빈발패턴을 이용한 스키마 매핑

채덕진* · 반경진** · 김응곤***

Schema Mapping Method using Frequent Pattern Mining

Duck Jin Chai* · Kyeong-jin Ban**, Eung-Kon Kim***

요 약

현재, 두 스키마 속성 사이의 효율적인 스키마 매핑 방법을 통해 메타데이터간의 상호운용성을 확보하기 위한 많은 연구가 진행되고 있다. 그러나 기존의 스키마 매핑 연구의 대부분은 몇몇 유사도 값들을 단순히 계산하여 매핑을 선택하기 때문에 정확률이 비교적 낮아 문서변환이나 시스템 통합을 위한 스키마 매핑에 적합하지 않다. 본 논문에서는 데이터 마이닝의 빈발패턴탐사 방법을 이용하여 대화식으로 스키마 매핑을 수행할 수 있는 알고리즘을 제안한다. 메타데이터 표준을 이루는 각 스키마 요소에 정의된 의미 부분을 이용하기 때문에 그 스키마 요소가 어떤 정보를 나타내는지를 알 수 있기 때문에 대화식으로 더 정교한 매핑 처리가 가능하게 된다. 제안하는 방법의 성능 평가를 위해 메타데이터 표준들을 이용하여 정확도에 대한 성능평가를 수행한다.

ABSTRACT

Currently lots of studies to solve meta-data interoperability in between schema attributes are conducted. But the accuracy in previous schema mapping studies is low since the studies just use the similarity in between attributes. So the studies are not suitable for the schema mapping such as document conversion, system integration, etc. In this paper, we propose a method which can conduct the schema mapping interactively using frequent pattern mining. The method can conduct more accurate mapping process because the method use the description element which is an element among each schema element for the metadata standard. A performance study has been conducted to compare the accuracy performance of the method using metadata standards.

키워드

Frequent Pattern Mining, Schema Mapping, Association Rule Mining

Ⅰ. 서 론

현재 다양한 응용 영역에서 메타데이터 표준을 이용한 효율적인 데이터 관리가 이루어지고 있다. 메타데이터 표준은 IPTV, VOD 등과 같은 동영상 서비스를 위한 메타데이터 표준에서부터 MARC, DC, ONIX 등과 같이 도서관 분야에서 사용되는 메타데이터 표준까지 여러 분야에서 다양하게 사용되고 있다. 이러한 메타데이터는 같은 종류의 서비스와 데이터에도 다양한 메타데이터 포맷들이 존재하고 서로 상이한 환경과 요구에 따라 개발되기 때문에 메타데이터간의 상호운용성을 저해하는 요소로 작용한다. 이러한 환경에서 메타데이터를 사용할 경우 데이터 요소와 표현이 불일치하므로 동일한 내용의 자원이 상이한 형식

* 전남대학교 데이터베이스연구실(djchai@nate.com)
** 순천대학교 컴퓨터과학과(multiwave@sunchon.ac.kr)
*** 교신저자 : 순천대학교 컴퓨터과학과(kek@sunchon.ac.kr)
접수일자 : 2010. 1. 28
심사완료일자 : 2010. 2. 18

으로 표현되며 다른 형식의 메타데이터에는 접근할 수 없다. 또한, 메타데이터를 구성하는 속성(element or attribute)들은 속성의 이름은 다르지만 같은 정보를 저장하게 된다.

메타데이터 상호운용성을 확보하기 위한 대표적인 연구로써, 스키마 속성들 사이의 매핑과 관련된 연구는 스키마 매핑(schema mapping)과 메타데이터 레지스트리(metadata registry)가 존재한다. 스키마 매핑은 두 스키마의 속성들 사이에 의미적(semantic) 유사성을 찾는 것으로 전자상거래, 데이터웨어하우스, 데이터 통합 등 여러 응용 분야에서 필요로 하고 있다[1, 2]. 메타데이터 레지스트리는 서로 다른 데이터베이스가 같은 개념에 대해 서로 다른 식별자나 서로 다른 단어를 사용할 경우 공유되는 개념들을 정형적, 명시적으로 명세화한 온톨로지가 집합되어 있는 메타데이터 관리시스템이라 할 수 있다. 그러나 효과적인 매핑 기법을 개발하는 것은 매우 어려운 문제로서 실제적으로 대부분의 응용 분야에서는 스키마 전문가에 의한 수작업으로 매핑이 이루어지고 있다[3, 4].

스키마 매핑이 어려운 이유는 두 스키마 사이의 구문적 차이, 의미적 차이, 계산의 복잡성 등과 같은 이유로 볼 수 있다[5]. 기존의 스키마 매핑 연구의 대부분은 정확률이 비교적 낮아 문서변환이나 시스템 통합을 위한 스키마 매핑에 적합하지 않다. 예를 들어, 기존의 스키마 매핑 연구는 몇몇 유사도 값들을 단순히 계산하여 매핑을 선택하기 때문에 잘못된 매핑을 계산할 수 있다. 이러한 어려움들을 극복하고 효율성과 정확성을 유지하도록 스키마 매핑의 품질을 높이는 것은 여전히 연구가 필요하다[6].

본 논문에서는 메타데이터 스키마 설계할 때 정의한 각 속성의 구성요소(예를 들어, 속성 이름, 속성의 데이터형, 속성값 크기, 속성의 계층적인 경로, 속성에 대한 의미 등)중에서 비 구조적 데이터인 속성에 대한 의미에 대해서 불용어들을 제거한 후, 연관 규칙 마이닝(association rule mining)의 빈발패턴탐사(frequent pattern mining)를 이용하여 대화식(interactive) 방법으로 정확도가 높은 매핑 처리를 할 수 있는 알고리즘을 제안한다.

속성에 대한 의미 정보는 속성 이름, 속성의 데이터형, 속성값의 크기, 그리고 속성의 계층적인 경로 등과 같은 정형화된 구조적 데이터가 아닌 그 속성이 표현하기 위한 데이터가 무엇인지를 정의하고 있는 데이터이다. 예를 들어, 방송과 관련된 메타데이터로 TV-Anytime과 DR Metadata 가 있다. 이 두 메타데이터에서는 방송 콘텐츠에 등장하는 사람의 이름을 표현하기 위해 각각 character와 name 이라는 이름의 속성을 사용한다. TV-Anytime의 character 속성은 “Specifies the name of a character played by an actor"로 정의되어 있고 DR Metadata 표준의 name 속성은 "The Person entity stores basic person information such as name and character"로 정의되어 있다. 이와 같이 비구조적인 정보를 이용하면 직접 어떤 속성이 어떤 정보를 포함하는지를 명확히 알 수 있기 때문에 단순히 속성의 구조적인 정보에 대한 유사도 계산보다 대화식 방법으로 계속적인 학습(training)을 통해 더 정확한 매핑을 수행할 수 있다. 또한, 사람의 이름과 동물의 이름을 저장하기 위한 두 개의 속성이 각각 name 이라는 같은 속성 이름을 사용할 때, 단순한 유사도 계산에 의해서 서로 다른 정보를 저장해야하는 속성임에도 같은 정보를 저장하기 위한 속성으로 인식될 수 있다. 이와 같은 오류도 각 속성에 정의된 의미 정보를 이용하면 정확률을 올릴 수 있다. 본 논문에서 제안하는 대화식 방법은 기존의 유사도 기반의 스키마 매핑에 보완적으로 사용할 수 있기 때문에 기존의 스키마 매핑보다 보다 정교한 수준의 매핑이 가능해진다. 또한, 보다 정교한 스키마 매핑을 통해서 얻어진 결과를 이용하여 온톨로지를 자동적으로 생성하고 갱신할 수 있게 된다.

본 논문의 구성은 다음과 같다. 2절에서는 관련연구로써, 연관규칙 마이닝을 위한 빈발패턴탐사와 메타데이터 상호운용성을 위한 방법인 스키마 매핑과 메타데이터 레지스트리에 대해서 살펴본다. 3절에서는 본 논문에서 제안하는 대화식 방법의 스키마 매핑 알고리즘에 대해서 기술하고 스키마 매핑 알고리즘에 의해 계산된 결과를 이용하여 온톨로지를 자동적으로 구축할 수 있는 방법에 대해서 기술한다. 4절에서는 제안된 방법을 실험을 통해 평가하고 마지막으로 5절에서는 결론 및 향후연구 방향에 대해서 기술한다.

II. 관련연구

본 절에서는 스키마 속성에 대한 비구조적 데이터인 속성의 설명을 이용하여 스키마 매핑을 수행하기 위해 사용되는 빈발패턴탐사에 대해 기술하고 기존의 스키마 매핑 연구에 대해서 소개한다.

2.1. 연관규칙 마이닝과 빈발패턴탐사

연관규칙은 1993년에 처음 소개된 것으로 $X \rightarrow Y$의 형태를 갖는 패턴이다. 이때 X와 Y는 항목(item)의 집합이다. 이 $X \rightarrow Y$ 형태의 연관규칙이 갖는 의미는 X 항목집합이 나타날 때는 Y 항목집합도 동반하여 나타나는 경향이 있다는 뜻이다[7, 8].

연관규칙이란 특정 사건의 발생이 다른 사건의 발생을 암시하는 경향을 표현하는 규칙으로 다음과 같이 정의할 수 있다. $I=\{a_1, a_2,..., a_n\}$를 데이터 항목들의 집합이라고 하고, 트랜잭션들로 구성된 데이터베이스를 $<T_1, T_2,..., T_m>$라고 하자. T_i를 트랜잭션 T_i가 접근하는 데이터 항목들의 집합을 표현한다고 하면 $T_i \subseteq I$ 이다. 이때, 트랜잭션은 한 명의 고객이 구입한 품목들의 집합이라고 정의할 수 있다. $X(X \subseteq T_i)$와 $Y(Y \subseteq T_i)$의 항목집합에 대한 연관규칙은 $X \rightarrow Y$로 표현되며, X를 규칙의 조건부(antecedent), Y를 결과부(consequent)라고 하고 $X \cap Y = \varnothing$이다.

연관규칙에는 지지도(support)와 신뢰도(confidence) 라는 두 가지 중요한 척도가 있다. 예를 들어, “빵을 구매하는 고객의 40%는 우유도 구매하며, 전체 트랜잭션의 2%는 빵과 우유를 포함하고 있다.” 여기서 40%는 이 규칙의 신뢰도라고 하고, 2%는 이 규칙의 지지도라고 한다. 주어진 데이터베이스에서 탐사되는 연관규칙이 사용자가 정의한 최소지지도와 최소신뢰도 이상의 값들을 가져야 하므로, 연관규칙을 탐사하는 문제는 기본적으로 다음의 두 단계로 구성된다[8, 9, 10].

- 단계 1 : 빈발항목집합들(large itemsets)을 찾아낸다. 미리 결정된 최소지지도(minimum support) S_{min} 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들을 빈발항목집합이라 한다.

- 단계 2 : 데이터베이스로부터 연관규칙을 생성하기 위하여 빈발항목집합을 사용한다. 모든 빈발항목집합 L에 대해서 L의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 그러한 부분집합 A에 대하여, A의 지지도 sup(A)에 대한 sup(L)의 비율이 최소신뢰도 C_{min} 이상이면 (sup(L)/sup(A)$\geq C_{min}$), $A \rightarrow (L-A)$ 형태의 규칙을 생성할 수 있다. 이 규칙의 지지도는 sup(L)이고 신뢰도는 sup(L)/sup(A)이다.

연관 규칙 탐사에서 위의 두 단계를 통해 원하는 최종 규칙들을 찾아낼 수 있다. 연관 규칙 탐사에 대한 연구는 주로 첫 번째 단계에서의 빈발 항목집합들을 데이터베이스로부터 빠르게 추출하는 방법에 대해 이루어지고 있다. 두 번째 단계는 첫 번째 단계에서 발견된 빈발 항목집합들이 최소신뢰도를 만족하는지를 검사하기 때문에 한 번의 데이터베이스 스캔으로 쉽게 문제를 해결할 수 있다.

2.2 메타데이터 매핑과 레지스트리

메타데이터 상호운용성은 두 개 이상의 시스템이 정보를 교환하거나 서로 다른 형식의 메타데이터 스키마를 사용할 경우 데이터 요소와 표현의 불일치를 해결하여 상이한 어휘로 정의된 요소들의 의미를 이해할 수 있는 기능을 제공하는 것을 말한다[11]. 현재 메타데이터의 상호운용성을 해결하기 위한 많은 연구

표 1. 메타데이터 속성에 대한 의미 정보의 예
Table 1. An example of semantic description attribute for the metadata standard

표준(standard)	속성(attribute)	의미(semantic)
TV-Anytime	language	Describes one spoken language for the program. There may be more than one spoken language specified for a program
TV-Anytime	captionLanguage	Describes on language of the caption information included with the program
DR Metadata	language	A list of valid codes and names for language, which describes the primary spoken language of the program

가 진행되고 있다. 상호운용성을 해결하기 위한 대표적인 방법은 크게 세 가지로 나눌 수 있다. 첫째, 대표적인 메타데이터 표준을 정의하고 그 표준에 맞추어 모든 스키마를 통합하는 것과 둘째, 메타데이터의 요소와 표현을 이해하고 스키마 매핑을 이용하는 방법[1, 2, 3, 4], 그리고 셋째, 메타데이터 레지스트리를 이용하는 것이다[12]. 그러나 하나의 메타데이터 표준을 제정하여 모든 데이터들을 수용하는 것은 현실적으로 어려운 문제이기 때문에 스키마 매핑을 이용하는 방법과 메타데이터 레지스트리를 이용하는 방법이 주로 쓰인다.

스키마 매핑은 소스 스키마와 타겟 스키마를 입력으로 받아 의미적인 연관 관계를 계산하고 매칭을 선택하는 것이다. 스키마 매핑에 관한 연구는 상호참조 테이블 방식과 범용 메타데이터 통합 방식이 있다[3, 11]. 상호참조 테이블 방식에는 USMARC와 DC, TEI, EAD의 매핑테이블, DC와 다른 메타데이터 형식과의 참조 테이블, 공통 메타데이터 기술 집합과의 매핑 등이 있고 범용 메타데이터 통합 방식으로는 W3C에서 개발한 개념적 차원의 통합구조인 RDF가 대표적이다[11, 13].

이러한 방법들은 크게 보면 스키마 매핑과 온톨로지를 이용한 매핑과 같이 매핑을 이용하여 상호운용성을 확보하기 위한 방법들이다. 그러나 지금까지 스키마 매핑 연구가 단순한 유사도에 의한 매핑 처리로 인해 정확률이 낮다는 문제점을 가지고 있기 때문에 정확한 매핑을 계산하기 위해서 의미적 및 구조적으로 스키마 매핑 관계를 계산하여야 한다.

III. 빈발패턴탐사를 이용한 스키마 매핑 방법

이 절에서는 빈발패턴탐사를 위한 트랜잭션 데이터를 구성하기 위해서 메타데이터 스키마의 각 속성에 대한 의미 정보를 이용하는 방법과 구성된 트랜잭션 데이터를 이용하는 빈발패턴탐사과정에 대해서 기술한다.

3.1 트랜잭션 데이터베이스 구축

메타데이터는 데이터를 설명하기 위한 구조화된 데이터를 의미한다. 메타데이터가 가지고 있는 정보는 크게 의미 정보, 구조 정보, 표현 정보로 분류된다. 의미 정보는 메타데이터가 가지고 있는 데이터 요소의 의미 정보를 나타내고 구조 정보는 각 데이터 요소간의 연관성을 나타내는 구조 정보를 나타낸다. 마지막으로 표현 정보는 값 영역, 데이터 형, 측정 단위 등과 같은 정보를 나타낸다.

일반적으로 스키마 매핑은 주로 구조 정보와 표현 정보가 이용된다. 이 두 가지 데이터에 대해서 각 스키마 요소간의 매핑 유사도를 계산하여 스키마 매핑을 수행한다. 본 논문에서는 기존의 유사도 계산에 의한 매핑의 정확률을 높이기 위해서 메타데이터의 의미 정보를 이용한다. 제안하는 기법은 먼저 메타데이터 속성에 대한 의미 정보를 구문분석, 어근분석, 불용어 제거 등의 순서로 전처리를 하고 빈발패턴탐사를 수행하기 위한 트랜잭션을 구성하기위해 의미 있는 키워드들의 집합을 생성한다. 각 속성의 의미 부분에서 키워드들의 집합은 트랜잭션에서 항목의 집합으로 생각할 수 있다. 트랜잭션 데이터의 생성 과정을 설명하기 위해서 표 1과 같이 TV-Anytime[14]의 language, captionLanguage 속성에 대한 의미 정보와 DR Metadata[15]의 language 속성에 대한 의미 정보를 이용한다.

TV-Anytime의 language 속성은 콘텐츠에서 이용되는 언어를 표현하기 위한 속성이고 captionLanguage는 language 속성의 하위 속성으로써 콘텐츠에서 이용되는 자막 언어를 표현하기 위한 속성이다. DR Metadata의 language 속성은 TV-Anytime의 language 속성과 같은 정보를 표현하기 위한 속성이다. 표 1의 의미 정보들로부터 구문분석, 어근분석, 불용어 제거 등과 같은 전처리를 통해 알고리즘 수행에 필요한 의미 있는 키워드들의 집합을 추출할 수 있다.

첫 번째 단계인 구문 분석과 불용어 제거가 끝나면 트랜잭션 데이터베이스가 생성되고 이 데이터베이스를 이용하여 빈발패턴탐사를 수행한다. 표 2는 전처리를 통해 추출된 의미 있는 키워드들의 집합으로 빈발패턴탐사를 위한 트랜잭션 데이터베이스이다.

3.2 전처리된 트랜잭션 데이터베이스에 대한 빈발 패턴 탐사

표 2. 전처리 후에 트랜잭션 데이터베이스의 예
Table 2. An example of transaction database after pre-processing

표준(standard)	속성(attribute)	의미(semantic)
TV-Anytime	language	describe, speak, language, program, specify
TV-Anytime	captionLanguage	describe, language, caption, information, include, program
DR Metadata	language	list, valid, code, name, describe, primary, speak, language, program

본 절에서는 전처리 후에 생성된 트랜잭션 데이터베이스에 대해서 빈발항목집합을 추출하는 과정을 기술한다. 현재 매우 다양한 빈발항목집합 탐사 알고리즘이 제안되어 있지만 본 절에서는 비교적 이해하기에 효율적인 Apriori 알고리즘[8]을 적용한다.

트랜잭션 데이터베이스를 처음 스캔하여 1-빈발항목을 추출할 수 있다. 그림 1은 표 2에 정의된 키워드 들을 하나의 트랜잭션으로 보고 3개의 트랜잭션에 대해서 빈발항목집합을 추출한 예이다. 빈발항목집합은 사용자가 정의한 최소지지도를 만족하는 항목집합을 가리킨다. 그림 2에서 최소지지도를 50%로 정의했을 때, 후보항목들(candidate items)로부터 {describe, speak, language, program}과 같이 4개의 항목이 1-빈발항목으로 추출된다. 1-빈발항목들은 다시 2-빈발항목집합들을 추출하기 위해 후보항목집합으로 생성되고 후보항목집합들의 지지도를 계산하여 최소지지도를 만족하는 2-빈발항목집합들을 추출하게 된다. 그림 1에서 보이듯이 {(describe, speak), (describe, language), (describe, program), (speak, language), (speak, program), (language, program)}과 같이 6개의 2-빈발항목집합이 추출된다. 이와 같이, 빈발패턴 탐사 알고리즘을 수행하면 최종적으로 4-빈발항목집합인 {(describe, speak, language, program)}이 추출되고 알고리즘은 종료된다.

3.3 대화식 스키마 매핑

대화식이 의미하는 것은 스키마 속성에 정의된 의미 정보를 이용하여 스키마 매칭을 수행한다는 것이다. 즉, 표 1에서 볼 수 있듯이, 스키마 속성에 정의된 의미 정보를 보고 우리는 그 속성이 어떤 정보를 표현하는 속성인지를 알 수 있다. 따라서 두 스키마 속성들의 의미 정보를 비교하면 두 스키마 속성의 유사도를 평가할 수 있다. 또한, 같은 이름의 속성이 서로 다른 정보를 표현해야 하는 경우에 발생할 수 있는 오류로부터 벗어날 수 있다. 앞 절의 예에서 빈발항목집합 탐사과정을 통해 4-빈발항목집합인 {(describe, speak, language, program)}이 추출되었다. 4-빈발항목집합이 의미하는 것은 두 스키마 속성이 모두 4개의 키워드들을 포함하고 있다는 것이다. 표 2에서 볼 수 있듯이, 일반적으로 서로 상이한 환경과 요구에 따라 개발된 메타데이터 스키마라 할지라도 같은 정보를 표현 하는 스키마 속성들을 정의할 때, 비슷한 의미를 가진 키워드들의 집합이나 동일한 키워드들의 집합으로 해당하는 속성의 의미가 정의된다. 그러므로 두 스키마 속성에 대한 의미 정보에서 빈발항목집합을 추출하게 되면 그 빈발항목집합에 포함된 항목들의 개수에 의해서 같은 의미의 정도를 평가할 수 있다.

$$\text{의미유사도}(A_1, ..., A_i) = \frac{\frac{|f|}{|t_1|} + \frac{|f|}{|t_2|} + \cdots + \frac{|f|}{|t_i|}}{|T|} \cdots\cdots (1)$$

본 논문에서는 이러한 의미의 정도를 의미유사도(semantic similarity)라 정의한다. 의미유사도는 각 속성들의 매핑관계를 평가하기 위해서 식 (1)과 같이 전체 트랜잭션에서, 각 트랜잭션에 포함된 항목들 중에 빈발항목이 각각 얼마나 포함되어 있는지로 정의한다. 여기서 $|f|$는 각 트랜잭션에서 비교되는 속성들에 공통으로 포함된 빈발항목의 개수를 의미하고 $|t_i|$는 각 트랜잭션에 포함된 항목의 개수를 의미한다. $|T|$는 전체 트랜잭션의 수를 의미한다.

그림 1에서, 4-빈발항목집합 {(describe, speak, language, program)} 을 모두 포함하고 있는 속성은 TV-Anytime의 language 속성과 DR Metadata 의

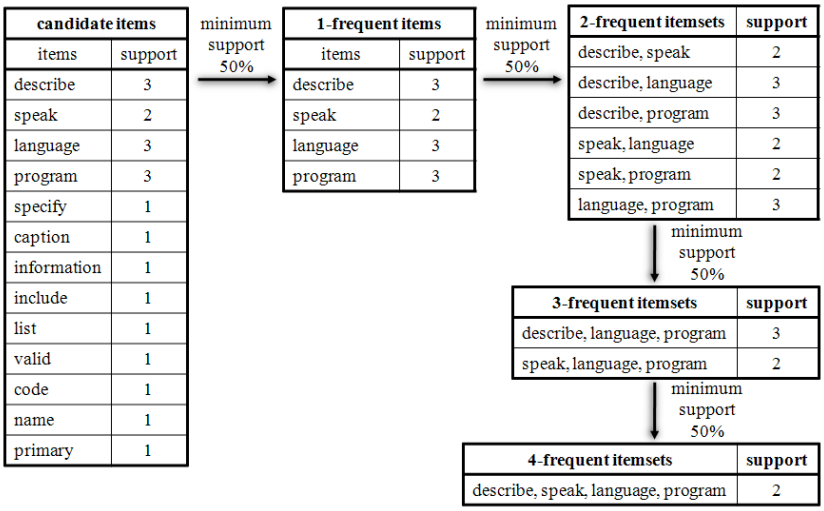

그림 1. 빈발항목집합 탐사의 예
Fig. 1 An example of frequent pattern mining

language 속성이다. 이 두 속성은 실제로 콘텐츠에 사용된 언어가 무엇인지를 표현하기 위한 속성들로써, 같은 정보를 표현하는 속성이므로 다른 속성과 비교해서 가장 많은 빈발항목집합을 포함하고 있어두 스키마 속성은 의미적으로 유사한 속성들이라고 유추할 수 있다. 식 (1)을 이용하여 두 스키마 속성의 의미유사도는 두 속성이 모두 4개의 빈발항목을 포함하고 있기 때문에 {(4/5)+(4/9)}/2 = 0.62를 구할 수 있다. 또한 TV-Anytime의 captionLanguage는 language 속성의 하위 속성이다. 그러므로 DR Metadata의 language 속성과도 유사성을 가지고 있다고 할 수 있다. 두 속성간의 의미유사도는, 두 속성이 공통적으로 포함하고 있는 빈발항목의 수는 3이기 때문에 의미유사도는 {(3/6)+(3/9)}/2 = 0.42가 된다. 따라서 language 속성간의 의미유사도보다 낮기 때문에 language 속성간의 관계보다는 좀 더 낮은 유사한 속성이라고 할 수 있다. 물론, 의미유사도는 여러 가지 환경과 데이터에 따라 변하기 때문에 과거에 축적된 학습 결과와 응용 분야에 맞게 설정하는 것이 타당하다.

본 논문에서 제안하는 대화식 스키마 매핑 방법은 기존의 데이터 타입과 경로 유사도에 의한 스키마 매핑 방법을 병행하여 사용함으로써 좀 더 정교한 스키마 매핑을 수행할 수 있다.

3.4 의미유사도에 의한 관계 설정 및 온톨로지 작성

온톨로지를 구축하기 위해서는 각 개념들간의 관계를 설정해야한다. 일반적인 온톨로지에서 각 개념들간의 관계는 Is-A, Similar, 그리고 PartOf 관계로 나눌 수 있다.

그림 2는 위의 예에서 보였던 빈발항목집합 탐사와 의미유사도를 이용하여 온톨로지를 구축하는 과정을 보여준다. 그림 2에서 TV-Anytime 표준에는 language 속성과 language 속성의 하위 속성으로 captionLanguage이 존재한다. DR Metadata 속성에는 language 속성이 존재한다. 빈발항목집합 탐사와 의미유사도의 결과에 의해서 TV-Anytime 표준의 language 속성과 DR Metadata 표준의 language 속성은 Is-A 관계로, TV-Anytime 표준의 captionLanguage 속성과 DR Metadata 표준의 language 속성은 Similar 관계로 온톨로지를 구축할 수 있다. 여기서 Is-A 관계는 동일한 의미를 가지고 있는 속성들이라는 것을 나타내고 Similar 관계는 동일한 정보를 표현하는 속성은 아니지만 하위 속성이나 비슷한 의미의 속성을 나타낸다.

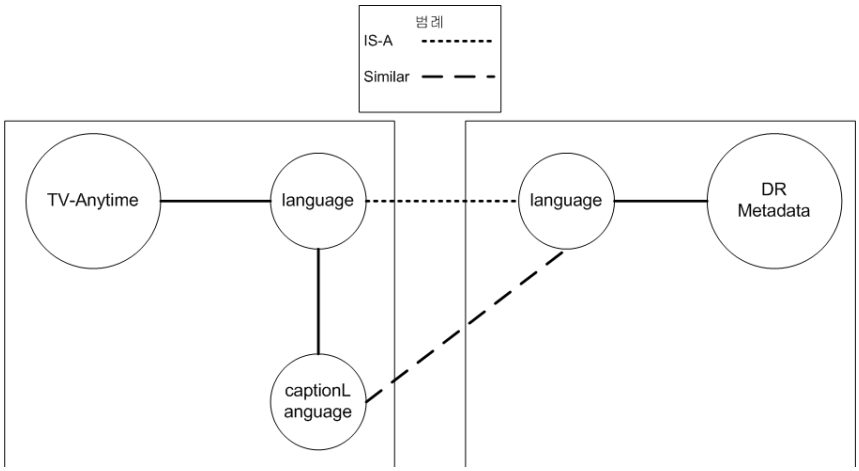

그림 2. 각 속성에 대한 온톨로지 구축의 예
Fig. 2 An example of ontology generation

Ⅳ. 성능평가

제안하는 빈발 패턴을 이용한 대화식 스키마 매핑 기법의 성능을 평가하기 위해서 메타데이터 표준인 RDF, TV-Anytime와 실제 방송국에서 사용하고 있는 DR Metadata 의 스키마 속성들을 대상으로 하였다. 실험을 위해 같은 분야에 대한 메타데이터 표준들을 대상으로 하였다.

빈발 패턴 탐사와 같은 데이터 마이닝에서는 많은 양의 데이터가 필요하다. 또한, 많은 양의 데이터에서 얻어지는 결과가 높은 신뢰성을 가질 수 있다. 그러므로 실험 평가를 위해서 실제적으로 많은 양의 메타데이터 표준들을 이용해서 각 표준의 속성들이 얼마나 높은 정확도로 매치가 되는지를 확인해야 한다. 그러나 빈발 패턴 탐사를 수행할 만한 데이터를 구하기 어렵기 때문에 RDF, TV-Anytime, 그리고 DR Metadata 에 존재하는 각 속성들 중에 몇 개의 단어가 서로 매치가 되는지 확인해 보았다. 이에 대한 실험 절차는 다음과 같다.

1) 세 개의 표준에서 대표 속성 50개를 추출한다.
2) 각 속성들에 대한 설명 속성들에서 불용어들을 제거하여 실험 데이터를 추출한다.
3) 추출한 50개 속성들에 대한 실험 데이터에서 유사한 의미를 가지고 있는 속성들간 몇 개의 단어들이 매치가 되는지를 계산한다.

표 3은 세 개의 표준에서 추출한 50개의 속성들에서 불용어를 제외한 나머지 단어들이 각 속성 간 얼마나 매치되는지를 나타내는 평균값이다. 각 속성에서 불용어들을 제외한 단어의 개수는 평균 10개를 형성하였다. 표의 수치는 그 중에서 각 속성간 매치되는 단어의 개수를 나타낸다.

TV-Anytime과 DR Metadata 표준은 디지털 방송을 위한 메타데이터이므로 RDF 보다는 좀 더 높은 매치율을 나타낸다.

표 3. 각 속성간 매치되는 단어의 평균 개수
Table 3. The number of average matching for each attribute

메타데이터 표준	RDF	TV-Anytime	DR Metadata
RDF	-	5	6
TV-Anytime	5	-	7
DR Metadata	6	7	-

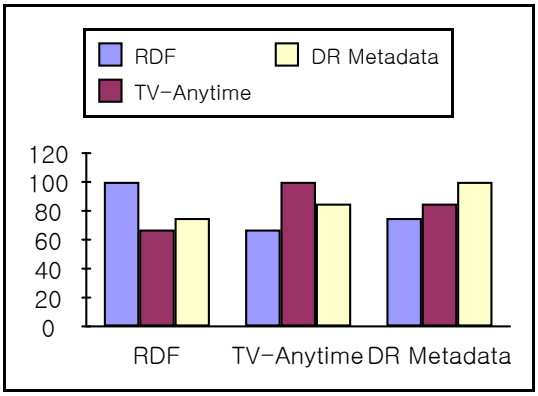

그림 3. 메타데이터 속성간 매치율
Fig. 3 Match ratio between each attribute

그림 3은 최소지지도를 50%로 정의했을때, 각 메타데이터 사이에 매치되는 정확률을 측정한 것이다. 각 속성에 존재하는 단어의 개수가 평균 10개이기 때문에 최소지지도 50%의 의미는 각 속성 사이에 단어 매치가 5개 이상이면 그 속성들은 유사하다고 할 수 있다는 것을 의미한다.

실험 결과는 좀 더 많은 데이터를 이용하여 빈발패턴 탐사를 통해 좀 더 높은 정확률을 얻을 수 있다. 만약 기존의 메타데이터 속성 매치 방법과 병행하여 사용한다면 훨씬 효율적으로 사용될 것이다.

V. 결 론

본 논문에서는 메타데이터 스키마 설계할 때 정의한 각 속성의 구성요소 중에서 비 구조적 데이터인 속성에 대한 의미에 대해서 불용어들을 제거한 후, 빈발패턴탐사를 이용하여 대화식으로 스키마 매핑 처리를 할 수 있는 기법을 제안하였다.

본 논문에서 제안하는 대화식 방법은 기존의 유사도 기반의 스키마 매핑에 보완적으로 사용할 수 있고 데이터마이닝의 특징인 데이터의 양이 증가하면 할수록 더욱 정확도가 높은 결과를 얻을 수 있기 때문에 점진적으로 데이터의 양이 증가하면 할수록 스키마 매핑에 대한 높은 정확률을 기대할 수 있다. 또한, 보다 정교한 스키마 매핑을 통해서 얻어진 결과를 이용하여 온톨로지를 자동적으로 생성하고 갱신할 수 있다.

앞으로 온톨로지를 자동적으로 구축할 수 있는 모델과 좀 더 정확률을 높일 수 있는 기법에 대한 연구가 필요하다. 또한, 각 메타데이터 속성과 불용어들을 제거한 후의 속성 데이터를 데이터베이스에 저장하고 이를 이용하여 빈발패턴탐사를 수행하고 그 결과를 점진적으로 이용하기 위한 데이터베이스 시스템 모델의 개발도 수행되어져야 한다.

감사의 글

본 지식재산권은 지식경제부 및 정보통신산업진흥원의 지원을 받아 수행된 연구결과임
(09-기반. 산업원천기술개발사업)

참고 문헌

[1] F. Giunchiglia and P. Shvaiko, "Semantic Matching," In The Knowledge Engineering Review Journal, Vol.18, No.3, 2004.

[2] J. Madhavan, P. Bernstein, and E. Rahm, "Generic Schema Matching with Cupid," In Proceedings of VLDB, 2001.

[3] E. Rahm and P. Bernstein, "On Matching Schemas Automatically," VLDB Journal, Vol.10, No.4, 2001.

[4] S. Sun and E. Rose, "Automated Schema Matching Techniques: An Exploratory Study," Res, Lett. Inf. Math. Sci., pp.113-136, 2004.

[5] P. Shvaiko and J. Euzenat, "A Survey of Schema-based Matching Approaches," Technical Report DIT-04-087, University of Trento, Italy, 2004.

[6] H. Do, S. Melnik, and E. Rahm, "Comparison of Schema Matching Evaluations," In Proceedings of the 2nd Int. Workshop on Web Databases, 2002.

[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of ACM SIGMOD Conference on Management of Data, Washington D.C., pp. 207-216, May 1993.

[8] R. Agrawal, R. Srikant, "Fast Algorithms for

Mining Association Rules", In Proceedings of the 20th VLDB Conference, Santiago, Chile, Sept., 1994.
[9] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000.
[10] 박종수, 유원경, 홍기형, "연관 규칙 탐사와 그 응용", 한국정보과학회논문지 제16권 제 9호, pp. 37-44, 1998.
[11] 백두권, "정보통신 및 표준화 기술 동향: 데이터 표준화와 메타데이터 레지스트리", TTA 저널 통권 71호, 2000.
[12] C. Blanchi, J. Petrone, "Distributed Interoperable Metadata Registry," D-Lib Magazine, December 2001.
[13] http://www.w3.org/RDF/
[14] http://www.tv-anytime.org/
[15] http://www.dr.dk/metadata

저자 소개

채덕진(Duck Jin Chai)

1999년 2월 동신대학교 컴퓨터학과 졸업 (이학사)
2001년 2월 전남대학교 대학원 전산통계학과 졸업(이학석사)
2006년 2월 전남대학교 대학원 전산학과 졸업(이학박사)
※ 관심분야 : 데이터마이닝, 스트림 데이터 마이닝

반경진(Kyeong-jin Ban)

2003년 2월 : 순천대학교 컴퓨터과학과 (이학사)
2005년 2월 : 순천대학교 컴퓨터과학과 (이학석사)
2007년 8월 : 순천대학교 컴퓨터과학과 박사수료
※ 관심분야 : 컴퓨터 그래픽스, RFID, USN

김응곤(Eung-Kon Kim)

1980년 2월 : 조선대학교 전자공학과 (공학사)
1986년 2월 : 한양대학교 컴퓨터공학과 (공학석사)
1992년 2월 : 조선대학교 컴퓨터공학과 (공학박사)
1993년 3월 ~ 현재 : 순천대학교 컴퓨터과학과 교수
※ 관심분야 : 영상처리, 컴퓨터 그래픽스, 멀티미디어, HCI