오류 데이터로부터의 데이터 품질 메트릭의 정립

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

Establishing Data Quality Metric from Dirty Data

Soo-Kyung Kim° Byoungju Choi
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

소프트웨어 제품의 품질을 보증하는 일은 매우 중요하며, 국제 표준인 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 오류 데이터 형태를 분류하고, 이를 기반으로 데이터 품질 특성을 추출한다. 추출된 데이터 품질 특성을 측정하기 위해, 오류 데이터를 품질 속성으로 하는 데이터 품질 메트릭을 제안한다. 본 논문에서 제시하는 데이터 품질 메트릭은 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 적용되기 어렵다. 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 오류 데이터를 분류하여 그것으로부터 데이터의 품질을 측정 가능토록 하는 데이터 품질 특성을 파악한다. 데이터 품질 특성을 측정하기 위해 오류 데이터를 품질 속성으로 하는 데이터 품질 메트릭을 제안하며 이것은 지식 공학 시스템에서의 데이터 품질 측정 및 제어를 목적으로 한다. 본 논문에서 제시하는 데이터 품질 메트릭은 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 오류 데이터의 형태별 분류를 제안한다. 4장에서는 3장에서 제시된 분류에 ISO/IEC 9126에 적용하여 데이터 품질 특성의 분류를 제안한다. 5장에서는 4장에서 제안된 데이터 품질 특성을 측정하기 위한 오류 데이터를 품질 속성으로 하는 데이터 품질 메트릭을 제안한다. 6장에서는 데이터 품질 메트릭을 적용하여 고안된 데이터 품질 측정 컴포넌트를 제안한다. 7장에서는 결론과 향후 연구 과제를 제시한다.

2. 정의

(1) 오류 데이터

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. 즉, 데이터 자체의 오류로 인해 예상치 못한 잘못된 연산이나 연산 자체가 불가능하게 된 것을 오류 데이터라고 한다. 오류 데이터가 생기는 원인은 실세계의 데이터를 시스템에서 인식 가능한 형태로 변환하여 입력하기 때문이며, 또한 실세계의 데이터는 끊임없이 변화하기 때문이다.

(2) 오류 데이터 항목

본 연구에 앞서 *"successive hierarchical refinement"* 방식을 이용하여 총체적인 오류 데이터의 분류[3](이하 Kim's et al 분류라 칭함)를 구축하였다. 오류 데이터의 항목이란 실제적인 오류 데이터 이름을 뜻하며 Kim's et al분류의 경우 분류 구조의 단말노드(leaf node)에 나타나있다. 품질 특성을 추출하기 위한 첫번째 단계로써, 품질 특성의 관점에서 오류 데이터의 종류를 분류할 필요가 있다. 따라서, 본 연구에서는 Kim's et al분류의 각 단말노드의 오류 데이터 항목들을 데이터 품질 특성 관점에서 다시 분류하였다.

(3) 데이터 품질

데이터 사용자의 요구사항을 충족하는 데이터를 말한다. 즉, 사용자가 특정 작업을 수행할 때, 계속적으로 요구되고 기대되어지는 데이터를 제공해 줄 수 있는 것을 데이터 품질 이라고 정의한다.

(4) 데이터 품질 특성 및 부특성

데이터 품질 특성은 데이터 사용자의 요구사항을 충족하는 각각의 데이터 특성이다. 데이터 품질 부특성은 데이터 품질 특성들이 자세히 세분화되어진 것이다. 데이터 품질 특성의 결합된 결과가 데이터 사용자를 위해 제공되는 데이터의 품질이다.

3. 품질 특성 관점의 오류 데이터 분류

오류 데이터는 ①데이터가 지닌 데이터 자체의 값과 ②그 값이 지닌 각각의 의미, ③데이터가 존재하는 방식의 표현에 관한 세 가지 관점에 의해 나뉜다. 데이터의 값에 관련된 오류의 형태는 내용(content)으로 정의한다. 이 데이터의 값이 어떻게 해석될 지와 연관되어 발생되는 오류의 형태는 정의(definition)로 정의한다. 이 데이터가 어떠한 형태로 표시되는 가에 관련되어 발생하는 오류의 형태는 표현(presentation)으로 정의한다. 따라서, 오류 데이터는 그림 1과 같이 내용, 표현, 정의의 세 가지로 분류될 수 있다.

내용의 경우 데이터의 구조적 측면에서 봤을 때, 데이터의 '구성(syntactic)'과 '의미(semantic)' 오류로 나뉘어진다. 오늘날의 관계형 데이터베이스를 기준으로 구성과 의미의 오류와 동일한 개념에서 발생하는 데이터 오류를 구성과 의미로 구분한다. 따라서, 데이터 내용에 관한 오류로 구성과 의미 오류이외의 데이터 오류는 없으므로 분류는 완전하다.

표현의 경우, 데이터 값들의 표현에서 발생하는 오류를 그 대상으로 한다. 즉, 결과물로 기대되어지지 않은 데이터의 잘못된 표현(wrong representation)을 본 연구에서는 표현(presentation)이라고 정의한다. 표현은 '없어진 데이터(missing)'와 '요구되어지지 않는 데이터의 표현(non-missing, but wrong representation)'으로 나뉘어 진다. 표현의 경우도 없어진 데이터와 없어지지 않은 데이터로 완전히 분류될 수 있다.

정의의 경우, 데이터 값이 원래의 의도에 맞게 사용되어 질 수 있는지에 관한 문제이다. 각각의 데이터가 잘못된 데이터가 아니더라도 원래의 의도에 맞게 사용되어 질 수 없다면 그것은 사용할 수 없는 데이터가 되어 데이터 오류라고 할 수 있다.

품질 특성 관점의 오류 데이터의 항목들은 그림 1에서처럼 이미 연구된 Kim's et al분류[3]의 단말 노드와 일치한다. 데이터 품질 특성 추출을 위해서는 *"successive hierarchical refinement"*방식으로 분류한 Kim's et al분류는 적합치 않다. 따라서, 본 논문은 이미 검증된 Kim's et al분류의 단말 노드를 품질 특성 추출을 위해 품질 특성 관점에서 위에 기술한 오류 데이터 분류 형태에 따라 재배치한다.

데이터 품질 특성 관점에서 분류된 오류 데이터는 그림 1과 같으며, Kim's et al 분류 및 각 오류 데이터 항목별로 방지하는 기술을 같이 대응시켰다.

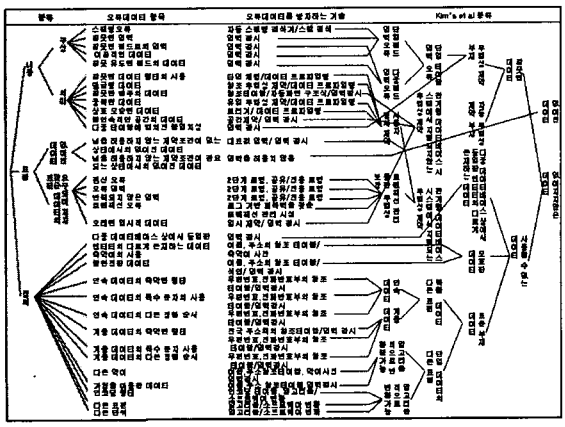

그림 1. 데이터 품질 특성 관점에서의 오류 데이터의 분류

4. 데이터 품질 특성 분류

데이터 품질 특성은 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[4]와 Wang and Strong[5]의 연구를 들 수 있다. 이들 연구 결과를 분석했을 때, 데이터 품질에는 대표적으로 4가지 차원, 즉 정확성(accuracy), 적시성(timeliness), 완료성(completeness), 일관성(consistency)으로 분리하여 품질관련 연구가 이루어짐을 알 수 있다. 그러나, 이 네 가지 특성이 데이터의 품질을 보장하는 모든 특성들이라고 하기에는 너무 제한적이라고 할 수 있다. 따라서 본 연구에서는 오류 데이터와 ISO/IEC 9126의 소프트웨어 품질 특성을 토대로 한 새로운 데이터 품질 특성을 제안한다.

4.1 오류 데이터 항목들부터 데이터 품질 부특성의 추출

데이터 품질 부특성은 직접적으로 오류 데이터 항목으로부터 도출된다. 분류된 데이터 오류 항목 중 구성의 다섯 가지 항목인 스펠링 오류, 잘못된 입력, 잘못된 필드로의 입력, 이질적인 데이터, 잘못 유도된 필드의 데이터 등이 데이터 품질 부특성인 객관성이 된다. 또, 의미와 정의에 해당하는 오류 데이터 항목들로부터 데이터 품질 부특성인 접근성의 품질 특성을 추출할 수 있다. 접근성은 사용자가 특정 데이터를 요구 할 때 제대로 제어되는 지에 관한 데이터 품질 부특성이다. 따라서, 오류 데이터 분류 중 표현과 정의에 관한 항목들로부터 접근성이라는 데이터 품질 부특성을 추출할 수 있었다.

오류 데이터 항목으로부터 직접적으로 추출되어진 부특성들은 객관성, 정확성, 완료성, 접근성, 관련성, 신용성, 이해성, 해석성, 간결한 표현성, 적시성, 일관된 표현성, 검증가능성, 조작 편리성의 13가지이다. 데이터 품질 부특성을 추출하는 오류 데이터 항목은 그림 2와 같으며 각각의 데이터 품질 부특성의 정의는 다음과 같다.

- 객관성(objectivity) : 특정 작업이나 사용자의 목적을 만족시킬 수 있는 데이터의 능력
- 정확성(accuracy) : 요구되는 정확한 데이터를 제공하는 능력
- 완료성(completeness) : 특정 변수에서 요구되는 모든 값을 제공하는 데이터의 능력
- 접근성(accessibility) : 요구되는 모든 데이터를 제어할 수 있는 능력, 즉 사용자가 특정 데이터를 요구할 때 제대로 제어되는 지에 관한 특성
- 관련성(relevance) : 사용자의 특정요구사항을 충족하는 데이터의 능력
- 신용성(believability) : 데이터의 신뢰도 측면에서 믿을 수 있는 데이터
- 이해성(understandability) : 데이터가 사용자가 요구하는 적합한 데이터 인지를 사용자가 이해할 수 있도록 만드는 데이터의 능력
- 해석성(interpretability) : 데이터가 다른 시스템에서 쉽게 사용되어 질 수 있는 능력
- 간결한 표현성(concise representation) : 내용을 충분히 뒷받침하면서 요구되는 크기에 적합한 데이터
- 적시성(timeliness) : 기록된 값이 시간적으로 데이터 사용자가 요구하는 값과 일치하는 데이터의 능력
- 일관된 표현성(consistent representation) : 중복되어지거나 분산된 데이터베이스 환경에서 기록된 값의 표현이 모든 경우에 항상 같은 데이터의 능력
- 검증 가능성(testability) : 유효 검사를 위해 요구되는 데이터의 능력
- 조작 편리성(easy of manipulation) : 데이터를 처음 만들어 사용하는 것의 편리성뿐만 아니라, 그것을 조작하는 것의 편리성의 정도, 즉 오늘날의 관계형 데이터베이스 시스템에서 지원되는 능력을 모두 가진 데이터

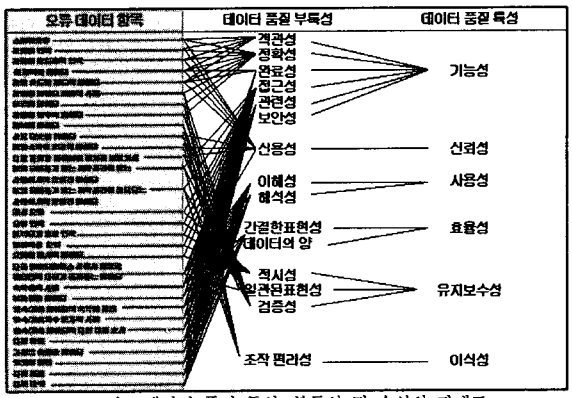

그림2. 데이터 품질 특성, 부특성 및 속성의 관계도

4.2 ISO/IEC 9126 소프트웨어 품질 부특성으로부터 데이터 품질 부특성 추출

앞 절의 13가지 품질 부특성에 ISO/IEC 9126을 토대로 하여 부특성을 보완한다. 앞에서 유도된 부특성들은 오류 데이터 항목들로부터 유도되어진 것이다. 따라서, 오류 데이터는 아니지만 데이터의 품질을 측정하기 위해서 고려되어야 하는 부특성은 제외되어 있다. 이에 대한 보완을 위해 위의 13가지 데이터 품질 부특성과 소프트웨어 품질 부특성을 대응시킨다. 대응의 방법은 각각의 부특성의 정의를 고려하여 유사 정의를 대응하는 방식을 택했다. 대응된 데이터 품질 부특성과 소프트웨어 품질 부특성은 그림 3과 같다.

소프트웨어 특성과 데이터 특성의 차이로 인해 각각의 대응이 1:1로 이루어지지는 않는 점을 고려한다고 해도 오류 데이터 항목으로부터 도출한 위의 13가지 데이터 품질 부특성들로 모든 데이터 사용자의 요구사항을 포함하는 것은 어렵다. 데이터의 품질을 위해 요구되는 사항 중에 소프트웨어 품질 부특성의 보안성(security), 자원 이용도(resource behavior)에 관한 사항이 보완되어야 한다.

오류 데이터 항목으로부터 유도된 13가지 품질 부특성에 소프트웨어 품질 부특성으로부터 보완한 2가지 품질 부특성(보안성, 데이터의 양)을 추가하여 15가지 데이터 품질 부특성을 완성한다. 추가된 두 가지 부특성의 정의는 다음과 같다.

- 보안성(security) : 기밀 문서를 권한이 없는 제어를 하려는 시도를 막거나, 의도되어지지 않은 제어를 방지하려는 데이터의 능력
- 데이터의 양(amount of data) : 데이터의 양

4.3 데이터 품질 특성의 도출

완성된 품질 부특성을 ISO/IEC 9126의 소프트웨어 품질 특성에 대응시켜 그림3과 같은 데이터 품질 특성, 부특성의 구조를 제안한다. 도출된 데이터 품질 특성의 정의는 다음과 같다.

- 기능성(functionality) : 명시된 조건하의 데이터 사용 시, 진술되거나 암시된 기능을 제공하는 데이터의 능력
- 신뢰성(reliability) : 명시된 조건하의 데이터 사용 시, 시스템의 성능 레벨을 만족시키는 데이터의 능력
- 효율성(efficiency) : 명시된 조건하의 데이터 사용 시, 사용되는 자원량과 관련지어서 필요한 성능을 제공하는 능력
- 사용성(usability) : 명시된 조건하의 데이터 사용 시, 사용자가 이해하고 사용하고 좋아할 수 있도록 하는 데이터의 능력
- 유지보수성(maintainability) : 수정시, 유지될 수 있는 데이터의 능력
- 이식성(portability) : 어떤 환경 하에서 다른 환경으로 이식해서 사용할 수 있는 데이터의 능력

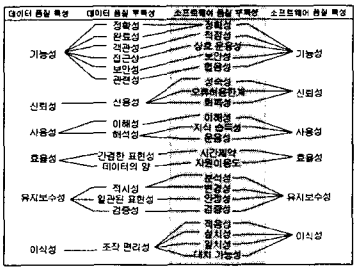

그림 3. 데이터 품질 특성 및 부특성의 구조 및 소프트웨어 품질 특성으로의 대응

5. 데이터 품질 메트릭

소프트웨어 품질 메트릭과 데이터 품질 메트릭은 측정하는 대상이 다르다. ISO/IEC 9126의 소프트웨어 품질 메트릭은 시스템의 프로세스에 초점을 맞추어 품질 메트릭을 만든데 반해, 데이터 품질 메트릭은 시스템을 구동하는 데이터에 초점에 맞추어져야 한다. 예를 들어, 소프트웨어 품질 메트릭에서는 시스템을 구성하는 함수를 측정의 대상으로 한다면, 데이터 품질 메트릭에서는 데이터 측정의 대상이 될 것이다. 데이터 품질 메트릭은 ISO/IEC 9126을 기반으로 만든 품질 메트릭스이다. 그러나, 측정의 대상이 다르기 때문에 품질을 보는 관점은 ISO/IEC 9126과는 달라야 한다.

본 논문에서는 품질 메트릭을 측정하는 속성의 종류에 따라 세가지 형태의 품질 부특성 메트릭이 있다. 첫째, 오류 데이터 항목을 품질을 측정하는 속성으로 사용하여 데이터의 품질을 측정하는 메트릭이다. 둘째, 보안성을 측정하는 메트릭과 마지막으로 데이터 양을 측정하는 품질 메트릭의 세가지로 나눌 수 있다. 품질을 측정하는 데 사용되는 속성인 오류 데이터 항목은 서로 다른 부특성에 중복되어 사용되어 질 수 있다. 데이터 품질 측정 메트릭은 다음과 같이 정의한다. 데이터 품질은 그림 3에서처럼 6개로 품질특성으로 세분화된다. i번째 데이터 품질 특성을 C_i라 할 때, C_i에 대한 품질 측정 메트릭 C_i에 대한 정의는 다음과 같다.

<정의> 데이터 품질 특성 C_i에 대한 품질 측정 메트릭 C_i

S_{ij}	:데이터 품질 특성 C_i에 속하는 품질 부특성 j,
$m_i = \|S_{ij}\|$	:데이터 품질 특성 C_i 의 총 품질 부특성 개수,
a_{ijk}	: S_{ij} 에 속하는 오류 데이터 항목 k,
$\|a_{ij}\|$	: S_{ij} 에 속하는 총 오류 데이터 항목의 개수,
N	:총 품질측정 대상 데이터 개수,
n_{ijk}	: a_{ijk} 의 오류 데이터 개수
L	:불법 제어 기록
K	:불법 제어 시도
S	:사용된 데이터의 크기
T	:전체 데이터의 크기

$$C_i = w_j \sum_{j=1}^{m_i} S_{ij}, \quad w_j = \frac{1}{m_i}$$

$$S_{ij} = \frac{1}{|a_{ij}|} \sum_{k=1}^{|a_{ij}|} \left(1 - \frac{n_{ijk}}{N}\right)$$ (단, S_{ij}는 오류 데이터가 품질속성인 품질 메트릭)

$$= \frac{L}{K}$$ (단, S_{ij}가 보안성을 측정하는 품질 메트릭)

$$= 1 - 2 \times \left| \frac{S}{T} - \frac{1}{2} \right|$$ (단, S_{ij}가 데이터의 양을 측정하는 품질 메트릭)

표1. 사용성의 데이터 품질 특성의 측정

품질 부특성	오류 데이터 항목	오류 데이터수
이해성	다중 데이터베이스 상에서 동일한 엔티티의 다르게 존재하는 데이터	0
	축약어의 사용	1
	불완전한 데이터	2
	다른 약어	0
	가정을 이용한 데이터	1
	인코딩 형태	3
	다른 표현	0
	다른 단위	0
	연속/계층 데이터의 축약된 형태	5
	연속/계층 데이터의 특수 문자 사용	4
	연속/계층 데이터의 다른 정렬 순서	3
해석성	축약어의 사용	1
	가정을 이용한 데이터	1
	다른 약어	0
	인코딩 형태	3
	다른 표현	0
	다른 단위	0

데이터 품질 특성 가운데 사용성 (즉, $C_{사용성}$)의 품질 측정치를 측정하는 경우를 예를 들어보자. 품질 측정 대상 데이터웨어하우스의 총 데이터 개수를 N = 100개라고 하자. 그림 3에서 사용성은 이해성, 해석성의 총 2개의 부특성을 가지고 있으므로 m_i = 2이다. 이중 두 개의 각 부특성에 속하는 모든 오류데이터 항목에 대한 오류데이터 개수를 표 1과 같다고 하면 $C_{사용성}$의 품질 측정치는 메트릭 $C_{사용성}$ 에 의해 다음과 같이 계산되어 98.5%가 된다:

$$C_{사용성} = 1/2 \times (0.98 + 0.99) = 98.5\ \%$$

6. 데이터 품질 측정 컴포넌트

본 연구에서 개발하는 데이터 품질 측정 컴포넌트, DQMC(Data Quality Measuring Component),는 Charmois 지식 공학 시스템에 포함될 예정이다. 이중, 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 구현되며, 그림 4와 같이 ETL 도구인 Power Mart의 COM 외부 처리(external procedures)라는 타 개발 측에서 개발한 제품을 소비자가 Power Mart의 라이브러리에 끼워서 쓸 수 있도록 하는 기능을 이용하여 개발된다.

DQMC의 알고리즘은 다음과 같다. 일단 품질 측정에 대한 요구가 들어오면 일단은 오류 데이터의 각각의 항목에 대해 데이터 품질 부특성을 계산된다. 다시 계산된 데이터 품질 부특성을 가지고 데이터 품질 특성을 수치화 시킬 수 있다.

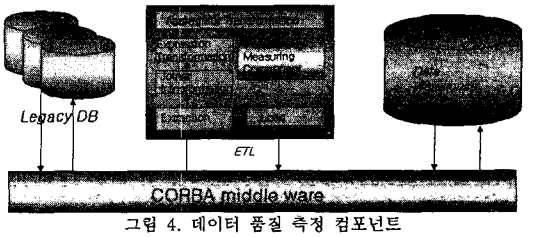

그림 4. 데이터 품질 측정 컴포넌트

7. 결론 및 향후 연구 과제

데이터 품질 자체에 대한 연구를 처음으로 체계적으로 진행시켰다는 점에서 이 논문의 의의를 찾을 수 있다. 이 논문은 향후 데이터 품질에 대한 표준을 정립하는 것을 목표로 한다. 본 논문은 다양한 데이터 소스들로부터 의미 있는 데이터나 나아가 지식들을 추출하는 지식 공학 시스템에서의 데이터 품질을 보장하는 것을 목적으로 한다. 따라서, 본 논문에서는 먼저 품질 특성 관점에서 오류 데이터를 분류하고, 이를 기반으로 데이터 품질 특성을 분류하였다. 분류한 각 오류 데이터의 항목에서 추출한 품질 부특성에 ISO/IEC 9126에 정의한 소프트웨어 품질 특성 및 부특성을 대응시킴으로써 데이터 품질 특성 및 부특성을 추출할 수 있었다. 데이터 품질 측정을 위해서 분류한 각 데이터 품질 특성 및 부특성에 대한 메트릭을 구축하였다. 향후 이 데이터 품질 메트릭을 적용하여 현재 프레임만 완성된 상태의 데이터 품질 측정 컴포넌트(DQMC)의 도구 개발을 완성할 예정이다. DQMC는 소프트웨어 공학의 컴포넌트 구조의 개념을 도입하여 체계적으로 컴포넌트의 역할을 할 수 있도록 구현할 것이다. 그러나, 이 모든 연구의 최종 목표는 동 산업계의 표준으로 사용될 수 있는 데이터 품질 메트릭의 정립이며, 이를 위해 실제 데이터 웨어하우스에의 사례 구축을 통하여 본 연구에서 제시한 데이터 품질 측정 방안을 적용한 결과를 통하여 검증할 계획이다.

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

[4]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

[5]Diane M. Strong, Yang W. Lee, and Richard Y. Wang, "Data Quality in context," *Comm. of the ACM*, vol. 40, no. 5, May 1997