SOA에서의 오류 데이터 정제를 위한 서비스 개발

지은미*, 최병주*, 이정원**
*이화여자대학교 컴퓨터학과
**아주대학교 전자공학부
e-mail: jjiem@ewhain.net, bjchoi@ewha.ac.kr, jungwon@ajou.ac.kr

Developing the SOA-based Dirty Data Cleansing Service

Eun-Mi Ji*, Byoung-Ju Choi*, Jung-Won Lee**
*Dept. of Computer Science & Engineering, Ewha Womans University
**Dept. of Computer Electronic & Engineering, Ajou University

요 약

최근 e-Business 어플리케이션을 통합하기 위한 개념으로 서비스 지향구조 (Service Oriented Architecture)에 기본 원리를 둔 분산 소프트웨어 통합 기술이 널리 확산되고 있다. 따라서 각 서비스간의 데이터 정제기법을 통한 신뢰성 있는 데이터 교환은 필수적 요소로 자리 잡고 있다.

본 논문에서는 시스템에 상호작용 시 교환되는 데이터의 오류를 탐지하고 정제하기 위한 서비스로 사용자의 데이터 제약조건을 결합 시키는 변환 과정, 오류를 탐지하는 탐지과정, 탐지된 오류를 정제하고, 정보를 보여주는 정제과정으로 이루어진 오류 데이터 정제 서비스(DDCS; Dirty Data Cleansing Service)를 구현하고, 이를 이용하여 SOA기반 ESB상에서 통합된 시스템들 간에 상호 작용하는 오류 데이터 정제를 보장하는 서비스를 개발한다.

1. 서론

기존의 컴포넌트 기반 소프트웨어 개발 방법론이 분산 컴퓨팅 환경에서의 대규모 시스템 통합 요구의 증가로 서비스-지향 구조(Service-Oriented Architecture)로 변화하고 있다. [1] 이러한 움직임은 경영환경의 변화·협업의 증가로 인한 이기종 시스템 간의 호환, 고객의 요구에 맞는 다양하고 새로운 상품의 빠른 개발 등에 따른 내, 외부 환경 변화에 유연하게 변화할 수 있는 능력이 기업 경쟁력의 핵심으로 자리 잡고 있음을 보여준다.

따라서 SOA의 느슨한 연결(loosely-coupled), 서비스와 정보기술의 연계, 재사용 및 공동 활용에 기반 한 정보자원의 효율적 사용 및 관리 등은 이러한 조직의 요구사항을 만족시켜 줄 수 있는 새로운 패러다임으로 인식되고 있다.[2]

SOA를 구현하기 위한 통합 기술인 ESB에서의 시스템 간에 상호 작용 시 교환되는 데이터의 품질을 보장할 수 있다면 더 나은 서비스를 제공할 수 있다. 즉, 서비스 품질을 보장하는 동시에 서비스를 작성, 통합하고 관리하는 것[3]도 SOA에서의 중요한 부분이지만, 서비스들이 정확한 작업 수행이 가능하도록 서비스 간에 전송되는 데이터의 품질을 보장하는 것 또한 필수요소이다.

본 논문에서는 SOA를 지원하기 위한 최신 구현 기술인 ESB에서의 서비스 통합 과정에서 실제 데이터에 대한 품질을 보장해주는 오류 데이터 정제 서비스인 DDCS서비스를 개발하였다. 이 DDCS 서비스는 SOA의 기본 원리를 지원하기 위한 ESB의 독립적인 기능으로 사용될 수 있으며, XML을 이용한 DDCS 인터페이스의 형식을 정의하여 자동화된 데이터 오류 탐지 및 정제 기법을 통해 서비스의 품질을 높이고, CRM, ERP 등의 시스템과 같이 상호작용이 많은 시스템의 데이터를 효과적으로 관리할 수 있게 한다.

2. DDCS 서비스

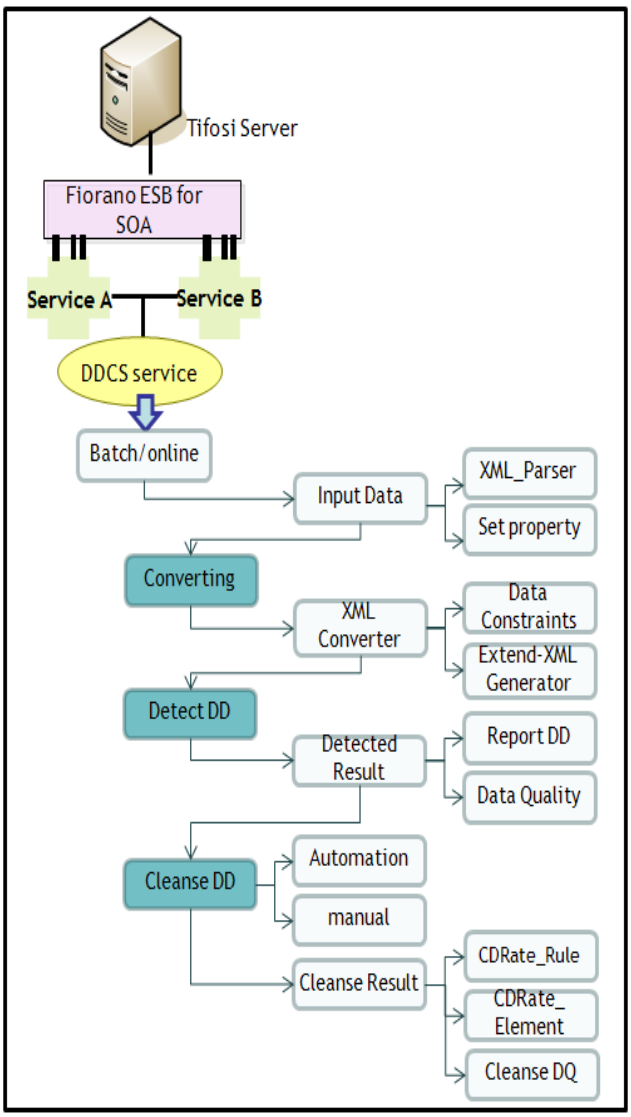
DDCS 서비스가 ESB상에서 제공되기 위하여 Fiorano사의 Fiorano Business Integration Suite인 FioranoESB™을 사용하여 구축하였다.

선행 연구[4] 로서 데이터베이스로의 데이터 수집, 통합, 저장 등에서 발생할 수 있는 오류들을 33가지로 분류하였고, 본 논문에서 요구되는 서비스간의 상호작용 시 발생할 수 있는 오류를 위해서 33가지의 오류를 재정리 하여 6가지 상호작용 데이터로 분류 하였고, 재정리한 오류 데이터 탐지 및 정제 규칙을 이용하여 SOA상에서 오류 데이터 정제를 위한 기본적인 서비스를 설계한다.

DDCS 서비스는 다른 서비스와 서비스를 연결하는 중간에 사용되며, 이는 한 서비스에서 다른 서비스로 데이터가 넘어갈 때 오류가 발생하는지 아닌지를 검사하고 정제

하는 서비스이다.

그림1은 오류 데이터 정제 서비스의 모듈의 흐름을 나타낸다.



(그림 1) 오류 데이터 정제 서비스 모듈

오류 데이터 정제 서비스는 SOA를 기반으로 하는 ESB 상에서 서비스를 개발하고 시스템을 통합하는 과정에서 이벤트가 발생하기를 기다리고 이벤트 발생 후 서비스의 입력으로 XML_Parser 모듈을 사용하여 데이터들을 분류한다. 서비스 사용자에게는 각 데이터에 따라 오류 데이터 측정 기준을 입력 받은 후, XML 문서로 변환한다. 사용자 입력 데이터와 측정 기준이 포함되어 있는 XML 문서를 가지고 오류 데이터를 탐지하고 그 결과를 보여주고, 서비스 사용자가 오류 데이터를 정제할 수 있는 환경을 제공한다. 오류 데이터를 정제한 후, 현재 입력된 데이터의 오류 측정 기준 및 데이터에 따른 오류 데이터 발생률과 서비스 적용 후 오류 데이터 발생률을 비교해 줌으로써 오류가 얼마나 정제 되었는지를 보여준다.

오류 데이터를 정제하는 서비스는 아래에 기술된 기능을 수행하는 모듈들로 구성되어 있다.

● 서비스는 Windows 2000 server 환경에서 Java 2 Platform, Enterprise Edition 1.4.2를 기반으로 구현되었고, SOA를 지원하는 ESB 상에서 서비스로 제공되기 위하여 Fiorano사의 Fiorano Business Integration Suite인 Fiorano ESB^{TM}을 사용하였다.

● 일괄처리/온라인 단계(Batch/Online): DDCS 서비스가 무엇을 대상으로 오류 정제를 측정할 것인지를 선택하는 모듈로써 오프라인의 데이터 품질에서 온라인 데이터 품질까지 그 대상을 확장하기 위한 기능을 제공한다.

● Input Data(서비스 입력 값) : 시스템을 구성하는 서비스의 입력으로써, XML 데이터가 이벤트를 통해 입력되는 지를 확인한다.

● 변환(Converting): 입력데이터와 사용자의 데이터제약조건을 결합시키는 과정을 제공한다.

● XML 변환(XML Converter): 입출력한 데이터를 데이터의 교환의 표준과 확장을 위해 XML로 명세화 한다.

● 오류 데이터 탐지 단계(Detect DD): 사용자에 의해 정의된 속성과 오류 데이터 분류법에 의하여 데이터 정제를 측정하여 오류 데이터를 탐지를 한다.

● 오류 데이터 탐지 결과 단계(Detected Result): 분류된 오류데이터를 리포트하고, 전체 데이터의 정제와 컬럼별 데이터 정제를 그래프로 보여준다.

● 오류 데이터 정제 단계(Cleanse DD): 분류 결과에 따라 사용자가 오류 데이터를 정제할 수 있는 환경을 제공한다.

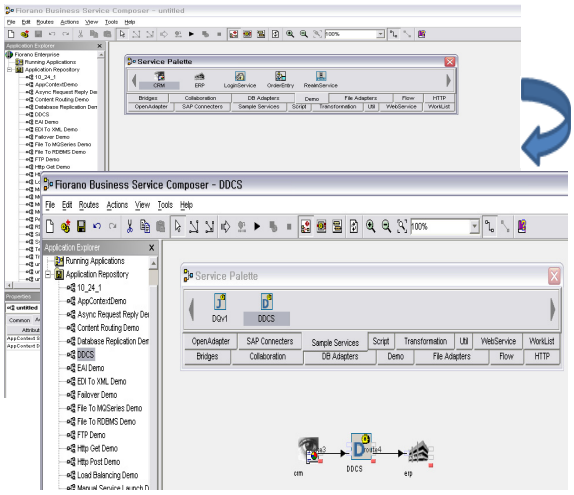
- 정제 단계(Automation): 사용자 요구에 따라 자동적으로 오류 데이터를 정제.
- 수동 정제 단계(Manual): 사용자에게 다시 정보를 입력 받아 오류 데이터를 수동적으로 정제.

● 오류 데이터 정제 결과 단계(Cleansed Result): 사용자로 하여금 오류데이터가 얼마나 정제되었는지를 쉽게 파악할 수 있도록 소스 데이터와 정제된 데이터의 오류 데이터 발생률을 비교함으로써 오류 데이터의 정제률을 보여준다.

3. DDCS 서비스 적용 사례

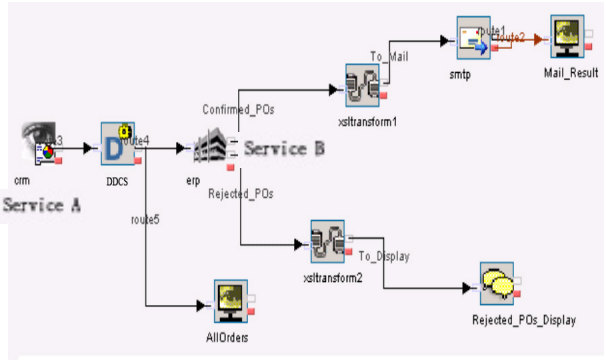
이 장에서는 본 논문에서 제안한 DDCS 서비스를 실제 분산 시스템의 통합에 적용한 결과, 본 서비스가 오류 데이터 정제에 얼마나 효과적인지 보일 것이다.

아래 그림 2에서처럼 Fiorano ESB^{TM} 를 사용하여 필요한 서비스를 Service Palette에서 선택하여 구성할 수 있다. Service를 끌어다 화면에 놓고 각 서비스 사이의 input/output port를 맞추어서 연결, 실행을 해볼 수 있고, 생성된 서비스는 Fiorano Business Service Composer를 이용 개발하여 Palette에 포함된다.



(그림 2) Fiorano ESB™ 서비스 통합화면

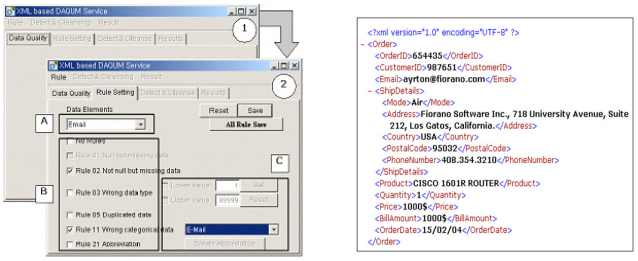
본 논문에서는 DDCS 서비스 활용사례로써 그림 3과 같이 SOA 시스템에 적용한 환경은 물품 거래 시스템으로 제품을 주문하는 CRM 서비스(Service A), 주문에 대한 승인과 거절을 결정하는 ERP 서비스(Service B)을 중심으로, 거래 정보를 전달하고 보여주는 여러 서비스들을 통합하여 서비스를 구성하였다. 이때 Service A와 Service B를 결합할 때 두 시스템의 상호 작용하는 데이터의 오류 관리를 위해 본 논문에서 개발된 서비스인 “오류 데이터 정제 서비스(DDCS)”가 오도록 하였다.



(그림 3) CRM 서비스 와 ERP 서비스 사이의 오류 데이터 정제 DDCS 서비스 적용 예

물품 구매자가 고객의 정보를 담당하는 CRM 서비스를 통하여 개인 정보와 함께 구매 정보를 제품 관리 서비스인 ERP 서비스에 전송한다. ERP 서비스는 전달 받은 정보를 가지고 주문을 승인할 것인지 거절할 것인지를 결정한 후, 구매자에게 이메일을 통해서는 승인된 결과를 전송하기도 하고 대화창을 통해 거절된 결과를 전송하기도 한다. 이러한 예는 여러 시스템들이 통합되고, 시스템 간에 대규모 데이터가 이동하는 대표적인 시스템으로 볼 수 있다.

구현된 DDCS 서비스가 CRM 서비스와 ERP 서비스 사이에 독립적으로 삽입되어 사용자가 지정한 목적에 따라서 사용자의 관점에서 데이터의 오류를 탐지하여 구매자의 잘못된 선택, 서비스 간 데이터의 처리 범위가 다른 경우, 네트워크 전송 장애로 인한 데이터 전송 오류 등을 포함한 CRM과 ERP간의 상호 작용 데이터의 오류들을 탐지해 낼 수 있다.

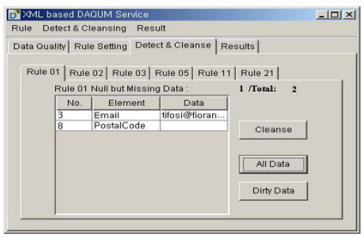


(a) 서비스 실행 (b) 서비스 간에 전송되는 실 데이터

(그림 4) 규칙 설정

CRM 서비스로부터 데이터 전송의 이벤트가 발생하면 위의 그림 4(a) ①에서와 같이 이벤트가 발생하기를 기다리고 ② 이벤트 발생 후 XML기반 데이터를 처리하여 엘리먼트 별로 오류 데이터를 측정하기 위한 기준을 설정한다. A부분은 입력된 데이터를 분류하여 사용자에게 엘리먼트만을 제시하여 각 데이터에 적합한 오류 데이터 탐지 규칙을 지정하도록 한다. B 부분은 선택한 엘리먼트에 어떤 규칙을 적용할 것인지를 사용자가 지정해주는 부분이다. C 부분은 오류 데이터를 탐지하는 규칙들 중에서 추가적인 정보가 필요한 경우 그에 해당하는 입력을 주기 위한 부분으로써, 그림4(a)에서는 Email이라는 이름을 가진 엘리먼트의 경우 오류 데이터 탐지 규칙 2와 11을 적용하고 규칙 11의 E-Mail 카테고리를 선택하는 것을 볼 수 있다. 탐지 규칙을 입력 받은 후 입력된 데이터와 오류 데이터를 탐지하는 규칙을 결합시킨 오류 데이터 정제를 위한 문서를 생성한다.

그림 4(b)는 CRM 서비스에서 ERP 서비스로 전송되는 데이터로 정제대상으로 선정한 총 15개의 엘리먼트를 가지는 XML데이터 이다.

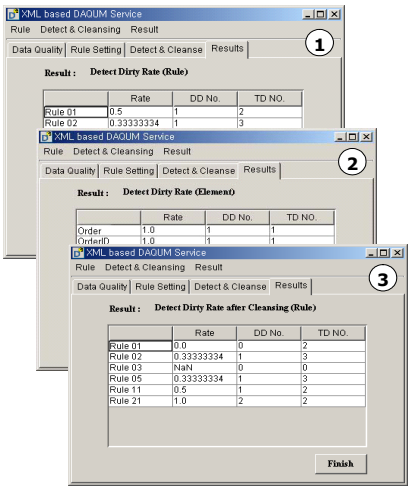


(그림 5) 탐지와 정제

다음 단계인 위의 그림 5에서는 데이터와 오류 데이터 탐지 규칙이 저장된 오류 데이터 정제 문서를 바탕으로 오류 데이터를 분류하고, 내부적으로는 extended-XML이 생성되고 이는 탐지 과정으로 넘어가 오류 데이터를 분류

하고, 사용자에게 보여줌으로써 사용자가 오류 데이터를 정제할 수 있는 환경(그림 5)을 제공한다.

사용자에게 규칙 1, 즉 'Null'도 허용하는 상태에서 empty인 데이터'는 오류 탐지를 위한 제약 조건을 적용한 데이터를 모두 보여 주는 화면이다. 이 규칙을 검사한 데이터는 총 2개의 데이터로 2번째 데이터인 'Email'과 8번째 데이터인 'PostalCode'를 보여주고 있다. 그 중, 하나만이 데이터가 위배되었음을 알려 주고 (즉 Rule 01 Null but Missing Data: 1/Total:2), 사용자는 이 탐지 결과를 보고 'Email'은 empty가 아니므로 empty 데이터인 'PostalCode'를 정제하게 된다. 그렇다면 'PostalCode'에 필요한 데이터를 사용자가 입력하여 오류 데이터를 정제한다. 이 외에도 사용자에게 데이터에 대한 모든 정보를 보여주거나('All Data') 오류 데이터만(Dirty Data)을 보여줄 수도 있다.



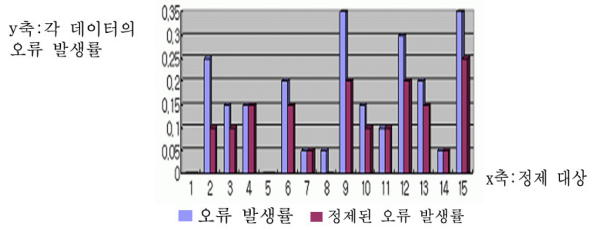
(그림 6) 결과

오류 데이터를 탐지하고 정제하는 과정이 끝나면, 그림 6과 같이 오류 데이터에 대한 결과를 보여 준다. ①은 측정 기준에 따른 오류 데이터의 발생률 ②는 데이터의 엘리먼트에 따른 오류 데이터의 발생률 ③은 정제된 데이터에 대해 기준과 같은 오류 데이터 측정 기준을 적용한 결과를 보여준다.

4. 적용 사례 결과 및 결론

오류 데이터를 정제하는데 효과적인지 보이기 위해 CRM에서 ERP로 전송되는 데이터 이벤트를 총 200회 발생 시켰다. 다음 그림 7을 보면 첫 데이터인 엘리먼트 1번과 엘리먼트 5번은 'Order'와 'ShipDetails'라는 XML문서의 계층 구조를 표현하기 위한 형식 데이터일 뿐 데이터 값을 가지지 않는다. 2번인 OrderID는 서비스를 사용하여 규칙 2, 3, 5에 의해 오류 데이터가 탐지되고 오류 데이터의 피드백을 통한 데이터 정제가 잘 이루어져 25%의 오류를 보이던 것이 본 서비스를 사용하여 정제한 후 10%로 감소된 결과를 보인다.

반면 입력된 데이터 중에서 4(Email), 7(Address), 11(Product), 13(Price), 그리고 14(BillAmount)번 데이터들은 상대적으로 낮은 정제율을 보였다. 그 이유는 4와 11, 13 데이터들은 탐지 규칙 11과 21에 의해 오류로 판정되고 정제되어야 하는데, 오류로 판단은 되었으나 정제 되어야 할 올바른 데이터를 알 수 없는 경우에 해당한다. 예를 들어 'Email'의 경우 'tifosiafiorano.com'의 경우 'string@string(.string)'에 위배되어 오류로는 탐지 된다. 이 데이터를 잘 아는 경우 'tifosi@fiorano.com'이라고 재입력 할 수 있지만 도메인 전문가조차도 확신할 수 없기 때문에 바로 정제할 수 없는 것이다.



(그림 7) 오류 발생률 비교

전체적으로 초기에 입력된 전체 데이터의 오류 발생률과 서비스 적용 후 정제된 데이터 비율은 18.08%에서 12.31%로 감소된 결과를 보였다. 이는 정제 비율로만 보았을 때, 오류 데이터를 약 31.91% 감소시킨 결과이다. 더 높은 정제 율을 보일 수 없는 것은 DDCS 서비스가 서비스 간 상호 작용 데이터에 국한된 것이며 자연어 처리 기법 혹은 상업적인 도구를 고용하고 있지 않다는 데 있다.

이렇게 SOA를 기반으로 서비스를 작성하고 통합하여 새로운 서비스를 만드는 경우, 본 논문에서 개발한 서비스 간의 상호 작용하는 데이터의 오류 탐지 및 정제 서비스를 이용한다면, 사용자에게 피드백을 줌으로써 최대한 유효한 데이터로 만들 수 있다는 것을 보였다.

개발된 서비스는 오류 데이터의 정제율을 더 높이기 위해 다양한 실험과 자동화 영역을 확장 시킬 것이다.

참고문헌

[1] 문은영, 이정원, 최병주, "ESB상에서 데이터 품질관리를 위한 서비스 개발", 한국정보과학회 가을 학술발표 논문집, Vol,31, No.2, pp517-519, 2004

[2] J.W.Lee, E.Y.Moon, and B.J.Choi, "Data cleansing for Service-Oriented Architecture",LNCS, Vol87-97,2005

[3] M.P.Papazoglou and D.Georgakopoulos, "Service-Oriented Computing", Communication of the ACM, Vol.46, No.10, pp25-28, 2003.10

[4] 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, Vlo7 No.1, PP81-99, 2003.1