소프트웨어 사용자 관점의 데이터 품질 측정 방안

양자영[0] 최병주
이화여자대학교 컴퓨터학과
{komi, bjchoi}@ewha.ac.kr

A Methodology of Measuring Data Quality from Viewpoint of Software user

Jayoung Yang[0] Byoungju Choi
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

소프트웨어 제품의 품질을 보증하는 일은 중요하며, 이를 위해서는 실제 소프트웨어 제품이 실행될 때 최적의 결과에 영향을 주는 데이터, 즉 데이터의 품질이 보증 되어야 한다. 그러나 대부분의 소프트웨어 품질 관련 연구에서는 소프트웨어 품질 측정에 대한 모형만을 제시할 뿐 데이터 품질에 대해서는 다루어지고 있지 있다. 본 논문에서는 데이터 품질 평가를 위하여 데이터 품질을 측정하는 메트릭을 제안한다. 제안한 메트릭은 전체 데이터베이스에서 오류 데이터가 발생한 비율과 데이터 사용 목적에 따라 데이터 항목마다 다른 가중치를 적용하여 구해진다. 본 논문에서 제안하는 데이터 품질 메트릭은 특히 데이터를 주로 처리하는 소프트웨어 시스템의 품질 측정에 기여할 수 있다.

1. 서론

ISO/IEC 9126[1]은 소프트웨어 제품의 품질 평가를 위한 국제 표준이며 이는 소프트웨어 품질을 측정하기 위한 품질 모형만을 제시한다. 현재 데이터 품질에 관한 연구는 꾸준히 이루어 지고 있으나 소프트웨어 품질 연구와는 달리 아직 표준이 정립되어 있지 않으므로 그 표준이 되는 메트릭의 부재로 인해 품질의 정도를 측정하는 것은 사실상 어렵다.

실제 소프트웨어를 실행할 때 요구되는 데이터는 소프트웨어 품질에 영향을 준다. 즉, 소프트웨어 시스템에서 최적의 결과를 얻기 위해서 데이터 품질 측정이 요구된다. 따라서 데이터 품질 측정은 소프트웨어 품질 측정 표준인 ISO/IEC 9126의 데이터 관련 부분을 보완해주므로 소프트웨어 품질의 상세 측정에 기여할 수 있다.

본 논문은 데이터 품질 측정을 위해서 정량적 평가가 가능하도록 메트릭을 제시한다. 제안한 메트릭은 전체 데이터베이스에서 오류 데이터가 발생한 비율과 데이터 사용 목적에 따라 컬럼별로 다른 가중치를 적용하여 구해진다. 따라서 오류 데이터의 비율을 측정하기 위하여 오류 데이터를 발견하는 알고리즘을 기술한다. 데이터를 사용하는 목적에 따라 오류 데이터의 심각도가 달라지므로 데이터 항목마다 다른 가중치를 주는 방법을 제안한다. 데이터 품질을 측정할 수 있는 시스템 아키텍처(DAQUM Architecture)를 설계한 후 제안한 메트릭을 적용하여 실제로 사례연구를 통해 데이터 품질을 측정해 본다. 본 논문에서 제시하는 데이터 품질 측정 메트릭은 최종 사용자에게 제공되는 정보와 지식의 가치를 평가할 수 있는 기준을 제공해준다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 다루어지는 용어의 정의 및 데이터 품질 측정에 필요한 관련연구를 기술한다. 3장에서는 데이터 품질 측정 메트릭과 이를 적용한 전체적인 아키텍쳐 설계방안을 제안한다. 구현된 데이터 품질 측정 도구를 실제로 사례연구를 통해 살펴본다. 4장에서는 데이터 품질 측정 컴포넌트를 제안한다. 5장에서는 결론과 향후 연구 과제를 제시한다.

2. 정의 및 관련연구

(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 [2].

데이터 오류는 시스템의 데이터 관점과 실세계의 데이터를 비교하여 입력해야 하지만 이 실세계의 데이터는 끊임없이 변화하기 때문에 발생한다.[3].

(2) 오류 데이터 항목

본 논문은 이미 검증을 받은 " successive hierarchical refinement" 방식을 이용하여 구축된 총체적인 오류 데이터의 분류[4](이하 Kim's et al분류라 칭함)를 기반으로 연구했다. Kim' s et al분류의 경우 분류 구조의 단말노드(leaf node)가 실제적인 오류 데이터의 이름을 뜻하며 이를 오류 데이터 항목이라고 정의했다. Kim's et al분류는 오류 데이터가 생기는 이유를 없어진 데이터로 인해 생기는 오류 데이터, 없어지지는 않았지만 잘못된 오류 데이터 잘못되지는 않았지만 사용할 수 없는 데이터의 세가지로 나누었다. 본 연구는 이 분류체계의 오류 데이터 항목을 이용하여 각 항목에 따른 오류 데이터 발견 알고리즘을 제안했다.

3. 데이터 품질 측정 방안

본 논문에서는 사용자 중심의 데이터 품질 평가를 위해 전체적인 데이터 품질 측정하는 DAQUM(Data Quality Measurement) 아키텍처를 설계하고 여기에 구축된 데이터 품질 측정 메트릭을 적용한다. 메트릭을 구축하기 위해서는 다음 두가지 단계를 수행한다. 먼저 오류 데이터를 발견하는 알고리즘을 통해 오류 데이터를 발견하고 다음단계는 발견된 오류 데이터를 데이터 사용하는 목적에 따라 다른 가중치를 준다.
즉, 데이터 품질을 측정하는 알고리즘은 아래와 같다.

<데이터 품질 측정 알고리즘>

step 1. 오류데이터의 자동 발견
  1.1 비즈니스 도메인 분석 후 요구사항에 맞는 비즈니스 규칙 및 제약사항 결정
  1.2 오류 데이터 항목별로 규칙을 위반한 제약조건을 SQL query로 변환
  1.3 컬럼별로 SQL query를 실행하여 총 오류 데이터의 개수 파악
step 2. 데이터 사용목적에 따른 가중치 부여
  1.1 사용목적과 관련이 있는 컬럼중 우선 순위가 높은 순으로 컬럼 추출
  1.2 추출된 컬럼의 중요도에 따라 가중치 부여
step 3. step1과 step2를 적용한 데이터 품질 측정 메트릭 구축
step 4. 데이터 품질 측정 도구에 제안한 메트릭을 적용하여 데이터 품질 측정값을 수치화

3.1 오류 데이터 발견 알고리즘

데이터 품질 측정을 위해 오류 데이터 항목을 발견하는 알고리즘은 오류 데이터 항목별로 도메인에 정의된 제약조건을 위배한 규칙을 SQL Query로 변환하는 기법으로 작성했다. 예를 들어 그림 1에서 도메인에 주민등록번호가 앞자리수가 6자리이고 다음에 대쉬기호가 오며 그 뒷자리수는 7자리가 되는 규칙이 정의 되었다면 이 규칙에 맞게 SQL을 작성한 후 이 조건을 만족하지 않는 데이터를 검색하여 오류 데이터를 발견했다.

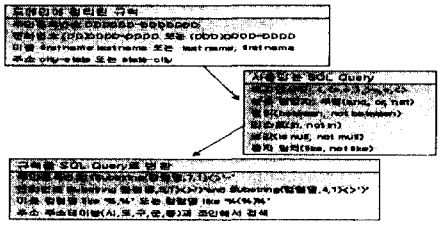


그림1. 도메인의 제약조건을 위배한 규칙을 SQL Query로 변환

오류 데이터 분류체계[5,6]의 오류 데이터 33가지 항목 중 SQL 및 알고리즘을 통해 오류 데이터를 자동으로 발견될 수 있는 항목은 그림 2처럼 18가지 항목이다. 나머지 항목은 도메인 전문가의 중재가 필요한 항목들이다.[5,6]

본 연구에서는 오류 데이터가 자동으로 발견 가능한 18개의 항목 중 7개 항목은 값, 타입 체크를 통한 알고리즘으로 오류 데이터가 검색이 가능하게 했다. 나머지 11개 항목은 오류 데이터 검색시 이름, 주소, 전화번호부, 우편번호, 범주 등의 참조 테이블과 약어, 축약어가 저장된 사전들을 필요하므로, 이를 데이터베이스 안에 구축하여 조인을 통해 오류 데이터 검색이 가능하게 했다.

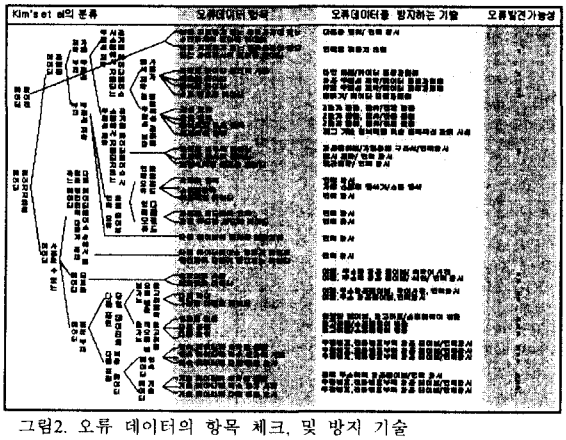
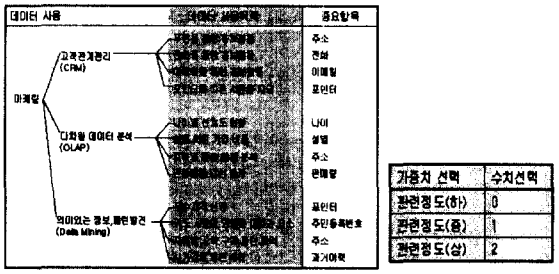


그림2. 오류 데이터의 항목 체크, 및 방지 기술

3.2 사용목적에 따라 중요한 항목에 가중치 부여

데이터 품질은 데이터 사용 목적에 따라 달라지므로 본 연구에서는 그림 3처럼 데이터 사용목적과 그에 따른 중요한 항목을 추출하여 분류체계를 만들었다. 현재 가중치 주는 기준에 대한 표준이 마련되어 있지 않아 가중치 부여는 주관적인 평가에 의해서 맞길 수 밖에 없다. 가중치를 주는 방법은 이미 구축된 분류 체계를 고려하여 중요항목으로 선정된 데이터 항목을 다른 항목과는 다른 penalty를 주도록 한다. 예를 들면 고객에게 홍보용 자료를 우편으로 보내는 목적으로 데이터가 사용되었다면 주소항목이 가장 중요한 항목이 되며, 주소 항목에서 오류가 발생을 했다면 다른 필드보다는 높은 penalty를 줄 수 있도록 한다.



가중치 선택	수치선택
관련정도(하)	0
관련정도(중)	1
관련정도(상)	2

그림3. 데이터 사용목적과 중요항목의 분류체계와 가중치

3.3 데이터 품질 측정 메트릭

데이터 품질 메트릭은 전체 데이터베이스에서 오류 데이터가 발생한 비율과 데이터 사용 목적에 따라 컬럼별로 다른 가중치를 적용하여 구해진다.

데이터 품질 메트릭은 다음과 같이 정의한다.

<정의> 데이터 품질 측정 메트릭 Q

Q : 데이터 품질 측정값
N : 총 품질 측정 대상 데이터 개수
T : 컬럼별로 가중치를 부여한 총 오류 데이터 개수
m_i : 총 컬럼수
c_{ij} : 각컬럼에 해당되는 오류 데이터의 개수
w_j : 각 컬럼에 해당하는 가중치
a_{ij} : 데이터 품질 측정에 필요한 총오류 데이터 항목의 개수
n_{ijk} : a_{ijk}의 오류 데이터 개수
a_{ijk} : 데이터 품질 특정에 필요한 오류 데이터 항목 k

$$Q = 1 - \frac{T}{N}$$

$$T = \sum_{j=1}^{m_i} c_{ij} \times w_j$$

$$c_{ij} = \sum_{k=1}^{a_{ij}} n_{ijk}$$

3.4 데이터 품질 측정 아키텍처

데이터 품질 측정 도구 DAQUM은 JDK1_3_1와 SQL Server 2000을 사용하여 구현했다. 데이터 품질 측정 아키텍처는 그림 4처럼 SQL Server 2000을 이용하여 세가지로 구분되어 운영되는 데이터베이스와 데이터 품질 측정 메트릭을 적용하여 구현된 DQAUM도구로 구성된다. 오류데이터 검색은 JDK1_3_1로 구현된 DAQUM도구에서 source database를 불러와서 SQL Query에 의해 오류 데이터가 자동으로 검색이 된 후 repository database에 오류 데이터 항목별, 측정 테이블의 컬럼별로 수치값이 저장 된다. 그리고 DAQUM도구에 제안한 메트릭을 적용하여 데이터 품질을 측정한 후

수치화된 결과를 사용자에게 보여주고 target database에 저장한다.

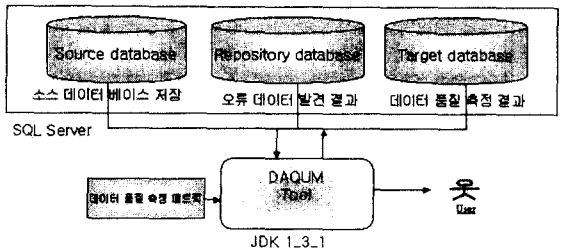


그림 4. 데이터 품질 측정 아키텍처(DAQUM)

3.5 데이터 품질 측정 도구(DAQUM) 구현 및 사례연구

데이터 품질 측정 아키텍처에 실질적으로 물품관리 회사의 직원 데이터를 적용시켜 사례연구를 통해 데이터 품질을 측정했다. 오류 데이터 발견 알고리즘에 의해 오류 데이터 타입별, 컬럼별로 오류 데이터를 발견해 이를 수치화한 결과는 그림5와 같다. 그림6은 사용 목적에 따라 데이터 품질 측정을 2가지 타입으로 선택하여 측정할 수 있게 했다. Type1는 임의로 데이터 품질을 측정할 수 할 수 있고 Type2는 사용자가 직접 사용 목적과 가중치를 직접 선택하여 데이터 품질을 측정할 수 있게 했다. 데이터 품질 측정값은 그림 6 처럼 같은 테이블에서 데이터 사용목적에 따라 측정값이 달라진다. 또한 오류 데이터가 발생한 중요한 항목에 가중치를 큰 값으로 부여할수록 데이터 품질 측정값은 낮아진다. 즉 데이터 품질 측정값은 데이터 사용목적에 따라 달라지는 오류 데이터의 심각도를 반영해 주는 결과값이다.

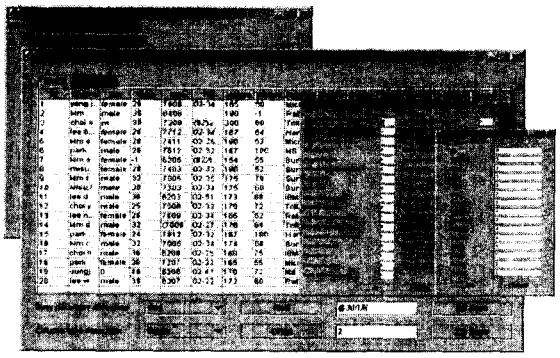


그림5. 오류 데이터 알고리즘을 통해 오류 데이터 발견 결과

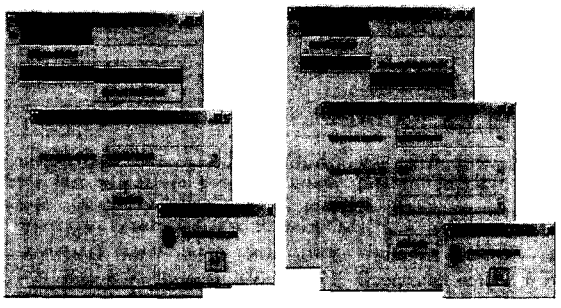


그림6. 사용목적에 따라 임의로 측정한 데이터 품질 측정값과 사용자가 중요항목과 가중치 선택한 데이터 품질 측정값

4. 데이터 품질 측정 컴포넌트

Chamois[5,6]는 현재 이화여대 컴퓨터학과에서 컴포넌트 기반 소프트웨어 개발 기술을 적용하여 개발하고 있는 지식 공학 시스템이다. 지식 공학 시스템의 품질을 보장을 위해서는 먼저 데이터의 품질이 보장되어야 한다. 본 연구에서 개발하는 DAQUM은 그림 7 처럼 향후 지식 공학 시스템에 포함될 모듈이다. 지식 공학 시스템 내에는 여러 곳의 데이터 품질 측정 대상이 있으며 이를 컴포넌트로 개발을 하면 동일한 기능을 재사용할 수 있어 소프트웨어 개발 시 드는 비용과 노력을 줄일 수 있다.

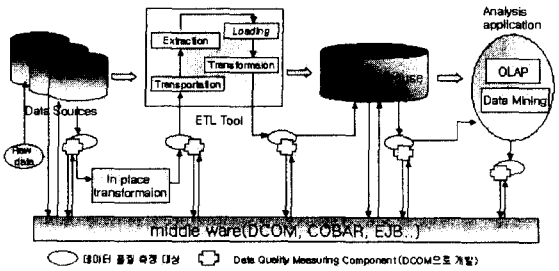


그림 7. 데이터 품질 측정 모듈

5. 결론 및 향후 연구 과제

소프트웨어 시스템의 품질을 보장하기 위해서는 먼저 데이터의 품질이 보장되어야 한다. 본 논문은 의미 있는 지식을 추출할 수 있는 원천인 데이터의 품질을 보장하기 위해 데이터 품질 측정 방안을 제안했다.

본 논문에서 제안한 품질 측정 메트릭은 최종 사용자로 하여금 데이터 품질을 평가하고 제어하는 기준을 제공해주는데 의의가 있다. 데이터 품질 측정 메트릭은 오류 데이터가 발생한 정도와 데이터 사용 목적에 따라 컬럼별로 다른 가중치를 적용하여 구해진다. 따라서, 본 논문에서는 오류 데이터의 비율을 측정하기 위하여 오류 데이터를 발견하는 알고리즘을 기술했다. 데이터를 사용하는 목적에 따라 오류 데이터의 심각도가 달라지므로 데이터 항목마다 다른 가중치를 주는 방법을 제안했다. 결국 데이터 품질 측정방안은 데이터 품질이 영향을 미치는 소프트웨어 품질 측정에 기여할 수 있다.

이 논문은 향후 데이터 품질 측정에 대한 표준을 정립하는 것을 목표로 하며, 데이터 품질 측정 메트릭이 측정에 요구되는 모든 사항들을 반영하고 있는지를 검증할 계획이다. 현재 개발된 데이터 품질 측정 도구는 JDK1_3_1로 구현된 응용 프로그램이다. 향후 지식 공학 시스템에 포함되는 데이터 품질 측정 컴포넌트 도구 개발을 완성할 예정이다.

6. 참고 문헌

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

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

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

[4] Won Kim, Byoungju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, Doheon Lee, "A Taxonomy of Dirty Data," *Data Mining and knowledge Discovery*, 2000, accepted

[5] 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

[6] 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, "Chamois: A Component-Based Knowledge Engineering Framework,"*JOOP*, 2001,accepted