Research of Industrial Technology
Vol. 22, No. 1, 2004. 6

정보과학회지
22/4, 2004. 4

정보추출의 기술 현황

한국해양대학교 김재훈*

1. 정보추출이란?

최근 인터넷이 급속히 발전하면서 우리는 정보홍수 속에서 살아가고 있다. 넘쳐나는 정보 속에서 유용한 정보를 찾는다는 것은 쉬운 일이 아니다. 이런 어려움 때문에 원하는 정보를 찾기 위해 하루에도 많은 시간을 인터넷과 씨름하면서 보낼 것이다. "여러분은 인터넷에서 원하는 정보를 쉽게 찾을 수 있었나요?" 모르긴 해도 하루 종일 원하는 정보를 찾기 위해 이 누리집 저 누리집을 누비고 다녔을 것이다.

요즘 인터넷을 능숙하게 잘 다루는 젊은이들을 N세대라고 한다. 이런 N세대의 실업 문제가 신문이나 여러 대중 매체에서 연일 크게 보도되고 있다. N세대의 대부분은 구직이나 채용 정보를 인터넷을 통해서 구할 것이다. N세대들은 과연 자신들이 원하는 정보를 얼마나 쉽게 찾을까? 아마도 그렇게 쉽지는 않을 것으로 생각된다. 검색 엔진에서 "시스템 프로그래머 직원 채용"라고 질의했다고 가정하자. "첫 화면에서 원하는 정보를 찾을 수 있었나요?" 아마도 이런 방법으로 원하는 정보를 쉽게 찾을 수 없을 것이다. 설령 찾더라도 이미 많은 시간을 낭비했거나, 마감시간이 지난 정보일 수도 있다[1]. 아마도 요즘 N세대는 이런 방법으로 채용 정보를 찾지 않고, 채용 정보 전문 누리집(예: ㈜ 잡코리아[2])을 먼저 찾을 것이다. 이런 전문 누리집들은 일반적으로 인터넷이나 신문지상 혹은 채용 회사 등의 채용 공고로부터 필요한 정보를 데이터베이스화 하여 사용자들에게 유용한 정보를 제공한다. N세대들은 바로 이런 누리집들을 통해서 자신들이 필요한 정보를 찾을 것이다. 더 필요한 정보가 있을 때에는 관련 회사의 누리집이나 기타 통신 수단으로 직접 문의하여 필요한 정보를 구할 것이다. 이처럼 ***정보추출 시스템은 신문 기사, 웹 문서, 전자우편 등과 같이 정형화되지 않은 문서를 입력으로 받아서 미리 정해놓은(찾기를 원하는) 정보를 찾아내어 주어진 문서를 요약하는 시스템이다.*** 다시 말해서 비구조화된 혹은 반구조화된 문서로부터 구조화된 데이터베이스를 구축하는 시스템을 말한다. 정보추출 시스템은 채용 정보 뿐 아니라 가격 비교(MORPHEUS[1]) 부동산 정보(XTROS[2]), 법률 정보[3], 의료 정보(Medstract[4]), 도서 정보(ResearchIndex/CiteSeer[5]) 등 매우 다양하게 응용되고 있다.

정보추출은 자연언어이해 기술이 필요하다. 현재 자연언어처리 기술로는 자연언어를 완전히 이해할 수 없다. 따라서 모든 자연언어 문서로부터 원하는 정보를 찾는 것은 다소 무리가 따른다. 그러나 신문 기사로부터 사건사고 일지, 구인구직 광고로부터 채용분야 및 연봉, 세미나 공고로부터 날짜와 주제 정보, 학술대회 공고로부터 날짜와 마감시간, 웹 페이지로부터 주식 정보와 날씨 정보 그리고 가격 정보 등은 현재의 기술로도 충분히 실용화할 수 있으며(가격비교: ENURI[3], 채용 정보:Flip Dog[4], 논문 정보:ResearchIndex/Citeseer[5]), 일부의 기술은 상용화되어 사용되고 있다(Junglee[6], AeroText[7], Jango[8], mySimon[9]. 이와 같이 정보를 추출하는 데는 완전한 수준의 자연언어 이해가 필요하지 않으며, 부분적으로 필요한 정보만 정확히 분석하면 된다. 이런 기술은 부분 분석이라고 하는데 이 기술을 사용함으로써 실제 응용 분야에서 고질적인 견고성 문제도 자연스럽게 해결할 수 있다.

정형화되지 않은 일반 자연언어 문서에서 원하는 정보를 찾아내는 것은 쉬운 일은 아니다. 비록 자연언어

*종신회원

1) 검색 엔진의 색인은 일정한 주기로 갱신되기 때문에 매일매일 발생되는 정보를 바로 찾을 수 없다.

2) http://www.jobkorea.co.kr/

3) http://www.enuri.com/

4) http://flipdog.monster.com/

5) http://citeseer.nj.nec.com

6) http://www.norvig.com/junglee/overview.html

7) http://mds.external.lmco.com/products/gims/aero/

8) http://www.jango.com

9) http://www.mysimon.com

문서가 아니더라도 다양한 영역에 다양한 사용자와 다양한 정보를 제공하기 위한 정보추출 시스템을 구축하는 것은 쉬운 일은 아닐 것이다. 이를 더욱 어렵게 하는 것은 일반적으로 자연언어 문서들이 특정 역역에 따라 서로 다른 형식을 가지고 있다는 것이다. 따라서 특정 영역을 대상으로 개발된 시스템이 다른 영역에 그대로 적용되지 않으며, 경우에 따라서는 새로운 시스템을 구축해야 한다. 새로운 영역에서 시스템을 구축하기 위해서는 고도로 숙련된 정보추출 전문가가 필요하다.

지난 10여 년 동안은 HTML과 같은 반구조화된 문서에서 전자우편이나 뉴스 기사와 같은 자연언어 문서에 이르기까지 다양한 문서를 대상으로 일반적인 정보추출에 대한 연구가 활발하게 진행되어 왔다. 그러나 최근에 인터넷의 발달로 웹 문서가 대량으로 생산되면서 정보추출의 대상도 자연스럽게 웹 문서로 이동하게 되었다. 웹 문서의 경우는 일반 문서와는 달리 그 내용이나 형식이 매우 자유로울 뿐 아니라 매우 빠르게 변하기 때문에 이에 능동적으로 대처해야 한다. 이런 환경의 변화를 반영하여 최근 정보추출에 관한 연구는 기계학습, 지식관리, 에이전트, 자료 융합, 정보 수집 등의 기술과 결합하여 적응 가능한 혹은 학습 가능한 정보추출 분야로 발전해가고 있다[6].

본 논문은 정보추출에 대한 최근 기술 동향에 대해서 기술하며, 다음과 같이 구성된다. 2절에서 정보추출 시스템의 간단한 역사에 대해서 살펴보고, 3절에서는 일반적인 정보추출 시스템의 구조에 대해서 살펴본다. 4절에서는 웹 정보추출의 기술 동향과 웹 정보추출 도구에 대해서 기술한다. 끝으로 5절에서 앞으로 정보추출의 전망을 살펴보고 결론을 맺고자 한다.

2. 정보추출 시스템의 역사

정보추출은 자연언어처리와 그 역사를 같이 하고 있으나, 80년대 말 미국 국방성에서 실용적인 자연언어처리 시스템을 개발하기 위한 목적으로 개최된 MUC(Message Understanding Conferences)[10]을 계기로 급속도로 발전하게 되었다. 정보추출 시스템을 크게 웹 전 시대와 웹 시대로 나눈다[7].

표 1 정보추출 및 시스템 평가와 관련된 학술대회

MUC (Message Understanding Conference)[11]	1987(MUCK1), 1989(MUCK2), 1991(MUC3), 1992(MUC4), 1993(MUC5), 1995(MUC6), 1998(MUC7)	미국방성 (DARPA)	• 대상: 테러 정보(MUC3-4), 벤처기업(MUC5), 전자회로 조립(MUC5) 기업경영관리(MUC6), 항공기엔진(MUC7) • 객관적인 평가 수행(NE, CO, TE, TR) • 자연언어처리 응용 가능성 보임.
TISPTER Workshop[12]	1단계 (1991-1994) 2단계 (1994-1996) 3단계 (1996-1998)	미국방성 (DARPA), CIA 협력기관: NIST, SSC	• 목적: 텍스트 처리의 산학연 공동연구 • 연구분야: 문서검출, 정보추출, 요약 • 1단계: MUC와 TREC를 통해 활동 • 2단계: 기술표준화, 다국어 개체명 인식(MET) 시작 • 3단계: 요약, 상호참조 기술에 중점
MET (multilingual Entity Task Conference)[13]	1996(MET1) 1998(MET2)		• MUC에서 다국어 개체명 인식을 별도의 작업으로 지정해서 수행 • 영어, 중국어, 일본어, 스페인어
TREC (Text REtrieval Conferences)[14]	1992-2000(TREC1~9) 2001-현재(TREC2001~3)	NIST, DARPA	• TISPTER 프로그램의 일환으로 시작 • 대규모 정보검색 시스템 평가 • 산학연 공동연구 • 기술 이전 속도 증가 • 적절한 평가 기술 개발

웹 전 시대에는 주로 신문 기사 등과 같은 자연언어 문서를 대상으로 연구가 진행되었으며 대부분의 시스템은 수동으로 정보추출 규칙을 작성하였다. 그 시작은 Wilks의 논문 "Text Searching with Templates"[8]에서 찾아볼 수 있으나, 그 당시 컴퓨터의 계산 능력이 충분하지 못해 크게 성공할 수 없었다. 최초로 성공한 정보추출 시스템으로는 Sager의 연구팀에 의해서 개발된 것[9]을 들 수 있다. 이 시스템은 의학 분야에 적용되었으며, 손으로 작성된 구문 규칙과 템플레이트를 결합하는 간단한 기술을 사용하였으나, 효과적으로 정보를 추출할 수 있었다. FRUMP[10]은 Schank의 SCRIPT 이론[11]을 적용하여 AP 뉴스 기사를 대상으로 원하는 정

10) http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
11) http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
12) http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
13) http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/
14) http://trec.nist.gov

보를 추출하고자 하였다. 이 시스템은 AP 뉴스에서 테러에 관련된 정보를 추출하는 초보 단계의 시스템이었으며, 구체적인 방법으로 평가되지는 않았으나, 이 시스템은 최초의 상용 정보추출 시스템인 ATRANS[12]에 크게 영향을 주었다. 그 후 실험실 수준뿐 아니라 상업용 수준에 이르기까지 매우 다양한 영역에서 다양한 형태의 시스템들이 등장하게 되었다[13-14]. 한편 이들 시스템에 대한 객관적인 평가를 위해 미 국방성의 지원으로 MUC, TISPTER 등의 학술대회 및 평가대회(표 1 참조)가 개최되면서 여러 가지 다양한 방법으로 정보추출 시스템을 평가하였다. 그 결과로 정보추출 분야는 괄목할만 하게 성장하였다. 이런 대회를 통해서 주목할 만한 시스템으로는 SRI의 FASTUS[15]를 들 수 있다. 이 시스템은 대부분의 모듈이 유한상태 오토마타로 작성되었다는 큰 특징을 가지고 있으며, 현재 연구되는 대부분의 시스템의 기본 모델이 되고 있다.

웹 시대의 가장 큰 특징은 처리 대상이 웹 문서라는 점과 다양한 기계학습 기법들이 사용되고 있다는 점이다. 1994년 AAAI 주최로 개최된 "소프트웨어 에이전트"에 심포지움[16]를 시작으로 웹 문서를 대상으로 기계학습 방법이 적용되게 되었다. 그 후로 기계학습과 정보추출에 관련된 워크샵이나 학술대회가 꾸준히 개최되어 이 분야의 연구가 어느 정도 성숙하게 되었다(표 2).

기계학습을 위해서는 많은 학습 자료나 평가 자료들이 필요하다. 이를 위해서 CMU대학의 WebKB 프로젝트[17]나 NIST의 TREC 학술대회 등에서 제공된 자료들은 이 분야를 더욱 발전하게 하였다. 최근에는 레퍼 추론에 대한 기술이 발전하면서 정보추출 규칙조차도 자동으로 생성하고 있다[18-19]. 이런 기술을 이용해서 구축된 누리집으로는 논문을 대상으로 하는 ResearchIndex/Citeseer, 채용 정보를 대상으로 하는 FlipDog 등이 있으며, 최근에 의학 및 생물학 관련 자료를 통한 정보추출 연구(GENIA[15], Medstract 등)가 매우 활발하게 진행되고 있다.

표 2 최근 개최되었거나 개최될 정보추출 관련된 학술대회

AAAI-04 Workshop on Adaptive Text Extraction and Mining	www.ai.sri.com/~muslea/ATEM-04.html
ECML-03 Workshop on Adaptive Text Extraction and Mining	www.dcs.shef.ac.uk/~fabio/ATEM03/
IJCAI-03 Workshop on Information Integration on the Web	www.isi.edu/info-agents/workshops/ijcai03/iiweb.html
ICAPS-03 Workshop on Planning for Web Services	www.isi.edu/info-agents/workshops/icaps2003-p4ws/
IJCAI-01 Workshop on Adaptive Text Extraction and Mining	www.smi.ucd.ie/ATEM2001/
ECAI-00 Workshop on Machine Learning for Information Extraction	www.dcs.shef.ac.uk/~fabio/ecai-workshop.html
AAAI-99 Workshop on Machine Learning for Information Extraction	www.isi.edu/info-agents/RISE/ML4IE/
AAAI'98 Workshop on AI and Information Integration	www.isi.edu/ariadne/aiii98-wkshp/proceedings.html

3. 정보추출의 기본 구조

초기 정보추출 시스템을 구조적인 면에서 분류하면 크게 두 부류로 나눌 수 있다. 하나는 형태소 분석, 구문 분석, 의미 분석, 문맥 분석으로 구성된 전통적인 자연언어 처리 시스템과 같은 구조를 가진 시스템들이고, 다른 하나는 언어 정보를 거의 사용하지 않고 단순한 문자열 검사를 통해서 필요한 정보를 추출하는 시스템이다[13, 20]. 정보추출 시스템은 적용 분야나 용도에 따라 조금씩은 차이를 보이겠지만, 최근 정보추출 시스템의 구조는 대체로 그림 1과 같은 시스템 구조를 갖는다[15,20].

이하에서는 그림 1의 각 단계에 대해서 좀더 구체적으로 설명하고자 한다. 먼저 문서추출 단계는 초기 정보추출 시스템에서 크게 다루지 않았으나, 최근 웹 기반 정보추출 시스템에서는 매우 활발하게 연구되고 있다[7,21]. 원하는 문서를 추출하는 위해서 이용되는 기술로는 정보검색, 정보여과, 문서 군집화 및 분류 등이 있다. 정보수집은 그림 1에서 보듯이 여러 형태의 정보원을 대상으로 이루어진다. 일반적으로 정보수집 과정은 여러 가지 방법이 있을 수 있으나 최근에는 주로 웹을 대상으로 로봇(robot, crawler)이나 레퍼(wrapper)을 이용해서 문서를 수집한다. 이렇게 수집된 웹 문서는 여

15) http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

러 종류의 태그가 포함되어 있는데 이를 제거하거나, HTML 혹은 XML 파서와 같은 웹 문서 분석 도구를 이용해서 DOM 트리를 생성하여 필요한 정보가 포함된 텍스트나 표 등을 추출한다(4절 참조)

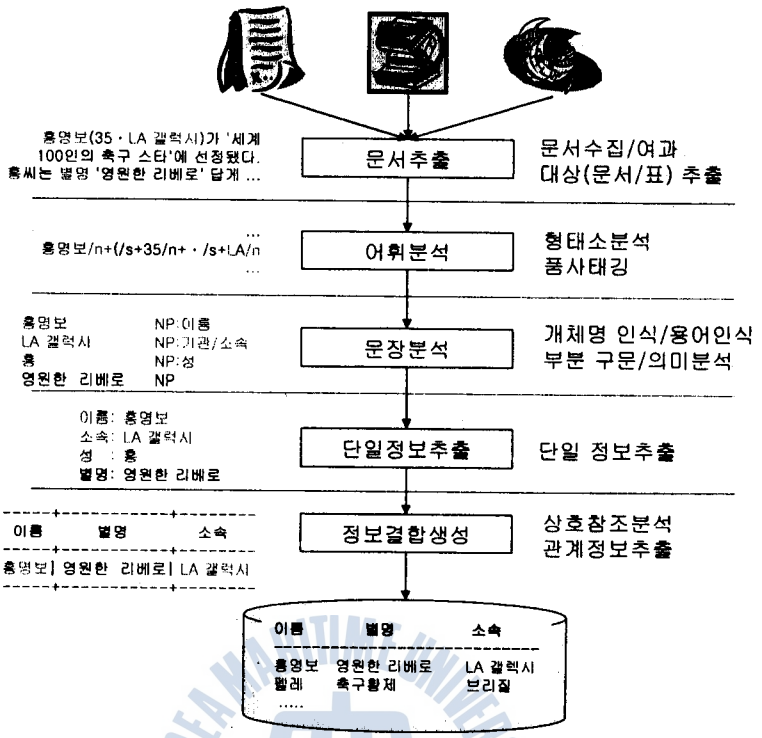

그림 1 정보추출 시스템의 일반적인 구조

어휘 분석 단계는 일반적으로 형태소 분석과 품사태깅으로 구성된다. 즉 단어의 품사를 결정하는 일이 이 단계에서 이루어진다. 예를 들면 그림 1의 예에서 "홍명보"를 명사(n)로 인식하는 과정이다. 일반적으로 정보추출의 기본적인 대상은 명사나 명사구(복합명사구 포함)이므로 품사를 정확하게 결정하는 일은 대단히 중요한 일이다. 최근에는 은닉마르코프 모델과 같은 기계학습 방법을 어휘 분석 단계에 적용하여 적용 영역 및 분야의 변화를 매우 효과적으로 대처하고 있다[22].

문장 분석 단계는 개체명 인식, 부분구문 분석, 부분의미 분석으로 구성된다[23-24]. 이름인식은 문서 내에 사용된 단어 혹은 복합단어가 사람이름(예: "홍명보"), 기관명(예: "LA 갤럭시"), 지명(예: "LA"), 수식표현(예: "100인") 등 고유명사적인 성질을 가진 단어인지를 결정하는 과정이다. 부분구문 분석은 문장의 온전한 구문 구조를 파악하지 않고 부분적으로 명사구, 동사구 등의 구조를 분석하는 과정이다. 예를 들면 그림 1의 예에서 작은따옴표로 둘러싸인 "영원한 리베로"를 하나의 명사구로 인식한다. 경우에 따라서는 부분구문 분석의 결과를 이용해서 완전한 구문 분석을 할 수 있다. 이 과정의 장점은 사전, 문법 등의 지식 부족으로 인해 완전한 구문 분석을 수행하기 어려운 상황에서 널리 사용될 수 있다. 특히 정보추출 분야는 자연언어의 완전한 이해를 필요로 하지 않기 때문에 매우 유용한 과정이다. 부분의미 분석은 적용 영역에 국한된 의미 분석을 말한다. 예를 들면, 최근 활발히 연구되고 있는 생의학 혹은 유전자 관련 정보추출 시스템에서는 생의학에 관련된 단백질 이름, 질병명 등에 관한 의미 분석을 수행한다[25]. 일반적인 의미 분석은 개념 구조와 같은 잘 정의된 언어 지식이 필요하나, 이와 같은 지식을 구하는 것은 쉽지 않다. 따라서 정보추출에서는 정해진 영역에서 꼭 필요한 의미 정보만을 분석하기 위한 목적으로 부분의미 분석을 수행한다[26]. 정보추출에서 사용되는 문장분석 방법도 어휘분석 방법과 마찬가지로 은닉마르코프 모델이나 결정트리 등과 같은 기계학습 방법을 이용하여 적용 영역 등의 변화에 효과적으로 대처하고 있다.

단일 정보추출 단계는 미리 정해진 필드에 단순한 정보를 채우는 과정이다[15,20]. 예를 들면 그림 1에서 정보추출의 대상은 이름, 별명, 소속 기관이므로 이에 관련된 "홍명보"를 이름란에, "LA 갤럭시"는 소속란에, "영원한 리베로"를 별명란에 채운다. 또한 "홍"의 경우에도 이름이 될 수 있기 때문에 필요한 정보로 간주하고 성란에 채운다. 이렇게 채워진 정보로는 "홍명보"의 소속이 "LA 갤럭시"이고 별명이 "영원한 리베로"인지를 정확하게 말할 수 없다. 적용 분야의 빠른 변화에 효과적으로 대처하기 위해서 이 단계에서도 마찬가지로 은닉마르코프 모델

과 같은 기계학습 방법이 주로 이용되고 있다[27].

정보결합 및 생성 단계는 단일 정보추출 단계에서 추출된 각 필드의 정보의 상관관계를 분석하는 단계이다. 예를 들면 그림 1의 예에서 "홍명보"의 소속과 별명을 상황 문맥을 통해서 "LA 갤럭시"와 "영원한 리베로"로 결정하는 과정이다. "홍명보"의 소속을 "LA 갤럭시"로 결정하는 것은 그다지 어려워 보이지 않는다. 그러나 별명을 "영원한 리베로"로 결정하는 것은 약간의 추론과 추가적인 분석이 필요하다. 즉 성으로 인식된 "홍"이 "홍명보"라는 사실을 추론하고 "홍"의 별명이 "영원한 리베로"이므로 "홍명보"의 별명이 "영원한 리베로"로 결정해야 할 것이다. 이런 분석에는 대용어 분석, 상호참조 분석, 동의어처리 등이 있다. 이 단계에서도 마찬가지로 적용 분야의 빠른 변화에 효과적으로 대처하기 위해서 여러 가지 기계학습 방법들이 이용되고 있다[28-29].

4. 웹 정보추출

4.1 웹 정보추출의 동향

최근 정보추출에 관한 연구의 대부분은 웹을 대상으로 하고 있으며 그 연구 영역도 단순한 정보추출에 국한되는 것이 아니라, 정보 통합, 정보 수집, 텍스트 마이닝, 레핑, 적응형 정보추출 등의 분야로 확대되었다[30-32]. 본 절에서는 웹 정보추출과 밀접한 관계를 가지는 레핑에 대해서 좀더 자세히 언급하고자 한다. 웹을 대상으로 할 경우에 가장 큰 문제는 문서 형식이 매우 다양하다는 것이다. 예를 들면 서적의 최저 가격을 추출하는 정보추출 시스템을 구축한다고 할 때, 웹 사이트(예: 아마존[16], Yes24[17] 등)에 따라 서로 다른 형식을 가지고 있다. 이런 다양한 형식을 구조화된 형식으로 바꾸는 일을 레핑이라 하며, 레핑은 레퍼라고 하는 프로그램 혹은 규칙들의 집합에 의해서 수행된다. 레퍼의 주된 역할은 주어진 웹 사이트에 질의를 보내서 결과로 찾아진 페이지로부터 질의의 답에 해당하는 튜플을 추출하는 것이다. 웹 페이지는 인터넷 기술의 빠른 발전에 따라서 매우 빠르게 변하고 있다. 이런 변화 때문에 개발된 혹은 추론된 레퍼도 따라서 자주 변해야 한다. 정보추출 시스템은 일반적으로 여러 개의 레퍼를 사용하는데 이들이 항상 정확한 정보를 추출할 수 있도록 지속적인 관리가 필요하다. 그림 2에서 레퍼를 관리하는 절차를 보여준다[33].

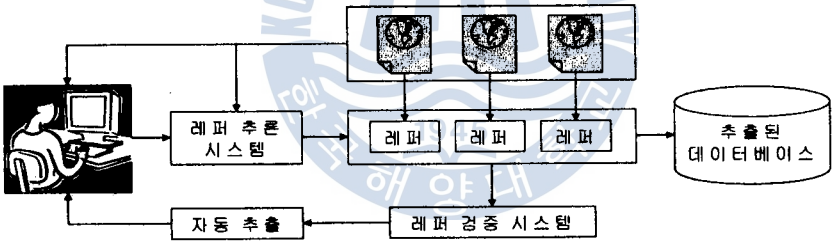

그림 2 레퍼 관리 시스템의 개관

레퍼 검증 시스템은 레퍼가 웹 문서로부터 정확한 정보를 추출하고 있는지를 검증한다[34]. 만약 정확한 정보를 추출하지 않을 경우는 레퍼를 재생성해야 한다. 레퍼의 구축에 가장 어려운 점은 일반적으로 웹 페이지는 사람들이 보기에는 편하도록 되어 있지만, 프로그램에 의해서 필요한 정보를 잘 추출할 수 있도록 되어 있는 것은 아니다. 따라서 페이지 내에서 많은 자연언어 문장이 포함되어 있을 뿐 아니라 이미지와 같은 그래픽 문자들도 많이 포함되어 있다. 더구나 IT 기술의 지속적인 변화에 따라서 이들 페이지는 지속적으로 변하기 때문에 레퍼를 더욱 어렵게 한다. 레퍼는 두 가지 종류가 있다. 하나는 페이지의 구조나 구성을 추출하기 위한 특별한 형태의 문법을 정의하는 것이고[35], 레퍼가 자동적으로 학습할 수 있도록 귀납적 학습 기술을 개발하는 것이다. 최근에는 주로 후자의 방법에 대해서 활발하게 연구가 진행되고 있다. 그 연구의 시작은 [36]이며, HTML 태그의 사용 패턴을 추출하여 문법 형태로 생성한다. [37]에서는 태그 정보와 자연언어의 언어 정보를 사용하여 학습을 시도하였다. XML이 등장하면서 웹 페이지들은 많은 의미 정보를 지니고 있으며 HTML보다 기계가 잘 처리할 수 있어서 레퍼의 구축 및 학습을 어느 정도 용이하게 한다. 더구나 시맨틱 웹이 어느 정도 활성화되면 레퍼의 구축을 더욱 용이하게 할 것이다[38].

4.2 정보추출 도구

웹의 자료들이 기하급수적으로 늘어나면서 웹으로부터 정보추출에 대한 요구도 점차 늘어나면서 이를 지원하는 도구 또한 매우 다양하게 등장하게 되었다. 본 절에서는 이들의 현황을 분석하여 분류하고자 한다.[32]에서는 웹 기반 정보추출 도구를 크게 레퍼 개발용 언

16) http://www.amazon.com/
17) http://www.yes24.com/

어, HTML 인식 도구, 자연언어처리 기반 도구, 레퍼 추론 도구, 모델링 기반 도구, 온토로지 기반 도구로 나누었다.

```
WHERE <bib><book>
   <publisher><name>"Addison-Wesley"
   </name></publisher>
   <title>  $t </title>
   <author> $a </author>
</book></bib> IN "bib.xml"
CONSTRUCT $a
```

그림 3 레퍼 개발용 언어의 예(XML-QL)

레퍼 개발용 언어는 레퍼를 기술하기 위해서 특별히 고안된 언어이며, Minerva[39], TSIMMIS[35], Web-OQL[40], XML-QL[41], WebL[42], WIDL[43] 등이 있다. 그림 3은 XML-QL로 기술된 질의이며, XML 문서 "bib.xml"에서 출판사가 "Addison-Wesley"인 모든 책의 저자를 출력하라는 의미이다.

HTML 인식 도구는 HTML 문서를 파싱하여 구조화된 트리로 표현하고, 반자동 혹은 자동으로 생성된 규칙을 트리에 적용하여 레퍼를 생성한다. 이런 부류의 도구에는 W4F[44], XWRAP[45], RoadRunner[46], Lixto[47], JEDI[48] 등이 있다. W4F를 통해서 어떤 방법으로 레퍼의 생성 과정을 간단하게 살펴보면 그림 4와 같다. W4F는 레퍼 구축을 용이하게 하기 위해서 각종 마법사(form wizard, extraction wizard, mapping wizard)를 두고 있으며, 작성된 레퍼를 검증할 수 있도록 검증 도구(wizard applet)도 지원한다. W4F를 이용해서 레퍼를 설계하는 방법은 크게 세 단계로 구분된다. 첫 번째 단계(RETRIEVAL_RULES)는 접근 단계로서 어떻게 원하는 정보를 가지고 오는가를 지정하는 단계이다. 이 단계는 주로 URL을 지정하여 원하는 페이지를 가져올 수 있으나 경우에 따라서는 스크립트를 이용해서 페이지를 가져오는 경우도 있다. 다음은 추출 단계(EXTRACTION_RULES)로서 어떤 정보를 어떻게 추출할 것인가를 지정한다. 대표되는 몇몇 페이지를 이용해서 레퍼를 학습하면 다른 페이지에 대해서는 일반적으로 그대로 적용될 수 있다. 아래의 예는 영화 제목을 추출하기를 원할 경우에 그 추출 방법으로 명시한 것이다. 마지막 단계(SCHEMA)는 검증 단계로서 앞 단계에서 정의된 레퍼들을 검증하고 수정하는 단계이다. 그림 5는 HTML 인식 도구인 X4F에 대한 예로서 URL "http://us.imdb.com/M/multi-search"에서 영화 "As Good As It Gets"에서 배우와 역할을 추출하기 위한 레퍼이다. 이를 위해 사용하는 규칙으로 그림 5에서 SCHEMA, RETRIEVAL_RULES, EXTRACTION_RULES이 있으며, 이들은 각각 그림 4에서 "mapping rule," "retrieval rule," "extraction rule"에 대응된다. 그림 5에서는 HEL(HTML Extraction Language)라고 하는 레퍼 개발용 언어를 사용한다.

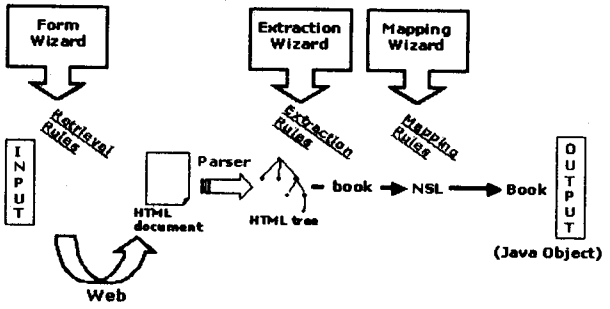

그림 4 W4F의 레퍼 생성과정
(http://cheops.cis.upenn.edu/W4F 참조)

자연언어처리 기반 도구는 자연언어로 쓰여진 문서에서 관련된 자료를 추출하기 위한 규칙을 학습하는 도구이다. 여기서 학습될 추출규칙은 구문 및 의미적인 제약조건이다. 따라서 일반적인 자연언어 처리 도구인 형태소 분석, 품사 태깅 등의 자연언어 처리 도구들이 여기에서 사용된다. 일반적으로 자연언어처리 기반 도구는 구조화된 문서보다는 순수한 자연언어로 구성된 문서에 더 적합하다. 대표적인 도구로 RAPIER[49], SRV[50], WHISK[18], BWI[51], LP2[52] 등이 있으며 예로서 RAPIER에 대해서 간단히 살펴보고자 한다. 그림 6은 RAPIER에 의해서 학습된 규칙이며, 규칙의 형식은 pre-filler, filler, post-filler로 나누어 정의하고 있다. 여기서 pre-filler과 post-filler는 찾기를 원하는 정보의 주변 문맥이다. 그림 6에서 보아 알 수 있듯이 RAPIER 시스템은 품사 태깅, 의미 분석 등 자연언어처리시스템을 이용하고 있으며 이들에 의해서 분석된 정보를 정보추출의 제약조건으로 사용하고 있다. RAPIER의 학습은 귀납적인 방법을 이용하고 있다.

```
SCHEMA { }
RETRIEVAL_RULES {
    get() {
        METHOD: POST;
        URL: "http://us.imdb.com/M/multi-search";
        PARAM: "for" = "As+Good+As+it+Gets",
               "type" = "title";
    }
}
EXTRACTION_RULES {
    movie = html.body( ->h1.txt, match /(.*?) [(]([0-9]+)[)]/ );
}
JAVA_CODE{
    public static void main(String args[])
    throws Exception  {
        IMDB movie = IMDB.get();
        System.out.println(movie);
    }
}
```

그림 5 W4F에서 정의된 레퍼 규칙

```
pre-filler:
    tag: {nn, nnp}
    list: length 2
filler :
    word : undisclosed
    tag : jj
post filler
    sem:price
```

그림 6 Rapier에 의해서 학습된 레퍼 규칙

레퍼 추론 도구는 학습 예제로부터 추출 규칙을 생성한다. 이 도구는 언어적인 제약들을 사용하지 않는다는 점에서 자연언어처리 기반 도구와 구별된다. 대표적인 도구로서 IEPAD[21], WIEN[53], SoftMealy[54], STALKER[55] 등이 있다. 이 분야는 최근에 들어와서 매우 활발하게 진행되고 있는 한 분야이며 일반적으로는 웹 페이지에 표현되어진 정보들을 규칙 추론이 가능한 다른 형태의 자료 구조인 PAT tree[21], FST[54] 등으로 표현하고 이와 같은 자료구조에서 문법 추론과 같은 방법으로 레퍼 규칙을 자동으로 추론하는 방법이며, 추론에 사용되는 기계학습 방법은 매우 다양하다[6]. 예로서 WIEN을 간단하게 소개하고자 한다. 그림 7에서 ㄱ)과 같은 페이지는 ㄴ)과 같은 HTML문서에 의해서 표현되었으며, 레퍼에 의해서 ㄷ)과 같은 정보를 추출하려고 한다고 가정하자. ㄴ)의 HTML 문서를 관찰해 보면 국가명은 〈B〉와 〈/B〉로 둘러싸였고, 국가 코드는 〈I〉와 〈/I〉로 둘려쌓였다. 이와 같이 일반적으로 유용한 정보 혹은 데이터베이스화 할 수 있는 정보들은 일정한 규칙을 가지고 표현된다. 레퍼 추론에서는 이와 같은 규칙을 효과적으로 찾아내는 기술이다. 이 예에서 정보의 분리자로서 〈B〉, 〈/B〉, 〈I〉, 〈/I〉만으로는 부족한 것 같다. 예를 들면, 이 규칙만 사용하면 "〈B〉Some Country Codes〈/B〉"도 추출될 수 있다. 이런 문제를 해결하기 위해서 좀더 자세히 관찰해 보면 원하는 정보가 시작하는 위치에 태그 "〈P〉"가 있고 끝나는 위치에 "〈HR〉"이 있다. 추가적으로 이들 두 경계를 이용해서 원하는 튜플들의 경계를 정확히 추정할 수 있다. 예에서는 "〈B〉Congo 〈/B〉 〈I〉242〈/I〉〈BR〉"이 하나의 튜플이며, 이와 같은 여러 개의 튜플을 귀납적으로 추적하여 국가명과 그 코드를 추정하는 레퍼(규칙)를 추론할 수 있다. WIEN의 경우는 아니지만 이해를 돕기 위해 국가명을 추출하기 위한 레퍼를 정규 표현으로는 "〈B〉([^〈]*)〈/B〉"와 같이 표현하여 〈B〉와 〈/B〉를 제외하고 출력하면 된다. 좀더 구체적인 추론 알고리즘은 [19,53]을 참조하라. 레퍼를 추론할 때 사용되는 기계학습 알고리즘으로는 귀납추론 혹은 관계학습[18,50,56], HMM[27,57,58], 이론정제[59], 여러 방법을 결합하는 부스팅[51] 등 매우 다양한 방법들이 이용되고 있다. 이 방법들의 대부분은 기존에 제안된 여러 가지의 기계학습 방법들을 이용하지만, 레퍼 추론을 위해서 특별히 제안된 방법도 있다.

ㄱ)가상의 웹 페이지

```
<HTML><TITLE> Some Country Codes</TITLE>
<BODY><B>Some Country Codes</B><P>
<B>Congo</B> <I>242</I><BR>
<B>Egypt</B> <I>20</I><BR>
<B>Belize</B> <I>501</I><BR>
<B>Spain</B> <I>34</I><BR>
<HR><B>End</B></BODY><HTML>
```

ㄴ) ㄱ)을 표현하는 HTML 문서

```
(
<Congo, 242>,
<Egypt, 20>,
<Belize, 501>,
<Spain, 34>,
)
```

ㄷ) 추론된 레퍼에 추출된 투플

그림 7 WIEN을 이용한 레퍼 추론을 위한 예[60]

모델링 기반 도구과 온톨로지 기반 도구는 그다지 활발하게 연구되지는 않았다. 본 논문에서는 이들에 대해서 간단하게 소개만 하고자 한다. 모델링 기반 도구는 관심 있는 구조가 주어지면 그 구조에 일치하는 웹 페이지를 찾는 도구이다. 모델에 기본이 되는 요소로는 표나 리스트 등이 있다. 이렇게 관심 있는 구조가 찾아지면 레퍼 추론 도구와 같은 방법으로 원하는 정보를 추출한다. 여기에 속하는 도구로는 NoDoSE[61], DEByE[62] 등이 있다. 온톨로지 기반 도구는 레퍼를 위한 규칙이나 패턴을 추출할 때 온톨로지 정보를 사용하는 경우를 말하며, 대표적인 도구로는 Brigham Young University의 Data Extraction 그룹[62]에서 개발된 도구가 있다.

5. 앞으로의 전망 및 결론

본 논문에서 최근 정보추출의 동향을 기술하였다. 최근 정보추출에 대한 연구의 상당 부분은 웹 문서를 대상으로 하고 있다. 웹 정보 추출에서 가장 큰 문제는 정보

의 출처가 매우 다양하고, 정보원이 자주 변한다는 것이다. 이를 효과적으로 대처하기 위해서 레핑 기술이 이용되며, 이 기술을 통해 레퍼라고 하는 규칙 혹은 프로그램을 반자동 혹은 자동으로 생성한다. 레퍼들은 특정 누리집에 따라 고유한 방법으로 학습되어야 하고, 그 누리집이 변하면 레퍼도 따라서 변해야 한다. 따라서 레퍼가 효과적으로 관리되지 않으면 정확한 정보를 추출할 수 없다[33]. 오늘날 레퍼 관리의 대부분은 정보추출 전문가에 의해서 수동으로 이루어지고 있으나, 웹의 변화에 능동적으로 대처하려는 노력이 꾸준히 진행되고 있다. 이런 연구 분야를 적응형 정보추출이라고 하며, 최근 이를 주제로 한 워크샵(ATEM-2001[18], ATEM-2003[19]) 이 두 번 개최되었으며, 올해도 다가오는 7월에 미국 캘리포니아 산호세에서 ATEM-2004[20]를 개최할 계획이다. 여러 가지의 기계학습 도구와 기술들이 적응형 정보추출 분야의 발전에 큰 도움을 주고 있으나 아직 완전한 적응형 정보추출 시스템 개발은 좀더 연구가 진행되어야 할 것이다.

이런 연구의 일환으로 시맨틱 웹에 대한 연구도 매우 활발히 진행되고 있으나, 적응형 정보추출에 대한 요구는 결코 사라지지 않을 것이다. 왜냐하면, 기존의 수많은 웹 문서를 모두 시맨틱 웹으로 쉽게 바꿀 수 없으며, 또한 모든 사용자들의 요구를 충분히 수용할 수 있는 시맨틱 웹을 구현한다는 것을 거의 불가능하기 때문이다.

적응형 정보추출이 성공하기 위해서는 두 가지 문제가 해결되어야 할 것이다[Cirvegna, 2001]. 첫째, 다양한 분야나 영역에 자동적으로 적응할 수 있어야 한다. 즉, 새로운 영역 정보, 다양한 전문영역 언어, 다양한 장르의 문장, 다양한 형식의 문서를 처리할 수 있어야 한다. 기계학습 기술이 이 문제를 어느 정도 완화시키기는 하지만 기계학습을 위해서는 작은 규모이기는 하지만 학습 자료를 만들어야 하며, 기계학습 방법이 항상 모든 영역에서 만족할 만한 결과를 보이는 것은 아니다. 또한 일반적으로 학습 자료를 만드는 일은 매우 많은 시간과 연구비가 소요된다. 이런 이유로 자율학습에 대한 연구도 함께 진행되고는 있으나 만족할 만한 결과는 아직 없는 편이다. 둘째, 일반 사용자들도 자연스럽게 사용할 수 있도록 인터페이스가 개선되어야 할 것이다. 즉, 정보추출 분야의 비전문가가 손쉽게 새로운 정보추출 시스템을 구축할 수 있도록 사용자 인터페이스가 개선되어야 한다. 오늘날 사용되고 있는 대부분의 정보추출 시스템이 새로운 영역에 적용될 경우, 정보추출 전문가의 손을 거쳐야 한다. 이 문제를 해결하기 위해서는 적절한 설계 시나리오가 개발되어야 하고, 새로운 영역에서 학습된 정보추출 시스템의 결과를 항상 믿을 수 있어야 하며, 사용자의 요구에 따라 추출하기를 원하는 정보를 손쉽게 변경할 수 있어야 한다.

앞에서 언급한 정보추출의 어렵게 하는 요인으로 정보추출의 대상이 되는 누리집의 다양한 형식과 빈번한 수정을 들었다. 여기서 정보추출을 어렵게 하는 요인들을 좀더 살펴보고자 한다. 첫째, 인간은 사용하기 편할 뿐 아니라 화려하고 멋진 웹 누리집을 원한다. 이렇게 되면 될수록 그 누리집에는 더 많은 그래픽, 이미지 등과 같은 멀티미디어 정보를 포함하게 되어 사실상 기계가 인식하기 점점 더 어려워진다. 둘째, 많은 누리집은 스크립트나 프로그램에 의해서 생성되기 때문에 일반적인 방법으로 쉽게 접근할 수 없는 "숨겨진 웹"이다. 이는 텍스트에 대해서 중의성을 줄이기 위해서 XML과 시멘틱 웹 기술이 발전시키려는 노력에 대치되며, 이 기술의 발전 속도보다 "숨겨진 웹"의 성장 속도가 훨씬 더 빠르게 성장하고 있다.

정보추출 분야의 발전 전망은 매우 밝다. 그 이유를 몇 가지만 언급하고자 한다. 첫째, 연구와 실제 응용과의 거리가 멀지 않다는 것이다. 정보추출 기술은 앞에서 언급했듯이 가격 비교, 부동산 정보, 채용 정보 등과 같이 매우 다양한 분야에서 손쉽게 상용화할 수 있기 때문에 연구기관은 물론이고 회사에서도 특별한 관심을 가지고 많은 연구를 진행하고 있다. 둘째, 레퍼 추출에 밀접하게 연관되어 있는 기계학습 알고리즘은 인공지능 분야에서 오래 전부터 연구되어 응용할 수 있는 방법이 매우 다양하다. 셋째, XML이나 시멘틱 웹 기술의 발전으로 웹 문서의 상당 부분은 의미적인 중의성이 현재보다는 훨씬 줄어들 것이다. 넷째, 자연언어처리 기술도 기계학습과 접목하여 영역의 변화나 환경의 변화에 쉽게 적응할 수 있게 될 것이다.

참고문헌

[1] 양재영, 김태형, 최중민 (2001). "MORPHEUS: 확장성이 있는 비교쇼핑 에이전트", 정보과학회 논문지: 소프트웨어 및 응용, 28(2) : 179-191.

[2] 서희경, 양재영, 최중민, "준구조화 정보소스에 대한 지식기반 Wrapper 학습 에이전트", 정보과학회 논문지: 소프트웨어 및 응용, 29(1-2) : 42-52, 2002.

[3] Jackson, P., Al-Kofahi, K., Tyrrell A. and Vachher, A., "Information extraction from case law and retrieval of prior cases,"

18) http://www.smi.ucd.ie/ATEM2001/
19) http://www.dcs.shef.ac.uk/~fabio/ATEM03/
20) http://www.ai.sri.com/~muslea/ATEM-04.html

Artificial Intelligence, 150(1-2) : 239-290, 2003.

[4] Pustejovsky, J., Castaño, J., Saurí, R., Rumshisky, A., Zhang, J., and Luo, W., "Medstract: Creating large-scale information servers for biomedical libraries," Proceeding of Workshop on Natural Language Processing in the Biomedical Domain, ACL-2002, 2002.

[5] Lawrence, S. and Giles, C. L. and Bollacker, K., "Digital Libraries and Autonomous Citation Indexing," IEEE Computer, 32(6) : 67-71, 1999.

[6] Kushmerick, N. and Thomas, B., "Adaptive information extraction: Core technologies for information agents," Intelligent Information Agents R&D in Europe: An AgentLink perspective, Klusch, Bergamaschi, Edwards and Petta, eds. Lecture Notes in Computer Science 2586, Springer, 2003.

[7] Cohen, W. and McCallum, A., "Information extraction from the world wide web," Tutorial Note of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), 2003.

[8] Wilks, Y., Text Searching with Templates, Cambridge Language Research Unit Memo, ML.156, 1964.

[9] Hirschman, L., Grishman, R., and Sager, N., "From Text to Structured Information: Automatic Processing of Medical Reports," AFIPS Conference Proceedings, 45 : 267-275, 1976.

[10] de Jong G., "An overview of the FRUMP system." Strategies for Natural Language Processing, W. G. Lehnert and M.H.Ringle (eds), Lawrence Erlbaum Associates, pp. 149-176, 1982.

[11] Schank, R. C. and Abelson, R. P., Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures, L. Erlbaum, 1977.

[12] Lytinen, S. and Gershman, A., "ATRANS: Automatic processing of money transfer message," Proceedings fo the 5th National Conference of the American Association for Artificial Intelligence, pp. 93-99, 1993.

[13] Cowie, J. and Lehnert, W., "Information extraction," Commun. ACM, 39(1):80-91, 1996.

[14] Seitsonen, L., Extracting information from the World-Wide Web: A Case Study of Fuel Cell Contracts, MS Thesis, Department of Engineering and Physics and Mathematics, Helsinki University of Technology, 1999.

[15] Appelt, D., Hobbs, J., Israel, D., and Tyson, M., "FASTUS: A finite-state processor for information extraction from real-world text," Proceedings of IJCAI-93, 1993.

[16] AAAI. Working Notes of the AAAI Spring Symposium on Software Agents, 1996.

[17] Mitchell, T. et al., The World Wide Knowledge Base Project. http://cs.cmu. edu/~WebKB, 1998.

[18] Soderland, S., "Learning information extraction rules for semi-structured and free text." Machine Learning, 34(1-3) : 233-272, 1999.

[19] Kushmerick, N., "Wrapper induction: efficiency and expressiveness," Artificial Intelligence, 118 : 15-68, 2000.

[20] Cardie, C., "Empirical methods in information extraction." AI Magazine, 18(4) : 65-79, 1997.

[21] Chang, C.-H., Hsu, C.-N. and Lui, S.-C., "Automatic information extraction from semi-structured web pages by pattern discovery," Decision Support Systems Journal 35(1) : 129-147, 2003.

[22] Brill, E., "Tranformation-based error driven learning and natural language processing," Computational Linguistics, 21(4) : 543-565, 1995.

[23] Abney., S., "Partial parsing via finite-state cascades," Journal of Natural Language Engineering, 2(4) : 337-344, 1996.

[24] Bikel, D., Schwartz, R., and Weischedel, R., "An algorithm that learns what's in a

name," Machine Learning, 34(1-3) : 211-231, 1999.
[25] Kim, J.-D. and Tsujii, J., "Corpus-based approach to biological entity recognition." Proceedings of the Second Meeting of the Special Interest Group on Text Data Mining of ISMB 2002, 2002.
[26] Shatkay, H. and Feldman, R., "Mining the biomedical literature in the genomic era: An overview." Journal of Computational Biology, 10(6) : 821-856, 2003.
[27] 엄재홍, 은닉마르코프 모델을 이용한 정보추출, 서울대학교 컴퓨터공학과 석사학위 논문, 2000.
[28] Soon, W., Ng ,H. and Lim, D., "A machine learning approach to coreference resolution of noun phrases," Computational Linguistics, 27(4) : 521-544, 2001.
[29] Ng, V. and Cardie, C., "Improving machine learning approaches to coreference resolution," Proceedings of 40th Annual Meet ing of the Association for Computational Linguistics, 2002.
[30] Florescu, D., Levy, A., and Mendelzon, A., "Database techniques for the world-wide web: Survey," SIGMOD Record, 27(3) : 59-74, 1998.
[31] 윤보현, 황이규, 정의석, 임수종, 왕지현, 임명은, "웹 정보 추출의 동향", 한국인터넷정보학회 2(3) : 3-17, 2001.
[32] Laender, A., Ribeiro-Neto, B., da Silva, A., and Teixeira, J., "A brief survey of web data extraction tools," SIGMOD Record, 31(2) : 84-93, 2002.
[33] Lerman, K., Minton, S. N. and Knoblock, C. A., "Wrapper maintenance: A machine learning approach," Journal of Artificial Intelligence Research, 18 : 149-181, 2003.
[34] Kushmerick, N., "Wrapper verification," World Wide Web Journal, 3(2) : 9-94, 2000.
[35] Hammer, J., Breunig, M., Garcia-Molina, H., Nestorov, S., Vassalos, V., and Yerneni, R., "Template-based wrappers in the TSIMMIS system," Proceedings of the Twenty-Sixth SIGMOD International Conference on Management of Data, 1997.
[36] Riloff, E. M., Information Extraction as a Basis for Portable Text Classification Systems, PhD thesis, University of Massachusetts Amherst ,1994.
[37] Craven, M., Dipasqua, D., McCallum, A., and Mitchell, T., "Learning to extract symbolic knowledge from the world wide web," Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, pp. 509-516, 1998.
[38] 최중민, "시맨틱 웹의 개요와 연구동향", 정보과학회지, 21(3) : 4-10, 2003.
[39] Crescenzi, V., and Mecca, G., "Grammars have exceptions," Information Systems, 23(8) : 539-565, 1998.
[40] Arocena, G. and Mendelzon, A., WebOQL: Restructuring Documents, Databases and Webs, Proceedings of 14th. International Conf. on Data Engineering (ICDE 98), 1998.
[41] Florescu, D., Deutsch, A., Levy, A., Suciu, D. and Fernandez, M. "A query language for XML," Proceedings of 8th International World Wide Web Conference, 1999.
[42] Marais, H., Compaq's Web Language A Programming Language for the Web, Compaq Systems Research Center (SRC), http://research.compaq.com/SRC/WebL/, 1999.
[43] Allen, C. A., "Automating the web with WIDL," XML: Principles, Tools, and Techniques, Dan Connolly, ed. O'Reilly, 1997.
[44] Sahuguet, A. and Azavant F., "Building intelligent web applications using lightweight wrappers," Data Knowledge Engineering, 36(3) : 283-316, 2001.
[45] Liu, L., Pu, C. and Han W., "XWRAP: An XML-enabled wrapper construction system for Web information sources," Proceedings of the 16th International Conference on Data Engineering (ICDE'2000), 2000.
[46] Crescenzi, V. and Mecca, G., On Automat-

ic Information Extraction from Large Web Sites, TR DIA-76-2003, Dipartimento di Informatica e Automazione, Università degli Studi Roma Tre, 2003.

[47] Baumgartner, R., Flesca, S. and Gottlob, G., "Visual web information extraction with Lixto. Proceedings of 27th International Conference on Very Large Data Bases (VLDB 001), pp. 119-128, 2001.

[48] Huck, G., Fankhauser, P., Aberer, K., and Neuhold, E. J., "Jedi: Extracting and synthesizing information from the web," Proceedings of the Conference on Cooperative Information Systems, pp. 32-43, 1998.

[49] Califf, M. and Mooney, R., "Relational learning of pattern match rules for information extraction," Working Papers of the ACL-97 Workshop in Natural Language Learning, pp. 9-17, 1997.

[50] Freitag. D., "Machine learning for information extraction in informal domains," Machine Learning 39(2-3) : 169-202, 2000.

[51] Freitag, D. and Kushmerick, N., "Boosted wrapper induction," Proceedings of the 17th National Conference on Artificial Intelligence AAAI-2000, pp. 577-583, 2000/

[52] Ciravegna, F., "Adaptive information extraction from text by rule induction and generalisation," Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), 2001.

[53] Kushmerick, N. and Grace, B., "The Wrapper Induction Environment," Proceedings of Workshop on Software Tools for Developing Agents, AAAI-98, 1998.

[54] Hsu, C.-N. and Chang, C.-C., "Finite-state transducers for semi-Structured text mining," Proceedings of IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications, 1999.

[55] Muslea, I., Minton, S., and Knoblock. C. A., "STALKER: Learning extraction rules for semistructured, Web-based information sources." Proceedings of AAAI-98 Workshop on AI and Information Integration, Technical Report WS-98-01, AAAI Press, 1998.

[56] Califf, M. and Mooney, R., "Bottom-up relational learning of pattern matching rules for Information Extraction." Journal of Machine Learning Research, 4:177-210, 2003.

[57] Leek T., Information Extraction Using Hidden Markov Models, Masters Thesis, Department of Computer Science & Engineering, University of California, San Diego, 1997.

[58] Freitag, D. and McCallum, A. "Information extraction with HMM structures learned by stochastic optimization," Proceedings of the 17th National Conference on Artificial Intelligence AAAI-2000, pp. 584-589, 2000.

[59] Eikvil, L., Information extraction from world wide web - A survey. Technical Report #945, Norweigan Computing Center, 1999.

[60] Kushmerick, N., Weld, D., & Doorenbos, B., "Wrapper Induction for information extraction." Proceedings of International Joint
Conference on Artificial Intelligence, 1997.

[61] Adelberg, B., "NoDoSE - A tool for semi-automatically extracting structured and semistructured data from text documents." Proceedings of SIGMOD'98, pp. 283-294, 1998.

[62] Laender, A. H. F., Ribeiro-Neto, B. and da Silva, A. S., "DEByE - Data extraction by example." Data and Knowledge Engineering, 40(2) : 121-154, 2002.

[62] Liddle, S. W., Hewett, K. A., and Embley, D. W., "An integrated ontology development environment for data extraction." ISTA2003, 2003.

[63] Ciravegna, F., "Challenges in information extraction from text for knowledge management." IEEE Intelligent Systems and Their Applications, 16(6) : 88-90, 2001.

김 재 훈

1986 계명대학교 전자계산학과(학사)
1988 한국과학기술원 전산학과(공학석사)
1996 한국과학기술원 전산학과(공학박사)
1988~1997 한국전자통신연구원 선임연구원
2000~2002. 2 한국과학기술원 첨단정보기술연구센터, 연구원
2001~2002. 2 USC, Information Sciences Institute, 방문연구원
1997~현재 한국해양대학교 컴퓨터공학과 부교수
관심분야: 자연언어처리, 한국어 정보처리, 정보검색, 정보추출
E-mail: jhoon@mail.hhu.ac.kr