한국지능정보시스템학회논문지 제11권 제1호 2005년 6월(pp.123-153)

웹 페이지의 내재 규칙 습득 과정에서 규칙식별 역할에 대한 효과 분석

강주영
한국과학기술원 내 국자전자상거래연구센터
책임연구원
(jykang@gsm.kaist.edu)

이재규
한국과학기술원 테크노경영대학원 교수
(jklee@kgsm.kaist.ac.kr)

박상언
한국과학기술원 테크노경영대학원
경영공학 박사과정
(mascon@kgsm.kaist.ac.kr)

오늘날 자원의 보고라 할 수 있는 웹에는 자연어로 표현된 텍스트와 테이블들로 구성된 무수히 많은 문서들이 존재하고 있다. 이러한 웹문서들로부터 규칙을 습득하고 습득된 규칙과 웹문서간의 일관성을 유지하기 위해, 본 논문에서는 확장형 규칙 표식 언어 (eXtensible Rule Markup Language, XRML) 체계를 개발하였다. XRML은 웹페이지에 내재되어 있는 규칙을 식별하여 자동으로 정형화된 규칙을 생성할 수 있도록 지원하는 규칙 식별 표식 언어 (Rule Identification Markup Language, RIML)와 구조화된 규칙 표현을 위한 규칙 구조 표식 언어 (Rule Structure Markup Language)로 구성된다. 특히, RIML은 HTML안에 내재되어 있는 규칙을 HTML 문서에 직접 명시할 수 있도록 설계되었기 때문에, RIML을 통해 웹페이지에 있는 규칙들을 식별하고 이 식별된 규칙은 RSML으로 표현된 정형화된 규칙으로 자동 변환될 수 있다. 본 논문에서는 RIML의 설계 시 웹페이지로부터 규칙을 식별하는 과정에서 발생하는 공유되는 변수 (variables) 및 값 (values), 생략된 어구, 동의어와 같은 몇가지 중요한 현상들을 발견하고 이를 해결하고자 하였다.

제안된 XRML 접근 방법의 성능을 측정하고자, 3개의 대표적인 온라인 서점인 Amazon.com, BarnesandNoble.com, Powells.com의 실제 웹페이지들로부터 배송 및 환불과 관련된 규칙을 습득하여 XRML의 효과를 측정하는 실험을 수행하였다. 실험 결과에 따르면, 웹페이지로부터 규칙은 97.7%의 매우 높은 정확성을 가지고 습득되었으며, 생성된 규칙의 완전성은 88.5%로 측정되어, XRML이 특정 주제에 관한 전문가 시스템을 구축하기 위해 웹페이지로부터 규칙을 추출할 때 효율적인 도구가 될 수 있음이 예시되었다.

논문접수일 : 2005년 3월 게재확정일 : 2005년 6월 교신저자 : 강주영

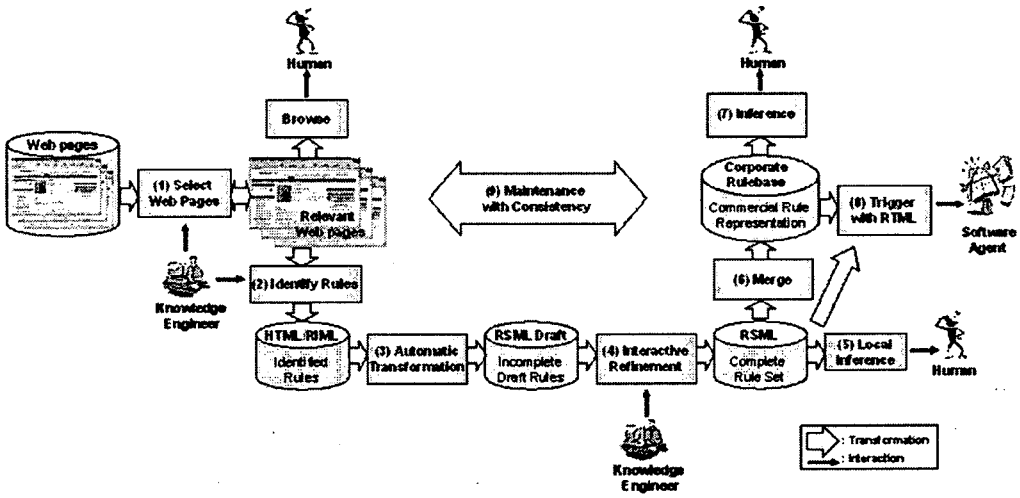
1. 서론

우리의 일상생활을 비롯하여 경제, 문화 등 모든 분야에서 웹의 중요성이 커져감에 따라 웹의 표현력을 증대시키고자 하는 노력도 증가하고 있다. 이러한 의도에 따라 근래까지의 웹 기반 기술은 인간에게 다양한 정보를 보다 효율적으로 보여주기 위한 방향으로 발전해왔다. 반면 차세대 웹으로 거론되고 있는 시맨틱 웹 (Miller et al., 2001)은 웹 페이지에 있는 다양한 데이터와 규칙을 기계가 이해할 수 있는 형태로 표현하고 추출하기 위해 많은 노력을 기울이고 있다. 웹 상에서 데이터를 구조적으로 표현하기 위하여 XML (Bray et al., 2000; van der Alast and Kumar, 2003)이 점차 널리 사용되고 있으며, 규칙을 표현하기 위하여 다양한 규칙표식언어에 대한 연구 (RuleML, 2003)가

진행 중에 있다. 규칙표식언어의 주목적은 추론 가능한 규칙을, XML 기반의 표준언어로 표현하여 웹 상에서 상호호환성을 갖도록 하는데 있다. 그러나 규칙을 XML로 표현하여 메타규칙의 표현을 효과적으로 할 수 있는 공통의 플랫폼을 제공했다고 해서 규칙의 추론능력이 향상되었다고 볼 수는 없다.

규칙표식언어를 활성화하는데 있어 매우 중요한 논점은 자연어와 표를 중심으로 표현되어 있는 웹 페이지로부터 어떻게 보다 효율적으로 규칙을 습득하고 이를 관리할 것인가 하는 것이다. 규칙습득과 관련해서 Hulth (2001)는 자연어로 표현된 문장을 구조화된 규칙으로 바로 전환하는 것은 매우 복잡하기 때문에 지식관리자의 개입이 불가피하다고 지적하였다. 따라서, 본 논문에서는 중간단계를 거쳐 지식을 습득하는 단계적인 방법론으로 XRML(eXtensible Rule Markup Language)을 이용한 방법론을 제안하고자 한다. XRML의 구성요소 중 하나인 규칙 식별 표식 언어(Rule Identification Markup Language, RIML)는 웹 페이지로부터 규칙과 규칙구성요소를 식별하기 위해 설계되었다. 식별된 규칙구성요소들은 RIML의 형태로 표현되고, 이는 규칙 구조 표식 언어 (Rule Structure Markup Language, RSML) 에서 정의하고 있는 추론 가능한 규칙표현 양식으로 자동변환된다. 그러나 자동 생성된 규칙은 초안으로서 아직 완전한 규칙은 아니다. 초안을 보완하여 완성된 RSML 문서는 일반적인 규칙기반시스템에서 사용되는 규칙으로 자동변환이 가능하다. [그림 1]은 이와 같은 규칙습득과정을 보여주고 있다.

전통적인 방법으로 규칙 기반 시스템을 구축하는 경우에는 규칙의 출처와 규칙간에 연관관계 없이 규칙을 추출하였다. 그러나, XRML 환경에서는 먼저 RIML을 사용해서 웹 페이지로부터 규칙을 식별한 후에 RIML을 RSML으로 변환한다. 따라서, RIML은 웹 페이지로부터 규칙을 식별하는 역할을 하는 동시에 RSML과 웹 페이지와의 연결고리 역할을 수행한다. 이와 같이 본 연구의 기여이자 타 연구와의 가장 큰 차이점은 규칙 식별의 결과인 RIML에 있다.



[그림 1] XRML 접근방법에 의한 규칙 습득의 절차

RIML을 이용한 규칙식별을 통해 얻을 수 있는 혜택은 규칙식별의 결과로부터 규칙에 가까운 규칙 초안을 자동으로 생성할 수 있다는 점이다. 또한, 자동 생성된 규칙 초안은 규칙을 수정 및 보완하여 완전한 규칙으로 변환하여야 하는 지식관리자의 부담을 크게 줄일 수 있다. 더 나아가 RIML과 RSML간의 연결성을 이용하여 웹 페이지에 변경이 발생했을 때 이를 쉽게 규칙베이스에 반영할 수 있도록 지식관리자에게 적절한 정보를 제공할 수 있다.

따라서, XRML 접근 방법은 웹 페이지 상의 지식을 기반으로 한 규칙기반 시스템의 구축에 매우 유용하다고 할 수 있다. XRML을 활용하여 지금까지 개발된 응용으로는 워크플로우 (Workflow) 환경에서의 자동화된 전자결재 처리시스템 (이재규 외, 2002), 인터넷 쇼핑몰에서의 약관감사 시스템 (양성병 외, 2003), 서로 다른 조직간 지식 공유 시스템 (김우택 외, 2002) 등이 있다.

XRML 연구의 초기 단계 (Lee and Sohn, 2003)는 규칙과 변수 및 변수값을 표현할 수 있는 태그의 설계에 초점을 두었다. XRML 초기 연구의 다음 단계로서, 본 연구는 자연어 문장과 표로 구성된 웹 페이지로부터 규칙을 습득하여 결합하는데 목표를 두고 있다. 예를 들어 Amazon.com과 같은 온라인 서점에서 물품을 구입할 경우, 자신이 선택한 물품들의 총 배송비용을 계산하기 위해 사용자는 표와 자연어를 통해 기술된 배송관련 규칙들을 적절히 결합하여 사용해야 한다.

본 연구에서는 먼저 규칙식별을 위해 필요한 RIML 태그들을 설계하였다. 그리고, 설계된 RIML 태그들을 이용하여 웹 페이지로부터 규칙을 습득하는 단계를 제안하고 Amazon.com 웹 페이지에서 규칙을 습득하는 예를 이용해 각 단계들을 상세히 설명하였다. 또한, 규칙 습득단계 중 규칙 식별과정에서 발생하는 몇 가지 이슈들을 분석하고 이를 해결하는 방법론을 제안하였다. 여기서 본 연구에 대한 중요한 물음은 "웹 페이지로부터 규칙을 습득하고 유지보수하는 과정에 규칙식별이 어느 정도의 효과를 가지는가"라는 것이다. 이러한 물음에 답하기 위해 Amazon.com, BarnesandNoble.com, Powells.com의 세 온라인 서점으로부터 실제로 배송관련 규칙을 습득하는 실험을 했다. 이 실험을 통해 규칙식별이 웹 페이지로부터의 규칙습득에 미치는 영향을 효과(Effectiveness)와 효율성(Efficiency) 측면으로 나누어 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 규칙표식언어와 지식 획득에 관련된 연구들을 XRML 연구와 비교하여 설명하였다. 3장에서는 웹 페이지에 있는 자연어 문장과 표로부터 규칙을 습득하기 위해 필요한 RIML과 RSML의 표현방법에 대해 정의하고, RIML을 이용해 규칙을 습득하는 과정을 기술하였다. 그리고, Amazon.com으로부터 실제로 규칙을 습득하는 예제를 이용하여 규칙습득과정을 상세히 설명하였다. 4장에서는 RIML을 이용한 규칙식별과정에서의 이슈와 해결방안을, 5장은 RIML의 한계점에 대해 기술하였다. 6장에서는 세 개의 온라인 서점으로부터 규칙을 식별하고 완성하는 실험을 통해 규칙식별의 효과를 보이고자 하였다. 마지막 장에서는 결론과 본 논문의 기여에 대하여 정리하였다.

2. 이론적 배경

XRML의 역할에 대한 이해를 돕기 위하여 본 장에서는 규칙표식언어와 지식 획득에 관련된 연구들을 조사하고, XRML의 목표와 이들 연구의

목표를 비교하여 설명하였다.

2.1 규칙표식언어 및 XRML

규칙표식언어와 관련된 연구활동을 지원하기 위해 만들어진 온라인 컨소시엄인 RuleML (RuleML, 2003) 홈페이지를 보면 다양한 규칙표식언어 관련연구가 활발하게 진행되고 있음을 볼 수 있다. 규칙표식언어의 주 목적은 추론이 가능한 논리적 규칙을 웹 상에서 원활히 교환할 수 있도록 XML을 이용하여 표현하는데 있다. 대표적인 규칙표식언어로 DARPA Agent Markup Language (DAML) - Rules (Grosof, 2002)가 있으며 그 외에 Predictive Model Markup Language (Grossman et al., 1999), Attribute Grammars' semantic rules (Psaila and Crespi-Reghizzi, 1999; Levy and Rousset, 1998), Mathematical Markup Language (Carlisle et al., 2003) 등이 있다. 최근 Semantic Rule Markup Language (SWRL) (Horrocks et al., 2004)이 W3C에 표준안으로 제안된 상태이다. 이러한 규칙표식언어의 공통적인 특징은 XML을 기반으로 만들어진 온톨로지 언어인 RDF(S) (Brickley and Guha, 2000; Lassila and Swick 1999), DAML+OIL (Connolly et al., 2001; Horrocks, 2002), OWL (Horrocks et al., 2003; Smith et al., 2003)과 호환될 수 있도록 표현되었다는 점이다.

XRML도 넓은 의미에서 보면 XML을 기반으로 규칙을 표현한다는 점에서 규칙표식언어의 일종으로 볼 수 있다. 그러나 기존의 규칙표식언어가 상호 이질적인 지식기반 시스템간의 규칙 교환에만 초점을 맞춘 반면, XRML은 규칙 교환뿐만 아니라, 웹 페이지에 내포되어 있는 암묵적인 지식을 습득하여 지식기반 시스템에서 이용될 수 있도록 하며 사람과 기계가 일관성을 유지하면서 추론 기능을 제공해준다는 점에서 여타의 규칙표식언어와는 확연히 구별된다.

2.2 지식 습득 관련 연구

웹 페이지로부터의 규칙 습득은 자연어로부터의 지식 획득과 많은 유사한 점을 갖고 있다. 기존의 지식 획득 방법론을 크게 지식분석도를 이용한 지식 획득 방법론, 지식획득을 위한 자연어 처리, 자동학습으로 나누어 조사하고 본 연구의 접근 방법과 비교하였다.

2.2.1 전문가 다이어그램 접근 방식

오랜 기간 동안 지식의 습득은 전문가 시스템의 구축과정에서 병목으로 자리해왔다. 특히 규칙 전문가와 도메인 전문가 사이의 의사소통이 원활하지 않을 경우에는 더욱 어려운 작업이었는데, 전문가 다이어그램 (Lee et al., 1990)은 도메인 전문가가 도메인 지식을 구조화된 형태로 표현하는 것을 돕기 위해 많이 사용되었다. 전문가 다이어그램 외에도 개념 그래프(Conceptual Graph) (Amati and Ounis, 2000; Plant, 1994; Szpakowicz, 1990; Way, 1994), 의사결정 테이블(Decision Table) (Plant, 1994; Seagle and Duchessi, 1995), 인플루언스 다이어그램(Influence Diagram) (Boose et al., 1993; Kim et al., 1988; McGovern et al., 1991) 등이 지식을 구조화하는데 이용되어 왔다. 이들 다이어그램은 추론과정을 시각적으로 표현함으로써 논리적 오류 없이 지식을 구체적으로 표현할 수 있도록 지원하였다. 또한, 전문가의 지식과 실제 추론에 사용되는 추론 규칙의 중간 매체로서 사용되었으며, 추론규칙으로 자동 변환될 수 있기 때문에 쉽게 규칙과의 연관관계를 보일 수 있다. 그러나,

XRML의 목적과 같이 원본 문서와 규칙을 연결시키지는 못한다.

2.2.2 온톨로지(Ontology) 획득을 위한 자연어 처리

Wetter and Nűse(1992)가 지적한 바와 같이 자연어는 그 중의성으로 인해 기계에 의해 완전하게 해석하는 것이 거의 불가능하다. 따라서 자연어처리를 이용하여 웹으로부터 규칙을 습득하는 것 역시 매우 어려운 작업이다. 그럼에도 불구하고 최근 웹에서 특정 도메인에 대한 지식을 획득하기 위한 도구로 온톨로지가 활발하게 사용되고 있다(Guarino, 1997; van Heijst et al., 1997). 도메인이 한정되어 있는 경우에, 온톨로지는 적절한 어휘를 선택하는데 도움을 줄 수 있다(Crow and Shadbolt, 2001; Hulth et al., 2001; Vargas-Vera et al., 2001). 또한 지식이 서술되어 있는 문법적 패턴을 분석하여 온톨로지 스키마로 정의하고 온톨로지의 형태로 지식을 자동 추출하는 연구에서 자연어 처리가 활발하게 사용되고 있다(Babowal and Joerg, 1999; Maedche and Stabb, 2000; Rau et al., 1989; Ruiz-Sánchez et al., 2003). 그러나 자연어로부터 추출된 지식의 완성도가 아직은 만족할 만한 수준이 아니며 따라서 지식관리자의 수정 및 보완이 필수적으로 요구되고 있다(Schmidt and Wetter, 1998; Wetter and Nűse, 1992).

XRML 접근방식의 기본적인 아이디어는 자연어 처리가 아직은 완전하기 못하다는 가정 아래 지식관리자의 규칙습득과정을 보조하고자 하는 것이다. 그러나 자연어 처리의 기본적인 기술들이 웹 페이지의 문장으로부터 규칙을 식별하는 과정에서 규칙을 구성하는 단어나 동의어를 검색하거나 규칙을 편집하는데 사용되었다.

2.2.3 자동학습과 웹 마이닝

연역적 학습, 인공신경망 그리고 통계적 모델과 같은 자동학습 기법은 웹 페이지로부터 수집된 로그 데이터들에 대해 웹 마이닝 (Web Mining)을 사용함으로써 적용될 수 있다 (Jicheng et al., 1999; Kim et al., 2003). 데이터의 구조적인 집합이 일정한 틀로 표현되어 있을 때, 이러한 데이터의 틀을 이용하면 보다 일반화되고 추상화된 지식들을 추출하는 것이 가능하다. 이러한 방법론을 이용하여 많은 수의 기법과 도구가 개발되었다(Alani et al., 2003; Apte et al., 1994; Craven et al., 2000; Soderland, 1999). 그러나 자연어로 된 문장과 표로부터 추론이 가능한 규칙을 습득하고자 하는 경우에는 연역적인 학습 (Apte et al., 1994; Craven et al., 2000; Soderland, 1999)을 이용한 이러한 방법론이 적절하지 않다. 왜냐하면, 대부분의 학습방법론은 추론 가능한 규칙보다는 간단한 개념이나 데이터에 가까운 지식 획득을 목표로 하고 있기 때문이다. 따라서 자동학습을 이용한 지식의 추출과 XRML을 이용한 규칙의 추출은 그 추출 대상이 다르다고 할 수 있다.

3. 확장형 규칙표식언어(XRML) 접근 방법

3장에서는 웹 페이지로부터 규칙을 습득하기 위한 전반적인 과정과 이 과정에서 규칙 식별의 역할에 대해 설명하고자 한다. 이해를 돕기 위해 이들 과정을 Amazon.com의 배송 및 환불과 관련된 규칙을 습득하는 예제와 함께 설명하였다.

3.1 규칙 식별 언어(RIML)의 정의

여기서는 서론에서 설명된 RIML의 역할에 따라 RIML의 문법을 정의하고자 한다. RIML 버전 1.0에서는 문제를 최대한 단순화하기 위하여 *RuleGroup, Rule, variable, value*와 같은 기본적인 요소만을 포함하였다. 이러한 요소들은 XML 문법에 따라 '< >'기호로 표현되었다. 예를 들어 *Delivery Method* 와 같은 변수는 RIML에서 *<variable>Delivery Method </variable>*와 같이 표현되었다. RIML 2.0에서는 RIML의 표현의 범위를 확장하기 위하여 *RuleTable, IF, THEN*, 그리고 *AND, OR, NOT* 과 같은 접속사를 포함하였다. 또한 (*GT, GE, LT, LE, +, -, *, /*) 와 같은 연산자를 포함하도록 하였다. 표로부터 추출된 규칙들은 독립적인 규칙으로 사용될 수도 있지만, 때로는 일반 자연어 문장으로부터 추출된 규칙들과 결합되어야 완전한 규칙이 된다. RIML 2.0의 DTD는 부록 A에 기술되어 있다.

RIML에서 정의된 태그는 표현의 통일성을 위해 RSML에서도 최대한 동일하게 사용하고자 하였다. 그러나, RSML이 규칙을 표현하기 위한 모든 요소를 다 사용하는 반면, RIML은 웹 페이지로부터 규칙을 식별하기 위해 유용한 태그만을 정의해서 사용하고 있다. 예를 들어 수학 연산자의 경우, 표현방식이 매우 다양하고 복잡해서 이를 식별단계에서 완전하게 표현하는 것이 매우 어렵다. 따라서 이런 경우는 5장에서 지적된 바와 같이 규칙을 구조화하고 편집하는 단계인 RSML 생성단계로 지연하는 것이 보다 바람직하다.

3.2 XRML을 이용한 규칙습득과정

[그림 1]에서 설명한 바와 같이 XRML을 이용한 규칙습득과정은 RIML을 이용하여 웹 페이지로부터 규칙을 식별하는 과정과 식별된 규칙들을 RSML 문법에 맞도록 구조화하는 과정으로 이루어져 있다. 둘째 단계인 규칙의 구조화 단계는, XRML 편집기에 의해 자동으로 규칙의 초안을 생성하는 과정과 지식관리자가 수동으로 규칙을 완성하는 과정으로 나누어진다. 다음은 이 과정을 보다 자세하게 정리한 것이다.

Ⅰ. 규칙베이스의 설계 및 웹 페이지의 선정
- 규칙 베이스의 목표와 주제를 선정하고 규칙 베이스를 구성하는 규칙의 그룹들을 설계한다.
- 규칙베이스와 관련된 웹 페이지를 선정한다.

Ⅱ. RIML을 이용한 규칙의 식별
- 웹 페이지로부터 *RuleGroups, Rules, variables, values, IF-THEN* 관계 및 *AND, OR*와 같은 접속사를 식별한다.
- HTML 문서 위에 RIML 태그들을 추가하는 형태로 RIML 문서를 생성한다.

Ⅲ. RIML 문서로부터 RSML 문법을 따르는 규칙 초안을 자동생성
- 웹 페이지로부터 식별된 규칙들은 RSML 문법의 규칙들로 자동 변환한다. 이렇게 자동으로 변환된 결과는 아직 불완전한 형태의 규칙 초안이 된다.

Ⅳ. 완전한 규칙 집합을 만들기 위해 규칙 초안을 보완하고 새로운 규칙을 추가
- 규칙식별 과정에서 완전하게 식별되지 못한 규칙구성요소들을 식별.
- 추론이 완결될 수 있도록 RIML로부터 생성된 규칙과 추론의 결론을 연결하는 규칙을 추가.

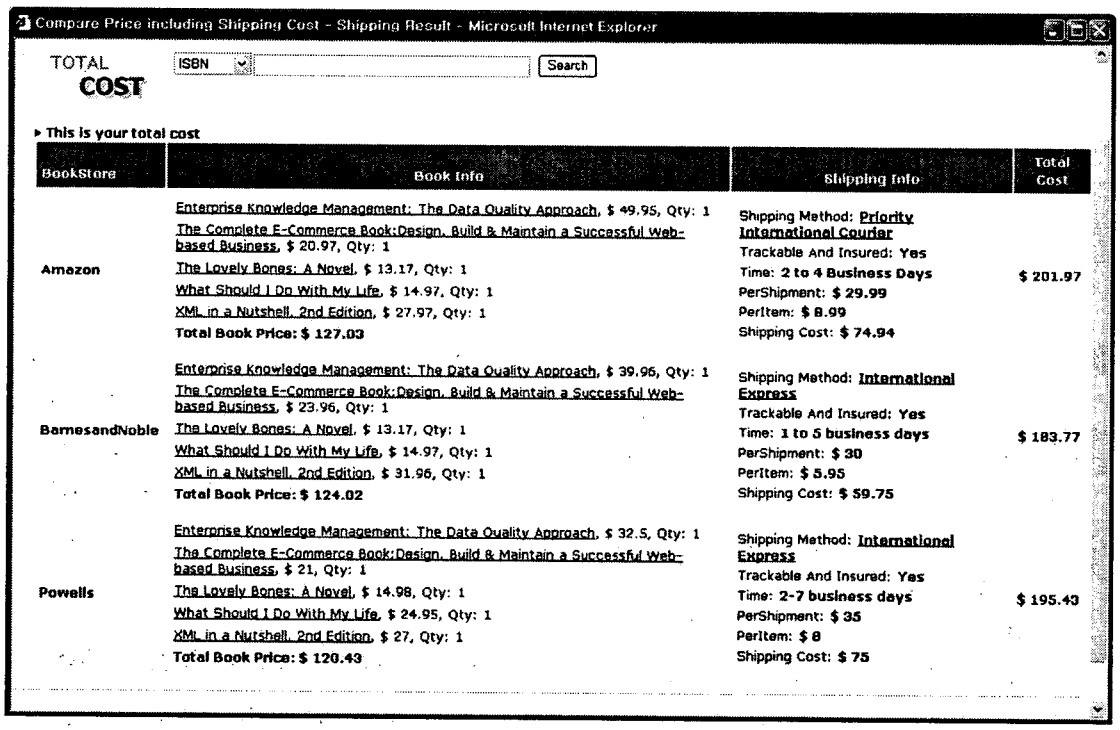
이상에서 제시된 규칙 습득 과정 중에서 지식관리자는 규칙베이스의 설계, 규칙의 식별, 규칙의 보완 과정을 담당하고 있다. 규칙 초안의 생성과정이 자동으로 이루어지는 반면, 이 세 과정은 지식관리자가 XRML 편집기에 의해 도움을 받으며 진행하도록 되어 있다.

3.3 XRML을 활용한 규칙습득의 예

여기에서는 Amazon.com으로부터 규칙을 습득하는 예를 통해 XRML을 이용한 규칙습득과정에 대해 상세히 설명하고자 한다. 습득된 규칙은 [그림 2]와 같은 배송과 교환 및 환불에 대한 전문가시스템을 구축하기 위해 이용되는데, 이렇게 구축된 시스템은 물품의 가격뿐만 아니라 배송비까지도 함께 비교해주는 가격비교사이트에 통합될 수 있다. [그림 2]를 보면 전문가시스템은 *Priority International Courier*라는 배송방법을 사용하여 서울로 보낼 책들에 대해, 각 서점의 책가격과 함께 배송비까지 비교해주고 있다. 결과를 보면 Powells.com이 책 가격은 가장 저렴하나 배송비까지 고려한 경우 BarnesandNoble.com이 가장 저렴한 가격으로 책을 구입할 수 있는 인터넷 쇼핑몰이 된다.

3.3.1 규칙베이스의 설계

Amazon.com 웹 페이지로부터 규칙을 습득하여 배송관련 전문가시스템을 구축하고자 할 때 가장 먼저 할 일은 배송과 교환 및 환불정책에 대해 설명하고 있는 웹 페이지를 Amazon.com 웹 사이트


Compare Price including Shipping Cost - Shipping Result - Microsoft Internet Explorer

TOTAL COST ISBN Search

▸ This is your total cost

BookStore	Book Info	Shipping Info	Total Cost
Amazon	Enterprise Knowledge Management: The Data Quality Approach, $ 49.95, Qty: 1 The Complete E-Commerce Book:Design, Build & Maintain a Successful Web-based Business, $ 20.97, Qty: 1 The Lovely Bones: A Novel, $ 13.17, Qty: 1 What Should I Do With My Life, $ 14.97, Qty: 1 XML in a Nutshell, 2nd Edition, $ 27.97, Qty: 1 **Total Book Price: $ 127.03**	Shipping Method: **Priority International Courier** Trackable And Insured: **Yes** Time: **2 to 4 Business Days** PerShipment: **$ 29.99** PerItem: **$ 8.99** Shipping Cost: **$ 74.94**	**$ 201.97**
BarnesandNoble	Enterprise Knowledge Management: The Data Quality Approach, $ 39.96, Qty: 1 The Complete E-Commerce Book:Design, Build & Maintain a Successful Web-based Business, $ 23.96, Qty: 1 The Lovely Bones: A Novel, $ 13.17, Qty: 1 What Should I Do With My Life, $ 14.97, Qty: 1 XML in a Nutshell, 2nd Edition, $ 31.96, Qty: 1 **Total Book Price: $ 124.02**	Shipping Method: **International Express** Trackable And Insured: **Yes** Time: **1 to 5 business days** PerShipment: **$ 30** PerItem: **$ 5.95** Shipping Cost: **$ 59.75**	**$ 183.77**
Powells	Enterprise Knowledge Management: The Data Quality Approach, $ 32.5, Qty: 1 The Complete E-Commerce Book:Design, Build & Maintain a Successful Web-based Business, $ 21, Qty: 1 The Lovely Bones: A Novel, $ 14.98, Qty: 1 What Should I Do With My Life, $ 24.95, Qty: 1 XML in a Nutshell, 2nd Edition, $ 27, Qty: 1 **Total Book Price: $ 120.43**	Shipping Method: **International Express** Trackable And Insured: **Yes** Time: **2-7 business days** PerShipment: **$ 35** PerItem: **$ 8** Shipping Cost: **$ 75**	**$ 195.43**

[그림 2] 배송비 비교를 위한 전문가시스템 화면의 예

에서 찾는 것이다. 실제로 Amazon.com에서 배달과 관련된 웹 페이지 21개를 찾을 수 있었다. 이들 웹 페이지는 배송이 가능한 지역, 배송 요금, 무료 배송조건, 배송에 소요되는 시간, 그리고 교환 및 환불이 가능한 조건을 명시하고 있다. 여기 설명된 규칙은 *Shipping Rates, Modifying Orders, Shipping Guide, Returns and Refunds*의 4개 범주로 분류될 수 있었다.

3.3.2 RIML을 이용한 규칙의 식별

다음 단계는 선택된 웹 페이지에서 규칙의 존재를 식별하는 것이다. 예를 들어, [그림 3]과 같이 아시아 지역으로의 배송요금을 표시하고 있는 웹 페이지를 살펴보면 선택된 웹 페이지에 규칙을 설명하고 있는 자연어 문장들과 수식, 그리고 *Standard International Shipping, Expedited International Shipping, Priority International Shipping*로 이루어진 세 개의 표가 있음을 볼 수 있다. 이 웹 페이지에 해당되는 HTML문서는 [그림 4]에 나타나 있다.

[그림 3]에 나타나 있는 웹 페이지를 보면, 배송요금이 배송 지역과, 구매한 품목의 종류, 구매한 품목의 수량, 배송방법에 따라 결정되는 것을 알 수 있다.

웹 페이지로부터 규칙을 습득하기 위해 할 일은 *RuleGroup, Rules, variables, values, IF-THEN* 관계와 *AND, OR* 같은 접속사를 식별하는 것이다. 규칙을 설명하고 있는 표는 *RuleTable*로 식별한다. [그림 5]에서는 이러한 식별의 결과로서 HTML 문서에 부가하여 쓰인 RIML 태그들을 이탤릭체로 보여주고 있다.

본 예제에서는 웹 페이지 전체가 배송비용에 대해 설명하고 있기 때문에 전 범위를 '*title* = *Shipping Rates*'인 하나의 *RuleGroup*으로 식별하였다. 웹 페이지에서 처음으로 식별된 규칙은 *<Rule rid=1>*로 표기되었으며, 첫 변수는 *<variable vid=1>able to ship</variable>*로 식별되었다. 이 변수의 변수값인 *True*는 웹 페이지 상에 표시되지 않고 생략되었기 때문에 지식관리자에 의해 *<value vid=1 name="True">*로 식별되었다. 첫째 변수와 변수값은 규칙에서 *THEN* 절에 속하기 때문에 다시 *<THEN>*으로 식별된다. 둘째 변수인 *Items*는 *books, CDs, DVDs, VHS videotapes, music cassettes, vinyl*의 여섯 개의 변수값에 의해 공유되고 있다. 공유된 변수인 *<Items>*가 *Rule 1* 과 *2*에서 각 변수값마다 반복적으로 사용되고 있음을 [그림 6]의 RSML 문서에서 볼 수 있다. RIML에서는 *Items*가 공유된 변수임을 나타내기 위해 여섯 개의 변수값에 *Items*의 *vid*를 동일하게 부여하였다.

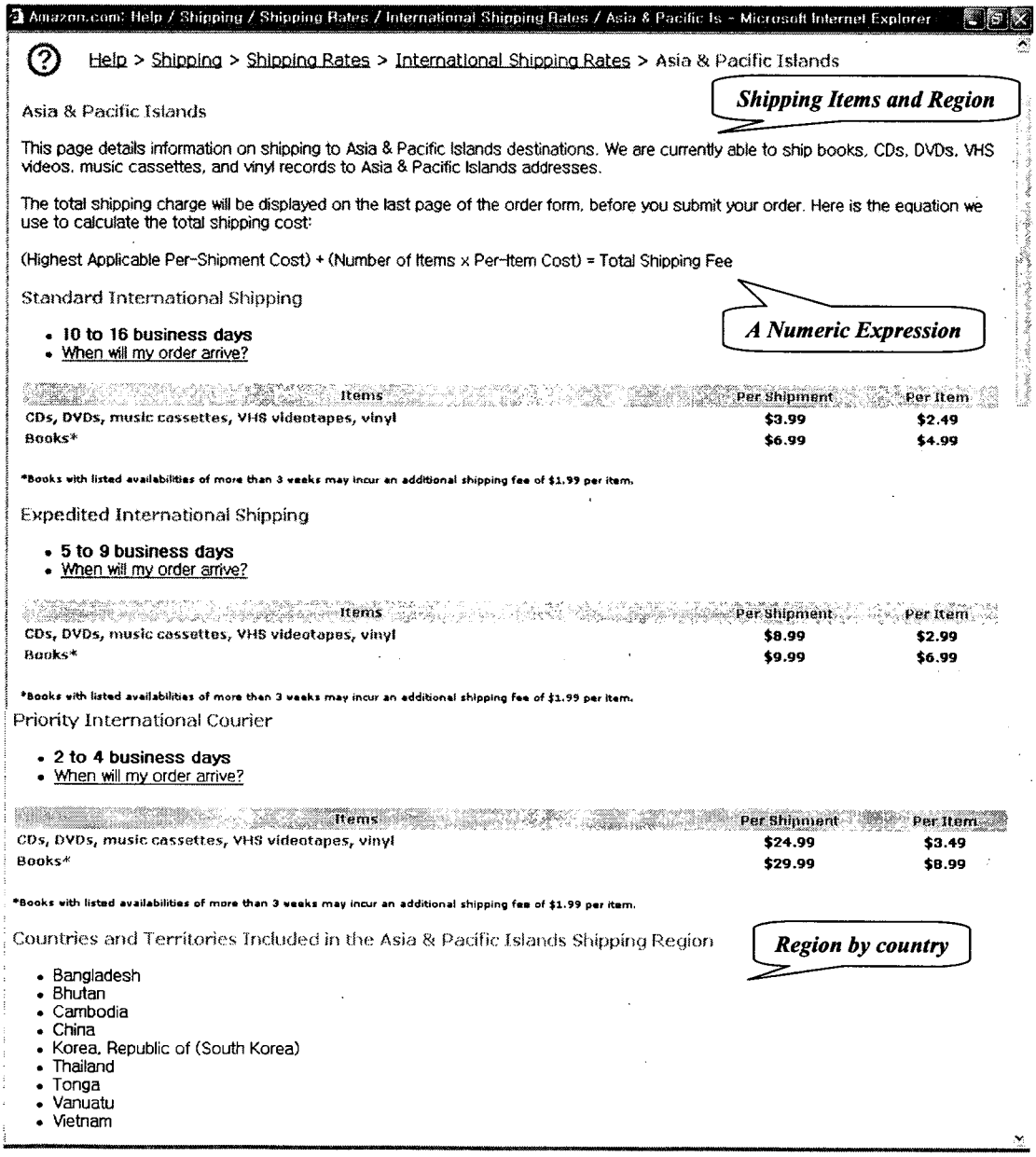
지식관리자가 HTML을 직접 들여다보며 RIML 태그들을 삽입하는 것이 쉽지 않기 때문에, XRML 편집기는 웹 브라우저 모양에서 마우스와 GUI를 이용해 보다 쉽게 규칙구성요소를 식별할 수 있도록 지원하여야 한다. XRML 편집기가 식별을 통해 실제로 생성하는 것은 [그림 5]와 같은 RIML 문서이다.

이상에서 설명한 방식에 따라 Amazon.com으로부터 4개의 *RuleGroup*, 120개의 *Rule*, 35개의 *RuleTable*, 313개의 *variable*, 808개의 *value*, 13개의 *operator*, 119개의 *IF* 절, 120개의 *THEN* 절, 107개의 접속사를 식별하였다. 다 합하면 총 1,635개의 규칙구성요소를 식별하였다.

3.3.3 RSML 문서 초안의 자동생성 과정

RIML 문서의 첫째 목적은, RIML 태그로 식별

된 부분들을 RSML 형태의 규칙 초안으로 변환하는 것이다. 둘째는 RIML 태그들을 걸러낸 HTML 파일을 통해 사람이 보통의 웹 브라우저로 규칙에 대한 설명을 볼 수 있도록 하는 것이다. [그림 5]의 RIML 문서는 자동생성을 통해 [그림 6]의 규칙 초안으로 변환된다.



[그림 3] Amazon.com에서 배송규칙을 설명하는 웹 페이지의 예

```
<HTML><HEAD><TITLE>Amazon.com: Help / Shipping / Shipping Rates / International Shipping
Rates / Asia & Pacific Islands</TITLE></HEAD>
<BODY>
<P><H3>
<img src="help.gif"> Help > Shipping > Shipping Rates > International Shipping Rates > Asia & Pacific
Islands </H3></P>
<P> <H3><font color=#cc6600>Asia & Pacific Islands</font></H3> </P>
<P> This page details information on shipping to Asia & Pacific Islands destinations. We are currently
able to ship books, CDs, DVDs, VHS videotapes, music cassettes, and vinyl records to Asia & Pacific
Islands addresses. </P>
<P> The total shipping charge will be displayed on the last page of the order form, before you submit
your order. Here is the equation we use to calculate the total shipping cost: </P>
<P> (Highest Applicable Per-Shipment Cost) + (Number of Items x Per-Item Cost) = Total Shipping Fee
</P>
.....

<P>Priority International Courier</P>
<UL>
<LI>2 to 4 business days
  .....
</UL>
<TABLE>
<TR><TD>Items</TD>
    <TD>Per Shipment</TD>
    <TD>Per Item</TD></TR>
<TR><TD>CDs, DVDs, music cassettes, VHS videotapes, vinyl</TD>
     <TD>$24.99</TD>
     <TD>$3.49</TD></TR>
<TR> <TD>Books</TD>
      <TD>$29.99</TD>
      <TD>$8.99</TD></TR>
 </TABLE> </P>

.....

<P> <H3><font color=#cc6600>Countries and Territories Included in the Asia & Pacific Islands
Shipping Region</font></H3>
<UL>
<LI>Bangladesh</LI>
<LI>Bhutan</LI>
.....
<LI>Cambodia</LI>
<LI>China</LI>
.....
<LI>Korea, Republic of (South Korea) </LI>
.....
<LI>Vanuatu</LI>
<LI>Vietnam</LI>
</UL> </P>
.....
</BODY></HTML>
```

Shipping Items and Region

Table in HTML

Region by country

[그림 4] [그림 3]에 해당되는 HTML 문서

```
<HTML>...<BODY>
<RIML version="2.0">
<RuleGroup title="Shipping Rates">
<P><H3><img src="help.gif"> Help > Shipping > Shipping Rates > International Shipping
Rates > Asia & Pacific Islands </H3></P>
<URL rsml="http://xrml.kaist.ac.kr/amazon/Shipping_Rates.rsml">
.....

<P>
<Rule rid=1>
This page details information on shipping to Asia & Pacific Islands destinations. <THEN>We
are currently <variable vid=1>able to ship</variable> <value vid=1 name="True"/></THEN>
<IF><OR><variable vid=2 name="items"/><value vid=2>books</value>,<value vid=2>CDs</value>,
<value vid=2>DVDs</value>, <value vid=2>VHS videotapes</value>, <value vid=2>music
cassettes</value>, and <value vid=2>vinyl</value></OR> records to <value vid=3>Asia & Pacific
Islands</value> <variable vid=3 name="Shipping Region">addresses</variable></IF>.
</Rule></P>
.....

<RuleTable>
<P><IF rid="2, 3"><variable vid=4 name="Delivery Method"/>
<value vid=4>Priority International Courier</value> </IF></P>
.....

<Table>
<tr><IF rid="2, 3"><td><variable vid=5>Items</variable></td></IF>
    <THEN rid="2, 3"><td><variable vid=6>Per Shipment </variable></td>
                    <td><variable vid=7>Per Item</variable></td>
    </THEN></tr>
<Rule rid=2>
<tr><IF><td><OR><value vid=5>CDs</value>,
                <value vid=5>DVDs</value>,
                <value vid=5>music cassettes</value>,
                <value vid=5>VHS videotapes</value>
                <value vid=5>vinyl</value></OR></td></IF>
    <THEN><td><value vid=6>$24.99</value></td>
          <td><value vid=7>$3.49</value></THEN></tr>
</Rule>
<Rule rid=3>
<tr><IF><td><value vid=5>Books</value>*</td></IF>
    <THEN><td><value vid=6>$29.99</value></td>
          <td><value vid=7>$8.99</value></td></THEN></tr>
</Rule>
</Table>
</RuleTable>
</P>
.....

<P>
<Rule rid=4><H3><font color=#cc6600>
<IF><variable vid=8>Countries</variable></IF> and Territories Included in the
<IF rid="2, 3"><THEN><value vid=9>Asia & Pacific Islands</value>
<variable vid=9>Shipping Region</variable></THEN></IF></font></H3>
<IF><OR>
<UL>
  <LI><value vid=8>Bangladesh</value></LI>
  <LI><value vid=8>Bhutan</value></LI>
  .....
  <LI><value vid=8>Cambodia</value></LI>
  <LI><value vid=8>China</value></LI>
  .....
  <LI><value vid=8>Korea, Republic of (South Korea) </value></LI>
  .....
  <LI><value vid=8>Vanuatu</value></LI>
  <LI><value vid=8>Vietnam</value></LI>
</UL></OR></IF></Rule>
</P>
.....
</URL></RuleGroup></RIML> </BODY></HTML>
```

Rule 1 Identified

RuleTable in HTML/RIML

Rule 2 Identified

Rule 3 Identified

Rule 4 Identified

[그림 5] [그림 4]에 RIML이 추가된 HTML/RIML 파일

```
<RSML version="2.0">
……
<RuleGroup title="Shipping Rates">
<URL riml="http://xrml.kaist.ac.kr/amazon/AsiaPacific.riml">
<Rule id=1>
 <IF>
  <AND>
    <OR><Items>books</Items>
        <Items>CDs</Items>
        <Items>DVDs</Items>
        <Items>VHS videotapes</Items>
        <Items>music cassettes</Items>
        <Items>vinyl</Items>
    </OR>
    <Shipping_Region>Asia & Pacific Islands</Shipping_Region>
  </AND>
 </IF>
 <THEN><Able_To_Ship>True</Able_To_Ship></THEN>
</Rule>
```

Rule 1

```
<Rule id=2>
 <IF>
  <AND><Delivery_Method>Priority International Courier</Delivery_Method>
       <OR><Items>CDs</Items>
           <Items>DVDs</Items>
           <Items>music cassettes</Items>
           <Items>VHS videotapes</Items>
           <Items>vinyl</Items>
       </OR>
       <Shipping_Region>Asia & Pacific Islands</Shipping_Region >
 </AND>
 </IF>
 <THEN><AND><Per_Shipment>$24.99</Per_Shipment >
            <Per_Item>$3.49</Per_Item></AND>
 </THEN>
</Rule>
```

Rule 2

```
<Rule id=3>
 <IF>
  <AND><Delivery_Method>Priority International Courier</Delivery_Method>
       <Items>Books</Items>
       <Shipping_Region>Asia & Pacific Islands</Shipping_Region >
  </AND>
 </IF>
 <THEN><AND><Per_Shipment>$29.99</Per_Shipment >
            <Per_Item>$8.99</Per_Item></AND>
 </THEN>
</Rule>
```

Rule 3

```
<Rule id=4>
<IF>
 <OR>
  <Countries>Bangladesh</Countries>
  <Countries>Bhutan</Countries>
  …
  <Countries>Cambodia</Countries>
  <Countries>China</Countries>
  …
  <Countries>Korea, Republic of (South Korea)</Countries>
  …
  <Countries>Vanuatu</Countries>
  <Countries>Vietnam</Countries>
 </OR>
</IF>
<THEN><Shipping_Region>Asia & Pacific Islands</Shipping_Region></THEN>
</Rule>
……
</URL></RuleGroup> …… </RSML>
```

Rule 4

[그림 6] [그림 5]로부터 생성된 RSML 초안

이러한 방식으로 Amazon.com에서 식별된 RIML 문서로부터 총 2,520개의 규칙구성요소를 생성하였다. <표 1>에서와 같이 규칙구성요소 중에서 *Rule, RuleTable, IF, THEN, operator*는 규칙식별과정에서의 식별된 수(*IC*)와 규칙 초안에서 자동 생성된 수(*GC*)가 동일하였다. 그러나 변수(*variable*)와 변수값(*value*)은 각기 614개(66.2%)와 119개(12.8%)가 자동으로 추가되었다. 또한 접속사 중에서 *AND*는 152개(57.1%)가 추가되었다.

3.3.4 최종 RSML 문서의 완성

[그림 6]과 같은 규칙 초안으로부터 문법적으로 뿐만 아니라 의미적으로 완성된 규칙을 만들어 내기 위해서는 규칙들을 수정 및 보완하고 새로운 규칙을 추가하여야 한다(Lee et al., 1990; Liebowitz, 1998; Nguyen et al., 1987). 이 작업은 지식관리자에 의해 XRML 편집기 위에서 수동으로 이루어진다. XRML 편집기는 문법적으로 완전하지 않은 *IF* 절이나, *THEN* 절을 찾아서 지식관리자에게 알려줄 수 있다. <표 1>의 *GC* 행에서와 같이 생성된 규칙들 중 어떤 규칙들은 완전하지 않거나, 또는 전혀 식별되지 않는 규칙들이 존재할 수도 있다. <표 1>의 *rC* 행은 규칙을 완성하기 위하여 지식관리자에 의해 기존의 규칙에 추가된 규칙구성요소의 수를 보여주고 있다. 세 개의 산술연산자와 다섯 개의 접속사가 기존 규칙을 완성하기 위해 추가되었다.

그 밖에도 추론의 결론에 도달할 수 있도록 하는 규칙과 규칙을 서로 연결하기 위해 필요한 절은 웹 페이지에 명시되지 않기 때문에, 지식관리자는 RSML 단계에서 이러한 절들을 추가해 주어야 한다. 예를 들어, [그림 6]의 *Rule 2*와 *3*은 *<Set_Shipping_Rates>Computed</Set_Shipping_Rates>*가 각 규칙의 THEN 절에 추가되어 완성된다. 이와 같이, 본 예에서는 전체 규칙구성요소의 7.5%에 해당되는 206개의 규칙구성요소가 추가되었으며 <표 1>의 *lC* 행이 각 구성요소 별로 추가된 수를 보여주고 있다. 이 예에서는 완전한 추론을 위해 필요한 규칙구성요소의 92.2%가 규칙식별단계에서 식별되었다.

<표 1> Amazon.com으로부터 생성된 규칙구성요소의 통계

Rule Component	Identified at RIML Stage (IC)	Automatically Generated from Shared Components (sC)	Automatically Generated from Default Operators (dC)	All Generated from RIML Statements (GC=IC+sC+dC)	Interactively Added for Rule Refinement (rC)	After Refinement (RC=GC+rC)	Effectiveness (%)=[GC/RC]*100	Added to Provide Linkages (lC)	Total in Complete Rules (TC=RC+lC)	Linkage Effort (%)=[lC/TC]*100	Overall Effectiveness (%)=[GC/TC]*100
RuleTable	35	0	0	35	0	35	100.0%	0	35	0.0%	100.0%
Rule	120	0	0	120	0	120	100.0%	4	124	3.2%	96.8%
Variable	313	614	0	927	0	927	100.0%	94	1,021	9.2%	90.8%
Value	808	119	0	927	0	927	100.0%	94	1,021	9.2%	90.8%
Operator	13	0	0	13	3	16	81.3%	1	17	5.9%	76.5%
IF	119	0	0	119	0	119	100.0%	5	124	4.0%	96.0%
THEN	120	0	0	120	0	120	100.0%	4	124	3.2%	96.8%
Connectives	107	4	148	259	5	264	98.1%	4	268	1.5%	96.6%
Total	1,635	737	148	2,520	8	2,528	99.7%	206	2,734	7.5%	92.2%

4. 규칙식별 과정에서의 문제점 및 해결방안

본 장에서는 규칙식별 과정에서 발생하는 문제점들과 그 해결방안을 설명하고자 한다. 이와 같은 문제점에는 표로부터의 규칙 습득, 공유된 규칙구성요소의 식별, 생략된 규칙구성요소의 식별, 동의어의 식별과 사용의 문제가 있다. RIML은 규칙 식별과정에서 발생할 수 있는 문제점을 고려하여 이를 해결할 수 있도록 설계하였다.

4.1 표로부터의 규칙 습득

RIML의 초기연구는 웹 페이지에 있는 자연어로부터 규칙을 식별하는 것에 초점을 맞추었다. 그러나 실제 웹 페이지에서는 비슷한 형태를 가진 일련의 규칙들을 사용자가 알기 쉽게 효율적으로 표현하기 위해 표가 많이 사용되고 있다. 표를 이용해 표현된 규칙들 중 일부는 자연어로 된 문장을 통해 보완되기도 한다. 따라서 표를 이용해 표현된 규칙들을 표현하고 이를 자연어로 표현된 규칙과 결합시키기 위한 규칙표현방안을 제안하였다. RIML 태그 중 <*RuleTable*>은 이를 목적으로 추가되었다.

표를 식별하는 과정을 살펴보면 다음과 같다. 일반적으로 표에서 첫 행에 있는 각 열의 제목은 변수로 식별되고 나머지 행의 내용들은 변수값이 된다. 또한 첫 열의 변수와 변수값들은 *IF* 절에 해당하게 되며 나머지 열의 변수와 변수값들은 *THEN* 절에 해당하게 된다. 그러나 표를 이용한 규칙표현방식이 다양하기 때문에 항상 이 규칙을 따르는 것은 아니므로 지식 관리자의 역할이 중요하다. 특히 표를 이용해 표현된 규칙들은 표에서 부족한 부분을 표 외부의 자연어 문장으로 표현하는 경우가 있으므로, 표가 외부에 연결된 문장들과 결합해야 완전한 규칙의 형태를 갖출 수 있는 경우가 종종 발생한다. 따라서 외부 문장과 표가 연결되어 규칙을 만들 수 있도록 표로부터 규칙을 식별해 낼 수 있어야 한다.

예를 들어 [그림 7]에서와 같이 *Priority International Courier*는 표의 제목 역할 뿐만 아니라 표의 각 행의 조건부로 추가되어야 한다. 또한 [그림 7]의 표는 아랫부분의 *Asia & Pacific Region*일 경우에 해당되는 *Shipping Rates*이기 때문에 아랫부분의 지역을 설명하는 부분과 표를 적절히 연결할 수 있어야 한다. 표 밖의 문장이 표에서 조건으로 쓰이는 경우에 표와 표 밖의 문장을 연결하기 위해서는 그 조건절이 표의 어느 규칙들과 연결되는 지를 <*IF*>문의 *rid* 속성을 사용하여 표시하도록 한다. 예를 들어, [그림 5]에서의 마지막 문단을 보면 *Asia & Pacific Islands*라고 표시된 부분의 앞부분에 <*IF rid="2,3"*>으로 표현되었는데, 이 의미가 해당 절이 표에 사용된 *Rule 2*와 *3*에 연결되어 있음을 보여주는 것이다.

표에서 규칙을 식별할 때 XRML 편집기는 규칙표의 식별, 행과 열을 이용한 변수 및 변수값의 식별 및 *IF*절과 *THEN*절의 식별을 지원함으로써 지식관리자가 보다 쉽게 규칙을 식별할 수 있도록 한다.

4.2 공유된 규칙구성요소의 식별

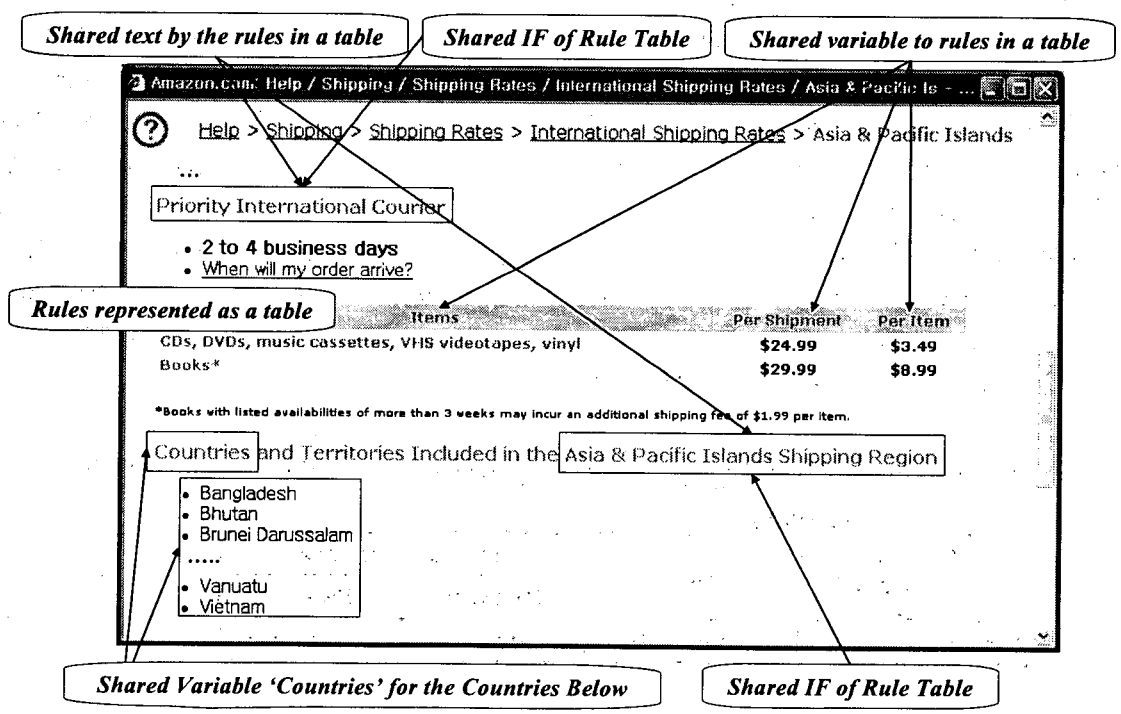
자연어의 특성상 웹 페이지에서는 특정 단어의 반복을 피하여 설명하는 경우가 많이 있다. 이 때 반복을 피해 사용된 단어가 규칙에서 변수의 역할을 하는 경우, 이 변수와 연결된 변수값들은 하나의 변수를 공유하게 된다. 드물게 여러 개의 변수가 하나의 변수값을 공유하는 반대의 경우도 있을

수 있다.

예를 들어 [그림 7]에서와 같이 표의 머리 부분에 있는 *Per Shipment*는 한번 적혔지만 그 값은 변수 *Items*의 종류에 따라 두 번 적힘을 알 수 있다. 또 다른 예는 표 밖의 *Priority International Courier*로서 이 변수값은 표의 모든 규칙의 조건으로 공유된다. 마지막으로 아랫부분의 *Countries* 라는 변수 역시 아랫부분에 열거된 모든 나라이름의 값에 공유됨을 알 수 있다.

웹 페이지에서 공유된 변수를 식별함으로써 지식관리자가 해야 하는 작업을 상당 부분 감소시킬 수 있다. 공유된 변수에 대한 식별과정 없이 지식관리자가 직접 규칙을 작성하는 경우, 지식관리자는 여러 규칙에서 동일한 변수를 반복하여 적어야 하지만 공유변수의 식별을 이용하면 XRML 편집기가 이러한 단순반복과정을 규칙 초안을 작성하는 과정에서 대신 해 줄 수 있다. 공유된 규칙구성요소의 식별은 공유된 변수와 이 변수를 공유하고 있는 변수값들에 속성으로 동일한 *vid*를 부여함으로써 이루어진다. 예를 들어 [그림 5]의 11번째 줄에 있는 생략된 변수 *Items*는 *<variable vid=2 name="Items/">*와 같이 표현된다. 이 때 *books*와 *CDs*는 이 *Items*의 변수값으로서 *Items*라는 변수를 공유하고 있음을 *<value vid=2>book </value>*, *<value vid=2>CDs</value>*와 같이 나타낸다.

이와 같이 공유된 구성요소를 식별하는 것은 규칙을 생성할 때 코딩의 노력을 감소할 뿐만 아니라 RSML에서 변경이 발생했을 때 RIML에서는 한번만 변경하면 되므로 일관성 유지에서도 도움



[그림 7] 공유된 규칙구성요소의 예

을 줄 수 있다.

4.3 생략된 규칙구성요소의 식별

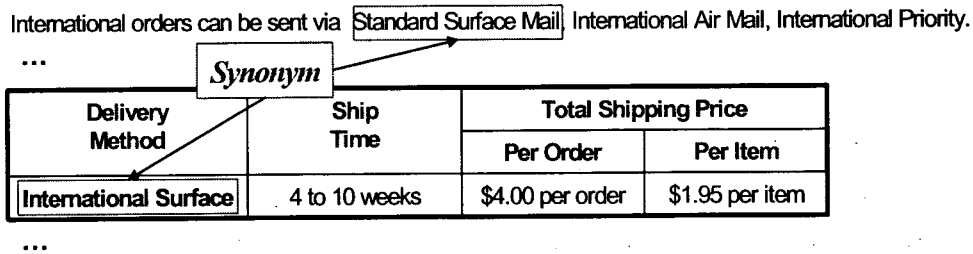
규칙식별에 있어 또 다른 논점은 문맥상 굳이 명시하지 않아도 이해할 수 있는 단어들에 대해서는 웹 페이지에서 생략된 경우가 가끔 있다는 것이다. 생략된 단어가 규칙에 포함되어야 하는 경우에는 규칙식별과정에서 생략된 규칙구성요소를 식별하는 것이 더 바람직하다. 왜냐하면, 웹 페이지를 보면서 생략된 규칙구성요소를 식별하는 것은 어렵지 않지만 나중에 규칙을 구조화하면서 완성하는 단계에서 생략된 규칙구성요소를 알아내는 것은 쉽지 않기 때문이다. 또한 공유된 구성요소와 마찬가지로 생략된 규칙구성요소는 규칙 초안생성의 자동화를 높이는 역할을 한다. 생략된 규칙구성요소를 식별하기 위하여 RIML 2.0에는 <*variable*>과 <*value*>의 속성으로 *name*을 추가하였다.

예를 들어 [그림 7]에서 *Priority International Courier*와 *2 to 4 business days*의 변수에 해당되는 *Delivery Method*와 *Delivery Time*은 생략되어 있다. 각각의 변수는 *name* 속성을 사용하여 <*variable name="Delivery Method"/*>와 <*variable name="Delivery Time"/*>으로 표시하여 웹 페이지 안에 표시해 둠으로써 RSML을 생성할 때 이용할 수 있도록 한다.

4.4 동의어의 식별과 사용

자연어의 특성상 웹 페이지에서는 문장을 매끄럽게 하기 위해 많은 동의어가 사용된다. 그러나, 동의어가 규칙에 포함될 경우, 혼란을 없애기 위해서 각 동의어는 표준단어로 대체되어야 한다. 이러한 동의어 문제를 다루기 위해서는 동의어 사전을 구축하고 동의어가 웹 페이지에서 발견될 때마다 지식관리자에게 이를 알리고 표준단어로 대체할 수 있어야 한다. 그러나 HTML 원문을 손상하지 않도록 하기 위해서는 식별과정에서 동의어를 유지하면서 동의어의 표준단어를 명시할 필요가 있다. RIML 2.0에서는 생략된 규칙구성요소를 표현하는 name 속성을 이용하여 표준단어를 명시하였다.

예를 들어, BarnesandNoble.com 서점에서는 [그림 8]에서와 같이 *International Surface*와 *Standard Surface Mail*이 같은 의미로 쓰고 있다. 이 예제에서는 *Standard Surface Mail*이 표준단어로 사용되고 *International Surface*는 동의어가 된다. 따라서 *International Surface*는 <*variable vid=2 name="Standard Surface Mail"*>*International Surface*</*variable*>과 같이 식별된다.



[그림 8] 동의어 사용의 예

5. 규칙식별의 지연전략

웹에 있는 상당수의 규칙구성요소가 효과적으로 식별될 수 있다 하더라도 어떤 규칙구성요소에 대해서는 웹 페이지로부터 바로 식별하기가 매우 까다로울 때가 있다. 이처럼 식별이 어려운 규칙구성요소들은 RSML 생성단계에서 식별하는 것이 보다 효율적일 수 있다. RIML 단계에서 규칙을 식별함으로써 얻어지는 장점은 HTML에서 식별된 규칙구성요소와 실제 추론 가능한 규칙에 있는 규칙구성요소 간에 일관성을 유지할 수 있다는 것이다. 따라서 식별단계에서 최대한 많은 요소를 식별하는 것이 바람직하다. 그러나 규칙식별을 위해 과도한 노력이 필요한 경우, 오히려 전체적인 성과는 감소할 수 있기 때문에, 지식관리자는 규칙식별을 위해 필요한 노력과 규칙구조화 및 보완에 필요한 노력 간에 적절한 균형을 유지하여야 한다.

규칙식별의 성과는 웹 페이지의 복잡성, 도메인 지식과 지식관리에 대한 지식관리자의 친숙도, 그리고 XRML 편집기의 성능에 따라 달라진다. 이러한 점들을 고려했을 때, 규칙식별을 규칙의 구조화 및 보완단계로 지연하는 것이 바람직한 전형적인 상황은 규칙과 결론의 연결, 복잡한 수식의 표현, 여러 페이지에 산재된 규칙, 외부 출처로부터의 규칙으로 각각의 내용은 다음과 같다.

5.1 규칙과 결론의 연결

모든 규칙기반의 추론은 추론이 결론에 이르렀을 때 추론의 진행을 멈추게 하는 규칙을 필요로 한다. 그러나 이러한 결론 규칙은 웹 페이지 내에 명시되어 있지 않은 것이 대부분이다. 그리고 이 결론 규칙과 웹 페이지에 명시된 규칙을 연결하는 부분도 명시되어 있지 않다. 또한, 규칙들 간을 연결해주는 절 역시 웹 페이지에 명시되어 있지 않은 것이 대부분이다. 이런 경우, 식별과정에서 웹 페이지에 없는 규칙과 규칙구성요소를 식별하고자 노력하기보다는 기본적인 규칙 초안이 생성된 다음 결론 규칙 및 관련 규칙구성요소를 생성하는 것이 더 바람직하다.

예를 들어 [그림 7]의 *Rule 2*와 *3*에서 결론 규칙과 각각의 규칙을 연결하기 위해 *THEN* 절에 ***<Set_Shipping_Rates>Computed</Set_Shipping_Rates>*** 절이 추가되어야 한다. Amazon.com의 경우 연결을 위해 4개의 규칙이 추가되었다.

5.2 복잡한 수식의 표현

웹 페이지에서 복잡한 수식을 표현하고 있는 경우, 규칙식별과정에서 식별의 대상은 수식 자체보다는 수식을 구성하고 있는 변수와 변수값들이다. 따라서 다양한 수학연산자를 포함한 복잡한 수식 전체에 대한 식별은, 식별과정이 지나치게 복잡하지 않도록 하기 위해 식별단계보다 RSML 단계에서 하는 것이 더 바람직하다. [그림 3]을 보면 전체 배송 비용을 계산하는 식이 있는데 이러한 식은 마크업 태그를 이용하여 표현하기 쉽지 않다는 것을 알 수 있다.

5.3 여러 페이지에 산재된 규칙

하나의 규칙을 표현하는 문장들이 여러 웹 페이지에 걸쳐 산재해 있는 경우, 각각의 문장들을 연결하여 규칙을 식별하는 것은 쉽지 않은 일이다. 이런 경우, 각각의 웹 페이지에서 식별된 규칙구성요소들을 완전하게 연결하는 것은 RSML 단계로 지연하는 것이 바람직하다. 이를 위해 지식관리자

는 규칙이 아직 불완전함을 표시하여 RSML 단계에서 이를 지식관리자에게 알려주도록 할 수 있다.

5.4 외부 출처로부터의 규칙

웹 페이지에 명시된 규칙들 중에는 웹 페이지 외에 다른 출처의 문장들을 필요로 할 때가 있다. 그러나 규칙의 식별과정에서 다른 출처의 문장까지 식별하는 것은 식별과정을 지나치게 복잡하게 만든다. 따라서 지식관리자는 식별된 규칙에 대해 규칙의 일부가 다른 외부 출처에 있음을 표시하고 RSML 단계에서 이를 완성하는 것이 바람직하다.

6. 규칙 식별 성능의 평가

6.1 실험의 설계

이 장에서는 RIML을 이용한 규칙 식별의 효과를 평가하고자 한다. 평가를 위하여 웹 상에서 널리 알려진 Amazon.com, Barnes & Noble.com (이상 BN.com), Powells.com, 이 세 온라인 서점을 선택하였다. 각 서점들로부터 배송과 교환 및 환불 정책을 기술해둔 36개의 관련된 웹 페이지를 분석하여 규칙 식별을 통해 필요한 규칙들을 습득하였다. 이렇게 습득된 규칙들은 [그림 2]와 같은 책값뿐만 아니라 배송비용까지 함께 포함하여 가격을 비교해주는 비교쇼핑 사이트를 만드는데 사용되었다.

본 연구에서 규칙 식별의 효과를 평가하는데 있어 고려한 사항은 다음과 같다.

1. 규칙 식별단계에서 식별되는 규칙구성요소가 전체 규칙구성요소 중에서 차지하는 비율은 얼마인가?
2. 공유된 변수와 변수값이 규칙 습득에 어떤 효과를 미치는가?
3. 생략된 변수와 변수값이 어느 정도의 비율로 규칙 식별단계에서 식별될 수 있는가?
4. 웹 페이지에 얼마나 많은 동의어가 사용되었는가?

규칙 식별의 효과를 측정하기 위해 규칙 수준과 규칙구성요소 수준에서 필요한 변수를 정의하면 다음과 같다.

IR: RIML 단계에서 식별되는 규칙의 개수
nR: RSML 단계에서 추가되는 규칙의 개수
TR: 완전한 규칙베이스를 구축하는 데 필요한 전체 규칙의 개수

규칙 수준에서는, 식별된 규칙이 완전한 형태의 규칙이 아닌 경우도 있다. <표 2>에 나타나있는 규칙 식별의 효과(***Effectiveness of Rule Identification***)는 (1)과 같이 정의된다.

$$\textit{규칙 식별의 효과(Effectiveness of Rule Identification) (\%)} = IR / TR \quad (1)$$

위에서 언급한 대로 식별된 규칙은 완전한 형태의 규칙이 아닌 경우가 있기 때문에, 좀 더 정확하게 규칙 식별의 성능을 측정하고자 규칙구성요소 수준에서 효과를 살펴보았다. 규칙구성요소는 *RuleTable, Rule, Variable, Value, Operator, IF, THEN, Connectives* 들로 구성되어 있다.

IC: RIML 단계에서 식별된 규칙구성요소의 개수
sC: 공유 구성요소로부터 자동으로 생성된 규칙구성요소의 개수

dC: 디폴트 연산자로부터 자동으로 생성된 규칙 구성요소의 개수
GC: RIML로부터 생성된 규칙구성요소의 개수

$$GC = IC + sC + dC \quad (2)$$

rC: RSML 단계에서 규칙 보완(Rule Refinement)을 위해 추가된 개수
RC: 규칙 보완 후 규칙구성요소의 개수

$$RC = GC + rC \quad (3)$$

이들 표현을 이용하여 식별된 규칙이 전체 규칙에서 차지하는 비율을 측정하기 위해 규칙구성요소 식별의 효과(***Effectiveness of Rule Component Identification***)를 정의하면 (4)와 같다.

$$\textit{효과(Effectiveness)}\ (\%) = [GC/RC] * 100 \quad (4)$$

식별된 규칙구성요소의 개수에 대해 자동으로 생성된 규칙의 구성요소의 비율을 측정하기 위해 공유에 의한 효율성(***Efficiency Ratio by Sharing***)을 정의하면 (5)와 같다.

$$\textit{공유에 의한 효율성(Efficiency Ratio by Sharing)} = sC/IC \quad (5)$$

본 연구에서 효율성(Efficiency)이 의미하는 바는 공유된 구성요소로부터 자동으로 생성된 규칙구성요소로 인해 규칙을 작성하는 노력을 줄일 수 있다는 것이다. 따라서, 효율성이 높다는 것은 많은 숫자의 규칙구성요소가 식별된 구성요소로부터 자동으로 생성되었다는 것을 의미한다. 같은 의미로 디폴트에 의한 효율성(***Efficiency Ratio by Default***)은 (6)과 같이 정의된다.

$$\textit{디폴트에 의한 효율성(Efficiency Ratio by Default)} = dC/IC \quad (6)$$

공유와 디폴트의 효율성을 합하여 자동화에 의한 효율성(***Efficiency Ratio by Automation.***)으로 정의하면 (8)과 같다.

$$aC = sC + dC \quad (7)$$

$$\textit{자동화에 의한 효율성(Efficiency Ratio by Automation)} = aC/IC \quad (8)$$

완전한 규칙베이스를 완성하기 위해서는 지식관리자는 규칙들 간을 연결하기 위한 연결절(linkage statement)들을 추가하는 작업을 해야한다. 이러한 연결절들을 추가하는 작업의 정도를 측정하기 위해 전체 규칙구성요소 중 규칙간 연결을 위해 필요한 규칙구성요소의 비율로 연결노력(***Linkage Effort***)을 (10)과 같이 정의한다. 이러한 연결절의 종류로는 규칙과 규칙들을 연결하거나, 규칙과 결론을 연결하거나 결론을 위해 새로운 규칙들이 추가된 경우가 있다.

lC: 연결절을 구성하는 규칙구성요소의 개수
TC: 전체 규칙구성요소의 개수

$$TC = RC + lC \quad (9)$$

$$\textit{연결 노력(Linkage Effort)}\ (\%) = [lC/TC] * 100 \quad (10)$$

연결절은 규칙식별로 할 수 없는 부분이기 때문에, 이를 규칙식별의 한계정도(***Limit of***

Identification)로 (11)과 같이 정의한다.

규칙 식별의 한계(Limit of Identification) (%)
= *100* - *연결 노력(Linkage Effort)* (11)

(2)와 같이 정의된 Effectiveness는 규칙식별의 한계를 반영하여 (12)와 같이 전체효과(*Overall Effectiveness*)로 정의될 수 있다.

전체효과(Overall Effectiveness) (%)
= *GC / TC* (12)

지금까지 정의된 식을 이용해 세 온라인 서점에 대해 XRML 접근 방법론을 실험을 통해 평가하였다.

6.2 실험의 결과

전체적인 실험결과는 <표 2>에 요약되어 있다. Amazon.com은 21개의 배송관련 웹 페이지를 갖고 있었으며, BN.com은 8개, Powells.com는 7개를 갖고 있었다. 관련 웹 페이지 내에서 *Rulegroup*의 수는 각기 4개씩이었으며, 웹 페이지로부터 추출된 규칙의 수는 각기 123, 101, 258개였다. Powells.com의 웹 페이지 숫자는 세 서점 중 가장 작으나 규칙 개수는 가장 많은데, 이는 Powells.com이 배송비용을 기술한 표의 크기가 크기 때문이다.

Amazon.com의 경우, 35개의 표와 표에 관련된 문장으로부터 79개의 규칙이 추출되었으며, 일반 자연어 문장에서는 전체 규칙의 34.2%에 해당하는 41개의 규칙이 추출되었다. 이처럼 많은 수의 규칙이 표로부터 생성된 이유는 배송비용에 대한 설명이 주로 표를 통해 이루어졌기 때문이다. 120개의 규칙이 RIML 단계에서 식별되었으며 RSML 단계에서는 겨우 4개의 규칙만이 추가되었다. 이것은 96.8%의 규칙이 웹 페이지로부터 직접 식별될 수 있었음을 의미한다. 새로 추가된 4개의 규칙은 기존의 규칙들과 추론의 결론을 연결하기 위해 필요한 규칙들이었다. 이와 비슷하게 BN.com에서는 4개의 규칙이 추가되었고 Powells.com에서는 3개가 추가되었다. 전반적으로 세 사이트에서 97.7%의 규칙이 웹 페이지로부터 직접 식별되었다. 이는 규칙습득이 규칙식별에 매우 높게 의존하고 있음을 보여준다.

그러나, 모든 규칙이 식별과정에서 전부 식별된 것은 아니기 때문에 각 규칙구성요소별로 식별정도를 파악하기 위하여 <표 3>과 같이 규칙구성요소별로 단계별 식별개수를 정리하였다. Amazon.com의 경우, RIML 단계에서 1,635개의 규칙구성

<표 2> 웹 페이지로부터 식별된 규칙의 성능

Book Store	No. of Web Pages	No. of Rule Groups	No. of Tables	No. of Identified Rules			No. of New Rules for Linkage (nR)	No. of Total Rules (TR)	Percentage of IR/TR
				from Tables	from Texts	Total (IR)			
Amazon	21	4	35	79	41	120	4	124	96.8%
BN	8	4	7	54	43	97	4	101	96.0%
Powells	7	4	3	242	13	255	3	258	98.8%
Total	36	12	45	375	97	472	11	483	97.7%

요소가 식별되었으며, 공유된 변수, 변수값으로 인해 737개의 구성요소가, *AND* 접속사로 인해 148개의 규칙구성요소가 자동으로 생성되어 추가되었다. 이는 전체 규칙구성요소의 35.1%에 해당하는 값이다. 이렇게 RIML 단계에서 생성된 규칙구성요소는 2520개이다. 여기서 8개의 규칙구성요소가 RSML 초안을 만들기 위해 수동으로 추가되었다. 따라서 규칙구성요소의 효과(Effectiveness)는 99.7%(=2520/2528)가 된다. 그러나, 결론과 연결하기 위해 전체 7.5%에 해당하는 206개의 규칙구성요소를 추가해야하므로 식별로 최대한 습득할 수 있는 규칙구성요소는 92.5%가 되고, 전체 규칙식별효과는 92.2%가 된다. 이러한 방법으로, 세개의 온라인 서점에 대해 연결절을 고려하지 않은 평균 효과(GC/RC)는 99.7%가 되고, 연결절을 고려한 평균 효과(GC/TC)는 88.5%가 된다. 이처럼 규칙 식별효과는 본 실험에서 높은 수치를 나타내고 있다.

전체 세개의 서점에 대한 공유에 의한 효율성은 0.445(=1961/4402)로 나왔고, 디폴트에 의한 효율성은 0.168(=741/4402)로 나왔다. 이처럼 식별된 규칙구성요소의 61.3%가 규칙생성단계에서 자동으로 생성되기 때문에 규칙작성에 소요되는 노력도 61.3%로 줄어들었다.

생략된 규칙구성요소에 대해 분석한 결과 자연어 문장으로부터 생략된 규칙구성요소를 식별한 비율은 Amazon.com, BN.com, Powells.com에 대해 각각 13.6%, 25.3%, 24.7%이었고 표로부터는 18.6%, 2.0%, 0.4%였다. 표가 자연어 문장보다 비율이 적은 이유는 표의 경우 어느 정도 정형화되어있기 때문에 용어가 생략되는 경우가 적었기 때문이다. 이와 같이 생략된 규칙구성요소를 식별함으로써 규칙을 생성하는 과정을 도울 수 있을 뿐만 아니라, 후에 웹 페이지와 규칙베이스간 일관성 유지에 도움을 줄 수 있다.

동의어의 사용에 대해 분석한 결과 자연어 문장으로부터는 각각 13.0%, 14.8%, 38.3% 였고, 표로부터는 3.5%, 12.0%, 1.3%였다. 여기서도 자연어 문장이 좀 더 동의어를 많이 사용함을 알 수 있었는데, 이는 자연어의 경우 같은 단어를 사용하는 걸 회피하거나 대명사를 사용하는 경우가 많았기 때문이다.

실험 결과를 보면, 생략된 구성요소와 동의어의 효과적인 식별이 규칙 식별을 좀 더 완전하게 하는데 중요한 역할을 하고 있음을 알 수 있다.

6.3 실험의 한계

본 연구의 실험은 XRML 접근방법의 효용성을 어느 정도 예를 통해 보여주었으나 많은 한계점을

<표 3> 웹 페이지로부터 식별된 규칙구성요소의 성능

Book Store	Identified at RIML Stage (IC)	Automatically Generated from Shared Components (sC)	Automatically Generated from Default Operators (dC)	All Generated from RIML Statements (GC=IC+sC+dC)	Efficiency Ratio by Sharing (sC/IC)	Efficiency Ratio by Default (dC/IC)	Interactively Added for Rule Refinement (rC)	After Refinement (RC=GC+rC)	Effectiveness (%)= [GC/RC] *100	Added to Provide Linkages (lC)	Total in Complete Rules (TC=RC+lC)	Linkage Effort (%)= [lC/TC]*100	Overall Effectiveness (%)= [GC/TC]*100
Amazon	1635	737	148	2520	0.451	0.091	8	2528	99.7%	206	2734	7.5%	92.2%
BN	1151	494	109	1754	0.429	0.095	8	1762	99.5%	175	1937	9.0%	90.6%
Powells	1616	730	484	2830	0.452	0.300	3	2833	99.9%	519	3352	15.5%	84.4%
Total	4402	1961	741	7104	0.445	0.168	19	7123	99.7%	900	8023	11.2%	88.5%

가지고 있다. 먼저 본 실험은 웹사이트의 특성에 많은 영향을 받으며 실험을 수행한 지식관리자에 많이 의존하고 있다. 또한 실험에서 사용한 XRML 편집기의 성능에 따라 실험 결과가 달라질 수 있는 여지가 있다. 따라서, 이러한 상황들을 모두 통제할 수 있도록 웹사이트 및 지식관리자의 숫자를 충분히 하는 것이 XRML 접근방법의 효과를 정확하게 측정하기 위해 요구된다. 그러나, 본 연구에서는 XRML 접근 방법에 대하여 효과를 보이기 위해 정확하고 실용적인 효용성이 아닌 잠재적인 효용성을 보이는 데 초점을 맞추었다. 이는 본 연구에서 새로운 방법론을 제안하였기 때문에, 잠재적인 효용성이라 할 지라도 이 방법론이 어느 정도의 가치가 있는 지를 보이고자 하는 목적으로 기본적인 환경에서 실험을 수행하였다.

본 연구의 방법론은 응용 도메인, 지식관리자의 능력, 지식관리자의 XRML에 대한 친숙도 정도, 웹 페이지의 구조화 정도, XRML의 편집기 능력에 따라 다양한 결과가 나올 수 있기 때문에 이러한 관점에서 다양한 실험을 수행하는 것을 추후 연구로 계획 중에 있다. 또한, 본 연구에서 주장한 일관성 유지 효과도 함께 측정할 수 있도록 실험을 설계하고 있다.

7. 결론

지식 획득 및 지식 원본과의 유지 보수 문제는 지식공학 분야의 가장 근본적인 장애물로 취급되어 왔다. 오늘날 자원의 보고라 할 수 있는 웹에는 자연어로 표현된 문장과 표로 구성된 무수히 많은 문서들이 존재하고 있다. 이러한 웹 페이지들로부터 규칙을 습득하고 습득된 규칙과 웹 페이지 간의 일관성을 유지하는 방법론이 개발된다면, 지금의 웹을 지능화된 웹으로 발전시킬 수 있는 원동력이 될 것이다.

이상의 목적을 위해, 본 논문에서는 확장형 규칙 표식 언어 (eXtensible Rule Markup Language, XRML) 체계를 개발하였다. XRML은 웹 페이지에 내재되어 있는 규칙을 식별하여 자동으로 정형화된 규칙을 생성할 수 있도록 지원하는 규칙 식별 표식 언어 (Rule Identification Markup Language, RIML)와 구조화된 규칙 표현을 위한 규칙 구조 표식 언어 (Rule Structure Markup Language)로 구성된다. 특히, RIML은 RSML과 유사한 형태로서 HTML안에 내재되어 있는 규칙을 HTML 문서에 직접 명시할 수 있도록 설계되었기 때문에 표나 자연어 문장 형태로 표현된 규칙을 효율적으로 식별할 수 있도록 지원한다. 또한, 이렇게 식별된 규칙은 자동으로 정형화된 RSML 문서로 변환될 수 있다. 이에, 본 논문에서는 RIML을 설계하는 과정에서 웹 페이지로부터의 규칙식별과정과 관련된 문제점인 공유된 변수 및 변수값, 생략된 어구, 동의어와 같은 몇 가지 중요한 현상들을 발견하고 이를 해결하고자 하였다.

제안된 XRML 접근 방법의 성능을 측정하고자, 본 논문에서는 3개의 대표적인 온라인 서점인 Amazon.com, BarnesandNoble.com, Powells.com의 실제 웹 페이지들로부터 배송 및 환불과 관련된 규칙을 습득하여 XRML의 효과를 측정하는 실험을 수행하였다. 실험 결과를 보면, 웹 페이지로부터 규칙의 습득은 97.7%로 매우 높은 정확성을 보였으며, 생성된 규칙의 완전성은 88.5%로 측정되었다. 이러한 실험결과를 통해 XRML이 특정 주제에 관한 전문가 시스템을 구축하기 위해 웹 페이지로부터 규칙을 추출할 때 매우 효율적인 도구가 될 수 있으며, 또한 추출된 규칙과 웹 페이지 간의 일관성이 효과적으로 유지될 수 있음을 알

수 있었다.

그러나 RIML로부터 규칙의 많은 부분이 자동으로 추출되었음에도 불구하고, 아직 규칙을 식별하는 단계는 지식관리자에게 많은 부분을 의존하고 있다. 따라서 향후 연구로 유사한 규칙들 혹은 이 규칙들로부터 추출된 온톨로지를 활용하여 규칙식별을 현재보다 자동화하기 위한 방안을 연구중에 있다. 향후 연구와 본 연구가 결합되면 "지식획득의 병목현상"의 해결에 보다 다가설 수 있을 것으로 기대된다.

본 연구는 다양한 분야에서 적용될 수 있을 것으로 기대된다. 웹 페이지를 통해 자사의 지식을 설명하고 있는 모든 온라인 업체는 그와 동시에 웹을 통해 상담을 해주는 전문가 시스템의 도입을 고려할 수 있을 것이다. 또한 일반 쇼핑몰, 보험, 금융 등의 전자상거래 분야에서 규칙에 기반한 지능적인 비교 쇼핑을 제공하는 데 활용될 수 있을 것으로 기대된다.

참고문헌

[1] 김우택, 이재규, 강주영, "XRML를 활용한 정부와 기업간의 지식 공유체계", 한국경영정보학회 춘계학술대회, 2002, pp. 706-715.

[2] 양성병, 이재규, "XRML 기반의 인터넷 쇼핑몰 약관 감사체계", 한국경영정보학회 춘계학술대회, 2003, pp. 1085-1094.

[3] 이재규, 손미애, 강주영, "확장형 규칙 표식언어(eXtensible Rule Markup Language): 설계 원리 및 응용", 한국지능정보시스템학회논문지, 제8권 제1호, 2002.

[4] Alani, H., Kim, S., Millard, D. E., and Weal, M. J., "Automatic Ontology-Based Knowledge Extraction from Web Documents", IEEE Intelligent Systems, Vol. 18, No. 1, 2003, pp. 14-21.

[5] Amati, G., and Ounis, I., "Conceptual Graphs and First Order Logic", Computer Journal, Vol. 43, No. 1, 2000, pp. 1-12.

[6] Apte, C., Damerau, F., and Weiss, M.S., "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems, Vol. 12, No. 3, 1994, pp. 233-251.

[7] Babowal, D., and Joerg, W., "From Information to Knowledge: Introducing WebStract's Knowledge Engineering Approach", Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering, Edmonton, Alberta, May 1999, pp. 1525 -1530.

[8] Boose, J.H., Bradshaw, J.M., Koszar, J.L., and Shema, D.B., "Knowledge Acquisition Techniques for Group Decision Support", Knowledge Acquisition, Vol. 5, No. 4, 1993, pp. 405-448.

[9] Bray, T., Paoli, J., Sperberg-McQueen, C.M., and Maler, E., "eXtensible Markup Language (XML) 1.0", 2nd edition of W3C, <http://www.w3.org/TR/RECxml>, 2000.

[10] Brickley, D., and Guha, R.V., "Resource Description Framework(RDF) Schema Specification 1.0", W3C Recommendation, <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>, Mar. 2000.

[11] Carlisle, D., Ion, P., Miner, R., and Poppelier, N., "Mathematical Markup Language (MathML) Version 2.0 (Second Edition)", W3C Recommendation, <http://www.w3.org/TR/2003/REC-MathML2-20031021/>, Oct. 2003.

[12] Connolly, D., van Harmelen, F., Horrocks, I., McGuiness, D.L., Patel-Schneider, P.F., and

Stein, L.A., "DAML+OIL Reference Description", W3C Note <http://www.w3.org/TR/2001/NOTE-daml+oil-reference-20011218>, 2001.

[13] Craven, M., DiPasco, D., McCallum, A., and Quek, C.Y., "Learning to Construct Knowledge Bases from the World Wide Web", Artificial Intelligence, Vol. 118, No. 1-2, 1999, pp. 69-113.

[14] Crow, L. and Shadbolt, N., "Extracting Focused Knowledge from the Semantic Web", International Journal of Human-Computer Studies, Vol. 54, 2001, pp. 155-184.

[15] Grosof, B., "DAML Rules Phase II", < http://www.daml.org/rules/>, Oct. 2002.

[16] Grossman, R. L., Bailey, S., Ramu, A., Malhi, B., Hallstrom, P., Pulleyn, I., and Qin, X., "The Management and Mining of Multiple Predictive Models using the Predictive Modeling Markup Language (PMML)", Information and Software Technology, 1999.

[17] Guarino, N., "Understanding, building and using ontologies", International Journal of Human and Computer Studies, Vol. 46, 1997, pp. 293-310.

[18] Horrocks, I., "DAML+OIL: a description logic for the Semantic Web", IEEE Data Engineering, Vol. 25, No. 1, 2002, pp. 4-9.

[19] Horrocks, I., Patel-Schneider, P.F., and van Harmelen, F., "From SHIQ and RDF to OWL: the making of a Web Ontology Language", Journal of Web Semantics, Vol. 1, No. 1, 2003, pp. 7-26.

[20] Horrocks, I., Patel-Schneider, P.F., Boley, H., Grosof, B., Dean, M., "SWRL: A Semantic Web Rule Language Combining OWL and RuleML", <http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/>, 2004.

[21] Hulth, A., Karlgren, J., Jonsson, A., Boström, H., and Asker, L., "Automatic Keyword Extraction using Domain Knowledge", Proceedings of the Second Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, 2001, pp. 472-482.

[22] Jicheng, W., Yuan, H., Gangshan W., and Fuyan, Z., "Web Mining: Knowledge Discovery on the Web", IEEE Systems, Man and Cybernetics Conference Proceedings, Vol. 2, 1999.

[23] Kim, D., Jung, H., and Lee, G., "Unsupervised Learning of mDTD Extraction Patterns for Web Text Mining", Information Processing & Management , Vol. 39, No. 4, 2003, pp. 623-637.

[24] Kim, J.D., and Courtney, J.F., "A Survey of Knowledge Acquisition Techniques and Their Relevance to Managerial Problem Domains", Decision Support Systems, Vol. 4, No. 3, 1988, pp. 269-284.

[25] Lassila, O., and Swick, R.R., "Resource Description Framework(RDF) Model and Syntax Specification", <http://www.w3.org/TR/REC-rdf-syntax/>. Feb. 1999.

[26] Lee, J.K., and Sohn, M., "Extensible Rule Markup Language - toward Intelligent Web Platform", Communications of the ACM, Vol. 46, May 2003, pp. 59-64.

[27] Lee, J.K., Lee, I.K., Ahn, S.M., and Choi, H.R., "Automatic Rule Generation by the Transformation of Expert's Diagram : LIFT", International Journal of Man-Machine Studies, Vol. 30, 1990.

[28] Levy, A., and Rousset, M.C., "Combining Horn Rules and Description Logics in

CARIN", Artificial Intelligence Journal, Vol. 104, Sep. 1998.

[29] Liebowitz, J., "Foundation and Application of Expert System Verification and Validation", The Handbook of Applied Expert Systems, pp. 111-151, CRC Press LLC, 1998.

[30] Maedche, A., and Stabb, S., "Mining Ontologies from Text", Proceedings of the European Knowledge Acquisition Workshop, Lecture Notes in Artificial Intelligence, Vol. 1937, 2000.

[31] McGovern, J., Samson, D., and Wirth, A., "Knowledge Acquistion for Intelligent Decision Systems", Decision Support Systems, Vol. 7, No. 3, Aug.1991, pp. 263-272.

[32] Miller, E., Swick, R., Brickley, D., McBride, B., Hendler J. , and Schreiber, G. , "Semantic Web Introduction, Specifications and Related Works", <http://www.w3.org/2001/sw/>, 2001.

[33] Nguyen, T.A., Perkins, W.A., Laffey, T.J., and Pecora, D., "Knowledge Base Verification", AI Magazine, Vol. 8, No. 2, 1987, pp. 69-75.

[34] Plant, R.T., "Techniques for Knowledge Acquisition from Text", The Journal of Computer Information Systems, Vol. 35, No. 1, 1994, pp. 64-70.

[35] Psaila, G., and Crespi-Reghizzi, S., "Adding Semantics to XML", Proceedings of the Second Workshop on Attribute Grammars and their Applications, Mar. 1999, pp. 113-132.

[36] Rau, L.F., Jacobsa, P.S, and Zernika, U., "Information Extraction and Text Summarizatrion using Linguistic Knowledge Acquisition", Information Processing & Management, Vol. 25, No. 4, 1989, pp. 419-428.

[37] Ruiz-Sánchez, J.M., Valencia-García, R., Fernández-Breis, J.T., Martínez-Béjar, R., and Compton, P., "An Approach for Incremental Knowledge Acquisition from Text", Expert System with Applcations, Vol. 25, No. 1, 2003, pp. 77-86.

[38] RuleML, "The Rule Markup Initiative", <http://www.dfki.uni-kl.de/ruleml/>, 2003.

[39] Schmidt, G. and Wetter, T., "Using Natural Language Sources in Model-based Knowledge Acquisition", Data & Knowledge Engineering, Vol. 26, 1998, pp. 327-356.

[40] Seagle, J.P., and Duchessi, P., "Acquiring Expert Rules with the Aid of Decision Tables", European Journal of Operational Research, Vol. 84, No. 1, Jul. 1995, pp. 150-162.

[41] Smith, M.K., Welty, C., and McGuinness, D., "OWL Web Ontology Language Guide", W3C Working Draft, <http://www.w3.org/TR/2003/WD-owl-guide-20030331/>, 2003.

[42] Soderland, S., "Learning Information Extraction Rules for Semi-structured and Free Text", Machine Learning, Vol. 34, No. 1, 1999, pp. 233-272.

[43] Szpakowicz, S., "Semi-automatic Acquisition of Conceptual Structure from Technical Texts", International Journal of Man-Machine Studies, Vol. 33, No. 4, 1990, pp. 385-397.

[44] van der Alast, W.M.P., and Kumar, A., "XML-Based Schema Definition for Support of Interorganizational Workflow", Information Systems Research, Vol. 14, No. 1, 2003, pp. 23-46.

[45] van Heijst, G., Schreiber, A.T., and Wielinga, B.J., "Using Explicit Ontologies in KBS Development", International Journal of Human-Computer Studies, Vol. 45, 1997, pp. 183-292.

[46] Vargas-Vera, M., Motta, E., Domingue, J., Shum, S.B., and Lanzoni, M., "Knowledge Extraction by using an Ontology-based Annotation Tool", Proceedings of the Knowledge Markup and Semantic Annotation Workshop, Canada, 2001.

[47] Way, E. C., "Conceptual Graphs - Past, Present, and Future", Lecture Notes in Computer Science, Vol. 835, Springer-Verlag, 1994.

[48] Wetter, T., and Nűse, R., "Use of Natural Language for Knowledge Acquisition: Strategies to Cope with Semantic and Programtic Variation", IBM J. Res. Develop., Vol. 36, No. 3, May 1992.

부록 A . The RIML version 2.0 Document Type Definition

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- ENTITY Declarations -->
<!ENTITY % op_type "(LE|LT|GE|GT)">

<!-- ELEMENT and ATTLIST Declarations -->
<!-- ELEMENT: RIML, RuleGroup, URL, RuleTable, Rule
     Declarations-->
<!ELEMENT RIML (#PCDATA)>
<!ATTLIST RIML  version CDATA #REQUIRED>

<!ELEMENT RuleGroup (URL+)>
<!ATTLIST RuleGroup title CDATA #REQUIRED>

<!ELEMENT URL (Rule | RuleTable | IF | THEN | variable |
     value | operator)+>
<!ATTLIST URL  rsml CDATA #REQUIRED>

<!ELEMENT RuleTable ( (Rule| IF | THEN)*, (value |
     variable | operator)*)>

<!ELEMENT Rule ( (IF | THEN)*, (value | variable |
     operator)*)>
<!ATTLIST Rule rid ID #REQUIRED>
               title CDATA #IMPLIED>

<!-- ELEMENT: IF, THEN, AND, OR, NOT Declarations-->
<!-- rid attribute: rule id-->
<!ELEMENT IF (THEN*|AND | OR | NOT | (variable | value
     | operator)+)>
<!ATTLIST IF rid IDREFS #IMPLIED>

<!ELEMENT THEN (IF*|AND | OR | NOT | (variable | value
     | operator)+)>
<!ATTLIST THEN rid IDREFS #IMPLIED>

<!ELEMENT AND (AND | OR | NOT | variable | value |
     operator)+>

<!ELEMENT OR (AND | OR | NOT | variable | value |
     operator)+>

<!ELEMENT NOT (AND | OR | NOT | variable | value |
     operator)>

<!-- ELEMENT: variable, value, opeartor Declarations-->
<!-- vid attribute: variable id->
<!-- name attribute: keyword of variable or value->
<!ELEMENT variable (#PCDATA)>
<!ATTLIST variable vid ID #REQUIRED
                   name CDATA #IMPLIED>

<!ELEMENT value (#PCDATA)>
<!ATTLIST value vid IDREF #REQUIRED
                name CDATA #IMPLIED>

<!ELEMENT operator (#PCDATA)>
<!ATTLIST operator vid IDREF #REQUIRED
                   type %op_type; #IMPLIED>
```

Abstract

Effect of Rule Identification in Acquiring Rules from Web Pages

Juyoung Kang* · Jae Kyu Lee* · Sang-un Park*

In the world of Web pages, there are oceans of documents in natural language texts and tables. To extract rules from Web pages and maintain consistency between them, we have developed the framework of XRML (eXtensible Rule Markup Language). XRML allows the identification of rules on Web pages and generates the identified rules automatically. For this purpose, we have designed the Rule Identification Markup Language (RIML) that is similar to the formal Rule Structure Markup Language (RSML), both as parts of XRML. RIML is designed to identify rules not only from texts, but also from tables on Web pages, and to transform to the formal rules in RSML syntax automatically. While designing RIML, we considered the features of sharing variables and values, omitted terms, and synonyms. Using these features, rules can be identified or changed once, automatically generating their corresponding RSML rules.

We have conducted an experiment to evaluate the effect of the RIML approach with real world Web pages of Amazon.com, BarnesandNoble.com, and Powells.com. We found that 97.7% of the rules can be detected on the Web pages, and the completeness of generated rule components is 88.5%. This is good proof that XRML can facilitate the extraction and maintenance of rules from Web pages while building expert systems in the Semantic Web environment.

Key words : 규칙식별(Rule Identification), 규칙습득(Rule Acquisition), 지식공학(Knowledge Engineering), 확장형 규칙 표식 언어(XRML), 전문가시스템(Expert System), 지식획득(Knowledge Acquisition)

* Kaist Graduate School of Management