엔지니어링 문서 스키마의 효율적 매칭을 위한 데이터마이닝 기법의 활용방안

A data mining approach for efficient matching of engineering document schemata

박 상 일* • 안 현 정** • 김 효 진*** • 이 상 호†
Park, Sang Il • An, Hyun Jung • Kim, Hyo-Jin • Lee, Sang-Ho

요 약

본 연구에서는 데이터 저장의 질적 향상을 도모하는 XML 스키마 매칭의 효율적 활용방안을 제시하였다. 이를 위하여 매칭의 가중치의 변화에 따라 달라지는 정확도 데이터를 수집하고, 수집한 데이터를 활용하여 데이터 마이닝 기법 중 하나인 의사결정나무 모델을 수립하였다. 수립모델을 응용하여 구현한 가중치 자동선정 모듈은 설명변수인 교량의 형식, 문서가 포함하고 있는 요소의 수, 문서를 작성한 회사 등의 값에 따라 의사결정나무 모델의 목표변수인 정확도뿐만 아니라, 가장 높은 정확도를 보일 수 있는 가중치까지 간접적으로 제안가능하다. 본 연구로 구현한 모듈을 통해 제안된 XML 스키마 매칭 가중치를 활용하면 그렇지 않은 경우에 비하여 약 10% 정확도 상승효과가 있음을 알 수 있었다.

Keywords: 데이터 마이닝, 의사결정나무, 엔지니어링 문서, XML 스키마 매칭

1. 서 론

사회기반시설물을 효과적이고 안전하게 운영/관리하기 위해서는 생애주기동안 발생하는 관련 엔지니어링 정보를 체계적으로 저장하고, 정확하게 분배하는 기술을 통한 의사결정지원 시스템의 구축이 필수적이다. 한편, 구조계산서나 안전진단보고서 등은 시설물의 생애주기동안 생산되는 중요한 엔지니어링 문서들로, 국토해양부나 한국시설안전기술공단(2004) 등의 공공기관에서 관련 해당문서를 보관하고 있어 활용 가능한 정보의 양은 급격히 팽창하였다. 그러나 Liu *et al.* (2006)이 지적한 바와 같이 여러 비슷한 문서에서 일부분에 대한 정보만을 필요로 하는 대부분의 엔지니어링 업무에서, 수집한 정보를 보다 효율적으로 활용하기 위해서는 정보의 생산 및 저장의 차원을 넘는 통합 및 분배 기술의 필요성이 제기되었다. 정보의 정확한 통합 및 분배는 수집 데이터의 품질과 구조화가 선행되어야 하며, 이는 이종의 스키마간 의미적 관계를 식별하는 대표적인 기술인 스키마 매칭을 통한 표준화된 구조에 따른 정보 변환으로 구현 가능하다(Rahm and Bernstein, 2001). 이러한 점에 착안하여 Lee *et al.* (2006)은 XML 스키마 매칭 기법을 활용한 구조계산서가 포함하는 정보항목에 대한 질적 향상방안 연구를 수행하였지만, 매칭의 정확도와 관련한 가중치 선정에 대한 부분은 포함하지 않아 다양한 형태의 엔지니어링 문서에 그대로 적용하기는 힘들다. 본 연구에서는 시설물의

* 학생회원 • 연세대학교 토목환경공학과 박사과정 si@csem.yonsei.ac.kr
** 학생회원 • 연세대학교 토목환경공학과 박사과정 ahj38@csem.yonsei.ac.kr
*** 정회원 • 연세대학교 토목환경공학과 연구교수 jinski@yonsei.ac.kr
† 정회원 • 연세대학교 토목환경공학과 교수 lee@yonsei.ac.kr (교신저자)

생애주기동안 발생하는 정보의 질적 향상을 위해 박상일 등(2009)이 제안한 방법을 통해 구조화환 대표적 엔지니어링 문서인 교량 구조계산서를 대상으로 데이터마이닝 기법의 하나인 의사결정나무를 활용하여 XML 스키마 매칭의 정확도를 효율적으로 향상시킬 수 있는 방안을 제시하였다.

2. 의사결정나무를 활용한 매칭 가중치의 간접적 선정방안

2.1 XML 응용 스키마 매칭 기법 및 매칭 가중치에 따른 스키마 매칭의 정확도 변화

Yi *et al.* (2005)가 제안한 XML 스키마 매칭은 XML 요소의 데이터 타입을 배제하면서 두 스키마의 언어적-구조적 유사성을 정량적으로 나타낼 수 있기 때문에 교량의 구조계산서와 같이 임의로 작성한 문서 적용하기가 한층 용이하다. XML 스키마 매칭은 요소 간 유사성을 측정하는 과정과 문맥의 제약에 기반을 두면서 유사성 측정과정의 신뢰도를 향상시켜 주는 릴렉세이션 레이블링의 두 가지 과정을 거친다. 이 중 매칭 항목 간의 가중치 설정으로 정확도의 변화를 일으키는 부분은 유사성 측정과정으로 식 (1)과 같이 XML 스키마 a(원시 스키마)와 b(표적 스키마)의 자기 자신의 요소(NE), 형제 요소들(S), 자식 요소들(C), 부모 요소(P)의 각 유사성의 합으로 산정된다.

$$S_T(E^a, E^b) = \rho_{NE} S_m(E^a_{NE}, E^b_{NE}) + \rho_S S_m(ES^a_S, ES^b_S) + \rho_C S_m(ES^a_C, ES^b_C) + \rho_P S_m(ES^a_P, ES^b_P) \quad (1)$$

식 (1)에서 ρ_X는 X 요소 각각의 매칭 유사성 가중치를 나타내는 것으로, $X \in \{NE, S, C, P\}$이이며, $\sum_X \rho_X$ = 1이다. 그림 1은 단어 간 유사성 가중치 ρ_{NE}, ρ_S 및 ρ_C의 비를 모두 1로 고정한 후 ρ_P 값의 변화에 따라 도출되는 매칭 정확도를 나타낸 것으로, 가중치의 변화에 따라 정확도가 약 83% ~ 94%의 범위에서 (평균 정확도: 약 86%) 변할 수 있음을 나타내고 있다. 특히, 각각의 가중치 ρ_X는 그 합이 1이라는 것 이외에는 아무런 연관성이 없으며, 이는 높은 정확도의 매칭 결과를 얻기 위해서는 문서의 각 종류에 따라 다른 가중치를 정해주어야 함을 의미한다. 따라서 본 연구에서는 적절한 매칭 가중치를 산정하기 위하여 대상으로 하는 문서인 교량 구조계산서의 포함 항목 수, 교량의 형식, 문서를 작성한 회사를 설명변수로 교량의 상, 하부 구조계산서 20종, 매칭결과 데이터 580개를 활용하여 의사결정나무(Decision tree)를 활용한 XML 스키마 매칭에 필요한 가중치를 간접적으로 제안하는 연구를 수행하였다.

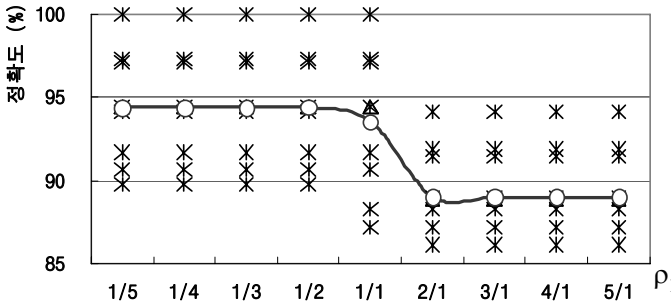

그림 1 ρ_P에 따른 매칭 정확도 변화

2.2 기존 데이터를 활용한 의사결정나무 모델의 생성

의사결정나무는 대상이 되는 집단을 소집단의 나뭇가지 형태로 구분하는 데이터 마이닝(Data mining)의 한 기법으로, 해석의 용이성, 변수간의 상호작용의 효과, 비모수적(Nonparametric) 모형이라는 점에서 통계모델 구성에 활용성이 높다. 본 연구에서는 표 1과 같은 설명변수를 활용하여 범주형 정확도 범위를 산정하는 의사결정나무 모델을 통계분석 프로그램인 SAS 9.1을 활용하여 구성하였다. 모델 구성을 위한 학습용 데이터(Training data set)와 수립모델 검증을 위한 검증용 데이터(Validation data set)는 6대 4의 비율로 산정하였고, 최종 모델은 대표적인 의사결정나무 구현 모델 CHAID, C4.5, CART 중 검증용 데이터의 유효 오분류

율이 가장 낮은 CART 모델을 사용하였다. 수립모델은 최종 선택의 경우를 나타내는 나뭇잎(leaf) 35개로 나타났으며, 생성된 종단마디(node)는 대체적으로 교량의 형식, 문서 작성회사, 문서 요소의 수, 매칭에 사용하는 가중치의 순으로 나타났다.

표 1 의사결정나무 모델 구성에 사용한 목표변수와 설명변수

변수종류	내용	변수 범위	
목표변수	매칭 정확도	A: 100% B: 95% ~ 99% C: 90% ~ 94%	D: 85% ~ 89% E: 80% ~ 84% F: <= 79%
설명변수	ρ_{NE}	연속형 변수	
	ρ_S	연속형 변수	
	ρ_C	연속형 변수	
	문서 요소 수	연속형 변수	
	교량의 형식	SA: 사장교 KB: 강박스교 KP: 강플레이트교	SUB_V: V형 교각 SUB_T: T형 교각
	문서 작성 회사	C_D: D 엔지니어링 C_Y: Y 엔지니어링 C_M: M 엔지니어링	C_S: S 엔지니어링 C_K: K 엔지니어링

2.4 수립한 의사결정나무 모델을 활용한 XML 스키마 매칭 가중치의 산정방안

의사결정나무 모델의 활용은 설명변수를 활용하여 새로운 목표변수를 도출해내는데 주로 사용하지만 본 연구에서는 최적의 목표변수를 나타낼 수 있는 설명변수를 선정하기 위해 의사결정나무 모델을 활용하기 때문에 설명변수에 따라서 순차적으로 종단마디를 추적하는 일반적인 방법은 사용하기 힘들다. 따라서 본 연구에서는 종단마디에서 확실히 판단 가능한 설명변수 내용을 포함하고 있으면 해당 가지(branch)를 따라가되, 판단가능하지 않는 설명변수인 매칭 정확도와 관련한 종단마디에서는 모든 경우의 가지를 추적하는 "선택과 보류"의 방법을 활용하였다. 그림 2는 수립 모델의 활용성을 검증하기 위하여 시험용 데이터(Test data) {C_S 엔지니어링 회사에서 문서를 생산한 상부구조 사장교 형식의 1045개의 문서 요소 수}를 포함하고 있는 구조계산의 가중치 선정 프로세스를 나타낸 것이다. "선택과 보류"의 방법을 활용하여 선정된 최종후보 나뭇잎은 그림 2에서와 같이 ①, ②, ③, ④의 네 가지 경우이고, 이때 나타난 정확도는 ①인 경우에 {B: 12.5%, C: 62.5%, D: 25%}, ②인 경우에 {B: 25%, C: 75%}, ③인 경우에 {B: 33.3%, C: 66.7%}, ④인 경우에 {B: 75%, C: 25%}로 ④인 경우의 정확도가 가장 높으며, 따라서 {사장교, C_S, 1045개 문서 요소 수}의 시험 데이터에서는 ρ_C >= 0.3229, ρ_{NE} >= 0.1742의 값으로 간접인 가중치가 선정이 된다. 이를 활용하여 ρ_{NE} = 0.2, ρ_S = 0.23 ρ_C = 0.35, ρ_P = 0.22 일때의 XML 스키마 매칭 정확도를 산정해보면 가중치 선정 프로세스를 거치지 않았을 때의 평균 매칭 정확도 약 86%에 비하여 10% 이상 상승한 약 98.08%의 정확도를 보인다.

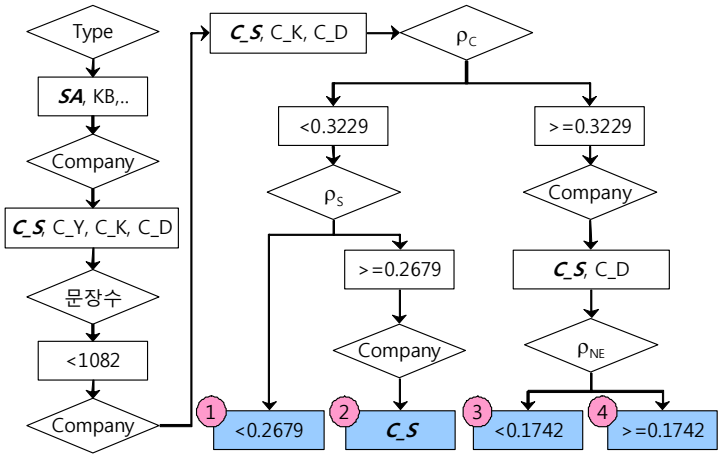

그림 2 {사장교, C_S, 1045개의 문서 요소 수} 경우의 매칭 가중치 선정 프로세스

2.5 수립한 의사결정나무 모델을 활용한 XML 스키마 매칭 가중치 산정 모듈의 구현

본 연구에서는 수립한 의사결정모델을 데이터베이스에 저장하고, 자동으로 매칭 가중치를 제안하는 시범적 모듈을 구현하였다. 그림 3은 .NET Framework 3.0과 Cubrid DBMS를 활용하여 Windows 기반 자동 가

중치 선정 인터페이스를 나타낸 그림으로, {사장교, C_S, 1045개 문서 요소 수}의 설명변수인 경우에 제안된 XML 스키마 매칭의 가중치를 나타낸다. 제시한 모듈을 활용하면 시험용 데이터 {강플레이트교, C_M, 문서 요소 수 832}의 경우에는 정확도 B 100%의 확률로 ρ_S >= 0.3229, ρ_C <= 0.3229가 제안되며, ρ_{NE} = 0.18, ρ_S = 0.35 ρ_C = 0.3, ρ_P = 0.17을 활용하여 실제 매칭 정확도를 구해보면, 약 95.13% 매칭 정확도가 산출된다. {사장교, C_D, 문서 요소 수 1248}의 경우에는 정확도 B 약 66.7%, C 약 33.3%의 확률로 ρ_{NE} >= 0.2111, ρ_S >= 0.2020이 제안되며, ρ_{NE} = 0.25, ρ_S = 0.23 ρ_C = 0.26, ρ_P = 0.26을 활용하여 실제 매칭 정확도를 구하면, 약 92.124%의 매칭 정확도가 산출된다.

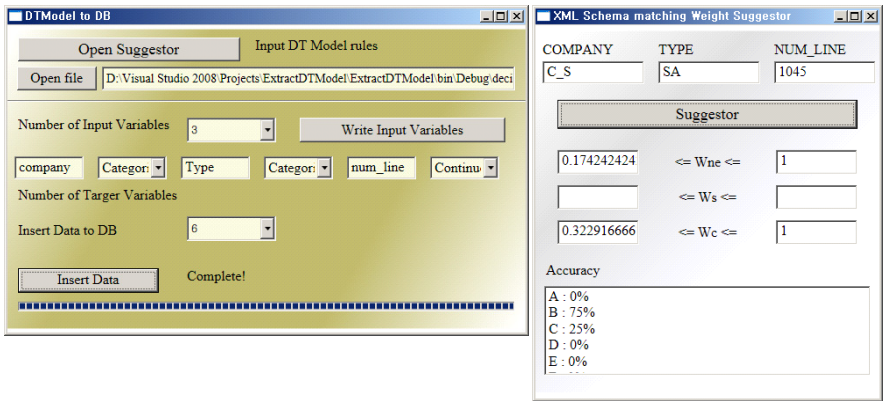

그림 3 의사결정나무 모델 DB를 활용한 자동 스키마 매칭 가중치 선정 모듈

3. 결론

본 연구에서는 데이터의 통합, 비교 등에 활용되는 XML 응용 스키마 매칭 기법을 엔지니어링 문서에 효과적으로 적용하기 위한 방법을 제시하고 시범적 모듈을 구현하였다. XML 응용 스키마 매칭 기법은 구성요소간의 매칭 가중치에 따라 다양한 매칭 정확도의 결과를 나타내는데, 본 연구에서는 20종 구조계산서 580개의 XML 스키마 매칭 결과 데이터를 바탕으로 의사결정나무 모델을 수립하여 교량의 형식, 문서가 포함하고 있는 문장요소의 수, 문서를 작성한 회사에 따라 최적의 매칭 가중치를 효율적으로 선정할 수 있는 방법을 제시하였다. 의사결정지원나무를 활용하여 선정한 매칭 정확도 약 95%정도를 나타내고 있어 특별히 가중치를 선정하지 않은 상태의 평균 매칭 정확도 약 86%보다 약 10% 상승한 매칭 정확도 값을 얻을 수 있음을 알 수 있었다.

감사의 글

본 연구는 2009년도 중소기업청의 기술혁신개발사업(과제번호: S1061715)과 교육인적자원부 BK21사업의 일환인 연세대학교 사회환경시스템공학부 미래사회기반시설 산학연공동사업단의 연구비 지원에 의해 수행됨.

참고문헌

박상일, 김봉근, 김경환, 이상호 (2009) **엔지니어링 문서의 문장 자동 계층정의 방법론**, 한국전산구조공학회 논문집, 22(4), pp. 323-330.

한국시설안전기술공단 (2004) **설계도서 등의 사본작성 및 관리지침**, 한국시설안전기술공단.

Lee, S.-H., Kim, B.-G., Kim, D.-H. and Jeong, Y.-S. (2006) Development of standardized semantic model for structural calculation documents of bridges and XML schema matching technique, Proceedings of the 3rd IABMAS, pp. 633-634.

Liu, S., McMahon, C.A., Darlington, M.J., Culley, S.J., and Wild, P.J. (2006) A computational framework for retrieval of document fragments based on decomposition schemes in engineering information management, Advanced Engineering Informatics, 20, pp. 401-413.

Rahm, E. and Bernstein, P.A. (2001) A survey of approaches to automatic schema matching, The VLDB Journal, 10, pp. 334-350.

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