개별 인용 필드의 인용 매칭에 대한 영향력 평가

Evaluating an Influence of Individual Citation Field on Citation Matching

구희관, 강인수, 정한민, 이승우, 성원경
한국과학기술정보연구원, 정보기술개발단

HeeKwan Koo, In-Su Kang, Hanmin Jung,
Seung-Woo Lee, Won-Kyung Sung
Korea Institute of Science and Technology Information

요약

인용 매칭(Citation Matching, CM)은 동일한 논문을 지칭하는 인용레코드(Citation Record)를 군집화하는 방법이다. 일반적으로, 저자, 논문제목, 게재지명이나 출판연도 등의 인용 필드로 구분하는 인용 필드 분해가 인용 매칭 보다 선행하게 된다. 상당히 많은 연구가 인용 매칭과 인용 필드 분해의 문제를 해결하고자 했지만, 인용 필드 분해와 인용 매칭과의 상관관계에 대한 연구는 부족하였다. 인용 매칭에 대한 인용 필드 분해의 여러 측면 중에, 본 논문은 인용 매칭에 가장 영향력이 있는 인용 필드를 밝히고자 한다. 첫 번째 시도로, 수작업으로 인용 필드 분해를 수행한 다양한 크기의 인용 필드 집합에 대하여 인용 매칭의 성능을 비교하였고, 그 결과 많은 인용 필드를 사용한 인용 매칭이 인용 레코드를 더 잘 군집화 할 수 있다는 것을 확인하였다.

Abstract

Citation matching (CM) is a method for clustering citation records that refer to the same paper. Normally, CM is preceded by citation field segmentation (CFS) which divides a citation record into its fields such as author(s), a title, a title of publication, year, etc. Although many studies have attacked CFS and CM, the relationship between CFS and CM was not sufficiently explored. Among many aspects of the effect of CFS on CM, this study concentrates on what citation fields should identify for CM. As its first attempt, we compared CM performances over different sets of citation fields manually segmented, and confirmed that the use of more citation fields help CM to cluster citation records better.

I. 서 론

연구자가 필요한 논문을 찾으려 할 때, 논문의 선택에 다양하게 활용될 수 있는 정보가 인용 정보이다. 문헌과 문헌과의 인용은 최근 연구의 흐름을 추적할 수 있게 하는 도구 일 뿐만 아니라 연구자가 연구 분야에 대한 변화된 연구의 흐름을 파악하거나, 정해진 분야의 범주를 넘어 다양한 연구를 수행하는 연구자에게 꼭 필요한 정보가 된다.

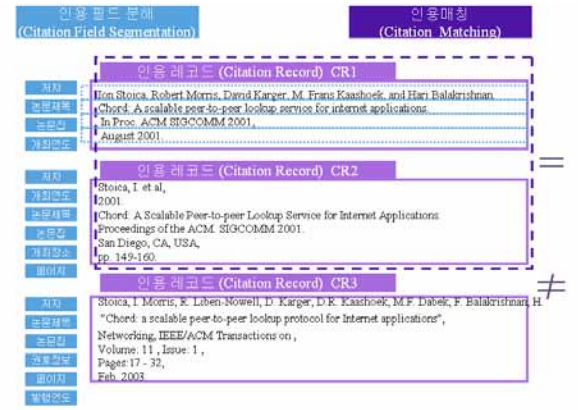
인용 정보를 생성하기 위해서 필요한 단계 중 가장 기본 적인 단계는 문헌 간의 인용관계를 파악하는 것이다. 문헌간의 인용 관계를 형성하는 단계로서, 인용 매칭(Citation Matching, CM)은 동일한 논문을 지칭하는 인용레코드(Citation Record)를 군집화(Clustering)하는 것을 말한다. 인용 레코드는 논문에서 기술된 하나의 인용을 말한다. 인용 레코드는 여러 인용 필드로 구성이 되는데, 대표적인 필드로 저자, 논문제목, 게제지명 등을 예로 들 수 있다.

그림 1은 인용 필드 분해와 인용 매칭의 예를 보이고 있다. 자동적인 인용 필드 분해(Citation Field Segmentation)는 인용 레코드를 각각의 필드(e.g., 저자, 논문제목, 게재지 등)로 구분하는 것이다. 그림 내의 인용 레코드 CR1은 인용 필드 분해 후 대응하는 필드가 점선으로 나누어져 있는 것을 볼 수 있다. 인용 레코드 CR1, CR2은 인용 레코드의 구조가 다르다. 인용 레코드 CR1은 저자, 논문제목, 논문집, 개최연도로 구성이 되며, 인용 레코드 CR2는 저자, 개최연도, 논문제목, 논문집, 개최장소, 페이지 등으로 구성이 된다. 인용 필드 분해를 인용 레코드에 적용하게 되면, 인용 필드 간 매칭을 통해 인용 매칭을 수행할 수 있는 장점을 가진다. 인용 매칭은 그림 1에서 굵은 점선으로 표시되어 있는데, 인용 레코드CR1과 인용레코드 CR2는 인용 매칭이 완결된 후, 하나의 군집으로 생성된다. 인용 레코드 CR3은 필드 별로 나누어 인용 매칭을 고려한다면, 인용 레코드 CR1과 CR2와 저자, 논문제목 등에서는 상당히 유사하게 보이나 기술된 논문집, 발행연도, 페이지 등이 다르기 때문에 생성된 인용군집에 포함이 되지 않는다.

본 논문에서는 인용 필드를 고려한 인용 매칭 성능 측정을 통해 우선적으로 인용 매칭에 고려해야 할 인용 필드를 찾아내고자 한다. 인용 필드 분해를 통해 인용 매칭을 수행하는 경우, 사용되는 필드에 따라 인용 매칭의 성능은 다르게 나타날

것이다. 인용 매칭에 사용된 인용 필드가 인용 매칭에 미치는 영향에 대해 평가를 수행하여, 이를 기준으로 필수적인 인용 필드를 분해 할 때, 우선적으로 추출하거나, 인용 매칭에서 인용 필드의 가중치 부여의 근거가 될 수 있으리라 기대된다.

논문은 2장에서 관련연구를 기술하고, 3장 및 4장에서 실험 방법 및 결과에 대해 설명하고, 최종적으로 5장에서 결론 및 향후 연구에 관해 언급한다.



▶▶ 그림 1. 인용 필드 분해와 인용 매칭의 예

II. 관련연구

CiteSeer는 인용 매칭 알고리즘으로 4가지 군집화 방법을 제안하였으며, 이중 가장 높은 성능을 보인 방법은 단어와 구 비교 알고리즘(Word and Phrase Matching Algorithm)「2」「3」이었다. 단어와 구 비교 알고리즘은 인용 레코드(Citation Record)에 일치하는 단어 수와 연속으로 발생하는 단어에 대해 가중치를 부여하여, 군집화하는 알고리즘이다. 이 알고리즘은 몇 가지 문제점을 포함할 수 있는데, 인용 레코드에서 연속된 단어의 일치가 발생하는 제목에 의존하는 인용 매칭 알고리즘이라는 한계를 가진다.

최근 CiteSeer는 일괄적인 인용 매칭 알고리즘의 해결을 위해 검색엔진을 이용한 인용 매칭 방법을 제안하였다.「1」 제안된 방법은 임의의 인용을 선택한 후, 검색엔진에 질의하여, 검색된 일정한 편집 거리(Edit distance) 내에 인용들을 대상으로 인용 매칭을 수행하는 방법이었다. 이 방법은 이미 CiteSeer에서 일차적으로 가공된 인용을 수집하여 실험을 진행하였기 때문에, 기존 CiteSeer 인용 매칭 방법에 의존적이다. 또한, 검색엔진에 사용된 인용 필드가 저자와 논문 제목만으로 한정되게 사용되었기 때문에, 인용된 논문의 게재지 명(논문지명, 학술대회 논문집명) 및 권호 정보, 발간연도(개최연도) 등의 정보를 활용하지 않는다.

III. 실험방법

본 장은 실험 데이터 셋과 전체 실험에 관해 기술한다. 1절은 실험에 사용된 데이터 셋에 대해 설명하며, 2절은 전체 실험에 사용된 절차에 관해 설명한다.

1. 실험데이터 셋

본 논문의 실험 데이터로는 McCallum의 인용 매칭 테스트 셋을 사용하였다. 이 테스트 셋은 인공지능 관련 논문들의 인용 레코드(Citation Record)를 수집하여 인용 필드 분해 및 인용 군집을 수작업으로 수행하여 만들어진 테스트 셋이다. 하나의 인용 레코드는, 저자, 논문 제목, 게재지 명칭(논문지/학술대회 논문집), 연도, 권호정보, 페이지 번호 등의 주요 인용 필드뿐만이 아니라, 출판사, 학술대회 개최 지역/장소, 편집자 등 실제 인용 레코드를 총 16개의 필드로 구분하며, 1,879개의 인용 레코드가 포함되어 있다. 이 중, 본 논문에서는, 논문 제목이 누락된 인용 레코드를 제외한 1,838개의 인용 레코드에 대해 실험을 수행하였다. 인용 군집의 수는 187개이며, 평균 인용 군집의 크기는 10개 정도였다. <date>필드와 <year>가 일관성있게 기술되지 않아, 실험을 위해 “<year>”필드가 존재하지 않고“<date>”로 연도가 기술 되어 있는 경우에는, 따로 <date>필드에서 연도정보를 추출하여 “<year>”필드로 추가하였다. 인용레코드에 대한 예는 다음과 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

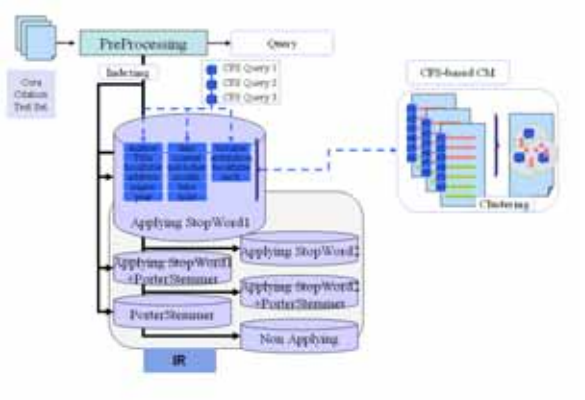
2. 인용 매칭 및 군집화

실험에 사용한 검색엔진은 자바기반 오픈소스 검색엔진 Lucene으로, 가장 최근 버전인 2.2 버전을 사용하였다. 이미 인용 매칭 실험에 Lucene이 이용되었기 때문에, 기존 실험과의 도구적인 차이는 크게 없으리라 여겨진다.「1,4,6」

그림 2는 인용 매칭 실험에서 사용된 전체적인 절차를 설명한다. 첫 번째, 인용 매칭 테스트 셋을 색인을 이용하여 구성한다. 두 번째, 구성된 색인에 테스트 셋 내의 인용을 질의하고, 검색이 수행되면, 검색 결과의 유사도 값을 이용하여 군집을 생성한다. 최종적으로 군집의 성능을 측정한다.

단계별로 나누면, 첫 단계는 색인 단계이다. 문자와 숫자를 제외한 모든 기호를 제거하고 문자를 소문자로 변환하는 전처리 단계를 거친다. 인용 매칭 테스트 셋을 색인으로 구성할 때

는, 전체 테스트 셋 내의 모든 필드에 대해 색인을 생성한다. 색인 생성 방법은 공백으로 분리하는 모든 문자와 숫자를 색인으로 생성하였기 때문에, 숫자로 구성되어 있는 페이지 정보나 권호 정보를 질의해도 검색 결과를 생성할 수 있다.



▶▶ 그림 2. 인용 매칭 절차

그림 2에서 필드별 질의(CFS Query)를 생성하여 색인에 검색을 수행하고, 군집화하는 결과를 보여주고 있다. 인용 필드를 조합하는 질의하는 다중 필드 질의는 Lucene이 필드별 가중치를 따로 부여하지 않고 조합하여 질의한다.

검색에 이용된 검색 모델은 벡터 모델인데[5], Lucene이 사용하는 유사도 계산 방법은 다음과 같다.

$$score(q,d) = coord(q,d) \times queryNorm(q) \times \left(\sum_{t \in q} (tf(t \in d) \times idf(t)^2 \times Boost(t.field \in d) \times Norm(t,d)) \right)$$

질의(q)와 문서(d)의 유사도는 tf(term frequency)와 idf (inverse document frequency)의 제곱값에, 색인할 때 설정한 가중치(Boost(t.field ∈ d))와 색인할 때 필드에 포함된 단어 개수 및 길이를 정규화한 값(Norm(t,d))을 곱한 값의 합에 문서 내에 포함된 개별 검색어의 개수에 대한 조절값(coord(q,d))과 개별검색어의 가중치 값에 제곱의 합을 정규화한 값(queryNorm(q))의 곱으로 구성된다.

생성된 결과 내에 임계치를 기준으로 군집을 생성하여 이를 평가한다. 즉 특정 임계치 값의 결과는 같은 인용 군집 내에 포함되어 군집화를 수행한다.

그림 3은 검색 결과 내에 특정 임계값 이상의 인용에 대해 군집화(Single-link Agglomerative Clustering)를 수행하는 것을 보여준다. 이것은 검색 결과 내에 인용에 대해 관계를 설정하게 함으로써, 군집화를 수행한다. 예를 들어, 그림 3의 좌측의 첫 번째 단계에서 임계치 이내의 검색 결과인 R1의 "C2, C3, C4"는 인용 군집 1에 포함되게 된다. 두 번째 단계의 인용 군집 1과 인용 군집 2는 "C4"라는 인용을 함께 포함하고 있기 때문에 하나의 군집이 되며, 이와 마찬가지로 군집 N과 군집 2는 "C4, C5"라는 인용을 공통으로 포함하고 있기 때문에 하나의 군집으로 형성이 된다.

▶▶ 그림 3. 군집화 절차(Single-link Aggomerative Clustering)

인용 매칭 성능은 각각의 군집의 F1 성능을 이용한다. 군집의 정확률은 정답 군집의 요소가 얼마나 정확하게 해당 군집에 포함되어 있는 개수를 의미하고, 군집의 재현률은 검색 결과를 이용해 생성된 인용 군집이 많은 정답 인용을 포함하고 있는 가로 측정하였으며, 이를 이용하여 F1을 계산하여 군집의 성능을 측정한다.

IV. 실험 결과

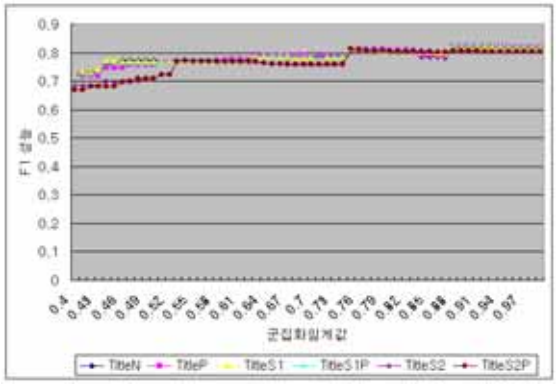
표 1은 각각의 개별 필드가 가지는 인용 매칭의 영향력을 평가한 것이다. 평가 방법은 각 단계마다 가장 영향력이 있는 필드를 선택하고 그 외의 필드를 이용해서 가장 좋은 성능을 보이는 필드를 추가하는 방식으로 진행했다. 다시 말하면, 처음 1레벨에서는 논문 제목 필드를 사용하여 최대 성능을 측정하였다. 논문 제목은 거의 변하지 않고 가장 많은 단어를 포함하고 있기 때문에 인용매칭에 가장 큰 영향을 가진다. 이렇게 레벨 1을 측정하고 나면 나머지 필드들과 결합하여 성능을 측정하고 가장 좋은 성능을 보이는 저자 필드를 추가하는 방식으로 실험을 진행했다.

그림 4는 논문 제목을 이용하여 성능을 측정한 것이다. 그림 4의 도표에서 X축은 군집에 사용된 인용 검색의 임계값이며, Y축은 F1의 성능이다. 레벨 1이외의 모든 레벨에서의 임계값은 모든 필드에게 적절한 임계값처럼 보이는 군집화 임계값 0.7을 사용하여 실험하였다. 그림 4에서 군집화의 임계값이 증가함에 따라 성능이 미세하게 증가하다 임계치 0.9주변에서 미세하게 감소하는 것을 볼 수 있다. 즉 논문 제목 필드 만을 이용하여 인용 매칭을 수행하게 된다면 0.9가 가장 적절한 임계치라 할 수 있다. 0.9 이상의 임계치값은 재현율이 낮게 나타나 성능이 떨어진다. 0.9 이내에서는 일반적인 경향은 임계치 값이 증가함에 따라 정확율이 증가하면서 성능이 조금씩 증가하는 것을 알 수 있다.

대상 필드는 인용 매칭의 주요 필드인 저자(Author), 논문 제목(Title), 게재지명(PubName), 권호정보(Vol), 발행연도

(Year), 페이지(Page)를 선택하여 그 성능을 측정하였다. 게재지명은 논문지 이름과 학술대회 논문집 이름 중에 기술되어 있는 것을 선택했다. 따라서 논문의 종류에 상관없이 하나의 필드로 측정이 가능하도록 했다.

표 1의 결과에서 보여주듯이 하나의 필드가 선택되고 나서 다른 필드가 가지는 인용 매칭의 성능은 이전 단계에서 보여주는 성능과는 조금 다르게 보이는 데, 2단계에서 출판명이 두 번째의 높은 성능을 보이지만 이후에는 가장 마지막 단계까지 좋지 않는 성능을 보이는 것으로 나타났다. 정리하면, 인용 매칭에 가장 높게 기여를 하는 필드는 논문제목, 저자, 권호정보, 페이지, 발행연도, 게재지명이었다.



▶▶ 그림 4. 레벨 1 논문제목에 따른 성능 측정

[표 1] 필드별 인용 매칭 성능

Level	Default Field	Append Field	F1
1		Title	0.76
2	Title	Author	0.806
		PubName	0.803
		Page	0.794
		Year	0.791
		Vol	0.78
3	Title Author	Vol	0.85
		Page	0.826
		Year	0.84
		PubName	0.827
4	Title Author Vol	Page	0.864
		Year	0.856
		PubName	0.832
5	Title Author Vol Page	Year	0.86
		PubName	0.828
6	Title Author Vol Page Year	PubName	0.881

V. 결 론

본 논문은 인용 매칭에서 인용 필드의 영향력을 알아보기 위해, 인용 필드 영향을 분석하였다. 인용을 구성하는 각 인용 필드의 인용 매칭에 관한 영향력을 밝혀 해당 필드 별 중요성을 평가하였다. 또한 필드를 많이 조합하면 할수록 인용 매칭의 성능이 증가한다는 것을 실험적으로 보였다.

인용 매칭은, 하나의 실체를 가진 여러 형태적 변이형을 어떻게 형태적 군집화를 통해 이를 하나의 실체를 갖는 변이형인가를 판별하여 같은 인용으로 묶어 내는가에 대한 문제이다.

예전에는 인용의 대상이 주로 논문 등을 대상으로 수집되고 있으나, 오늘날에는 웹저널이나 블로그 등 다양한 매체에서의 인용이 발생하기 때문에 자동적인 인용 매칭의 방법은 통합적인 인용을 측정할 수 있는 방법으로도 새로운 의미를 가질 수 있을 것이다.

더욱이, 문헌의 인용이 잘 구성이 된다면, 저자 별로 인용을 측정해 저자의 전문성 평가나 공저자 네트워크에 인용을 반영하는 등의 다양한 연구에 적용 가능할 것이다.

▌참 고 문 헌▐

[1] Council, G., Li, H., Zhuang, Z., Debnath, S., Bolelli, L., Lee, W., Sivasubramaniam, A., Giles., C. L., "Learning metadata from the evidence in an on-line citation matching scheme," Joint Conference on Digital Libraries 2006 (JCDL 2006),pp.276-285, 2006.

[2] Lawrence, S., Giles, C. L., Bollacker, K. D., "Autonomous citation matching," Proceedings of the third annual conference on Autonomous Agents, Seattle, Washington, United States, pp. 392 - 393, 1999.

[3] Lawrence, S., Giles, C. L., and Bollacker, K., "Digital libraries and autonomous citation indexing." IEEE Computer, Vol. 32, No.6, pp. 67-71, 1999.

[4] Mansuri, I. R., Sarawagi, S., "Integrating Unstructured Data into Relational Databases," Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), pp-29, 2006

[5] McCallum, A., Nigam K., and Ungar, L., "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," In Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000). 2000.

[6] Sarawagi, S., Vydiswaran, V. G. V., Srinivasan, S., Bhudhia, K., "Resolving citations in a paper repository," SIGKDD Explorations Vol. 5, No. 2, pp.156-157, 2003.