인용 필드 정규화와 인용매칭의 관계 연구

구희관* 강인수** 정한민* 성원경*
* 한국과학기술정보연구원 정보서비스연구팀
** 경성대학교 컴퓨터정보학부
{hkkoo, jhm, wksung@kisti.re.kr}, dbaisk@ks.ac.kr

Study on the Relation of Field Normalization with Citation Matching

HeeKwan Koo* In-Su Kang** Hanmin Jung* Won-Kyung Sung*
KISTI (Korea, Institute of Science and Technology Information), ISRLab
KyungSung Univ., Department of Computer Science

요 약

본 논문은 인용필드 정규화와 인용매칭의 관계에 대한 분석을 제시한다. 인용매칭은 논문에서 수집된 인용레코드의 인용필드들 간의 비교 결과를 조합하여 동일 논문의 참조여부를 판별하여 인용레코드를 군집화한다. 따라서 인용매칭에 성능을 높일 수 있는 인용필드와 인용매칭 성능의 관계에 대한 연구가 필요하다. 본 논문에서는 인용필드 정규화 및 필드 별 결합에 의하여 인용매칭 성능이 변화하는 것을 보였다. 또한, 인용매칭 성능을 인용필드 유사도와의 관점에서 분석하였다. 앞으로, 인용필드 정규화 및 특성이 인용매칭에 미치는 영향에 대한 이해를 넓혀, 이를 인용매칭에 활용할 수 있으리라 여겨진다.

1. 서 론

인용매칭(citation matching)은 논문의 참고문헌 영역에서 추출된 인용레코드(citation record)를 대상으로 동일한 논문을 참조하는 인용레코드를 군집화하는 것이다. 그림 1은 인용매칭의 대상이 되는 인용레코드의 예를 보여준다. 인용레코드는 그림에서 보이는 바와 같이 저자명, 논문제목, 게재지명, 게재페이지 등의 인용필드(citation field)들로 구성된다.

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

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 문제를 어렵게 만든다.

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[1] 인용스타일과 저자명, 게재년도, 논문제목 순의 기술 순서를 요구하는 APA[2] 인용스타일의 상이함을 예로 들 수 있다.구조적 불일치를 다루기 위해, 기존 연구에서는 인용매칭의 첫 단계로 인용레코드를 인용필드들로 분해하는 인용필드분해(citation fields segmentation: CFS)를 사용하여 구조적인 불일치문제를 해결하려하였다[3, 5].

형태적 불일치의 예로는, 저자명 기술 방식의 차이로 인한 변이형(예: 'David Jones', 'D. Jones')이나, 축약어 기술 방식의 차이로 인한 변이형(예: pp., p., vol., v.) 등을 대표적인 예로 들 수 있다. 따라서 대응되는 인용필드들 사이의 형태적 불일치를 해소하기 위해 많은 인용매칭 연구자들은 인용필드 정규화와 유사문자열매칭(approximate string matching) 기법들을 병행 활용하였다[1, 7, 9, 10, 11].

인용필드 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. 또한, 통계적 유사성에 의존하는 학습기반 유사문자열 매칭에 비해 정규화 결과에 대한 예측성과 통제성이 높다는 장점이 있다. 따라서 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, proc. -> proceedings), 불용어류 단어(예: pp., pages) 삭제 등의 정규화를 적용하였다[9]. McCallum은 저자명, 논문제목, 게재년도, 게재지명의 네 필드를 사용한 인용매칭을 시도하였으며, 각 필드는 소문자화하였고, 논문제목과 게재

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

지명은 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[10]. 상기의 연구들은 정규화를 전처리 기법으로 다루었으나 정규화가 인용매칭에 미치는 영향에 대한 분석은 제시하지 않았다.

본 논문에서는 정규화를 통한 인용필드의 형태적 불일치 해소 및 인용 필드의 결합이 인용매칭 성능에 미치는 영향을 분석하였다. 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[4, 6, 10].

2. 인용필드 정규화 및 인용매칭 방법

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.

검색엔진 기반 인용매칭의 기본 절차는 다음과 같다. 먼저 인용매칭 대상이 되는 n개 인용레코드를 입력으로 받아 각 인용레코드를 하나의 문서로 고려하여 색인을 수행한다. 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 연구들[4, 6, 10]에서 사용된 자바 기반 오픈 소스 검색엔진인 Lucene[3]을 이용하였고, 군집화 방법으로 단일링크 응집형 군집법[1, 2]을 적용하였다. 색인은 벡터공간모델을 적용하였으며, 질의 방식은 Lucene에서 제공되는 다중 필드 검색을 사용하였다.

테스트 세트는 인용매칭 연구에서 일반적으로 많이 사용되는 테스트 세트인 McCallum의 인용매칭 테스트 세트를 사용하였다[1, 2, 8]. 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 인용군집은 평균적으로 10개의 인용레코드를 포함하고 있다. <year>필드의 값이 없을 경우, <date>필드에서 연도정보를 추출하여 <year>필드를 생성하였다.

```
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 셋[4]의 16개 인용필드를 필드 발생 비율을 고려하여 저자명, 논문제목, 게재지명, 권/호, 게재페이지, 게재년도, 기타의 7개 필드로 재구성하여 인용매칭을 수행하였다.

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

3. 실험결과

3) http://lucene.apache.org/

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

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

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

4.1 인용필드 정규화와 인용매칭 성능

● 논문제목 필드 정규화와 인용 매칭 성능

그림 3, 표 1은 논문제목 필드에 적용한 여섯 가지 정규화 기법들의 인용매칭 성능을 보여준다. 이들 6가지 방법 중 가장 높은 평균 성능을 보인 정규화 방법은 불용어목록1을 적용한 *TitleS1*이었으며, 가장 낮은 성능을 보인 정규화 방법은 *TitleS2*였다. 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 *TitleN*대비 1% 내외의 성능 차이를 보인다.

표 1 논문제목 필드 정규화와 인용매칭 성능

	TitleN	*TitleP*	*TitleS1*	*TitleS1P*	*TitleS2*	*TitleS2P*
0.40	0.7192	0.7188	0.7220	0.7188	0.6837	0.6674
0.45	0.7619	0.7481	0.7713	0.7565	0.6971	0.6819
0.50	0.7733	0.7568	0.7716	0.7550	0.7113	0.7092
0.55	0.7728	0.7696	0.7731	0.7753	0.7693	0.7716
0.60	0.7727	0.7751	0.7722	0.7735	0.7681	0.7687
0.65	0.7879	0.7892	0.7877	0.7893	0.7591	0.7607
0.70	0.7846	0.7841	0.7844	0.7844	0.7582	0.7596
0.75	0.8114	0.8121	0.8117	0.8153	0.8077	0.8132
0.80	0.8082	0.8092	0.8060	0.8079	0.8021	0.8052
0.85	0.7851	0.7883	0.8005	0.8034	0.8027	0.8043
0.90	0.8202	0.8226	0.8242	0.8264	0.8044	0.8066
0.95	0.8172	0.8204	0.8215	0.8244	0.8035	0.8057
0.99	0.8172	0.8195	0.8215	0.8244	0.8035	0.8057
평 균	0.7871	0.7857	0.7898	0.7888	0.7670	0.7661
분 산	0.0286	0.0317	0.0288	0.0326	0.0441	0.0501
표준편차	0.0008	0.0010	0.0008	0.0011	0.0019	0.0025

● 단일 필드 정규화와 인용 매칭 성능

그림 4는 개별 인용 필드 별 인용매칭 성능을 보인다. 각 필드 별로 정규화 기법의 적용 유무는 실선과 점선으로 표시되었으며, 인용매칭의 성능이 정규화가 적용되면서 성능이 향상됨을 보이고 있다. 표 1은 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 보이기 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 분석된다.

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

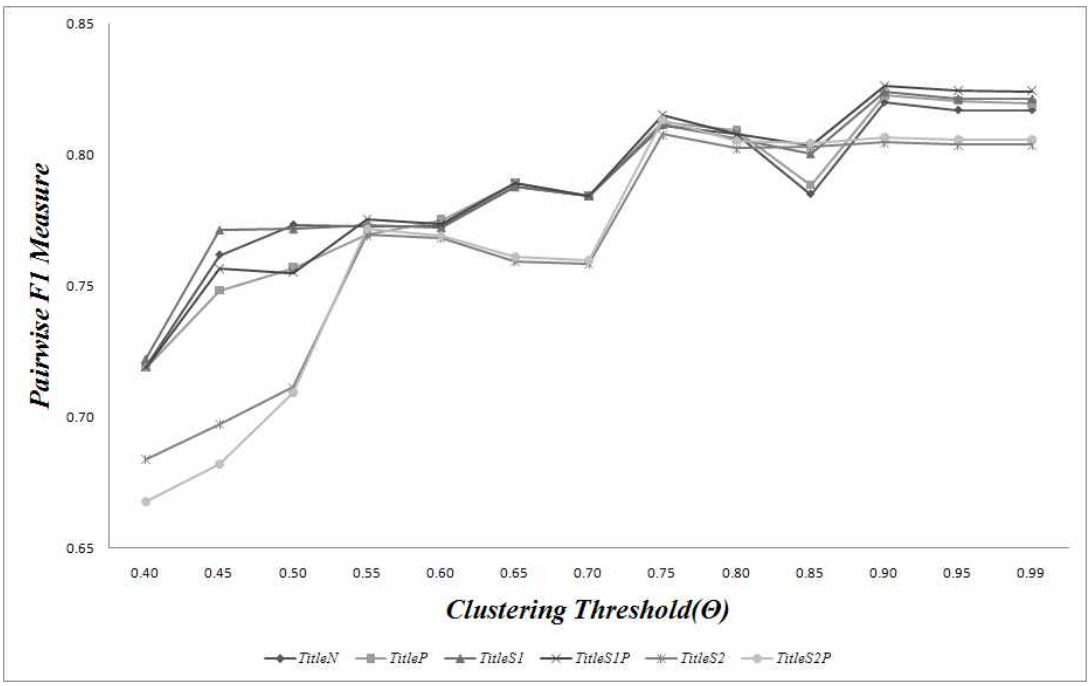

그림 3 논문제목 필드 정규화와 인용매칭 성능

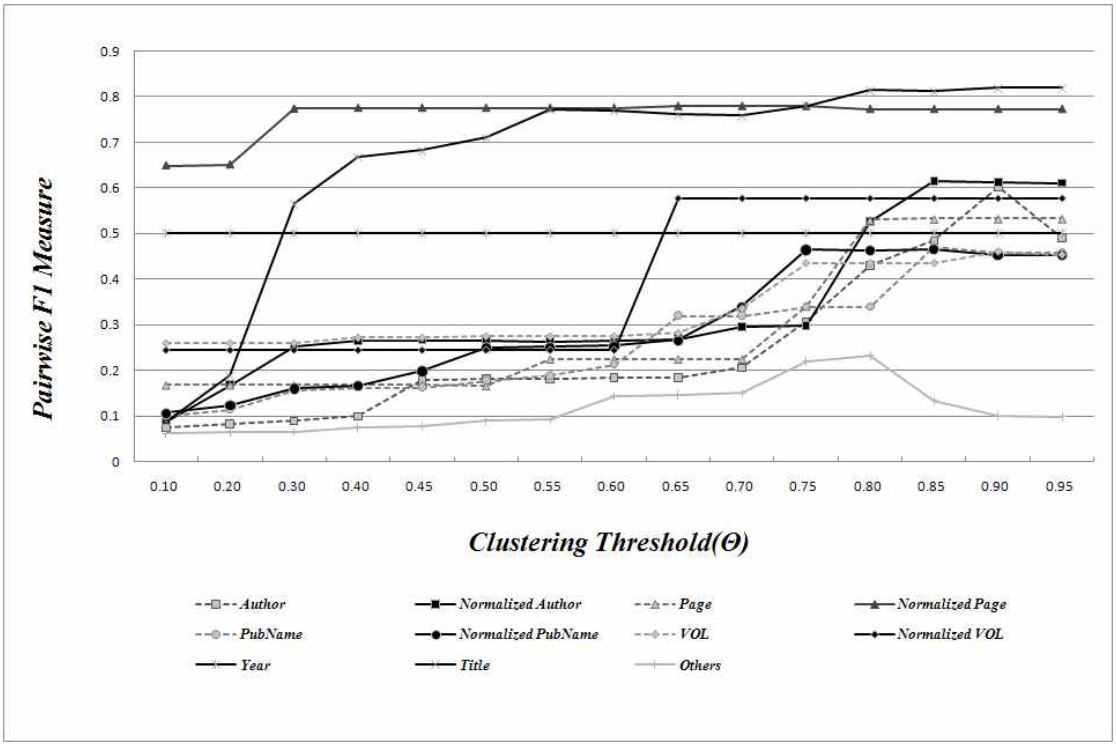

그림 4 단일 필드를 이용한 인용매칭 성능

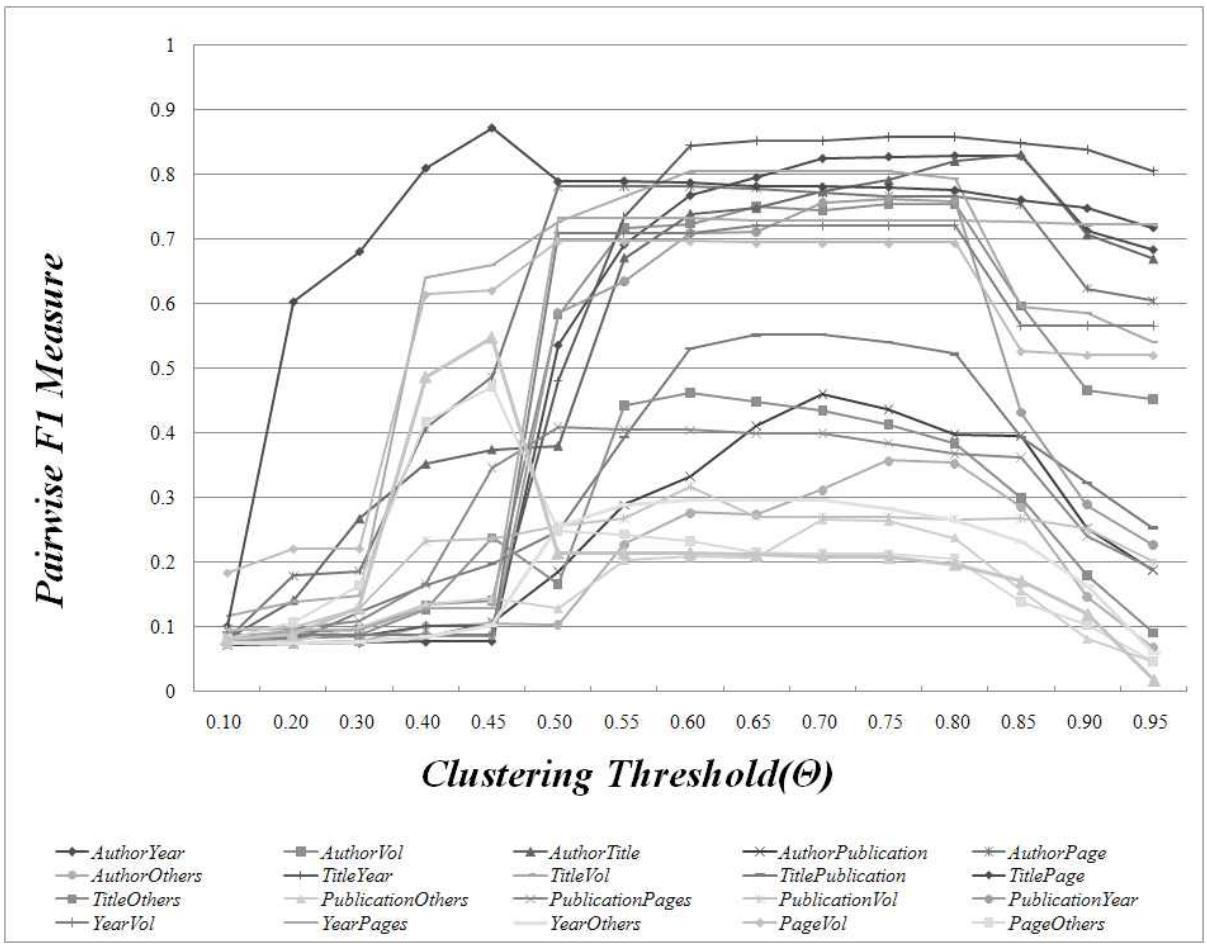

그림 5 이중 인용필드를 이용한 인용매칭 성능

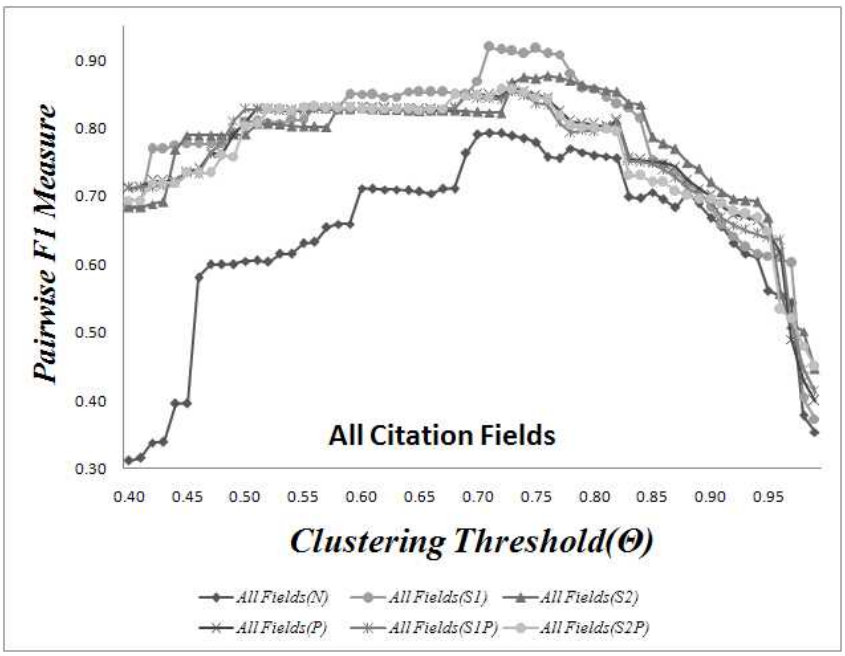

그림 6 전체 인용필드를 이용한 인용매칭 성능

● 다중 필드 정규화와 인용 매칭 성능

그림 5는 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 6은 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 3장에 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 불용어 목록2와 포터스테머를 적용하고 다른 필드들에 대해서는 정규화 기법을 적용한 인용매칭의 성능이다. 전반적으로 정규화를 적용한 인용매칭 성능이 소문자화와 기호제거를 이용한 전처리만 적용한 인용매칭 성능보다 증가된 것을 볼 수 있다.

4. 결론

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 보였다. 그리고 개별 필드의 결합을 이용한 인용매칭 성능을 살펴보았다. 또한, 인용필드들의 유사도가 인용매칭 성능에 미치는 영향력을 측정하여 인용필드 유사도의 정도가 인용매칭의 성능에 어떤 영향을 미치는지를 밝혔다. 향후 연구로는 인용 필드의 결측비를 보완하여 이를 인용매칭에 적용할 수 있는 방법을 찾는 연구가 수행되어야 할 것이다.

5. 참고문헌

[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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] M. Richardson and P Domingos, "Markov logic Networks," J. of Machine Learning, Vol.62, pp.107-136, 2006.

[9] 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.

[10] 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.

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