규칙 탐사규칙을 이용하여 마이닝된 정보의 측정 지표는 그림 3과 같이 지지도(Support)와 신뢰도(Confidence)로 표현된다. 지지도는 전체 트랙잭션 N의 경우에서 연관 규칙을 만족하는 트랜잭션의 비율이며 지지도는 자주 발생되는 패턴 혹은 규칙의 빈도를 표시하므로 해당 패턴이나 규칙의 유용성을 나타낸다.

기본가정

Itemset I의 부분집합 X에 대해, X ⊆ T이면 T는 X를 만족한다고 정의한다.

Itemset X (X ⊆ I)를 만족시키는 D의 트랜잭션 수를 |X|로 표기한다.

X, Y ⊆ I에 대한 R은 X∩Y = ∅ 의 특성을 갖는다.

Measure

Support(연관 규칙 X ⇒ Y에 대한 지지도)

$$S = \frac{|X \cup Y|}{N}$$

Confidence(연관 규칙 X ⇒ Y에 대한 신뢰도)

$$C = \frac{|X \cup Y|}{|X|}$$

그림 3 연관규칙을 위한 지지도 및 신뢰도 측정

신뢰도는 같은 상황에서 X가 발생한 모든 트랙잭션의 경우에 대하여 Y가 발생한 트랙잭션의 비율이며 규칙의 실행 시 그 정확도를 표시한다. 이때 지지도(S)는 최소지지도(S_{min})보다 크거나 같고, 신뢰도(C)는 최소신뢰도(C_{min})보다 크거나 같아야만 규칙 R 이 N개의 트랜잭션 T로 구성된 집합 D 에서 성립된다

3. 검색 엔진의 설계 및 구현

3.1 로봇 에이전트의 구조

자바 프로그램으로 구현된 로봇 에이전트는 방문한 웹 사이트의 검색 패턴 정보를 분석하여 웹 페이지로부터 URL 주소, 제목, 내용 등의 정보를 추출하여 색인 데이터베이스에 저장하는 기능을 담당한다.

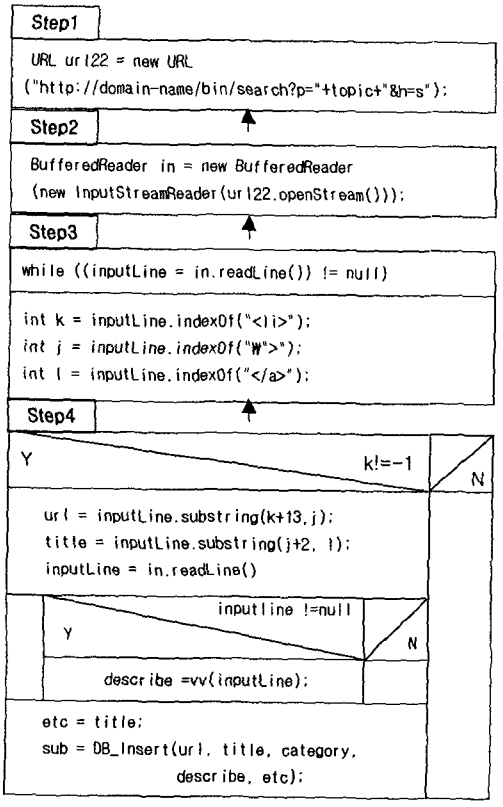

그림 4 색인 데이터 추출 및 저장 과정

그림 4는 로봇 에이전트를 이용한 색인 데이터베이스 구축 과정을 나타낸다. Step 1에서 URL 정보에 등록된 도메인 이름을 통해 URL 클래스 객체를 생성한다. Step 2에서는 그 객체의 논리적인 입력 스트림(stream)을 이용하여 버퍼링이 가능한 InputStreamReader 클래스를 생성하는 과정이다. Step 3에서는 입력 스트림의 값이 Null일 때까지 반복적으로 문서 내의 패턴을 분석하여 문자열 인스턴스 값의 위치정보를 검출해 낸다. Step 4에서는 Step 3에서 검출한 위치 정보 값을 이용하여 문자열 인스턴스를 서버 문자열로 잘라내어 url, title 등의 저장 정보를 색인 데이터베이스에 저장한다. 이와 같은 과정을 통해 주제어와 카테고리 별로 데이터를 분류하여 색인 데이터베이스를 구축한다.

3.2 검색 에이전트의 구조

검색 에이전트는 색인 데이터베이스와 연동하기 위해 IP 주소(IP-Address)와 포트 번호(port number)를 객체화하고 사용자가 질의한 검색어를 SQL언어로 캡슐화(encapsulation)함으로써 원격에 분산된 데이터베이스 서버와 통신을 수행한다.

색인 데이터베이스 드라이버와 내부 연결에 이용되는 JDBC 드라이버는 웹 환경을 지원하는 네트워크 프로토콜 기반의 Oracle용 Thin Driver를 이용하였고 색인 데이터베이스와 사용자간의 실제적인 연결은 "java.sql" 패키지와 "java.net" 패키지에 포함된 여러 클래스의 API 메소드를 통하여 이루어진다.

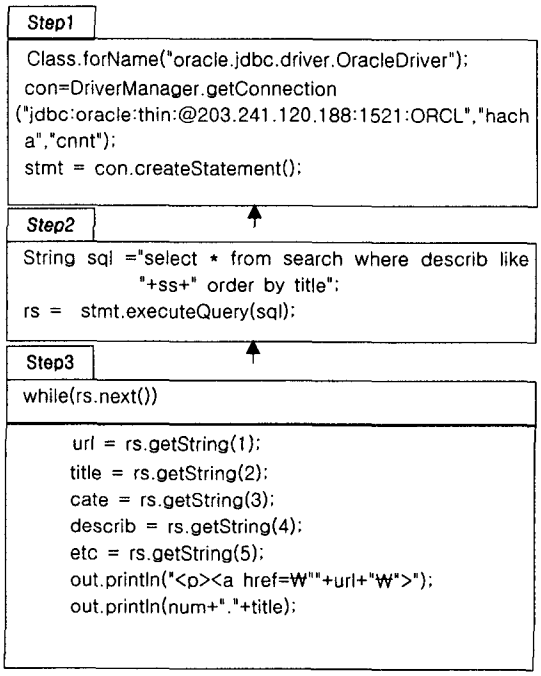

그림 5 검색 에이전트의 질의 처리 과정

그림 5는 JDBC 드라이버 통하여 색인 데이터베이스에 검색어를 질의하여 관련된 레코드를 검출하고 HTML 문서 형식으로 결과를 출력하는 과정을 보여주고 있다. 먼저 Step 1에서는 Class의 forName() 메소드로 연동에 필요한 드라이버를 로드한 후 DriverManger 클래스의 getConnection() 메소드를 통해 원격의 데이터베이스와 연결하여 실제 데이터베이스와 연결을 담당할 Connection 객체를 생성한다.

Step 2에서는 executeQuery() 메서드를 통해 데이터베이스의 테이블에 저장된 데이터를 검색하는 SQL 문장을 실행하여 ResultSet 클래스 타입의 결과 값을 rs변수에 대입한다. Step 3에서는 ResultSet 객체인 rs에 저장된 정보를 HTML 코드 형태로 웹 브라우즈에 출력하는 과정을 수행한다.

3.3 연관규칙 탐사기법의 구현

연관 규칙 탐사 기법은 기본적으로 2 단계로 구성된다[10]. 제 1 단계에서는 사용자가 미리 정의한 최소 지지도를 만족하는 데이터 항목 집합을 탐사하는 단계로써 각각의 데이터 항목에 대하여 지지도를 계산한 후 최소 지지도를 만족하는 데이터 항목들만 추출하다. 제 2 단계에서는 1 단계에서 얻은 데이터 집합들 중에서 데이터의 부분 집합에서 생성된 규칙 중 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 탐사하여 최종 규칙으로 정하게 된다. 연관규칙의 탐사기법의 성능은 1단계에서 결정되며 1단계에서 추출하는 빈발 항목 집합(large itemsets)을 확인한 후에 해당되는 연관규칙을 단계 2에서 쉽게 유도할 수 있다. 본 연구에서는 그림 6와 같이 빈발 항목 집합을 생성하기 위해 검색창에서 주어진 키워드를 테이블에 설정된 기본키와 외래키의 관계를 따라 레코드들을 셀프조인(self-join) 연산을 시행하고 임계값으로 주어진 최소지지도에 따라 후보 항목 집합이 결정된다. 이때 전체 트랜잭션의 크기와 개수를 줄이기 위해 전체 데이터베이스를 탐색 대상으로 하지 않고 해당 사용자의 탐색 패턴이 저장된 로컬 데이터베이스로 영역을 국한 시켰다.

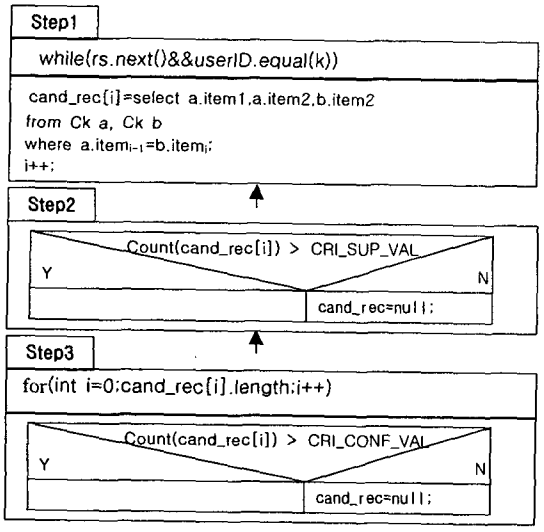

그림 6 후보항목과 연관규칙 생성 과정

또한 연관규칙을 결정하기 위해 1단계에서 생성된 후보 항목집합을 대상으로 임계값으로 주어진 최소 신뢰도에 따라 주어진 키워드와의 연관성을 결정하여 의미론적 정보 검색 기능을 구현하였다.

4. 실험 및 결과

본 연구에서 구축한 "AISE"라는 선박 부품 전문 검색 엔진은 프로그램 개발 언어로 자바1.3을 사용하고 색인 데이터베이스는 오라클 8.1을 사용하여 구현하였다. 그림 7는 선박 부품 전문검색엔진의 초기화면으로 검색어를 텍스트 박스에 입력하면 검색어와 매치(match)된 주제어 정보를 색인 데이터베이스에서 찾아 하이퍼텍스트 문서로 결과를 작성하여 그림 8과 같이 보여주고 있다. 표 1은 선박부품 관련 검색어 기준으로 일반 상용 검색엔진과 AISE를 비교 실험한 결과표이다. 결과표에서 나타나는 것과 같이 AISE가 다른 검색엔진에 비해 상대적으로 낮은 불량 링크률을 보이고 있다.

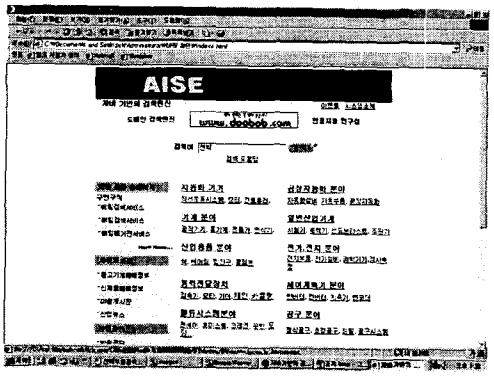

그림 7 선박부품 전문검색엔진의 초기화면

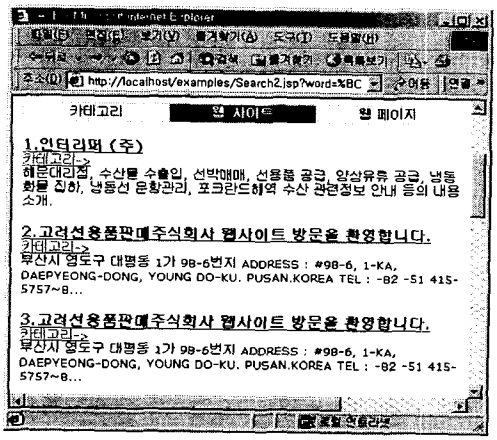

그림 8 검색 결과 출력 화면

표 7 기존 검색 엔진과의 링크 불량률 비교표

검색엔진	불량 링크률
AISE	12.4 %
S 검색엔진	18.43 %
K 검색엔진	20.6 %

5. 결 론

검색 엔진은 사용자의 검색요구에 대해 정보량과 검색속도 뿐만 아니라 정보의 정확성에 의해서도 평가를 받는다. 본 선박 부품 전문 검색 엔진은 데이터마이닝 기법 중 연관 규칙 탐사 기법을 이용하여서 검색어 패턴 매치 방식의 일반 검색 엔진에 비해 사용자의 검색 질의어에 대한 정확성

이 많이 증가되었다. 하지만 본 연구에서 구현한 검색 엔진 역시 불량 링크률이 아직은 높은 편이다. 이것은 사용자의 접근 패턴만을 고려한 탐색 결과로 보인다. 따라서 색인 데이터베이스 구축시 무결성을 보다 향상시키기 위해 사용자의 검색 경험을 참조한 지식 베이스를 구성하여 검색에 반영한다면 보다 검색 질의어와 관련성이 높은 정보를 사용자에게 제공할 수 있을 것으로 사료된다.

참 고 문 헌

[1] 고휘정, "인터넷 웹사이트의 사용자 인터페이스 분석 및 평가에 대한 연구", 홍익대학교, 석사논문, 1998.

[2] 구홍서, "www과 데이터베이스 연동기술의 조사 분석", 정보과학회지 제18권 4호, 2000.

[3] 박종수외2, "연관규칙탐사와 그응용",정보과학회지 제16권 pp.37-44, 1998.

[4] 이도헌, "데이타마이닝 : 개념 및 연구동향", 데이터베이스연구회지 제13권, pp.122-137, 1998.

[5] 이원휘, "Java를 이용한 인터넷 정보검색 엔진의 설계 및 구현", 전주대학교 석사논문, 1999.

[6] 하창승, "웹 인터페이스 개발을 위한 JDBC-ODBC Brideg 기법과 ISAPI 확장기법에 관한 비교 연구", 한국해양정보통신학회 제5권 제3호, pp 493~501, 2001.6.

[7] George Reese, "Database Programminzg with JDBC and JAVA", O'RILLY, 1996.

[8] Michael J.A.Berry, Gordon Linoff, Data Mining Techniques-For Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997.

[9] Pieter Adrians, Dolf Zantige, Data Mining, Addison Wesley, 1996.

[10] Rakesh Agrawal, R.Srikant, "Fast Algorithms for Mining Association Rules", *Proceedings of VLDB Conference*, pp.487-499, 1994.