제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

TMDR 기반의 키워드 모호성 해결을 위한 질의 기법에 관한 연구

정계동*, 황치곤*, 신효영**, 최영근*
*광운대학교 컴퓨터과학과
**경복대학 인터넷정보과
e-mail:duck1052@kw.ac.kr

The study of Query Method for keyword disambiguation based on TMDR

Gye-Dong Jung*, Chi-Gon Hwang*, Hyo-young Shin**, Young-Gun Choi*
*Dept of Computer Science, Kwang-woon University
**Dept of Internet Information

요 약

키워드의 모호성으로 인한 검색 결과가 부정확하게 되는 문제를 해결하기 위한 질의기법을 제안한다. 제안하는 질의 기법은 TMDR을 이용한다. TMDR은 로컬 데이터베이스를 통합하기 위한 스키마 정보의 통합 관리하기위한 MDR과 데이터 접근을 위해 온톨로지 지식 저장소로 토픽맵으로 구성된다. 토픽맵은 연관관계 분석을 통한 데이터 모호성 해결을 지원한다. 이를 이용하여 기존 시스템의 이질적 문제를 해결한다. 토픽맵은 지식을 제공하고, 지식 간의 관계성을 제공하므로 키워드의 모호성을 해결할 수 있다. 본 논문에서는 이러한 TMDR을 이용하여 키워드의 모호성과 기존 시스템의 이질적 환경을 적응하기 위한 질의기법을 제안한다.

1. 서론

최근 많은 정보가 발생하고 소멸되고 있으며, 이에 따른 방대한 지식도 제공되고 있다. 일반적으로 정보나 지식은 찾지 못할 뿐 없는 경우는 적다. 예를 들어, 특정 저자의 책을 검색하고자 할 때, 저자 이름만 같으면 모두 검색된다. 물론 동명이인(검색한 사람과 다르지만 이름이 같은 사람)이 쓴 책도 검색이 된다. 이러한 사실을 알고 있는 사람은 구분해서 정보를 확인할 수 있지만, 그렇지 못한 사람은 이를 구분하기 힘들다. 이 사례와 같이, 사용자가 아는 정보만을 통해서 모호한 검색결과를 제거하는 방안이 필요하다. 이를 해결하기 위한 대표적 기술이 시멘틱 웹이 있다.

시멘틱 웹의 특징은 웹 컨텐츠에 대해서 기계가 처리할 수 있는 형식으로 정보가 구축된다는 점이다. 따라서 시멘틱 웹은 일반 정보에 정보가 갖는 의미를 추가함으로써 이용성이 향상된다. 많은 연구자들은 웹 페이지에 시멘틱 정보를 추가할 수 있는 방법을 연구해오고 있다. 시멘틱 정보의 추가에서 객체의 이름은 매우 중요한 요소가 된다.[1][2]

이를 위해 ISO/IEC 11179 MDR(Meta-Data Registry)에 의해 지속적으로 관리되는 메타데이터들은 정보 시스템의 설계, 구현뿐만 아니라 기존 시스템 간의 데이터 공유 및 교환에 중요한 역할을 할 수 있다.[3]

객체의 이름은 특정 사물이나 단체의 명칭과 같이 단어의 조합으로 구성된다. 단어의 조합 때문에 시멘틱 정보의 모호성이 발생된다.[4] 모호성은 동명이의(명칭은 같으나 뜻이 다른 경우, 이름은 같으나 다른 사람인 경우 등)와 동의이어(뜻이나 사람은 같으나 불리는 명칭이 다른 경우) 때문에 검색을 위한 정보자체의 모호성이 검색의 성능을 떨어뜨리는 특성이 있다.

이에 따라 본 논문에서는 온톨로지 표현 기법 중 정보간의 연관관계를 지원하는 토픽맵을 이용하여 모호성 문제를 해결할 수 있는 방안을 제안하고, 새로운 시스템을 개발하는 것 보다 기존 시스템의 수정없이 접근할 수 있도록 하기 위해 MDR을 적용한 질의 방안을 제안한다.

2. TMDR

TMDR은 지식표현을 위한 토픽맵과 데이터 통합을 위한 MDR을 결합한 것이다.[5] 이를 통해 효율적인 데이터 통합과 모호한 키워드에 대한 정확성을 부여한다.[5]

2.1 토픽맵

토픽맵은 기본적으로 토픽, 연관관계(association), 어커런스(occurrence)로 이루어진다. 토픽은 주제를 표현하는 이름이다. 주제가 포함된 영역, 기본이름, 가변이름들을 포함하며, 토픽클래스가 되는 상위 토픽을 가진다. 연관관계는 토픽간의 상호 연관성을 표현한다. 연관관계는 토픽에서 중요한 개념 중 하나로 일반적인 계층인 트리 구조로 표현된 지식을 접근할 때 발생하는 문제점을 해결하기 위한 해법이 된다. 각 토픽은 토픽에 대한 지식을 표현하는 위치 및 참조할 자원을 연동해야 하는데, 이러한 정보자원

<표 1> 데이터 충돌의 유형과 사례

종류	설명
값 충돌	동일 값에 따른 다양한 해석, 동일 값에 대한 다양한 표현 (예. 눈(目)과 눈(雪))
형식 충돌	데이터 표현 형식의 차이로 발생하는 문제 (예. 'yy/mm/dd'과 'mm/dd/yy')
단위 충돌	동일 데이터에 대한 단위 차이로 발생하는 문제 (예. 미터와 인치)
정밀도 충돌	동일 데이터에 대한 정밀도 표현의 차이 (예. π 값→3.14와 3.14159)

의 위치에 대한 링크인 어커런스가 있다.

일반적으로 발생하는 데이터 충돌을 표 1과 같이 정리할 수 있다. 이런 충돌 문제는 토픽맵의 토픽과 연관관계를 통하여 해결할 수 있다.

TMDR에서 토픽맵은 원래의 개념을 적용하고 어커런스 참조 방법을 본 시스템에 적합하게 수정한다. MDR은 표준메타데이터를결정하고, 결정된 항목에 대한 레거시 시스템의 메타데이터를 매핑하여 이질적인 데이터베이스 시스템을 통합하기 위한 기술이다. 여기서 말하는 표준 메타데이터는 이질적 데이터베이스를 대표하는 메타데이터로 각 테이블의 대표 필드명이다. 토픽맵에서 사용하는 토픽 중 스키마토픽(schema topic)은 표준 메타데이터를 이용한다. 그리고 토픽맵의 정보자원을 연결하는 어커런스가 있는데 이는 MDR의 레거시 접근 정보를 이용하도록 하여 토픽맵에서는 어커런스 부분은 표현하지 않고 토픽맵과 MDR의 결합에서 다루도록 한다.

3.2 MDR(Meta-Data Registry)

MDR는 데이터의 상호운용성을 확보하기 위한 목적으로 고안되었다. MDR은 메타데이터 등록과 인증을 통하여 표준화된 메타데이터를 유지 관리하며, 메타데이터의 명세와 의미의 공유를 통해 메타데이터 집합 또는 메타데이터 요소 간의 호환성을 유지시킨다. 이를 활용한 메타데이터 관리 시스템은 서로 다른 데이터베이스가 같은 개념에 대해 서로 다른 식별자 혹은 서로 다른 언어를 사용할 경우 이를 해결해 주기 위해 공유되는 개념화를 정형적, 명시적으로 명세화한 결과물인 온톨로지가 결합되어 있다. 일반 사용자에게 보다 편리한 사용자 인터페이스 환경을 제공하기 위해서는 현재의 윈도우즈의 기반 사용자 인터페이스의 차원을 넘어서 사용자의 작업을 대행해 줄 수 있는 에이전트 시스템이 제공되어야 한다. 또한 에이전트 시스템서비스 확장과 사용보급을 위하여 응용을 위한 미들웨어 플랫폼에 대한 연구개발이 이루어져야 한다.

MDR은 메타스키마와 메타로케이션으로 구성된다. 메타스키마는 글로벌 스키마와 로컬 스키마 사이의 매핑정보를 제공하는 온톨로지이다. 메타로케이션은 로컬시스템의 접근 정보를 저장 관리한다.

3.3 결합

MDR은 표준 항목(가상 스키마)과 레거시 시스템의 로컬 스키마(실제 스키마) 간의 매핑정보와 레거시 시스템을

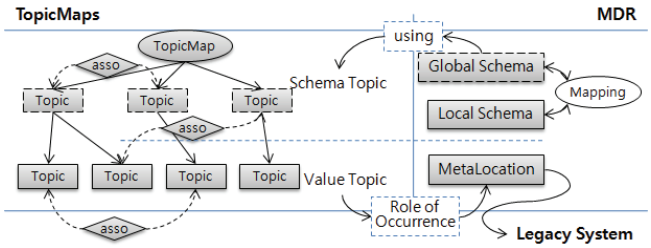

그림 1 TMDR의 구성 및 운용

접근하기 위한 접근정보를 관리한다. 토픽맵은 표준 항목과 실제 데이터 키워드를 이용하여 지식을 계층적으로 표현하고, 연관관계를 통한 계층 간의 접근이 가능하도록 표현한 온톨로지 지식 표현이다.

이 두 개념은 두 가지 방법을 통해 결합한다.

1. MDR은 레거시 시스템의 접근정보를 관리하기 위해서 위치정보를 관리하고, 토픽맵은 표현된 지식에 대한 정보의 위치를 제공하기 위한 어커런스를 요구한다. 이 둘을 MDR의 메타로케이션으로 통합관리한다.

2. 토픽맵에서 사용하는 스키마 토픽은 모든 레거시 시스템에 적용할 수 있도록 되어야 한다. 그 이유는 어떤 레거시 시스템에서 요구하는 지식을 검색하기 위해 해당 레거시 만의 지식이 아닌 통합된 지식표현을 제공하고, 이 표준은 MDR을 통해서 확인할 수 있다. 즉, 토픽맵의 스키마 토픽은 MDR의 글로벌 스키마를 통해서 수행한다.

이러한 결합에 대한 개괄적인 개요는 그림 1에서 표현된 것과 같다. 토픽의 스키마 토픽은 글로벌 스키마를 이용하고 글로벌 스키마는 로컬스키마와 매핑 되어있다. 그리고 토픽맵에 대한 어커런스는 MDR의 메타로케이션을 어커런스 역할로 이용한다는 것을 보여준다.

4장. 질의처리

4.1. TMDR의 접근

TMDR의 접근은 검색을 위한 사용자의 필드 선택과 키워드 입력으로 발생한다. 스키마 정보에 해당하는 필드명과 해당 필드의 데이터를 이용하여 MDR을 접근한다. 필드명은 시스템 생성시 구성된 MDR의 표준항목이므로 토픽맵의 상위토픽에 해당한다. 해당 상위토픽을 가지는 하위토픽 중 사용자가 입력한 데이터를 검색하고, 검색된 결과를 토픽맵의 계층구조와 연관관계를 사용자에게 제공하여 정확한 요구사항을 선택할 수 있도록 지원한다. 토픽맵의 연관관계에 따른 질의의 확장도 추가된다.

그림 2는 사례를 이용한 질의과정이다. 질의과정을 요약하면 먼저, 표준항목에 해당하는 특정 키워드를 가지는 논문을 검색하기위해 사용자에게 토픽맵의 연관관계 정보를 제공한다. 이 과정은 키워드의 명확성을 높이기 위한 과정이다. 둘째, 선택된 연관관계를 통한 질의를 확장한다. 이때 생성된 질의는 초기 사용자의 키워드에 의한 질의문보다 복잡한 질의문으로 변환되지만, 검색결과의 정확성을 향상시킬 수 있다. 셋째, MDR을 이용하여 로컬에 전송하

1. 표준항목명 '키워드'를 선택하고 키워드 입력
2. '키워드'와 데이터를 통한 TMDR 검색
   - 기본 표준 질의 생성
3. 토픽맵의 계층 트리와 데이터의 연관관계를 통한 인터페이스 구성
4. 사용자의 연관관계 선택
   - 표준 질의의 확장1
5. 유사관계 질의에 확장
   - 표준 질의의 확장2
6. MDR과 ML을 통한 표준질의를 전송을 위한 질의로 변환
   - XML 쿼리
7. 로컬 질의 변환
   - 각 로컬 데이터베이스 스키마와 표준 스키마간의 매핑 정보 이용
8. 질의 수행

그림 2. 질의 변환과정

그림 2.의 질의 변환 과정에서 질의 변환을 수행한 사례는 그림 3과 같이 표현된다.

그림 3에서 과정에 따라 질의 변환 사례를 적용한 것으로 ①은 표준항목에 따라 사용자가 입력한 키워드를 통하여 생성된 표준 질의문이며, 로컬정보를 포함하지 않은 글로벌 질의 형태로 from문을 가지지 않는다. ②는 ①의 키워드 정보를 토픽맵의 연관관계에 따라 조건이 추가되어 복잡해진다. 토픽맵의 온톨로지 정보에 의해 수행된다. ③은 ②의 확장된 질의를 로컬의 스키마 구조에 따라 변환된 질의문이다. 이는 MDR을 통해서 수행된다.

이상의 질의 변환 방법에 따라 시스템에 수정없이 기존의 환경을 적응하고, 기존 키워드가 가지는 모호성을 해결할 수 있다.

5. 결론

본 논문에서는 TMDR을 기반 기술로 하여 분산된 데이터베이스 환경하에서 데이터를 상호운용할 수 있는 시스템을 제안했다. 클라이언트가 데이터베이스에 접근하여 데이터를 사용하기 위해 질의할 때, 질의문은 TMDR 내에 있는 MDR의 표준 속성을 이용하여 생성하고, 분산된 각 시스템내부를 제외하고는 모두 표준 속성으로만 접근할 수 있다.

TMDR은 로컬 데이터베이스를 통합하기 위한 스키마 매핑 저장소 역할을 수행하고, 토픽맵을 통한 지식 저장소 역할을 수행한다. 토픽맵의 연관관계 분석은 키워드의 모호성 해결을 지원한다.

본 시스템에서 제안하는 방식은 수평으로 분할된 데이터베이스들 사이의 상호운용뿐만 아니라 연관성을 있지만 서로 다른 데이터를 가진 수직 분할된 데이터베이스를 접근하기 위한 허브역할도 수행할 수 있다. 사례로 적용된 질의 변환도 수직 분할된 데이터를 접근하는 과정을 간략히 표현한 것이다.

그리고 데이터웨어하우스는 데이터를 추출하고 변형하

①	Select title, author, society, vol, page, dated WHERE keyword='metadata';
②	SELECT title, author, society, vol, page, dated WHERE keyword='metadata' AND position='kwu' AND dated BETWEEN '05' AND '09';
③	SELECT db1.name title, db2.author author, db1.society society, db1.volume vol, db1.pages page, db1.dated dated FROM paper db1, author db2, p2a db3 WHERE db1.keyword='metadata' AND db2.post='kwu' AND db1.aid=db3.aid AND db2.pid=db3.pid GROUP BY db1.name;

그림 3. 적용된 질의 기법에 따른 질의 변환 사례

고 전송하는 과정을 가진다. 이 과정을 ETL이라하는데. 이 ETL 과정은 효과적인 정제기법과 실시간 데이터 분석이 요구된다. 향후 본 질의변환 기법은 데이터웨어하우스의 정제기법 및 실시간 데이터웨어하우스 구축에 적용하도록 한다.

참고문헌

[1] [Joseph Hassell, Boanerges Aleman-Meza, I.Budak Arpinar, "Ontology-Driven Automatic Entity Disambiguation in Unstructured Text", 5th International Semantic Web Conference, Athens, GA 30602-7404, USA, 2006, pp. 44-57.

[2] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, Miroslav Goranov, "KIM - Semantic Annotation Platform", 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 834-849, Springer-Verlag Berlin Heidelberg 2003.

[3] Ray Gates, "Introduction to MDR-Tutorial on ISO/IEC 11179", Metadata Open Forum 2004, Xian, May 17, 2004.

[4] Hui Han, Lee Giles, Hongyuan Zha, "Two Supervised Learning Approaches for Name Disambiguation in Author Citations", JCDL'04, Tucson, Arizoin, USA, 2004. 6. 7. pp.296-305.

[5] 정계동, 황치곤, 신효영, 최영근, "TMDR 기반의 효율적인 키워드 검색 방안에 관한 연구", 제 32회 한국정보처리학회 추계학술발표대회 논문집 제16권 제2호, 2009. 11

[6] 정계동, 황치곤, TMDR 기반의 실시간 데이터 통합 환경 설계, 한국해양정보통신학회, Vol.13, No.9, 2009.