계층적 메타데이터 레지스트리 기반의 점진적 데이터 통합

신동길[0] 정동원 백두권
고려대학교
{dkshin[0], withimp, baik}@software.korea.ac.kr

Progressive Data Integration based on Hierarchical Metadata Registry

Dongkil Shin[0] Dongwon Jeong Dookwon Baik
Korea University

요 약

오랜 기간동안 메타데이터를 기반으로 한 데이터통합에 대한 많은 연구들이 진행되어 왔다. 그러나 기존 방법론들은 전역 뷰 또는 전역 스키마와 같은 초기 가이드라인을 구축하는데 많은 비용이 요구 된다는 단점이 있다. 이는 기존 연구들이 해당 도메인 특성들을 간과했기 때문이다. 예를 들어 과학 데이터의 경우 일반사용자들은 생물의 이름이나 모양 등과 같은 단순정보에 관심을 갖는 반면 과학자나 전문가들은 보다 상세하고 전문적인 데이터에 관심을 갖는다. 추가적으로 모든 데이터에 대한 초기 표준 가이드라인을 구축하는 것은 현실적으로 많은 어려움이 따른다. 본 논문에서는 이러한 도메인 특성을 고려하여 점진적인 통합방법론(LOG : LOcalization-based Global metadata registry)을 제안한다.

1. 서 론

최근 메타데이터를 기반으로 한 많은 연구들이 진행되고 왔으나, 실제 도메인에 적용하고자 할 경우 해결해야 하는 몇 가지 문제점을 지닌다. 가장 중요한 문제점은 기존의 연구들이 도메인의 특성을 간과해왔다는 점이다. 즉, 그 동안의 많은 연구 노력들이 데이터 통합을 위한 기술적인 요소에만 집중되어 왔을 뿐, 데이터 레벨, 사용 레벨 및 데이터량 등과 같은 도메인 특성을 고려하지 않았다는 것이다. 실질적으로, 이러한 특성들은 실제 데이터 통합을 위해 고려되어야만 한다.

기존의 데이터 통합 접근 방법들은 하위 데이터베이스 설계 및 분석을 통해 가이드라인을 생성하기 위한 필수적인 과정을 수행하며, 이 과정은 주어진 데이터베이스 내의 모든 데이터와 스키마를 분석하여 하나의 가이드라인을 정의하기 때문에 많은 비용이 요구된다.

그러나 일반적으로 모든 사용자들이 모든 데이터에 관심을 지니지는 않는다. 비전문적인 사용자들은 단지 이해하기 쉽고 일반적인 데이터에 관심을 지니고 전문가들은 보다 전문적인 데이터에 관심을 지닌다. 그러므로 초기 통합 단계에서 모든 데이터를 대상으로 하나의 가이드라인을 생성할 필요가 없다. 초기에는 가장 일반적인 데이터들을 통합하기 위해 필요한 최소의 전역 가이드라인을 구축한다. 그리고 나서 도메인특성에 따라서 가이드라인을 점진적으로 확장해 나가는 것이다.

이 논문에서는 도메인 특성을 반영하는 LOG (LOcalization-based Global metadata registry) 방법론을 제안한다. LOG 방법론은 점진적인 통합 메커니즘을 제공한다.

2. 관련연구

가장 일반적인 데이터 통합 방법은 온톨로지 기반 접근 방법이다. 온톨로지 기반 접근 방법은 모든 데이터베이스에 대한 분석과 설계 과정을 수행하고 난 후, 하나의 공통된 스키마를 추출하여 정의하여 한다.

가장 전통적인 접근 방법으로서, 많은 지역 스키마들간 이질적인 특징을 지니는 하나의 전역 스키마를 구축하여 이를 각 지역 데이터를 액세스 하는데 이용한다[1,2,3,4]. 두 번째 접근 방법은 연방 접근 방법으로서, 다중 통합 스키마를 이용하여 단일 전역 스키마로 나아가면서 지역 데이터베이스를 액세스한다[6]. 추가적으로, 분산 객체 접근 방법[7,8,9], 중계기 기반 접근 방법[10,11], 추론 기반 접근 방법[5] 등이 있다.

이러한 접근 방법들은 온톨로지 기반의 통합 방법론으로서 분류할 수 있다. 그러나 이들 방법들은 관리에 많은 비용이 소요되는데, 이는 하위 레벨에 있는 임의의 데이터베이스가 수정될 때마다 계속적으로 기존 온톨로지가 변경되어야만 하기 때문이다. 또한 통합의 처음 단계에서 모든 데이터들간 사상을 위한 단일의 통합 스키마를 필요로 하기 때문에 많은 초기 비용이 요구된다.

앞서 언급한 문제를 해결하기 위한 다른 접근 방법이하향식 접근법이다. 먼저, 표준화 된 가이드라인이 생성되고, 데이터베이스 개발자들은 제공된 가이드라인에 따라 데이터베이스를 구축하게 된다. 따라서 온톨로지 접근 방법과 같이 가이드라인 정의를 위한 분석 비용은 물론 이를 관리하는데 소요되는 비용을 절감할 수 있다. 지금까지 MARC[13], Dublin Core[14], ONIX[15], RDF[16] 등의 많은 가이드라인이 연구 개발되었다.

그러나 이러한 가이드라인들은 동적인 메타데이터 관리 방법을 제공하지 않는다. 이 문제점을 해결하기 위해 ISO와 IEC는 정보 기술분야 공동 기술 위원회인 ISO/IEC JTC 1을 구성하였다. 특히, ISO/IEC JTC 1은 전통적인 데이터 관리 방법론들이 보다 더 적은 시간과 노력으로 공유 데이터 환경을 생성할 수 있도록 하기 위하여 ISO/IEC 11179를 개발하였다[16,17].

실세계의 모든 사물들은 그들을 인식할 수 있는 이름과 같은 식별자, 색상 및 크기와 같은 특성들을 지닌다. 이러한 특성들은 데이터로 표현된다. 즉, 데이터는 자동적인 방법이나 인간에 의해 통신, 해석 및 처리가 적합하도록 형식화된 방법으로 사실, 개념 및 명령어 등을 표현한 것이다.

ISO/IEC 11179는 데이터를 쉽게 이해하고 공유할 수 있도록 하기 위해 데이터 요소를 표준화하고 등록하는 방법을 기술한다. 이는 데이터베이스들간의 공유 및 교환을 극대화하고,

표준방식으로 일관성 있게 데이터베이스를 표현하고 구축하기 위한 것이다. ISO/IEC 11179는 데이터 요소라는 중요한 개념을 지니고 있으며, 이는 정의, 식별, 표현 및 허용 가능한 값들을 속성 집합으로서 명세하기 위한 데이터 단위이다.

ISO/IEC 11179에서 제공하는 메타데이터 레지스트리 개념이 많은 장점을 지니고 있지만, 대부분의 실제 응용에 있어서, 기존 데이터베이스들의 통합을 위한 표준 가이드라인 정의가 필수적으로 요구된다. 이는 온톨로지 기반 접근 방법과 같이 많은 비용이 요구된다.

따라서 본 논문에서 비용과 활용성을 고려하여, 메타데이터 레지스트리 기반의 점진적인 데이터 통합 방법론을 제안한다.

3. 데이터 가시성

3.1 도메인 특성

기존 연구들은 도메인 특성을 고려하지 않고 단지 기술 개발에만 집중되어 왔다. 즉, 컴퓨터 과학 분야에 있는 대부분의 연구자들이 도메인 특성을 고려하여 방법론이나 시스템을 개발하는데 반영하고자 노력하지 않았다.

도메인의 특성은 데이터 레벨, 사용자 레벨, 데이터 이용성 등을 의미한다. 이 논문에서는, 데이터 이용성이 데이터 레벨과 사용자 레벨에 의존하기 때문에 이들과 데이터량과의 관계성에 초점을 둔다. 데이터 레벨은 데이터의 전문화 정도로서, 데이터가 전문화될수록 보다 더 상세하고 복잡하게 된다. 사용자 레벨은 해당 분야에대한 사용자들의 전문지식 정도를 의미한다. 사용자들은 전문지식 정도에 따라서 크게 두 그룹으로 분류된다.

첫번째 그룹은 전문가들로 구성되며, 다른 그룹은 일반 사용자들로 구성된다. 사실상, 일반 사용자들은 제한된 일반 정보에 관심을 지닌다. 반면, 일반적인 데이터를 포함하여 전문가들은 보다 깊이 있고 복잡한 데이터에 관심을 지닌다. 사용자 레벨과 데이터량과의 관계에서, 전문가들은 대부분의 데이터를 이용하기 때문에 일반 사용자들보다 많은 데이터를 이용한다. 결론적으로, 데이터량은 다른 도메인 특성들간의 관계성에 의해 계층적으로 그룹화될 수 있다. 이 논문에서는 이러한 개념을 데이터 가시성이라고 정의한다.

그림 1은 데이터 레벨, 사용자 레벨 및 데이터량과의 관계성을 나타내는 데이터 가시성(Data Visibility)의 개념을 보여준다.

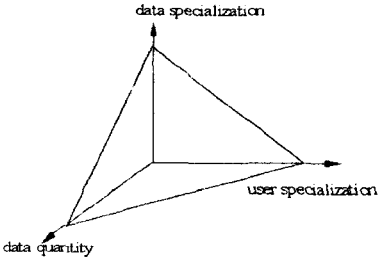

그림 1 데이터가시성의 개념적 다이어그램

위 그림에서, 데이터 전문성과 사용자 전문성은 각각 데이터 레벨과 사용자 레벨을 의미한다. 그림을 통해 알 수 있듯이, 보다 전문적인 사용자들, 즉 도메인 전문가들이 일반 사용자들보다 많은 데이터를 이용한다. 또한 보다 더 전문화된 데이터에 관심을 갖는다.

3.2 메타데이터 레지스트리와 데이터 가시성

이미 앞 절에서 기술하였듯이, 데이터의 집합은 데이터 가시성에 의해 계층적인 다수의 관점으로 분류된다. 따라서 메타데이터 레지스트리를 계층적으로 구축할 수 있다. 또한 상향식 접근 방법으로 최상위 레벨의 메타데이터 레지스트리를 점진적으로 확장해 나갈 수 있다. 그림 2은 데이터 가시성과 메타데이터 레지스트리와의 관계성을 보여준다.

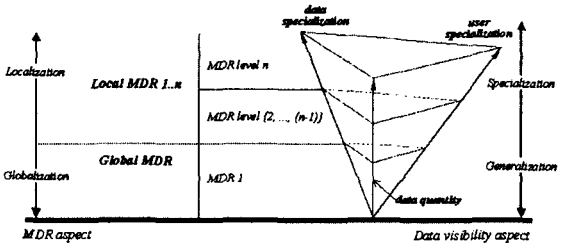

그림 2 데이터가시성을 이용한 계층적 MDR

데이터 레벨과 사용자 레벨, 즉 데이터 전문화와 사용자 전문화 정도에 따라 모든 데이터들은 계층적으로 형성된다. 이에 따라서, 메타데이터 레지스트리는 다수의 레이어를 지님으로써 계층적으로 구축된다. 그림에서, 전역 MDR은 모든 데이터 집합들에 대해 공통의표준적인 데이터 명세를 포함하는 전역 가이드라인으로서 이용된다. 지역 MDR은 일부 데이터집합에 대한 개별적인 공통 데이터 명세방법을 지니게 된다. 결과적으로, 전역 MDR은 해당 데이터집합의 가시성이 가장 높기 때문에 가장 우선적으로 생성되어 한다.

4. 데이터 통합 모델

4.1 점진적 통합을 위한 개념 모델

LOG 방법론의 개념 모델은 인터페이스 계층, 전역 MDR 계층, 지역 MDR 계층 및 데이터 자원 계층으로 구성된다. 이러한 분류는 이미 기술한 데이터 가시성에 기반을 두고 있다. 그림 3은 LOG 방법론을 개념적으로 보여준다.

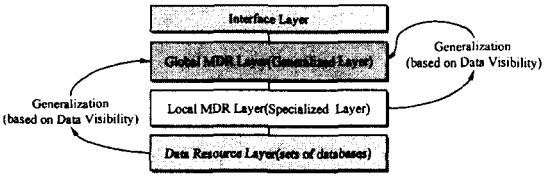

그림 3 LOG 방법론의 개념적인 모델

그림 3에서, 가장 상위 계층이 인터페이스 계층이며 지역 인터페이스 계층과 전역 인터페이스 계층으로 구성된다. 인터페이스 계층은 검색, 질의 및 뷰잉과 같은 서비스들을 제공한다. 사용자들은 이 계층을 통해 질의를 보내고 질의 결과를 얻을 수 있다.

두 번째 계층이 가장 일반화된 메타데이터 레지스트리를 관리하는 전역 MDR 계층이다. 세번째 계층인 지역 MDR 계층은 보다 복잡하고 자세한 데이터를 이용하는 전문가를 지원한다. 이 계층은 지역 메타데이터 레지스트리와 지역 저장소들을 관리한다. 마지막 계층은 데이터 자원 계층으로서, 각 도메인 전문가 또는 데이터베이스 관리자에 의해 각각 독립적으로 생성된 데이터베이스 집합을 포함한다.

새로운 표준 데이터 즉, 데이터 요소가 일반화될 때마다 가이드라인들을 확장하기 위하여 메타데이터 레지스트리들이 수

정되도록 해야만 한다. 즉, 모든 데이터의 점진적인 통합을 위하여 메타데이터 레지스트리에 변화된 상황을 반영하여야만 한다. LOG 방법론은 기존 메타데이터 레지스트리를 점차적으로 확장할 수 있는 메커니즘을 제공한다.

4.2 시스템 아키텍처

이 절에서 LOG 방법론을 위한 시스템 아키텍처의 각 계층별 주요구성요소를 기술하며 그림 4는 전체 시스템 아키텍처를 보여준다.

첫 번째 계층은 검색, 뷰잉 등과 같은 서비스를 제공하는 인터페이스 계층이다. 이 계층은 서브 계층인 전역 사용자 인터페이스와 지역 유저 인터페이스 계층으로 구성된다. 서브 계층인 전역 사용자 인터페이스의 역할은 지역 사용자 인터페이스의 기능과 유사하다. 그러나 전역 사용자 인터페이스는 전문가를 위해 서비스를 제공하고, 일반 사용자는 지역 사용자 인터페이스를 통해 서비스를 받는다

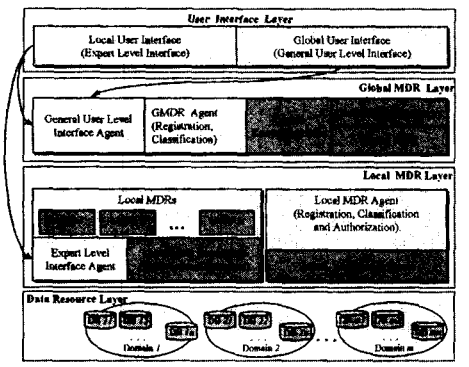

그림 4 LOG 방법론기반 시스템 아키텍처

전역 MDR 계층은 4개의 컴포넌트로 구성된다. 이 계층에서 가장 중요한 컴포넌트는GMDR(Global MDR) 에이전트이다. GMDR 에이전트는 GMDR(Global Metadata Registry)와 GMeta 레파지토리를 관리하고 제어한다. 즉, 데이터의 변경에 따라 GMDR과 GMeta 레파지토리를 변경시킨다. GMDR과 GMeta 레파지토리는 각각 전역 데이터 요소와 실제 메타데이터의 집합이다. 첫번째 단계로 새로운 데이터베이스 생성시 데이터베이스 관리자나 생성자는 적당한 가이드라인을 찾기위해 GMDR을 조사한다. 만약 적당한 가이드라인이 존재한다면 메타데이터 레지스트리 혹은 데이터베이스 생성자는 가이드라인에 따라 스키마를 생성해야만 한다. 그렇지 않다면 사용자 정의 데이터 요소를 이용하여 유사한 도메인과 관련있는 GMDR 혹은 LDMR(Local MDR)에 메타데이터를 등록한다. 등록된 데이터 요소는 GMDR 에이전트나 적당한 LMDR 에이전트에 의해 평가, 분류, 권한부여, 일반화된다.

지역 MDR 계층은 네 개의 컴포넌트로 구성되어 있다. 컴포넌트의 역할은 전역 MDR의 컴포넌트와 유사한 기능을 가진다. 그러나 이 계층에는 논리적으로 일반화된 데이터 요소를 관리하는 LMDR이 있다. LMDR은 계층적으로 구성될 수 있으며 전문가를 여러 그룹으로 분류하고 그룹에서 각 LMDR이 각각 관리될 수 있다. 이런 메커니즘은 데이터의 지역성을 증가시키며 앞에서 언급한 것처럼 많은 이점을 갖고 있다. 또한 향후, 전역 MDR 계층에서 GMDR이 그들을 관리하는 공통 전역 데이터 요소를 일반화시킬 수 있다. 결과적으로 점진적으로 통합범위를 증가시키고 GMDR을 확장시킬 수 있다

마지막 계층은 가장 낮은 레벨에서 데이터베이스 집합을 포함하는 데이터 자원 계층이다. 데이터베이스는 동일한 도메인으로 분류될 수 있고 독립적으로 관리된다.

5. 결론

본 논문에서는 점진적으로 데이터를 통합함으로써 기존 데이터 통합 접근의 문제를 극복하기 위한 방법론을 제시하였다. 이 방법론은 ISO/IEC 11179의 MDR에 기반을 두었으며 사용자 레벨 및 데이터 레벨과 같은 도메인을 반영하는 점진적인 통합 프레임워크를 제공한다. LOG 모델은 사용자 인터페이스 계층, 지역 MDR 계층, 전역 MDR 계층과 데이터 자원 계층으로 구성된다.

향후, 각 컴포넌트에 대한 정확하고 체계적인 설계를 통하여 LOG 모델 프레임워크를 상세히 설계하고 실제 도메인에 적용할 것이며 마지막으로 이 방법론의 일반화 알고리즘이 요구된다.

참고문헌

1. Ram, S., Special issue on heterogeneous distributed database systems, *IEEE Computer Magazine*, 24, 12 (December 1991).
2. Sheth, A., Special Issue in Multidatabase Systems, *ACM SIGMOD Record on Management of Data*, 20, 4 (December 1991).
3. Ahmed and et al, The Pegasus heterogeneous multidatabase system, *IEEE Computer*, 24, 12 (1991).
4. Kim, W. and et al. On Resolving Semantic Heterogeneity in Multidatabase Systems, *Distributed and Parallel Databases*, 1, 3 (1993).
5. Panti, M., Spalazzi, L., Giretti, A, A Case Based Approach to Information Integration, *Proceedings of the 26th VLDB Conference* (2000).
6. Sheth, A., Larson, J., Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, ACM Transaction on Database Systems, Vol. 22, No. 3, 1990.
7. Manola, F. and et al, Distributed object management, *International Journal ,of Intelligent and Cooperative Information Systems*, 1, 1 (March 1992).
8. Ozsu, T., Dayal, U. and Valduriez, P., Distributed Object Management, *Morgan Kaufmann*, San Mateo, CA (1993).
9. Cattell, R.G.G., The Object Database Standard ODMG-93, *Morgan kaufmann*, San Mateo, CA (1993).
10. Molina, H.G., Papakonstantinou, Y., Guass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., Widom, J., The TSIMMIS approach to mediation: data models and languages, *Journal of Intelligent Information Systems*, 8 (1997), 117-132.
11. Wiederhold, G., Mediators in the Architecture of Future Information Systems, *IEEE Computer Magazine*, 25 (March 1992), 38-49.
12. Uschold, M., Gruninger, M., ONTOLOGIES: Principles, Methods and Applications, *Knowledge Engineering Review*, 11, 2 (June 1996).
13. http://www.loc.gov/marc/
14. http://dublincore.org/
15. http://www.editeur.org/
16. http://www.w3.org/RDF/
17. http://www.jtc1sc32.org/
18. ISO/IEC JTC1 SC32, ISO/IEC 11179: Specification and standardization of data elements, *ISO/IEC JTC 1*, Part 1~6
19. Song, C.Y., Yim, S.B., Moon, C. J. and Baik, D.K., Design and Implementation of a Component Registry Using XML, *In Proceedings of the International Symposium on Future Software Technology*, Guiyang, China, 2000.