# 인용 필드 정규화와 타입이 인용매칭에 미치는 영향

## Influence of Normalization and Types of Citation Fields on Citation Matching

**구희관, 정한민, 성원경**
한국과학기술정보연구원 정보서비스연구팀

HeeKwan Koo(hkkoo@kisti.re.kr), Hanmin Jung(jhm@kisti.re.kr),
Won-Kyung Sung(wksung@kisti.re.kr)

**요약**

본 논문은 인용필드의 정규화와 타입이 인용매칭에 미치는 영향에 대한 분석을 제시한다. 인용매칭은 같은 논문을 참조하는 인용레코드를 군집화하는 일련의 과정을 지칭한다. 인용매칭은 인용레코드를 구성하고 있는 인용필드들 간의 비교 결과들을 조합하여 인용레코드의 일치 여부를 판별하는 것이다. 인용매칭 단계 내의 인용필드 간 비교를 위하여 인용필드 정규화 및 인용필드 타입에 대한 연구가 필요하였으나, 인용매칭 방법에 대한 연구에 비해 상대적으로 미흡하였다. 본 연구에서는 인용매칭 성능이 인용필드의 정규화 및 인용필드 타입에 따라 달라진다는 것을 보였다. 추가적으로, 정규화를 적용한 다중 필드 결합을 이용한 인용매칭 성능을 분석하였다. 실험결과에 의하면, 인용필드는 정규화를 통하여 전반적인 성능향상이 있었으며, 인용필드 타입에 따라 성능 양상이 다르게 나타났다.

■ 중심어 : | 인용레코드 | 인용필드 | 인용매칭 | 정규화 | 필드타입 |

**Abstract**

In this paper, we present the analysis of the fact that normalization and types of citation fields have an effect to the citation matching. Citation matching indicates the series of grouping process for the citation records referring to the same paper. The citation matching combines the comparison results of citation fields, and determines which citation records are the same. For the citation field comparison in the citation matching phase, studies on the normalization and types of citation fields are needed. But they are relatively insufficient when compared with the studies on citation matching methods. In this research, we showed that the citation matching performance was affected by the normalization and types of citation fields. Additionally, we also analyzed the combination of normalized multiple fields. According to the experimental result, the citation field had the overall performance improvement through a normalization, and the performance mode differently showed up at the citation field type.

■ keyword : | Citation Record | Citation Field | Citation Matching | Normalization | Field Type |

## 1. 서 론

인용매칭(citation matching)은 논문의 참고문헌 영역에서 추출된 인용레코드(citation record)를 대상으로 동일한 논문을 참조하는 인용레코드를 군집화하는 것

접수번호 : #080916-005
접수일자 : 2008년 09월 16일
심사완료일 : 2008년 10월 31일
교신저자 : 구희관, e-mail : hkkoo@kisti.re.kr

으로, 저자 및 논문 간 인용의 흐름 파악 및 피인용지수 계산의 자동화를 위한 핵심 기능 중 하나이다[2][9]. [그림 1]은 인용매칭의 대상이 되는 인용레코드의 예를 보여준다. 인용레코드는 그림에서 보이는 바와 같이 저자명, 논문제목, 게재지명, 게재페이지 등의 인용필드(citation field)들로 구성된다. 이러한 인용필드들은 다양한 인용레코드 기술스타일에 따라 서로 다른 순서와 형식으로 표현되며 일부 인용필드들은 저자의 편집오류 및 데이터 변환 과정 오류 등의 이유로 누락되기도 한다. 이는 동일 논문을 지칭하는 인용레코드들의 불일치를 야기함으로써 인용매칭 문제의 복잡도를 증가시킨다.

| *Aha, D. and Kibler, D.* | *Learning ... An Initial Case study* | *In Proceedings ... Machine Learning* | *pages 24-30* | ... |
|---|---|---|---|---|
| **Author** | **Title** | **PubName** | **Page** | ... |

**그림 1. 인용 필드로 구성된 인용 레코드의 예**

인용레코드불일치 문제는 동일한 인용레코드들임에도 불구하고, 인용필드 기술순서의 상이함에서 야기되는 구조적 불일치와, 개별 인용필드의 기술형식의 상이함에서 야기되는 형태적 불일치로 구분될 수 있다. 구조적 불일치의 예로는, 저자명, 논문제목, 게재지명의 순으로 인용레코드 기술순서가 요구되는 MLA 인용스타일[1]과 저자명, 게재년도, 논문제목 순의 기술 순서를 요구하는 APA 인용스타일[2]의 상이함을 예로 들 수 있다. 반면, 저자명 기술 방식의 차이로 인한 변이형(예: 'David Jones', 'D. Jones', 'Jones, D.')이나, 축약어 기술 방식의 차이로 인한 변이형(예: pp., p., vol., v.) 등이 형태적 불일치의 대표적인 예가 된다.

전술한 인용레코드의 구조적 불일치를 다루기 위해, 기존 연구에서는 인용매칭의 첫 단계로 인용레코드를 인용필드들로 분해하는 인용필드분해(citation fields segmentation: CFS)를 사용하였다[4][6]. 또한, 대응되는 인용필드들 사이의 형태적 불일치를 해소하기 위해 많은 인용매칭 연구자들은 인용필드 정규화와 유사문자열매칭(approximate string matching) 기법들을 병행 활용하였다[1][10][14][15][16]. 정규화는 일련의 변환 규칙 및 전거 파일(authority files)의 사용을 통해 이형태로 출현한 인용필드 값들을 일정한 형태로 변환함으로써 재현율을 증가시켜 인용매칭의 성능을 향상시키는 기법이다. 정규화는 통계적 유사성에 의존하는 학습기반 유사문자열 매칭에 비해 정규화 결과에 대한 예측성과 통제성이 높다는 장점이 있다.

본 논문에서는 정규화를 통한 인용필드의 형태적 불일치 해소가 인용매칭에 미치는 효과에 대한 평가 및 분석을 다뤘다. 기존 연구들은 형태적 불일치 해소를 위해 전처리 관점에서 정규화를 다루었으며 정규화와 인용매칭의 관계를 고려한 연구는 본 저자의 지식범위 내에서는 찾을 수 없다.

정규화 기법으로 이 연구에서는 인용필드의 특성에 따라 대소문자변환, 저자명 약어 처리, 불용어 및 스태밍(stemming) 처리, 축약어 표준화 등의 다양한 방법을 적용한다. 인용매칭 방법으로는 학습데이터 의존성이 없으며 최근 연구에서 높은 성능을 보인 정보검색 기반 방법을 사용한다[5][8][15]. 또한 실험에서 정규화를 통한 개별 인용필드의 인용매칭 기여도를 분석하고 인용필드들을 문자형과 숫자형 타입으로 구분하여 인용매칭 성능에 미치는 특징을 분석한다. 인용필드 분해단계에서 발생하는 오류가 점차적으로 다음 단계로 확산되는 특성을 감안한다면, 문자형 타입에 비해 상대적으로 인식이 용이한 숫자형 타입들의 인용매칭 효과에 대한 분석은 실용적 인용매칭 기법 개발에 큰 도움을 줄 것이다.

## 2. 관련연구

인용필드 정규화는 인용매칭을 연구하는 연구자에 의해 다양하게 수행되었다. 사례별로 살펴보면, CiteSeer 관련 연구에서는 인용레코드에 대해 소문자화, '-' 문자 제거, 인용레코드 선행태그(예: [3], [Giles92]) 제거, 축약어 확장(예: conf. -> conference,

1 MLA Style, style format of the Modern Language Association. http://en.wikipedia.org/wiki/MLA_style

2 APA Style, style format of the American Psychological Association. http://en.wikipedia.org/wiki/APA_style

proc. -> proceedings), 불용어류 단어(예: pp., pages) 삭제 등의 정규화를 적용하였다[14]. McCallum은 저자명, 논문제목, 게재년도, 게재지명의 네 필드를 사용한 인용매칭을 시도하였으며, 각 필드는 소문자화하였고, 논문제목과 게재지명은 60문자 이내의 길이로 제한하였다. 그는 저자명은 제1저자의 성을 이용하는 정규화 기법을 사용했다[1]. Sarawagi는 축약어(예: Phys., Math.), 숫자, 기호 등을 인용레코드에서 제거하는 정규화를 적용했다[15]. 상기의 연구들은 정규화를 전처리 기법으로 다루었으나 정규화가 인용매칭에 미치는 영향에 대한 분석은 제시하지 않았다.

인용필드로 출현한 문자열의 유사도를 계산하는 방법은 크게 문자기반 방법(character-based methods)과 단어기반 방법(token-based methods)으로 나눌 수 있다. 문자 기반 방법으로는 Soundex, Levenshtein/edit distance, Jaro/Jaro-Winkler등이 있으며, 단어 기반 방식은 TF/IDF기반의 Cosine-유사도 방식이나, Jaccard Coefficient 등이 있다[16]. 또한 문자 및 단어 기반 기법들을 동시에 적용하는 방법으로는 SOFT TF/IDF방식이 제안되었다[10]. 대량의 인용 레코드를 대상으로 인용매칭이 수행된다는 점을 감안할 때 단어기반 방법이 문자기반 방법에 비해 확장성이 높은 장점이 있다. 그러한 이유로 인용매칭을 위한 인용레코드 유사도 계산을 위해서는 단어 기반 비교 방법이 주로 사용되었다. Lawrence는 CiteSeer 서비스를 위한 인용매칭 방법으로 가중치가 부여된 단어 기반 비교 방법을 제안하였으며, TF/IDF를 이용한 인용매칭 방법을 사용하였다[5]. Baxter는 블록킹(blocking)이라는 기법을 통해 인용필드의 선택이나 다양한 인용 매칭 생성방법에 대한 연구를 하였다[13].

최근에는 단어기반 유사도 계산을 위해 통계적 방법을 적용하는 연구가 시작되었다. 그 대표적인 예로는, 마르코프 네트워크(Markov networks)와 형식논리(FOPL)을 결합한 방법이나, 인용레코드의 인용필드분해와 인용매칭을 동시에 수행하는 확률적인 모델을 적용한 방법, First-order logic에 확률적인 가중치로 학습하는 방법 등이 이에 해당한다[3][7][11][12]. 단어기반 통계적인 모델을 적용한 인용매칭 방법들은 학습된 가중치가 인용필드에 반영이 된다. 그러나 통계적 모델에 사용되는 가중치는 인용필드의 유사도와 인용매칭과의 관계에 대한 이해를 연구자에게 잘 드러내지 못한다.

## 3. 실험방법

검색엔진을 이용한 단어기반 인용매칭 방법은 간단한 임계값 설정으로 대량의 인용레코드에 대해 신속한 인용매칭 결과를 생성할 수 있는 장점이 있다. 또한, 통계 및 규칙을 적용하여 인용매칭 결과를 재생성 할 수 있다는 면에서 확장성 역시 높다.

3.1에서는 검색엔진 기반 인용매칭의 절차에 대해 기술하고, 3.2에서는 테스트 세트에 대한 소개를 하며, 3.3에서는 이 연구에서 적용한 정규화 방안에 대해 기술한다.

### 3.1 검색엔진 기반 인용매칭

검색엔진 기반 인용매칭의 기본 절차는 다음과 같다. 먼저 인용매칭 대상이 되는 n개 인용레코드를 입력으로 받아 각 인용레코드를 하나의 문서로 고려하여 색인을 수행한다. 다음, 색인된 n개 인용레코드 중 임의의 하나를 질의로 사용하고 나머지 n-1개 인용레코드를 검색 대상 문서 집합으로 고려하여 검색하는 과정을 서로 다른 n개 질의(인용레코드)에 대해 반복한다. 이러한 n번의 검색을 통해 임의의 두 인용레코드 간 유사도가 검색엔진을 통해 질의-문서 유사도의 형태로 얻어진다. 마지막으로 인용레코드 간 유사도를 기반으로 인용레코드들의 군집화가 수행된다. 이 연구에서는 검색엔진 기반 인용매칭 연구들에서 사용된 자바 기반 오픈소스 검색엔진인 [3]Lucene을 이용하였고, 군집화 방법으로 단일링크 응집형 군집법을 적용하였다[1][3][5][8][15]. 색인은 벡터공간모델을 적용하였으며, 질의 방식은 Lucene에서 제공되는 다중 필드 검색을 사용하였다.

### 3.2 실험데이터

3 http://lucene.apache.org/

본 연구에서는 인용매칭 연구에서 일반적으로 많이 사용되는 테스트 세트인 McCallum의 인용매칭 테스트 세트[4]를 사용하였다[1][3][11]. 이 테스트 세트는 인공지능 관련 논문들의 인용레코드를 수집하여 수작업에 의해 인용필드분해 및 인용군집을 생성한 것이다. 하나의 인용레코드는 저자명, 논문제목, 게재지명(논문지/학술대회논문집), 게재년도, 권/호(volume/issue), 게재페이지 등의 주요 인용필드 뿐만 아니라, 출판사, 학술대회 개최 지역/장소, 편집자 등의 필드를 포함하여 총 16개의 필드로 구성되어 있으며, 총 1,879개의 인용레코드가 포함되어 있다. 이 중, 논문제목이 누락된 인용레코드를 제외한 1,838개의 인용레코드에 대해 실험을 수행하였다. 테스트 세트 내의 인용 군집의 수는 187개이며 하나의 인용 군집은 평균적으로 10개의 인용레코드를 포함하고 있다. 테스트 세트에 대한 전처리로 <year>필드가 존재하지 않고 <date>필드가 존재하는 경우 <date>필드에서 연도정보를 추출하여 <year>필드를 자동 생성하였다. 실험에 사용된 인용레코드의 예는 [그림 2]와 같다.

aha1987
〈DocID〉1〈/DocID〉
〈author〉 Aha, D. and Kibler, D. 〈/author〉
〈title〉 Learning Representative Exemplars of Concepts: An Initial Case Study. 〈/title〉
〈booktitle〉 In Proceedings of the Fourth International Conference on Machine Learning, 〈/booktitle〉
〈pages〉 pages 24-30, 〈/pages〉
〈address〉 U. C. Irvine, CA, 〈/address〉
〈year〉1987. 〈/year〉
〈publisher〉Morgan Kaufmann. 〈/publisher〉

**그림 2. 인용 레코드의 예**

실험을 위하여, McCallum 테스트 셋의 16개 인용필드를 필드 발생 비율을 고려하여 저자명, 논문제목, 게재지명, 권/호, 게재페이지, 게재년도, 기타의 7개 필드로 재구성하여 인용매칭을 수행하였다. [표 1]은 전술한 7개 필드 중 기타를 제외한 6개 필드를 문자형과 숫자형 필드 타입으로 구분하고 각 필드의 값이 전체 인용 레코드 집합에서 출현한 비율을 보인 것이다.

**표 1. 인용매칭 테스트 세트의 타입별 필드구성 비율**

| 순위 | 문자형인용필드 | 숫자형 인용필드 |
|---|---|---|
| 1 | 논문제목(100%) | 게재년도(89%) |
| 2 | 저자(99%) | 게재페이지(67%) |
| 3 | 게재지명(81%) | 권/호(46%) |

## 3.3 정규화방법

전술한 바와 같이 정규화는 이형태로 출현한 인용필드 값들을 일정한 형태로 변환함으로써 재현율을 증가시켜 인용매칭의 성능을 향상시키는 효과가 있다. 실험에 적용한 인용필드 별 정규화는 다음과 같다. 전체 인용필드에 공통적으로 문자와 숫자를 제외한 모든 기호를 제거하고 문자를 소문자로 변환하는 기본적인 전처리를 수행하였다. 개별 인용 필드에 대한 정규화 기법들은 다음과 같다.

- 저자명 필드: 논문 저자의 이니셜 제거
  (예: D. Johnson -> Johnson).
- 논문제목 필드: 불용어(stopword) 제거와 포터스태머[5] (Porter stemmer) 적용
- 게재지명 필드: 학술대회논문집 및 논문지를 의미하는 단어의 변이형들 제거
  (예: proc., proceedings, j., journal)
- 게재페이지,권/호 필드: 숫자를 제외한 모든 문자 및 기호를 공백 처리, 로마자 숫자는 아라비아 숫자로 변환
  (예: page 120-130 -> 120 130, III -> 3)

논문제목 필드에 대해서는 다음 여섯 가지 정규화 기법들을 비교해 보았다. TitleN이외에 나머지 모든 기법들은 TitleN의 기본 정규화 처리를 포함한 것이다.

- TitleN: 소문자와 기호 제거만 적용
- TitleP: 포터스태머 적용
- TitleS1: 불용어로 “a, an, the”의 관사들만 사용한 “불용어목록1”을 적용

4 http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz

5 http://www.tartarus.org/martin/PorterStemmer/

- TitleS2: 429개의 불용어로 구성된 “불용어목록2[6]”를 적용
- TitleS1P: 불용어목록1과 포터스태머 적용
- TitleS2P: 불용어목록2와 포터스태머 적용

## 4. 실험결과

실험결과는 크게 두 개의 부분으로 나누어 정리하였다. 4.1절은 개별 인용 필드에 정규화를 적용한 인용 매칭 성능과 인용 필드의 결합에 관한 실험결과를 기술한다. 다만, 개별 필드의 성능 중 논문 제목은 인용 매칭에 중요한 영향을 미칠 수 있는 인용 필드에 해당하기 때문에 첫 실험 결과에 기술된다. 4.2절은 인용 필드의 타입을 문자와 숫자타입으로 나누어, 타입이 가지는 인용 매칭의 성능의 특성을 기술한다.

### 4.1 인용필드 정규화와 인용매칭 성능

- 논문제목 필드 정규화와 인용 매칭 성능

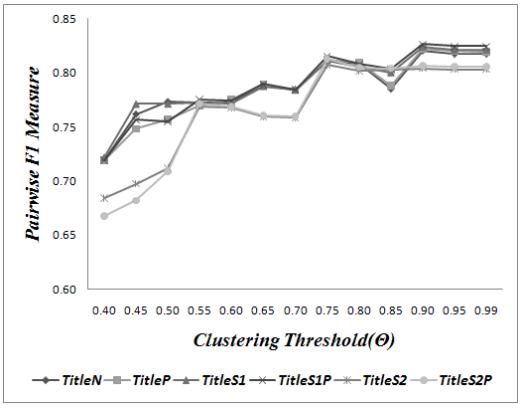


그림 3. 논문제목 필드 전처리와 인용매칭 성능

[그림 3]은 논문제목 필드에 적용한 여섯 가지 정규화 기법들의 인용매칭 성능을 보여준다. 이들 6가지 방법 중 가장 높은 평균 성능을 보인 정규화 방법은 불용어목록1을 적용한 TitleS1이었으며, 가장 낮은 성능을 보인 정규화 방법은 TitleS2였다. 불용어 목록1(TitleS1)은 논문 제목에 관사의 생략이 자주 발생하기 때문에, 관사를 제거함으로써 논문제목을 비교하는 정확률을 높임으로써 성능이 향상된 것으로 보인다. 다만 성능의 차이는 크지 않으며 TitleN대비 1% 내외의 성능 차이를 보인다.

- 단일 필드 정규화와 인용 매칭 성능

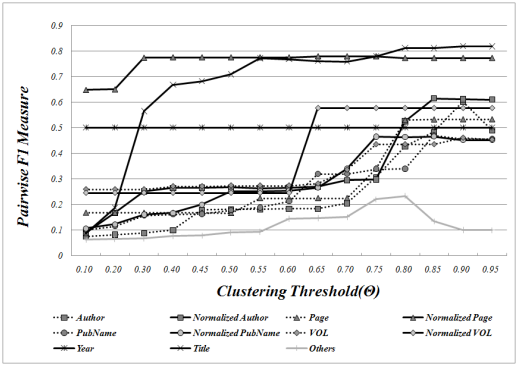


그림 4. 단일 필드를 이용한 인용매칭 성능

[그림 4]는 개별 인용 필드 별 인용매칭 성능을 보인다. 각 필드 별로 정규화 기법의 적용 유무는 실선과 점선으로 표시되었으며, 인용매칭의 성능이 정규화가 적용되면서 성능이 향삼됨을 보이고 있다. [표 2]는 개별 인용 필드를 사용한 인용매칭의 최대 성능을 정리한 것이다. 인용매칭의 성능은 이후 논의에서 정확률, 재현율, 그리고 필드의 값이 누락된 정도를 나타내는 결측비(missing ratio) 관점에서 분석된다.

인용필드는 정규화를 통해 전체적인 성능 향상 이외에 안정적인 성능을 보여주는 구간이 증가되었다. 특히 게재페이지 필드는 임계값과 상관없는 높은 성능을 보이는 구간이 넓어졌다. 또한, 게재페이지 필드는 0.4 이하의 낮은 임계값에서도 일정하게 높은 성능을 보여주기 때문에 인용매칭에 적용할 임계값을 다양하게 설정할 수 있는 장점이 있다. 권/호 및 저자명 필드도 임계값에 상관없는 성능구간이 정규화를 통해서 전체적으로 증가되었음을 확인할 수 있다. 이는 인용매칭을 수행하고자 할 때, 임계값 길이를 넓힐 수 있기 때문에 필드 별 결합을 용이하게 하는 효과가 있을 것으로 분석

6 http://rdsweb2.rdsinc.com/help/stopword_list.htm

된다.

**표 2. 인용필드 별 인용매칭의 최대 성능**

| | 임계값 | 정확률 | 재현율 | Pairwise F1 | Missing Rate(%) |
|---|---|---|---|---|---|
| 논문제목 | 0.90 | 0.7894 | 0.8509 | 0.8190 | 0.00 |
| 게재페이지 | 0.65 | 0.9613 | 0.6564 | 0.7801 | 33.29 |
| 저자명 | 0.85 | 0.4774 | 0.8634 | 0.6148 | 0.16 |
| 권/호 | 0.65 | 0.7706 | 0.4623 | 0.5779 | 54.41 |
| 게재지명 | 0.80 | 0.3260 | 0.8094 | 0.4648 | 17.84 |
| 게재년도 | 0.10 | 0.3347 | 0.8355 | 0.4779 | 10.99 |
| 기타 | 0.80 | 0.4609 | 0.1558 | 0.2329 | 0.00 |

● 다중 필드 조합과 인용 매칭 성능

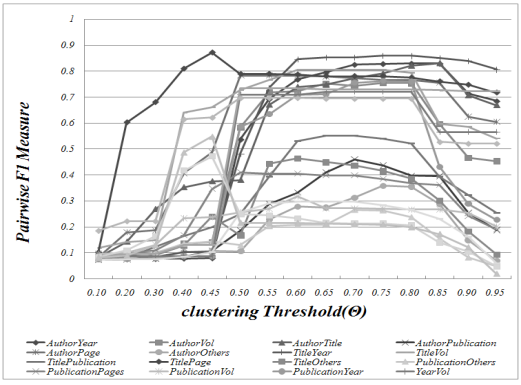


**그림 5. 이중 인용필드를 이용한 인용매칭 성능**

[그림 5]는 서로 다른 인용필드 2개를 조합하여 실험한 인용매칭 성능을 보인다. 이들 필드의 조합은 총 7개 필드를 이용한 21개의 필드 조합으로 생성된다. 최대 성능을 보이는 필드의 조합은 논문 제목과 게재페이지의 결합이었다. 단일 필드를 사용하였을 때보다 2개의 필드를 조합하였을 경우에 최대 성능이 10% 정도 상승하는 효과를 가진다. 이 두 필드의 조합은 단일 필드 실험 중에서 가장 높은 성능을 보였던 두 필드의 결합이었기 때문에 성능이 가장 높으리라 예상되었다. 그러나 성능은 최대 성능을 보였지만 최대 성능을 보인 구간이 매우 짧았으며 이후 임계값의 증가에 따라 급격히 낮은 성능을 보여 이를 보완할 수 있는 추후 연구가 필요하다. 그리고 그 외 20개의 필드 조합 중에서 게재페이지, 게재년도, 권/호, 저자명 중 하나와 논문제목을 결합한 경우와 저자명과 게재년도를 결합한 경우의 총 다섯 가지의 필드결합이 0.8 이상의 인용 매칭 성능을 보였다.

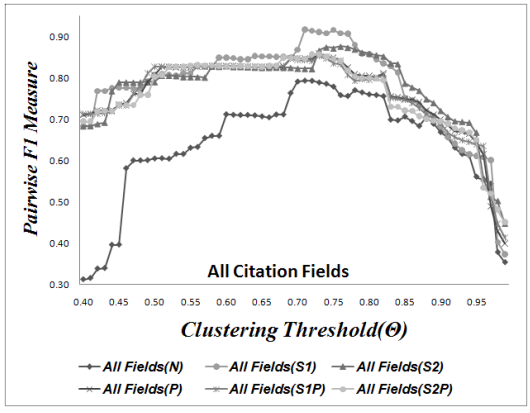


**그림 6. 전체 인용필드를 이용한 인용매칭 성능**

[그림 6]은 전체 인용 필드에 정규화 기법 적용 유무의 인용매칭 성능을 비교하여 보이고 있다. 그림에서 AllFields(N)은 전체 7개 인용필드에 대해 소문자화와 기호제거 이외에 어떤 정규화도 적용하지 않은 것이며 나머지는 전체 인용필드들에 대해 3.3절에 기술된 정규화를 적용한 것이다. 또한 AllFields(X)에서 (안의 X는 논문제목 필드에 적용한 서로 다른 정규화 기법을 표시한 것이다. 예를 들어 AllFields(S2P)는 논문제목에 불용어목록2와 포터스테머를 적용하고 다른 필드들에 대해서는 3.3절의 정규화 기법을 적용한 인용매칭의 성능이다. 전반적으로 정규화를 적용한 인용매칭 성능이 소문자화와 기호제거를 이용한 전처리만 적용한 인용매칭 성능보다 증가된 것을 볼 수 있다.

## 4.2 인용필드 타입과 인용매칭 성능

논문 제목 필드와 게재페이지 필드는 단일 인용 필드를 사용한 인용 매칭의 성능이 가장 높게 측정된다. 따라서 이 두 인용 필드는 각 인용 필드 타입을 대표하는 특성을 잘 드러낸다. [그림 7]은 두 인용 필드들의 인용매칭 성능의 특성을 보인다. 전반적으로 논문제목은 군집화 임계값이 상승함에 따라 재현율이 서서히 감소하면서 정확률이 상승하는 특징을 보이며, 게재페이지는

특정 군집화 임계값에서부터 높은 정확률을 보이지만 낮은 재현율로 인해 전반적인 인용 군집의 성능은 논문 제목 필드보다 낮았다.

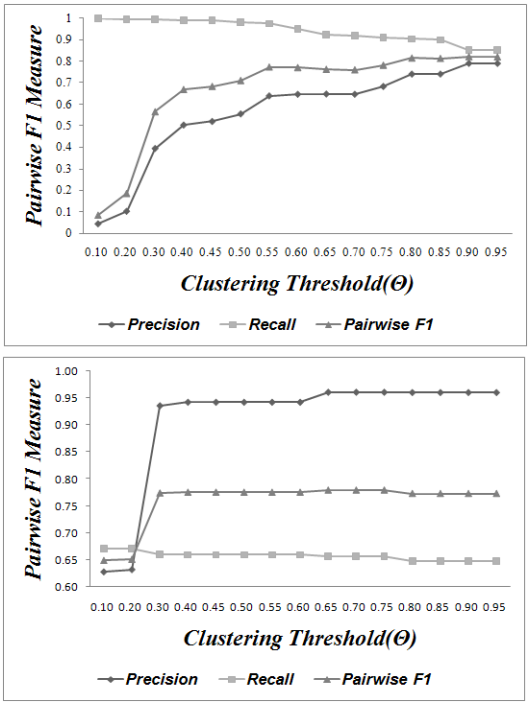


**그림 7. 필드타입 별 인용매칭 성능**

실험에 사용된 7개의 인용필드 중 문자형 타입의 필드는 논문제목, 저자명, 게재지명이다. 이 문자형 타입 인용필드들은 높은 임계값을 사용해 군집화를 수행할 때 일반적으로 높은 성능을 보였다. 이런 특징을 가장 잘 보여주는 필드가 논문제목이다. 그러나 0.95 이상의 임계값을 사용해 군집화를 수행하면, 논문제목이 정확하게 일치하는 인용레코드만을 군집화하기 때문에 미미한 문자 변이형들을 포함하는 인용레코드들을 정답 군집에 포함시키지 않음으로써 재현율을 떨어뜨리게 되어 인용매칭 성능이 낮아지게 된다. 그 외에 저자명과 게재지명 필드의 최대 성능을 보이는 임계값도 논문제목의 경우와 같이 특정 지점 이상의 임계값을 사용하면 다소 성능이 떨어지는 경향을 보였다.

인용필드 중 숫자형 타입의 필드에 해당하는 게재페이지 필드는 논문제목과 비교될만한 높은 성능을 보인다. 좀 더 자세히 기술하면, 숫자로 기술되었기 때문에 문자형 타입 필드보다 정확률이 높다. 그러나 게재페이지 필드는 인용레코드에서 누락되는 비율이 높아 상대적으로 결측비가 크다. 이러한 이유로 인해 게재페이지 필드의 재현율은 논문 제목에 비해 낮지만, 정확률이 매우 높아 전체적인 성능은 논문 제목필드의 인용매칭 성능과 큰 차이를 보이지 않는다. 따라서 기존에 크게 고려되지 않았던 숫자 필드를 이용하여 인용매칭을 설계한다면, 기존 인용매칭 성능을 보다 향상시킬 수 있을 것이다.

문자형과 숫자형 타입의 필드들을 구분하고 이에 대한 성능을 살펴보았으며, 동일한 문자형/숫자형 인용필드들 사이의 인용매칭 성능에도 차이가 있음을 알 수 있다. 이 성능의 차이의 상당부분은 결측비에서 기인하는 것으로 추정된다. 그러나 게재년도는 결측비가 가장 낮음에도 불구하고 성능은 게재페이지가 보이는 최대 성능에 비해 크게 낮은 것을 볼 수 있다. 따라서 결측비 이외에 인용매칭에 영향을 줄 수 있는 요소들을 고려해 볼 수 있다. 문자형 타입에 대해 살펴보면, 인공지능 관련 인용레코드들로 테스트 세트가 구성되었기 때문에 제한된 게재지명의 집합을 가진다. 따라서 저자명이 게재지명 보다 높은 성능을 보인다고 할 수 있다. 그러나, 대량의 인용레코드를 수집했다고 가정하더라도, 게재지명보다 저자수가 더 많을 것으로 예상되기 때문에 일반적으로 저자필드의 지시성이 게재지명보다 크다고 생각될 수 있다. 숫자형 타입에 대해 살펴보면, 가장 크게 기술된 숫자 길이가 다르기 때문에 인용매칭의 성능에 차이를 보인다고 할 수 있다. 따라서 권/호필드는 결측비이외에도, 한 단어를 구성하는 자릿수의 크기와 단어수에서 차이가 나기 때문에 게재년도보다 나은 성능을 보인다고 할 수 있다. 숫자형 타입의 성능을 정리하면 서로 다른 인용레코드 사이에 동일한 값을 공유할 확률이 숫자형 필드를 마다 서로 다르기 때문에 인용매칭의 성능에 차이를 보인다고 할 수 있다.

## 5. 결 론

본 논문에서는 인용필드 정규화와 인용매칭의 관계를 분석하였다. 이를 위해 단어기반 인용레코드 검색엔진을 이용한 TF/IDF를 기반으로 인용매칭을 수행하였다. 실험을 통해 개별 필드가 정규화를 통해 인용매칭 성능을 증가시킴을 다음과 같이 보였다. 첫째, 인용매칭에 대한 개별 필드의 특성을 살펴보았다. 인용필드의 특성을 살펴보고 인용매칭에 대한 이해를 넓혔다. 인용필드들의 유사도가 인용매칭 성능에 미치는 영향력을 측정하여 인용필드 유사도의 정도가 인용매칭의 성능에 어떤 영향을 미치는 지를 밝혔다. 둘째, 인용필드가 인용매칭에 미치는 영향력 측정을 통해 인용필드를 유사도의 특성에 따라 분류하고 인용필드의 관계를 기술했다. 인용필드 분해에서 상대적으로 용이하게 태깅될 수 있는 숫자형 타입의 필드를 규명하고 이들의 성능을 밝혀 인용매칭에 활용할 수 있는 방안을 찾아보았다. 향후 연구로는 인용 필드의 결측비를 보완하여 이를 인용매칭에 적용할 수 있는 방법을 찾는 연구가 수행되어야 할 것이다.

## 참 고 문 헌


[1] A. McCallum, K. Nigam, and L. Ungar, "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.169-178, 2000.

[2] A. Van Raan, "For Your Citations Only? Hot Topics in Bibliometric Analysis," Measurement Interdisciplinary Research and Perspectives, Vol.3, No.50, pp.50-62, 2005.

[3] B. Wellner, A. McCallum, F. Peng, and M. Hay, "An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching," Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp.593-601, 2004.

[4] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting, pp.329-336, 2004.

[5] G, Councill, H. Li, Z. Zhuang, S. Debnath, L. Bolelli, W. Lee, A. Sivasubramaniam, and C. Giles, "Learning Metadata from the Evidence in an On-line Citation Matching Scheme," Proceedings of Joint Conference on Digital Libraries, pp.276-285, 2006.

[6] H. Han, C. Giles, E. Manavoglu, Z. Hongyuan, Z. Zhenyue, and E. Fox, "Automatic Document Metadata Extraction using Support Vector Machines," Proceedings of Joint Conference on Digital Libraries, pp.37-48, 2003.

[7] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser, "Identity Uncertainty and Citation Matching," Advances in Neural Information Processing, pp.1401-1408, 2002.

[8] I. Mansuri and S. Sarawagi,."Integrating Unstructured Data into Relational Databases," Proceedings of the 22th International Conference on Data Engineering, p.29, 2006.

[9] K. Borner, J. Maru, and R. Goldstone, "The Simultaneous Evolution of Author and Paper Networks," Proceedings of the National Academy of Science of the United States, Vol. 101(suppl. 1), pp.5266-5273, 2004.

[10] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive Name Matching in Information Integration," J. of IEEE Intelligent Systems, Vol.18, No.5, pp.16-23, 2003.

[11] M. Richardson and P Domingos, "Markov logic Networks," J. of Machine Learning, Vol.62,

pp.107-136, 2006.

[12] P. Singla and P. Domingos, "Entity Resolution with Markov Logic," Proceedings of the 6th International Conference on Data Mining, pp.572-582, 2006.

[13] R. Baxter, P. Christen, and T. Churches, "A Comparison of Fast Blocking Methods for Record Linkage," Proceedings of the Workshop on Data Cleaning, Record Linkage and Object Consolidation at the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[14] S. Lawrence, C. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," J. of IEEE Computer, Vol.32, No.6, pp.67-71, 1999.

[15] S. Sarawagi, V. Vydiswaran, S. Srinivasan, and K. Bhudhia, "Resolving Citations in a Paper Repository," Proceedings of SIGKDD Explorations, Vol.5, No.2, pp.156-157, 2003.

[16] W. Winkler, "Overview of Record Linkage and Current Research Directions," Technical Report RRS2006/02, US Bureau of the Census, 2006.

## 저 자 소 개

**구 희 관(Heekwan Koo)** 정회원

- 2002년 2월 : 광운대학교 전자계산학과(공학사)
- 2004년 2월 : 광운대학교 콘텐츠학과(공학석사)
- 2004년 ~ 현재 : 과학기술연합대학원대학교 응용정보과학 박사과정

<관심분야> : 정보 추출, 인용 매칭, 인용 네트워크

**정 한 민(Hanmin Jung)** 정회원

- 1992년 2월 : 포항공과대학교 전자계산학과 (학사)
- 1994년 2월 : 포항공과대학교 전자계산학과 (석사)
- 2003년 8월 : 포항공과대학교 컴퓨터공학과 (박사)
- 1994 ~ 2000년 : 한국전자통신연구원 책임연구원
- 2000 ~ 2004년 : ㈜다이퀘스트 연구소장/기술이사
- 2004년 ~ 현재 : 한국과학기술정보연구원 선임연구원
- 2005년 ~ 현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 자연어처리, 시맨틱 웹, 정보 추출, 정보 검색

**성 원 경(Won-Kyung Sung)** 정회원

- 1987년 2월 : 연세대학교 불어불문학과(학사)
- 1989년 2월 : 연세대학교 불어불문학과(석사)
- 1996년 12월 : 프랑스 파리7대학교 언어학과(박사)
- 1997년 ~ 1998년 : 한국전자통신연구원 Post-doc
- 1998 ~ 2001년 : L&H Korea(주) 책임연구원
- 2001 ~ 2003년 : (주)보이스텍 연구개발본부장/상무이사
- 2004년 ~ 현재 : 한국과학기술정보연구원 정보서비스연구팀장/책임연구원
- 2004년 ~ 현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 자연어처리, 시맨틱 웹