참조연결을 위한 인용정보 자동추출에 관한 연구*

A Study on Automatic Extraction of Citation Information for Reference Linking

김 지 훈(Ji-Hoon Kim)**

목 차

초 록

최근 인터넷의 빠른 성장과 전자출판으로 인해 정보에 접근하는 방법이 상당히 변화하고 있다. 이와 함께, 디지털 문서에서 적합한 정보를 결정하고 효율적으로 추출할 필요가 있을 것이다. 이 연구는 디지털 문서에서 여러 가지 정보를 추출하기 위해 템플릿 마이닝이 사용될 수 있음을 제시하고, 참조연결시스템과 관련된 연구를 고찰하였다. 또한 실제적으로 논문 샘플을 분석한 결과를 이용하여 참조정보를 위한 템플릿을 구축하고, 이 템플릿을 이용하여 수작업으로 논문 샘플을 테스트한 결과, 템플릿 마이닝의 이용은 인용 데이터베이스를 자동으로 만들 수 있는 가능성을 보여주었다.

ABSTRACTS

Recently, the way information is accessed has changed significantly because of the rapid growth of the Internet and a move toward electronic publishing. With that condition, there will be a need to efficiently determine and extract relevant information from digital documents. This study has proposed that template mining can be used for extracting different kinds of information from digital documents, and described research of reference linking system and then built template for citation information using analysis result of sample articles. Also, this study has shown the potential of template mining to automatically create citation databases.

키워드: 참조연결, 인용연결, 인용색인, 인용정보추출, 템플릿마이닝, 인용데이터베이스

---

* 이 논문은 2002년도 한국문헌정보학회 추계학술발표회(2002. 10. 11-12)에서 발표된 내용을 보완한 것임.
** 계명문화대학 문헌정보과 부교수(jhkim@km-c.ac.kr)
논문접수일자 2003년 2월 28일
게재확정일자 2003년 3월 10일

1. 서 론

최근 전자출판과 인터넷의 발전으로 디지털 형태의 학술문헌이 증가하고 인터넷이 효율적인 정보자원이 됨에 따라 정보에 접근하는 방법이 상당히 변화하고 있다. 특히 연구자들이 학술잡지에 발표하거나 발표하지 않은 논문을 특정 웹사이트에 올려놓음으로써, 이용자들은 인터넷을 통해 원하는 학술문헌에 쉽게 접근할 수 있게 되었다. 하지만 인터넷상의 문헌은 체계적으로 조직되어 있지 않을 뿐 아니라, 일반적으로 사용하는 대부분의 탐색엔진이 이러한 문헌에 대해 최신성을 유지하거나 그 내용을 색인하고 있지 않기 때문에, 적합한 문헌을 찾아 이용하는 것이 그다지 쉽지 않은 실정이다.

이러한 문제점에도 불구하고, 인터넷은 이용자가 도서관에 직접 가는 수고를 상당히 감소시키고 있으며, 도서관에 직접 소장되지 않은 문헌에 접근할 수 있게 하여 학술문헌의 유통량을 매우 높이고 있다. 최근, 연구자들의 논문 인용 비율을 분석한 결과에서 오프라인 논문의 평균 인용횟수는 2.74개인데 비해 온라인 논문의 평균 인용횟수가 7.03개로 높게 나타나고, 특히 온라인 논문 중 무료로 입수한 논문이 높은 인용빈도를 나타내고 있는 것으로 조사된 것을 보면(Lawrence 2001), 접근이 편리하면 할수록 학술문헌의 이용률은 증가하는 것을 볼 수 있다. 물론 온라인 논문이 더욱 높게 인용되는 이유는 쉽게 접근해서 볼 수 있다는 면도 있지만, 한편으로는 품질이 우수한 논문이 온라인으로 더욱 입수 가능하게 되었기 때문이라고도 할 수 있다.

특히, 하이퍼링크를 제공하는 온라인상의 학술문헌은 인용한 논문에서 인용된 논문에 바로 접근하는 것을 가능하게 할 수 있는데, 최근 이러한 온라인상의 직접적인 접근을 실행하는 것을 '참조연결(Reference Linking)' 또는 '인용연결(Citation Linking)'이라 부르고 있다. 참조연결에 대한 논의는 오래 전부터 있어온 것으로, 전통적인 참조연결은 논문의 메타데이터와 그 논문의 참고문헌의 메타데이터를 수작업으로 조사하여 인용한 논문과 인용된 논문의 연결사항을 목록화한 '인용색인(Citation Index)'이다.

인용색인은 원래 학술논문의 탐색, 평가, 분석 등을 위해 주로 정보검색분야에서 이용되어 온 것으로, 주로 인용된 논문이 인용한 논문에 기여한 중요도와 그 논문의 유용성과, 특정 논문이 다른 논문에 얼마나 인용되는지를 조사하여 그 논문의 중요도를 파악할 수 있을 뿐만 아니라 나아가 연구경향을 분석하고, 새로운 연구분야를 확인하는데 이용되었다.

이처럼 인용색인이 문헌연구를 위한 중요한 기법임에도 불구하고, 현재 널리 이용되고 있는 ISI(Institute for Scientific Information)사에 제공하는 SCI(Science Citation Index)와 같은 인용색인은 인용정보를 수집하는 과정이 수작업으로 이루어져 인력과 시간이 많이 소요되고, 인쇄 출판된 학술잡지 중 ISI가 선택한 잡지의 논문만을 편향적으로 색인한다는 문제점을 가지고 있다(Cronin et al. 1997). 특히, 인용대상에 있어서 Lawrence의 조사에 의하면, WWW7 회의의 10개 논문을 샘플로 취하고 참고문헌의 분포를 분석한 결과에서 참고문헌의 17%만이 학술잡지의 논문인 반면, 30.3%는 회의논문이고, 18.0%는 도서

이고, 32%는 기술보고서, 논문, 웹 페이지로 나타나 인쇄된 학술잡지의 인용비율이 상대적으로 낮음을 보여주고 있으며(Lawrence 1999), 특히 최근 증가하고 있는 전자잡지, 웹 페이지 등 각종 디지털 형태의 학술문헌을 포함하지 않아 그 대상범위가 제한적이라 할 수 있다.

이러한 문제점을 해결하고 온라인 상에서 문헌의 용이한 접근을 위해 제시된 참조연결은 디지털 형태의 문헌에서 인용색인을 자동으로 만들어 이들 문헌을 서로 연결하는 것으로서, 〈그림 1〉과 같이, 온라인 문헌에 포함된 참고문헌을 바로 연결이 가능한 문헌으로 바꾸어 이용자가 모니터에서 문헌을 보는 동안 그 문헌내의 참고문헌 위치에서 다른 네트워크의 문헌으로 연결하여 바로 그 문헌을 보도록 한다(What is Refrence Linking?).

〈그림 1〉 참조연결 형태

참조연결을 하기 위한 과정은 온라인상의 논문을 자동으로 탐색하여 다운로드한 다음, 그 내용을 분석하여 인용정보를 추출하고 이를 연결시키는 것을 포함한다. 이러한 시스템을 구축하기 위해서는 여러 가지 기술이 필요한데, 디지털 형태의 논문에서 인용정보를 추출하는 방법으로는 주로 패턴인식과 패턴일치에 근거한 템플릿 마이닝(Template Mining)이 효과적인 것으로 제시되고 있다.

이 연구는 앞으로 많은 학술문헌이 온라인으로 입수가능 하리라 예견되는 상황에서 참조연결시스템을 구축하기 위한 선행연구로서, 지금까지의 참조연결에 대한 관련 연구를 고찰하고, 실제적으로 학술논문의 인용정보를 구축하기 위해 템플릿 마이닝을 이용하여 인용정보를 추출하는 가능성을 조사하였다.

2. 인용정보 추출기법

2.1 정보추출

정보추출(Information Extraction)은 규칙에 기반하여 자연언어를 처리하는 분야로서, 짧은 자연어 텍스트에서 미리 지정된 종류의 정보를 자동으로 추출하는 것, 즉 구조화되지 않은 텍스트로부터 구조화된 정보를 얻는 것을 의미한다. 구조화된 정보는 주로 인용색인 데이터베이스, 보고서 생성, 의사결정, 데이터 마이닝 및 인공지능 등을 만드는데 사용된다. 이 중, 참조연결을 위한 인용색인을 자동으로 만들기 위해, 온라인 문서에서 참조정보를 추출하는 것은 학술논문을 표제, 저자 등과 같은 인용한 문헌의 메타데이터, 참조연결점(reference anchors)과 참조문장을 포함하는 본문, 참고문헌등 세 부분으로 구분하여 다음과 같이 분석한다(Bergmark 2000).

1) 인용한 문헌의 메타데이터 추출

인용한 문헌의 메타데이터는 분석될 문헌에 대한 일반적인 데이터 즉 저자, 표제, 출판년도, 잡지명, 권, 호, 페이지 등을 추출하는 것으로, 이를 추출하기 위해 위치, 폰트, 구두점, 레이아웃 등의 단서를 이용한다. 보통 레이아웃 정보는 표제가 무엇인지 결정하는데 사용되는데, 표제는 논문의 앞부분에서 대개 큰 폰트로 나타난다.

표제가 파악되면 보통 저자가 그 다음에 나타나는데, HTML 형식에서 태그가 된 경우는 상대적으로 파악하기 쉽지만, 마크되지 않은 문서의 저자를 파악하기 어렵고, 특히 저자명과 기관명을 구분하는 것이 어려운 점이 있다. 또한 출판정보는 문헌 자체에 포함되지 않은 경우 파악하기 어렵고, 종종 그 문헌의 URL이나 DOI에서 결정될 수도 있다.

2) 본문 파악 및 참조연결점 조사

문헌의 본문을 파악하기 위한 가장 효율적인 알고리즘은 스캔된 텍스트를 체크하여 장 및 절의 제목을 찾고 초록, 서론 또는 목차부분을 구별하는 것이다.

본문의 참조연결점은 본문내의 문자열을 스캔하여 [10], [Lawrence 1999] 등을 포함하는 문장을 탐색하여 수집하고, 아울러 [1-3]은 [1][2][3]으로 대체하고, [smith, 1998; Jones, 1999]는 [Smith, 1998][Jones,

1999]로 하여 쉼표와 세미콜론으로 된 것을 개별적으로 나누는 것을 포함한다. 또한 참조형태가 저자 다음에 바로 출판년도가 있는 'Caplan and Guenther(1996) explored the difficulties..'와 같은 것은 [Caplan and Guenther, 1996]으로 분석하도록 한다.

3) 참고문헌분석

참고문헌은 일반적으로 논문의 마지막 부분에 '참고문헌', 'Reference' 등의 표제 이후에 나타나므로 그 위치를 파악하기는 쉬우며, 다만 목차에서 발견되는 것은 배제한다.

참고문헌분석은 기본적으로 참고문헌의 내용을 개별 요소로 분해하여, 주요 서지 데이터를 추출한다. 종종 나타나는 참고문헌 앞의 번호(e.g. 10.)가 있다면 이를 제거한 후, 나머지 참고문헌내용을 정해진 규칙에 의거하여 추출하고, 아울러 참조연결점과 참고문헌을 일치시킨다.

예컨대, OpCit 프로젝트 경우 Citation.pm이라는 메타데이터 추출 프로그램을 이용하여, 참고문헌에서 authors, first author, journal title, volume, issue or supplement, start page, year 등의 메타데이터를 〈그림 2〉와 같이 추출한다.

2. 2 템플릿 마이닝

정보추출에 사용되는 기법 중 하나인 템플릿 마이닝은 인식 가능한 패턴을 이용하여 텍스트에서 데이터를 추출하는 자연언어처리기법으로, 정보추출시스템은 텍스트가 미리 지정한 템플릿과 일치하면 그 템플릿과 연관된 지식에 의해 데이터를 추출하는 것이다(Lawson et al. 1996).

템플릿은 보도자료에서 회사 및 재정에 관한 사실정보, 과학논문초록, 새로운 생산물 정보요약, 특허에서 화학표시처리, 특허에서 서지정보 등 주로 특정주제분야의 자연언어텍스트에서 데이터를 추출하는데 이용되어 왔는데,

[1] Cho, M. R., D. W. Knowles, B. L. Smith, J. J. Moulds, P. Agre, N. Mohandas, and D. E. Golan. 1999. Membrane dynamics of the water transport protein Aquaporin-1 in intact human red cells. Biophys. J. 76:1136-1144.

↓

Authors : M.R.Cho:D.W.Knowles:B.L.Smith:J.J.Moulds:P.Agre:N.Mohandas:D.E.Golan
First Author : M.R.Cho
Journal : BIOPHYS.J.
Volume : 76
Start Page : 1136
Year : 1999

〈그림 2〉 참고문헌에서 메타데이터 추출(OpCit의 사례)

특히 디지털 형태의 문헌에서 여러 가지 종류의 정보를 추출할 때 매우 유용한 것으로 제시되었다.

예컨대, 온라인 뉴스 추출 시스템인 SCISOR은 온라인 뉴스에서 기업합병에 관한 정보를 추출하고(Jacobs and Latt 1990), JASPER는 뉴스 중 회사보도에 관한 텍스트에서 주요 사실정보를 추출하며(Andersen and Huettner 1994), LOLITA는 회사와 관련한 템플릿에 따라 재무정보를 추출하는 시스템으로, 특히 이 시스템은 이용자가 템플릿을 정의하여 새로운 템플릿을 구성할 수 있도록 하고 있다(Constantino et al. 1996). 또한 Paice는 'indicator phrases'라는 템플릿 마이닝을 이용하여 과학논문의 주제와 결과를 추출하여 자동으로 초록을 만드는 것을 제시하였고(Paice 1981), Lawson 등은 영어로 된 특허전문에서 서지정보를 자동으로 분리하여 추출하기 위해 템플릿 마이닝 방법을 이용하였다(Lawson et al 1996).

Chowdhury는 디지털 형태의 문서에서 가치 있는 정보를 추출하기 위한 방법으로 템플릿 마이닝을 제시하고, 이를 적용할 수 있는 분야를 다음과 같이 제시하였다(Chowdhury 1999).

- 디지털 문서의 인용 데이터베이스의 자동생성
- 전자잡지의 새로운 항목의 정보를 자동추출
- 연구의 후원자 정보 자동확인
- 메타데이터와 템플릿 마이닝을 이용한 정보추출

그는 디지털 문서의 인용 데이터베이스의 자동생성은 온라인 논문에서 SCI 데이터베이스와 유사한 인용데이터베이스를 자동으로 개발하는데 템플릿 마이닝이 사용될 수 있음을 제시하고, 1998년 출판된 전자잡지 D-Lib Magazine의 9개 호의 논문을 분석하여 〈그림 3〉과 같은 일반적인 기사구조를 조사하여 템

Journal title	D-Lib Magazine
Issue Date	Month Year
ISSN	ISSN 1082-9873
Title	
Author	
e-mail	
Abstract	
Keywords	
Text	
References	Referennces/Bibliography/Notes
Acknowledgments	Acknowledgments

〈그림 3〉 D-Lib Magazine의 논문기사의 일반적인 구조

플릿으로 만들 수 있음을 보여주었다. 이 그림에서 굵은 글씨는 텍스트가 일정함을 나타내고, 빈칸은 논문에 주어진 값이 있음을 의미하며, Issue Date 슬롯은 '월, 년' 형식으로 나타낸다. 또한 Reference 슬롯은 저자마다 사용된 표제가 다양함을 보여주고 있으며, 아울러 저자마다 여러 가지 인용스타일과 많은 불규칙을 가지고 있어 추출에 다소 어려움이 있음을 기술하였다.

전자잡지의 새로운 항목의 정보추출은, D-Lib Magazine의 경우, "Clips & Pointers" 이라는 항목에서 회의, 세미나 및 워크숍 등에 대한 정보를 제공하는 "Goings on"이라는 부분을 템플릿을 이용하여 개최 예정인 세미나 및 회의 등에 대한 정보를 추출하는 것이다. 또한 연구의 후원자 정보는 논문의 "Acknowledgment"에서 후원기관의 이름, 주소, 허가번호 등에 대한 정보를 추출하는 것이며, 메타데이터와 템플릿 마이닝을 이용한 정보추출은 전자문헌의 자동색인을 위해 Dublin Core와 같은 메타데이터 요소를 확인하여 템플릿을 만들어 메타데이터를 자동으로 추출하는 것을 의미한다.

3. 참조연결시스템 관련 연구

Cameron은 지금까지 출판된 모든 학술논문을 연결하여 인터넷을 통해 탐색 및 입수가 가능한 완전한 서지 및 인용데이터베이스의 구축에 대한 개념을 제안하였다(Cameron 1997). 그러나, 그는 다양한 형태의 문헌에 대한 구체적인 개념을 제시하지 않았을 뿐만 아니라 모든 문헌의 인용정보를 저자나 관리자가 수작업으로 관리하도록 하여, 단지 이용자 입장에서 문헌내의 인용된 문헌을 바로 입수하기를 원한다는데 초점을 둔 듯 하다.

실제적으로, 참조연결은 인용정보를 자동으로 추출하여 접근 가능하도록 하는 인용색인 과정을 완전히 자동화하는 것을 의미하는 것으로, 최근에 온라인으로 입수 가능한 학술문헌이 증가함에 따라 더욱 연구가 되고 있는 분야다. 참조연결은 디지털 형태의 문헌에서 인용색인을 자동으로 만들기 구축하기 위해 온라인상의 논문을 자동으로 찾아 인용정보를 추출하고, 관련내용을 바로 확인할 수 있게 하는 것이다.

참조연결의 구조는 먼저 인용한 문헌의 메타데이터와 참고문헌의 메타데이터를 추출하기 위해 대상 문헌을 분석하여 참조연결 데이터를 수집하여 저장하여 이들을 상호연결하는 것으로 구성된다.

최근 참조연결을 위한 시스템들이 특정 분야에서 개발되었는데, 그 사례는 NASA Astrophysics Data Systems, National Library of Medicine's PubMed/PubRef System 등이 있고, 상업적으로 대표적인 것인 ISI의 Web of Science(Atkins 1999)가 있으며, 국제 DOI 재단은 URN형식의 DOI를 이용한 시스템을 제시하였다(Paskin 1999).

또한 참조연결을 위한 프로젝트로서 CrossRef, ResearchIndex, OpCit 등이 있는데, 이들 프로젝트는 주로 텍스트의 인식, 일치, 연결 및 웹 상에서 인용된 문헌을 찾을 수 있는 소프트웨어에 관심을 두고 있으며, 그 내용은 다음과 같다.

3. 1 CrossRef

153개 학술 출판사들이 참여하고 있는 PILA(Publishers International Linking Association)가 운영하는 CrossRef는 약 6,400여 개 잡지의 490만 논문레코드를 보유하고 있으며, 참여 출판사가 제공하는 논문의 메타데이터를 공유하여 참조연결하고 있다. CrossRef는 저장소 이름, 구분자, 그 논문의 유일한 이름으로 구성된 DOI(Digital Object Identifiers)를 통해 참조연결을 제공하여, 이용자가 링크를 클릭 하면 그 논문의 전체 인용서지를 보여주는 출판사의 웹사이트의 페이지에 연결되어 바로 접근하는 서비스를 제공하고 있다.

1997년 AAP(Association of American Publishers)가 개발한 DOI는 변하지 않는 확실한 식별자(identifier)를 통해 인터넷상의 문헌을 찾을 수 있도록 하는 것인데, 예컨대, D-Lib 논문의 DOI는 10.1045/december99-miller와 같이 나타난다. 그러나, DOI는 아직 표준으로 받아들여지지 않은 상태이며 또한 개인 홈페이지와 같은 많은 온라인 문헌에 대해 DOI가 없다는 문제가 있다.

3. 2 ResearchIndex(CiteSeer)

ResearchIndex(Bollacker 1998; Giles 1998; Lawrence 1999)는 CiteSeer의 변경된 이름으로, 컴퓨터 분야의 대규모 온라인 인용정보 데이터베이스를 자동으로 만드는 참조연결시스템이다. 이것이 여러 시스템과 다른 점은 조직되거나 통제되지 않은 웹에서 적합한 문헌을 찾아 다운로드 하여 텍스트로 변환한 다음, 인용한 문헌과 그 논문의 본문에서 만들어진 문맥을 구문분석 하여 추출한 메타데이터를 데이터베이스에 저장하여 참조연결을 제공다는 점이다. 또한 Research Index는 같은 논문에 대한 여러 가지 변형된 인용정보를 확인하여 묶어주는 알고리즘을 포함하고 있으며, 특정 논문이 이후의 다른 논문에서 그 인용되는 문맥을 제시하고 있다.

특히 ResearchIndex는 ISI의 인용색인처럼 미리 선택된 학술잡지에만 국한하지 않고 학술잡지의 논문 뿐만 아니라 회의논문, 출판전 기사, 기술보고서 등을 색인하여 어떤 논문을 인용한 논문을 유형에 관계없이 바로 인용연결을 통한 검색을 제공하고 있으며, 어떤 논문이 다음의 출판에서 어떻게 인용되는지에 대한 인용문맥을 제시함과 동시에 인용문맥을 이용하여 논문의 상호작용적인 브라우징과 논문의 간략한 요약을 제공한다. 또한 ResearchIndex는 인용횟수에 근거한 논문, 저자 및 저널 등의 평가와 순위와 연구경향을 파악할 수 있도록 하고, 일반적인 인용정보와 단어벡터 유사도를 이용하여 논문이 서로 어느 정도 관련 있는지를 파악하는 등 전통적인 인용색인에 비해 많은 장점을 가지고 있다.

ResearchIndex 시스템에서 인용정보추출 부분은 문서위치파악과 문서처리 부분으로 구분된다. 문서의 위치파악은 주로 웹을 탐색하고, 메일링 리스트나 뉴스그룹을 체크하거나 직접 출판자와 연결함으로써 논문기사를 발견하는데, ResearchIndex는 웹을 탐색하는데 주로 휴리스틱을 이용한다. 예컨대, 단어 중 "publications," "paper," "postscript"를 포함

하는 페이지를 탐색하고, PDF 파일일 경우 텍스트로 변환하여 reference나 bibliography 부분이 존재하는지를 테스트하여 그 문서가 연구논문인지를 확인한다. 문서처리는 논문에서 인용된 문헌의 리스트를 확인한 다음, 개별 서지사항을 추출하여 기술하며, 저자, 논문명, 출판년, 페이지 등을 추출하기 위해 휴리스틱을 이용하여 각 서지사항을 분석한다. 예컨대, 본문에 있는 인용정보를 찾기 위해 "[6]," "[Giles97]," 또는 "Marr 1982"와 같은 인용사항을 이용한 후, 인용문맥을 추출한다.

3. 3 Open Journal(OpCit)

영국의 Electronic Libraries(eLib) programe의 지원 하에 수행된 Open Journal은 학술잡지의 논문을 참조연결하기 위한 프로젝트로서(Hitchcock et al. 1998), 그 목적은 ISI의 인용색인에 근거하여 ResearchIndex 방법과 유사하게 새로 입수된 문헌에서 인용정보를 추출하는 인용 소프트웨어를 개발하여 전자잡지에서 하이퍼링크를 자동으로 제공하여 잡지 및 다른 웹 자원에 빠른 접근을 지원하는 것이다.

OpCit는 1998년 마감한 Open Journals의 후속 프로젝트로서, 전자잡지를 포함한 전자문헌 모두를 대상으로 한 Open Archives의 참조연결을 위한 소프트웨어를 개발하는 프로젝트이다(Hitchcock et al. 1998; Hitchcock et al. 2000).

Open Journal은 전자잡지내의 논문간의 명백한 관계를 만들어 본문의 인용과 문헌의 끝의 참고문헌을 링크 한 다음, 외부의 서지데이터베이스와 참조링크를 하여 이용자가 직접 본문에서 다른 소스로 링크로 연결하는 것을 다룬다. Open Journal의 방법은 html 형식의 원 문헌에 링크 데이터를 삽입하는 대신, DLS(Distributed Link Service)라는 소프트웨어를 이용하여 그 링크 데이터를 'linkbase'라는 링크 데이터베이스에 각각 저장하고, 웹에서 보여질 때 링크데이터가 본문에 첨가되어 보여지도록 하는데, 그 사례는 〈그림 4〉와 같다(Hitchcock et al. 1997).

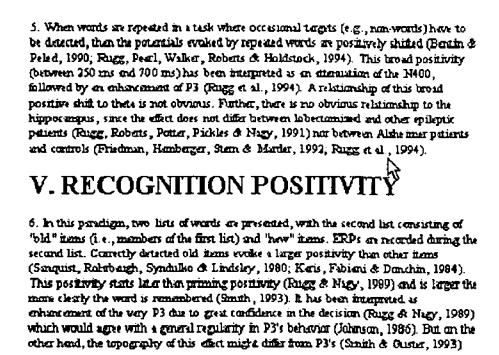
4. 템플릿 구성

이 논문은 참조연결을 위한 선행작업으로서, 템플릿 마이닝을 이용하여 인쇄형태 및 디지털형태의 연구논문에서 인용정보를 자동으로 추출하여 인용데이터베이스를 구축에 대한 연구이다. 이를 위해, 이 연구에서는 2001년에 발행된 문헌정보학 분야의 3개의 학술잡지 9권을 선택하여 샘플로서 사용하였으며, 선택된 학술잡지는 다음과 같다.

- 한국문헌정보학회지. 제35권 1-3호
- 정보관리학회지. 제18권 1-3호
- 한국도서관·정보학회지 제32권 1-3호

선택한 학술잡지가 비록 전자형식으로 출판되어 있지는 않지만, 많은 저자들이 특정 사이트에 자신의 논문을 올리고 있기 때문에, 앞서 고찰한 참조연결과정을 고려하여 샘플 논문들이 디지털화되어 있다는 가정을 전제로 한다.

이들 학술잡지의 각 권은 4개의 호로 구성되어 있는데, 이중 각 권의 1-2호 6개는 템플



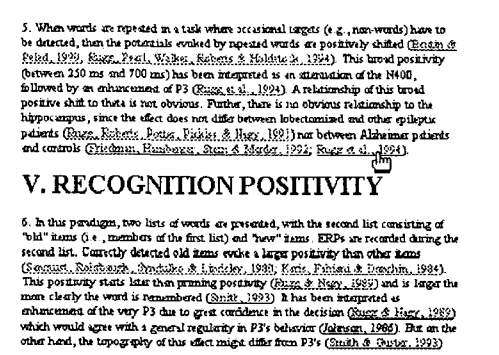
IV. PRIMING POSITIVITY

5. When words are repeated in a task where occasional targets (e.g., non-words) have to be detected, then the potentials evoked by repeated words are positively shifted (Bentin & Peled, 1990; Rugg, Pearl, Walker, Roberts & Holdstock, 1994). This broad positivity (between 250 ms and 700 ms) has been interpreted as an attenuation of the N400, followed by an enhancement of P3 (Rugg et al., 1994). A relationship of this broad positive shift to theta is not obvious. Further, there is no obvious relationship to the hippocampus, since the effect does not differ between lobectomized and other epileptic patients (Rugg, Roberts, Potter, Pickles & Nagy, 1991) nor between Alzheimer patients and controls (Friedman, Hamberger, Stern & Marder, 1992, Rugg et al., 1994).

V. RECOGNITION POSITIVITY

6. In this paradigm, two lists of words are presented, with the second list consisting of "old" items (i.e., members of the first list) and "new" items. ERPs are recorded during the second list. Correctly detected old items evoke a larger positivity than other items (Sanquist, Rohrbaugh, Syndulko & Lindsley, 1980; Karis, Fabiani & Donchin, 1984). This positivity starts later than priming positivity (Rugg & Nagy, 1989) and is larger the more clearly the word is remembered (Smith, 1993). It has been interpreted as enhancement of the very P3 due to great confidence in the decision (Rugg & Nagy, 1989) which would agree with a general regularity in P3's behavior (Johnson, 1986). But on the other hand, the topography of this effect might differ from P3's (Smith & Guster, 1993)

a. 링크가 없는 원래 논문



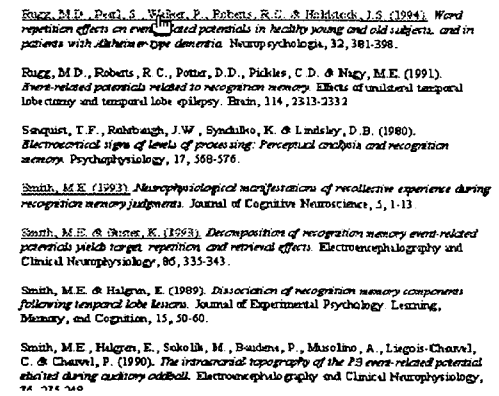
IV. PRIMING POSITIVITY

5. When words are repeated in a task where occasional targets (e.g., non-words) have to be detected, then the potentials evoked by repeated words are positively shifted (Bentin & Peled, 1990; Rugg, Pearl, Walker, Roberts & Holdstock, 1994). This broad positivity (between 250 ms and 700 ms) has been interpreted as an attenuation of the N400, followed by an enhancement of P3 (Rugg et al., 1994). A relationship of this broad positive shift to theta is not obvious. Further, there is no obvious relationship to the hippocampus, since the effect does not differ between lobectomized and other epileptic patients (Rugg, Roberts, Potter, Pickles & Nagy, 1991) nor between Alzheimer patients and controls (Friedman, Hamberger, Stern & Marder, 1992; Rugg et al., 1994).

V. RECOGNITION POSITIVITY

6. In this paradigm, two lists of words are presented, with the second list consisting of "old" items (i.e., members of the first list) and "new" items. ERPs are recorded during the second list. Correctly detected old items evoke a larger positivity than other items (Sanquist, Rohrbaugh, Syndulko & Lindsley, 1980; Karis, Fabiani & Donchin, 1984). This positivity starts later than priming positivity (Rugg & Nagy, 1989) and is larger the more clearly the word is remembered (Smith, 1993). It has been interpreted as enhancement of the very P3 due to great confidence in the decision (Rugg & Nagy, 1989) which would agree with a general regularity in P3's behavior (Johnson, 1986). But on the other hand, the topography of this effect might differ from P3's (Smith & Guster, 1993)

b. DLS에 의해 부가된 인용링크



Rugg, M.D., Pearl, S., Walker, P., Roberts, R.C. & Holdstock, J.S. (1994). *Word repetition effects on event-related potentials in healthy young and old subjects, and in patients with Alzheimer-type dementia.* Neuropsychologia, 32, 381-398.

Rugg, M.D., Roberts, R.C., Potter, D.D., Pickles, C.D. & Nagy, M.E. (1991). *Event-related potentials related to recognition memory: Effects of unilateral temporal lobectomy and temporal lobe epilepsy.* Brain, 114, 2313-2332

Sanquist, T.F., Rohrbaugh, J.W., Syndulko, K. & Lindsley, D.B. (1980). *Electrocortical signs of levels of processing: Perceptual analysis and recognition memory.* Psychophysiology, 17, 568-576.

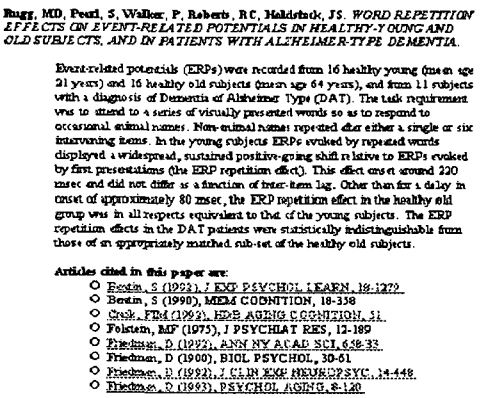
Smith, M.E. (1993). *Neurophysiological manifestations of recollective experience during recognition memory judgments.* Journal of Cognitive Neuroscience, 5, 1-13

Smith, M.E. & Guster, K. (1993). *Decomposition of recognition memory event-related potentials yields target, repetition, and retrieval effects.* Electroencephalography and Clinical Neurophysiology, 86, 335-343.

Smith, M.E. & Halgren, E. (1989). *Dissociation of recognition memory components following temporal lobe lesions.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 50-60.

Smith, M.E., Halgren, E., Sokolik, M., Baudena, P., Musolino, A., Liegois-Chauvel, C. & Chauvel, P. (1990). *The intracranial topography of the P3 event-related potential elicited during auditory oddball.* Electroencephalography and Clinical Neurophysiology,

c. 참고문헌에 이어진 인용링크



Rugg, MD, Pearl, S, Walker, P, Roberts, RC, Holdstock, JS. *WORD REPETITION EFFECTS ON EVENT-RELATED POTENTIALS IN HEALTHY YOUNG AND OLD SUBJECTS, AND IN PATIENTS WITH ALZHEIMER-TYPE DEMENTIA.*

Event-related potentials (ERPs) were recorded from 16 healthy young (mean age 21 years) and 16 healthy old subjects (mean age 64 years), and from 11 subjects with a diagnosis of Dementia of Alzheimer Type (DAT). The task requirement was to attend to a series of visually presented words so as to respond to occasional animal names. Non-animal names repeated after either a single or six intervening items. In the young subjects ERPs evoked by repeated words displayed a widespread, sustained positive-going shift relative to ERPs evoked by first presentations (the ERP repetition effect). This effect onset around 220 msec and did not differ as a function of inter-item lag. Other than for a delay in onset of approximately 80 msec, the ERP repetition effect in the healthy old group was in all respects equivalent to that of the young subjects. The ERP repetition effects in the DAT patients were statistically indistinguishable from those of an appropriately matched sub-set of the healthy old subjects.

Articles cited in this paper are:

- Bentin, S (1992), J EXP PSYCHOL LEARN, 18:1279
- Bentin, S (1990), MEM COGNITION, 18-358
- Craik, FIM (1992), HDB AGING COGNITION, 51
- Folstein, MF (1975), J PSYCHIAT RES, 12-189
- Friedman, D (1992), ANN NY ACAD SCI, 658-33
- Friedman, D (1990), BIOL PSYCHOL, 30-61
- Friedman, D (1992), J CLIN EXP NEUROPSYC, 14:448
- Friedman, D (1993), PSYCHOL AGING, 8-120

d. ISI 초록에 이어진 참고문헌연결

〈그림 4〉 Open Journal에서의 참조연결 사례

릿을 만들기 위해 사용하였으며, 3호는 그 템플릿의 유용성을 테스트하기 위해 사용하였다.

템플릿을 만들기 위해 선택된 잡지에 포함된 학술논문은 총 75개인데, 이중 참고문헌을 각주로 대신한 5건의 논문을 제외한 70건을 대상으로 조사하였다. 이들 논문이 포함하고 있는 전체 참고문헌 수는 1665개로, 논문 1건당 평균 23.8개의 참고문헌을 포함하고 있으며, 참고문헌 중 한글문헌은 687개, 영어문헌은 899개, 일어문헌은 20개, URL은 59개 조사되었다.

먼저 논문에서 인용정보를 추출하기 위해 샘플로 선택한 75개의 논문을 분석하여 나타난 패턴을 확인하여 두 개의 템플릿을 작성하였다. 하나는 인용한 논문 템플릿으로, 학술잡지에 게재된 논문의 서지정보를 대상으로 하였고, 또 하나는 인용된 논문 템플릿으로, 게재된 논문의 참고문헌에 있는 서지정보를 대상으로 하였다. 참고문헌의 서지정보는 보편적으로 학술잡지 발행기관이 내부편집규정을

정하여 저자가 따르도록 권고하고 있으며, 논문이 제출되면 편집자가 체크와 수정을 하기 때문에 다소 엄격하게 지켜지는 것으로 인식되고 있으나, 그 규정을 따르지 않아 많은 변형이 있는 것으로 조사되었다.

4. 1 인용한 논문 템플릿

인용한 논문 템플릿을 작성하기 위해, 각 논문에서 수집한 정보는 잡지정보, 논문명, 저자명, 저자소속, 전자우편, 초록, 키워드, 참고문헌 등이다. 인용한 논문은 이후에 다른 논문에서 인용되기 때문에 여기서 정확한 패턴으로 서지사항을 조사해 놓으면 나중에 인용된 논문과 정확한 연결을 보장할 수 있다. 이 연구에서 인용한 논문의 서지사항은 대체로 논문의 본문에 나타난 모든 내용이 디지털화되었을 때도 그대로 있다는 가정 하에 조사한 것이다. 3개의 학술지 샘플논문에서 조사한 인용한 논문의 템플릿에 대한 단서정보는 〈표 1〉과 같다. 물론 이들 논문들이 디지털화되었을 때는 이들 단서정보뿐만 아니라 휴리스틱스를 동시에 이용하는 것이 효과적일 것이다. 예컨대, 만약 논문이 HTML형식이라면, 논문명의 글자크기가 다른 것보다 크기 때문에 파악하기 쉽지만, 만약 PDF와 같은 형식을 텍스트 형식으로 변환했을 때는 휴리스틱스가 이용되지 않을 것이다.

이러한 단서정보에 근거하여 작성하는 인용한 논문의 템플릿 마이닝 알고리즘은 〈표 2〉와 같다.

4. 2 인용된 논문 템플릿

인용된 논문 템플릿은 앞의 인용한 논문 템플릿 마이닝에서 '참고문헌' 항목이 있으면 호출된다. 참고문헌은 일반적으로 인용한 논문의 마지막에 제시되는데, 그 인용패턴이 단

〈표 1〉 인용한 논문 템플릿의 단서정보

정보단위	단서정보	구두점
잡지명	a; a · a	
권	'제d권'	'('
호	'제d호'	')'
출판년	1900-2002	
논문명	(한줄 또는 여러줄) a ; a:a ; a+-a+-	줄바꿈
감사표시	(논문명끝의 * 와 주석연결) *+a	*
저자명	(영어명칭 무시) a	줄바꿈; '·'; ','
저자소속	(저자명끝의 * 갯수와 주석연결) a	*
전자우편	(저자소속에서) '@'	공란
초록	'초록', 'abstracts'	줄바꿈
키워드	'키워드', 'keywords'	줄바꿈; ':'; ','
참고문헌	'참고문헌', 'References'	공란

(a: 문자열, c:문자, d: 숫자, +:바로 뒤따름, - ' '안의 문자: 내용이 일정함을 의미)

〈표 2〉 인용한 논문의 템플릿 마이닝 알고리즘

```
START:
    IF 머리말에(저널정보) 존재
        EXTRACT 잡지명
        EXTRACT 출판년
        EXTRACT 권
        EXTRACT 호
        EXTRACT 페이지
    ENDIF
    EXTRACT 논문명
    IF(저자정보) Exist
        WHILE(.NOT. 마지막 공저자)
            EXTRACT 저자명
                IF 저자명 끝에 참조연결점(*) 존재
                    EXTRACT 소속
                    EXTRACT 전자우편
                ENDIF
        END WHILE
    ENDIF
    IF(키워드)
        EXTRACT 키워드
    ENDIF
    IF(초록)
        EXTRACT 초록
    ENDIF
    IF(Abstracts)
        EXTRACT abstracts
    ENDIF
    IF(참고문헌)
        CALL 인용된 논문 템플릿
    ENDIF
END
```

순한 것으로 보이지만 자세히 보면 대단히 불규칙하게 나타나고, 저자마다 여러 가지 다른 인용표현법을 이용하고 있음을 볼 수 있어 인용한 논문 템플릿을 작성하는 것 보다 상당히 어려운 부분이다. 특히 온라인 문헌이나 신문, 인터뷰 및 학위논문 등도 포함되어 있는데, 이 논문에서는 신문이나 인터뷰는 생략하고, 학위논문은 같은 패턴으로 사용하는 것으로 가정하고, 나머지 참고문헌의 인용패턴을 분석하여 〈표 3〉과 같은 인용된 논문의 템플릿을 위한 단서정보를 조사하였다.

① 저자명

저자명은 일반적으로 참고문헌의 제일 앞부분에 〈표 4〉와 같이 여러 가지 변형되어 나타나긴 하지만, 몇 가지 고정된 패턴을 유지하고 있는데, 이 부분은 주로 구두점을 이용하여 파악된다.

한글 개인명의 경우, 성과 명의 순서대로 나타나고, 공저인 경우 구두점을 이용하여 구분하여 나타내며, 또한 영어처럼 이름에 약어가 사용되지 않기 때문에 영어보다 파악하기 쉽다. 영어 개인명의 경우, 성과 명을 순서대

〈표 3〉 인용된 논문 템플릿의 단서정보

정보단위	단서정보	구두점 및 기호
저자명	(한) a; a+등; a+등편; a+외; a+역; a+외역; a+저; a+공역; a+지음; a+(+편+); a+옮김; (영) a; a,+c.; a,+c.+c.; a,+a+c; a,+c.; a,+a; a+a; a+a+a; a+a+c.; a+c.+a; a+a; c.+c.+a; c.+a; a+ +et al.; a+ +ed.; a+ +(+Ed.+)	(한) ','; '.'; ';'; ' '; '·' (영) ','; ';'; ' '; 'et al.'; 'and'; 'ed.'
출판년	(한) 1900-2002+d(월은 무시) (영) d.+ 1900-2002; a,+ 1990-2002(월 또는 월명은 무시)	'.'; '('; ','
논문명	(한) a; a+:+a; a+-a; a+-a-; (영) a; a:a 도서명은 아래의 표제에서 다름	(한) '.'; '"'; '.'; '".'; '",' (영) '"'
잡지명	(한) a; a · a (영) a;	(한) ' '; '.'; '「'; '」'; '『'; '』'; '≪'; '≫' (영) '.'; ':'
도서, 학위논문, 보고서 등의 표제	(한) a; a+:+a (영) a; a+:+a; 'In+a'; 'in+ a'; in +저자명+(eds.또는 ed)+s, 'Proceedings+a' 단행본의 경우 판사항 포함(표제와 출판사항 사이)	(한) ' '; '.'; ','; '「'; '」'; '『'; '』' (영) '␣'
출판사항	(한) a(출판지)+:+a(출판사); a(출판사); a:a,d(d는 출판년) (영) a+:+a; a+,+a	(한) ':'; '.'; '['; ']'; ','; '(주); '(' ; ')' (영) ':'; ','
Vol	(한) d; d+:; d+(; 통권d회; 제d집; d집; 제d권; 통권d; 창간호 (영) 'Volume'; 'Vol.'; 'V.'; d; d+(; d+:	(한) ' '; ','; ':' (영) ' '; ','; ':'
Issue	(한) (+d+); 제d호; .+d; :+d (영) 'Issue'; 'No.'; 'no.'; (+d+); (+d+/+d+); :+d;	'('; ')'; '.'; ','
페이지	'pp.'; 'p.'; ':+d-d'; 'd+:+d'; d-d; d	'pp.'; 'p.'; ':'
url	(한) 'http' (영) 'http'; 'ftp'; 'available'	'<'; '>'; '('; ')'

(a: 문자열, c:문자, d: 숫자, +:바로 뒤따름, - ' ' 안의 문자: 내용이 일정함을 의미)

로 나타내기 위해 쉼표로 구분하고 있으며, 공저인 경우 첫 번째 저자는 성명이 순서대로 나타나지만, 두 번째 저자부터는 성과 명이 순서대로 제시된 경우도 있고, 그렇지 않은 경우도 있어 다소 파악하기 어렵다.

한편, 저자명이 없어 표제가 바로 나타나는 경우는 일반적인 패턴으로 바로 파악하기 어렵기 때문에 알고리즘에서 휴리스틱스를 이용하는 것이 좋을 것이다. 또한 공저자일 경우, 동일 저자가 연이어 있는 경우, 두 번째 참고

〈표 4〉 저자명 기술 유형

- 이용남, 홍현진. 1999. ...
- 한국교육학술정보원. 2000. ...
- 박준식. 1988. ...
- _____. 1996. ...
- Argyri, C. 1964. ...
- Cameron, Kim and D. A. Whetten. 1983. ...
- Kirriemuir, J. et al. 1998. ...

문헌에서 저자명 대신 '_____'와 같이 표현하므로 앞의 저자를 참조하도록 하고, 참고문헌 앞에 [1]과 같이 번호를 붙이거나 다른 분야의 논문에서 나타나는 (Harter, 1999)와 같은 것은 미리 배제한다.

② 출판년도

출판년도는 네 개의 연속된 정수로 나타나기 때문에 쉽게 확인할 수 있지만, 그 위치가 일정하게 나타나지 않아 위치만으로 파악하기는 쉽지 않다. 예컨대, 대부분의 참고문헌에서 출판년도는 저자명 다음에 나타나지만, 참고문헌의 중간 또는 끝에 나타나기도 한다. 그러나 출판년도가 어디에 나타나든지 간에 네 자리 정수로 파악할 수 있으며, 출판년도와 같이 나타나는 월(e.g. 1999. 3 또는 March 1999)은 배제한다.

③ 논문명

논문명은 대체적으로 출판년도나 저자명 다음에 나타나는데, 따옴표로 묶거나 아니면 따옴표 없이 나타난다. 따옴표로 묶이는 경우는 파악하기 쉬우나, 그렇지 않은 경우는 이 논문이 학술잡지, 도서, 회의, 보고서 등에 포함된 것과 구분하기 위해 뒤에 나오는 정보를 먼저 파악한 다음 고려해야 한다. 특히 도서명이 바로 나오는 경우는 바로 다음에 출판사항의 유무를 판단하여 파악한다.

④ 잡지명

잡지명은 일반적으로 논문명 다음에 나타나며, 샘플논문에서 한글의 경우, 각괄호 등의 기호로 묶는 경우가 많으며, 영어는 이텔릭체로 기술하고 있어 휴리스틱을 동시에 이용하여 파악한다.

일반적으로 학술잡지의 경우, 잡지명은 논문명 다음에 나타나 비교적 파악하기 쉬운 편이지만, 약어가 포함된 경우에는 학술잡지명 데이터베이스가 따로 구축하여 이를 비교하여 전체 잡지명이 나타나도록 할 필요가 있을 것이다.

⑤ 도서, 학위논문, 회의, 보고서 등의 표제

참고문헌에서 도서나 회의자료는 출판년도 다음에 바로 나타나는 경우도 있지만, 특정 논문의 출처를 나타내기 위해 논문명 다음에 제시되기도 한다. 도서나 회의자료내의 논문이 아닌 일반단행본 도서의 경우, 표제 다음에 출

판사항이 뒤따르므로 이를 통해 파악하며, 특정 문헌내의 하나의 논문인 경우는 다음과 같이 'In'과 같은 단서정보나 논문명 다음에 다시 저자명이 뒤따르는 경우가 대부분이다.

Ferguson, N. 1995. "Subject-based Services: Orgines and Futures." in *Proceedings of the UK Office for Library and Information Networking(UKOLN) Conference: Networking and the future of Libraries 2*: 131-135.

⑥ 출판사항

출판사항은 주로 위의 도서, 학위논문, 회의, 보고서 등의 표제 다음에 나타나는데, 일반적인 패턴은 '출판지:출판사', '출판사', '출판지: 출판사, 출판년'의 형식이다. 따라서 위의 자료의 표제 다음에 이러한 패턴에서 출판지와 출판사에 대한 정보를 추출한다. 한편 출판지가 두 개 이상인 경우는 구두점을 경계로 모두 추출하며, 출판사에 대해 약어를 사용한다면 출판사 데이터베이스를 따로 구축하여 비교할 필요가 있을 것이다. 출판년는 앞서 파악되었으므로 무시한다.

⑦ 권·호

잡지의 권·호는 일반적으로 잡지명 다음에 나타나는데, 위에 제시한 단서정보로 쉽게 파악할 수 있다. 특히 한글은 영어에 비해 다양한 형태로 나타나는데, '창간호'와 같은 것은 1호와 별개로 되지 않게 관리할 필요가 있다.

⑧ 페이지

페이지는 잡지의 권·호 다음이나 단행본의 출판사항 다음에 나타나는데, 위에 제시한 단서정보로 파악한다.

⑨ URL

최근 논문의 인용은 온라인 문헌의 이용빈도가 증가하고 있다. 온라인 문헌은 인터넷의 URL을 이용하는데, 이 또한 위에 제시한 단서정보에 근거하여 쉽게 파악할 수 있다.

위의 여러 가지 인용패턴을 통해 구축한 템플릿에 근거한 인용된 논문 템플릿 마이닝 알고리즘은 〈표 5〉와 같다.

4. 3 템플릿 평가

앞서 제시한 템플릿의 효율성을 평가하기 위해 2001년 발행된 3개의 학술잡지 각 3호 3권에 수록된 논문 41개 논문 중 참고문헌을 각주로 대신하는 5개를 제외한 36건의 논문과 그 논문들의 참고문헌에 수록된 인용된 문헌 910개를 제시한 템플릿에 근거한 알고리즘에 의거하여 수작업으로 분석하였다. 이들 인용된 문헌 중 인쇄된 참고문헌은 754개로 83%이고, 온라인으로 입수한 참고문헌은 156개로 약 17%를 차지하는 것으로 나타났다.

1) 인용한 논문 템플릿 평가

인용한 논문 템플릿에 근거한 인용정보는 현재의 인쇄본이 디지털형태로 되더라도 형태를 그대로 유지한다는 가정 하에서 볼 때 단서정보에 고정되어 있기 때문에 추출비율은 거의 100% 이를 것으로 여겨진다. 예컨대 키워드에

〈표 5〉 인용된 논문의 템플릿 마이닝 알고리즘

```
START:
    IF (저자정보)
        WHILE (.NOT. 마지막 공저자)
            EXTRACT 저자명
        END WHILE
    ENDIF
    IF(출판년)
        EXTRACT 출판년
    ENDIF
    SWITCH(학술잡지/단행본/회의/전자저널/전자문서)
    CASE 학술잡지:
        EXTRACT 논문명
        EXTRACT 잡지명
        EXTRACT 권
        EXTRACT 호
        EXTRACT 페이지
    CASE 단행본:
        EXTRACT 서명
        EXTRACT 출판지
        EXTRACT 출판자
        EXTRACT 출판년
    CASE 회의:
        EXTRACT 서명
        EXTRACT 회의명
        EXTRACT 회의주소
        EXTRACT 페이지
    CASE 전자저널:
        EXTRACT 표제
        EXTRACT 전자저널명
        IF 권
            EXTRACT 권
        ENDIF
        IF 호
            EXTRACT 호
        ENDIF
        EXTRACT URL
    CASE 전자문서:
        IF 표제
            EXTRACT 표제
        ENDIF
        EXTRACT URL
    END SWITCH
END
```

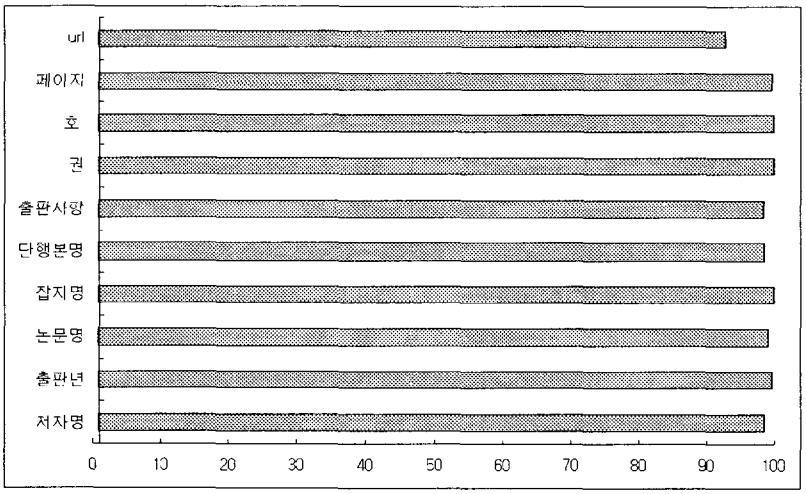
있어서, 시스템은 '키워드' 또는 'Kewords'와 같은 단서정보를 텍스트에서 찾아 분석하면, 그 부분의 모든 정보가 키워드 항목에 채워질 것이다. 즉 고정된 단서정보를 가진 정보항목은 높은 추출빈도를 보여준다.

2) 인용된 논문 템플릿 평가

인용한 논문에 비해 인용된 논문은 많은 변형을 포함하고 있기 때문에 템플릿에 근거한 추출율이 상대적으로 떨어질 것으로 보여진다. 비록 각 학회마다 편집위원회에서 인용문의 기술형식과 범례를 제공하고 있지만, 저자들이 이를 정확히 따르고 있지 않을 뿐만 아니라 같은 논문에서 여러 인용형식을 이용하고 있는 것으로 조사되었다. 또한 편집상 실수로 구두점이나 기호가 누락되거나 잘못 사용되는 경우도 있었다. 비록 이러한 많은 문제점이 있음에도 불구하고 템플릿에 근거하여 인용된 문헌 911개에 대한 분석 결과 〈그림 5〉와 같이 앞서 인용된 논문 템플릿의 평가 결과는 만족스런 수준으로 나타났으며, 실패사례는 다음과 같다.

저자명 항목에서의 추출하는데 실패 사례는 〈표 6〉과 같이 주로 편집지침에 따르지 않아서 생긴다.

논문명 항목이나 단행본 등의 표제항목에서



〈그림 5〉 인용된 논문의 각 항목에서 추출비율

〈표 6〉 추출 실패 사례 (저자명)

- 이용숙 & 김영천 편. 교육에서의...
- _____Diss. Syracuse...
- 「... Librarians」. Ed. G. E. Gorman...
- 도서관메일링리스트. 김동명. 경기도...
- *Informaion.... Research* Edited by Oddy, R. N., ...

주로 야기되는 추출실패는 구두점을 정확하게 기술하지 않은 경우 발생하고, 특히 단행본의 경우 총서사항이 부가될 경우를 고려하지 않아 발생하며, 그 사례는 〈표 7〉과 같다.

출판사항 항목에서의 추출실패는 주로 편집지침에 따르지 않거나 부가정보를 기입함으로서 야기되며, 그 사례는 〈표 8〉과 같다.

url 항목은 인용된 문헌전체에 대한 비율이 아니고 url이 기입된 156개 인용정보에서의 비율이라 다른 항목에 비해 상대적으로 낮은 편이나 대체적으로 템플릿의 단서정보에 의해 100% 추출되나, 조사된 논문 중 하나가 〈표 9〉와 같이 url을 기술하면서 인용날짜를 계속해서 기입함으로서 추출실패로 나타났다.

〈표 7〉 추출 실패 사례 (논문명)

- 정연경. 1999. 인터넷 ... 관한 연구. 제 7회 『정보관리학회 학술대회 논문집』.
- Metacrawlers and metasearch engines, [engines, [online],...
- Irwin, Kenneth R. "Professional... with "Freeie" Librarians." *Searcher* ...
- ... . 멀티미디어... 적용. 『춘계학술발표논문집』(정보처리학회 학술발표), 2:...
- Welsch, Wolfgang. 1998. Unsere postmoderne Moderne. 2. Aufl. ...
- Fowler, C. J. and Murray, D. (1987). "Gender ... Interface," *Human-Computer Interaction: INTERAT 87: Proceeding of the 2nd IFIP Conference*: 709-714
- Fowler, W. (1980). "Cognitive ... learning." In: ... eds., Advances in Child Development and Behaviour; 15: 163-206. New York: Academic Press.

〈표 8〉 추출 실패 사례 (출판사항)

- 『Ccollection ... Librarians』. Ed. G. E. Gorman, Ruth H, Miller. Westport, Greenwood press. 3-25
- AAAI Workshop ... Systems, Madison, WI: 81-83.
- 한인규. 2001. SCI ... 현주소. 『제4차 대덕과학포럼세미나』. 2001.6.21. [대전: 롯데 대덕 호텔]
- 이순철. 1999. 『사례로 본 ... 방법론』. 삼성경제연구소. 서울.
- In: M.E. Williams ed. *Annual... Technology*, 19: 33-64. Knowledge Industry Publications, Inc. for the American Society for Information Science.

〈표 9〉 추출 실패 사례 (URL)

- Content...System. The ... Press. http://www.press.umich.edu/jep/03-04/kartchner.html2001/03/03. 02:45

5. 결 론

지금까지 인용색인은 거의 수작업으로 인쇄 출판된 학술잡지를 대상으로 만들어왔으나, 최근 전자출판과 인터넷의 발전으로 인해 디지털 형태의 학술문헌이 증가함에 따라 온라인상에서 인용한 논문과 인용된 문헌을 상호 연결하여 바로 접근이 가능하도록 하는 '참조연결'에 대한 관심이 증가하고 있다.

디지털 형태의 문헌에서 인용색인을 자동으로 만드는 참조연결시스템은 온라인상의 논문을 자동으로 탐색하여 다운로드한 다음, 전체 내용을 분석하여 인용정보를 추출하고, 이를 연결시키는 과정으로 구성된다. 이중 문헌에서 인용정보를 추출하는 방법으로 패턴인식과 패턴일치에 근거한 템플릿 마이닝이 효과적인 것으로 제시되었다.

정보추출기법 중 하나인 템플릿 마이닝은 인식 가능한 패턴을 이용하여 텍스트에서 데이터를 추출하는 것으로, 텍스트가 미리 지정한 템플릿과 일치하면 그와 연관된 지시에 따라 데이터를 추출하는 것이다.

이 논문은 이러한 참조연결에 대한 내용과 기존의 시스템을 살펴보고, 학술논문을 대상으로 인용색인을 자동으로 만들기 위해 템플릿 마이닝을 이용하여 그 가능성을 보고자 프로토타입 형태의 실험을 수행하였다. 인용문헌간에 존재하는 불규칙과 비일관적인 스타일로 인해 인용정보를 자동으로 구축하기 위한 템플릿을 완전하게 구축하기는 어렵겠지만, 이 실험의 평가결과로 볼 때 가능성이 있을 것으로 판단된다. 다만, 앞으로 인용정보를 정확하고 효율적으로 추출하기 위해서는 참고문헌에 대한 엄격한 표준인용 규정을 만들어 부과하거나, 자료형태에 따라 각 인용형태를 위한 표준 템플릿을 준비하여 저자가 이를 따르도록 한다면 효과적일 것이다.

참 고 문 헌

Andersen, P. M. and Huettner, A. K. 1994. Knowledge Engineering for the JASPER Fact Extraction System. *Integrated Computer-Aided Engineering.* 1(6):473-493.

Atkins, Helen. 1999. The ISI Web of Science - Links and Electronic Journals. *D-Lib Magazine.* 5(9). 〈http://www.dlib.org/dlib/september99/atkins/09atkins.html〉.

Atkins, Helen; Lyons, Catherine; Ratner, Howard; Risher, Carol; Shillum, Chris; Sidman, David and Stevens, Andrew. 2000. Reference Linking with DOIs: A Case Study. *D-Lib Magazine.* 6(2). 〈http://www.dlib.org/dlib/february00/02risher.html〉

Bergamark, Donna. 2000. *Automatic Extraction of Reference Linking*

Information from Online Documents. Technical Report CSTR 2000-1821, Cornell University, Digital Library Research Group. 〈http://www.cs.cornell.edu/cdlrg/Reference Linking/extraction.pdf〉

Bergmark, Donna and Lagoze, Carl. 2001. *Reference Linking the Web's Scholarly Papers.* CSTR 2001-1835, Cornell University, Digital Library Research Group. 〈http://www.cs.cornell.edu/cdlrg/Reference Linking/LinkingTheWeb.pdf〉

Bollacker, Kurt D.; Lawrence, Steve and Giles, C. Lee. 1998. CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. In Katia P. Sycara and Michael Wooldridge (eds.). *Proceedings of the Second International Conference on Autonomous Agents.* New York: ACM Press. 116-123.

Cameron, R. D. 1997. A Universal Citation Database As a Catalyst for Reform in Scholarly Communication. *First Monday.* 2(4). 〈http://www.firstmonday.dk/issues/issue2_4/cameron/index.html〉

Caplan, Priscilla and Arms, William Y. 1999. Reference Linking for Jorunal Articles. *D-Lib Magazine.* 5(7/8). 〈http://www.dlib.org/dlib/july99/caplan/ 07caplan.html〉

Chowdhury, Gobinda G. 1999. Template Mining for Information Extraction from Digital Documents. *Library Trends.* 48(1): 182-208.

Constantino, M.; Morgan, R. G. and Collingham, R. J. 1996. Financial Information Extraction using Predefined User-definable Templates in the LOLITA System. *Journal of Computing & Information Technology.* 4(4):241-255.

Cronin, Blaise; Snyder, Herbert and Atkins, Helen. 1997. Comparative Citation Rankings of Authors in Monographic and Journal Literature: a Study of Sociology. *Journal of Documentation.* 53(3):263-273.

CrossRef: The Central Source for Reference Linking. 〈http://www.crossref.org〉

Ding, Ying; Chowdhyry, Gobinda and Foo, Schubert. 1999. Template mining for the extraction of citation from digital documents. *Proceedings of Second Asian Digital Libraries Conference, National Taiwan University, November 8-9, 1999.* 〈http://www.cs.vu.nl/~ying/download/adl9911.pdf〉

Hitchcock, S.; Carr, L.; Harris, S.; Hey, J.

M. N. and Hall, W. 1997. Citation Linking: Improving Access to Online Journals. In Robert B. Allen and Edie Rasmussen. (eds.). *Proceedings of the 2nd ACM International Conference on Digital Libraries.* New York: ACM. 115-122. 〈http://journals.ecs.soton.ac.uk/acmdl97.htm〉

Hitchcock, Steve. Carr, Les. Hall, Wen-dy. Harris, Steve. Probets, Steve. Evans, David and Brailsford, David. 1998. Linking Electronic Journals: Lessons from the Open Journal Project. *D-Lib Magazine.* December 1998. 〈http://www.dlib.org/dlib/december98/12hitchcock.html〉

Hitchcock, Steve; Carr, Les; Jiao, Zhuoan; Bergmark, Donna; Hall, Wendy; Lagoze, Carl and Harnad, Stevan. 2000. Developing Services for Open Eprint Archives: Globalisation, Integration and the Impact of Links. In *ACM Proceedings of Digital Libraries, San Antonio, Texas, June 2 - June 7, 2000.* 〈http://opcit.eprints.org/dl00/htdl00.pdf〉

Jacobs, P. and Ratt, L. F. 1990. SCISOR: Extracting Information from Online News. *Communications of the ACM.* 33(11):88-97.

Giles, C. Lee; Boolacker, Kurt D. and Lawrence, Steve. 1998. CiteSeer: An Automatic Citation Indexing System. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, (eds). *Digital Libraries 98 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA, June 23-26 1998.* Pittsburgh: ACM Press. 89-98.

Lawrence, Steve; Bollacker, Kurt and Giles, C. Lee. 1999. Indexing and Retrieval of Scientific Literature. *Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2-6.* 139-146.

Lawrence, Steve; Giles, C. Lee and Bollacker, Kurt. 1999. Digital Libraries and autonomous citation indexing. *IEEE Computer.* 32(6): 67-71.

Lawrence, Steve. 2001. Online or invisible?. *Nature.* 411(6837):521.

Lawson, M.; Kemp, N.; Lynch, M. F. and Chowdhury, G.G. 1996. Automatic extraction of citations from the text of English language patents: An example of template mining. *Journal of Information Science.* 22(6): 423-436.

NASA Astrophysics Data System. 〈http://adswww.harvard.edu〉

The NLM PubMed Project. 〈www.ncbi.nlm.nih.gov/PubMed/overview.html〉

An Open Journal Framework: Integrating Electronic Journals with Networked Information Resources. 1995. 〈http://journals.ecs.soton.ac.uk/flyer.html〉

Paice, C. D. 1981. The Automatic Generation of Literature Abstracts: an Approach Based on the Identification of Self-indicating Phrases. In R. N. Oddy, S. E. Robertson, C. J. Van Rijsbergen and P. W. Williams (eds.) *Information Retrieval Research.* London: butterworths. 172-191.

Paskin, Norman. 1999. DOI: Current Status and Outlook. *D-Lib Magazine.* 5(5). 〈http://www.dlib.org/dlib/may99/5paskin.html〉

ResearchIndex. 〈http://www.neci.nec.com/~lawrence/researchindex.html〉

What is Reference Linking?. 〈http://arabica.ecs.soton.ac.uk/ref__linking.html〉