단계적 데이터 품질 모델링 방법론과 스키마

나관상°, 백두권
한국통신 연구개발본부, 고려대학교 컴퓨터학과
e-mail: ksna@kt.co.kr, baik@swsys2.korea.ac.kr

A schema and stepwise methodology for modeling the data quality

Kwan-Sang Na°, Doo-Kwon Baik
R&D Group, Korea Telecom / Dept. of Computer Science & Engineering, Korea University

요 약

고객에게 원하는 정보를 제공하기 위해서는 데이터의 설계, 개발 및 이용에 있어 최적화된 데이터의 모델링 및 구조화가 매우 중요하며, 이를 통해 사용자에게 적기에 고품질의 데이터를 제공하는 것이 무한 경쟁시대에서 생존을 위한 핵심 요소이다. 특히, 우리는 인터넷의 출현으로 오프라인 기업에서 온라인 기업으로 급속한 전환과 기업간, 기업과 고객간, 기업과 정부간 보다 넓게는 전세계의 국가를 하나로 엮는 정보유통 시대에 살고 있다. 인터넷 상거래의 활성화와 전자정부 구현 등에서 기업 생존의 핵심 요소는 방대한 양의 데이터를 어떻게 공유하고 유통시키며, 양질의 데이터를 구축 하느냐 이다. 본 고에서는 기존 시스템의 컨버젼이나 마이그레이션 또는 이질적 시스템 통합과정에서 그리고 데이터베이스 설계과정에서 데이터의 품질을 향상시키기 위해 필요한 데이터 품질문제를 알아보고, 체계적으로 데이터 품질을 추출 및 표현하기 위한 방법론을 제안하며, 이를 개선 발전시키기 위한 연구방향을 소개한다.

1. 서 론

음성중심의 통신시대에서 멀티미디어 통신시대로 급속히 변화하고 있고, 미래에는 VR 기반 사이버 통신이 중심이 되는 사회로 전환될 것이다. 특히 인터넷이 비즈니스의 중심축으로 등장한 현 시대에서 기업의 생존력은 의사결정, 고객관리, 내부 혁신 등을 위한 데이터를 어떻게 하면 정확하고 완전한 양질의 데이터를 구축, 유통시키고 이를 유지 및 관리하느냐가 핵심 요소로 부각되고 있다. 특히, e 비즈니스의 급속한 성장으로 인해 대다수의 기업들이 데이터 과부하와 데이터 관리의 어려움을 경험하고 있으며, 이로 인한 데이터 품질의 관리는 더욱 어려워지고 있는 현실이다. 즉, 과거에는 운영자로부터 모든 데이터를 추출함에 따른 데이터 확보의 지연과 부정확한 데이터 존재로 대 고객서비스 품질과 투자정책간 연계 부재가 기업성장의 한계로 작용되었다. 현재는 Online 으로 정확하고 완전한 고품질의 데이터를 실시간으로 접근, 확인 할 수 있어야 하며, 이에 따른 정보의 공유에 의한 지식경영으로 기업경쟁력 뿐만 아니라 국가 경쟁력 향상의 핵심 기술 기반으로 데이터 품질향상이 중요하게 부각되고 있다.

실제로 1999 년 3/4 분기에 Information Week 에서 300 명의 IT 분야 최고경영자를 대상으로 2000 년 이후의 기술 우선순위 조사결과에 의하면 고객의 데이터 품질향상이 가장 중요한 요소(82%)로 나타나 있다[1]. 예를 들면 Telcordia (구 벨코어)의 경우 데이터를 회사의 귀중한 자산으로 분류하여 빌딩이나 장비와 같은 자산으로 관리하고 있다. 특히, 고품질의 데이터를 기업전체에서 공유하기 위해 엄격한 표준에 따라 관리하고 시스템에 독립적으로 데이터를 구축하여 데이터가 효율적으로 저장되고 검색될 수 있도록 운영하고 있다. 본 논문에선 데이터의 품질 문제를 알아보고 체계적으로 품질 데이터를 추출 및 표현하기 위한 방법론을 제안한다. 개발된 방법론은 데이터의 품질 요구사항을 수집하고 이를 구체화하여 실무에 적용하기 위한 절차를 제안하는 것을 목적으로 한다.

2. 데이터 품질의 문제 및 개선모델

2.1 품질의 정의

M.H.Brackett 은 데이터 품질을 데이터의 무결성과 데이터의 정확성 및 데이터 완전성으로 정의하고 있다[2]. 데이터 무결성은 데이터가 자원으로 저장되어 무결하게 유지되는 정도로 데이터의 값, 데이터 구조(관계 등), 데이터 갱신 및 삭제 등에 의한 데이터 보유상태로 정의된다. 데이터 정확성은 데이터 자원이 현실세계를 얼마나 정확히 나타내는가에 중점을 둔다. 이는 유무형의 항목 즉, 데이터의 크기, 현행성, 데이터 실례의 정확성, 개인적인 데이터 입출력 표현 등이 있다. 데이터 완전성은 정보에 대한 사용자의 요구를 만족시키기 위해 데이터가 저장되어 이용 가능한 정도를 나타낸다. 이는 사용자 요구를 만족시키기 위해 필요한 데이터가 무엇인지를 조사, 결정하고 이를 저장 유지하여 필요할 때 이용 가능하도록 지원하는 것이다.

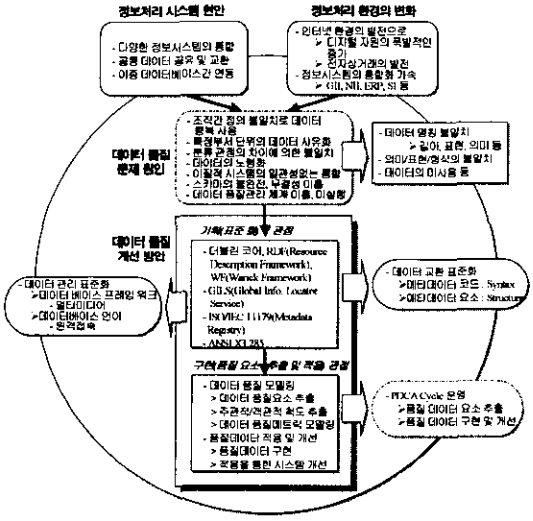
Bordie 는 데이터 품질을 의도된 어플리케이션의 필수적인 특성을 데이터베이스가 정확히 나타내는 정도로 정의하였으며 이를 위해 3 가지 특성 즉, 데이터 신뢰성, 논리적 또는 의미적 무결성, 물리적 무결성(구현의 정확성)으로 정의하고 있다[3]. 데이터 품질은 데이터의 적용유형에 따라 여러 형태로 정의할 수 있는데 문맥 내에서 데이터 품질문제는 본질적인 품질과 접근품질 및 문맥상의 품질, 표현품질로 구분할 수 있다[4]. Delone 과 McLean(1992)에 따르면 데이터 품질의 가장 중요한 7 가지 항목을 정보의 정확성,

출력의 시의성, 신뢰성, 완전성, 관련성, 정밀도 및 정확성으로 정의하고 있다[5]. 특히, 정보의 품질은 데이터의 품질과 동등한 품질(Garbage in Garbage out)을 갖는다고 일반적으로 알려져 있다.

본 고에서는 데이터 품질을 저장된 데이터가 얼마나 관련성이 있고 정확하며, 유용하고 이해할 수 있으며, 시기 적절한 데이터가 유지되는지를 참조하는 것으로 정의한다. 아울러 데이터 품질 모델링은 데이터 모델링과 유사하게 데이터 적용 Domain 으로부터 데이터의 품질을 추상화하는 과정으로 정의한다.

2.2 데이터 품질 문제 및 개선모델

일반적으로 기업내에서 다양한 시스템들이 독립적으로 존재한다. 예를들면 운용중인 시스템이 서로 다른 시점에 서로 다른 목적으로 개발되어 전사적 관점의 분석시 데이터 활용도가 떨어지거나 전사적 데이터가 아니라 특정부서 또는 담당자의 데이터로 인한 사유화와 분류관점의 차이 즉, A 부서 관점의 상품과 B 부서 관점의 상품이 서로 상이함에 따른 데이터 불일치 등의 품질문제의 주요 원인이다[6]. 그 밖에 그림 1에서와 같이 데이터 노령화, 스키마 설계 미흡, 관리체계 부실 등이 있다.

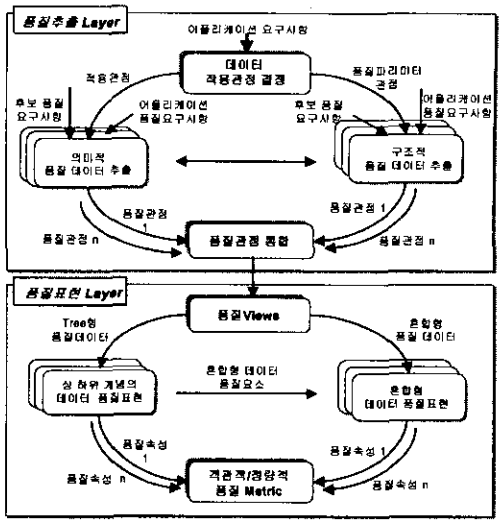


(그림 1) 품질문제의 원인 및 개선모델

데이터 품질개선 모델은 기획, 구현, 운영 관점으로 구분할 수 있다. 운영관점의 경우 조직(조직구조와 데이터 소유권 등 책임)과 관리정책 및 실행력(데이터 정책, 데이터 원칙, 방법론 등), 기술적 능력(측정, 저장) 과 같은 항목들의 유기적인 관리 및 유지가 필요하다. 즉, 시의적절하고 완전하며, 정확한 고품질의 데이터 유지를 위해 기획과 구현, 운영 즉, 3 박자가 유기적으로 실행되어야 한다.

3. 데이터 품질 모델링 방법론

데이터 품질 모델링의 목적은 데이터 품질의 측정 및 개선에 있다. 본 고에서는 그림 2 와 같이 체계적인 품질모델링을 위해 품질추출과 품질표현 단계를 제안한다. 품질추출 Layer 는 데이터 표현에 중점을 둔 데이터 품질향상 요소를 추출하는 구조적 품질데이터 모델링 단계와 데이터의 의미기반 데이터 품질요소를 모델링하는 의미적 품질데이터 추출 단계로 구성된다. 품질표현 Layer 는 품질추출 단계에서 모델링한 구조적, 의미적 품질 데이터를 통합한 품질관점으로부터 상하위 및 혼합형 품질데이터를 표현하는 단계이다. 품질표현 단계를 통해 추출된 품질 데이터의 최종결과는 객관적/정량적 메트릭으로 나타난다.



(그림 2) 품질데이터 모델링 개념도

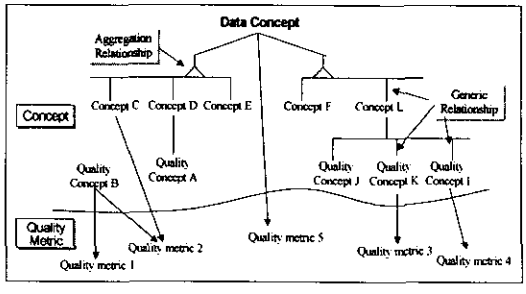
3.1 데이터 품질 추출단계

정보유통을 위한 시스템 통합이나 개선시 적정 수준의 데이터 품질을 보장하기 위해서는 리엔지니어링이나 리버스 엔지니어링 등 시스템을 모델링하는 과정에서 데이터 구축 과정에 대한 검토 및 관리와 통제가 매우 중요하다. 구조적 모델링은 데이터 표현모델에 종속적이고, 의미적 모델링은 표현모델에 다소 독립적인 특징을 갖는다.

3.1.1 의미적 데이터 품질 모델링

시스템을 재 구성하거나 새로운 정보시스템을 구축할 경우 관계형 데이터 모델링에서 구조적 표현은 쉽게 표현할 수 있으나 의미 표현은 매우 어렵다. 이를 위해 다양한 Semantic 모델이 등장 하였는데 의미적 표현의 예를 들면 Classification(instance/occurrence-of), Generalization/Specialization(is-a), aggregation(part-of), association (member-of) 등이다. 의미적 데이터 품질의 모델링은 Semantic 모델을 기반으로 수행하며, 그림 3과 같이 새롭게 구축될 시스템의 데이터 적용영역을 분석하여 최상위의 데이터 개념으로부터 하위 개념의 데이터를 추출하고, 이의 분석을 통해 의미적인 최종 품질데이터(메트릭)를 모델링해야 한다. 시스템의 마이그레이션이나 리엔지니어링 또는 이질적인 시스템의 통합 등 품질요소 추출을 위해 적용할 수 있는 기준은 다음과 같은 경우가 있다.

- 의미유사 즉, 동의어(의미와 표현이 동일한 경우) 이용 품질요소 추출
- 의미모호 즉, 동음 이의어(표현은 같으나 의미가 다른) 구조 이용 품질요소 추출
- 구조적 차이 즉, 의미는 같으나 표현구조가 다른 경우를 이용 품질요소 추출



(그림 3) 의미적 데이터 품질 모델링 체계

3.1.2 구조적 데이터 품질 모델링

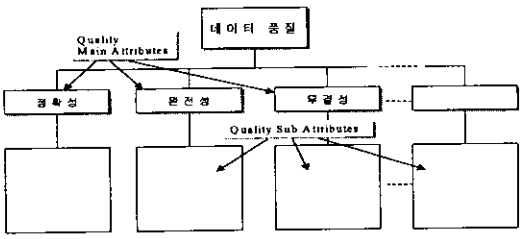
데이터 표준화 관점에서 데이터 품질을 모델링하고 정리하는 것이 매우 중요하다. 일반적인 데이터 품질문제의 근본원인은 서로 다른 부서에서 동일 데이터를 다르게 정의하여 중복 사용하는 경우나 시스템의 고령화시 또는 이질적인 시스템의 원칙없는 통합 등 구조적 모순에 의해 발생한다. 특히, 데이터 관련 표준의 부재나 존재하는 표준의 미준수 또는 표준 자체의 오류 등에 의해 구조적인 품질문제가 발생하며, 이들 구조적 오류정정에 의해 총 품질오류의 대부분을 해결할 수 있다. 데이터의 의미적 품질 대비 구조적 품질 오류는 Domain 에 따라 차이는 있으나 3:7 정도로 많은 부분이 데이터의 구조적 오류에 의해 발생된다. 구조적 품질데이터는 데이터의 구조, 형식, 교환 표준 등에 의해 도출한다.

3.2 데이터 품질 표현단계

데이터를 관리 및 통제하기 위해서는 컨트롤 Point 즉, 객관적인 메트릭이 표현되어야 한다. 예를들면 피사의 사탑이 더 이상 기울어져가는 것을 막기 위해선 어디를 통제해야 하는지 알아야 하는데 이와 같은 통제 점을 찾아가는 과정 즉, 데이터 적용분야에 따라 데이터의 품질향상을 위한 제어 Point 를 찾아 표현하는 과정이 본 단계의 목적이다.

3.2.1 상위개념의 데이터 품질 표현구조

상위개념의 품질 표현체계는 데이터 품질을 품질 주속성(Quality Main attributes)과 품질 하부속성(Quality Sub attributes) 들의 조합으로 표현하며, 데이터 품질의 주속성과, 하부속성 간에는 그림 4와 같은 계층적 관계가 있다.

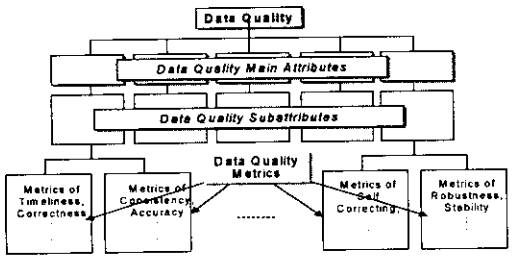


(그림 4) 상위 데이터 품질 표현구조

데이터의 품질 주속성 간의 구체적인 관계는 정형화시키기 어려운데 이는 데이터 품질이 적용되는 Domain 에 따라 다르게 정의될 수 있기 때문이다.

3.2.2 하위개념의 데이터 품질 표현구조

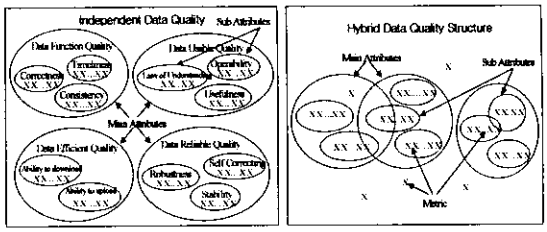
하위개념의 데이터 품질은 상위 개념의 품질데이터로부터 추출되며 최종결과는 객관적으로 데이터 품질을 측정, 개선시키기 위한 메트릭으로 나타난다. 그림 5는 상위 개념의 품질 데이터로부터 하위 개념의 품질 데이터 표현구조를 가시적으로 나타낸 것으로 품질 하부속성별로 메트릭 집합을 정의하며, 이들은 독립적이고 대칭적인 구조를 갖는다.



(그림 5) 하위 데이터 품질 표현구조

3.2.3 혼합형 데이터 품질 표현구조

데이터 품질의 상하위 속성 간에 일반적으로 독립적인 구조를 가지나 실제로 데이터 품질 측정시 품질 부속성과 세부 메트릭간 Overlap 되는 품질데이터를 상하위 품질 표현 구조로는 나타낼 수가 없다. 따라서 상호 중첩되는 혼합 속성의 품질데이터를 추출· 표현하기 위해 그림 6 과 같은 구조를 제안한다. 즉, Overlap 되는 혼합형 데이터 품질 표현을 위해 공통되는 속성을 정의함으로써 보다 빠른 시간에 정확한 데이터 품질 데이터를 예측할 수 있고 중복 정의되는 것을 예방할 수 있다.



(그림 6) 혼합형 데이터 품질 표현구조

4. 결론 및 향후 연구 과제

본 논문에서는 정보유통의 핵심인 데이터 품질향상을 위한 체계적인 품질모델링 방법론을 제안하였다. 이것은 ISO/IEC 9126 에서 제안한 소프트웨어 품질향상을 위한 품질특성 구조와 유사하게 데이터 품질을 계층적인 구조를 이용하여 표현하였다. 본 논문에서는 데이터 품질속성 들을 상하위 품질 표현뿐 아니라 혼합형 데이터 품질표현 체계를 제시하여 기존의 소프트웨어 품질표현 체계에서 나타낼 수 없었던 품질속성간 상호 Overlap 되는 품질데이터를 표현할 수 있는 체계를 제시하였다.

향후엔 본 논문에서 제안한 단계적 모델링 체계를 활용하여 네트워크 정보시스템에 적용할 예정이다. 또한 현재 제안된 단계적 방법의 효과적인 구현을 위해선 일반적인 즉, 공통적으로 적용할 수 의미적/구조적 품질데이터의 제시를 필요로 하는데 향후에는 이들 품질속성 및 메트릭 표준에 대한 연구도 필요하다.

[참고 문헌]

1. Information Week, Post-Year 2000 Technology Priorities for IT Ecutives, third-quarter, 1999
2. M.H. Brackett, The Dataware house Challenge : Taming Data Chaos, 1996
3. M.L. Bordie, Data quality information systems, information and management, Vol 3, pp245-258, 1980
4. Diane M. Strong, Yang W.Lee, and R.Y. Wang, Data Quality in context, Communications of the ACM, 1997
5. Delone, W.H., McLean, Information Systems Success : The Quest for the depended variable, Information Systems Research, 3:1, pp60-95
6. 고기원, 이응록, 정보공유 기반 구축을 위한 데이터 표준화, 정보통신연구, 1999.9
7. Jeff Rothenberg, Metadata to Support Data Quality and Longevity, IEEE, 1996
8. T.C. Redman, Data Quality for the Information Age, Arttech House, 1996
9. Donald P. Ballou and Giri Kumar Tayi, Enhancing Data Quality in Dataware House Environment, Communications of the ACM, 1999.1
10. ISO/IEC 9126-1: Information Technology-Software quality characteristics and metrics-Part 1: Quality characteristics and sub-characteristics, Feb. 4, 1997
11. Isabelle Mirbel, Semantic integration of conceptual schemas, Data & Knowledge Engineering, 1997
12. T.C. Redman, Data Quality for Telecommunications, IEEE, 1994
13. Ken Orr, Data Quality and Systems Theory, Communications of ACM, 1998
14. L.D. Paulson, Data Quality : A Rising E-Business Concern, IT Pro July/August, 2000