제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

온톨로지 기반의 쇼핑 사이트 고객을 위한 검색 시스템

구미숙*, 황정희*, 류근호*
*충북대학교 전자계산학과
e-mail:gumisug@dblab.chungbuk.ac.kr

Ontology based Retrieval System for Shopping Sites Customer

Mi-Sug Gu*, Jeong-Hee Hwang*, Keun-Ho Ryu*
*Dept of Computer Science, Chung-Buk University

요 약

시멘틱 웹은 기존의 웹과는 달리 정보의 의미가 정의되고, 이들 간의 의미적 연결을 지원한다는 특징이 있어서, 최근 차세대 웹으로 부각되고 있다. 이러한 의미적 연결을 위해서 시맨틱 웹의 기반인 온톨로지가 필요하다. 온톨로지는 리소스에 대한 메타데이터를 정의하여 의미적 연결이 가능하게 하므로 효율적인 정보 검색이 가능하다. 이 논문에서는 정보 검색의 효율을 증가시키기 위해서 시맨틱 웹의 핵심인 온톨로지 기반의 정보 검색 시스템을 제안한다. 쇼핑 사이트에서 효율적인 마케팅을 위해 사용자의 구매 패턴을 조사하여 고객에게 알맞은 정보 추천을 하기 위한 것을 목적으로 한다. 온톨로지의 구축은 XTM을 기반으로 토픽맵을 이용하였다. 그리고 온톨로지를 기반으로, 사용자의 구매패턴을 찾아서 정확한 정보 전달을 위해서 데이터 마이닝 기법을 이용하였다. 빈발패턴 트리 기법을 기반으로 하는 멀티 레벨 멀티 디멘션 빈발 패턴 마이닝 알고리즘을 이용하여 사용자 패턴을 분석하여 정보 검색에 효율을 기하였다.

1. 서론

시멘틱 웹(Semantic Web)은 기존의 웹과는 달리 정보의 의미가 정의되고, 이들 간의 의미적 연결을 지원한다는 특징이 있어서, 최근 차세대 웹으로 부각되고 있다[1]. 온톨로지는 리소스에 대한 메타데이터를 정의하여 의미적 연결이 가능하게 하므로 효율적인 정보 검색이 가능하다. 이 논문에서는 정보 검색의 효율성을 증가시키기 위해서 시맨틱 웹의 기반인 온톨로지를 이용한 정보 검색 시스템을 제안한다. 온톨로지란 어떤 특정 도메인의 정보들과 그 정보들 간의 관계를 정의해 놓은 것으로써, 특정 분야의 지식을 표현하기 위한 기본지식 체계를 제공하므로 사용자가 원하는 정보에 대해 정확한 검색 결과를 제시 해 준다[2]. 이 논문에서는 이러한 온톨로지를 기반으로 데이터 마이닝 기법을 적용하여 효율적인 정보 검색 기법을 제시하고자 한다. XTM을 이용하여 온톨로지를 설계, 구축하였는데, 토픽맵은 토픽, 연관 관계, 리소스 정보인 어커런스 등으로 구성된다[3]. 온톨로지는 쇼핑사이트가 효율적인 마케팅을 위해 고객의 구매패턴에 대한 정보를 찾아내어 고객에게 알맞은 정보 추천을 목적으로 하기 때문에 상품 온톨로지를 구축하였다. 온톨로지는 기존의 택소노미와 같이 계층 구조를 가지고 있다. 그리고 이와 유사한 계층 구조를 가지고 있는 데이터 마이닝 기법인 멀티-레벨 멀티-디멘션 빈발 패턴(Ada-FP) 알고리즘을 적용하여 고객의 쇼핑 구매패턴을 찾아낸다[5]. Ada-FP 알고리즘은 FP-growth 알고리즘을 확장한 데이터 마이닝 기법으로, Apriori 알고리즘 기법에 비해서 데이터베이스 스캔을 적게 하고, 지지도 제약조건을 유동적으로 적용시키는 장점이 있다. 데이터 마이닝 과정을 통해서 조사한 고객의 구매패턴을 이용하여, 토픽맵을 이용하여 만든 온톨

이 연구는 한국전자통신연구원의 정보통신 서비스 연구단의 연구비 지원으로 수행되었음

로지를 통해서 사용자의 구매패턴에 알맞은 상품의 추천이 가능하다. 이와 같이 시맨틱웹은 사용자가 원하는 정보에 대해서 정확한 검색결과를 제시해 준다.

2. 관련 연구

2.2 XTM(XML Topic Map)

토픽맵(Topic Map)은 주제 중심으로 개념을 명세화하고 개념들 간의 연관 관계를 정의한 모델로서 ISO의 표준안으로, 지식 관리 시스템의 지식맵, 컨텐츠 관리 시스템의 컨텐츠 맵 그리고 시맨틱 웹 온톨로지 등의 데이터 모델로 사용되고 있다. 토픽맵 모델은 토픽(Topic), 어커런스(Occurrence), 연관관계(Association)등 세가지 핵심 요소로 구성되어 있다. 토픽은 주제, 어소시에이션은 주제와의 연관 관계, 어커런스는 주제에 대한 리소스가 위치한 정보를 나타낸다[3].

2.2 FP-growth Algorithm

트랜잭션 데이터베이스에 대해서 빈발 패턴을 요약하여 중요한 정보를 저장한 prefix 트리에 기반한 빈발 패턴 트리(FP-tree) 구조를 가지고 있다. Apriori기법보다 후보 집합 생성 비용이 적게 들어 효율적이다. 빈발 패턴 트리에 기반 한 FP-growth를 발전시켜 패턴 확장에 의해 빈발 패턴 집합을 마이닝 하는 방식으로 다음 세 가지 효율성이 있다.

(1) 대량의 데이터베이스를 축약된 데이터 구조로 만들어서 반복된 데이터베이스 스캔을 적게 한다.

(2) 빈발 패턴 트리에 기반을 둔 데이터 마이닝은 고비용의 후보 집합 생성을 피하기 위해 패턴 확장 방식을 사용한다.

(3) 나누기(partition-based), 분할과 통합(divide and conquer)기법을 사용하여, 작은 집합으로 분해하여 조건부 데이터베이스에 제한된 마이닝 패턴을 만들어서 검색 공간을 줄여준다[4].

2.3 멀티-레벨 멀티-디멘션 빈발 패턴 알고리즘

멀티-레벨 멀티-디멘션 빈발 패턴(multi-level multi-dimension frequent pattern) 알고리즘은 다음 세 단계로 이루어져 있다[5].

첫째, 하나의 아이템과 디멘션으로 이루어진 카운트 값을 계산하기 위해서 데이터베이스를 한번 스캔한다. 빈발 1-아이템이나 1-디멘션은 그 카운트 값이 프린팅 임계값(printing threshold)을 통과한 것이다.

둘째, 빈발 패턴 트리를 구축하기 위해서 데이터베이스를 다시 스캔한다. 아이템이나 디멘션은 그 카운트 값이 패시지 임계 값(passage threshold)을 통과해야 빈발 패턴 트리에 나타난다.

셋째, 모든 빈발 패턴을 생성하기 위해서 조건부 빈발 패턴 트리를 반복적으로 마이닝하여, 빈발 패턴 아이템이 전체 개념 계층구조를 확장한다.

Ada-FP 알고리즘은 불가능한 더 긴 패턴을 걸러내는 패시지 임계값과, 현 단계에서 평가되는 빈발 패턴을 측정하는 프린팅 임계값을 가지고 있다.

이 알고리즘은 다음 세 가지의 특징이 있다.

- FP-growth 알고리즘의 확장된 형태로서 패턴 나누기와 분할과 통합을 한다.
- 트랜잭션 데이터베이스를 읽을때 아이템과 디멘션에 대한 정보를 계산하고, 정해진 임계값을 만족하면 빈발한 것으로 간주한다.
- 지지도 제약조건을 유연하게 적용하는데, 기존의 정해진 지지도 임계 값의 적용으로 인한 유용하지 못한 패턴을 만들어 내거나, 잠재하고 있는 유용한 패턴을 빠트릴수 있는 단점을 보완한 방식이다.

3 쇼핑 사이트의 상품 온톨로지 구축

이 논문에서는 XTM을 이용하여 상품 온톨로지를 설계 구축하였다.

상품 온톨로지는 쇼핑 사이트에서 취급하고 있는 상품을 조사하여, 쇼핑 사이트의 메뉴에 보이는 각 계층적인 관계를 이용하여 상품 온톨로지를 구축하였다. 상품 아이템을 루트로 하여 각 상품의 이름인 가전제품, 컴퓨터 제품, 의류, 식품류 등을 그 서브 계층으로 만든다. 그리고 그 하부에는 구체적으로 상품을 구성하는 품목을 위치시켰다. 예를 들어, 가전제품의 경우 냉장고, 텔레비전, 에어컨등이 온다. 그리고 그 아래에는 상품의 구체적인 이름이 오도록 하여, LG 에어컨, Samsung TV, 금성 냉장고 등을 위치시킨다. 이와 같이 각 아이템에 대한 계층구조를 가진 상품 온톨로지를 구축하였다.

아래 그림은 상품 온톨로지에 대한 계층구조이다.

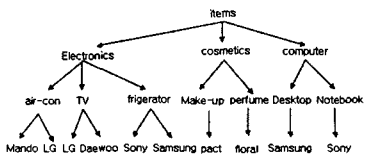

그림 1. 상품의 계층구조

이런 계층구조의 온톨로지 문서인 XTM 문서를 온토피아 사(http://www.ontopia.net)의 토픽맵 툴인 옴니게이터 (Omnigator)를[6] 이용하여 유효성 검사를 한다. 온톨로지는 XML형식의 문서이므로 온톨로지 데이터베이스에 저장하기 위해서는 Saxparser와 TMParser를 통해 파싱 과정을 거치게 된다. 파싱이 된 온톨로지는 하이버네이트(hibernate)를 사용하여 객체관계형(object relation)인 클래스 상태로 온톨로지 데이터베이스에 저장한다.

아래 그림2는 옴니게이터를 이용한 유효성 검사 화면이다.

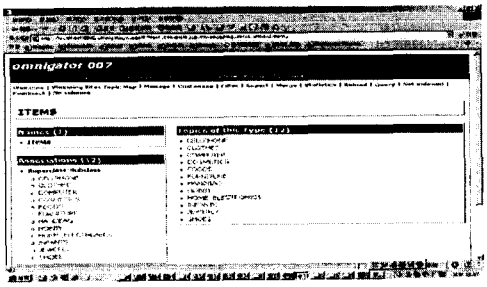

그림 2 . 옴니게이터의 유효성 검사

온톨로지 데이터베이스 스키마 설계는 각 상품명과 그 상품을 취급하고 있는 쇼핑 사이트를 토픽 클래스로 저장한다. 그리고 상품이 가지고 있는 상하 계층구조와 그 상품이 판매되고 있는 쇼핑 사이트에 대한 관계 설정을 superclass와 subclass, be_sold라는 관계를 토픽맵의 어소시에이션 클래스에 저장한다. 그리고 어커런스 클래스는 쇼핑 사이트의 주소가 된다. 이러한 구조인 온톨로지는 시맨틱웹의 기반이며 계층구조를 이용하여 정보검색에 효율적인 결과를 가져다준다.

4. 데이터 마이닝 기법에 의한 사용자 구매패턴

이 논문에서는 쇼핑 사이트 고객의 구매패턴을 알아내기 위해서 마이닝 기법인, 멀티-레벨 멀티-디멘션 빈발 패턴 알고리즘을 사용하였다[5].

쇼핑 사이트의 효율적인 마케팅에 이용하기 위해 고객의 구매패턴을 조사하여 온톨로지에 적용하여 효율적인 정보를 검색하는 것을 목표로 한다. 고객의 구매 패턴을 조사하기 위해서는 다음과 같은 과정으로 수행 된다.

첫째, 쇼핑 사이트의 고객의 상품 구매 패턴을 찾아서 자주 방문하는 쇼핑 사이트와 품목에 대한 정보를 찾아내어 그 상품과 쇼핑 사이트를 데이터베이스에 저장한다. 아래 테이블은 데이터베이스에 저장된 트랜잭션 데이터베이스이다.

TID	쇼핑사이트	아이템
T1	daum	Samsung 냉장고,LG TV,Sony
T2	yahoo	Mando 에어컨,Sony 노트북
T3	Empas	Sony TV,팩트,Mando 에어컨
T4	CJmall	Daewoo TV,LG 에어컨
T5	KTmall	Samsung 데스크탑

표 1. 트랜잭션 데이터베이스

이와 같이 저장된 데이터에 대해서 임계치를 사용하여 각 아이템과 디멘션에 대한 카운트 값을 계산한다. 다음은 아이템과 디멘션에 대한 임계값 카운트 테이블이다.

	카운트	패시지임계값	프린트임계값
Samsung 냉장고	2	1	2
Mando 에어컨	1	2	2
Sony 노트북	1	1	2
LG TV	2	1	2
팩트	2	2	2
Daum	3	2	3
Yahoo	3	2	2
Empas	2	2	2

표 2. 아이템과 디멘션 임계값 카운트

둘째, FP-growth에 의한 구조인 prefix-트리 구조를 적용한 빈발 패턴 트리를 구축한다. 아래 그림은 트랜잭션 데이터베이스에 대한 빈발 패턴 트리이다.

셋째, 멀티-레벨 멀티-디멘션 빈발 패턴을 생성하기 위해, 계층 구조를 가진 빈발 패턴 트리를 반복적으로 방문하여 패턴 생성과정을 반복한다.

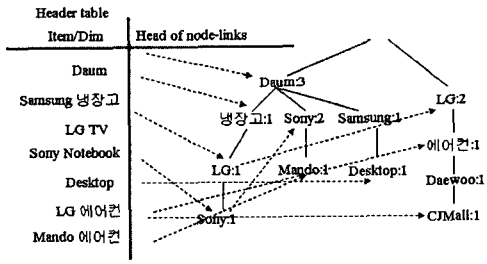

그림 3. 트랜잭션에 대한 빈발 패턴 트리

이와 같은 반복된 과정을 통해서 쇼핑 사이트 고객의 상품 구매 패턴을 알아낼 수 있다. 예를 들어, Daum사이트에 자주 들르는 어떤 고객이 의류에 관심이 있다는 것을 알아낼 수가 있으며, 그 고객에게

관심이 있는 상품을 추천할 수 있다.

5. 온톨로지를 이용한 시맨틱 웹 검색 시스템

빈발 패턴 트리 구조를 가진 쇼핑 사이트 고객의 상품 구매 패턴에 대한 빈발 패턴이 만들어지면 이것을 이용하여, 이미 구축된 상품 온톨로지를 이용하여, 사용자에게 알맞은 상품 품목과 쇼핑 사이트를 검색해 주는 검색 시스템을 구축한다. 온톨로지 데이터베이스에 저장된 온톨로지는 토픽 클래스로 상품 품목과 쇼핑 사이트가 저장되고, 그 연관관계와 쇼핑 사이트 주소인 어커런스도 클래스로 저장된다. 그리고 Ada-FP 알고리즘의 아이템인 각 상품 품목과 디멘션인 쇼핑 사이트는 토픽맵으로 구축된 온톨로지의 토픽과 매치를 시켜서, 빈발 패턴 트리의 상하 계층 구조는 품목간의 수퍼클래스와 서브클래스라는 연관 관계는 온톨로지의 어소시에이션이 된다. 그리고 쇼핑 사이트의 실제 주소가 온톨로지의 어커런스가 된다. 다음 그림4는 검색 시스템의 전체적인 처리 과정이다.

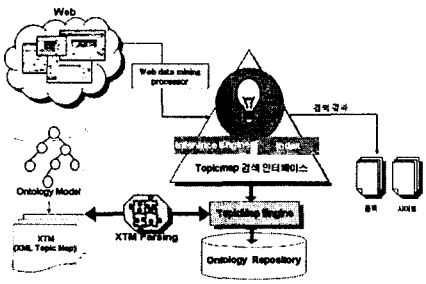

그림 4. 온톨로지 이용한 검색처리 과정

이와 같은 과정으로 수행되는 온톨로지를 이용한 정보검색 결과를 보여주기 위한 정보 검색 시스템의 사용자 인터페이스는, J2Sdk1.4.2_05, Oracle 9i와 공개 소스인 TM4J 토픽맵 엔진을 이용하였다.

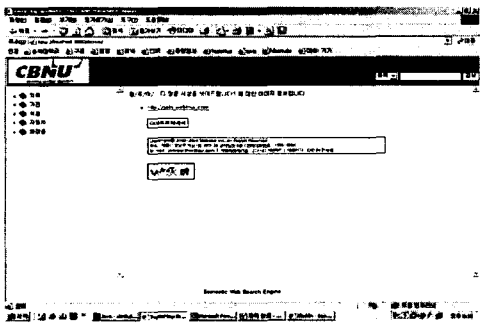

그림 5. 검색 인터페이스

위의 그림 5는 사용자에게 사용자가 원하는 정보를 온톨로지의 계층구조를 이용하여 검색한 결과를 보여주는 사용자 인터페이스이다.

5. 결론

이 논문에서는 대량의 다양한 정보를 사용자에게 좀더 효율적이고 정확한 검색 결과를 보여주기 위해서, 데이터 마이닝 기법인 멀티 레벨 멀티 디멘션 빈발 패턴 알고리즘과 토픽맵을 이용하여 구축된 온톨로지를 이용한 정보 검색 시스템을 제안한다. 기존의 웹의 키워드 검색에 의한 정보 검색이 사용자가 원하는 정보를 정확하게 전달하기 보다는 웹 페이지를 디스플레이 하는 정도의 단점을 가지고 있기 때문에, 시맨틱 웹의 장점인 사용자가 원하는 정보를 정확하게 전달하기 위해서 온톨로지를 구축하여 시맨틱 웹을 구현하였다. 앞으로는 RDF(S), DAML, OWL등과 같은 온톨로지 언어를 이용한 온톨로지 구축과, 다양한 데이터 마이닝 기법을 적용하는 연구를 계속 할 것이다.

참고문헌

[1] 서명희, 안재용, 민준기, 정진완, "시멘틱 웹상의 RDF 데이터 관리 시스템",정보과학회 춘계학술대회,2003

[2] 이정원, 방건동, 박세형, 백두권 "온톨로지 기반 설계 문서관리 시스템 설계 및 구현",한국정보 과학회,2001

[3] S. Pepper, B. Moore, "XML Topic Maps(XTM) 1.0", TopicMaps.Org.

[4] Jiawei Han, Jian Pei, Yiwen Yin, "Minig Frequent Patterns without Candidate Generation", ACM SIGMOD,1999

[5] Runying Mao,"Adaptive-FP: an Efficient and Effective Method for Multi-Level Multi-Demensional Frequent Pattern",Thesis Science In The School Of Computing Science,2001

[6] http://www.ontopia.net