한국컴퓨터정보학회 동계학술대회 논문집 제24권 제1호 (2016. 1)

단어 공기 통계 정보 기반 색인어 추출을 활용한 문서 유사도 검사 알고리즘

김진규[O], 이승철[*], 박기봉[*], 허덕행[*]
[O]창원문성대학교 빅데이터센터
[*]창원문성대학교 빅데이터센터
e-mail:{jinqkim1O, num1ysc[*]}@gmail.com, drebong2@hotmail.com[*], dhhuh01@cmu.ac.kr[*]

Document Content Similarity Detection Algorithm Using Word Cooccurrence Statistical Information Based Keyword Extraction

Jinkyu Kim[O], Seungchul Yi[*], Kibong Park[*], Huhduck Haing[*]
[O]Bigdata Center, Changwon Moonsung University
[*]Bigdata Center, Changwon Moonsung University

● 요 약 ●

빠른 속도로 쏟아지고 있는 각종 발행물, 논문들에 대한 표절 검토는 표절 검출 알고리즘을 통해 직접적인 복제, 짜깁기, 말 바꾸어 쓰기 등을 검토하거나 표절 검토자가 직접 해당 문서의 키워드를 검색하여 확인하는 방식으로 이루어지고 있다. 하지만 점점 더 늘어나는 방대한 양의 문서들에 대한 표절 검토 작업은 더욱 정교한 검토 방법론을 필요로 하고 있으며, 이를 돕기 위해 문서의 직접적인 단어나 복제 비교에서 더 나아가 문서의 내용을 비교하여 비슷한 내용의 문서들을 필터링 및 검출할 수 있는 방법을 제안한다. 문서의 내용을 비교하기 위해 키워드 추출 알고리즘을 선행하며, 이를 통해 문서의 핵심 내용을 비교할 수 있는 기반을 마련하여 표절 검토자의 작업의 정확성과 속도를 향상시키고자 한다.

키워드: 색인어 추출(keyword extraction), 표절 검사(plagiarism detection), 텍스트마이닝(textmining)

I. Introduction

표절은 타의 저작권 등 지적재산권을 침해하는 비윤리적인 행위이며, 기술 발전을 저해하는 행위이기 때문에 이러한 행위를 저지하기 위해 예로부터 꾸준히 방법론이 제기되어 왔다. 현재는 크게 두 가지 방법이 사용되고 있으며, 이는 표절 검출 알고리즘을 통한 직접적인 복제, 짜깁기, 말 바꾸어 쓰기 등을 검출하는 방법과 표절 검출자의 키워드 검색이다. 표절 알고리즘 같은 경우에는 문서의 내용보다는 문서 안의 단어 배열의 유사성 등에 집중하는 경향이 있으며, 표절 검출자가 직접 해당 문서의 키워드를 검색하여 표절을 확인하기 위해서는 문서 전체를 다시 읽어야 하는 불편함이 있기 때문에 이 두 방법론의 장점을 혼합하여 문서 내용을 비교할 수 있는 새로운 방법론이 필요하다.

현재까지 많은 연구가 진행되고 있는 문서 표절 검출 알고리즘의 방법론으로는 문법 기반 방법론(Grammar-based method), 의미론 기반 방법론 (Semantics-based method), 문법 의미론 혼합 방법론 (Grammar semantics hybrid method), 외부 표절 검출 방법론 (External plagiarism detection method), 내부 표절 검출 방법론 (Internal plagiarism detection method) 등이 있으며, 대부분의 알고리즘들은 외부 표절 검출 방법론을 활용하고 있다[1]. 본 논문에서는 문서의 주요 내용을 나타내는 주요 키워드들을 자동으로 추출하고 이에 따른 문장 레벨 전수 검색을 통해 표절 문서를 검출하는 일종의 내부 외부 혼합 표절 검출 방법론을 제안하며, 이를 통상 활용되는 외부 표절 검출 방법론 중 하나인 지문법 (Fingerprinting method) 알고리즘과 비교 분석함으로써 제안하는 기법의 효과를 증명하였다.

II. Preliminaries

1. Related works

1.1 지문법 알고리즘

지문법 알고리즘은 연속되어 있는 글자들을 이용하여 n-gram을 형성하고 각 n-gram에 대한 hashcode를 추출하여 활용하는 방법론이다. 이 hashcode 값은 각 n-gram의 지문(fingerprint)라고 칭하며, 전수 서치를 피하기 위하여 이 hashcode 셋의 일부만을 활용한다. 지문들의 일부만 활용하기 위해 통상 사용되는 방법은 0 mod p

(p는 고정값)인 hashcode값만을 추출하는 것이다. 이러한 방식으로 여러 문서에서 동일한 p값으로 지문을 추출하고, 같은 지문을 가진 문서들을 비교하는 방법론이 지문법 알고리즘이다. 지문법의 단점은 적절한 p값 선정을 하지 못할 경우에 표절을 잡아낼 수 없으며, 문서의 상당 부분이 표절인 경우에도 표절 부분에 속하는 단어들의 지문값이 0 mod p 이 아니라면 유사도 측정에도 큰 오류가 생기게 된다는 점이다[2].

1.2 단어 공기 통계 정보 기반 색인어 추출 알고리즘

색인어 추출은 문서 검색, 웹 페이지 검색, 문서 클러스터링, 텍스트 마이닝 등 여러 기술에 있어서 중요한 기술이다. 적절한 색인어를 추출함으로써 읽고자 하는 문서를 골라낼 수 있으며, 특정 문서와 다른 문서들의 관계도 알 수 있다. 색인어 추출을 위해서 통상 활용되는 tfidf 값은 한 문서에서는 높은 빈도수를 가지지만 나머지 문서들에서는 낮은 빈도수를 가지는 단어들을 색인어로 지정한다. 하지만 여러 문서에 의존하지 않고 한 문서 내에서의 색인어를 추출할 수 있는 방법은 문서 전체를 읽지 않고도 문서의 내용을 파악할 수 있는 일종의 요약 기술에 활용이 될 수 있으며, 물론 단순한 word count 방법을 활용할 수 있지만 정확한 색인어 추출을 위해서는 이보다 더 정교한 기술을 필요로 한다.

먼저, 높은 빈도수를 가진 단어들을 추출하고, 추출된 단어들과 다른 단어들의 공기 통계를 추출한다. 만약 어떤 단어가 높은 빈도수를 가진 특정 단어들과 같이 사용된다면 문서 내에서 주요한 의미를 가진 단어일 확률이 높다는 사실에 기반하여 문서 corpus에 의존하지 않는 색인어 추출을 할 수 있으며[3], 이 기법을 적절하게 활용할 경우, 문서의 핵심 내용을 비교하여 표절 검출에 도움을 줄 수 있다.

III. The Proposed Scheme

본 알고리즘은 크게 세 과정으로 나뉜다. 첫 번째는 형태소 분석기를 활용하여 문서 내 단어들과 문장들을 추출하기 위한 전처리 과정이며, 두 번째는 단어 공기 통계 정보 기반 색인어 추출 알고리즘을 통해 타겟 문서에 대한 색인어들을 추출하는 것이다. 마지막으로는 추출된 색인어들과 문서 내 문장들을 기반으로 문서 유사도를 추출한다. 문서 유사도 추출은 색인어 비교와 문서 비교로 나뉘며, 한 문서 쌍에 대해 색인어 유사도 값과 문서 유사도 값을 가지게 된다.

공기 통계 정보 기반 색인어 추출을 할 경우, 색인어는 측정된 문서 내 중요도에 따라 고유의 수치를 가지게 되며, 이를 기준으로 색인어를 정렬하였다. 문서 쌍에 대한 색인어 유사도 값 계산 시, 색인어 정렬 기준으로 한 문서의 상위 50%에 속하는 색인어가 다른 문서의 상위 50%에 속할 시 유사도 계산에 가중치를 주며, 하위 50%에 대해서도 동일한 방식으로 가중치를 부여했다. 색인어의 정렬 순서는 문서 내에 있어서의 각 색인어들의 중요도를 나타내기 때문에 두 문서가 완벽하게 같은 색인어 셋을 가지더라도 정렬 순서에 따라 문서의 내용은 달라질 수 있다는 점에 입각하였다. 즉, 이 계산 방식은 비교하고자 하는 문서 쌍의 색인어들의 문서 내 중요도를 고려하기 위해 추가되었다.

지문법 유사도 값			
	문서1	문서2	문서3
문서1		0.4557	0.6492
문서2			0.3484
문서3			
색인어 유사도 값 (본 논문 제안)			
	문서1	문서2	문서3
문서1		0.375	0.6667
문서2			0.25
문서3			

Fig. 1. Keyword similarity and fingerprint similarity

알고리즘의 정확성을 평가하기 위해 특정 분야에 대한 세 논문을 무작위로 선정하였다[4][5][6]. 선정된 세 개의 문서는 주로 슬관절에 관한 전문적인 내용으로 이루어져 있어, 비슷한 단어들을 활용하면서도 다른 내용을 가진다는 특성을 지닌다. 문서 유사도 측정 방식은 지문법과 본 알고리즘과 동일하기 때문에 본 논문에서는 생략하며, 색인어 유사도 값과 지문법을 활용한 유사도 값을 비교했다. [Fig. 1]은 비교 결과를 볼 수 있다.

문서1은 '새로이 개발된 재치환용 인공슬관절의 생체역학적 안정성 평가: 유한요소 해석', 문서2는 '고굴곡 구현을 위하여 새로이 개발된 인공슬관절의 생체역학적 안정성 평가: 보행 하중 조건', 문서3은 '슬관절 재전치환술용 경골삽입물 형상이 접촉 압력 분포에 미치는 영향' 으로 내용을 정밀히 살펴보지 않는 이상 실질적으로 표절 여부 판단이 어려운 상태이다. 이 같은 조건에서 문서 유사도를 측정하였으며, 유사도의 최대값은 동일하게 1로 설정한 후에 유사도를 비교하였다.

문서1과 문서3의 유사도는 색인어 유사도 값이 0.6667로 지문법 대비 0.0175 높게 측정되었으며, 나머지 문서 쌍들에 대해서는 지문법이 더 높은 유사도를 보였다. 문서1, 3 쌍은 실제로 유한요소 해석 또는 모델 개발에 관한 내용을 다량으로 포함하고 있어 문서 유사도가 높아 정밀검사가 필요한 문서 쌍으로 확인되었다. 문서1과 문서2의 경우, 인공슬관절의 생체역학적 안정성 평가라는 특성 상 동일하거나 비슷한 단어들을 활용하였기 때문에 지문법으로 검사했을 경우, 0.4557이라는 높은 유사도가 측정되었던 것으로 확인되었다. 하지만 색인어 추출 기법은 단순한 단어 활용도에 대한 유사도 측정이 아니기 때문에 유한요소 해석과 보행 하중 조건 평가라는 다른 내용을 파악할 수 있었기 때문에 동일 문서 쌍에 대해 0.375라는 상대적으로 낮은 유사도가 측정되었다.

IV. Conclusions

문서 유사도는 중복 연구나 표절 논문을 최종적으로 가려내는 정밀 검사를 위한 표절 후보군들을 선정하는데 있어서 활용되는 표절 검출 과정에서 중요한 필터링 인덱스이기 때문에 단순히 단어들의 배열과 빈도수에 의존할 수 없으며, 이와 같은 문제를 해결하기 위해 문서의 내용을 자동으로 판단 및 비교해 줄 수 있는 단어 공기

통계 정보 기반 색인어 추출 알고리즘을 추가적으로 개발 및 실험하였다. 색인어 추출의 정확도를 더욱 향상시키기 위해서는 사전 도입 및 고급 형태소 분석을 필요로 한다. 본 논문에서는 색인어 추출 방식이 문서 내용 비교 자동화에 얼마나 효과적인지를 판단하기 위하여 실험하였기 때문에 추가적인 사전 도입이나 형태소 분석을 행하지 않았지만, 각종 사전들을 도입할 경우에는 형태소 분석을 행할 시에 더욱 정교한 단어 추출이 가능하여 색인어 추출에 있어서 도움이 될 것으로 보인다. 또한, 고급 형태소 분석을 하여 불완전한 문장을 파악하고 품사를 정확하게 파악하게 한다면 보고서 형태에서 대제목, 소제목과 본문을 구분하는데에 도움이 될 것이며, 이 역시 색인어 추출의 정확성에 영향을 미칠 것으로 판단된다. 이러한 추가 작업을 통해 색인어 추출 정확도를 향상한다면, 기존의 문서 유사도 측정 알고리즘에 앞서 선행하여 향후 표절 검사 및 중복 연구 검사에 소비되는 인력 및 시간을 단축할 수 있을 것이다.

References

[1] A. M. E. T. Ali, H. M. D. Abdulla and V. Snasel , "Overview and comparison of plagiarism detection tools" , Proceedings of the Dateso 2011: Annual International Workshop on Databasis. Texts, Specifications and Objects

[2] Saul Schleimer , Daniel S. Wilkerson , Alex Aiken, Winnowing: local algorithms for document fingerprinting, Proceedings of the 2003 ACM SIGMOD international conference on Management of data, June 09-12, 2003, San Diego, California

[3] Matsuo, Y., Ishizuka, M. Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information. Proc. 16th Intl. Florida AI Research Society, 2003, 392--396.

[4] Y. W. Jang, J. H. Yum, J. Y. Lee, O. S. Yoo, J. S. Kim, D. H. Lim, "Evaluation of Biomechanical Stability of Total Knee Arthroplasty developed newly for Realization of High Deep Flexion : Loading Condition of Walking" , Proceedings of the conference of Korean Society Of Precision Engineering, 2014.5, 208-208 (1 pages)

[5] Paul Han, Young Woong Jang, Sumin Park, H. S. Kim, Dohyung Lim, "Evaluation of Bio-mechanical Stability of Newly Developed Revision Type Total Knee Arthroplasty: Finite Element Analysis" , Proceedings of the conference of Korean Society Of Precision Engineering, 2012.5, 1213-1214 (2 pages)

[6] Y. H. Kim, K. M. Koo, O. S. Kwon, "Effect of stem design on contact pressure distribution of end-of-stem in revision TKR" , Proceedings of the conference of Korean Society Of Precision Engineering, 2006.5, 179-180 (2 pages)