제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

샤모아 컴포넌트 시스템에서의 DAQUM 컴포넌트 설계 및 구현

김은희, 최병주
이화여자대학교 컴퓨터학과
e-mail : {ehkim, bjchoi}@ewha.ac.kr

DAQUM Component Design and Implementation in Chamois Component System

EunHee Kim, ByoungJu Choi
Dept. of Computer Science & Engineering, Ewha Womens University

요 약

샤모아 프레임워크(Chamois Framework)는 독특한 컴포넌트를 지니고 있는 새로운 지식공학 프레임 워크이다. 이러한 대용량의 데이터 소스로부터 의미 있는 지식을 추출하는 지식공학 시스템에서 소스 데이터의 품질을 보장하는 일은 매우 중요하다. 본 논문에서는 데이터의 품질 측정 도구인 DAQUM(DAta QUality Measurement) 컴포넌트의 설계 및 구현에 관한 주요 내용을 기술하고, 컴포넌트 기반의 구조를 가지는 샤모아 프레임워크에서 DAQUM 의 역할 및 동작에 대해 기술한다.

1. 서론

Client-Server, internet 등의 여러 IT 기술들이 기업, 정부기관이나 교육기관 등에서 유용하게 사용되고 있다. 이들 IT 인프라에 추가되는 새로운 분야로써, 최근 지식공학이 등장하였다. 지식공학은 고부가가치 창출을 위해 데이터를 체계적으로 수집, 저장,관리, 분석하여 지식을 추출하는 기술이다[1].

이러한 지식공학 시스템에서는 대용량의 데이터 소스로부터 의미 있는 지식을 추출하므로, 소스 데이터의 품질을 보장하는 일이 매우 중요하다. 만약, 이러한 지식공학 시스템에서 데이터 품질을 제어하는 기술이 제공되지 못한다면, 사용자에게 제공하는 데이터나 지식을 신뢰할 수 없으므로, 지식공학 시스템 자체의 존재가 무의미하게 될 것이다[2,3].

따라서, 지식공학 시스템인 샤모아[4] 컴포넌트 시스템에서도 데이터의 품질의 보증하는 것은 필수적이다.

본 논문에서는 지식공학 시스템에서 데이터 품질을 측정하는 도구인 DAQUM 컴포넌트의 설계 및 구현에 대해 설명한다. DAQUM 은 지식공학 시스템인 샤모아 컴포넌트 시스템에서, 처리되는 데이터들의 품질을 측정하고, 데이터의 품질을 보장할 수 있도록 한다. 이를 통해, 샤모아 프레임워크 내에서 동작하는 다른 컴포넌트들이 사용하는 데이터의 신뢰성을 높여주며, 궁극적으로 샤모아 컴포넌트 시스템의 품질 향상에 공헌 할 수 있게 한다.

본 논문은 2 장에서 관련연구를 기술한다. 3 장에서는 DAQUM 컴포넌트의 설계 및 구현, 4 장에서는 구현된 DAQUM 컴포넌트의 샤모아 컴포넌트 시스템에서의 역할에 대해 기술하다. 5 장에서는 본 논문의 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

2.1 샤모아 프레임워크

샤모아[4]는 이화여자대학교 과학기술대학원에서 수행중인 IKEA(Integrated Knowledge Engineering Architecture) 프로젝트로써, 컴포넌트 기반 지식공학 아키텍처를 구축하고자 하는 프로젝트이다. 샤모아 프레임워크는 본 대학원에서 이루어지는 여러 연구를 하나로 묶어 서로의 연구가 상호 시너지 효과를 낼 수 있도록 하며, 이를 통해 기존의 상업적인 지식공학 프레임워크보다 더 큰 규모의 독특한 컴포넌트를 지

니는 지식공학 프레임워크를 구축한다. 그림 1 은 전제적인 샤모아 프레임워크를 보여준다.

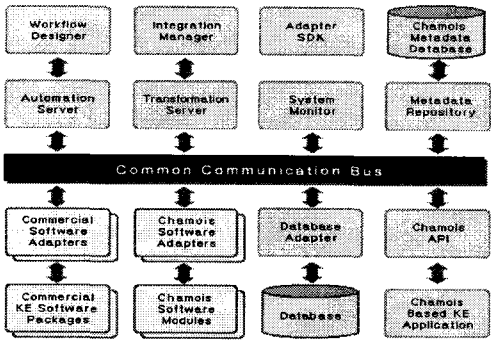

그림 1 샤모아 프레임워크(Chamois Framework)

샤모아에서 개발한 컴포넌트 기술들은 다음과 같은 내용을 포함한다.

- Data warehousing and metadata management
- OLAP and data/text/Web mining
- XML server, querying, mining
- Networking (QoS for multimedia delivery)
- Security (encryption, group communications, secure networks)
- E-commerce architecture, e-payment, personalization, m-commerce, voice-based access to the Internet
- Component-based software design, development, and testing
-

개발된 모듈들은 COM, Web Service 와 같은 컴포넌트 아키텍처 기술들을 사용한다.

2.2 오류데이터

논문[2]에서 오류데이터를 "successive hierarchical refinement" 방식으로 분류하였다. 오류데이터가 발생하는 이유를 "Missing data", "Not missing, but wrong data", "Not wrong, but unusable data" 세 가지로 나누고, 이를 계층적으로 분해하여, 그림 2 와 같이 총 33 개의 오류데이터로 분류하였다. 그림 2 의 오류데이터 분류 구조(taxonomy of dirty data)의 단말노드가 실제적인 오류데이터의 이름이다.

논문[2]에서 다른 관점으로부터 시작한 오류 데이터의 분리는 다른 구조를 가질 수 있으나, 각 분류의 오류데이터 형태를 구체적으로 나타내는 단말 노드는 동일함을 다른 분류 구조의 전개를 통해 증명하였다. 이들 오류데이터는 실제 가능한 오류데이터의 종류에 대해 95%이상의 정확도를 가진다[6].

논문[7]에서는 그림 2 의 오류데이터 종류별로 오류데이터를 파악하여 데이터 품질을 측정하였다.

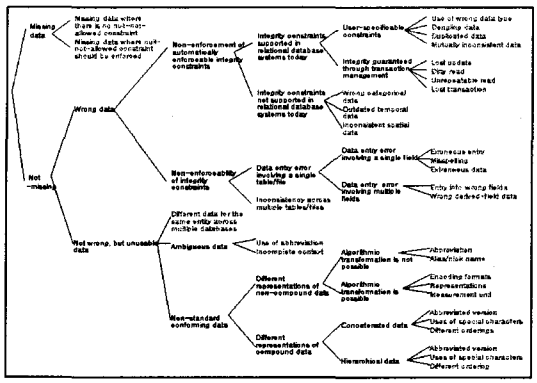

그림 2 오류데이터 분류 구조(Taxonomy of Dirty Data)

2.3 데이터 품질 측정

데이터 품질에 관한 연구는 소프트웨어 품질 연구와 달리 아직 표준이 정립되지 않았다. 데이터 품질에 대한 필요성에 따라 진행된 연구 중 대표적인 것으로 Wang 의 연구[8]가 있다. Wang 은 품질 관리 방법론인 TQM(Total Quality Management)을 데이터에 적용한 TDQM(Total Data Quality Management)을 제시하였다. Wang 의 TDQM 생명주기는 크게 정의, 측정, 분석, 향상의 단계로 구성된다.

TDQM 에 따른 정의단계를 수행하기 위해서 사용자가 데이터 품질에 대해 요구하는, 데이터 특성에 관한 연구가 요구되며, 이에 관한 대표적인 연구는 Ballou 의 연구[9]와 Wang 의 연구[8]가 있다. 이들의 연구 결과 분석을 통해, 데이터 품질의 대표적 특성을 4 가지: 정확성, 적시성, 완료성, 일관성으로 구분할 수 있다.

논문[7]에서는 TDQM 의 생명 주기에서 정의 및 측정 단계까지 실제 데이터를 사용하는 사용자가 어떤 목적으로 데이터를 사용했는가를 고려하여 동일한 데이터라 할지라도 사용자의 관점에서 품질이 다르게 평가되도록 하는 데이터 품질 평가 도구를 구현하였다.

3. DAQUM(DAta QUality Measurement) 컴포넌트

3.1 DAQUM 컴포넌트의 구조

DAQUM 컴포넌트의 구조는 그림 3 과 같으며, "interface", "Data Management", "Data Quality Measure", "Analyzer" 4 개의 패키지로 구성된다.

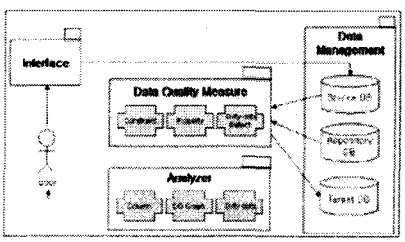

그림 3 DAQUM 의 구조

(1) 인터페이스(Interface) 패키지
데이터베이스관리, 데이터 제약사항입력, 데이터 사용목적입력 등의 데이터 품질 측정과 관련된 사용자와 DAQUM 간의 인터페이스를 제공한다.

(2) 데이터 품질 측정(Data Quality Measurement) 패키지
데이터 제약사항, 데이터 사용목적, 오류데이터검색을 통해 데이터의 품질을 측정한다.

(3) 품질 측정 결과 분석(Analyzer) 패키지
측정된 데이터의 품질을 오류데이터 분류별, 해당 컬럼 별로 보여주고, 사용목적의 관련 여부에 따라 두개의 그룹으로 측정한 데이터의 품질 측정결과와 전체 데이터 품질 측적 결과를 분석하여 보여준다.

(4) 데이터 관리(Data Management) 패키지
3 개의 데이터베이스를 가지고, 이들을 통해 데이터 사전, 데이터 프로파일 등을 관리한다. "Source DB"는 데이터 품질 측정 대상 데이터와 오류검색에 필요한 데이터 프로파일을 저장한다. "Repository DB"는 검색된 오류데이터를 저장하며, "Target DB"는 데이터 품질 측정 메트릭스[7]에 따른 오류데이터 품질 측정값을 저장한다.

3.2 DAQUM 컴포넌트의 기능

데이터 품질 측정 도구인 DAQUM 컴포넌트는 샤모아 컴포넌트 시스템의 한 컴포넌트로써, 동작할 수 있도록 하기 위해, Microsoft VisualBasic6.0 을 이용하여 COM(Component Object Model)으로 구현하였다.

(1) 측정 대상 데이터베이스 및 테이블 선택기능
샤모아 프레임워크 상의 DBMS 에 접속하고, 접속한 DBMS 의 데이터베이스와 테이블 중, 측정하고자 하는 대상 데이터베이스와 테이블을 선택한다.

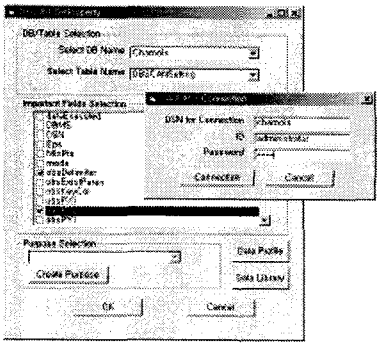

그림 4 품질측정 대상 데이터베이스 및 테이블선택

(2) 데이터 라이브러리 생성 및 관리기능
DAQUM 컴포넌트가 오류 데이터를 검출하도록 하기 위해, 사용자가 선택된 컬럼의 카테고리, 약어, 축약어[7]에 대한 테이블을 생성하고, 이를 관리할 수 있도록 한다.

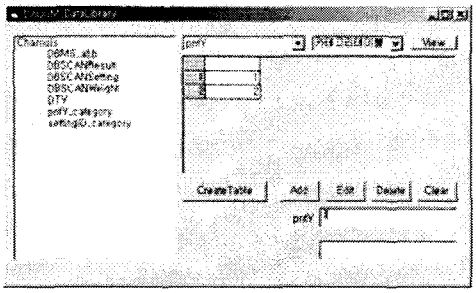

그림 5 데이터 라이브러리

(3) 데이터 프로파일 관리기능
데이터 프로파일은 품질 측정 대상이 되는 테이블의 컬럼에 대한 제약 사항이다. 사용자의 편의를 높이기 위해 사용자의 T/F 체크 만으로 query 를 자동 생성한다. 이렇게 정의된 데이터 프로파일은 각 컬럼 별로 테이블을 생성하여 관리되며, 이는 오류데이터 검출에 사용된다.

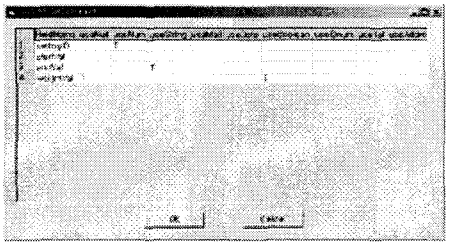

그림 6 데이터 프로파일

(4) 사용목적 생성 마법사
DAQUM 컴포넌트는 사용목적에 따라 데이터 품질을 측정할 수 있다[7]. 사용목적은 여러 형태의 데이터를 취급함에 있어, 다양한 목적이 요구되어진다. 이를 위해 DAQUM 컴포넌트는 사용목적을 사용자가 직접 생성할 수 있도록 한다. 이때, 사용목적 생성을 마법사 형태로 구현하여 사용자의 편의를 높일 수 있다.

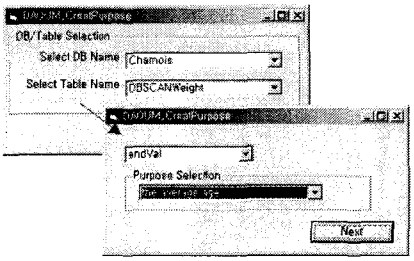

그림 7 사용목적 생성 마법사

(5) 오류데이터 결과 출력
오류데이터의 결과는 그림 8 과 같이 오류데이터

별(①), 컬럼 별(②), 사용목적에 따른 결과(③) 3 가지의 결과 화면을 보여준다.

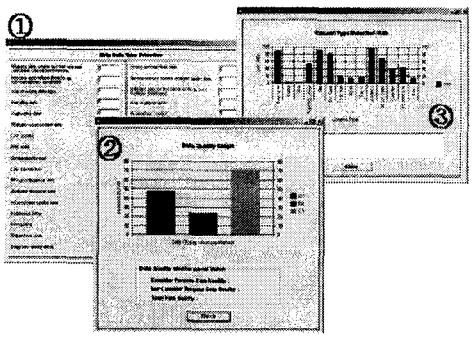

그림 8 오류데이터 결과 출력

4. 샤모아 컴포넌트 시스템에서의 DAQUM

DAQUM 컴포넌트 개발의 목적은 지식공학에서 사용되는 데이터의 품질을 측정하고, 이를 통해 데이터의 신뢰성을 높이고, 나아가서는 이 데이터를 사용하는 지식공학 시스템의 품질 향상을 이끌어내고자 하는 것이다.

컴포넌트 기반 지식공학 시스템인 샤모아 컴포넌트 시스템에서도 다양한 지식 및 데이터가 사용되고 있으며, 이에 대한 품질을 DAQUM 컴포넌트를 통해 측정한다. 이를 통해 샤모아 컴포넌트 시스템 내의 컴포넌트들이 사용하는 데이터의 품질을 보증할 수 있고, 따라서, 컴포넌트들의 품질도 향상될 수 있다. 이는 결국, 샤모아 컴포넌트 시스템 자체의 품질 향상을 가져올 수 있는 것이다.

그림 4 는 샤모아 프레임워크의 전체적인 구조에서 DAQUM 컴포넌트의 위치를 보여주고 있다. DAQUM 컴포넌트는 샤모아 프레임워크의 MS-SQL, COM+ 환경에서 동작하고 있으며, 샤모아 프레임워크의 여러 컴포넌트들이 데이터를 사용하기에 앞서, DAQUM 컴포넌트를 통해 품질이 측정되고, 보증 된 지식 및 데이터를 사용할 수 있도록 한다.

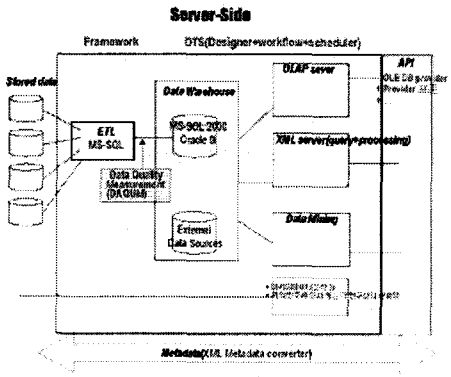

그림 9 샤모아 프레임워크에서의 DAQUM

5. 결론 및 향후 연구 과제

대용량의 데이터 소스로부터 부가적인 가치가 있는 데이터로 가공하여 의미 있는 지식을 추출하는 지식공학 시스템에서의 시스템의 품질을 보장하기 위해서는 먼저 데이터의 품질이 보증 되어야 한다. 이를 위해 실제적인 컴포넌트 기반 지식공학 프레임 워크인 샤모아 컴포넌트 시스템에서 사용되는 데이터 및 지식들의 품질을 측정하는 도구인 DAQUM 컴포넌트를 구현하였다. 본 논문에서는 DAQUM 의 설계 및 구현에 대해 설명하고, 샤모아 컴포넌트 시스템에서 DAQUM 의 동작 및 역할에 대해 기술하였다. 샤모아 컴포넌트 시스템에서 DAQUM 컴포넌트를 통해 데이터의 품질을 측정하여, 데이터의 신뢰성을 높여줌으로써, 샤모아 컴포넌트 시스템 자체의 품질 향상에도 기여할 수 있다.

향후, 본 논문에서 제안한 DAQUM 컴포넌트가 기타 다양한 컴포넌트 기반의 지식공학 시스템에서 동작할 수 있도록 기능을 보완, 확장할 계획이다.

참고문헌

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

[2] 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

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

[4] Won Kim, Ki-Joon Chae, Dong-Sub Cho, Byoungju Choi, Anmo Jeong, Myung Kim, Ki-Ho Lee, Meejeong Lee, Sang-Ho Lee, Seung-Soo Park, Hwan-Seung Young, "The Chamois Component-Based Knowledge Engineering Framework", IEEE Computer Journal, May 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] Won Kim, Byoung-Ju Choi, Eui-Kyeoung Hong, Soo-Kyoung Kim, Doheon Lee, "A Taxonomy of Dirty Data", Data Mining and Knowledge Discovery, 2002

[7] 양자영, 최병주, "소프트웨어 사용자 관점의 데이터 품질 측정 방안", 제 28 회 정보과학회 추계학술대회, pp436-438, Oct 19-20, 2001

[8] Richard Y.Wang, "A Product Perspective on Total Data Quality Management", Communication of the ACM, vol.41, on.2, pp.56-65, 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