GSN 기반 DB통합 모델에서의 data value 이질성 해결 기법

홍종하[0] , 박성공 , 이정욱, 백두권
고려대학교 컴퓨터학과 소프트웨어 시스템 연구실
{sam, skpark, ljo, baik}@swsys2.korea.ac.kr

A Data Value Heterogeneity Solving Method In A GSN Based DataBase Integration Model

Jong-Ha Hong[0] , Sung-Kong Park , Jeong-Oog Lee , Doo-Kwon Baik
Software System Lab., Dept. of Computer Science & Engineering, Korea University

요 약

분산되고 이질적인 환경에서의 정보 소스들을 통합하려는 노력은 끊임 없이 계속되어 왔다. 이질적인 다중 정보소스로부터 추출된 정보를 통합하는 도구를 개발하는 것은 인터넷 기반에서 다양한 정보들을 실시간으로 사용할 수 있다는 측면에서 아주 흥미로운 일이다. 이러한 도구를 개발하는데 있어서의 주된 문제점은 서로 다른 정보소스에 존재하지만 실제적으로는 같은 실세계의 개념을 가지고 있는 정보를 어떻게 효과적으로 표현할 것인가 하는 것이다. 이러한 의미적 이질성을 해결하기 위해서 WordNet이나 Common Thesaurus 등을 이용한 개념 기반의 접근방법이 많이 제안되었다. 하지만 이들은 스키마 이질성을 해결하는 방법을 제시 할 뿐, 데이터의 이질성을 해결 하는 방법은 보여주지 않는다.

본 논문에서는 GSN(Global Semantic Network)을 이용해서 스키마 이질성을 해결하는 데이터베이스 시스템에서 발생하는 데이터 이질성의 예를 제시하고 이러한 데이터 이질성을 해결할 수 있는 기법을 제안한다.

1. 서론

이질적인 다중 정보소스로부터 추출된 정보를 통합하는 지능적인 도구를 개발하는 것은 인터넷 기반에서 다양한 정보들을 실시간으로 사용할 수 있다는 측면에서 아주 흥미로운 일이다. 이러한 도구를 개발하는데 있어서의 주된 문제점은 의미적으로 연관된 정보들의 분류와 연결이다. 다시 말해서, 서로 다른 정보소스에 존재하지만 실제적으로는 같은 실세계의 개념을 가지고 있는 정보를 어떻게 효과적으로 표현할 것인가 하는 것이다. 전역적인 정보 시스템에서 사용 가능한 정보 소스들은 이미 존재하는 것들로 서로 다른 위치정보를 가지며 독립적으로 개발된 것들이다. 이러한 측면에서 볼 때 용어, 구조, 정보의 의미 등과 관련된 의미적 이질성의 문제가 발생하게 되며 이러한 정보들을 효과적으로 사용할 수 있도록 하기 위한 적절한 방법이 요구된다.

이질적인 정보의 통합은 데이터베이스 분야에서 아주 중요한 문제이다[3]. 최근에는 이러한 이질적인 데이터베이스에서의 정보 통합 중에서도 특히 의미적인 통합과 관련된 연구들이 많이 진행되고 있다. 의미 이질성(semantic heterogeneity)은

스키마 이질성(schema heterogeneity)과 데이터 이질성(data heterogeneity)으로 나눌 수 있는데,

본 논문에서는 GSN(Global Semantic Network)을 기반으로 스키마 이질성을 해결하는 데이터베이스 시스템에서 발생하는 데이터 이질성의 예를 제시하고 이러한 데이터 이질성을 해결할 수 있는 기법을 제안한다.

2. 관련연구

2.1 WordNet[4]

WordNet은 영어 어휘 지식을 모델링하기 위하여 시도된 프린스턴 대학의 연구 프로젝트의 산물이다. 이 시스템은 온라인 시소러스와 온라인 사전의 기능과 더불어 그 이상의 기능을 가지고 있다. 영어의 명사, 동사, 그리고 형용사 등은 동의어 집합(synonym set;synset)으로 구성되고 이들 동의어 집합 각각은 하나의 어휘 개념을 표현한다.

WordNet은 이러한 synset 이라고 불리 우는 논리적 그룹으로 구성되며 각 synset은 같은 뜻의 단어형태와 현재

synset과 다른 synset들간의 관계성을 가리키는 의미 포인터(semantic pointer)로 구성된다. 의미 포인터는 synonymy, antonymy, hyponymy, 그리고 meronymy를 포함하여 여러 형태가 있다. WordNet은 현재까지 95,600 개의 서로 다른 단어 형태(word form)를 포함하며, 이들은 70,100 개의 단어 의미(word meaning) 또는 동의어 집합으로 구성되어 있다. WordNet의 가장 주된 특징은 어휘 정보를 단어 형태가 아닌 단어 의미에 의해 구성하고 있다는 것이다.

2.2. Classic Approach[6,9]

이 방법은 상이한 다중 정보 소스들을 포함하는 하나의 전역 스키마를 구축하는 것이다. 전역 스키마에서 지역 스키마로의 매핑은 HOSQL, SQL/M과 같은 언어들을 이용하여 표현된다. 이 방법의 초점은 관계 데이터베이스와 객체지향 데이터베이스의 통합 시에 발생하는 스키마 이질성, 구조적 이질성의 해결에 맞추어져 있다. 그러므로 의미적 이질성을 해결하는 데는 한계성을 가지고 있다. 또 이 방법은 스키마가 통합 과정 초기에 정적으로 생성 된다는 점

에서 이질적이고 동적인 다양한 사용자의 요구를 지원하지 못한다는 단점도 가지고 있다.

2.3 Federated Approach[8]

이 방법은 이전의 방법에서 이질적인 사용자의 요구를 다루는 방법을 개선한 것으로 다중 통합 스키마에 기반 한다. 이 방법은 이질성의 문제점들이 스키마 통합 단계에서 해결되므로 정적인 통합 스키마를 가진다.
그러므로 새로운 사용자의 요구나 정보 소스들이 추가되거나 제거될 필요가 있을 때 확장성의 문제를 발생시킨다.

2.4 Concept-based Approach[2,4]

이 방법은 DB 통합 시에 발생하는 의미의 이질성을 해결하기 위해서 Common Thesaurus나 WordNet과 같은 어휘 정보 시스템을 사용하여 의미는 같고 표현 형태가 다른 엔티티들을 통합시켜 준다. 이들 어휘 정보 시스템에는 어휘들간의 상호 관계가 동의어(synonymy), 상위어(hypernymy) 등의 형태로 관계성이 표현되기 스키마 레벨에서의 엔티티들에 대한 의미적 통합을 위한 reference로 사용된다.

3. Data value 이질성 해결 기법

3.1 데이터 값의 이질성

데이터베이스는 스키마와 데이터에 의해 정의되기 때문에 의미 이질성은 스키마 이질성과 데이터 이질성으로 분류될 수 있다. 데이터의 이질성은 스키마 이질성이 해결된 상태에서 존재하는 데이터 포맷의 이질성 등으로 인해 발생한다.

Client table

name	sex	ssn	job	nation	weight
박성공	남	1223809	학생	KOR	70
권도훈	여	2138220	모델	KOR	65
김종환	남	1214990	회사원	KOR	55

Request table

oid	ssn	Pname	date	Day
1	1223809	책	06/05	월
2	2138220	장난감	07/21	수

(CDB 1)

Customer table

UP_ID	Name	Address	Job	weight
1630233	Jong H. H	Kimpo	회사원	140
1224539	Park S.K	Socho	대학생	154
2138220	Kwon D.h	Dukso	모델	143

Order table

O_Num	Customer_ID	Product	weekday
1	2138220	shampoo	1
2	2138220	lipstic	2

(CDB 2)

<표 1> 동일 도메인에 속하는 두 개의 CDB

본 논문에서는 데이터 값의 이질성 해결에 초점을 맞추고 있기 때문에 스키마적 이질성은 해결이 된 상태라고 가정한다. 스키마 이질성의 해결은 개념 기반 의미망[2]에 기초한다. 데이터 값의 이질성으로 인해 발생할 수 있는 문제점을 본 논문에서는 크게 5가지로 구분했다.

1> 서로 다른 문자셋(character set)으로 인한 이질성

이러한 충돌은 테이블 내의 동일한 속성에 대해서 서로 다른 문자 셋을 사용하였을 경우에 일어난다. 예를 들어 그림[1]에서 CDB 1의 이름 필드와 CDB 2의 Name 필드는 같은 이름(박성공)에 대해서 각각 한글(박성공)과 영어(Park.S.K)로 표현하고 있다.

2> 의미는 같고 표현이 다른 데이터 값에 대한 이질성

이러한 충돌은 데이터 값 간에 의미적 계층 구조가 존재하는 경우에 발생한다. 직업이 "학생"인 모든 사람의 이름을 반환하는 질의에 대해서 CDB1의 job 속성의 학생과 CDB2의 Job 속성의 대학생은 서로 동일한 의미 관계에 놓이게 된다.

3> numeric vs string(eq. Meaning)이질성

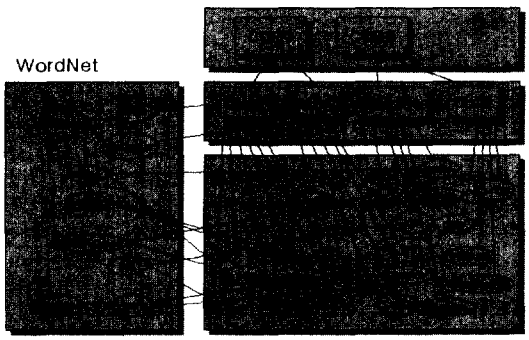
이러한 충돌은 numeric value와 string value가 동일하거나 유사한 의미로 사용될 경우에 발생한다. 실세계의 DB에서 이러한 현상은 자주 발생한다. CDB1의 day 속성 값 "월","수" 는 CDB2의 weekday의 1, 2,.. 등의 값과 매핑 된다. Grade{A,B,C,...}나 grade{90,80,70,...}도 비슷한 예라 할 수 있겠다.

4> 자율적 표현상의 이질성(esp. in english)

CBD2의 Park S.K라는 사람의 이름이 CBD3에서는 Song-Kong Park로 표현될 경우 이러한 충돌이 발생한다.

5> 단위의 이질성

이러한 충돌은 흔히 서로 다른 도량형을 사용할 경우 발생하는 문제이다. CBD1의 weight 속성에서는 kg을 도량형 단위로 사용하였고 CDB2의 weight 속성에서는 파운드(lb)가 단위로 사용되었다.



<그림 1> GSN에서의 두 데이터 베이스 통합 모습

<그림 1>은 <표 1>의 두 CDB를 GSN을 이용하여 스키마 이질성을 해결한 모습을 보여준다. 두 CDB의 통합 정보는 시스템에서는 실제로 매핑 테이블의 형태를 가진다. 데이터 값의 이질성은 value사이의 concept hierarchy, functional dependency 정보 등을 이러한 매핑 테이블에 추가함으로써 해결 할 수 있다.

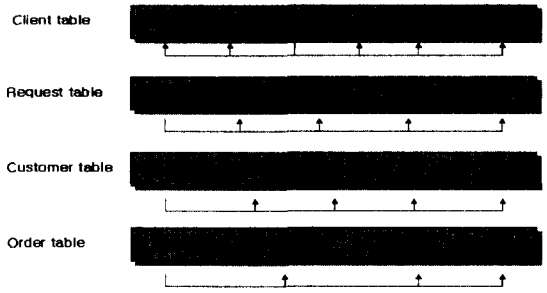
3.2 이질성 해결 단계

3.2.1 테이블 필드의 특성 추출과 매핑 테이블 작성

① numeric vs string value conflict를 가지는 필드에

대해 매핑 테이블 구성(예:entity1.attribute1.월= entity2.attribute1.1)
② 테이블에서의 PK와 Candidate Key(후보키)를 검색한다. Candidate Key는 테이블 내에서 unique한 값들만을 가지는 필드로, (unique values/total value)가 threshold 내에 존재하면 Candidate Key로 선정한다. functional dependency를 테이블로 표현
③ value내에 존재 할 수 있는 concept hierarchy를 구성하기 위해 string 타입필드에 대해서 의미적 연결 관계를 구성
④ 단위 이질성이 발생하는 필드에 대해 단위 매핑을 테이블을 구성(예:entity1.attribute1=entity2.attribute1*1.2)

<그림 2>는 <표 1>의 테이블에 Primary Key를 중심으로 나타나는 필드간의 functional dependency를 보여준다.



<그림 2>

<표 2>는 그림에 나타나 있는 functional dependency를 테이블의 형태로 나타낸 것이다. Cadkey는 후보키로 Pkey와 함께 필드 내에서 유일한 데이터 값들만을 가지고 있는 필드의 이름이다. 질의 처리 과정에서 값의 이질성을 해결하는데 <표 2>의 테이블이 참조된다.

Concept	Entity	Attribute	Pkey	Cadkey
name	Client	name	ssn	..
name	Customer	name	UP_ID	..
job	Client	job	ssn	..
Job	Customer	job	UP_ID	..
weekday	Request	day	oid	..
weekday	Order	weekday	O_Num	..
Weight	Client	weight	ssn	..
weight	Customer	weight	UP_ID	..

<표 2> 테이블에 표현된 Functional Dependency

3.2.2 이질성 해결

다음은 위에 나열한 이질성 문제에 대한 해결 방법이다.

1> 질의가 요청되었을 때 where [concept 조건]절에 해당하는 결과 값과 PK 값, CadK 값을 가져온다. 해당 PK값이 속한 테이블에 동일한 concept이 존재하는지 검사한다. 동일한 concept이 존재하면 이에 해당되는 나머지 select절의 동일 필드까지 가져온다.

2> value들 사이의 concept hierarchy에 대해 만든 매핑 테이블을 참조하여 질의를 수정한다.

3> 조건 값을 value 매핑 테이블의 하나로 통일하여 질의

4> 1>과 동일.

5> 매핑 테이블을 이용하여 결과에 대한 단위 변환

4. 결론

본 논문에서는 테이블 내의 데이터 값들에 대한 특성을 조사하여 이러한 정보를 바탕으로 데이터베이스 통합 시에 발생할 수 있는 데이터 값의 이질성을 해결하는 한 가지 방법을 제안하였다.
정보 통합 분야에서 이러한 의미적 통합을 이루려는 노력은 지금까지 많이 진행되어 왔다. 하지만 서로 이질적인 정보 소스의 완전한 의미적 통합을 이루기란 아주 어렵다. 본 논문에서 제안한 방법은 테이블에 존재하는 functional dependency, 이질적인 단위체계 매핑 등 비교적 간단한 방법을 이용하여 값에 대한 정보 통합을 하였다. 이렇게 해서 박성공, park s.h 와 같은 서로 다른 문자 코드를 가지고 있는 값에 대해서도 동일한 질의가 적용된다 이 방법의 단점은 도메인이 제한 되어 있고 전문가나 DB 설계자의 노력이 어느 정도 포함되어야 한다는 것이다. 자동화가 되지 못했다는 점도 또한 문제점으로 남겨져 있다. 현재는 제안된 도메인에 대해서만 성능 실험이 수행된 상태이다.
Mediator 를 이용한 자동화, 일반화된 룰(rule) 구축, 정보 추가 또는 삭제시의 효율성 증대 방안 등이 추후 연구 과제로 남아 있고 현재 연구 중이다.

참고 문헌

[1] Maurizio Panti, Luca Spalazzi, Alberto Giretti. A Case-Based Approach to Information Integration. Proceedings of the 26th VLDB Converence, 2000
[2] Sonia Bergamaschi, Silvana Castano, Maurizio Vincini. Semantic integration of heterogeneous information sources. DKEE, 2001
[3] R.Hull, Managing semantic heterogeneity in databases:A theoretical perspective, in: ACM Symposium on Principles of DB system, 1997
[4] 이정욱, 백두권. 멀티데이터베이스 시스템의 정보 공유를 위한 개념 기반 의미망. 2001
[5] Weiyi Meng, Clement Yu. Query Processing in Multidatabase Systems. Modern Database Systems. p551-p569
[6] Won Kim, Injun Choi, Sunit Gala, Mark Scheevel. On Resolving Schematic Heterogeneity in Multidatabase Systems. Distributed and Parallel Databases, 1993
[7] William Kelley, Sunit Gala, Won Kim, Bruce Graham. Schema Architecture of the UniSQL/M Multidatabase System. Modern Database Systems, p621-p646
[8] A. Sheth and J. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Transaction on Database Systems, 1990
[9] Ahmed and et al. The Pegasus heterogeneous multidatabase system. IEEE Computer, 1991