데이타 품질 측정 도구
(A Data Quality Measuring Tool)

양 자 영 † 최 병 주 ††
(Jayoung Yang) (Byoungju Choi)

요 약 소프트웨어 제품을 실행시키기 위해 요구되는 데이타의 품질은 소프트웨어 품질에 영향을 미치고 있다. 특히 대용량의 데이타로부터 의미 있는 지식을 추출하는 지식공학 시스템에서 원시 데이타의 품질을 보장하는 일은 매우 중요하다. 본 논문에서는 데이타의 측정 도구인 DAQUM도구를 설계 구현하였다.본 논문에서는 DAQUM도구의 설계 및 구현에 관한 주요내용을 기술하고, 사례연구를 통하여 DAQUM도구가 오류데이타를 검색하여 데이타 사용자 관점에서 데이타의 품질을 정량적으로 측정 가능하도록 함을 나타낸다. DAQUM도구는 데이타의 품질 측정 및 품질 제어를 가능하게 함으로써 데이타를 주로 처리하는 소프트웨어 제품의 품질 향상에 기여할 수 있다.

키워드 : 품질측정도구, 오류데이타, 데이타품질

Abstract Quality of the software is affected by quality of data required for operating the actual software. Especially, it is important that assure the quality of data in a knowledge-engineering system that extracts the meaningful knowledge from stored data. In this paper, we developed DAQUM tool that can measure quality of data. This paper shows: 1) main contents for implement of DAQUM tool; 2) detection of dirty data via DAQUM tool through case study and measurement of data quality which is quantifiable from end-user's point of view. DAQUM tool will greatly contribute to improving quality of software product that processes mainly the data through control and measurement of data quality.

Key words : Quality Measurement Tool, Dirty Data, Data Quality

1. 서 론

소프트웨어 제품이란 사용자에게 인도할 것으로 지정된 컴퓨터 프로그램, 절차와 관련 문서 및 데이타의 전체 집합[1]이다. IEEE과 ISO 등의 표준에서는 소프트웨어를 프로그램, 절차, 규칙 및 관련문서로 한정하고 있으나, 실제로는 소프트웨어 제품을 실행시키기 위해 요구되는 데이타의 품질 또한 소프트웨어 품질에 영향을 미치고 있다. 즉, 소프트웨어 시스템에서 얻어지는 최상의 결과를 얻기 위해서는 데이타 품질의 측정이 요구된다. 그러나, 아직까지 이를 정량적으로 측정하기 위해 제시된 데이타 품질 측정 모형 및 도구에 관한 연구는 미비할 뿐만 아니라, 현재 데이타 품질 측정과 관련된 국제적인 표준도 없는 실정이다.

인터넷/웹 기술이 급속히 발달함에 따라 방대한 양의 정보 수집이 가능하게 되었고, 이러한 정보로부터 고부가가치를 신속하게 창출하는 일은 기업의 장래를 위해 필수가 되었다. 고부가가치 창출을 위해서는 데이타를 체계적으로 신속하게 수집, 저장, 관리, 분석하여 지식을 추출할 수 있어야 한다. 이러한 역할을 하는 것이 지식공학 시스템[2,3,4]이다. 지식공학 시스템의 초기 입력이 되는 데이타의 품질은 사용자에게 제공되는 데이타나 지식의 품질을 결정하는 중요한 요인이 된다. 만일 지식공학 시스템에 데이타 품질을 제어하는 기술이 없다면, 신뢰할 수 없는 데이타나 지식을 사용자에게 제공하게 되므로, 지식 공학 시스템 자체의 존재가 무의미하게 될 것이다[5,6].

본 논문에서는 지식공학 시스템에서의 소프트웨어 품질의 데이타를 이용한 상세 측정에 기여할 수 있는 데

· 본 연구는 대학 IT연구센터 육성. 지원 사업의 연구결과로 수행되었음.
† 비 회 원 : 이화여자대학교 컴퓨터학과
komi@ewha.ac.kr
†† 종신회원 : 이화여자대학교 컴퓨터학과 교수
bjchoi@ewha.ac.kr
논문접수 : 2002년 9월 12일
심사완료 : 2003년 3월 10일

이타의 품질 측정 도구: DAQUM(DAta QUality Measurement)도구의 설계 및 구현에 관하여 기술한다. 본 연구진은 본 연구에 앞서 오류 데이타(Dirty Data)를 "successive hierarchical refinement" 방식으로 분류하였다. 오류 데이타가 생기는 이유를 "Missing data", "Not missing, but wrong data", "Not wrong, but unusable data"의 세 가지로 분류한 후, 이들을 계층적으로 분해하는 방식으로 총 33개의 오류 데이타를 분류하였다. DAQUM도구는 오류데이타를 검색하여 데이타의 품질을 정량적으로 측정할 수 있도록 한다. 측정된 데이타 품질 값은 사용자에게 대용량의 데이타로부터 추출된 지식에 대한 신뢰성을 판단하는데 도움을 줄 수 있도록 하며, 궁극적으로는 소프트웨어 제품 품질 향상에 공헌할 수 있게 한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이타 품질 측정 관련연구를 기술하며, 3장에서는 DAQUM도구 구현에 필요한 주요 내용을 기술하며, 4장에서는 DAQUM도구의 설계와 구현 및 실행사례를 기술한다. 5장에서는 DAQUM도구를 이용한 품질 측정 사례연구를 기술하고, 마지막으로 6장에서는 결론과 향후 연구 과제를 제시한다.

2. 관련 연구

본 절에서는 데이타 품질 측정에 필요한 요소인 오류 데이타 정의 및 오류 데이타 분류 체계에 관하여 수행했던 연구에 대해서 기술하고, 데이타 품질 측정에 관한 기존 연구로써 데이타 소비자의 업무 상황(task context)을 고려한 데이타 품질 평가와 데이타웨어하우스 환경을 고려하는 데이타 품질 평가 방법에 대해서 기술한다.

2.1 오류 데이타

본 논문에 앞서, 오류 데이타를 "successive hierarchical refinement" 방식으로 분류[5]하였다. 오류 데이타가 생기는 이유를 "Missing data", "Not missing, but wrong data", "Not wrong, but unusable data"의 세 가지로 나누어 그림 1에서처럼 총 33개의 오류 데이타를 분류하였다.

그림 1의 오류 데이타 분류 구조(taxonomy of dirty data)의 단말노드가 실제적인 오류 데이타의 이름을 뜻한다. 다른 관점으로부터 시작한 오류 데이타의 분리는 다른 구조를 지닐 수 있으나, 각 분류의 오류 데이타의 형태를 구체적으로 나타내는 단말노드는 동일함을 다른 분류 구조의 전개를 통해 증명하였다. 또한 [7]에서는 그림 1의 오류 데이타는 실제 가능한 오류 데이타의 종류에 대하여 95%이상 확실함을 보였다. 본 논문에서는 그림 1의 오류 데이타 종류 별로 오류 데이타를 파악하여 데이타 품질을 측정한다.

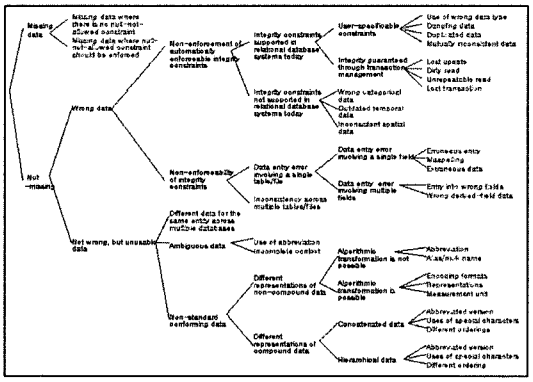

그림 1 오류 데이타의 분류 구조

2.2 데이타 품질 측정

데이타 품질 연구는 소프트웨어 품질 연구와는 달리 아직 표준이 정립되어 있지 않고, 각기 필요성에 따라 조금씩 연구가 진행되어 왔다. 이중 가장 대표적인 것으로 Wang의 연구[8]가 있다. Wang은 품질 관리 방법론인 TQM을 데이타에 적용한 TDQM(Total Data Quality Management)을 제시하였다. Wang의 TDQM 생명주기에 따르면 TDQM단계는 크게 네 가지: 정의, 측정, 분석, 향상의 단계로 이루어 진다. 그러나, 현재 데이타 품질 관련 연구는 정의단계도 확립하지 못한 상태이다.

TDQM에 따라 정의단계를 수행하기 위해서는 사용자가 데이타 품질에 대해 요구하는 데이타의 특성에 관한 연구가 요구된다. 이에 관한 연구 중에서 가장 대표적으로 것으로 Ballou의 연구[9]와 Wang의 연구[10]가 있다. 이들의 연구 결과를 분석했을 때, 데이타 품질에는 대표적으로 파악되는 4가지 특성: 정확성, 적시성, 완료성, 일관성이 있다. 그러나, 이들 데이타 품질 특성이 사용자가 데이타에 대해 요구하는 품질을 보장하는 모든 특성들을 총 망라하고 있다고 하기에는 그 범위가 너무 제한적이라고 할 수 있다.

이들 기존연구[6,8,9,10,11,12]에서는 업무 상황(task context)이나 업무 프로세스 및 데이타 사용이 사용자마다 다르므로, 데이타 사용자를 고려한 품질 평가를 중요성을 제시하였다. 그러나 사용자 관점에서의 데이타 품질 측정의 필요성을 이론적으로만 제시하였다.

데이타베이스안에 저장되어 있는 정보의 정확성, 일치성, 완전성, 무결성등을 관리해주는 데이타 품질 관리 도구는 Dataflux, Firstlogic, Trillium, Assential 사등

다양한 제품이 있다. 이들 도구들의 공통점은 전체 데이타베이스를 검색하여 설정된 값 및 규칙에 어긋난 오류 데이타를 검색하여 발생 건수를 리스트로 보여준다. 즉, 이 도구들에서는 업무상황 및 사용자의요구사항이 다르더라도 오류 데이타 발견 건수 리스트는 동일하다. 이 도구들은 사용자의 관점을 고려하지 않고 품질을 측정하였다.

본 논문은 TDQM의 생명 주기에서 정의 및 측정 단계까지 실제 데이타를 사용하는 사용자가 어떤 목적으로 데이타를 사용했는가를 고려하여 동일한 데이타라 하더라도 사용자의 관점에서 품질이 다르게 평가될 수 있도록 DAQUM도구를 설계 구현하였다.

3. 데이타 품질 측정

본 절에서는 데이타 품질 측정을 위한 절차에 대해서 기술한다. 특히 이 절차에 포함되어 있는 핵심 부분인 DAQUM도구의 설계 및 구현에 필요한 오류데이타 검색 기술과 데이타 품질 측정 메트릭스에 대해 상세히 기술한다.

3.1 데이타 품질 측정 절차

데이타 품질 측정 절차는 그림 2와 같다. 먼저 도메인 전문가가 데이타 웨어하우스의 요구 사항에 따라 데이타 구조 및 내용에 대한 제약사항을 결정한다. 데이타 정보가 담겨 있는 데이타 프로파일을 구축한다. 다음 단계는 이를 참고하여 컬럼별로 오류 데이타 발견 알고리즘에 의해 오류 데이타를 검색한다. 다음 단계는 데이타 사용 목적에 따라 이와 관련 있는 컬럼과 관련 없는 컬럼을 구분한다. 마지막으로 데이타 품질 메트릭스에 의해 이 두 그룹에 대해 각각 컬럼별로 전체 데이타 총수에서 오류 데이타가 발생한 비율을 계산하여 전체 데이타 품질을 파악한다.

데이타 품질 측정 절차에 대한 정보는 다음과 같다.

Input : 도메인 요구사항
Output : 오류 데이타 수, 데이타 품질 측정값
Step
1. 비즈니스 도메인 분석
   - 요구사항에 맞는 비즈니스 규칙 및 제약사항 결정
   - 데이타 프로파일 구축
2. 오류 데이타 항목 발견
   - 컬럼별로 오류 데이타 수 파악
3. 데이타 사용 목적 선택에 따란 관련 항복과 관련 없는 항목 결정
   - 사용 목적과 관련 있는 컬럼과 관련 없는 컬럼을 나누어 그룹화
4. 데이타 품질 측정
   - 메트릭스를 통해 데이타 품질 측정값을 수치화

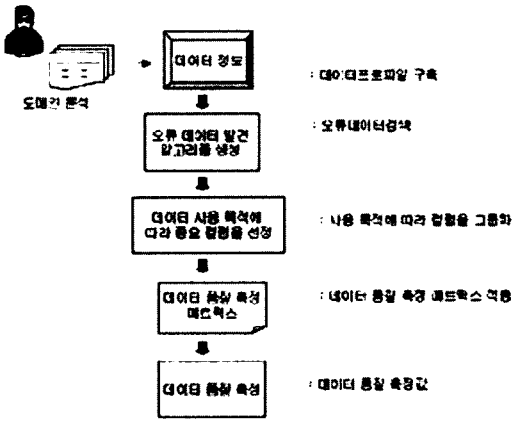

그림 2 데이타 품질 측정 절차

3.2 오류데이타 검색

오류 데이타 검색 프레임웍은 그림 3과 같다. 도메인 전문가가 데이타 베이스의 데이타 내용과 구조에 대한 정보 및 사용자 비즈니스 규칙에 대한 정보를 저장하기 위해 데이타 프로파일을 구축한다. 오류 데이타 검색시 필요한 카테고리, 스펠링 체크, 축약어 사전, 약어 사전, 백과 사전, 이름, 주소, 전화번호, 우편번호 등의 참조 테이블을 미리 구축한 후 품질 측정 대상 데이타 베이스와 참조 테이블간의 참조 관계를 설정한다. 데이타 프로파일에 저장된 데이타 정보와 참조 테이블의 관계를 참고하여 오류 데이타 항목별로 오류데이타를 검색한다.

그림 1의 33가지 오류 데이타는 그림 4처럼 1) 오류 데이타를 도메인 전문가의 중재에 의해 수동으로 검색

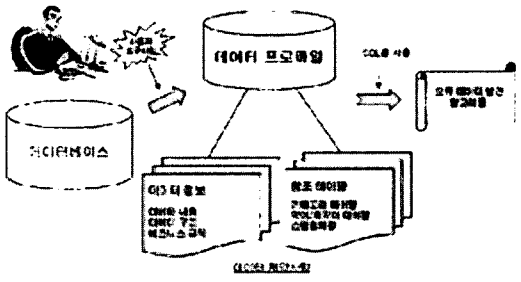

그림 3 오류데이타 검색 프레임웍

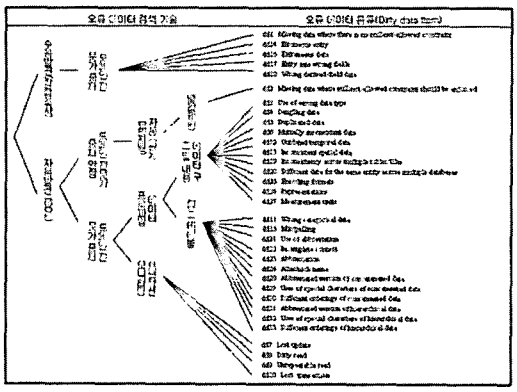

그림 4 오류 데이타 검색 방법

하는 경우와 2) SQL을 사용하여 오류 데이타를 자동으로 검색할 수 있는 경우로 나누어진다.

도메인 전문가가 직접 데이타를 조사하여 오류 데이타를 수동으로 검색해야 하는 데이타 오류는 5가지 종류(dd1, dd14, dd16, dd17, dd18)이다(번호는 그림 1의 오류데이타 종류 번호와 일치한다).

오류 데이타를 자동으로 검색할 수 있는 오류데이타 종류는 총 28가지이다. 이들은 1)도메인 전문가의 중재 없이 자동으로 검색할 수 있는 경우 2)도메인 전문가의 중재에 의해 자동으로 검색할 수 있는 경우로 분류된다.

도메인 전문가의 중재 없이 오류데이타를 자동으로 발견할 수 있는 경우는 1가지(dd2)이며, 다음과 같이 관계형 데이타 베이스의 "널값 체킹"으로 오류 데이타의 발견이 가능하다.

Select count(*) from <테이블명> where <컬럼명> is null

도메인 전문가에의한 자동 검색은 그림 3에서 나타내었듯이 구축된 데이타에 대한 내용, 구조 정보, 비즈니스 규칙, 참조 테이블 등을 저장한 "데이타 프로파일"과 "트랜잭션 모티터링" 등을 참조하여 이루어진다. 도메인 전문가에 의한 자동 검색이 가능한 오류 데이타는 27가지이다. 이것은 다시 1)데이타에 대한 내용 및 구조에 대한 제약사항 및 비즈니스 규칙을 기술한 데이타 프로파일 참조에 의한 검색(dd3, dd4, dd5, dd6, dd12, dd13, dd19, dd20, dd25, dd26, dd27), 2) 카테고리, 스펠링 체킹, 축약어 사전, 약어 사전, 백과 사전, 이름, 주소, 전화번호, 우편번호 등의 참조 테이블에 의한 검색(dd11, dd15, dd21, dd22, dd23, dd24, dd28, dd29, dd30, dd31, dd32, dd33), 3)트랜잭션 모니터링에 의한 검색(dd7, dd8, dd9, dd10)으로 나눌 수 있다. 이들 세 가지 도메인 전문가 중재에 의한 데이타 오류 검색은 다음과 같다.

(1) 내용 및 구조에 관한 데이타 프로파일에 의한 검색

11가지(dd3, dd4, dd5, dd6, dd12, dd13, dd19, dd20, dd25, dd26, dd27)의 오류데이타의 검색이 해당한다. 예를 들면, "Employee" 테이블의 "age" 컬럼에서 타당한 데이타의 범위를 [18-65]라고 하면, 도메인 전문가가 이 값의 범위 제약사항을 데이타 프로파일에 저장한다. 이 데이타 프로파일을 참조하여 범위 체크 연산자에 의해서 제약사항을 위배한 오류 데이타들을 다음과 같이 자동으로 검색한다.

Select count(*) from <테이블명> where <컬럼명1> not like '[18-65]'

(2) 참조 테이블 데이타 프로파일에 의한 검색

참조 테이블을 사용하여 검색하는 경우는 조인 방법에서 차이가 있으므로 이 가운데 오류데이타 dd3, dd21, dd23의 예를 들어 기술하면 다음과 같다.

오류데이타 3: "Wrong categorical data" 경우, US의 수입상태는 "super-rich", "middle-income", "poor", "poverty-stricken" 등의 카테고리 데이타를 갖는 경우를 예를 들면, 도메인 전문가는 분류된 데이타들에 대한 카테고리 참조 테이블을 저장한다. 이 카테고리 참조 테이블과 데이타 품질 측정 대상 테이블과의 조인을 통해 카테고리에 해당 되지 않는 오류 데이타를 다음과 같이 검색한다.

Select count(*) from <테이블명> where <테이블명>.<컬럼명> NOT IN(Select <컬럼명> from <카테고리테이블명>)

오류 데이타21: "Use of abbreviation"의 경우, 예를 들어 "Dr"이라는 축약어가 사용 되었을 경우 이 "Dr"이 "Doctor"를 의미하는지 "Drive"를 의미하는지 모호하기 때문에 발생되는 오류를 검색하기 위해서 축약어 테이블을 만들어 놓는다. 이 축약어 테이블과 측정 대상 테이블과의 조인을 통해 이 축약어 때문에 발생되는 오류 데이타를 다음과 같이 검색한다.

Select count(*) from <테이블명> where <컬럼명> IN(select <컬럼명> from <축약어 테이블>)

오류 데이타23: "Abbreviation"의 예를 들면, "high-way"에 대해 "hwy"이라는 약어를 사용하여 발생되는 오류를 발견하기 위해서 약어 테이블을 만들어 놓는다. 이 약어 테이블과 데이타 품질 측정 대상 테이블과의 조인을 통해 약어 때문에 발생되는 오류 데이타를 다음과 같이 검색한다.

Select count(*) form <테이블명> Inner Join <약어테이블> on <테이블명>. <컬럼명> =<약어테이블>.<컬럼명>

(3) 트랜잭션 모니터링에 의한 검색

DAQUM도구 사용자가 미리 데이타 베이스에서 지원하는 트랜잭션 처리에 대한 제약 사항을 설정토록 하여 트랜잭션 처리 모니터링을 이용하여 트랜잭션의 상태를 파악하여 오류를 검색한다.

3.3 데이타 품질 측정 메트릭스

DAQUM도구는 사용목적에 따라 데이타 품질을 측정할 수 있도록 한다(사용목적에 따른 데이타 품질 측정에 대한 예제는 4.2절과 5.2절을 참고하도록 하자). DAQUM도구는 데이타의 사용목적에 관련 있는 컬럼과 관련 없는 컬럼으로 구분하여 이 두 그룹에 대해 각각 컬럼 별 전체 데이타 총수에서 오류 데이타가 발생한 비율을 계산하여 전체 데이타의 품질을 파악한다.

데이타 사용 목적과 관련이 있는 컬럼에 대한 데이타 품질 측정 메트릭스를 R-Metric이라 하고, 데이타 사용 목적과 관련이 없는 컬럼에 대한 데이타 품질 측정 메트릭스를 NR-Metric라고 했을 때, 전체 데이타 품질 측정 메트릭스 Q-Metric는 다음과 같다.

[R-Metric]

$$R=1-\left(w_i \sum_{k=1}^{m} \frac{D_{ik}}{C_{ik}} \right),\ w_i=\frac{1}{m}$$

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

[NR-Metric]

$$NR=1-\left(w_j \sum_{k=1}^{n} \frac{ND_{ik}}{C_{ik}} \right),\ w_j=\frac{1}{n}$$

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

[Q-Metric]

$$Q=R\times NR$$

Q : R그룹과 NR그룹의 품질값을 반영한 전체 데이터 품질 측정값(0<= Q <=1)
R : 데이터 사용 목적과 관련 있는 컬럼들의 데이터 품질 측정값(0<= R <=1)
NR : 데이터 사용 목적과 관련 없는 컬럼들의 데이터 품질 측정값(0<= NR <=1)
D_{ik} : 데이터 사용 목적과 관련 있는 컬럼의 오류 데이터 개수($D_{ik} < C_{ik}$)
ND_{ik} : 데이터 사용 목적과 관련 없는 컬럼의 오류 데이터 개수($ND_{ik} < C_{ik}$)
C_{ik} : 측정 대상 총 데이터 개수
m : 데이터 사용 목적과 관련 있는 컬럼 수
n : 데이터 사용 목적과 관련 없는 컬럼 수
a_{ij} : 데이터 품질 측정에 필요한 총오류 데이터 항목의 개수
a_{ik} : 데이터 품질 측정에 필요한 오류 데이터 항목 k

4. DAQUM도구

데이타 품질 측정 도구인 DAQUM(DAta QUality Measurement)도구는 Windows Professional 2000 환경에서 데이타베이스는 SQL Server2000을 이용하였고 개발 언어로는 Java2 JDK 1.3.1(Java Development Kit)으로 구현하였다.

4.1 DAQUM도구의 구조

DAQUM도구의 구조는 그림 5와 같으며 "Source DB", "Repository DB", "Target DB"의 데이타베이스와 "인터페이스", "데이타 관리", "데이타 품질 측정", "품질 측정 결과 분석기"의 패키지로 구성한다.

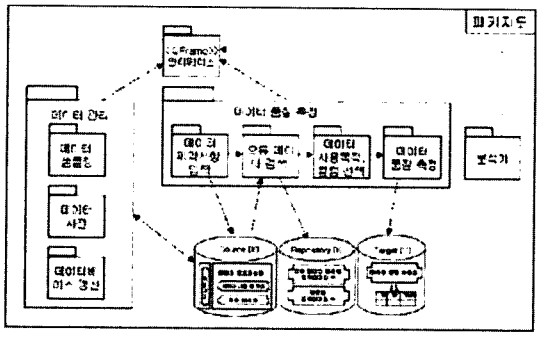

그림 5 DAQUM 도구의 구조

(1) 데이타 베이스

DAQUM도구는 그림 5처럼 3개의 데이타베이스, "Source DB", "Repository DB", "Target DB"로 구성한다. "Source DB"는 데이타 품질 측정 대상 데이타와 3.1절의 오류 검색에 필요한 데이타 프로파일을 저장한다. DAQUM도구는 3.1절의 데이타 오류 검색 방법에 따라 오류 데이타를 검색하여 "Repository DB"에 저장하고, 3.2절의 데이타 품질 측정 메트릭스에 따라 측정한 오류 데이타 품질 측정값을 "Target DB"에 저장한다.

(2) 패키지

- 인터페이스(Interface) 패키지
  데이타베이스 갱신, 샘플링, 사전 등의 데이타베이스 관리, 데이타 제약사항의 입력, 데터 사용 목적과 관련 있는 컬럼의 선택 등의 데이타 품질 측정과 관련한 사용자와 DAQUM도구와의 인터페이스를 제공한다.
- 데이타 품질 측정(Data Quality Measure) 패키지
  데이타 제약사항의 입력, 오류 데이타의 검색, 데이타 사용목적 관련 컬럼의 선택, 데이타 품질의 측정을 한다.
- 품질 측정 결과 분석기(Analyzer) 패키지
  오류 데이타 수의 비율, 사용 목적과 관련 있는 컬럼과 관련 없는 컬럼을 구분하여 두개의 그룹으로 측정한 데이타 품질 측정 결과와 전체 데이타 품질 측정 결과를 분석하여 보여준다.
- 데이타 관리(Data Management) 패키지

데이타 샘플링, 데이타 사전관리, 데이타 베이스 갱신 등 데이타 관리를 한다.

4.2 DAQUM도구 주요기능 및 실행 사례

(1) 측정 대상 소스테이블 및 저장테이블 선택 기능

사용자는 DAQUM도구의 초기화면에서 데이타베이스에 연결하기 위해서 그림 6의 "Connect" 메뉴를 선택하여 테이블이 저장되어 있는 SQL Server 접속 정보를 입력한다. 데이타베이스에 연결되면 "Source DB"의 측정 대상 테이블, "Repository DB"의 오류 데이타 종류별 및 컬럼 별로 검색한 오류 데이타 수를 저장할 테이블, "Target DB"의 데이타 품질 측정값을 저장할 테이블을 선택한다.

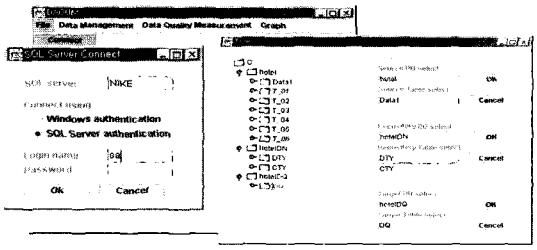

그림 6 측정 대상인 소스 테이블 및 저장 테이블 선택

(2) 데이타 제약 사항 입력 기능

사용자가 입력한 제약사항에 따라 오류 데이타 검색이 가능하도록 한다. 사용자는 그림 7처럼 사용자가 원하는 데이타 타입, 데이타 값, 널값 허용 여부 등의 정보("Value Setting"), 키관계 정보 ("Key Relationship"), 측정 대상 테이블과 참조테이블과의 조인 관계 등에 대한 정보 ("Join Relationship")를 입력한다. 이 제약사항들은 "Source DB"의 데이타 프로파일에 저장된다.

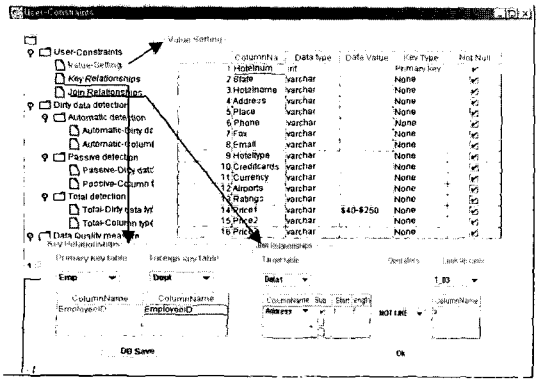

그림 7 데이타 제약사항 입력 화면

(3) 오류 데이타 검색 기능

오류 데이타는 3.1절에 기술한 것처럼 1) SQL을 사용해서 자동으로 검색되는 오류 데이타와 2) 도메인 전문가가 직접 조사를 하여 수동으로 검색되는 오류 데이타가 있다.

1) 오류 데이타의 자동 검색

사용자는 그림 8처럼 오류 데이타 검색 버튼을 선택하여 오류 데이타를 자동으로 검색한다. 사용자가 검색하고 싶은 오류 데이타 종류(①)나 컬럼(②)을 선택한 후, ①, ②버튼을 누르면 오류 데이타 및 오류 데이타 수를 파악 할 수 있다. 사용자가 ③, ④버튼을 선택하면 오류 데이타종류별(③-1), 컬럼별(④-1)로 오류 데이타 수를 파악할 수 있다.

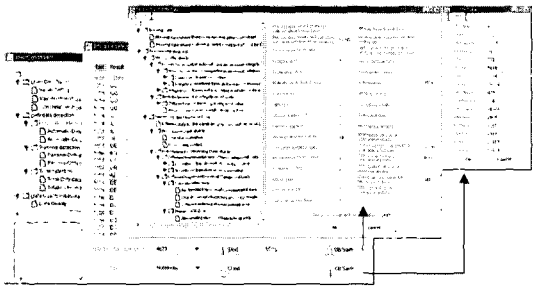

그림 8 오류 데이타 자동 검색 화면

2) 오류 데이타의 수동 검색

사용자는 그림 9처럼 직접 마우스로 데이타를 선택하여 오류 데이타를 수동으로 검색한다. 사용자가 오류 데이타를 선택하면 해당 컬럼에 오류 데이타의 수가 반영된다.(⑤) 사용자가 선택한 오류 데이타의 오류 데이타 종류를 반영하기 위해 사용자가 "dirty type"버튼 눌러 오류 데이타 종류를 선택(⑥)하면 오류 데이타 종류에 오류 데이타의 수가 반영된다.

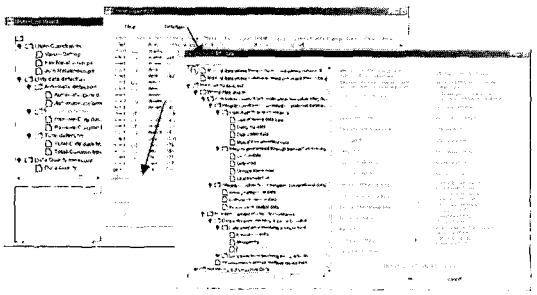

그림 9 오류 데이타 수동 검색 화면

3) 전체 오류데이타 현황 출력

전체 오류데이타의 총개수는 자동으로 검색된 오류데이타 개수와 수동으로 검색된 오류 데이타 개수의 합계이다. 사용자는 그림 10처럼 오류 데이타별로(⑦), 컬럼별로(⑧)로 전체 오류데이타 현황을 파악 할 수 있다.

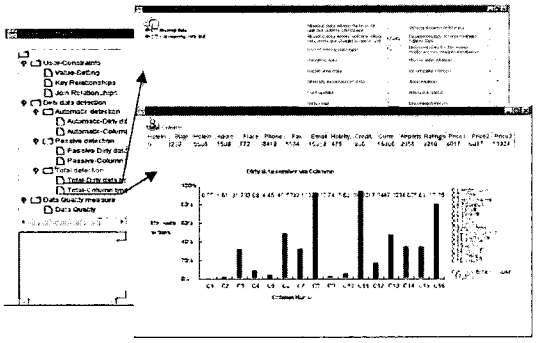

그림 10 전체 오류 데이타 검색 화면

(4) 데이타 사용목적 관련 컬럼 선택 기능

DAQUM도구는 데이타의 사용목적에 따라 데이타의 품질 측정이 가능하도록 한다. 사용자는 그림 11처럼 직접 데이타 사용 목적과 이와 관련 있는 컬럼을 선택한다.

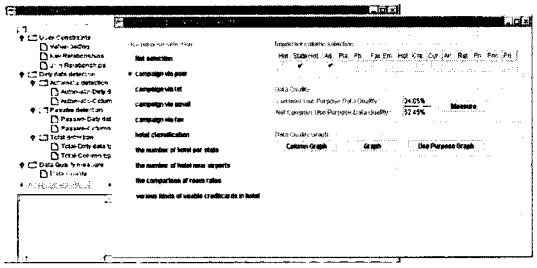

그림 11 사용 목적 및 관련 컬럼 선택

(5) 데이타 품질 측정 기능

3.2절에서 기술했던 데이타 품질 측정 메트릭스를 적용한 DAQUM도구는 데이타 사용목적과 관련 있는 컬럼과 관련 없는 컬럼을 구분하여 이들 각각에 대해 품질을 측정한다. 사용자는 그림 12처럼 그래프 버튼을 선택하여 각 컬럼별 오류 데이타 비율(①), 전체 데이타 품질 측정 값(②), 데이타 사용목적에 따른 데이타 품질 측정 값(③)을 파악할 수 있다.

(6) 데이타 관리 기능

DAQUM도구는 데이타 샘플링, 데이타 사전 관리, 데이타베이스 업데이트 등의 데이타 관리 기능을 제공한다.

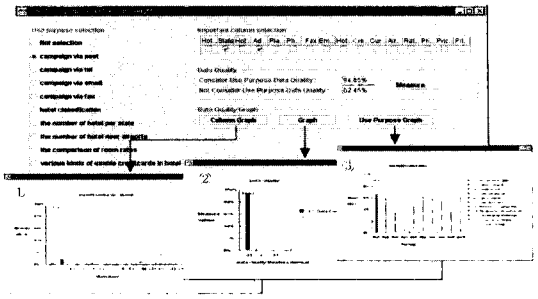

그림 12 데이타 품질 측정

1) 데이타 샘플링 기능

데이타웨어하우스의 전체 데이타 가운데 샘플링을 하여 데이타 품질을 측정할 수 있다. DAQUM도구는 랜덤(Random)함수를 이용하여 사용자가 측정하기를 원하는 샘플링할 개수만큼 레코드가 중복 선택되지 않으면서 랜덤하게 레코드를 추출한다.

2) 데이타 사전 관리 기능

사용자가 직접 사용자의 비즈니스 규칙에 맞게 비교할 수 있도록 데이타 사전 관리 기능을 제공한다

3) 데이타베이스 업데이트 기능

데이타 품질 측정 대상 데이타베이스에 새로운 데이타를 추가하거나 오류가 있는 데이타를 수정하거나 필요 없는 데이타를 삭제 할수 있도록 데이타베이스 업데이트 기능을 갖는다.

5. 사례연구

DAQUM도구의 사례연구를 위하여 웹사이트에서 받은 호텔정보에 대한 데이타를 대상으로 DAQUM도구를 실제 구동하여 데이타 품질 측정을 수행하였다. 호텔 데이타베이스는 16개의 컬럼과 17,350개의 레코드를 가지고 있다. 즉 전체 데이타의 개수는 16*17,350=277,600개이다. 사례연구로써 수행한 내용은 다음의 두 가지이다.

1) 오류 데이타 종류별로 검색한 오류 데이타 개수 분포

2) 다양한 데이타 사용 목적에 따른 품질 측정 값 비교

5.1 오류 데이타 종류별 오류 검색 데이타 개수

사례 대상 호텔 데이타베이스에서 검색한 오류 데이타 수는 표 1과 같다. DAQUM도구에 의해 전체 오류 데이타 수는 37,100개가 검색되었으며 이중에서 "Missing data where null-not-allowed constraint should be enforced"(dd2)오류는 17,349개가 검색되었다.

그림 13은 표 1의 오류 데이타 검색 개수를 오류 종

표 1 오류 데이타(dd: dirty data) 종류별 오류 데이타 개수

dd1	dd2	dd3	dd4	dd5	dd6	dd7
0	17,349	50	0	0	0	0
dd8	dd9	dd10	dd11	dd12	dd13	dd14
0	0	0	19	0	0	0
dd15	dd16	dd17	dd18	dd19	dd20	dd21
0	0	0	0	0	0	280
dd22	dd23	dd24	dd25	dd26	dd27	dd28
0	1505	0	0	0	0	0
dd29	dd30	dd31	dd32	dd33		
10,073	1,551	0	343	0		

류별로 막대 그래프로 나타낸 것이다. "Missing data where null-not-allowed constraint should be enforced"(dd2)오류와 "Uses of special characters of concatenated data"(dd29)오류가 차지하는 비율은 전체 검색된 오류 데이타 중 74%나 차지하였다. 그 이유는 호텔정보를 입력하는 웹사이트의 입력항목에 널값 제약을 걸지 않았거나, 전화번호를 입력하는 데이타베이스의 디자인이 연속 데이타를 처리하는 것을 고려하여 디자인되지 않았기 때문이며, 이로 인해 오류 데이타가 많이 발생하였다는 것을 알 수 있다. 따라서 호텔 데이타의 데이타 품질을 높이기 위해서는 각 입력항목에 제약사항을 걸어 널값의 입력을 방지하거나, 연속데이타의 처리를 고려하여 전화번호부 데이타베이스를 디자인하는 것이 요구된다.

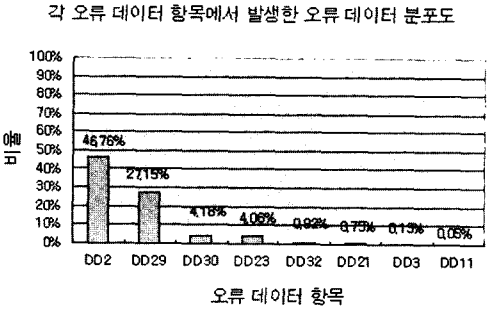

그림 13 오류 종류별 오류 데이타 검색 개수 분포도

5.2 데이타 품질 측정값

데이타 품질은 3.2절에서 기술하였듯이 데이타 사용목적에 따라 달라질 수 있으므로 DAQUM도구는 사용목적에 따라 사용자가 직접 데이타 사용목적과 그에 따른 중요항목을 선정하여 데이타 품질을 측정하도록 한다. 호텔 데이타베이스 사례연구에서는 표 2에서처럼 사용목적을 고려하지 않은 경우(Use1)와 9가지의 다양한 사용목적에 따른(Use2Use10) 데이타 품질 측정을 하였다. 일반적으로 가중치는 예를 들어, 0(하)-1(중)-2(상) 등 이러한 간격으로 결정하는 방법도 있지만, 가중치 부여 기준에 대한 메트릭스의 부재로 인해 분류가 어려운 실정이다. 본 논문에서는 데이타 사용자가 직접 사용목적에 따라 컬럼의 중요도 순으로 가중치(Weight)값을 달리 부여할 수 있도록 하였다. DAQUM도구 사용자는 사용목적에 따라 관련이 있는 컬럼에는 High값을(H)주고 관련 없는 컬럼에는 Low(L)값을 입력할 수 있다. 예를 들어 고객에게 홍보용 자료를 우편으로 보내는 목적으로 호텔 데이타베이스가 사용되는 Use2의 경우, 주소(Address) 컬럼 및 주(State) 컬럼을 참고하여 우편물을 보내므로, 이 주소 컬럼과 주 컬럼이 데이타 품질에 중요한 영향을 미친다.

호텔 데이타베이스에서 측정한 데이타 품질 측정값은 3.2절의 품질 측정 메트릭스에 따라 계산되며 표 2의 아래 줄에 그 측정값을 나타내었으며, 그림 14에 그래프로 표현하였다.

표 2의 데이타 품질 측정값은 사용자의 데이타 사용목적에 따라 이와 관련이 많은 컬럼들과 관련 없는 컬럼들을 구분하여 품질을 측정함으로써 중요한 컬럼의 데이타 중요도가 반영된 의미 있는 값이다. 품질 측정값은 데이타 품질 측정 메트릭스에 따라 다음과 같이 계산되었다. 예를 들어 고객에게 홍보용 자료를 우편으로 보내는 목적으로 데이타가 사용되며(Use2의 경우) 품질 측정 대상인 데이타웨어하우스는17,350개의 record와 번호, 주, 이름, 주소, 장소, 전화번호, 팩스, 이메일, 종류, 신용카드, 현금, 공항, 세율, 가격1, 가격2, 가격3 등등 16개의 컬럼을 가지고 있으며(즉 총 데이타 개수 N=277,600개) 이들 각 컬럼에서는 0개, 280개, 5,505개, 1,505개, 772개, 8,419개, 5,584개, 15,988개, 475개, 986개, 16,366개, 2,955개, 8,210개, 6,017개, 11,107개, 13,924개의 오류 데이타가 발견되었다. 이 경우 데이타 사용 목적을 고려하여 중요한 항목인 '주'컬럼과 '주소' 컬럼의 데이타를 가지고 R-Metric을 계산하고(R=1-(1/2(280/17,350+1,505/17,350))=0.9485) 나머지 컬럼들의 데이타를 가지고 NR-Metric을 계산한다(NR=1-(1/14(0/17,350+5,505/17,350+772/17,350+8,419/17,350+5,584/17,350+15,988/17,350+475/17,350+986/17,350+16,366/17,350+2,955/17,350+8,210/17,350+6,017/17,350+11,107/17,350+13,924/17,350))=0.6035). 이 두개의 R-Metric과 NR-Metric을 고려하여 계산된 전체 데이타 품질 측정값 Q는 57.25%(Q=0.9485*0.6035) 된다.

사용목적을 고려하지 않는 경우는 데이타 품질 측정값이 64.66%이다. 이메일로 홍보활동을 하는 목적으로

표 2 호텔 데이타베이스의 사용목적에 따른 관련 컬럼 및 데이타 품질 측정치

Colunm Name	데이타 사용 목적에 따라 컬럼을 그룹화									
	Use1: Not select Purpose	Use2: Campaign via post	Use3: Campagin via tel	Use4: Campaign via email	Use5: Campaign via tex	Use6: Hotel classification	Use7: The number of hoel per state	Use8: The number of hotel near airports	Use9: The comparision of room rates	Use10: Various kinds of usable creditcards in hotel
Hotelnumm	L	L	L	L	L	L	L	L	L	L
State	L	H	L	L	L	L	H	L	L	L
Hotelname	L	L	L	L	L	L	L	L	L	L
Address	L	H	L	L	L	L	H	L	L	L
Place	L	L	L	L	L	L	L	L	L	L
Phone	L	L	H	L	L	L	L	L	L	L
Fax	L	L	L	L	H	L	L	L	L	L
Email	L	L	L	H	L	L	L	L	L	L
Hoeltype	L	L	L	L	L	H	L	L	L	L
creditcard	L	L	L	L	L	L	L	L	L	H
currency	L	L	L	L	L	L	L	L	L	L
Airports	L	L	L	L	L	L	L	H	L	L
Ratings	L	L	L	L	L	L	L	L	L	L
Price1	L	L	L	L	L	L	L	L	H	L
Price2	L	L	L	L	L	L	L	L	H	L
Price3	L	L	L	L	L	L	L	L	H	L
데이타품질	64.66	57.25	33.74	5.37	43.71	60.78	57.25	52.64	28.36	59.12

데이타가 사용되는 Use4의 경우는 5.37%로 가장 낮은 품질 측정값을 가졌다.

Use4 경우가 가장 품질이낮게 측정된 이유는 데이타 사용 목적에 관련이 많은 이메일컬럼의 데이타에서 오류 데이타가 많이 발생했기 때문이다. 이는 호텔 데이타 관리자에게 이메일 주소 입력란에 제약 사항을 두어 데이타베이스가 디자인 되기가 요구된다는 것을 보여준다.

홍보용 자료를 보내는 목적으로 사용되는 Use2, Use3, Use4, Use5경우에서 데이타 품질값이 각기 다르게나타났다. 우편으로 보내는 Use2의 경우는 데이타 품질 측정값이 57.25%, 전화로 통화할 경우인 Use3의 경우는 데이타 품질 측정값이 33.74%, 이메일로 보내는 Use4의 경우는 데이타 품질 측정값이 5.37%, 팩스로 보내는 Use5의 경우는 데이타 품질 측정값이 43.71%로 나타났다. 이는 호텔 데이타 관리자가 홍보활동을 할 경우 이메일보다는 우편, 팩스, 전화를 통한 홍보 활동이 더 큰 효과를 준다는 것을 알 수 있다.

이 사례연구를 통하여 오류 데이타 수가 많을수록 데이타 품질은 낮아지는데, 오류 데이타 수가 같더라도 데이타 사용 목적에 따라 데이타 품질이 다르게 측정하는 것이 중요함을 파악 할 수 있다. 오류 데이타 종류별로 검색한 오류 데이타 개수와 측정된 데이타 품질 측정값을 제공함으로써 측정 대상 데이타베이스의 문제점을 분석하고 품질 개선을 위한 요구사항을 사용자에게 제시할 수 있다.

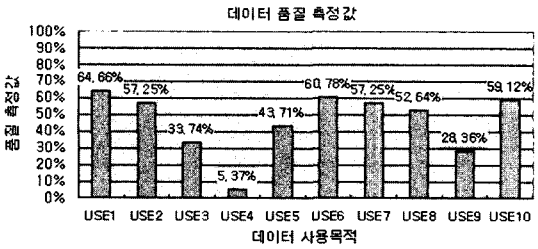

그림 14 데이타 사용 목적에 따른 데이타 품질

6. 결 론

대용량의 데이타로부터 부가적인 가치가 있는 데이타로 가공하여 의미 있는 지식을 추출하는 지식공학 시스템에서 원시 데이타의 품질을 보장하는 일은 매우 중요하다. 본 논문에서는 데이타의 품질을 측정할 수 있도록 품질 측정 도구인 DAQUM도구를 설계 개발하였다.

본 연구에 앞서 개발한 오류 데이타 분류에서 33개의 오류 데이타 종류를 검색할 수 있도록 DAQUM도구를

개발하였으며, 33개의 오류 데이타 종류별로 검색한 오류 데이타의 개수에 따라 데이타 품질 측정 메트릭스를 구축하였다. DAQUM도구 사용자가 데이타의 사용목적에 따라 데이타베이스의 중요 컬럼을 지정할 수 있도록 하여 데이타의 사용목적에 따라 데이타 품질이 다르게 측정될 수 있도록 하였다.

16개 컬럼 및 17,350개의 레코드인 총277,000개의 데이타 개수를 갖는 호텔정보에 대한 데이타베이스를 대상으로 DAQUM도구의 사례연구를 수행하였다. 호텔 데이타베이스 사례연구로부터 1) 오류 데이타 종류별로 검색한 오류 데이타 개수 분포와 2) 다양한 데이타 사용 목적에 따른 품질 측정 값 비교를 분석하였다. DAQUM도구로 검색한 전체 오류 데이타 수는 37,100개였으며 이중에서 "Missing data where null-not-allowed constraint should be enforced"(dd2)오류는 17,349개였다.

DAQUM도구는 사용목적에 따라 사용자가 직접 데이타 사용목적과 그에 따른 중요항목을 선정하여 데이타 품질을 측정하도록 하였다. 호텔 데이타베이스 사례연구에서는 사용목적을 고려하지 않은 경우(Use1)와 9가지의 다양한 사용목적에 따른(Use2Use10) 데이타 품질 측정치를 비교하였다. 사용목적에 따라 품질 측정값이 5.37%에서 64.66%으로 다양하게 나타났다. 이로부터 오류 데이타 수가 같더라도 데이타 사용 목적에 따라 데이타 품질이 다르게 측정하는 것이 중요함을 파악할 수 있었다.

기존 연구에서는 정의 단계에서 사용자 관점을 고려한 데이타 품질 평가의 중요성만을 언급만 했던 것에 비해 DAQUM도구는 데이타 사용자 관점에서 측정을 가능하도록 하였다는데 의의가 있다. DAQUM도구는 오류 데이타 종류별로 검색한 오류 데이타 개수와 측정된 데이타 품질 측정값을 제공함으로써 지식공학 시스템에서 도출된 지식의 유용성을 판단하는데 도움을 줄 수 있다. DAQUM도구는 데이타의 품질 제어를 가능하게 하며 데이타를 주로 처리하는 소프트웨어 제품의 품질 향상에 기여할 수 있다.

현재 DAQUM도구는 샤모아 지식공학 시스템[3,9]의 초기 단계인 데이타웨어우스에 대한 ETL과정에서 오류 데이타 검색 및 데이타 품질 측정 컴포넌트로써 개발되었다. 향후 DAQUM도구를 보완하여 실제로 데이타 품질 측정이 필요로 하는 지식공학 시스템의 다양한 단계에서 사용할 수 있도록 보안할 예정이다. 현재 왜곡된 사용 목적을 감지해 주는 장치는 현재 존재하지 않으므로, 향후 데이타 사용 목적이 왜곡되어 잘못된 경우에 미리 데이타 마이닝 알고리즘에 의해 미리 지시를 해주는 장치가 필요하며 이러한 기능을 DAQUM에서 제공할 계획이다. 현재 DAQUM도구는 오류데이타를 검색하여 사용자 관점에서 품질측정이 가능함을 일반적으로 제시한 것에 의의가 있으나, 이를 기반으로 바이오인포메틱스, 웹환경에서 추출한 데이타 등 대상데이타의 특성 및 환경에 따라 특화된 오류데이타 검색과 사용자관점의 데이타 품질 측정이 가능하도록 DAQUM도구를 확장할 계획이다.

참 고 문 헌

[1] ISO/IEC 14598-1,2,3,4,5,6, JTC 1 SC 7 Documents, 1999.

[2] Won Kim et "A Component-Based Knowledge Engineering Architecture," JOOP, vol.12, no.6, pp 40-48, 1999.

[3] Won Kim et al. "The Chamois component-based knowledge engineering framework," IEEE Computer, May 2002.

[4] Won Kim et al. "The Chamois Re-configurable Data-Mining Architecture," Journal of Object Technology, pp21-34 , June 2002.

[5] D. Ballou and G.K. Tayi "Enhancing Data Quality in Data Warehouse Environments," Communications of the ACM, vol. 42, no. 1, pp. 73-78, Jan. 1999.

[6] Amir Parssian, Sumit Sarkar, Varghese S. Jacob, "Assessing data quality for information products," Proceeding of the 20th international conference on Information Systems, p.428-433, January, 1999.

[7] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, Doheon Lee, "A Taxonomy of Dirty Data," Data Mining and Knowledge Discovery, 2002, Acceptedfor publication.

[8] Richard Y. Wang "A Product Perspective on Total Data Quality Management," Communication of the ACM, vol. 41, no. 2, pp. 58-65, Feb. 1998.

[9] 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.

[10] R. Wang, V. Storey and C. Firth "A Framework for Analysis of Data Quality Research," IEEE Transactions on Knowledge and Engineering, vol. 7, no. 4, pp. 623-640, Aug. 1995.

[11] Wang et al. "Data Quality in context," Communication of the ACM, vol. 40, no 5, May 1997.

[12] Ken Orr, "Data Quality and System Theory," Communications of the ACM, vol.41, no.2 Feb. 1998.

양 자 영

1995년~1999년 조선대학교 전산통계학과 학사. 2000년~2002년 이화여대 과학기술대학원 컴퓨터학과 석사. 현재 유리자산운용 전산지원팀. 관심분야는 소프트웨어 및 데이타 품질 측정, 데이타 품질 관리

최 병 주

정보과학회논문지 : 컴퓨팅의 실제
제 9 권 제 2 호 참조