SOA 기반 서비스간 상호 작용 데이터 관리를 위한 서비스

김은영° 이정원 최병주
이화여자대학교 컴퓨터학과
key-es@hanmail.net°, {jungwony, bjchoi}@ewha.ac.kr

A Service for Managing Interactive Data between Services based on SOA

Eunyoung Kim° Jung-Won Lee, Byoungju Choi
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 대규모의 분산 시스템을 통합, 구축하기 위한 S/W 설계 방법론으로 웹 서비스를 구현 기술로 한 서비스-지향 구조(Service-Oriented Architecture) 개념이 등장하였다. SOA를 기반으로 조립되는 서비스 간에 상호 작용하는 데이터는 인터페이스를 통한 형식적인 검증뿐 아니라 사용자의 의도에 맞게 사용될 수 있는지에 대한 실질적인 검증도 필요하다. 본 논문은 서비스-지향 구조에서 서비스간에 상호 작용하는 데이터의 오류를 실시간으로 탐지하고 데이터의 제약 조건을 학습시킴으로써 개발자의 수고를 덜고 e-business 시스템과 같이 상호 작용이 많은 시스템의 데이터를 효과적으로 관리할 수 있는 서비스를 개발한다.

1. 서 론

분산 시스템을 구축하기 위해 서비스를 기반으로 조립되는 웹 서비스 기술과 더 나아가 서비스-지향 구조(Service-Oriented Architecture)에서 각 서비스들은 다른 서비스에 독립적으로 작업을 수행하며 서로의 인터페이스를 통해서만 연결된다. 시스템을 구성한 후, 일반적으로는 시스템을 시뮬레이션 함으로써 각 서비스들이 잘 연결되었는지 확인하지만 이는 서비스 인터페이스에 국한된 부분으로 상호 작용하는 실제 데이터의 올바른 사용은 확인 할 수 없다. 따라서 서비스들이 정확하게 작업을 수행하는 지를 검사하기 위해서는 서비스 간에 흐르는 실제 데이터의 품질을 측정할 필요가 있다.

기존에 데이터베이스의 정적인 데이터를 대상으로 데이터 품질을 관리하는 몇몇 연구와 도구들이 있으나 이를 서비스간의 상호 작용하는 데이터에 적용시킨다면 실시간으로 데이터의 품질관리를 할 수 없으며 상호 작용 데이터의 특성이 조금만 달라져도 제약 조건을 만족시키기 위해 서비스 인터페이스뿐만 아니라 서비스 내부 도 변경해야만 한다. 또한 데이터의 제약조건을 단순한 매핑 방법을 이용하여 표현함으로써 오류 데이터를 걸러 낼 수도 있으나 이는 100% 인간의 개입을 전제로 한다. 거의 대부분의 데이터 품질 관리를 위한 오류 데이터의 탐지 및 정제과정에서는 인간의 개입을 전제로 하지만 오류데이터 탐지 및 정제 과정에서 인간의 입력을 학습시킴으로써 자동화를 꾀할 수 있다.

따라서 본 논문에서는 데이터베이스와 같이 정적인 데이터가 아닌 서비스간에 상호 작용하는 데이터의 품질 관리를 위해 오류 데이터를 탐지하고 탐지 과정에서 오류 데이터를 학습하여 인간 개입을 최소화 할 수 있는 방법을 제안하고자 한다. 개발된 오류 데이터 탐지 서비스는 SOA 기반에서 작성되는 서비스의 품질을 높이고, 서비스 작성시 개발자의 수고를 덜며, e-business시스템과 같이 상호작용이 많은 시스템의 데이터를 효과적으로 관리할 수 있게 한다.

2. 관련 연구

최근 분산 소프트웨어 통합기술은 기업 내에서 여러 소프트웨어들을 밀 결합(tightly-coupled)하는 차원이 아닌 e-business 어플리케이션들을 약 결합(loosely-coupled)으로 통합하는 데 있어서 웹 서비스 방식을 사용하기 시작하였다. 그러나 웹 서비스는 단순히 서비스를 어떻게, 어디에 기술하고, 어떻게 찾을 지를 기술하며 서비스-지향 구조의 구현 기술일 뿐이다[1]. 따라서 기업 내외 의 비즈니스 상호 작용을 지원하기 위해 서비스 작성(composition)과 관리를 위한 구체적인 기본 원리를 제공하고 있는 SOA와는 구별된다[2].

SOA를 기반으로 서비스들이 통합될 경우 전체 시스템의 동작을 시뮬레이션 함으로써 각 서비스들이 잘 연결되었는지 확인 할 수 있다. 하지만, 이때 연결된 각 서비스의 입력과 출력에 대한 명세가 올바르게 짝지어진 것인지를 확인할 뿐, 실제 데이터의 올바른 사용은 확인 할 수 없다. 즉, 서비스 품질을 보장하면서 서비스를 작성, 통합하고 관리하는 것이 SOA의 목표[2]이지만 서비스들이 정확하게 작업을 수행할 수 있도록 서비스 간에 전송되는 데이터의 품질을 보장하는 것은 개발자의 몫이다.

현재 데이터의 품질을 탐지하고 이를 정제하는 과정은

준비된 데이터의 품질이 결과의 정확도에 심각하게 영향을 미치는 데이터베이스와 데이터 마이닝과 같은 분야에서는 통계적인 방법을 이용하여 필수적으로 적용되고 있다[3]. 즉, 데이터베이스의 한 테이블 내에 중복된 값들을 찾는 문제, 스키마 매핑시 이름이나 구조상 충돌을 일으키는 데이터를 정제하는 문제 등으로 주로 데이터베이스의 오류 데이터를 찾아내는데 초점을 두고 있다. 한편 통계적인 방법을 이용하여 데이터 집합 내에서 의심스럽거나 잃어버린 데이터를 찾아내는 연구도 활발히 진행되고 있다.

그러나 이미 데이터가 모두 수집된 데이터베이스 내에서의 데이터 정제 기법만을 고려한다면 웹 서비스를 이용하는 SOA기반 시스템들에서 상호 작용하는 데이터의 품질을 고려할 수 있는 방법이 없다. 따라서 본 논문에서는 데이터베이스와 같이 정적인 데이터가 아닌 서비스간에 상호 작용하는 데이터의 품질 관리를 위해 오류 데이터를 탐지하고 탐지 과정에서 오류 데이터를 학습하여 인간 개입을 최소화 할 수 있는 방법을 제안하고자 한다.

3. 오류 데이터 탐지

3.1 오류 데이터 및 탐지 규칙 설정

우리는 선행 연구[4]로서 데이터 베이스로의 데이터 수집, 통합, 저장 등에서 발생할 수 있는 오류들을 33가지로 분류하였다. 본 논문에서 요구되는 서비스간의 상호작용 시 발생할 수 있는 오류를 위해서는 33가지의 오류를 재정리 하여 6가지 상호작용 데이터의 오류의 타입을 선별하였다. 다음 표 1은 데이터의 제약조건에 따른 탐지 규칙을 보여 준다.

표 1. 오류 데이터 타입에 따른 탐지 규칙

Rule	Error Type(E) / Detecting Rule(D)
1	(E) Null이 허용되지 않는 제약조건이 없는 상태의 결여 데이터
	(D) ' Null' 값은 허용하고 데이터 값이 존재하지 않는(empty) 데이터를 찾는다.
2	(E) Null이 허용되지 않는 제약조건이 있는 상태의 결여 데이터
	(D) ' Null' 값을 포함하여 데이터 값이 존재하지 않는(empty) 데이터를 찾는다.
3	(E) Wrong data type의 사용 (value range를 포함하는 data type 위반)
	(D) 설정된 value range의 허용 범위를 벗어난 데이터를 찾는다.
5	(E) duplicated data (non-null & uniqueness 조건 위반)
	(D) 이미 수집된 데이터 집합에서 중복되면 안 되는 제약 조건을 가진 데이터의 중복 여부를 확인한다.
11	(E) wrong categorical data (wrong abstraction level, out of category range data)
	(D) category lookup table 혹은 pre-built computerized abstraction hierarchy를 사용하여 category range에 벗어난 데이터를 찾는다. * Category lookup table (1) Arithmetic Expression (2) Number Expression (3) Date (4) Postal Code (5) Phone Number (6) E-Mail Address (7) Home Page (8) URL (9) File
21	(E) abbreviation
	(D) 사용자-정의 용어 사전에 그 데이터의 약어가 있는지 찾는다.

33가지의 오류 중 6가지의 오류 데이터만을 비즈니스 트랜잭션의 상호작용 데이터로서 빈번히 발생할 수 있는 상호작용 데이터로 취급하였는데 여기서 배제된 오류들은 주로 데이터베이스에서 발생하는 오류들이거나 다른 데이터와의 집단 비교를 통해서만 오류로 판단될 수 있는 것들이다. 또한 서비스 개발자의 제약 조건 명세에 따라 탐지 가능한 것만을 선택한 것으로 misspelling, extraneous data, 잘못된 필드에 입력, 모호한 데이터, incomplete context와 같은 데이터 오류는 특별한 종류의 도구, 예를 들면 철자 교정기나 데이터 수집 도구 등과 같은 상업적인 도구를 고용해야 하기 때문에 본 논문의 연구 범위에서는 제외한다.

여기에서 Rule 11은 특별한 도구를 사용하거나 시소러스를 고용한 상태에서 오류를 걸러 내는 것이 아니므로 오류를 완벽하게 찾아 낼 수는 없다. 그러나 정규 표현(regular expression)을 이용하여 Date는 yy/mm/dd, yyyy/mm/dd, 우편번호나 전화번호 자릿수, 메일 주소는 string@string(.string)+, 홈페이지나 URL은 http://(string.)+ string(/string) 에 맞추어 오류를 걸러 낼 수 있다.

3.2 탐지 규칙 설정의 자동화

서비스 개발자가 입력하는 데이터 제약 조건을 학습함으로써 유사한 데이터가 사용되었을 때, 설정될 수 있는 규칙을 추천해 주며 동시에 설정될 수 있는 규칙의 우선순위를 줌으로써 개발자의 탐지 규칙 설정을 자동화 할 수 있다. 이를 위해 데이터 제약 조건의 학습은 다음 두 단계를 거친다.

- 데이터 확장 (확장-엘리먼트 벡터 생성): 먼저 서비스간에 상호 작용하는 데이터를 유사한 단어들로 확장한 다음 정의 1에서 정의된 ' 확장-엘리먼트 벡터' 를 생성한다. 예를 들어 탐지 서비스가 ' quantity' 라는 데이터를 받아 들여 quantity→(QT, amount)로, ' order' 를 (ordering, purchase) 등으로 확장함으로써 유사 데이터에 설정될 규칙을 추천해 줄 수 있다.

(정의 1. 확장-엘리먼트 벡터)

// SynVect_UDL and SynVect_WN: WordNet과 사용자-정의 용어 사전의 유사어, 생략어, 합성어

// MaxSize : 확장-엘리먼트 벡터의 최대 길이

// <T> 는 엘리먼트, t 는 확장-엘리먼트 벡터의 용어

$\mathbf{DataSet}_i = (<T^i_1>, <T^i_2> \ldots, <T^i_m>)$, $T^i_p = (t^i_{p1}, t^i_{p2}, \ldots, t^i_{pa})$

when $a \leq$ MaxSize and p: random numbers, $t^i_{ma} \in$ SynVect_UDL or SynVect_WN

- 규칙의 학습 (규칙-빈도수 생성): 같은 데이터라 하더라고 조립되는 시스템의 목적에 따라 데이터의 제약 조건은 달리 설정될 수 있다. 예를 들어 ' orderID' 에 대해 규칙 2(not null), 3(wrong data type), 5(uniqueness)를 모두 적용하기도 하고 때로는 값의 범위를 주지 않고 규칙 2와 5만 적용할 수도 있다. 따라서 규칙-빈도수(RF: Rule Frequency)를 설정하고 데이터에 설정되는 규칙의 수를 카운트한다. 다음 정의 2는 정의 1에 덧붙여 RF를 가지는 확장-엘리먼트

벡터를 정의한다. 각 데이터의 규칙 별로 rf가 높으면 높을수록 가장 높은 설정 우선 순위를 갖는다.

(정의 2.) 규칙-빈도수를 가지는 확장-엘리먼트 벡터
// RF는 규칙-빈도수, 총 6개의 계수기 포함

$$\mathbf{RF}(T^i_p) = (rf^i_{p1}, rf^i_{p2}, \ldots, rf^i_{p6})$$

4. 서비스 개발 및 적용

SOA를 위한 서비스 개발 절차 및 방법은 우리의 선행연구로서 [5]에 기술되어 있다. 본 서비스는 windows 2000 server 환경에서 Java 2 Platform, Enterprise Edition 1.4.2를 기반으로 구현하고 SOA를 지원하는 ESB(Enterprise Service Bus) 상에서 서비스로 제공되기 위하여 Fiorano사의 Fiorano Business Integration Suite인 FioranoESB™을 사용하였다. 서비스는 다음 그림 1과 같이 크게 변환, 탐지, 정제와 규칙의 학습 과정으로 구성된다. 서비스 내부에는 입력 데이터와 사용자의 데이터 제약조건을 결합시키는 ' 변환' 과정, 오류 탐지에 필요한 정보를 추출하고 규칙에 따라 오류를 탐지하는 ' 탐지' 과정, 탐지된 오류 정보를 보여 주는 브라우저를 통해 데이터를 정제하고 오류 통계를 보여주는 ' 정제' 과정, 그리고 입력 데이터의 제약조건을 학습하는 ' 학습' 과정으로 이루어 진다.

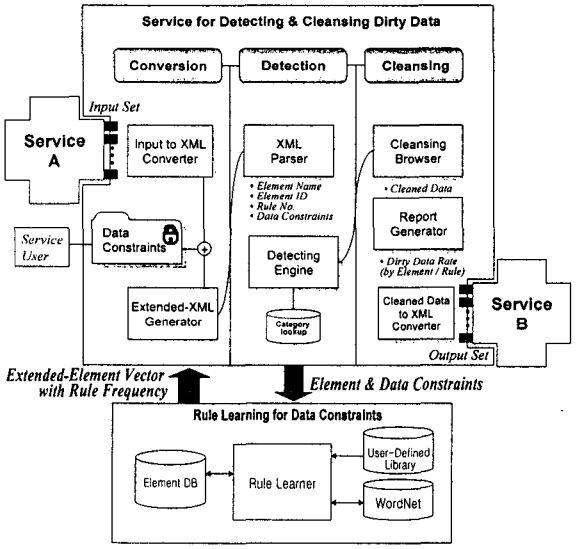

그림 1. 상호작용 데이터 오류 탐지 서비스

- *From Input to XML Converter*: 서비스 개발 시 입출력에 관련된 모든 명세를 XML로 명세화 한다.
- *Data Constraints*: 먼저 서비스 사용자로부터 입력 데이터들에 대해 적용되어야 할 규칙과 필요에 따라 값의 범위 및 카테고리에 대한 제약 조건을 입력 받고 규칙 학습기로부터 추천 받은 규칙들을 참고로 사용자가 직접 입력하거나 자동으로 설정할 수 있다.
- *Generating Extended-XML with Data Constraints*: 데이터 제약 조건은 XML로 변환된 입력 문서와 결합하여 각 데이터 별 제약 조건이 명시된 XML 문서로 확장된다.
- *Parsing Extended-XML*: 확장-XML 문서를 XML 파서를 이용하여 각 데이터 별로 element name, element ID, 처리되어야 할 규칙 번호, 카테고리 번호, 규칙에 필요한 제약 조건을 추출한다.
- *Detecting Dirty Data*: 파싱후 얻은 데이터 정보를 토대로 3.1절에서 제안한 탐지 규칙을 적용한다.
- *Cleansing*: 정제를 위한 정보는 오류로 탐지된 데이터 이름과 값을 동시에 제공한다.
- *Report Generation*: 오류 율에 따른 통계를 위주로 보고서를 생성한다.
- *Converting Cleaned Data to XML*: 정제를 거친 데이터가 있다면 정제된 값으로 교체한 XML을 생성한다.
- *Rule Learner*: XML 문서 파싱후 얻은 데이터를, WordNet과 사용자-정의 사전을 토대로 확장하고 제약 조건을 학습하여 규칙-빈도수를 생성한다. 규칙-빈도수는 차후에 입력되는 데이터가 기존의 학습 데이터와 유사할 때 설정해 두어야 할 데이터의 제약 조건을 추천해 주거나 자동 설정하는데 이용된다.

개발된 서비스를 물품 거래 시스템으로 제품을 주문하는 CRM 시스템과 주문에 대한 승인과 거절을 결정하는 ERP 시스템을 통합하는데 적용되어 사용자가 지정한 목적에 따라서 사용자의 관점에서 데이터의 오류를 탐지하여 구매자의 잘못된 선택, 서비스간 데이터의 처리 범위가 다른 경우, 네트워크 전송 장애로 인한 데이터 전송 오류 등을 포함한 CRM과 ERP간의 상호 작용 데이터의 오류들을 걸러내고 유사 데이터 사용시 규칙을 추천해 줄 수 있다.

5. 결론 및 향후 연구

본 논문은 기존의 정적인 데이터 베이스의 데이터 오류를 재정의하고 오류 탐지 기법을 적용하여 서비스간에 상호 작용하는 데이터의 품질 관리를 위한 서비스를 개발하였다. 또한 사용자의 개입을 최소화 하기 위한 방법으로 데이터의 제약 조건을 학습시킴으로써 서비스 개발자가 데이터의 제약 조건 설정 시 설정 규칙을 추천 받을 수 있게 하였다. 개발된 서비스는 현재 e-business 시스템 구축에 적용되고 있으며 다양한 실험을 통해 오류 데이터 탐지 율을 높이고 자동화 영역을 확장시킬 것이다.

6. 참고 문헌

[1] 이경하, 이규철, "웹 서비스의 표준화 동향과 발전 방향", 한국정보과학회 데이터베이스 연구회지, 제 19권 제 1호, pp80-87, 2003.3

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

[3] Theodore Johnson, and Tamraparni Dasu, "Data Quality and Data Cleaning", Tutorials of 10th SIGKDD, 2004.8

[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, Vol7 No.1, pp81-99, 2003.1

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