한국전산구조공학회
2005년 봄 학술발표회 논문집

강박스교 구조계산서 XML 시맨틱 모델의 스키마 매칭 기법 적용

Applying the Schema Matching Method to XML Semantic Model of Steelbox-bridge's Structural Calculation Reports

양 영 애* 김 봉 근** 이 상 호***
Yang, Yeong-Ae Kim, Bong-Geun Lee, Sang-Ho

ABSTRACT

This study presents a schema matching technique which can be applied to XML semantic model of structural calculation reports of steel-box bridges. The semantic model of structural calculation documents was developed by extracting the optimized common elements from the analyses of various existing structural calculation documents, and the standardized semantic model was schematized by using XML Schema. In addition, the similarity measure technique and the relaxation labeling technique were employed to develop the schema matching algorithm. The former takes into account the element categories and their features, and the latter considers the structural constraints in the semantic model. The standardized XML semantic model of steel-box bridge's structural calculation documents called target schema was compared with existing nonstandardized structural calculation documents called primitive schema by the developed schema matching algorithm. Some application examples show the importance of the development of standardized target schema for structural calculation documents and the effectiveness and efficiency of schema matching technique in the examination of the degree of document standardization in structural calculation reports.

1. 서 론

건설사업을 수행하면서 작성된 문서는 업무수행 사이 또는 사업 참여주체 사이에 정보의 전달을 위한 중간 매개체로서 중요한 역할을 담당하고 있다. 그러나 워드프로세서 등과 같은 특정 응용 프로그램을 이용하여 작성한 문서의 이용은 사업 참여주체들이 사용하는 응용 프로그램의 통일화를 필요로 하며, 특히 긴 공용기간을 가지는 토목시설물의 유지관리단계에서 여러 기관이 방대한 양의 문서를 신속하게 검색하여 참조하고 재활용하기에는 한계가 있다. 이러한 한계를 극복하기 위해 XML(extensible markup language) 기술을 건설분야 문서정보화에 도입한 연구가 최근 국내외로 활발히 진행되고 있다[(1)-(3)]. 이와 같은 기존 건설분야 문서정보화에 관한 연구는 각 사업단계에서 필요한 문서정보를 명세화하는 일종의 온톨로지 구축에 초점이 맞추어져 왔으며, 건설분야 문서정보를 통합하여 운영하기 위해 필요한 구체적인 응용 모델 및 기법을 제시하고 있지는 않다. 한

* 연세대학교 건설공학연구소 연구원
** 정회원 · 연세대학교 사회환경시스템공학부 박사과정
*** 정회원 · 연세대학교 사회환경시스템공학부 부교수

편, 전통적으로 다중데이터베이스 시스템에서 데이터를 통합하기 위해 논의된 스키마 매칭 기법이 최근에는 온톨로지 관점에서 카테고리간 혹은 각 객체간의 의미론적 유사성을 측정하는데 적용되고 있다[4]-[7]. 특히 정보산업 분야에서 데이터의 교환에 XML을 적극적으로 적용함에 따라 XML의 스키마 매칭에 관한 연구 또한 활발히 진행되고 있다[5]. 본 연구는 강박스교의 구조계산서를 대상으로 데이터베이스화하는데 적합하도록 상세화된 정보모델을 개발하며, 여러 기관으로부터 생성된 XML 형태의 구조계산서 정보를 통합하기 위해 필요한 구조계산서 XML 스키마 매칭 알고리즘을 개발하고 그 적용성을 테스트 하는데 목적이 있다. 이를 위해 기존에 작성된 강박스교 구조계산서들을 분석하여 표준화한 구조계산서의 시맨틱 모델을 도출하였으며, 도출된 문서구조는 XML DTD(document type definition)의 단점을 개선한 XML Schema를 이용하여 정의하였다. 또한 XML 스키마 매칭 알고리즘을 개발하기 위해 본 연구에서는 유사성 측정 기법과 릴렉세이션 레이블링(relaxation labeling)기법을 적용하였다. 개발된 구조계산서 XML 스키마 매칭 알고리즘은 본 연구에서 개발한 표준화된 강박스교 구조계산서 모델과 임의의 특정 강박스교 구조계산서를 대상으로 적용되었으며, 서로 다른 XML 문서구조를 가지는 구조계산서의 의미적 유사성을 효율적으로 측정할 수 있었다.

2. 강박스교 구조계산서의 XML 정보모델링

본 장에서는 건설CALS/EC 단체표준에서 권고하는 구조계산서 XML DTD가 데이터베이스화 하여 운영하기에는 의미론적인 측면에서 부족한 점을 개선하여 보다 상세화된 구조계산서의 정보모델을 개발하는 과정을 서술하였다. 본 연구에서는 대표적인 교량구조물인 강박스교의 구조계산서를 대상으로 XML 전자문서 모델링을 수행하였다.

2.1 강박스교 구조계산서의 XML 정보모델링 방법론

강박스교 구조계산서 전자문서모델을 개발하기 위해 본 연구에서는 서로 다른 기관에서 작성한 여러 종의 강박스교 구조계산서의 문서구조를 분석하였다. 기존 구조계산서는 구조물 설계기준과 해석방법, 해석결과분석 등 업무수행 절차에 따라 구성해 놓은 형태로 정형화되어 있었다. 그러나 큰 관점에서 업무수행 절차에 따라 정형화된 구조계산서이지만 작성자에 따라 작성항목 표현이나 세부항에 대한 문서구조의 차이로 모든 문서의 구조가 일치되어 있지는 않았다. 이에 따라 본 연구에서는 그림 1과 같은 방법론을 통하여 표준화된 강박스교 구조계산서의 XML 정보모델을 개발하였다.

그림 1에 나타낸 바와 같이 본 연구에서는 기존 강박스교 구조계산서를 분석하여 문서의 구성항목을 추출하고 각 구성항목의 속성을 추출하였다. 분석된 구조계산서의 문서구조에서 공통항목을 정의하기 위하여 공통항목을 그룹화하고 대표되는 항목명을 정한 다음 공통으로 활용할 가치를 갖는 문서정보에 대해요소와 속성을 정의하고 요소 사이의 계층 구조의 관계를 정의하였다. 이러한 공통항목의 정의 과정 이후 해당 교량의 특수성 등에 따라 사용되는 비 공통항목들은 공통항목에서 정의한 내용에 해당 문서에서 추가적으로 필요한 요소와 속성을 정의하였으며, 공통항목과의 관계에 따라 해당 데이터의 위치와 필수여부를 정의하였다. 본 연구에서는 요소명을 정의함에 있어서 도로교 설계기준에서 사

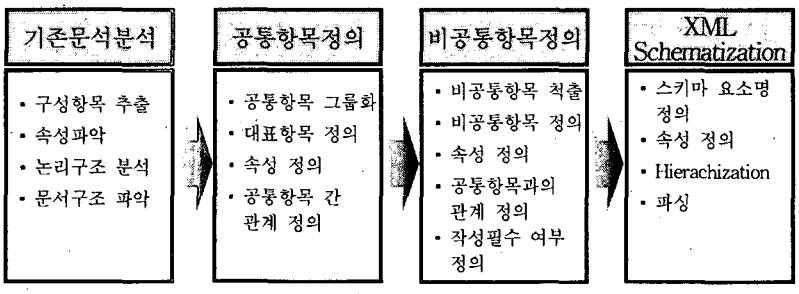


그림 1. 강박스교 구조계산서의 표준 문서구조 도출 과정

용되는 용어를 표준으로 하여 이용하였다. 최종적으로 이러한 일련의 과정에서 산출된 결과를 XML Schema를 이용하여 XML 문서모델링을 수행함으로써 표준화된 강박스교 구조계산서 XML 정보모델을 개발하였다.

2.2 강박스교 구조계산서의 문서구조와 항목정의

표준화된 구조계산서의 문서구조를 이루기 위해서는 최상위 위계에서 최하위 위계까지 포함하여 구조화할 수 있는 구조체계를 가져야 한다. 본 연구에서는 최상위요소 '강박스교구조계산서'의 1차 하위요소, 2차 하위요소의 해당 요소는 모든 강박스 구조계산서에 필수적으로 작성되어야하는 구조로 항목을 구성하였으며, 3차 하위요소 이하 최하위요소까지는 특정 정보항목에 제한되지 않는 특징을 가지도록 비 필수항목을 포함하여 정의할 수 있도록 하였다. 표 1은 기존 강박스교 구조계산서의 문서구조 분석을 통해 첫 단계에서 구성항목 추출과 논리구조 설정을 위해 정의한 문서구조 및 정보항목의 예를 나타낸 것이다.

2.3 강박스교 구조계산서의 데이터 속성 정의 및 강박스교 구조계산서의 XML 정보모델링

데이터 요소 분석 대상은 요소명과 요소명하위에 위계되는 자식요소와의 관계, 요소의 데이터형식, 요소의 작성여부, 요소의 작성 가능 개수 등이 있다. 표 2는 도출된 문서구조의 각 작성항목 중에서 '설계기준'의 1차하위 요소에 속한 일부분에 대하여 요소의 속성을 정의한 예이다. 각 단계별로 수행되는 요소에 해당하는 액션은 문자열, 숫자, 링크문서 등 문서 작성자의 적합한 데이터 형태로 입력하기, 모델에서 주어진 데이터 선택하기가 있다. 이에 따른 문서작성의 액션은 데이터 형태, 작성필수여부와 다중생성가능여부 등을 고려하여 각각의 정보형태에 따라 수행될 액션의 형태를 정해준다.

강박스 구조계산서의 작성항목에 대하여 최하위 요소에 대하여 데이터 분석을 수행한 요소들 사이에 위계표현이 될 수 있도록 XML Schema 표현기법을 이용하여 전체 강박스교 구조계산서의 각 해당 세부항목까지 고려한 모델을 개발하였다. 그림 3은 XML 스키마모델 표현기법을 이용하여 모델링한 강박스 구조계산서의 XML 표적스키마모델이다.

표 1. 문서구조 및 정보항목 정의

강박스교 구조 계산서	설계기준	…	…			
	바닥판설계	…	…			
	부재설계	주거더 설계	…	…		
			하중 산정	고정 하중	각 재료의 단위중량	…
					각 부위 작용하중	…
					각 부위 작용 팔길이	…
					각 부위 모멘트	…
					단면별 총 작용하중	…
					단면별 총 모멘트	…
				…	…	
			…	…		
		…	…			
	부대설계	…	…			
	사용성검토	…	…			

표 2. 전자문서모델 요소의 데이터 분석

어미요소	요소명	자료형태	작성여부
설계개요	…		
주구조 설계조건	교량상판정보	하위요소 소유	필수선택
	교량등급	정수	필수선택
	교량연장정보	하위요소 소유	필수입력
	교량폭원정보	하위요소 소유	필수입력
	…	…	…
	기초형식	문자열	비필수선택
	포장	문자열	비필수입력
	설계속도	실수형	비필수입력
	평면형상정보	하위요소 소유	비필수선택
	횡단구배	실수형	비필수입력
사용재료	…		
단면가정	…		

3. 유사도 평가 기법과 릴렉세이션 레이블링 기법을 이용한 XML 스키마 매칭 알고리즘

방대한 양의 정보를 다루는 다중데이터베이스 시스템에서 온톨로지의 정보 교류에서 생기는 두 개체간의 구조적 충돌이 생기는 문제점을 해결하기 위하여 응용된 기법이 스키마 매칭 기법이다. 그러나 정보산업 분야에서도 아직까지 스키마 매칭은 부분적으로 사람이 직접 수행하고 있어 시간이 지체되고 오류를 범하기 쉬워 비용이 많이 소비되고 있어 비효율적이다. 본 장에서는 서로 다른 기관에서 생성한 강박스교 구조계산서 XML 문서의 통합과정을 자동화하기 위해 필요한 스키마 매칭 알고리즘에 관하여 설명한다.

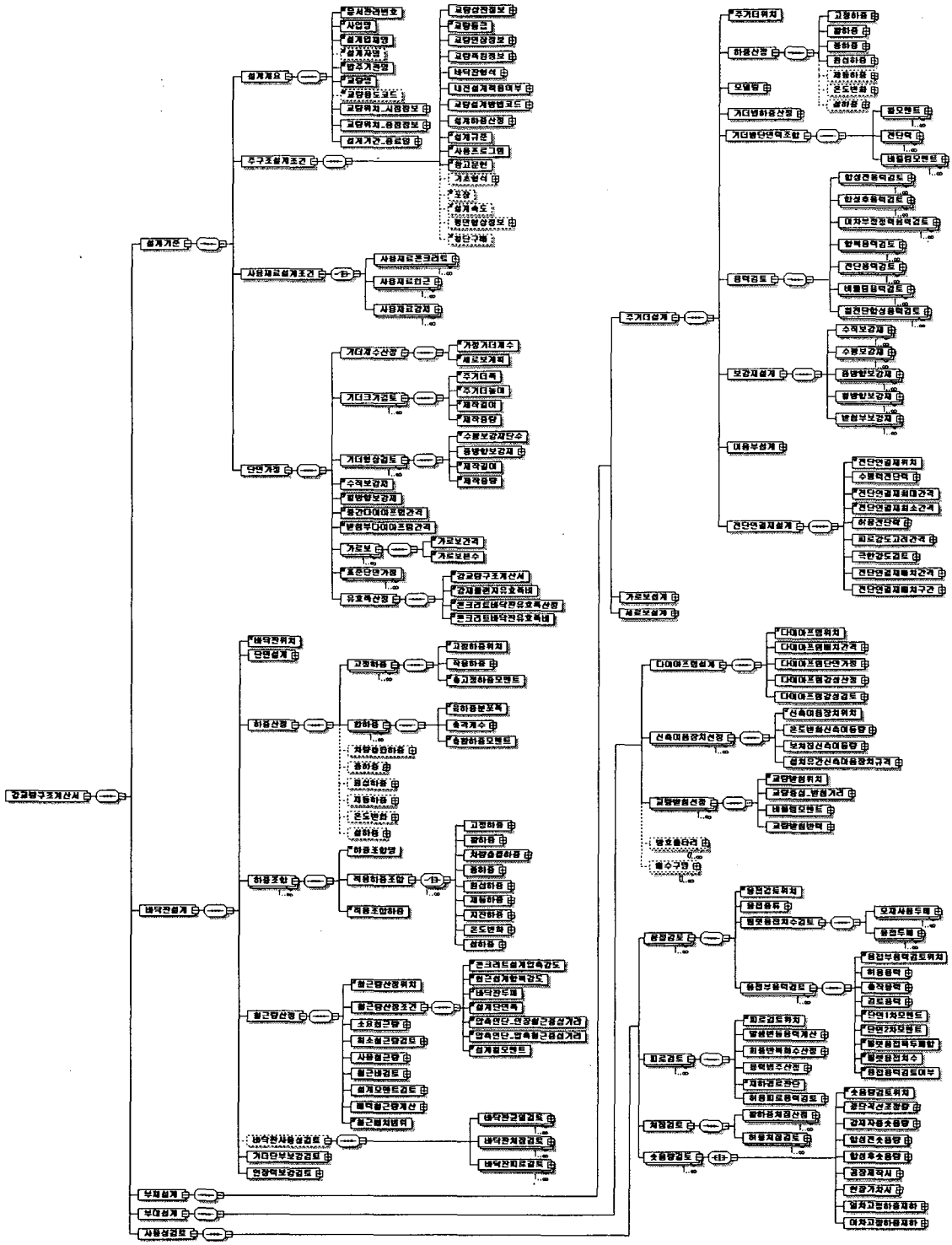


그림 2. 강박스교 구조계산서 XML 시맨틱 모델

3.1 XML 스키마의 유사도 평가 기법

본 연구에서 적용한 유사도를 평가하기 위한 기법은 Tversky와 Shafir[4]이 제안한 특성 매칭 모델(feature matching model)을 기반으로 한다. 본 연구에서는 XML Schema의 유사도 평가를 위해 각 요소의 특성을 요소명, 어미요소의 유무, 형제요소명, 자식요소명, 속성명의 5 가지로 구분하였다. 예로 비교대상이 되는 두 스키마 중에서 기준이 되는 표적 스키마의 요소 t_i와 비교의 대상이 되는 원시 스키마 요소 s_k의 요소명에 대한 유사도 $s_{name}(t_i, s_k)$는 식 (1)과 같이 나타낼 수 있다.

$$s_{name}(t_i, s_k) = \frac{N_{n(t_i)\cap n(s_k)}}{N_{n(t_i)\cap n(s_k)} + \alpha N_{n(t_i)/n(s_k)} + (1-\alpha) N_{n(s_k)/n(t_i)}} \tag{1}$$

여기서 $n(t_i)$와 $n(s_k)$는 각각 요소 t_i와 s_k의 요소명에 대한 단어들의 집합을 나타내며, N은 연산된 집합 성분의 개수를 나타낸다. 그리고 $n(t_i)\cap n(s_k)$는 집합 $n(t_i)$와 $n(s_k)$에 공통으로 갖고 있는 최소화 단어의 교집합을 의미하며, $n(t_i)/n(s_k)$는 집합 $n(t_i)$와 $n(s_k)$의 차집합을 의미한다. α는 기준 요소와 매칭 대상이 되는 요소사이의 관계성을 표현한 관계성 계수로서 0에서 1의 값을 가질 수 있으며, 본 연구에서는 0.5로 취하였다. 이와 같은 방법으로 각 요소의 속성명, 형제요소명, 자식요소명에 관한 유사도를 구할 수 있다. 다만 어미요소의 경우 해당 요소에 대하여 어미요소의 유·무 여부만을 고려하여 식 (2)와 같이 정의하였다.

$$s_{parent}(t_i, s_k) = \begin{cases} 1 \text{ if } parent(t_i) = parent(s_k) \\ 0 \text{ if } parent(t_i) \neq parent(s_k) \end{cases} \tag{2}$$

최종적인 두 요소 t_i와 s_k의 유사도는 식 (3)과 같이 각 요소의 특성들에 의하여 측정된 유사도의 합으로 정해진다.

$$\begin{aligned} s(t_i, s_k) = {} & w_n s_{name}(t_i, s_k) + w_a s_{attribute}(t_i, s_k) + w_b s_{brother}(t_i, s_k) \\ & + w_c s_{children}(t_i, s_k) + w_p s_{parent}(t_i, s_k) \end{aligned} \tag{3}$$

여기서 w_n, w_a, w_b, w_c, w_p는 각각 요소명, 속성명, 형제요소명, 자식요소명 및 어미요소에 대한 가중치를 의미하며, 각 가중치의 합은 1이다. 이들 가중치를 결정하기 위해서는 해당 스키마의 문서 특성과 스키마 매칭 목적을 고려하여 최적화된 가중치 적용을 위한 연구가 필요하다. 그러나 본 연구의 목적이 우선 스키마 매칭 기법의 적용성을 테스트 하는 정도이므로 본 연구에서는 모든 가중치를 0.2로 동일하게 두었다.

3.2 릴렉세이션 레이블링 기법을 적용한 스키마 매칭 알고리즘

릴렉세이션 레이블링은 제약조건을 이용하여 그래프상의 노드에 관련된 레이블을 할당하는 문제를 해결하는데 적용되어 왔다. $B = \{b_1, \cdots, b_m\}$ 및 $\Lambda = \{\lambda_1, \cdots, \lambda_n\}$가 각각 객체 집합과 레이블 집합이라 하면 정량화된 제약조건은 $R_{ij} = \{r_{ij}(k,l)\}$과 같이 실수를 가지는 적합강도로 이루어진 4차 메트릭스로 표현된다. 여기서 $r_{ij}(k,l)$은 레이블 λ_k가 객체 요소 b_i로 할당되고 레이블 λ_l이 객체 요소 b_j로 할당되기 위한 적합강도를 말한다. 두 스키마 중에서 기준이 되는 표적 스키마와 비교 대상이 되는 원시 스키마를 각각 S_T 및 S_S로 정의하면, 객체 집합 B는 S_T를 의미하며, 레이블 집합 Λ는 S_S를 의미한다. 3.1절에서 언급한 유사도 평가 결과는 S_T 및 S_S에 포함되는 비교 대상 요소만을 고려되는 반면에 릴렉세이션 레이블링 기법을 적용하면 전체 스키마의 구조와 이를 이루고 있는 제약조건을 고려하게 되므로 보다 최적화된 스키마 매칭 결과를 얻을 수 있다. 본 연구에서는 3.1절에서 언급한 유사도가 곧 두 스키마 요소 사이에 매칭이 될 수 있는 신뢰도를 나타내는 것으로 가정하였다. 이에 따라 릴렉세이션 레이블링 기법을 적용하여 개선되는 신뢰도는 식 (4)와 같다.

$$p_i^{(t+1)}(k) = \frac{p_i^{(t)}(k) q_i^{(t)}(k)}{\sum_{l=1}^{m} p_i^{(t)}(l), q_i^{(t)}(l)} \quad (4)$$

여기서 t는 반복되는 스텝이며, $p_i(k)$ 및 $p_i(l)$는 각각 λ_k가 b_i로 매칭될 수 있는 신뢰도 및 λ_l이 b_i로 매칭될 수 있는 신뢰도를 의미한다. 그리고 $q_i(k)$ 및 $q_i(l)$은 지지정도로서 $q_i(k)$의 경우식 (5)와 같이 나타낼 수 있으며, 앞서 가정한 사항에 따라 λ_k가 b_i로 할당될 초기 신뢰도 $p_i^{(0)}(k)$는 3.1절의 유사도 평가를 통해 최종적으로 얻은 $s(t_i, s_k)$를 이용하였다.

$$q_i^{(t)}(k) = \sum_{j=0}^{m-1} \sum_{l=0}^{n-1} r_{ij}(k,l) p_j^{(t)}(l) \quad (5)$$

그림 3은 유사도 평가 기법과 릴렉세이션 레이블링 기법을 도입한 스키마 매칭을 위한 알고리즘을 나타낸 것이다.

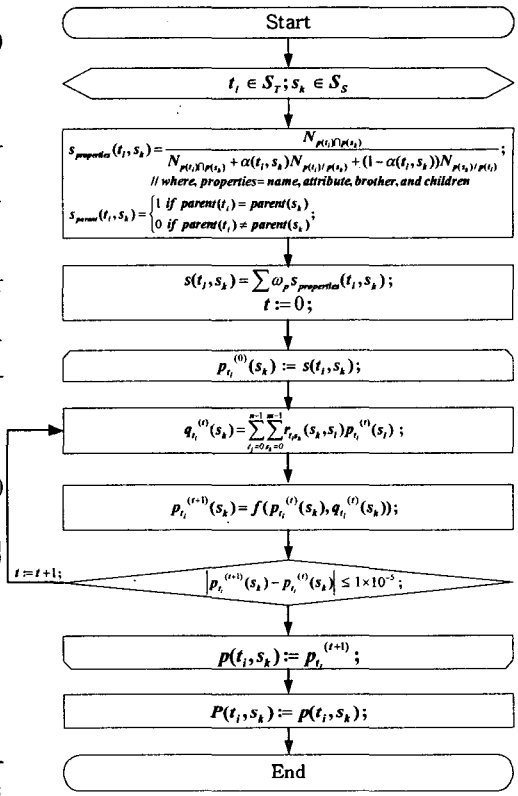


그림 3. XML 스키마 매칭 알고리즘

4. 강박스교 구조계산서 XML 스키마 매칭 수치예제

본 연구에서 개발한 강박스 구조계산서 XML 시맨틱 모델을 스키마 매칭 알고리즘을 적용하여 기존 구조계산서의 시맨틱 모델과 자동으로 비교하기 위한 수치해석을 수행하였다. 2장에서 기술한 구조계산서 XML 시맨틱 모델을 표적 스키마로 하고 그림 4와 같은 기존 강박스교 구조계산서를 원시 스키마로 하였다. 그림 5는 표적 스키마와 원시 스키마의 스키마 매칭 결과를 반복 차수에 따라 나타낸 그림이다. 그림 5의 표적 스키마의 요소번호 4, 6은 각각 '교량상판정보', '교량연장정보'이며 이들에 매치되는 원시 스키마는 각각 '구조형식', '지간구성'으로 정확히 표적 스키마의 요소와 원시 스키마의 요소가 매치되는 것이 아니라 표적 스키마의 요소에 포함되는 자식요소임을 알 수 있다. 표적 스키마의 원소번호 40과 49는 각각 '세로보설계', '교량받침선정'으로서 표적 스키마의 작성필수항목이고 원시 스키마에서 매칭되는 요소가 없기 때문에 원시 스키마에서 추가가 필요한 항목임을 알아낼 수 있다. 두 항목을 제외하고 위에서 기술한 낮은 수렴률을 보이는 항목이 비 필수항목이거나 새로운 문서항목임을 고려하면 그림의 결과에서 알 수 있는 것은 전체적으로 표적스키마의 높은 매치 신뢰도 값을 보임으로써 비교 대상이 구조계산서는 표적 스키마가 필요로 하는 정보를 적절히 포함하는 문서임을 판단할 수 있다. 그림 6은 각각 $P^{(t)}(t_0, s_0)$, $P^{(t)}(t_4, s_3)$, $P^{(t)}(t_1, s_{13})$, $P^{(t)}(t_2, s_8)$인 경우의 반복계산 경과에 따른 수렴되는 수치를 사람이 매칭을 실행했을 때의 판단기준과 비교한 그림이다. $P^{(t)}(t_0, s_0)$과 $P^{(t)}(t_1, s_{13})$은 2 번째 반복계산 경과 후 수렴하는 빠른 수렴률을 보이고 $P^{(t)}(t_4, s_3)$과 $P^{(t)}(t_2, s_8)$은 다소 느린 수렴률을 보임으로써 각 요소마다 매칭신뢰도의 수렴률은 다름을 알 수 있다. 그러나 그림 6의 세 경우 모두 매뉴얼매칭과 일치함으로써 알고리즘의 높은 정확도를 보였다.

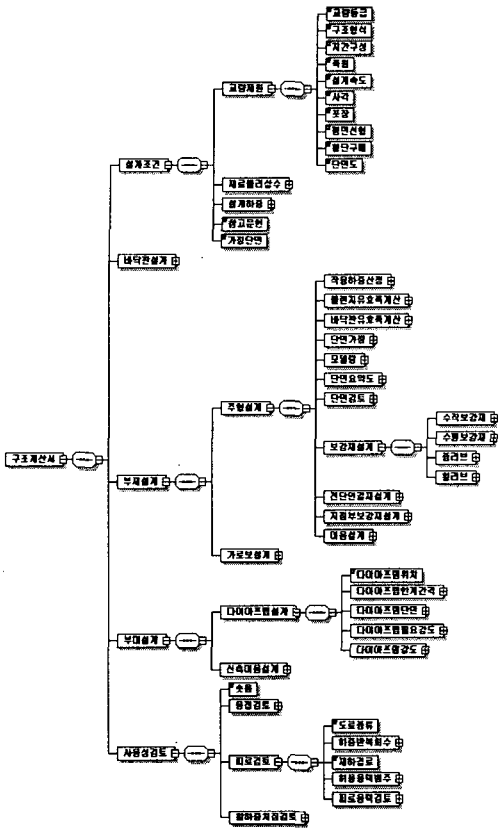
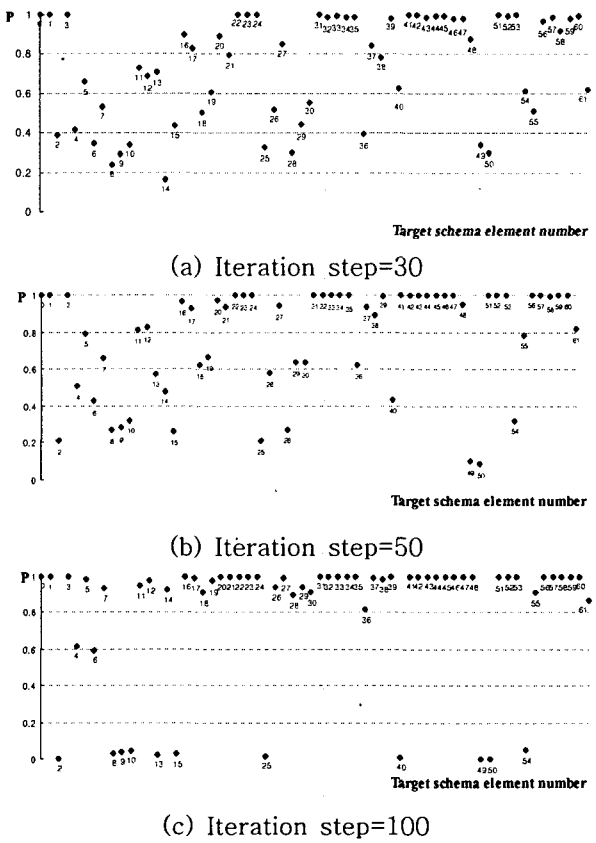


그림 4. 스키마 매칭 대상의 시맨틱 모델

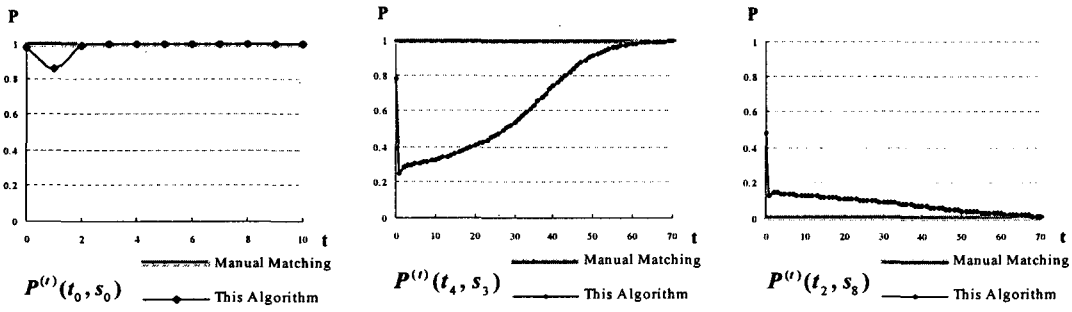


(a) Iteration step=30

(b) Iteration step=50

(c) Iteration step=100

그림 5. 반복 차수에 따른 전체 요소의 신뢰도 변화



(a) 최상위 요소 (b) t_4요소와 s_3요소 (c) t_2요소와 s_8요소

그림 6. 반복 차수에 따른 신뢰도 변화와 매뉴얼 매칭 결과와의 비교

5. 결 론

본 연구에서는 강박스교 구조계산서 XML 시맨틱 모델을 개발하고 다른 문서구조를 가지는 강박스교 구조계산서와 자동으로 비교 검토할 수 있는 스키마 매칭 알고리즘을 개발하였다. 본 연구에서 제안한 강박스교 구

조계산서 XML 시맨틱 모델은 기존의 여러 구조계산서를 분석하여 표준화된 문서구조를 가지며, 여러 형태의 데이터베이스나 응용 프로그램에서 활용될 수 있도록 플랫폼에 의존적이지 않은 XML Schema를 이용하여 모델링 되었다. 또한 본 연구에서 개발한 강박스교 구조계산서 XML 스키마 매칭 알고리즘은 비교 대상이 되는 스키마내의 각 요소들 간의 유사도를 평가하는 유사도 평가기법과 스키마의 전체 구조를 이루고 있는 제약조건을 고려한 릴렉세이션 레이블링 기법을 이용하여 개발되었다. 스키마 매칭 알고리즘을 적용한 수치해석 결과 패턴인식 등의 특별한 학습 알고리즘을 추가하지 않더라도 매뉴얼 매칭결과와 잘 일치하는 결과를 얻을 수 있어 스키마 매칭 기법이 여러 기관에 산재한 건설분야의 문서를 통합하는 데에도 적절히 이용될 수 있음을 확인하였다. 다만 본 연구에서는 스키마 매칭을 수행함에 있어 여러 파라매터 값을 평균치로 적용하였으므로 구조계산서에 타당한 패라매터를 결정하기 위한 연구가 필요할 것으로 판단된다. 본 연구에서 개발한 강박스교의 상세화된 구조계산서 시맨틱 모델은 범 국가적 차원에서 재난관리 및 효율적인 시설물관리를 위해 필요한 통합 데이터베이스 구축에 활용될 수 있으며, 구조계산서 XML 스키마 매칭 알고리즘은 서로 다른 관리기관에서 운영되는 XML 구조계산서의 통합을 자동화시키는데 활용될 수 있다.

감사의 글

본 연구는 건설교통부에서 실시한 건설핵심기술연구개발사업(교량설계핵심기술연구단)의 연구비 지원에 의해 수행되었으며, 이에 깊은 감사를 드립니다.

참고문헌

1. 박재원, 최재황, "건설공사정보 메타데이터의 XML 스키마 설계에 관한 연구", 한국문헌정보학회지, 제36권, 제3호, 2002, pp. 155-179.

2. Lee, S.-H., Kim, B.-G., Jeong, Y.-S., and Kang, H.T., "The Bridge Design Process with Web-based Documents", *Proceeding of The Third International Conference on Advances in Structural Engineering and Mechanics (ASEM'04)*, 2004, pp. 1198-1204.

3. Zhiliang, M., Heng, L., Shen, Q.P., and Jun, Y., "Using XML to support information exchange in construction projects", *Automation in Construction*, Vol. 13, Issues 5, 2004, pp. 629-637.

4. Tversky, A. and Shafir, E., *Preference, Belief, and Similarity*, A Bradford Book, The MIT Press Cambridge, 2004

5. Yi, S., Huang, B., and Chan, W.T., "XML application schema matching using similarity measure and relaxation labeling", *Information Sciences*, Vol. 169, 2005, pp. 27-46.

6. Pelillo, M., "The dynamics of nonlinear relaxation labeling processes", *Journal of Mathematical Imaging Vision*, Vol. 7, 1997, pp. 309-323.

7. Torsello, A. and Hancock, E.R., "Computing approximate tree edit distance using relaxation labeling", *Pattern Recognition Letters*, Vol. 24, 2003, pp. 1089-1097.