데이터 오류에서 추출한 데이터 품질 특성

김수경° 최병주
이 화 여 자 대 학 교 컴 퓨 터 학 과
{992COG05, bjchoi}@mm.ewha.ac.kr

Extraction of Data Quality Characteristics from Dirty Data

Sookyung Kim° Byoungju Choi
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

소프트웨어 제품의 품질을 보증하는 일은 매우 중요하며, 국제표준인 ISO/IEC 9126은 소프트웨어 품질 특성 및 측정 메트릭 표준을 제공하고 있다. 이때 ISO/IEC 9126에서는 소프트웨어를 프로그램, 절차, 규칙 및 관련문서로 한정하고 있기 때문에 데이터의 품질에는 적용할 수 없다. 본 논문에서는 데이터 품질 평가 및 제어를 위하여 데이터 오류 형태를 분류하고, 이를 기반으로 데이터 품질 특성 및 부특성을 분류한다. 데이터 품질 특성 분류는 ISO/IEC 9126에 정의한 소프트웨어 품질 특성을 데이터 오류 형태에 대응시켜 추출한다. 본 논문에서 제시하는 데이터 품질특성 분류는 지식 공학(knowledge engineering) 시스템이 최종 사용자에게 제공하는 데이터나 지식의 품질 측정 및 제어에 기준이 된다.

1. 서론

ISO 국제표준에 따르면, 소프트웨어 제품이란 사용자에게 인도할 것으로 지정된 컴퓨터 프로그램, 절차와 관련 문서 및 데이터의 전체 집합으로 정의하고 있다. ISO/IEC 9126 [1]은 소프트웨어 제품의 품질 평가를 위한 국제 표준으로, 소프트웨어의 품질을 정의하고 소프트웨어 품질을 측정하는 방법을 제시하여 소프트웨어 품질 향상을 위한 기반을 제공한다. 그런데, ISO/IEC 9126에서 대상으로 하는 소프트웨어란 "데이터 처리 시스템의 운영에 관계된 프로그램, 절차, 규칙 그리고 관련 문서로 된 지적 창작물"로 한정되어 있어, 소프트웨어 제품을 실제적으로 구동 시키는데 이용되어지는 데이터의 품질에 대해서는 다루어지고 있지 않다. 즉, ISO/IEC 9126은 데이터 품질분야에는 적용되지 못하고 있다. 표준이 확립되지 않은 연구는 그 응용에 한계가 있기 때문에 개발에 여러 제한이 생기리라는 것은 예측 가능한 사실이다.

지식 공학(knowledge engineering) 시스템[2]은 다양한 데이터 소스로부터 최종 사용자가 요구하는 의미 있는 데이터나 나아가 지식을 추출할 수 있도록 한다. 따라서 지식 공학 시스템에 초기 입력이 되는 데이터의 품질은 최종 사용자에게 제공되는 데이터나 지식의 품질을 결정하는 중요한 요인이 된다. 만일 지식 공학 시스템에 데이터 품질 제어 기술이 없다면, 신뢰할 수 없는 데이터나 지식을 최종 사용자에게 제공하게 되므로, 그 자체의 존재가 무의미하게 될 것이다. 이런 의미에서 "데이터 품질 평가 및 제어"의 중요성을 찾아볼 수 있다. 지식 공학 시스템을 구성하고 있는 데이터 웨어하우징, OLAP, 지식탐사 컴포넌트는 품질이 제어된 데이터 없이는 그 성공 여부가 불투명하다고 까지 말할 수 있다.

본 논문은 데이터 오류의 분류를 시작으로 데이터 품질 특성을 파악하여 지식 공학 시스템에서의 데이터 품질 측정 및 제어를 목적으로 한다. 데이터 품질 특성 분류는 ISO/IEC 9126에 정의한 소프트웨어 품질 특성을 데이터 오류 형태에 대응시켜 추출한다. 본 논문에서 제시하는 데이터 품질특성 분류는 지식 공학 시스템이 최종 사용자에게 제공하는 데이터나 지식의 품질 측정 및 제어에 기준이 된다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 다루어지는 용어의 정의를 기술한다. 3장에서는 데이터 오류의 형태별 분류를 제안한다. 4장에서는 3장에서 제시된 분류에 ISO/IEC 9126에 적용하여 데이터 품질특성을 분류한다. 5장에서는 4장에서 제안된 데이터 품질 특성을 적용하여 고안된 데이터 품질 측정 컴포넌트를 제안한다. 6장에서는 결론과 향후 연구 과제를 제시한다.

2. 정의

(1) 데이터 오류

없어진 데이터(missing data), 잘못된 데이터 (wrong data), 표준이나 형식이 없는 데이터(lack of standard)와 사용자의 특정 제한을 만족시키지 못하는 데이터를 모두 일컬어 데이터 오류(dirty data)라고 한다 [3,4,5].

데이터 오류가 생기는 가장 근본적인 문제는 시스템의 데이터 관점과 실세계의 데이터를 비교하여 입력해야 한다는 점과, 또 이 실세계의 데이터는 끊임없이 변화한다는 점이다[6]. 이로 인해 발생되는 부정확한 데이터, 데이터베이스 전체에 걸친 불일치, 나아가 불필요한 데이터등과 같은 문제들로 인해, 데이터 마이닝이나 OLAP 분석을 위한 데이터 웨어하우스가 구축되었을 때 과연 이 자료들을 믿을 수 있을 것인가 하는 신뢰성문제가 발생하게 된다.

(2) 데이터 품질특성

데이터 품질특성(data quality characteristics)은 소프트웨어 품질 특성[1]과는 달리 아직 표준이 정립되어 있지 않고 각기 필요성에 따라

조금씩 연구가 진행되어 왔다. 이 중에서 가장 대표적이라 할 수 있는 것으로 Ballou and Pazer의 연구[7]와 Wang[8,9]의 연구를 들 수 있다. 이들 연구 결과를 분석했을 때, 데이터 품질에는 4가지 차원, 즉 정확성(accuracy), 적시성(timeliness), 완료성(completeness), 일관성(consistency)으로 분리하여 품질관련 연구가 이루어짐을 알 수 있다. 따라서 본 연구에서는 이들 데이터 품질차원에 따라 데이터의 오류를 분류한다. 각 품질차원의 정의는 다음과 같다:

정확성:기록된 값이 실제 값과 일치하는 상태이다.
적시성:기록된 값이 시간적으로 실제 값과 일치하는 상태를 말한다.
완료성:특정 변수에서 요구되는 모든 값들이 기록되어진 상태이다.
일관성:기록된 값의 표현이 모든 경우에 항상 같은 상태이다.

(3) 데이터 품질 측정 대상

현재까지의 데이터 품질 연구는 legacy 데이터베이스에 국한되어있기 때문에, 지식 공학 전반에 대한 데이터를 대상으로 했을 때 이들 연구결과는 미완성이라 할 수 있다. 예를 들어, *Chamois*[2]는 이화여대 컴퓨터학과에서 5년간의 R&D프로젝트로써 개발하고 있는 지식공학시스템으로써, 그림1과 같이 '관계형 데이터베이스', '데이터 웨어하우스', 'OLAP', '데이터마이닝'이 주요 컴포넌트이며, 이들은 컴포넌트 기반 소프트웨어 개발 기술을 적용하여 개발되고 있다. 데이터 품질 연구는 지식 공학 시스템에서의 데이터의 출발에서부터 소멸에 이르는 전과정을 고려해야 한다. 그림 1에서 데이터의 변환이 이루어지는 곳은 동그라미로 표시된 곳으로써 모두 9군데이며, 이들 각 단계에서 데이터 오류가 발생 할 수 있다. 본 연구에서는 지식공학시스템의 데이터 변환이 일어나는 모든 장소를 데이터 품질 측정 대상으로 한다. 즉, 그림 1의 9군데의 데이터 변환 장소는 각각 다음과 같이 데이터 품질 측정 대상이 된다:

- Legacy DB
- In place transformation
- ETL_Ext (Extraction)
- Data cleansing tool
- ETL_J (Joiner transformation)
- ETL_Exp (Expression transformation)
- ETL_A (Aggregator transformation)
- ETL_L (Load)
- 데이터 웨어하우스

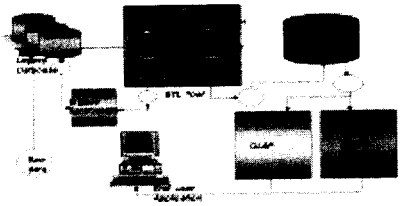

그림 1. 데이터 품질 측정 대상

3. 데이터 오류 분류

데이터 품질 측정을 위해서는 데이터 품질 세부 특성을 정확히 파악해야 하며, 품질특성에 따라 품질을 측정하기 위한 메트릭을 개발하여야 한다. 데이터 품질 세부 특성을 추출하기위한 전단계로써 본 연구에서는 앞 절에서 정의한 4가지의 데이터 품질 차원인 정확성, 적시성, 완료성, 일관성에 관한 데이터 오류 종류를 분류하였다.

3.1 데이터 오류 분류의 구성

데이터 오류 형태(dirty data type)는 크게 '데이터 자체 오류(inherent dirty data)'와 '사용 적합성(pragmatic dirty data)'에 관한 오류의 두 가지 형태로 나뉜다. 데이터 자체는 '데이터 내용(data content)', '데이터 표현(data presentation)', '데이터 정의(definition)' 오류로써 다음과 같이 세분화된다.

- **데이터 자체 오류 형태**
  - **데이터 내용 오류**: 데이터 내용상의 오류,
  - **데이터 표현 오류**: 데이터 존재방식의 오류,

표1. 데이터 품질 오류 분류

품질차원	데이터 오류 종류	오류 항목
정확성	데이터 내용 오류	▪유효하지 않은 문자 ▪유일 법칙에 어긋남 ▪존재법칙에 어긋남 ▪범위의 제약성 ▪타이핑 실수 ▪스펠링 오류 ▪부정형 문자 ▪동음이의어
	데이터 표현 오류	▪입력시에 잃어버린 데이터 ▪입력시간을 모른채 있는 정보
	데이터 정의 오류	▪다른 단위 ▪다른 해독 ▪다른 문자열 ▪다른 생략 ▪다른 가정 ▪공백 사용 ▪특수문자의 사용
	사용적합성	▪구조화 되지 않은 데이터 ▪문서화 되지 않은 데이터 ▪크기가 적합치 않은 데이터 ▪실행 효율성이 시간에 바탕을 두지 않음 ▪저장효율성이 없음 ▪잘못된 공유
적시성	데이터 내용 오류	▪동일한 id를 가진 과거의 내용을 추출 ▪최신 버전이 아닌 과거의 버전을 로드
	데이터 표현 오류	▪입력시간을 모른채 있는 정보
	데이터 정의 오류	
	사용적합성	▪구조화 되지 않은 데이터 ▪문서화 되지 않은 데이터 ▪크기가 적합치 않은 데이터 ▪실행 효율성이 시간에 바탕을 두지 않음 ▪저장효율성이 없음 ▪잘못된 공유
완료성	데이터 내용 오류	▪이상한 정보 ▪유효의 법칙에 어긋남
	데이터 표현 오류	▪입력시에 잃어버린 데이터 ▪입력시간을 모른채 있는 정보
	데이터 정의 오류	▪다른 단위 ▪다른 해독 ▪다른 문자열 ▪다른 생략 ▪다른 가정 ▪다른 공백 ▪다른 특수문자
	사용 적합성	▪구조화 되지 않은 데이터 ▪문서화 되지 않은 데이터 ▪크기가 적합치 않은 데이터 ▪실행 효율성이 시간에 바탕을 두지 않음 ▪저장효율성이 없음 ▪잘못된 공유
일관성	데이터 내용 오류	▪유일법칙에 어긋남
	데이터 표현 오류	
	데이터 정의 오류	▪다른 단위 ▪다른 해독 ▪다른 문자열 ▪다른 생략 ▪다른 가정 ▪다른 공백 ▪다른 특수문자 ▪다른 동의어
	사용 적합성	▪구조화 되지 않은 데이터 ▪문서화 되지 않은 데이터 ▪크기가 적합치 않은 데이터 ▪실행 효율성이 시간에 바탕을 두지 않음 ▪저장효율성이 없음 ▪잘못된 공유

  - **데이터 정의 오류**: 데이터의 정의나 의미가 달라 발생하는 오류
- **사용 적합성 오류 형태**
  - **사용 적합성 오류**: 주어진 목적에 적합하지 못한 데이터로 인한 오류

데이터 오류의 연구[3,10]는 이미 시작된 상태이며, 그 결과 Trillium[11,12]등의 데이터 정제(data cleansing) 제품이 상용화되고 있다. 그러나 이들의 연구가 사람의 이름이나 주소등의 단순한 경우와 Legacy DB에 한정되어 있고, 게다가 정제 제품을 통해 정제되는 데이터 만을

대상으로 하고 있기 때문에 지식 공학 시스템 전반에 걸쳐 적용하기는 어렵다고 할 수 있다.

본 연구에서는 위에서 기술한 데이터 오류 형태를 세분화하여 데이터 오류항목을 분류하였다. 데이터 오류 분류 대상은 데이터 품질 측정 대상인 지식 공학 시스템의 데이터가 변형되는 9가지 장소를 대상으로 한다. 본 연구에서 추출한 데이터 오류항목을 데이터 품질차원에 따라 각각 표1에 나타내었다.

본 연구에서는 표1에 나타낸 모든 데이터 오류 종류에 대한 사례를 식품회사의 고객 정보를 대상으로 Windows NT 서버에서 구축하였다. Legacy 데이터베이스의 작성은 Microsoft SQL Server 7.0으로, ETL도구로는 Power Mart 4.5를 사용하였다

4. 데이터 품질 특성 분류

ISO/IEC 9126에 따르면 소프트웨어 품질 특성은 다시 부특성(sub-characteristics)으로 세분화된다. 데이터 품질 특성도 이와 마찬가지로 세분화할 수 있다. 데이터 품질특성 및 품질 부특성을 분류하기 위하여, 표1에 기술한 각 데이터 오류항목을 그림 2에서처럼 ISO/IEC 9126의 소프트웨어 품질 특성 및 부특성에 대응시켰다. 대응된 오류항목으로부터 데이터 품질 특성 및 부특성을 분류할 수 있다. 세분화된 데이터 품질 특성은 표2와 같다.

예를들어, 그림 2에서 Legacy DB의 '데이터 내용 오류' 형태의 오류항목인 '유효하지않은 데이터', '잘못된 타이핑', '스펠링 에러', '잘못된 약어', '동음이의어' 등은 ISO/IEC 9126의 품질 부특성의 '정확성'에 대응되며, 정확성은 ISO/IEC 9126에서 '기능성' 품질 특성에 속한다. 따라서 표 2의 데이터 품질 특성 및 부특성에 해당하게 된다.

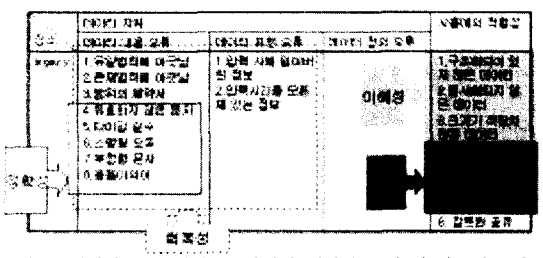

그림 2. 데이터 오류 항목으로부터의 데이터 특성 및 부특성 도출

표2. 데이터 품질 특성 및 부특성

특성	부특성	데이터 오류 종류
기능성	정확성	데이터 내용 오류
	적합성	사용적합성 오류
	내부동작성	사용적합성 오류
신뢰성	성숙성	데이터 내용 오류 데이터 표현 오류
	오류허용	데이터 내용 오류 데이터 표현 오류
	회복성	데이터 내용 오류 데이터 표현 오류
유지보수성	분석가능성	데이터 정의 오류
	변화가능성	데이터 정의 오류
	테스트가능성	데이터 정의 오류
이식성	적용성	데이터 정의 오류
	부합성	데이터 정의 오류
	대체가능성	데이터 정의 오류
효율성	시간작용	사용적합성 오류
	자원작용	사용적합성 오류
사용성	이해성	사용적합성 오류

5. 데이터 품질 측정 컴포넌트

본 연구에서 개발하는 데이터 품질 측정 컴포넌트, DQMC(Data Quality Measuring Component),는 *Charmois* 지식 공학 시스템에 포함될 예정이다. 데이터의 품질 측정은 지식 공학 시스템의 모든 품질 측정 대상에서 이루어 져야 한다. 현재는 1차적으로 그림 3에서처럼 ETL에서 사용할 수 있도록 하였다. 실제 데이터가 분석되어 사용되어 지는 곳은 데이터 웨어하우스이다. 이곳에 데이터가 저장되어 OLAP이나 데이터 마이닝으로 분석되어 사용되기 이전에 데이터 품질의 측정이 이루어질 수 있게 되어 데이터 웨어하우스에 대한 사용자의 신뢰성을 높일 수 있게 된다.

DQMC는 Visual C++으로 구현하였으며, ETL 도구인 Power Mart의 COM 외부 처리(external procedures)라는 타 개발측에서 개발한 제품을 소비자가 Power Mart의 라이브러리에 끼워서 쓸 수 있도록 하는 기능을 이용하여 개발하였다

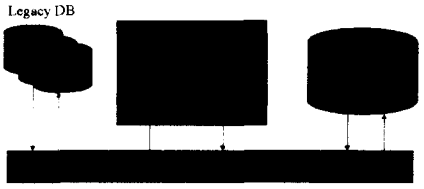

그림 3. Data Quality Measuring Component

6. 결론 및 향후 연구 과제

본 연구는 다양한 데이터 소스들로부터 의미 있는 데이터나 나아가 지식들을 추출하는 지식 공학 시스템에서의 데이터 품질을 보장하는 것을 목적으로 한다. 이를 위하여 데이터 오류를 분류하고, 이를 기반으로 데이터 품질 특성을 분류하였다.

정확성, 적시성, 완료성, 일관성 데이터 품질 차원에 따라 데이터 오류형태와 형태별 오류항목을 지식 공학 시스템에서의 품질 특정 대상에 따라 분류하였다. 분류한 각 데이터 오류항목에 ISO/IEC 9126에 정의한 소프트웨어 품질 특성 및 부특성을 대응시킴으로써 데이터 품질 특성 및 부특성을 추출할 수 있었다.

향후 분류한 데이터 오류 종류에 대하여 진행한 사례구축을 더욱 보강하여 데이터 오류 분류 및 데이터 품질 특성 분류를 보완할 예정이다. 한편, 데이터 품질 측정을 위해서는 분류한 각 데이터 품질 및 부품질에 대한 품질 측정 메트릭을 개발하여야 하며, 이 메트릭을 DQMC에 적용하여 데이터 품질 측정 컴포넌트의 도구 개발을 완성할 예정이다.

참고문헌

[1]ISO/IEC 9126 -1,2,3, JTC 1 SC 7 WG 6(Evaluation & Metrics) Documents, Nov 1996

[2]Won Kim, Ki-Joon Chae, Dong-Sub Cho, Byoungju Choi, Myung Kim, Ki-Ho Lee, Meejeong Lee, Sang-Ho Lee, Seung-Soo Park, Hwan-Seung Yong, "A Component-Based Knowledge Engineering Architecture," *JOOP*, vol.12, no.6, pp40-48, 1999

[3]"The Five Legacy Data Contaminants You Will Encounter in Your Warehouse Migration"
http://www.vality.com

[4]J. Williams," Tools for traveling Data," DBMS, Miller Freeman,Inc., Jun 1997

[5]Cutter Information Corporation. "Data Management Strategies Newsletter on The State of the Data Warehousing Industry", Vol. 2, N. 3, Mar. 1998

[6]Ken Orr, "Data Quality and System Theory," Communications of the ACM, Vol. 41, Num. 2 Feb. 1998

[7]Ballou, D. P. and Pazer, H.L, "Modeling Data and process Quality in multi-input, multi-output information systems," Management Science 31, pp 150-162, Feb.1998

[8]Diane M. Strong, Yang W. Lee, and Richard Y. Wang, "Data Quality in context," Communication of the ACM, Vol. 40 Num. 5, May 1997

[9]Richard Y. Wang, Veda C. Storey, and Christopher P. Firth, "A framework for Analysis of Data Quality Research," IEEE Transactions on Knowledge and Engineering, Vol. 7. Num. 4, pp.623-640, Aug. 1995

[10]Vality Technology, "Five Common Excuses for Not Re-engineering Legacy Data," DM Review

[11]Leonard Dubois - Trillium Software, Inc., "Achieving Enterprise Data Quality"
http://www.tdan.com

[12]Trillium Software product introduction paper

[13]Donald P. Ballou and Giri Kumar Tayi, "Enhancing Data Quality in Data Warehouse Environment," Communications of the ACM, Vol. 42. Num.1, pp73 – 78, Jan. 1999

[14]Larry P. English, "Improving data Warehouse and Business Information Quality - Methods for Reducing Costs and Increasing Profits," Wiley