URI 기반 저자 인용 네트워크 구축 및 활용
(Construction of Citation Network of Authors Using URI)

구 희 관 † 정 한 민 †† 강 인 수 †† 이 승 우 †† 성 원 경 ††
(Heekwan Koo) (Hanmin Jung) (Insu Kang) (Seungwoo Lee) (Wonkyung Sung)

요 약 과학기술 문헌에 대한, 정확한 인용 정보의 제공을 위해서는 인용 관계를 구성하는 인력에 대한 동명이인 문제가 우선적으로 해소되어야 한다. 본 논문에서는 정확한 저자 중심의 인용 관계 및 네트워크를 구축하기 위해 URI(Universal Resource Identifier)를 이용하는 방법을 제안한다. 본 연구의 특징은, 지속적으로 추가되고 갱신되는 인용 정보를 일관성 있게 유지하고 정확한 문헌으로의 접근성을 보장하기 위해, 시맨틱웹 기술의 하나인 URI를 문헌과 저자에 적용하여 저자 중심 인용 관계 및 네트워크를 구축한다는 것이다. 실험에서는, 국내 학술대회 발표 논문들로부터 2,872개의 저자 중심 인용 관계 쌍을 추출하였고, 이를 바탕으로 구축한 저자 인용 네트워크에서는 135개의 인용 네트워크 그룹을 발견할 수 있었다. 본 연구의 결과는 향후 국가과학기술인력 종합정보시스템에서 제공하는 인력 DB 및 과학기술정보 포털서비스인 YesKiSTi와의 연계를 통하여 새로운 개념의 다양한 연구자 네트워크 서비스를 제공할 수 있을 것으로 기대된다.

키워드 : 인용 네트워크, 저자 인용 네트워크, URI

Abstract For the construction of accurate scientific citation information, author disambiguation should be primarily resolved. This study proposes a method that utilizes URI(Uniform Resource Identifier) to create precise author citation networks. The adoption of URIs for representing authors and papers in this study enables us to maintain the integrity of constantly changing citation information and to guarantee the accessibility to the right literature. In experiments, we extracted 2,872 author-centric citation relation pairs from recent major IT-related proceedings written in Korean. From those, 135 citation network groups were discovered. The findings of this study are expected to be applied to a variety of researcher network services and scientific information portal services.

Key words : Citation Network, Author Citation Network, URI

1. 서 론

문헌에 대한 인용 네트워크는 지식의 구조와 흐름을 파악하는 방법으로 폭넓게 사용되고 있다. 인용과 피인용의 관계를 이용하여 정보의 흐름을 파악하고 이를 이용해서 특정 지역 및 특정 분야의 지식의 구조와 흐름을 파악하고자 하는 연구가 다양하게 시도되고 있다. 인용 네트워크와 관련하여 인용 분석의 적용 분야 또한 매우 다양하다. 예를 들어, 인용 분석은 이용자/과학사 연구, 정보 검색, 특정 분야의 문헌 형태나 이용 형태의 구조적 특성 규명, 특정 분야 연구자의 연구경향 파악 등의 연구에서 광범위하게 적용되고 있는 것으로 알려져 있다[1].

그러나 이러한 인용 정보는 주로 수작업으로 구축되기 때문에 비용 및 시간의 제약으로 인한 어려움을 겪으며, 이는 즉각적인 인용 정보 구축을 어렵게 만든다. 이 문제를 해결하기 위해 자동화된 인용 정보 구축 의 필요성이 제기되었다. 그러나, 대부분의 자동적인 인용 정보 구축 방법들은 문헌의 접근성 및 저자의 고유성을 확보하지 못하는 한계를 가진다. 문헌의 접근성 문제는 문헌의 직접적인 접근방법을 보장하는 하나의 식별체계를 이용하는 방법을 사용하여 해결될 수 있을 것이다. 저자의 고유성을 구별한다는 것은 동명이인의 문제를 어떤 방법으로 해결할 것인가의 문제이다. 저자에 대해 고유한 URI를 부여하게 되면 저자간의 인용 정보를 더

† 학생회원 : 과학기술연합대학원대학교 응용정보과학
hkkoo@kisti.re.kr
†† 정 회 원 : 한국과학기술정보연구원 정보시스템연구팀 연구원
jhm@kisti.re.kr
dbaisk@kisti.re.kr
swlee@kisti.re.kr
wksung@kisti.re.kr
논문접수 : 2006년 8월 13일
심사완료 : 2006년 8월 31일

욱 정확하게 구별해 낼 수 있으며, 이를 이용하는 인용 네트워크 서비스 시스템의 신뢰성을 보장할 수 있을 것이다.

한국과학기술정보연구원에서 개발 중인 국가과학기술 R&D 기반정보 온톨로지처럼, 문헌만이 대상이 아니라 과제, 지적재산권 등의 성과에 대한 인용 정보를 고려할 때 저자 간의 인용 정보는 더욱 필수적인 정보이다[2,3]. 저자 중심 인용 네트워크 구축은 저자 간의 인용을 통해 성과에 대한 질적인 판단의 기준을 제공해 주기도 하지만, 부가적으로 인용은 저자간의 사회적인 연결 네트워크를 보여주는 지표로서 사용될 수 있다. 만약 특정 주제 분야인 "한국어정보처리"에 대해, 저자 간 인용을 추적한다면, 인용 네트워크를 통해 해당 분야 전문가를 추천할 수 있게 될 것이다.

본 논문에서는 먼저 자동적인 방법을 이용하여 문헌 간의 인용 정보를 구축한다. 이후, 문헌 중심의 인용을 확장하여 저자 중심의 인용 네트워크를 구축한다. 이 과정에서 문헌과 저자를 각각 문헌 URI와 저자(인력) URI로 표현함으로써 문헌과 저자의 고유성을 확보하고자 시도하였다.

2. 관련 연구

2.1 개요

문헌 중심의 인용 관계에서 저자 중심 인용 관계로의 확장은 저자를 저자 개인이 아닌 저작물에 담긴 지식에 대한 소유자나 대표자로 고려하고자 하는 시도이다. 문헌이 아닌 저자를 인용의 기본 단위로 함으로써, 특정 주제 분야에 대한, 사람 중심의 실세계 지식 구조를 보다 정확하게 분석할 수 있을 것이다. 기존의 저자 중심 인용 네트워크에서는 대부분 제1저자만을 고려했다. 그러나, 제1저자 중심의 저자간의 인용관계를 공동 저자 전체 간의 인용관계로 확장함으로써, 학제간 복합적 연구나 대규모 컨소시엄 형태의 프로젝트와 같은 최근의 연구 현실을 잘 반영할 수 있을 것이다[4]. 이러한 현실을 반영하여 본 논문에서는 한 문헌의 공동 저자 모두를 해당 문헌에 내재된 지식 유통의 기여자로 고려하여 저자 중심 인용 네트워크를 구축하고자 시도하였다.

2.2 인용정보 구축

국내 학술 논문의 인용분석을 시스템적으로 해결하려는 방법에 대한 최근 연구로는, 최광남[5]이 제안한 Korean Science Citation Index(KSCI)[1]를 예로 들 수 있다. 이는 Thomson Scientific(TS)이 제안한 JCR 방식으로 영향력 지표와 즉시성 색인을 생성하여 이를 분석한 것으로 한국과학기술정보연구원(KISTI)에서 2001년 10월부터 2003년 8월까지 구축한 과학기술분야 학술지 247종 5,287권에 대한 인용 정보를 제공하고 있다. KSCI는 다양한 학술지에 대한 인용 정보를 구축하고 제공하지만 인용정보를 수작업으로 입력해야 하는 단점과 저자에 대한 동명이인의 해결방법을 제시하지 못하는 단점을 가진다.

자동적인 인용 정보를 생성하는 예는 CiteSeer[2]와 Google Scholar[3]를 들 수 있다. 현재 CiteSeer는 72만여 개의 논문과 800만여 개의 인용 정보를 구축하고 있으며 Google Scholar는 4억여 개 이상의 논문 및 인용 정보를 제공하고 있다[6]. 이 중 CiteSeer는 원문을 이용해 인용 정보를 구축한다. Postscript형식의 원문을 대상으로 하여 수집된 논문을 먼저 텍스트로 변환한 다음, 논문의 원문 정보에 대해 정규식을 이용하여 정해진 영역별(URL, Header, Abstract, Introduction, Citations, Citations context, Full text)로 추출해 낸다. 인용 정보를 구축하는 과정에서는, 기존에 구축된 논문정보들과 추출된 Citation 영역에서의 논문과의 일치 여부를 확인하는 단계를 거친다. 저장된 논문 정보에 대해 인용 관계를 파악하려면 변형적으로 나타나는 동일한 논문을 파악해야 한다. 이를 위해, 일련의 정규화 과정(소문자화, '-'제거, 인용 태그 제거, 일반적으로 사용되는 약어의 확장, 몇몇 기호 및 단어 제거)을 거쳐 참고문헌과 기존 논문과의 단어기반의 일치 여부와 단어 발생 순서를 기준으로 특정한 임계치 이상의 값을 가진 문헌에 대해 인용 관계를 설정한다[7]. 본 논문에서 사용하는 인용정보 추출 방식도 CiteSeer의 방법과 크게 다르지 않으나, 한글로 쓰여진 국내 논문들의 참고문헌에 대한 자동 정보 추출을 시도했다는 점에서 의의를 찾을 수 있을 것이다.

2.3 동명이인 문제 해소

Yoojin Hong 등[8]은 인용 관계 구축 시에 이름에 대한 Unique ID(예: DOI) 결정 시스템을 제안하였다. 저자 간에 인용 관계를 맺으려 할 때, 발생하는 문제를 Citation Matching이라 정의하고 이 문제를 변경(결혼으로 성이 바뀌는 경우), 분할(동명이인이 출현한 경우), 통합(동일한 저자이름이 다양하게 출현하는 경우)의 세 가지로 구분하여 각각에 대해 Unique ID를 이용하여 해결할 것을 제안하였다. 영어의 경우와 달리 한글로 쓰여진 참고문헌 내 저자 간 매칭에 있어서, 전술한 변경이나 통합의 경우는 제한적으로 발생하므로, 분할의 문제를 중점적으로 다룰 필요가 있을 것이다.

국외 학술 논문 인용에 관한 대표적인 예는 SCOPU

1) http://ksci.kisti.re.kr

2) http://citeseer.ist.psu.edu/

3) http://scholar.google.com/

S[4]를 들 수 있다. SCOPUS는 동명 저자에게 소속기관과 결합된 저자 ID를 부여함으로써 동명이인 문제를 해결하고자 시도했으나, 저자의 소속 변경과 같은 흔히 발생할 수 있는 현상에 대처하지 못하는 단점이 있다. 예를 들어 '포항공대'의 '정한민'과 'ETRI'의 정한민은 동일인물이지만 SCOPUS에서는 각각 다른 저자ID (예: AU-ID("Jung, Hanmin" 7403030093), AU-ID("Jung, Hanmin" 7403030150))를 부여함으로써 이 두 사람이 다른 인력으로 처리되는 문제점을 가진다.

3. 시스템 구성

본 연구에서 사용하는 저자 중심 인용 네트워크 구축의 핵심 과정은 그림 1에서 보는 것처럼 크게 인용정보 추출, 문헌 기반 인용 관계 구축, 저자 기반 인용 관계 구축, 그리고 인용 네트워크 구축의 네 단계로 나누어진다. 첫 단계인 인용 정보 추출에 앞서 먼저 전처리로 원문을 수집하고 텍스트 필터 소프트웨어를 이용하여 원문을 텍스트로 변환한다. 다음으로 CiteSeer와 유사하게 정규식을 이용하여 참고문헌으로부터 인용 정보를 추출한다. 두 번째 단계인 문헌 기반 인용 관계 구축은 앞 단계에서 추출된 참고문헌 부분에 기술된 문헌의 제목을 이용하여 개별 문헌에 문헌(객체) URI를 할당한다[9]. 세 번째 단계로 문헌 기반 인용관계를 이용하여 저자 중심 인용 관계를 구축한다. 마지막 단계로 저자 중심 인용 네트워크를 구축한다. 여기에서 저자 중심 인용 관계는 기존 방식에서와 같은 제1저자를 기준으로 한 인용관계가 아니라 전체 저자들 간의 인용 관계로 확장된 것이다.

이 시스템은 협업 연구를 위한 시맨틱 웹 기반 지식정보 공유·유통 플랫폼인 OntoFrame-K®의 연구자 네트워크 중 인용 기반 네트워크 구축 부분을 차지하고

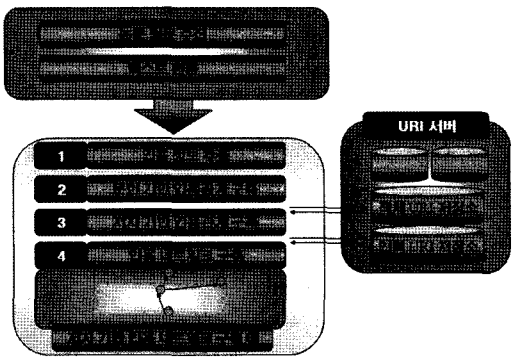

그림 1 저자 중심 인용 네트워크 시스템 및 구축 단계

4) http://www.scopus.com/

있다[10]. OntoFrame-K®은 새로운 개념의 정보유통 플랫폼으로서, 이미 검증된 전문성과 높은 부가가치를 지닌 개인 소장 과학기술 지식정보를 공유·유통 시킬 수 있는 "자발적 가상 협업 연구 커뮤니티(Self-Organizing Virtual Research Community)"의 구현을 지원하기 위한 시스템이다.

그림 1에서는 URI서버에서 문헌URI와 인력URI가 관리되고 있는 것을 보여주고 있는데, 문헌URI 저장소에서는 개별 문헌에 고유한 식별자를 자동 부여하여 해당 문헌의 메타 정보를 관리하며, 인력URI저장소에서는 인력에 대해 고유한 식별자를 반자동 방법으로 부여하여 해당 인력의 메타정보를 관리하게 된다. 본 연구에서는, 문헌URI와 인력URI가 이미 구축되어 있다고 가정하고 있다. 그러나, 인력 중심의 인용 관계 구축에서 인력URI가 핵심적인 부분이므로 본 연구의 기반이 되는, 인력에 대한 동명이인 해소 방법에 대해 간략히 기술하고자 한다.

본 연구팀에서는 동명이인의 문제를 해소하기 위한 단서로 동일명 연구자의 논문게재건들에 내재된 공저자 정보, 전자메일주소, 소속기관명, 출판년도 등을 복합적으로 고려하는 방법을 사용한다[3]. 이 논문에서는 그 방법의 핵심 아이디어만을 소개한다. 먼저 아래와 같은 동일명 연구자 A의 논문게재건들이 주어졌다고 하자.

논문 1의 저자명: A, B, C
논문 2의 저자명: A, C, D

위에서 논문 1과 2는 A의 공동저자 C를 공유하고 있는데, 이러한 상황에서 논문 1과 2에서의 두 A는 동명이인일 확률이 매우 낮을 것이다. 여기에 더하여 만약 논문 1과 2의 두 A의 전자메일 주소가 같고, 논문 1과 2의 두 C의 전자메일 주소도 같다면 논문 1과 2의 두 A가 동일인일 확률은 더 높아진다. 여기에 논문 1과 2의 두 A와 두 C가 각각 소속기관까지 같고, 두 논문의 출판년도마저도 같다면, 논문 1과 2의 A라는 동일이름의 두 저자는 실제로 같은 사람일 확률이 훨씬 더 높아질 것이다. 이러한 동명이인 해소 방법을 컴퓨터로 자동화시키는 것은 어렵지 않을 것이다

4. 인용 정보 추출

인용 정보 추출의 자동화로 얻을 수 있는 가장 큰 장점은 지속적으로 변화하며 생성되는 인용 정보를 시스템적으로 바로 반영할 수 있다는 것이다. 이로 인해 인용정보를 이용한 다양한 특정 지역 및 특정 분야의 지식의 구조와 흐름의 파악이 용이하게 되며, 이용자 연구, 과학사 연구, 특정분야 과학자의 커뮤니케이션 유형 규명, 과학적 영향 평가 및 생산성 평가의 측정 등의 연구에 기여할 수 있는 바가 크다고 할 수 있다.

그러나, 자동적인 인용 정보 추출의 어려운 점은 저자

들의 참고 문헌 기술 방식이 일정하지 않고 다양하다는 것이다. 다시 말하면 참고문헌은 개략적으로 "저자, 제목, 출처" 등의 순서로 기술되지만 참고문헌을 기술하는 방식은 원문에 따라, 저자에 따라, 그리고 게재지의 요구사항에 따라 다양한 형태를 띄고 있는 것이 현실이다.

그림 2에서는 인용 정보 추출의 세 단계를 그림으로 보여주고 있다. 인용 정보 추출은 참고문헌 영역 확인, 인용 정보 추출, 인용 정보 메타데이타 추출의 세 단계로 구성이 되어 있다. 그림에서는 참고 문헌 영역확인과 인용정보추출이 끝나고, 인용 정보 메타데이타를 추출하는 과정을 보여 주고 있다. 그림에서, 문헌에 할당된 객체 URI 식별자는 "KISTI1.PCD.0004407"이며, 저자 URI는 "0000003854"이다.

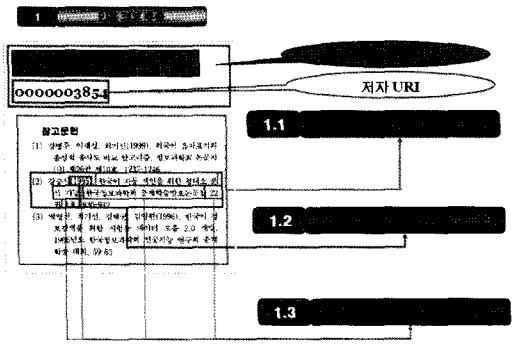

그림 2 인용정보추출

원문으로부터 인용 정보 추출의 첫 번째 단계는 참고문헌 영역을 결정하는 것이다. 원문의 참고문헌 영역은 동일한 학술대회일지라도 저자에 따라 "Reference", "References", "참 고 문 헌", "참고문헌" 등의 다양한 변형이 가능하기 때문에 이에 대해 적절한 정규식을 지정하는 것이 필요하다. 그리고 참고문헌 영역을 지정하는 또 다른 이유는 문헌 의 텍스트 변환 후에 참고문헌 영역은 문헌의 끝부분에 항상 위치하고 있는 것은 아니기 때문이다.

인용 정보의 메타데이타는 정규화된 필드로 구성되는데, 본 연구에서는 인용 문헌의 서지정보로부터 학회필드, 학술대회필드, 권(Volume) 필드, 호(Issue) 필드, 연도필드에 해당하는 정보를 추출한다. 인용 정보 메타데이타 추출은 그림 3과 같은 단계로 진행이 된다.

첫 단계인 공백제거 및 오류 제거는 추출되는 한글에 대한 처리를 용이하게 하기 위한 전처리 단계에 해당한다. 오류에 해당하는 문자열은 원문을 텍스트필터를 사용하여 변환할 때 생성되는 특정한 패턴을 가지는 문자열이기 때문에 쉽게 제거가 가능하다. 두 번째 단계는 문자열을 구분 기호로 분리하여 토큰을 생성하는 단계

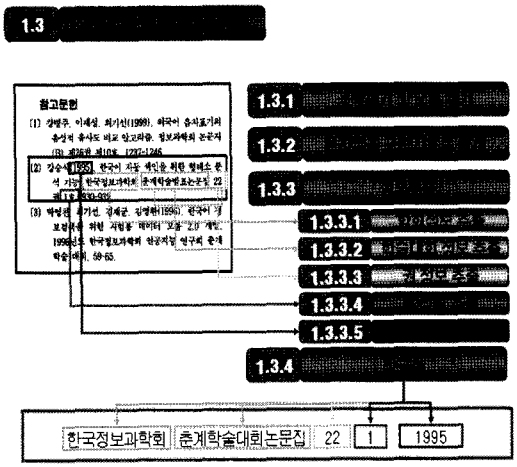

그림 3 인용정보 메타데이타 추출

이다. 이는 문자열을 특정한 단위로 처리하기 위함이며, 구분기호로는 ' ,', '.' 등이 사용된다. 세 번째 단계인 정보 추출에서는, 이전 단계에서 분리된 토큰들이 추출 조건에 부합되면 해당 정보를 추출하여 메타데이타의 필드를 채우도록 진행된다. 표 1은 정보 추출 메타데이타 필드들과 추출 정규식을 보여준다.

표 1 학술대회 추출 필드 및 추출 정규식

추출 필드	추출 정규식
학회필드	"학[ ]?회" 추출 후 적절한 형태 변환
학술대회 필드	"H[ ]?C[ ]?I[ ]?, 학[ ]?술[ ]?대[ ]?회[ ]?, 학[ ]?술[ ]?발[ ]?표[ ]?" 추출 후 삭제 및 변환 학술 대회 정보를 획득
권(Volume) 필드	[Vv][Oo][L].[0-9][0-9]*[-]*[A-L] *[Vv].[0-9][0-9]*[-]*[A-L]* [Vv][Oo][L][0-9][0-9]*[-]*[A-L] *[0-9][0-9 -]*[A-L]*권
호(Issue) 필드	[Nn][Oo][.][0-9][0-9 -a-lA-L]*, [Nn].[0-9][0-9 -a-lA-L]* [Nn][Oo][0-9][0-9 -a-lA-L]* [Nn][0-9][0-9 -a-lA-L]* [0-9][0-9 -a-lA-L]*호"
연도필드	[1][9][0-9][0-9] [2][0][0-9][0-9] [9][0-9], [9][0-9], [0][0-9], [0][0-9]

메타데이타 추출 이후에 정규화는 메타데이타의 필드를 하나의 대표명으로 변환하는 단계가 필요하다. 표 2는 메타데이타 학회필드의 변환테이블을 보여준다. 변환의 기준이 되는 대표 학회명은 한국과학기술정보연구원의 학회마을[5]의 명칭을 기준으로 생성한다. 보통 전거

5) http://society.kisti.re.kr

라 부르는 다음과 같은 변환테이블은 지속적으로 추가되어야 할 것이다. 이외에 고려해야 할 사항으로는 '춘계학술대회'와 '봄학술대회' 등 계절에 관련한 학술대회 명칭이 함께 출현하는 경우를 고려하여, 학술대회명에 대한 정규화의 과정도 필요하다. 이외에 권(Volume)과 호(Issue)의 정보는 저자의 기술이 일관되게 기술되지 않는 경우가 많기 때문에, 이를 권호정보 숫자로 변환하여 처리한다(예 : Vol.32, V.32, 32권 등). 마지막으로, 학술대회 연도정보도 일관되게 기술되지 않는 경우가 많으므로 정규화의 과정이 필요하다(예 : 1999, 99, '97).

표 2 학회명 변환 테이블의 예

대표 학회명	변환 대상 학회명
한국정보처리학회	한국정보처리학회 정보처리학회
한국정보과학회	한국정보과학회 정보과학회
대한전자공학회	전자공학회 대한전자공학회
한국멀티미디어학회	멀티미디어학회 한국멀티미디어학회
한국통신학회	한국통신학회 통신학회
한국방송공학회	한국방송공학회 방송공학회
한국음향학회	한국음향학회 음향학회
대한전기학회	전기학회 대한전기학회

논문지에 대한 메타데이타의 정규화 과정도 위의 과정과 유사한 과정을 거쳐 생성된다. 논문지의 메타데이타 정규화 필드는 학회명, 논문지명, 권, 호, 연도 필드들로 구성된다. 논문지 메타데이타 생성과정에서도 학술대회 메타데이타 생성과정에서처럼 논문지명에 대한 정규화의 과정이 필요하다. 표 3은 대표 논문지명으로 변환되는 대표 논문지명과 변환 대상 논문지명의 예를 보여준다.

표 3 논문지명 변환 테이블의 예

대표논문지명	변환 대상 논문지명
한국정보처리학회지	정보처리학회지, 한국정보처리학회지, 한국정보처리학회학회지, 정보처리학회학회지
한국정보처리학회논문지	정보처리학회논문지, 정보처리논문지
대한전자공학회학회지	전자공학회지, 대한전자공학회지, 전자공학회학회지, 전자공학학회지
대한전자공학회논문지	전자공학회논문지, 전자공학논문지, 전자공학회지논문지

이 장에서 기술한 것처럼, 현재 본 논문이 제안한 인용 정보 추출 방법은 CiteSeer의 방식과 유사하게 정규식을 이용하여 정보를 추출하고 있다. 그러나, 이러한 방식은 정규식 형태의 규칙 작성의 어려움과 복수 규칙들 간 적용에 있어서의 충돌과 우선순위 결정 등의 어려움 등 통상의 규칙기반 접근법의 단점을 안고 있다. 따라서, 앞으로는 다양한 통계기반 학습 방법을 적용하여 체계적으로 구축하는 방법에 대한 연구가 필요할 것이다.

5. URI 기반 논문(객체) 인용 관계 구축

문헌의 인용간의 관계를 단순히 문자열 기반으로 유사도를 이용하여 구축한, CiteSeer와 유사한 접근방법으로는 문헌 자체의 접근성을 보장해 내지 못하며, 이는 개별 문헌에 고유한 URI를 부여함으로써 해결될 수 있다. 또한 문헌에 부여되는 객체 URI는 저자들의 인용관계 생성을 위한 고유한 키값으로 유용하게 사용될 수 있다.

URI 기반 인용 관계를 설명하기 위해서 문헌 URI와 인력 URI의 식별체계를 기술할 필요가 있겠다. 문헌 URI 식별체계는 호환성을 고려해 KISTI의 KOI 생성규칙을 간략화하여 정의하였으며, 생성형식은 "등록관리기관코드.자료유형코드.일련번호"형태를 가진다. "등록관리기관코드"는 객체의 종류에 따라 학술지, 보고서, 개인공유자료, 기반정보 등으로 구분하며, "자료유형 코드"는 학술지, 학술대회, 학위논문, 특허, 연구보고서, 개인공유자료 등으로 구분한다. 마지막으로 "일련번호"는 7자리의 숫자를 사용한다[9].

인력 URI 식별체계는 국가과학기술인력 종합정보시스템에서 사용하는 10자리 인력ID값에 기반하여 설계되었다. 예를 들어, 인력 URI "http://www.kisti.re.kr/isrl#PER_7010186243"에서는 네임스페이스(http://www.kisti.re.kr/isrl), 자료유형(PER), 그리고 10자리 일련번호(7010186243)로 구성되어 있다.

그림 4는 문헌 중심의 인용 네트워크를 구축하기 위한 세 단계를 보여준다. 첫 단계는 인용 정보 중 제목을 추출한다. 다음 단계는 문헌제목을 이용하여 문헌 간에 인용 관계를 구축하는 것이다. 마지막 단계는 문헌에 해당하는 객체 URI 저장소 내에 저장되어 있는 객체 URI를 이용하여 문헌 중심의 인용 관계를 구축한다.

참고문헌 제목추출 단계에서 문헌 제목의 추출은 보

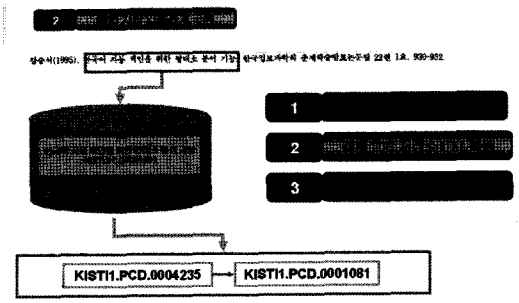

그림 4 문헌 기반 인용관계 구축 단계

통 이중인용부호(“)를 이용하여 그 경계를 인식할 수 있지만, 원문 출전에 따라 이중인용부호(“,”)와 단일인용부호(‘,’) 등의 여러 구분자들이 사용될 수 있으므로 다양한 형태적 특징을 고려하여 제목 추출 과정이 필요하다. 또한 “문헌제목,”과 같은 형태가 관습적으로 많이 사용되기 때문에 이에 대한 전처리 과정이 필요하다.

다음으로, 추출된 제목들을 이용하여 문헌 중심의 인용 네트워크를 구축한다. 인용 네트워크를 구축하기 위해 추출된 문헌 제목들을 인용과 피인용의 쌍으로 단순화한다. 본 논문이 제안한 시스템은 국내 논문들 간의 인용만을 대상으로 하였기 때문에 국외 논문들의 관계는 제외하였다. 마지막으로 인용과 피인용의 관계를 URI서버를 이용하여 객체 URI로 변환한다. 최종적으로 그림 4의 하단에 사각형 안에 객체 URI간의 인용관계 예를 확인할 수 있다.

6. URI 기반 저자(인력) 인용 관계 구축

그림 5는 문헌 인용 네트워크를 확장한 저자(인력) URI 기반 인용관계 구축의 주요 단계를 보여준다. 첫 단계는 인용-피인용으로 연결된 문헌URI 쌍으로부터 하나의 문헌URI를 획득한다. 다음으로, 개별 문헌URI의 저자에 해당하는 인력URI를 URI서버로부터 획득하고, 마지막으로 문헌URI 쌍으로 묶인 두 문헌의 저자URI 집합들간에 인용-피인용 관계를 생성한다.

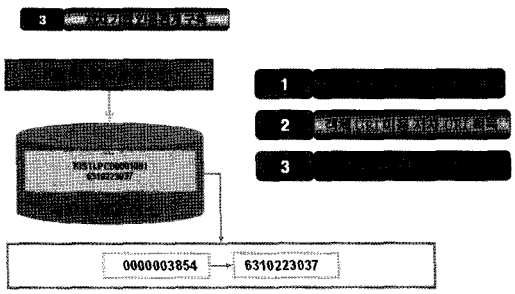

그림 5 저자 중심 인용관계 구축 단계

7. 인용 네트워크 구축

인용 네트워크는 지식의 구조와 흐름을 파악하는 방법으로 인용과 피인용의 관계를 이용하여 정보의 흐름을 파악하고 이를 이용해서 특정 지역 및 특정 분야의 지식의 구조와 흐름을 파악하고자 하는 연구에 사용된다. 특히, 저자 중심 인용 네트워크 구축은 저자간의 인용을 통한 성과에 대한 질적인 판단의 기준을 제공해 주기도 하지만 부가적으로 저자간의 사회적인 연결 네트워크를 보여주는 지표로서 사용될 수 있다.

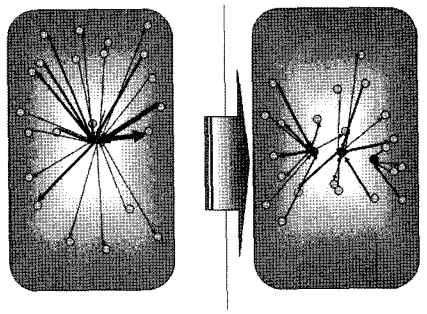

그림 6 저자 중심 인용 네트워크의 고유성 확보

그림 6은 저자 중심의 인용 네트워크가 인력 URI를 이용하여 저자의 고유성을 확보하게 되면 인용-피인용의 관점에서 동명이인으로 인한 애매성이 해소되어 정확한 저자 중심 인용 네트워크가 구축되는 예를 보여준다.

8. 실험 및 결과

저자 중심 인용 네트워크를 구축하기 위한 원문 수집은 국내 여러 학회에서 배포한 학술대회논문집 관련 CD-ROM에 수록된 원문을 이용하였다. 표 4에 보인 것처럼 4차에 걸쳐 원문 수집이 진행되었는데, 총 수집된 문서는 10,262개의 논문이고 텍스트 필터링되어 추출된 논문의 수는 7,237개였다. 1차 원문 수집 대상 중 정보과학회의 2005년 추계학술대회 이후 논문집은 원문으로 문서필터링이 이뤄지지 않았다. 1차 원문 수집 이후 학회의 출현 빈도 및 중요도에 따라 이후 차수의 대상 학회를 선정하여 추가하였다.

표 5는 인용정보추출의 성능을 보여주고 있는데, 학술대회와 논문지에 게재된 문헌을 인용하고 있는 서지정보 925건과 1,156건을 대상으로 하였고, 게재지명(학술대회명, 논문지명), 권/호정보, 연도정보를 추출하는 규칙기반 인용정보추출 방법의 성능을 재현율(Rec.), 정확

표 4 수집 문헌

학회명칭	발행연도	차수	총 문서	추출문헌
한국정보과학회(KISS)	2002/2004	1	3793	2673
대한전자공학회(IEEK)	2003/2005	1	1458	1193
한국HCI학회(HCI)	2003/2006	1	1300	722
한국정보처리학회(KIPS)	2004/2006	1.4	1267	1230
한국통신학회(KICS)	2003/2005	2	1559	995
한국멀티미디어학회(KMMS)	2005/2006	3	499	218
한국인터넷정보학회	2005/2006	3	245	125
한국과학기술정보인프라워크숍	2005	3	141	81
계			10262	7237

표 5 인용정보추출 성능

추출 필드	학술대회 (925건 대상)			논문지 (1156건 대상)		
	Rec.	Pre.	F	Rec.	Pre.	F
게재지명	100	98.7	99.3	98.1	98.3	98.2
권정보	29.9	98.9	45.9	71.7	99.5	83.3
호정보	27.9	98.8	43.5	75.8	99.7	86.1
연도정보	88.8	99.6	93.9	85.8	98.8	91.8

률(Pre.), 그리고 F-measure값으로 구분하여 표시하고 있다. 학술대회 게재 문헌의 경우 권/호 정보의 재현율이 낮은 것은, 논문지의 경우와 달리, 학술대회논문집의 경우 권/호 정보가 없는 경우가 대부분이기 때문이다.

인용정보추출을 위해 학술대회와 논문지에 대해 각각 63개와 47개의 정규식 형태의 규칙을 사용하였으며, 이의 구축을 위해 한 명의 대학원생이 일주일 정도의 시간을 투입하였다. 학회명과 논문지의 명칭 정규화를 위한 변환 테이블로는, 학회명과와 논문지에 대해 각각 16개와 25개의 매핑 정보를 사용하였다.

표 6은 수집된 메타데이타의 정규화된 정보를 이용하여 참고문헌의 출처를 분석한 것이다. 국외 참고문헌과 국내 참고문헌의 비중은 79%와 21%로 나타났다. 이는 최광남[11]의 KSCI 구축 결과와 유사한 결과이며, 국외 참고문헌 비중이 높은 현실을 반영하고 있다. 국내 참고문헌의 비율은 다시 논문지, 학술대회, 홈페이지, 학위논문이 각각 13%, 13%, 5%, 6%로 세분되었다. 그리고 앞의 4가지 경우에 해당되지 않는 참고문헌을 기타로 분류하였는데 63%의 높은 비율을 보였다. 기타의 유형을 살펴보면 일반적으로 단행본류의 서적인 경우가 제일 높았으며, 그 외에 학회 논문집이 아닌 논문집(eg. 대학교 정기 논문집) 등이었다. 이후 기타를 좀 더 세분화하여 분류하는 작업이 진행되어야 할 것이다.

표 7과 표 8은 1차 구축분을 대상으로 출현한 학회 인용 건수 및 인용문헌 출간년도를 보여주고 있다.

표 7 논문지 별 학회 출현 빈도 및 인용문헌 연도 변화

학회명칭	인용 추출건수	인용문헌 년도	인용 추출건수
한국정보과학회	333	1996	32
한국정보처리학회	303	1997	34
한국통신학회	85	1998	54
대한전자공학회	85	1999	64
한국정보보호학회	38	2000	118
한국멀티미디어학회	30	2001	207
한국전자파학회	14	2002	186
한국방송공학회	13	2003	146
한국음향학회	12	2004	83
한국컴퓨터보호학회	11	2005	8
계	924	계	932

표 8 학술대회 별 학회 출현빈도 및 인용문헌 연도변화

학회명칭	인용 추출건수	인용문헌 년도	인용 추출건수
한국정보과학회	474	1996	6
한국정보처리학회	168	1997	24
대한전자공학회	64	1998	37
한글 및 한국어정보처리학회	62	1999	60
HCI학회	59	2000	85
한국멀티미디어학회	37	2001	193
한국통신학회	30	2002	280
대한전기학회	9	2003	233
계	903	2004	89
		2005	7
		계	932

그림 7은 URI 기반 논문(객체) 인용 관계 구축 결과를 보여준다. 그림의 원에 들어 있는 숫자의 값은 각 단계마다 결과로 발생 건수들이며, 그 안에 사각형에 들어 있는 것은 앞 단계에서 현재 단계로 진행되는 동안 인용 관계가 줄어든 이유이다. 1,407건의 학술대회 논문의 인용건수 중 학술대회 논문의 메타데이타 정규화 형태

표 6 참고문헌 분석

학회명칭	참고문헌 추출건수	국외 참고문헌	국내 참고문헌	국내 논문지	국내 학술대회	홈페이지 링크	학위논문	기타
한국정보과학회(KISS)	18363	14632	3731	534	634	280	230	2053
대한전자공학회(IEEK)	7157	6010	1147	188	118	13	87	741
한국HCI학회(HCI)	6487	5120	1367	104	179	76	126	882
한국정보처리학회(KIPS)	9543	7025	2518	380	343	110	181	1504
한국통신학회(KICS)	7220	6208	1012	87	67	27	23	808
한국멀티미디어학회(KMMS)	1843	1445	398	46	28	27	13	284
한국인터넷정보학회	807	525	282	42	17	14	15	194
한국과학기술정보인프라워크숍	750	374	356	25	21	19	7	284
계	52150	41339 (79%)	10811 (21%)	1406 (13%)	1407 (13%)	566 (5%)	682 (6%)	6750 (63%)

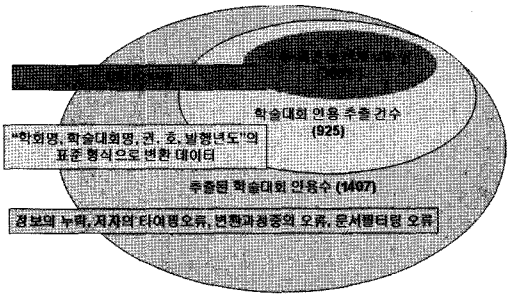

그림 7 문헌기반 인용 관계 구축 및 객체 URI 변환 결과

로 변환된 인용은 925건 이었으며, URI 서버 내의 객체 URI 저장소에 구축되어 있는 객체 URI쌍으로 변환된 URI쌍은 366쌍이었다. 366쌍이 최종 결과가 된 가장 큰 이유는 그림에서 보여주듯이 URI 서버에 논문 URI가 구축되지 않았기 때문이다. 그러나 이 문제는 지속적으로 문헌URI를 확장해 나감으로써 해결될 수 있다.

객체 URI 기반 인용-피인용 쌍 366쌍을 이용하여 URI 기반 저자(인력) 인용 관계 구축 한 결과는 2,872개의 저자 중심 인용-피인용 인용 관계 쌍이 추출되었다. 그리고 저자 중심 인용-피인용 관계 쌍 중 4%인 인용 정보 116건과 3.8%인 피인용 정보 110건에 대해 동명이인 문제를 해결할 수 있었다. 또한 135개의 인용 네트워크 그룹을 생성할 수 있었다

8.1 인용 네트워크의 특징

저자 기반 인용 네트워크는 저자 간 인용과 피인용의 관계를 구성단위로 사용하여 네트워크를 구성한다.

인용 출현 형태는 소수의 저자가 다수의 집중적인 인용을 받는 경향을 보인다. 이는 소수의 연구자가 많은 논문을 작성하여 많은 공저자 관계를 형성하는 것과 유사하다. 기존의 문헌간의 인용관계가 아닌 저자 중심 인용 네트워크 생성의 이점은 다음과 같다. 단순히 문헌에 대한 인용만을 파악하는 것이 아니라 인용의 수가 많은 논문의 저자는 다시 많은 인용의 논문을 재생산할 가능성이 높다고 예상할 수 있기 때문에, 이를 이용해 아직 인용이 발생하지 않은 새로운 논문들에 대한 평가도구로도 사용될 수 있을 것으로 예상된다.

전체 저자 기반 인용 네트워크를 표현하면 그림 8과 같이 표현된다. 앞서 언급한 것과 같이 국내 저자간 인용이 빈번하지 않게 출현하기 때문에 전체 인용 네트워크는 네트워크의 형태가 밀집되지 않은 형태를 보인다. 그림 8의 네트워크 표현을 위해서 네트워크 분석 도구인 Pajek[6]을 이용하여 표현하였다.

표 9는 인용네트워크의 구성 요소들을 보여준다. 이

6) http://vlado.fmf.uni-lj.si/pub/networks/pajek/

표 9 인용 네트워크 구성

그룹 개수	135개
전체 저자 수	827명
그룹 당 평균 저자 수	6.39명
가장 큰 그룹 저자 수	38명
가장 작은 그룹 저자 수	2명

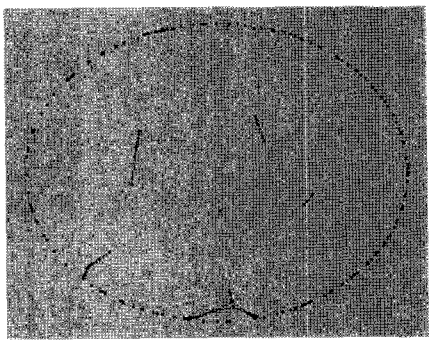

그림 8 전체 저자 중심 인용네트워크

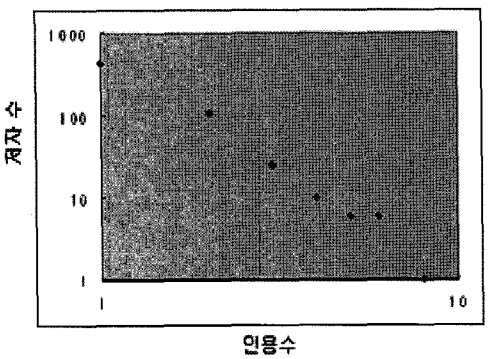

그림 9 인용 수와 저자 수의 히스토그래프
(x,y 축에 logarithmic scale 적용)

중 저자 수는 URI식별체계를 이용하여 중의성 문제를 해결한 저자의 수이다. 학술대회 논문이 국외 논문을 주로 인용하기 때문에 공저자 관계의 네트워크보다 그 구성 요소의 수가 작은 경향을 보인다. 그룹 개수는 인용의 방향성을 제거하고 그룹의 개수를 측정하였다.

그림 9는 저자 수와 인용의 관계를 보인다. 그림이 의미하는 바는 네트워크의 연결이 소수의 저자에게 집중되는 것을 보여주고 있으며, 이 둘의 관계는 소수의 노드에 집중적인 연결관계를 가지는 네트워크의 특징인 멱함수(Power-Law)의 특성을 보인다.

8.2 인용 네트워크의 실제 구현 예

협업 연구를 위한 시멘틱 웹 기반 지식정보 공유·유통 플랫폼인 OntoFrame-K®의 연구자간 네트워크(Communities of Practice: COP) 중 인용 네트워크 구축 부

분을 차지한다[10]. OntoFrame-K®은 새로운 개념의 정보유통 플랫폼으로서, 이미 검증된 전문성과 높은 부가가치를 지닌 개인 소장 과학기술 지식정보를 공유·유통 시킬 수 있는 "자발적 가상 협업 연구 커뮤니티(self-organizing virtual research community)"의 구현을 지원하기 위한 시스템이다.

그림 10은 표 9에서 보인 135개의 인용 네트워크 그룹이 OntoFrame-K®내에서 실제로 제공되는 인터페이스를 보인다. 이들 인용 네트워크 그룹 정보가 사용자에게 제공될 때, 주제 및 분야를 한정하여 인용 네트워크 정보를 제공한다. 주제 및 분야를 한정하는 방법은 다음과 같다. 문헌으로부터 추출한 색인어를 시소러스 개념어와 매칭시킴으로써 해당 문헌을 대표하는 자동 주제 및 분야 할당을 한다. 이후에 사용자가 선택한 주제 및 분야 인용 네트워크 그룹을 시스템이 보여주게 된다[11]. 정리하면, 전체 인용 네트워크를 특정 주제나 분야로 나누어 부분 네트워크를 보여준다. 또한, 위 그림의 우측에는 특정 주제 및 분야에 인용지수가 높은 저자가 상위 20명이 나열되어 있으며, 이중 가장 높은 피인용 저자의 인용 그룹을 선택한 화면이다.

그림 11은 그림 10에서 선택된 인용 그룹 내에 상세 인용 네트워크를 보여준다. 이 네트워크 내에서 노드를

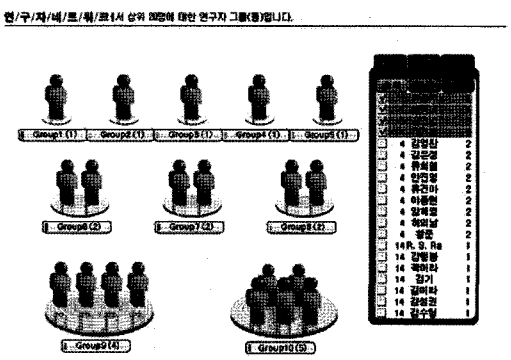

그림 10 저자 중심 인용네트워크 그룹 구현 예

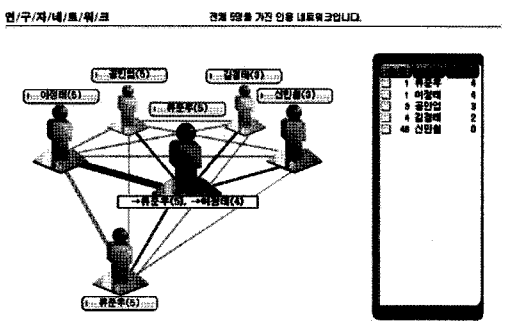

그림 11 저자 중심 인용 네트워크 구현 예

연결하는 링크는 인용의 방향과 빈도 정보를 함께 제공한다.

9. 결 론

기존 문자열 중심의 인용 네트워크는 동명이인의 문제를 해결하지 못했기 때문에 정확한 정보의 제공을 할 수 없었다. 본 논문에서는 지속적으로 추가되고 갱신되는 인용 정보를 일관성 있게 반영하고 정확한 문헌의 접근성을 보장하기 위해, 시맨틱웹 기술의 하나인 URI를 문헌(객체 URI)과 저자(인력 URI)에 적용하여 저자 중심 인용 관계 및 네트워크를 구축하였다.

향후 연구는 다양한 부분에서 확장해야 할 것이다. 자동적인 인용 정보 추출에서는 정규식의 사용 이외에도 [12]에서와 같은 Hidden Markov Models(HMM)을 이용하는 통계기반 방법의 사용을 고려할 수 있을 것이다. 또한, 기존에 구축된 논문의 국외 인용 문헌 정보를 바탕으로, DBLP와 CiteSeer내의 논문을 수집함으로써, 저자 중심 인용 네트워크를 국외로 확장할 필요도 있을 것이다[13].

참 고 문 헌

[1] 김홍렬 (2003). "과학기술문헌의 인용분석 연구", *정보관리학회지* 20(4):1-21, 2003.

[2] 강인수, 정한민, 이승우, 김평, 성원경 (2006). "국가과학기술 R&D 기반정보 온톨로지와 추론 모델링", *한국컴퓨터종합학술대회 논문집*, 제 33권 1(B)호, pp.13-15, 2006.

[3] 이승우, 정한민, 김평, 강인수, 성원경 (2006). "서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구", *한국컴퓨터종합학술대회 논문집*, 제33권 1(B)호, pp.10-12, 2006.

[4] 이은숙 (2002). "복수저자를 고려한 저자동시인용분석 연구: 정보학과 컴퓨터과학을 대상으로", *연세대학교 대학원 석사학위논문*, 2002.

[5] 최광남 (2004). "국내학술지 영향력 지표 분석을 위한 한국과학기술인용색인(KSCI) 연구", *한국문헌정보학회지* 38(4):271-289, 2004.

[6] Rahm, E., Thor, A. (2005). "Citation analysis of database publications," *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp.48-53, 2005.

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

[8] Hong, Y.J., On, B.W., Lee, D.W. (2004). "System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach," *Proceedings of the 8th European Conference on Digital Libraries*, 2004.

[9] 구희관, 정한민, 강인수, 성원경, 이승준, 심빈구(2006) "국가 과학기술 R&D 기반정보 온톨로지 구축을 위한

URI 관리 및 서비스 시스템 구현", *한국컴퓨터종합학술대회 논문집*, 제33권 1(B)호, pp.217-220, 2006.
[10] 성원경, 정한민, 박동인 (2006). "OntoFrame-K®: 협업 연구 지원을 위한 시멘틱 웹 기반 지식정보 공유·유통플랫폼", *정보과학회지* 24(4):65-72, 2006.
[11] 정한민, 강인수, 성원경 (2006). "시소러스와 분야분류체계를 이용한 과학기술문헌에의 주제 및 분야 할당", *한국언어정보학회 여름학술대회*, 2006.
[12] Borkar, V., Deshmukh, K., Sarawagi, S. (2001). "Automatic Segmentation of Text into Structured Records," *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp.175-186, 2001.
[13] Isaac G., Councill, Li, H., Zhuang, Z., Debnath, S., Bolelli, L., Lee, W.C., Sivasubramaniam, A., Giles, C.L. (2006). "Information Retrieval 2: Learning Metadata from the Evidence in an On-line Citation Matching Scheme," *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries(JCDL'06)*, pp.276-285, 2006.

구 희 관
1993년~2002년 광운대학교 전자계산학과(학사). 2002년~2004년 광운대학교 컴퓨터과학과(석사). 2004년~현재 과학기술연합대학원대학교 응용정보과학 박사과정. 관심분야는 정보 추출, 사회네트워크, 시멘틱웹

정 한 민
1988년~1992년 포항공과대학교 전자계산학과(학사). 1992년~1994년 포항공과대학교 전자계산학과(석사). 2000년~2003년 포항공과대학교 컴퓨터공학과(박사). 1994년~2000년 한국전자통신연구원 선임연구원. 2000년~2004년 ㈜다이퀘스트 연구소장/기술이사. 2004년~현재 한국과학기술정보연구원 선임연구원. 2005년~현재 과학기술연합대학원대학교 겸임교수. 관심분야는 자연어처리, 시멘틱 웹, 정보 추출, 정보 검색

강 인 수
1988년~1995년 경북대학교 컴퓨터공학과(학사). 1997년~1999년 포항공과대학교 컴퓨터공학과(석사). 2001년~2006년 포항공과대학교 컴퓨터공학과(박사). 1995년~1997년 (주)포스데이터 DBA. 1999년~2001년 포항공과대학교 학술정보원 연구원. 2006년~현재 한국과학기술정보연구원 초청연구원 관심분야는 자연어처리, 시맨틱웹, 정보검색

이 승 우
1997년 2월 경북대학교 컴퓨터공학과(공학사). 1999년 2월 포항공과대학교 컴퓨터공학과(공학석사). 1999년~2000년 포항공과대학교 정보통신연구소 연구원. 2005년 8월 포항공과대학교 컴퓨터공학과(공학박사). 2005년~2006년 대구가톨릭대학교 컴퓨터교육과 강의전담교원. 2006년~현재 한국과학기술정보연구원 선임연구원. 관심분야는 자연어처리, 시맨틱 웹, 정보검색

성 원 경
1987년 2월 연세대학교 불어불문학과(학사). 1989년 2월 연세대학교 불어불문학과(석사). 1996년 12월 프랑스 파리7대학교 언어학과(박사). 1997년~1998년 한국전자통신연구원 Post-doc. 1998년~2001년 L&H Korea(주) 책임연구원. 2001년~2003년 (주)보이스텍 연구개발본부장/상무이사. 2004년~현재 한국과학기술정보연구원 정보시스템연구팀장/책임연구원. 2004년~현재 과학기술연합대학원대학교 겸임교수 관심분야는 자연어처리, 시맨틱 웹