컴포넌트 기반 샤모아 데이터 정제 도구 개발

김은희° 최병주
이화여자대학교 컴퓨터학과
{ehkim, bjchoi}@ewha.ac.kr

Development of a Component-Based Chamois Data Cleansing Tool Suits

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

요 약

샤모아 지식공학 시스템(Chamois Knowledge Engineering System)은 대용량의 데이터 소스로부터 의미 있는 지식을 추출하는 시스템이다. 이러한 지식공학 시스템에서 데이터 소스의 품질을 보장하는 일은 매우 중요하다. 본 논문에서는 샤모아 지식공학 시스템에서의 데이터 정제관련 컴포넌트의 구조 및 동작에 대해 기술한다. 또한 이들 컴포넌트들이 동작할 수 있는 컴포넌트 프레임웍의 기능 및 동작에 대해 기술한다. 구현한 데이터 정제 관련 컴포넌트는 컴포넌트 기반의 시스템에서 데이터의 정제를 통해 신뢰성 있는 데이터를 제공하고, 이를 통해 개발하고자 하는 시스템의 품질을 향상 시킬 수 있다.

1. 서 론

Client-Server, internet 등의 기존 IT 인프라에 새로이 추가되는 분야로써, 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 데이터의 정제가 이루어지지 못한다면, 사용자에게 제공하는 데이터나 지식을 신뢰할 수 없으므로, 지식공학 시스템 자체의 존재가 무의미하게 될 것이다[2,3].

따라서, 지식공학 시스템인 샤모아[4] 지식공학 시스템에서 데이터의 정제는 필수적인 것이다.

본 논문에서는 샤모아 지식공학 시스템에서 사용되는 데이터 정제관련 컴포넌트의 구현에 대해 기술한다. 또한 이들 컴포넌트들이 동작할 수 있는 프레임웍을 개발함으로써, 구현된 컴포넌트간의 동작이 제대로 이루어져 있는지를 테스트할 수 있도록 한다. 이렇게 개발된 데이터 정제관련 컴포넌트들은 샤모아 지식공학 시스템 내에서 동작하는 다른 컴포넌트들이 사용하는 데이터의 신뢰성을 높여주고, 궁극적으로 샤모아 지식공학 시스템의 품질 향상에 공헌할 수 있도록 한다.

본 논문은 2장에서 관련연구를 기술한다. 3장에서는 데이터 정제관련 컴포넌트들의 구현 및 동작에 대해 설명하고, 4장에서는 구현된 데이터 정제관련 컴포넌트들이 동작하는 컴포넌트 프레임웍의 구조 및 동작에 대해 기술한다. 5장에서는 본 논문의 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

2.1 샤모아 지식공학 시스템

샤모아[4]는 이화여자대학교 과학기술대학원에서 수행중인 IKEA(Integrated Knowledge Engineering Architecture) 프로젝트로써, 컴포넌트 기반 지식공학 아키텍처를 구축하고자 하는 프로젝트이다. 샤모아 지식공학 시스템은 본 대학원에서 이루어지는 여러 연구를 하나로 묶어 서로의 연구가 상호 시너지 효과를 낼 수 있도록 하며, 이를 통해 기존의 상업적인 지식공학 프레임웍보다 더 큰 규모의 독특한 컴포넌트를 지니는 지식공학 프레임웍을 구축한다. 그림 1은 샤모아 지식공학 시스템의 구성을 보여준다.

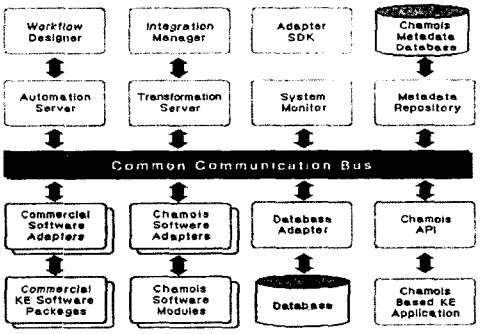

그림 1 샤모아 지식공학 시스템

2.2 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 Wang의 연구[5]가 있다. Wang은 품질 관리 방법론인 TQM(Total Quality Management)을 데이터에 적용한 TDQM(Total Data Quality Management)을 제시하였다. 이에 관한 대표적인 연구는 Ballou의 연구[6]와 Wang의 연구[5]가 있다. 이들의 연구 결과 분석을 통해, 데이터 품질의 대표적 특성을 4가지: 정확성, 적시성, 완료성, 일관성으로 구분할 수 있다.

논문[7]에서는 TDQM의 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. 데이터 정제 관련 컴포넌트

샤모아 지식공학 시스템의 데이터 정제를 위해 구현된 컴포넌트들은 다음과 같이 구성된다. 샤모아 지식공학 시스템은 COM(Component Object Model) 컴포넌트 아키텍쳐로 구축되어 있다. 따라서 이들 컴포넌트들은 Microsoft VisualBasic6.0을 이용한 COM으로 구현하였으며, MSSQL 2000을 DBMS로 사용한다.

3.1 DSNConnection

DSNConnection 컴포넌트는 DSN 연결 정보를 생성하는 컴포넌트이다. DSNConnection 컴포넌트는 사용자가 입력한 정보를 이용하여 DSN 연결 값을 생성하고, 연결 값을 요청하는 다른 컴포넌트에게 이를 제공하는 역할을 수행하는 컴포넌트이다.

3.2 DAQUM

DAQUM 컴포넌트는 샤모아 지식공학 시스템에서 사용되는 데이터들의 품질을 측정하는 컴포넌트이다[7]. DAQUM 컴포넌트의 개발 목적은 지식공학에서 사용되는 데이터 품질을 측정하고, 이를 통해 데이터의 신뢰성을 높이고, 나아가서는 이 데이터를 이용하는 지식공학 시스템의 품질 향상을 이끌어 내고자 하는 것이다.

DAQUM 컴포넌트는 크게 5가지의 기능을 수행한다.

(1) 측정 대상 데이터베이스 및 테이블 선정 기능
DSNConnection에서 연결정보를 받아 DBMS에 연결하고, 접속한 DBMS의 데이터베이스와 테이블 중에서 측정하고자 하는 대상의 데이터를 선택한다.
(2) 데이터 라이브러리 생성 및 관리기능
DAQUM 컴포넌트의 오류 데이터 검출을 위해, 사용자가 선정한 대상의 카테고리, 약어, 축약어[7]에 대한 테이블을 생성, 관리할 수 있도록 한다.
(3) 데이터 프로파일 관리기능
데이터 프로파일은 품질 측정 대상의 컬럼에 대한 제약 사항이다. 정의된 프로파일은 테이블로 생성하여 관리하며, 오류 데이터 검출에 사용된다.
(4) 사용목적 생성 및 관리기능
DAQUM 컴포넌트는 사용 목적에 따라 데이터 품질을 측정할 수 있다[7]. DAQUM 컴포넌트는 다양한 지식공학 데이터의 사용에 적용할 수 있도록 하기 위해, 이를 사용목적을 사용자가 생성할 수 있도록 한다.
(5) 오류데이터 측정 결과 출력 기능
오류데이터[8] 측정 결과는 오류데이터 별, 컬럼 별, 사용목적에 따른 결과의 3가지 형태로 출력된다. 또한 이들 측정결과는 날짜 및 시간으로 분류하여 저장 관리 함으로써, 오류데이터 측정 빈도 및 각 측정별 오류데이터 수를 사용자에게 제공한다.

3.3 DDCleaning

DDCleaning 컴포넌트는 DAQUM 컴포넌트에서 측정된 데이터들을 수정 및 삭제 하는 컴포넌트이다. 오류가 있음으로 처리된 데이터들을 사용자에게 제공하고, 이들 데이터를 개별적으로 수정, 삭제하여 데이터베이스에 저장한다. 뿐만 아니라, DDCleaning 컴포넌트는 DAQUM 컴포넌트가 처리하지 않은 데이터를 불러와 수정 및 삭제할 수 있도록 하는 기능을 제공한다.

3.4 OLAPPBrowser

OLAPPBrowser 컴포넌트는 DAQUM과 DDCleaning 컴포넌트를 통해 정제된 데이터를 사용자가 요구하는 다차원 및 구조적인 형태의 데이터 형태로 출력해주는 컴포넌트이다. 이 OLAPPBrowser 컴포넌트는 MS Analysis Services를 이용하여 구현하였다.

4. 데이터 정제 관련 컴포넌트를 위한 프레임웍

3절에서 설명한 데이터 정제 관련 컴포넌트들은 컴포넌트의 재사용의 측면을 고려한 크기로 세분하여 4개의 컴포넌트 DSNConnection, DAQUM, DDCleaning, OLAPPBrowser로 구현하였다. 이들 컴포넌트들은 컴포넌트 기반의 시스템에서 조립 혹은 개별적으로 지식공학 데이터의 정제라는 역할을 수행한다.

본 논문에서는 데이터 정제 관련 컴포넌트 뿐 아니라, 이들 컴포넌트들이 동작할 수 있는 컴포넌트 개발 개념에 따른 프레임웍을 구현하였다. 이들 데이터 정제관련 컴포넌트들은 프레임웍에 자유롭게 등록 및 삭제될 수 있다. 즉, 컴포넌트들은 제공된 API를 통해 프레임웍에 등록된 다른 컴포넌트들과 상호적으로 동작하거나 혹은 프레임웍 상에서 개별적으로 동작한다.

그림 2는 프레임웍에 컴포넌트 데이터 정제관련 컴포넌트들이 등록되는 것을 보여준다. 프레임웍에 컴포넌트를 등록하기 위해 사용자가 .dll 형태의 파일을 찾고(①②), 해당 컴포넌트를 사용하기 위해 컴포넌트가 제공하는 API를 입력한다(③). 이렇게 등록된 컴포넌트를 사용자에게 구조화 된 시각적 형태로 현재 프레임웍 상에 연결된 컴포넌트들을 보여준다(④).

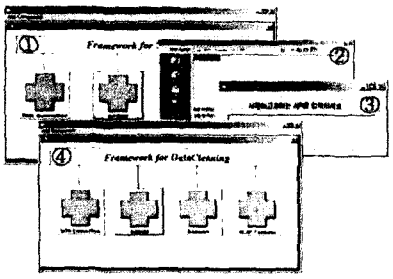

그림 2 프레임웍의 컴포넌트 등록

이렇게 프레임웍에 등록된 컴포넌트들은 프레임웍이 각 컴포넌트들의 연결을 버튼형태로 제공함으로, 해당하는 컴포넌트의 버튼을 클릭함으로써 컴포넌트를 실행시킬 수 있다. 앞서 구현한 데이터 정제 컴포넌트들은 컴포넌트 프레임웍 상에서 다음과 같이 동작한다.

(1) DSNConnection 실행
DSNConnection 컴포넌트는 그림 3과 같이 실행되며, DBMS 서버에 접속하고, 연결 값을 생성한다.

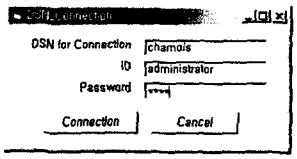

그림 3 DSNConnection 연결

(2) DAQUM 품질측정 및 결과 출력
DAQUM 컴포넌트를 통해 지식공학 데이터의 품질을 측정한다. 이 과정을 통해 지식공학 데이터의 오류를 찾아낼 수 있다. 우선 정제를 요하는 데이터베이스, 테이블 및 컬럼을 선정하고(그림 4①), 사용목적을 작성한다(②). 그리고 데이터의 오류를 측정하기 위해서 데이터 프로파일(③) 및 데이터 라이브러리(④)를 작성한다. 이러한 과정을 거친 데이터의 오류 측정 결과는 그림 5와 같이 오류데이터 별(①), 컬럼 별(②), 사용목적에 따른 결과(③) 3가지의 형태로 보여진다.

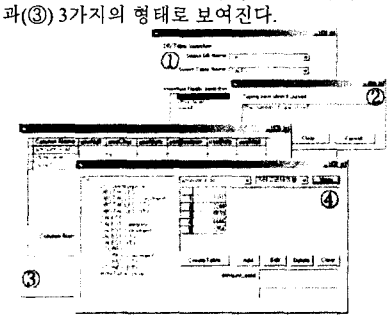

그림 4 DAQUM 품질측정

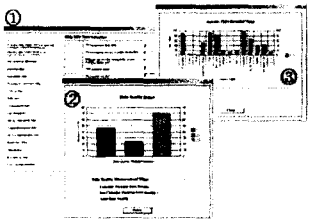

그림 5 DAQUM 품질 측정 결과

(3) DDCleaning 데이터정제
DAQUM 컴포넌트를 통해 분류된 오류 데이터는 DDCleaning 컴포넌트를 통해 수정 및 삭제과정을 거침으로써, 지식공학 데이터를 사용하고자 하는 사용 목적에 맞는 의미 있는 데이터로써 정제 된다. 그림 6은 리스트에 출력된 정제할 데이터들의 세부 정보를 수정할 수 있도록 동작하는 DDCleaning 컴포넌트를 보여준다.

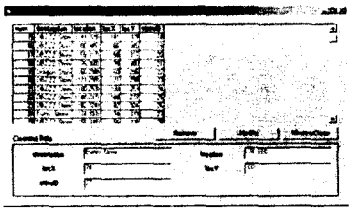

그림 6 오류데이터 정제

(4) OLAPPBrowser 처리
DAQUM과 DDCleaning을 거쳐 정제된 데이터를 이용하여, OLAP에서 사용자가 요구하는 다차원이나 계층적인 데이터로써 분류할 수 있도록 한다. 그림7은 선정한 측정값(①)에 대한 OLAP 처리 결과(②)를 보여주는 OLAPPBrowser의 동작을 설명한다.

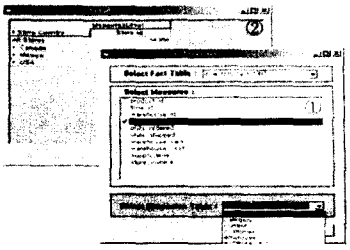

그림 7 OLAP을 이용한 정제된 데이터의 사용

5. 결론 및 향후 연구

대용량의 데이터를 가공하여 지식을 추출하는 지식공학 시스템에서, 개발된 시스템의 품질을 보증하고 가치를 높이기 위해서는 사용하는 데이터의 품질이 보증 되어야 한다. 지식공학 시스템으로 개발된 샤모아 지식공학 시스템도 데이터의 품질을 보증할 수 있는 데이터 정제 관련 컴포넌트가 필수적이다. 본 논문에서는 샤모아 지식공학 시스템에서 사용하는 데이터 정제와 관련된 4개의 컴포넌트 DSNConnection, DAQUM, DDCleaning, OLAPPBrowser의 기능 및 동작에 대해 기술하였다. 또한 구현된 컴포넌트들이 동작할 수 있는 프레임웍을 구현함으로써, 개발된 컴포넌트들의 구현을 테스트하기 위해, 실제 지식공학 시스템에서 동작시키는데 소요되는 시간과 비용을 줄일 수 있도록 하였다.

이러한 프레임웍에서의 동작과정을 거쳐 테스트 된 컴포넌트들은 샤모아 지식공학에 사용되는 데이터의 정제를 통해, 사용되는 데이터의 신뢰성을 향상시키고, 시스템 자체의 품질 향상에도 기여할 수 있다.

향후, 이들 컴포넌트를, 실제적인 샤모아 지식공학 시스템에 적용하여 다양한 어플리케이션에서 수행할 수 있도록 할 계획이다.

참 고 문 헌

[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", Computer, No.5, pp46-54, 2002.5.
[5] Richard Y.Wang, "A Product Perspective on Total Data Quality Management", Communication of the ACM, vol.41, on.2, pp.56-65, Feb. 1998
[6] 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
[7] 양자영, 최병주 "데이터 품질 측정 도구", 한국 정보 과학회 논문지: 컴퓨팅의 실제 제 9 권 제 3 호, pp278-288, 2003.6
[8] Won Kim, Byoung-Ju Choi, Eui-Kyeoung Hong, Soo-Kyoung Kim, Doheon Lee, "A Taxonomy of Dirty Data", The Data Mining and Knowledge Discovery Journal, Vol7 No.1, pp81-99, 2003.1