DOI: 10.3745/KIPSTB.2009.16-B.2.181

웹 검색과 문서 유사도를 활용한 2 단계 신문 기사 표절 탐지 시스템

조 정 현[†] · 정 현 기[†] · 김 유 섭[††]

요 약

최근 문서 저작권에 대한 관심과 중요도가 높아지고 있어 문서 표절에 관한 연구가 지속적으로 이루어지고 있다. 이러한 표절 문제는 신문 기사의 경우에서도 큰 관심을 끌고 있는데, 이는 상업적 가치가 큰 기사의 표절 또는 무단도용 문제가 적지 않게 발생하고 있기 때문이다. 현재까지의 문서 표절 관련 연구는 실시간 특성이 매우 강한 신문 기사의 표절 문제에 적용하기 어려웠다. 따라서 현재는 이러한 표절 기사를 가려내기 위해 수백 개의 신문사에서 하루 수천 건씩 올라오는 기사들을 눈으로 일일이 가려내는 상황이다. 본 논문에서는 이러한 시간과 비용의 문제를 줄이기 위해 네이버와 다음에서 제공하는 웹 검색 OpenAPI를 활용해 표절 가능성이 있는 기사들을 1차적으로 선별한 다음, 선별된 기사들과 원본 기사와의 문서 유사도를 측정하여 선별된 기사들의 표절 여부를 자동으로 판정할 수 있도록 하였다. 본 연구에서는 실험을 위하여 연합뉴스에서 제공되는 기사를 원본 기사로 활용하였고, 표절 가능성이 있는 기사는 네이버 및 다음의 뉴스 서비스에서 제공되는 모든 기사 중에서 선별하도록 하였다.

키워드 : 신문 기사 표절, 웹 검색, OpenAPI, 문서 유사도 측정, 표절 탐지 시스템, 연합 뉴스, 뉴스 서비스

A Two Phases Plagiarism Detection System for the Newspaper Articles by using a Web Search and a Document Similarity Estimation

Cho, Jung-Hyun[†] · Jung, Hyun-Ki[†] · Kim, Yu-Seop[††]

ABSTRACT

With the increased interest on the document copyright, many of researches related to the document plagiarism have been done up to now. The plagiarism problem of newspaper articles has attracted much interest because the plagiarism cases of the articles having much commercial values in market are currently happened very often. Many researches related to the document plagiarism have been so hard to be applied to the newspaper articles because they have strong real-time characteristics. So to detect the plagiarism of the articles, many human detectors have to read every single thousands of articles published by hundreds of newspaper companies manually. In this paper, we firstly sorted out the articles with high possibility of being copied by utilizing OpenAPI modules supported by web search companies such as Naver and Daum. Then, we measured the document similarity between selected articles and the original article and made the system decide whether the article was plagiarized or not. In experiment, we used YonHap News articles as the original articles and we also made the system select the suspicious articles from all searched articles by Naver and Daum news search services.

Keywords : Newspaper Plagiarism, Web Search, OpenAPI, Document Similarity Estimation, Plagiarism Detection System, Yonhap News, News Service

1. 서 론

최근 문서 저작권에 대한 기준이 엄격해 지면서 표절(plagiarism) 문제가 여러 분야에서 발생하고 있고 그 심각성도 날로 커지고 있다. 특히 인터넷 정보의 양이 기하급수적으로 늘어나고 개인이 운영하는 홈페이지나 블로그(blog) 등의 수가 급격히 증가함에 따라 인터넷상에서의 표절도 날로 증가하고 있다. 특히 신문의 경우에는 매일 엄청난 양의 기사가 지속적으로 웹에 올라오고 있으며, 전문가에 의하여 작성된 양질의 콘텐츠를 보유하고 있어 그 상업적 유용성이 매우 뛰어난 특징을 가지고 있기 때문에, 문서의 표절 또는 무단 도용 문제에 더욱 민감한 상황이다. 이에 본 논문에서는 매일 수만 건씩 웹을 통하여 발표되는 신문 기사의 표절을 자동으로 탐지할 수 있는 방법을 제시하고자 한다.

최근까지, 문서의 유사도 또는 표절에 관한 연구는 지속적으로 진행되고 있었는데, 주로 프로그램 소스 코드[1-3],

† 준 회 원 : 한림대학교 컴퓨터공학과 석사과정
†† 종신회원 : 한림대학교 컴퓨터공학과 부교수(교신저자)
논문접수 : 2008년 11월 18일
수 정 일 : 1차 2008년 12월 17일
심사완료 : 2008년 12월 17일

영문 문서[3-6], 또는 한글 문서[7-10] 등의 표절과 관련된 연구들이었다. 이 중에서 소스 코드 표절 검사 방법들은 그 특성상 신문 기사와 같은 텍스트 문서에 그대로 적용하기에는 매우 어렵다. 그리고 영문 문서 표절은 문장 단위의 코사인 유사도와 서열 정렬 기법을 혼합하거나[5] 헬름홀츠 머신에 기반한 의미커널을 이용하여 유사도를 측정하는[6] 방법 등을 통하여 탐지하고자 하였다. 하지만 이러한 방법들은 영어 문서들을 대상으로 하였기 때문에 한글의 특수성을 활용하지 못하였으며, 또한 표절 여부보다는 문서의 의미 유사도를 추정하고자 하는데 더 큰 목적이 있었다. 한편, 한글 문서 표절 검사 방법으로는 생물 서열 유사도 측정(BLAST: Basic Local Alignment Search Tool) 방식을 많이 활용하였는데, [7]에서는 문서 축약을 통하여, 그리고 [8]에서는 말뭉치의 확률 모델을 통하여 BLAST 알고리즘을 표절 탐지에 적용시켰다. 이밖에 벡터에 기반하거나[9], 인공신경망에 기반하여[10] 표절 여부를 판단하는 방법들이 있다. 이러한 연구들은 영어와는 다른 한국어의 특성을 반영하며 표절 여부를 판단하는데 있어 매우 정교한 알고리즘을 사용하여 정교한 표절 검사 및 부분 표절 검사 등에도 적용할 수 있는 장점을 가지고 있다. 그러나, 이러한 방법론들은 매일 수백 개의 신문사에서 각각 수백 건씩 새로이 게재되는 신문 기사의 표절 문제에 그대로 적용하기에는 적절치 않다. 왜냐하면 하나의 원본기사는 수만 건의 기사들에 대해 모두 정교한 표절 검사를 수행해야 하는데, 이러한 기사가 하루에 수백 건씩 발생하기 때문에 표절 검사의 빈도가 감당하기 어려울 정도로 증가하기 때문이다. 또한 표절 검사를 수행하기 위해서는 분 또는 초단위로 업데이트되는 모든 신문사의 기사를 텍스트 파일로 가지고 있어야 하는데 이 또한 실제로는 불가능하다.

현재는 이러한 표절 기사를 가려내기 위해서 수많은 신문사에서 하루 수천 건씩 올라오는 기사들을 눈으로 일일이 가려내고 있는 상황이다. 이러한 작업은 시간, 인력, 비용이 크게 소모될 뿐만 아니라 그 정확도 역시 담보할 수 없다.

이에 본 논문에서는 현재 인터넷으로 거의 모든 신문 기사를 볼 수 있고 또한 매일 실시간으로 기사가 업데이트 되고 있는 점을 고려하여, 효율적인 표절 기사의 탐지를 위해 네이버(Naver)[1]와 다음(Daum)[2]에서 제공하는 OpenAPI (Open Application Program Interface)를 활용하고 간단한 벡터 유사도[11]를 사용하여 두 기사간의 유사도를 측정하여 자동으로 표절 여부를 가려내는 표절 기사 탐지 시스템을 제안한다. 또한 표절 기사 탐지 정확성과 효율성을 입증하기 위해 실험에서는 실제 여러 원본의 기사들을 대상으로 검색되는 기사들의 표절 여부를 확인하여 그 성능을 측정하였다.

본 연구에서는 연합뉴스[3]에서 새로이 게재된 기사를 이후 게재된 타 신문사의 기사들이 단순히 참조를 했는지 아니면 표절을 했는지를 판단하는 시스템을 제안하였다. 이를 위하여 연합뉴스의 기사 중에서 5개의 구문을 추출하여 OpenAPI의 검색 질의로 선정하였다. 그리고 5번의 검색을 실행하여 검색된 기사들의 URL을 수집하였으며 이 URL 중에서 두 번 이상 검색된 기사들을 표절 가능성이 있는 기사로 선정하였다. 그 후에 표절 가능성이 있는 기사를 벡터 유사도를 사용해 원본 기사와의 유사도를 측정하여 최종적으로 표절 여부를 판별하였다.

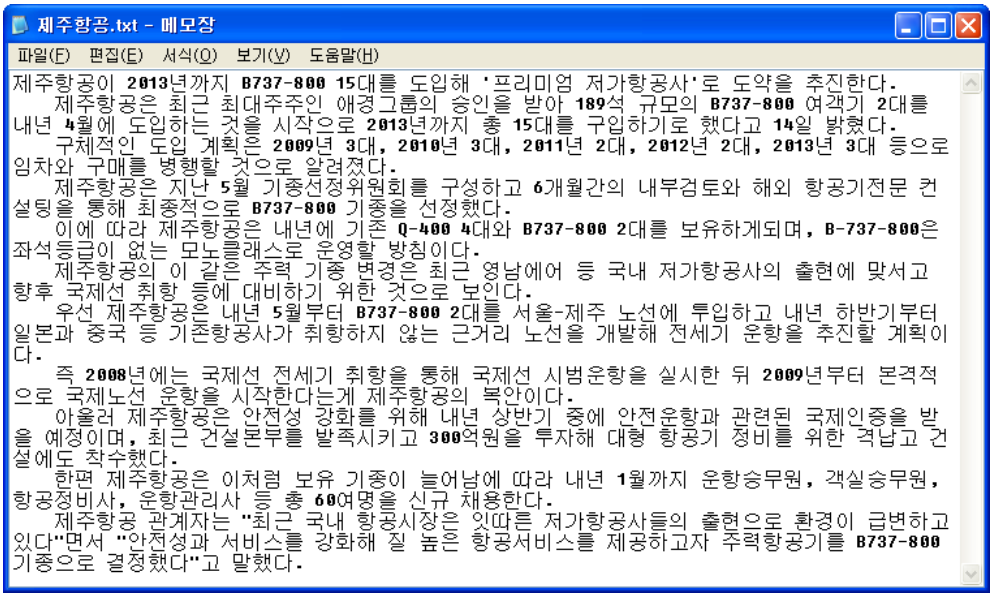
본 논문의 2장에서는 신문 기사들의 표절 유형에 대하여 분석하였고, 3장에서는 본 논문에서 제안한 표절 탐지 시스템의 전체적인 구조와 각 세부 과정 및 시스템 구현에 대하여 설명하였다. 그리고 4장에서는 실험을 통한 성능 평가 결과에 대하여 설명하고, 마지막 5장에서 결론 및 향후 연구 방향에 대하여 논하였다.

2. 기사의 표절 유형

본 논문에서는 기사의 표절 유형을 다음 4가지로 분류하였다. 첫째, 원본기사의 내용을 그대로 가져와 표절하는 유형이다. 이 유형은 출처를 알리는 크레딧 없이 원래 기사와 똑 같은 내용을 무단전재 하는 것으로, 작성된 기사가 마치 표절한 기자에 의하여 전적으로 작성된 것처럼 오인하도록 한다. 두 번째는 원본기사 중에 몇 개의 문장을 가져와 조합한 유형이다. 예를 들면 원본기사의 내용을 별도의 수정 없이 몇 개의 문장 만으로 축소하는 것이다. 보통은 마지막 문장 몇 개 또는 중간의 문장을 몇 개를 뺀다. 세 번째는 원본기사 중에 몇 개의 문장을 그대로 가져오거나 조금씩 고쳐서 사용하고 거기에 자신이 새로운 어휘나 문장을 조금씩 추가하는 유형이다. 마지막 네 번째는 원본기사의 부분 부분을 개조하고 자신이 쓴 내용과 섞어서 쓰는 유형이다. 그러나 이러한 유형은 실제 표절 여부를 판단하는데 있어 매우 전문적인 지식과 경험이 필요하기 때문에 본 연구에서는 이 유형의 표절 여부는 판단하지 않는다.

아래 (그림 1)은 원본 기사, (그림 2)는 첫 번째 표절 유형 사례, (그림 3)은 두 번째 표절 유형 사례, 그리고 (그림 4)는 세 번째 표절 유형 사례를 보여주고 있다. 먼저 (그림 2)는 첫 번째 표절 유형 사례로 (그림 1)의 원본기사와 처음부터 끝까지 완전히 같은 것을 볼 수 있다. 이는 출처를 알리는 크레딧 없이 똑 같은 내용을 무단으로 전제 한 것이다. 다음 (그림 3)은 두 번째 표절 유형 사례로 원본 기사에서 첫 번째, 두 번째, 여섯 번째 문장만을 가져와 축소하여 표절한 것이다. 마지막으로 (그림 4)는 세 번째 표절 유형 사례로 첫 번째 단락과 두 번째 단락은 원본 기사의 첫 번째 문장과 두 번째 문장에 어휘를 조금 추가하여 다르게 고치고 세 번째 단락은 원본 기사의 3, 4, 5번째 문장, 네 번째 단락은 원본 기사의 7, 8번째 문장, 마지막 다섯 번째 단락은 원본 기사의 9, 10번째 문장을 각각 조합하고 어휘를 추가하거나 조금 수정한 것이다.

1) http://openapi.naver.com
2) http://dna.daum.net/apis
3) http://www.yonhapnews.co.kr


제주항공.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

제주항공이 2013년까지 B737-800 15대를 도입해 '프리미엄 저가항공사'로 도약을 추진한다.
제주항공은 최근 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 총 15대를 구입하기로 했다고 14일 밝혔다.
구체적인 도입 계획은 2009년 3대, 2010년 3대, 2011년 2대, 2012년 2대, 2013년 3대 등으로 임차와 구매를 병행할 것으로 알려졌다.
제주항공은 지난 5월 기종선정위원회를 구성하고 6개월간의 내부검토와 해외 항공기전문 컨설팅을 통해 최종적으로 B737-800 기종을 선정했다.
이에 따라 제주항공은 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게되며, B-737-800은 좌석등급이 없는 모노클래스로 운영할 방침이다.
제주항공의 이 같은 주력 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 국제선 취항 등에 대비하기 위한 것으로 보인다.
우선 제주항공은 내년 5월부터 B737-800 2대를 서울-제주 노선에 투입하고 내년 하반기부터 일본과 중국 등 기존항공사가 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진할 계획이다.
즉 2008년에는 국제선 전세기 취항을 통해 국제선 시범운항을 실시한 뒤 2009년부터 본격적으로 국제노선 운항을 시작한다는게 제주항공의 복안이다.
아울러 제주항공은 안전성 강화를 위해 내년 상반기 중에 안전운항과 관련된 국제인증을 받을 예정이며, 최근 건설본부를 발족시키고 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 착수했다.
한편 제주항공은 이처럼 보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, 운항관리사 등 총 60여명을 신규 채용한다.
제주항공 관계자는 "최근 국내 항공시장은 잇따른 저가항공사들의 출현으로 환경이 급변하고 있다"면서 "안전성과 서비스를 강화해 질 높은 항공서비스를 제공하고자 주력항공기를 B737-800 기종으로 결정했다"고 말했다.

(그림 1) 원본 기사

제주항공, 2013년까지 B737 15대 도입

제주항공이 2013년까지 B737-800 15대를 도입해 '프리미엄 저가항공사'로 도약을 추진한다.

제주항공은 최근 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 총 15대를 구입하기로 했다고 14일 밝혔다.
구체적인 도입 계획은 2009년 3대, 2010년 3대, 2011년 2대, 2012년 2대, 2013년 3대 등으로 임차와 구매를 병행할 것으로 알려졌다.
제주항공은 지난 5월 기종선정위원회를 구성하고 6개월간의 내부검토와 해외 항공기전문 컨설팅을 통해 최종적으로 B737-800 기종을 선정했다.
이에 따라 제주항공은 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게되며, B-737-800은 좌석등급이 없는 모노클래스로 운영할 방침이다.
제주항공의 이 같은 주력 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 국제선 취항 등에 대비하기 위한 것으로 보인다.
우선 제주항공은 내년 5월부터 B737-800 2대를 서울-제주 노선에 투입하고 내년 하반기부터 일본과 중국 등 기존항공사가 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진할 계획이다.
즉 2008년에는 국제선 전세기 취항을 통해 국제선 시범운항을 실시한 뒤 2009년부터 본격적으로 국제노선 운항을 시작한다는게 제주항공의 복안이다.
아울러 제주항공은 안전성 강화를 위해 내년 상반기 중에 안전운항과 관련된 국제인증을 받을 예정이며, 최근 건설본부를 발족시키고 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 착수했다.
한편 제주항공은 이처럼 보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, 운항관리사 등 총 60여명을 신규 채용한다.
제주항공 관계자는 "최근 국내 항공시장은 잇따른 저가항공사들의 출현으로 환경이 급변하고 있다"면서 "안전성과 서비스를 강화해 질 높은 항공서비스를 제공하고자 주력항공기를 B737-800 기종으로 결정했다"고 말했다.

(그림 2) 첫 번째 표절 유형 사례

제주항공, 2013년까지 B737 15대 도입

제주항공이 2013년까지 B737-800 15대를 도입해 프리미엄 저가항공사로 도약을 추진합니다.
제주항공은 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 15대를 구입하겠다고 밝혔습니다.
제주항공의 이같은 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 국제선 취항 등에 대비하기 위한 것으로 보입니다.

(그림 3) 두 번째 표절 유형 사례

제주항공 2013년까지 B737-800 15대 도입

제주항공이 최대주주인 애경그룹의 지원을 등에 업고 2013년까지 B737-800 15대를 도입한다. 이를 통해 제주항공은 다른 저가항공사의 추격을 뿌리치고'프리미엄 저가항공사'로 자리매김한다는 방침을 정했다.

제주항공은 최근 애경그룹의 승인을 얻어 2008년 4월에 189석 규모의 B737-800 여객기 2대를 도입하는 것을 시작으로, 2013년까지 모두15대를 구입키로 결정했다고 14일 밝혔다.

구체적인 도입 계획은 2009년 3대, 2010년 3대, 2011년 2대, 2012년 2대, 2013년 3대 등으로 임차와 구매를 병행한다. 제주항공은 이에 따라 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게 된다. 앞서 제주항공은 지난 5월 기종선정위원회를 꾸려 6개월간의 내부검토와 해외 항공기전문 컨설팅을 통해 최종적으로 B737-800 기종을 선정했다.

제주항공은 내년 5월부터 B737-800 2대를 서울~제주 노선에 투입하고 하반기부터는 일본과 중국 등 기존항공사가 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진해 2009년부터는 본격적인 국제노선 운항에 들어갈 계획이다.

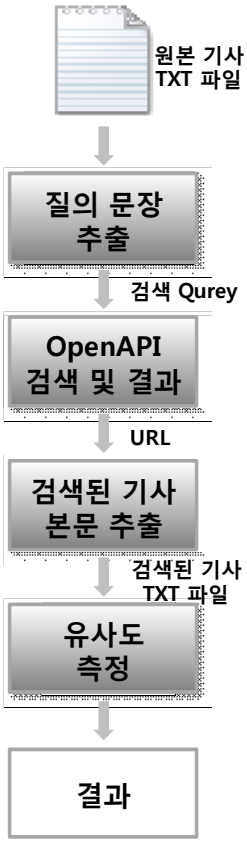
제주항공은 안전성 강화를 위해 내년 상반기 중에 안전운항과 관련된 국제인증을 받는 한편, 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 들어갔다. 제주항공 관계자는 "보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, 운항관리사 등 총 60여명을 신규 채용할 것"이라고 말했다.

(그림 4) 세 번째 표절 유형 사례

3. 기사 표절 탐지 시스템

본 시스템은 수많은 신문사에서 하루 수만 건씩 실시간으로 올라오는 신문 기사의 특성을 반영하기 위하여 웹 검색 업체에서 제공하는 OpenAPI에 기반하여 개발되었다. 시스템의 전체 구성은 아래 (그림 5) 에서 볼 수 있다.

본 시스템에서는 입력된 하나의 원본 기사에서 5개의 부분 문장을 질의어(Query)로 사용하기 위하여 효과적인 질의 문장 추출 실험을 통해 결정된 질의 문장 추출 방법을 사용하였다. 추출된 5개의 부분 문장은 포탈 업체에서 제공되는 OpenAPI의 검색어로 사용되고, OpenAPI는 뉴스 검색 방식을 사용하여 웹을 검색한다. 검색 결과로부터 뉴스 기사들의 URL이 추출되고 그 URL의 중복 횟수를 계산한다. 마지막으로 중복 검색된 URL에 해당하는 기사의 본문에서 기사 내용만을 추출한 뒤, 원본기사와의 유사도를 측정하여 표절 여부를 판단한다.



(그림 5) 시스템 전체 구성

3.1 질의 문장 추출

이 단계에서는 OpenAPI 검색을 위하여 사용될 5개의 검색어(Query)를 찾는다. 효과적인 검색을 위해서는 원본 기사에서 어떠한 방식으로 검색어를 추출하는가가 매우 중요한데, 본 연구에서는 신문 기사의 특징을 고려하여 몇 가지 방법을 미리 정해놓은 다음, 이 방법 중에서 가장 효과적인 방법을 실험을 통하여 결정하였다.

본 연구에서는 신문 기사의 특징을 고려하여 크게 세 가

지 방법을 고려대상으로 삼았는데, 그 첫 번째 방법은 원본 기사에서 무작위로 5개의 문장을 추출하는 것이다. (그림 6)은 첫 번째 방법의 사례를 보여준다. (그림 6)을 보면, 밑줄이 있는 부분이 있는데, 이 부분이 무작위로 추출되어 검색어로 사용될 부분이다. 이 방법은 표절을 탐지하는 데 있어 가장 기본적인 방법이라 할 수 있다. 그러나 일반적으로 기사의 중간 부분은 대부분 삭제 또는 편집되는 기사 표절의 특성을 반영하지 못하는 단점이 있다.

두 번째 방법으로는 원본기사에서 첫 문장부터 순서대로 5개의 문장을 각각 첫 어절부터 추출하는 방법이 있다. (그림 7)은 방법 2의 예이다. 이 방법은 기사의 표절이 기사의 모두 부분에 집중된다는 사실에 기반하였다. 그러나 모두 부분에 약간의 편집이 집중되면 표절여부를 판단하기 어려워지는 문제를 가지고 있다.

마지막 방법은 (그림 8)에서 보는 바와 같이 원본기사의 처음 문장과 끝 문장 그리고 중간에 무작위로 3개의 문장을 추출한다. 이는 기사의 모두 및 결론 부분을 표절 여부 판단에 활용하고자 하는 시도라 할 수 있다.

Article

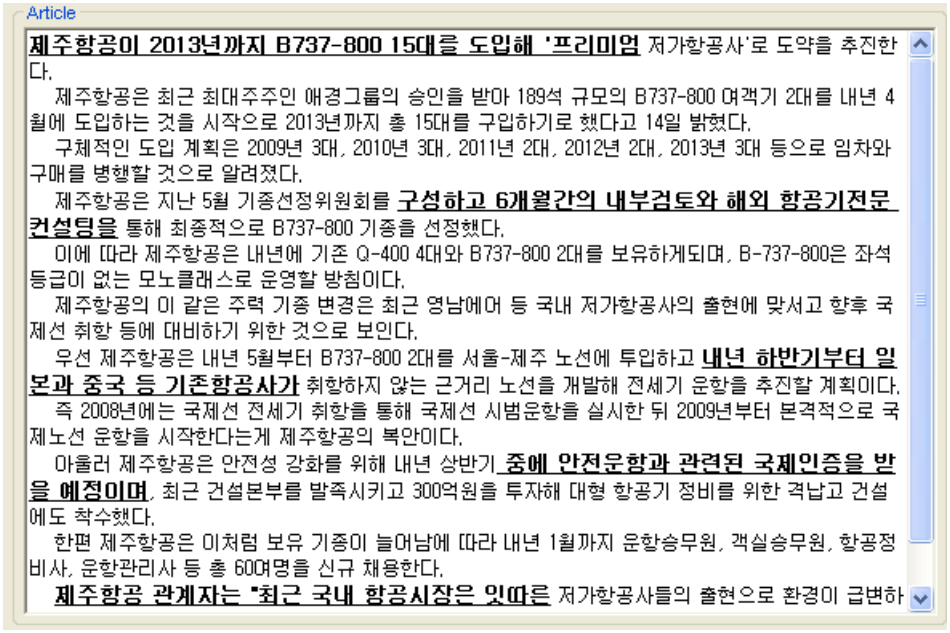
제주항공이 2013년까지 B737-800 15대를 도입해 '프리미엄 저가항공사'로 도약을 추진한다.
제주항공은 최근 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 총 15대를 구입하기로 했다고 14일 밝혔다.
구체적인 도입 계획은 2009년 3대, **2010년 3대, 2011년 2대, 2012년 2대,** 2013년 3대 **등으로 임차와 구매를 병행할 것으로 알려졌다.**
제주항공은 지난 5월 기종선정위원회를 구성하고 6개월간의 내부검토와 해외 항공기전문 컨설팅을 통해 최종적으로 B737-800 기종을 선정했다.
이에 따라 제주항공은 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게되며, B-737-800은 좌석등급이 없는 모노클래스로 운영할 방침이다.
제주항공의 이 같은 주력 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 **국제선 취항 등에 대비하기 위한 것으로** 보인다.
우선 제주항공은 내년 5월부터 B737-800 2대를 서울-제주 **노선에 투입하고 내년 하반기부터 일본과 중국** 등 기존항공사가 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진할 계획이다.
즉 2008년에는 국제선 전세기 취항을 통해 국제선 시범운항을 실시한 뒤 2009년부터 본격적으로 국제노선 운항을 시작한다는게 제주항공의 복안이다.
아울러 제주항공은 안전성 강화를 위해 내년 상반기 중에 안전운항과 관련된 국제인증을 받을 예정이며, 최근 건설본부를 발족시키고 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 착수했다.
한편 제주항공은 이처럼 보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, **운항관리사 등 총 60여명을 신규 채용한다.**
제주항공 관계자는 "최근 국내 항공시장은 잇따른 저가항공사들의 출현으로 환경이 급변하고 있다"면서 "안전성과 서비스를 강화해 질 높은 항공서비스를 제공하고자 주력항공기를 B737-800 기종으로 결정했다"고 말했다.

(그림 6) 무작위 추출 방법

Article

제주항공이 2013년까지 B737-800 15대를 도입해 '프리미엄 저가항공사'로 도약을 추진한다.
제주항공은 최근 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 총 15대를 구입하기로 했다고 14일 밝혔다.
구체적인 도입 계획은 2009년 3대, 2010년 3대, 2011년 2대, 2012년 2대, 2013년 3대 등으로 임차와 구매를 병행할 것으로 알려졌다.
제주항공은 지난 5월 기종선정위원회를 구성하고 6개월간의 내부검토와 해외 항공기전문 컨설팅을 통해 최종적으로 B737-800 기종을 선정했다.
이에 따라 제주항공은 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게되며, B-737-800은 좌석등급이 없는 모노클래스로 운영할 방침이다.
제주항공의 이 같은 주력 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 국제선 취항 등에 대비하기 위한 것으로 보인다.
우선 제주항공은 내년 5월부터 B737-800 2대를 서울-제주 노선에 투입하고 내년 하반기부터 일본과 중국 등 기존항공사가 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진할 계획이다.
즉 2008년에는 국제선 전세기 취항을 통해 국제선 시범운항을 실시한 뒤 2009년부터 본격적으로 국제노선 운항을 시작한다는게 제주항공의 복안이다.
아울러 제주항공은 안전성 강화를 위해 내년 상반기 중에 안전운항과 관련된 국제인증을 받을 예정이며, 최근 건설본부를 발족시키고 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 착수했다.
한편 제주항공은 이처럼 보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, 운항관리사 등 총 60여명을 신규 채용한다.
제주항공 관계자는 "최근 국내 항공시장은 잇따른 저가항공사들의 출현으로 환경이 급변하고 있다"면서 "안전성과 서비스를 강화해 질 높은 항공서비스를 제공하고자 주력항공기를 B737-800 기종으로 결정했다"고 말했다.

(그림 7) 모두 추출 방법


Article

제주항공이 2013년까지 B737-800 15대를 도입해 '프리미엄 저가항공사'로 도약을 추진한다.

제주항공은 최근 최대주주인 애경그룹의 승인을 받아 189석 규모의 B737-800 여객기 2대를 내년 4월에 도입하는 것을 시작으로 2013년까지 총 15대를 구입하기로 했다고 14일 밝혔다.

구체적인 도입 계획은 2009년 3대, 2010년 3대, 2011년 2대, 2012년 2대, 2013년 3대 등으로 임차와 구매를 병행할 것으로 알려졌다.

제주항공은 지난 5월 기종선정위원회를 **구성하고 6개월간의 내부검토와 해외 항공기전문 컨설팅을** 통해 최종적으로 B737-800 기종을 선정했다.

이에 따라 제주항공은 내년에 기존 Q-400 4대와 B737-800 2대를 보유하게되며, B-737-800은 좌석 등급이 없는 모노클래스로 운영할 방침이다.

제주항공의 이 같은 주력 기종 변경은 최근 영남에어 등 국내 저가항공사의 출현에 맞서고 향후 국제선 취항 등에 대비하기 위한 것으로 보인다.

우선 제주항공은 내년 5월부터 B737-800 2대를 서울-제주 노선에 투입하고 **내년 하반기부터 일본과 중국 등 기존항공사가** 취항하지 않는 근거리 노선을 개발해 전세기 운항을 추진할 계획이다.

즉 2008년에는 국제선 전세기 취항을 통해 국제선 시범운항을 실시한 뒤 2009년부터 본격적으로 국제노선 운항을 시작한다는게 제주항공의 복안이다.

아울러 제주항공은 안전성 강화를 위해 내년 상반기 **중에 안전운항과 관련된 국제인증을 받을 예정이며**, 최근 건설본부를 발족시키고 300억원을 투자해 대형 항공기 정비를 위한 격납고 건설에도 착수했다.

한편 제주항공은 이처럼 보유 기종이 늘어남에 따라 내년 1월까지 운항승무원, 객실승무원, 항공정비사, 운항관리사 등 총 60여명을 신규 채용한다.

제주항공 관계자는 "최근 국내 항공시장은 잇따른 저가항공사들의 출현으로 환경이 급변하

(그림 8) 모두-결론 추출 방법

검색어를 추출할 때에는 검색어의 위치뿐만 아니라 검색어의 길이도 고려하여야 한다. 즉 검색어를 구성하는 단어의 개수가 너무 적으면 내용이 유사한 많은 기사들이 표절 의심을 받게 되며, 또한 개수가 너무 많으면 표절 여부를 너무 엄격하게 판단하기 때문에 약간의 편집이 가미된 기사도 탐지할 수 없게 된다. 따라서 질의 문장 추출 방식 결정을 위한 실험에서는 검색 문장의 위치 및 길이가 고려대상이 되어야 한다.

실험은 총 8개의 원본 기사를 사용하였다. 먼저 세 가지 방법으로 나누어 실험을 하였는데, 검색된 모든 기사 중에 표절 또는 원본기사의 비율을 계산하여 세가지 방법 중에 가장 효과적인 방법을 결정했다. 그 후 결정된 방법에서 2개에서 10개까지 검색어의 길이를 변화시키면서 검색된 모든 기사 중에 표절 또는 원본 기사의 비율을 계산하여 효과적인 검색어의 길이를 결정하였다.

<표 1>은 세 가지 방법으로 추출된 검색어를 사용하여 검색한 결과, 실제 검색된 기사 중에서의 표절 및 원본 기사의 비율을 보여준다.

이 실험에서는 방법 2가 가장 높은 표절 또는 원본기사의 비율을 보여주었다. 또한, 방법 2는 비율뿐만 아니라 실제 검색에 성공한 표절/원본 기사의 개수 역시 가장 많았다. 이 실험의 결과 일반적으로 표절은 기사의 모두 부분에서 집중적으로 이루어지고 있음을 알 수 있었으며, 이 부분을 집중적으로 이용하는 것이 표절 탐지에 매우 유리하다 할 수 있다.

<표 2>는 방법 2의 검색어 어절 개수 별 검색 결과 및 실제 검색된 기사들의 표절 비율을 보여준다.

<표 2>에서 보면 어절의 수가 6개일 때 가장 표절/원본 기사의 비율이 높았다. 또한 검색된 표절/원본 기사의 절대 개수 역시 가장 많았다. 따라서 본 실험에서는 어절의 수가 6일 때 가장 효과적인 표절 탐지가 가능함을 보여주고 있다. 물론 어절의 개수가 10개일 때 비율이 가장 높게 나왔지만, 검색된 기사의 개수 자체가 너무 적기 때문에 6개일 때보다 효과적이라 보기 어려워 이를 제외한 것이다. 그리고 표를 보면 어절의 개수가 너무 작으면 전혀 상관없는 기사의 비율이 매우 커지고 반대로 어절의 개수가 너무 많아지면 약간 수정된 표절 기사도 검색되지 않아 너무 적은 수의 기사만이 검색되고 있음을 알 수 있다.

결과에 따라서 본 시스템에서는 원본기사에서 처음 5개의

〈표 1〉 방법에 따른 표절 기사 비율

실험 방법	총 검색기사개수	표절 or 원본기사개수	표절 or 원본기사비율
방법 1	231	68	29.44 %
방법 2	254	97	38.18 %
방법 3	344	87	25.29 %

〈표 2〉 어절 개수 별 표절 기사 비율

어절 개수	총 검색기사개수	표절 or 원본기사개수	표절 or 원본기사비율
2 개	73	6	8.21 %
3 개	45	10	22.22 %
4 개	30	13	43.33 %
5 개	29	15	51.72 %
6 개	19	15	78.95 %
7 개	19	10	52.63 %
8 개	17	12	70.58 %
9 개	12	7	58.33 %
10 개	10	9	90.00 %

문장을 대상으로 하여 첫 어절부터 연속한 6개의 어절을 추출한다.

3.2 OpenAPI 검색 및 결과

3.2.1 OpenAPI

OpenAPI는 사용자 및 개발자가 다양한 웹 서비스 및 응용을 개발할 수 있도록 기술과 서비스를 공유하는 프로그램이다. 웹 검색과 관련한 OpenAPI는 주로 구글[4], 네이버, 다음과 같은 검색 포탈 사이트에서 제공되고 있다. 본 연구에서는 AJAX 기반의 구글 API와는 달리 간단한 URL 조작에 기반한 네이버 및 다음 API를 활용하였다. 이러한 검색 API는 지식, 블로그, 웹문서, 뉴스 등 여러 종류의 문서들에 대한 검색 기능을 제공하고 있는데, 본 논문에서는 네이버와 다음의 OpenAPI 중에서 뉴스 검색 API를 사용하였다.

(그림 9)와 (그림 10)은 각각 네이버와 다음에서 제공하는 OpenAPI의 검색 Request URL을 보여준다.

Request URL의 변수를 살펴보면 네이버의 key와 다음의 apikey는 각각의 포털에서 발급받은 OpenAPI를 사용할 때 필요한 키(일종의 식별번호)를 말하고, query와 q는 추출문장이 들어갈 위치로 검색을 원하는 질의어를 말한다. 또한 start와 pageno은 검색의 시작위치를, sort는 정렬 방법을, 그리고 display와 result는 출력건수를 말한다. 마지막으로 네이버 Request URL의 target은 검색할 대상을 지정하는 것이다. 즉, 본 논문에서는 뉴스 기사를 검색하므로 news값을 사용한다. news이외에 blog, cafe등도 지정할 수 있다. 다음은 네이버와는 달리 검색이 뉴스 디렉토리 내에서 이루어지도록 정적 패스를 뉴스로 지정한다.

본 시스템에서는 검색시 Naver, Daum 중에 하나를 선택하거나 또는 둘 다 선택할 수 있도록 하였다. 둘 중 하나를 선택하면 원본 기사에서 추출한 문장이 총 5개 이므로 선택한 곳에 5번의 Request URL을 보내게 되고 둘 다 선택하게 되면 각각 5번씩 총 10번의 Request URL을 보낸다. 이렇게 Request URL을 요청하면 그 결과는 XML 문서 형태로 전달된다.

```
http://openapi.naver.com/search?key=test&query=추출문장
&target=news&start=1&sort=sim&display=100
```

(그림 9) 네이버 OpenAPI 검색 Request URL

```
http://apis.daum.net/search/news?q=추출문장&result=20
&pageno=1&sort=accu&condition=title&apikey=xxx
```

(그림 10) 다음 OpenAPI 검색 Request URL

3.2.2 결과 XML 문서

(그림 11)은 Request URL을 네이버 API를 통하여 요청하여 얻은 XML문서의 예이다.

XML 문서 각 부분을 살펴보면 다음과 같다. 먼저 <total> … </total> 부분은 주어진 검색어로 검색된 문서

```
<?xml version="1.0" encoding="UTF-8" ?>
- <rss version="2.0">
  - <channel>
      <title>Naver Open API - news ::'naver'</title>
      <link>http://search.naver.com</link>
      <description>Naver Search Result</description>
      <lastBuildDate>Wed, 16 Apr 2008 12:04:48
        +0900</lastBuildDate>
      <total>34608</total>
      <start>1</start>
      <display>100</display>
    - <item>
        <title>티베트 사태/Tibetan Activists &lt;YONHAP NO-
          0371&gt; (AP)</title>
        <originallink />
        <link>http://news.naver.com/main/read.nhn?
          mode=LSD&mid=sec&sid1=104&oid=077&aid=0001960176</link>
        <description>... Photo/Ric Francis)/2008-04-16 08:40:17/
          비비아나 런도란 이름의 한 여성이 15일 로스 앤젤레스 주재 중국
          영사관 앞에서 티베트에서의 자유와 정의를 촉구하는 티베트인들
          과 지지자들의 데모에 참여하고 있다(AP=연합뉴스).(hcs).
          (paulohan@<b>naver</b>.com). © 2007 .</description>
        <pubDate>Wed, 16 Apr 2008 12:03:00 +0900</pubDate>
      </item>
```

(그림 11) 네이버에 URL 요청 결과 XML 문서 사례

의 총 개수를 말하고, <start> … </start> 부분은 검색된 문서 중 문서의 시작점을, 그리고 <display> … </display>는 검색된 문서 중 XML 문서에서 보여주는 문서의 개수를 말하는데, (그림 11)에서는 전체 34,608건의 검색 결과 중에서 상위 1번째 문서부터 100개의 문서를 보여주고 있음을 이야기하고 있다. XML 문서에서 핵심 정보가 있는 <item> … </item> 은 검색된 문서의 제목(title), 링크(link), 본문요약(description), 날짜 정보(pubDate)가 들어 있다. 그 중에서 <originallink> … </originallink> 또는 <LINK> … </LINK> 에 있는 URL을 추출한다.

3.2.3 표절 의심 문서 추출

<originallink> … </originallink> 또는 <LINK> … </LINK> 에 있는 해당기사의 URL은 XML 문서 파싱을 통해 추출된다. 본 시스템에서는 5개의 검색어를 가지고 각각 OpenAPI를 통하여 검색하기 때문에 각 검색어 당 하나의 추출된 URL 집합이 생성된다. 이렇게 생성된 5개의 URL 집합에 포함되어 있는 모든 URL을 대상으로 각각의 URL이 몇 개의 URL 집합에 포함되어 있는지를 계산한다. 본 연구에서는 2개 이상의 집합에 포함되어 있는 URL의 문서를 표절이 의심되는 문서로 보는데, 다시 말하여 다섯 개의 질의 문장 중에서 2개 이상의 문장을 포함하는 기사를 표절 의심 기사로 보는 것이다.

(그림 12)는 Naver와 Daum 모두를 사용하여 탐지된 URL과 그 URL이 겹치는 횟수를 보여준다. 여기서 첫 번째 열 'URL'은 검색된 기사의 5개의 URL집합 중에서 2개 이상 겹치는 URL을 나타내고 두 번째 열 'Match Count'는 그 겹

Result

URL	Match Count
http://news.naver.com/main/read.nhn?mode=LSD&mid=...	5
http://newslink.media.daum.net/news/20071113101711...	5
http://www.heraldbiz.com/SITE/data/html_dir/2007/11/1...	4
http://gonews.freechal.com/common/result.asp?sFrstC...	4
http://newslink.media.daum.net/news/20071114025210...	3
http://www.seoul.co.kr/news/newsView.php?id=2007...	2

(그림 12) URL과 URL이 겹치는 횟수

4) http://code.google.co

치는 횟수를 2개에서 최대 5개까지 나타낸다. 즉 Match Count 수가 높을수록 표절 가능성이 큰 기사이고 그 수가 작을수록 표절 가능성이 낮은 기사라 할 수 있다. 5개의 집합에서 한번만 나온 URL은 우연히 겹치는 표현이 있는 것이라 판단하여 제외하였다. (그림 12)의 첫 번째 URL의 기사는 원본 기사에서 추출한 5개의 문장 모두를 포함하고 있기 때문에 표절 가능성이 매우 높다고 할 수 있으며, 세 번째 및 네 번째 기사의 경우에는 비록 원본의 5개 문장 중 하나의 문장을 포함하고 있지는 않으나 4개의 문장을 포함하고 있는데 이는 약간의 편집이 가미된 것이라 할 수 있기에 표절 가능성이 여전히 높다고 할 수 있다.

3.3 유사도 측정

본 연구에서는 표절이 의심되는 후보 기사들의 표절 여부를 자동으로 판단하기 위하여 벡터를 이용한 유사도 측정을 하였다. 즉 원본기사와 검색된 기사를 각각 두 개의 벡터로 변환시켜 두 벡터간의 유사도를 측정하게 된다.

유사도 측정을 하기 위해서는 검색된 기사 페이지에서 기사 본문을 추출해야 한다. 추출된 표절 후보 기사는 원본 기사와 함께 형태소 분석기를 통해 색인어와 그 빈도수를 추출한다. 이때, 원본 기사의 색인어를 기준으로 빈도수로 이루어진 벡터를 생성하고 두 벡터간의 유사도를 계산함으로써 표절 여부를 최종적으로 판정하게 된다.

3.3.1 검색된 기사 본문 추출

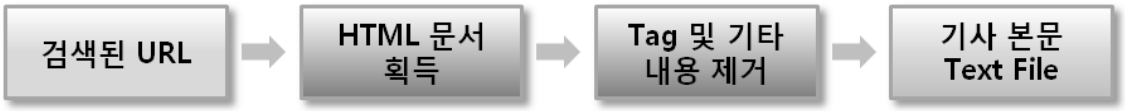
원본기사와 검색된 후보기사간의 유사도 측정을 하기 위해서는 먼저 검색된 모든 기사의 본문, 즉 기사내용을 추출해야 한다. 본문 추출은 3.2에서 OpenAPI로 검색된 기사의 URL을 이용한다. 추출방법은 URL을 요청하여 해당하는 모든 HTML 코드를 가져온 다음 기사내용을 제외한 모든 HTML 태그와 내용을 제거하는 것이다. (그림 13)은 기사 본문 추출과정을 나타낸 것이다.

(그림 14)는 HTML 코드 추출 결과 중 일부분 이고 (그림 15)는 (그림 14)에서 본문을 추출한 결과이다.

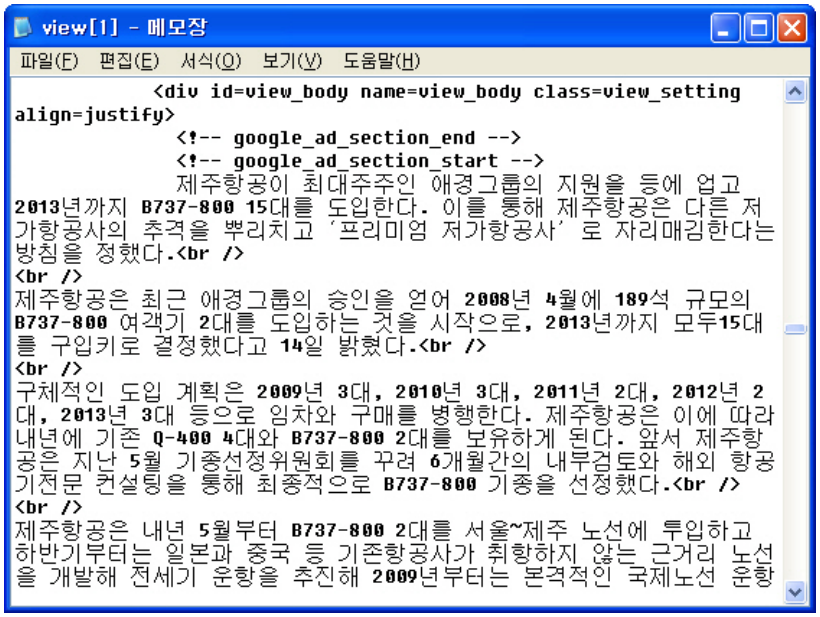
3.3.2 벡터를 이용한 유사도 측정

벡터를 이용한 유사도 측정을 하기 위해서는 먼저 원본기사와 검색된 기사의 색인어와 빈도수를 추출해야 한다. 본 연구에서는 형태소 분석기를 사용해 색인어와 빈도수를 추출하였다. 형태소 분석기는 국민대학교 한글공학-정보검색 연구실[5]의 연구용 형태소 분석기 KLT version 2.1.0를 사용했다. 본 연구에서는 형태소 분석기를 통해 나오는 여러 정보 중에서 색인어와 빈도수 정보만을 추출하여 사용하였다. <표 3>은 원본 기사에서 추출된 색인어 및 그 빈도의 일부를 보여준다.

먼저 원본 기사를 형태소 분석기를 이용하여 벡터로 변환시킨다. 물론 벡터의 원소는 추출된 색인어의 기사 내 빈도

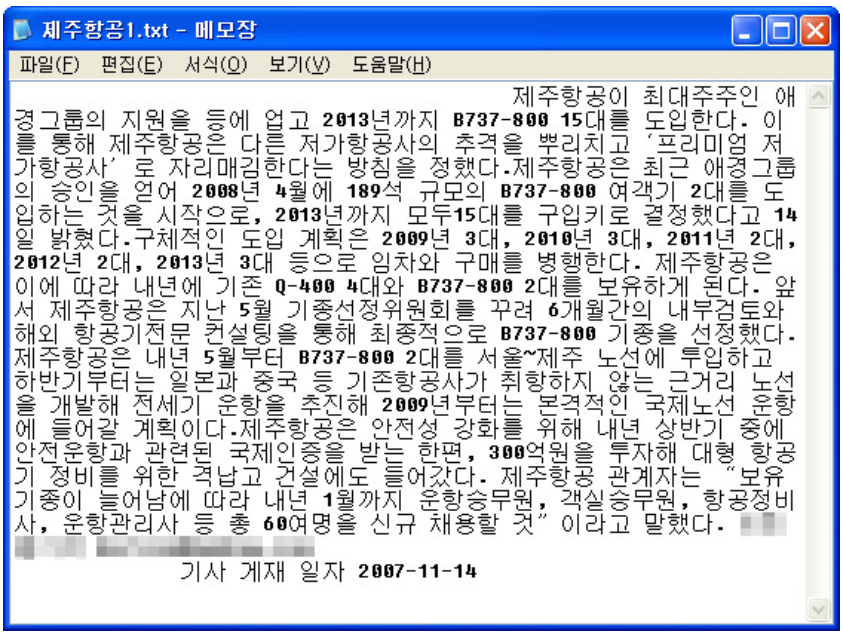


(그림 13) 기사 본문 추출과정



(그림 14) HTML 코드 추출 결과(일부)

5) http://nlp.kookmin.ac.kr



(그림 15) 본문 추출 결과

가 된다. 마찬가지의 방식으로 검색된 기사 또한 형태소 분석기를 이용하여 색인어와 해당 빈도수를 추출한다. <표 4>는 검색된 기사의 형태소 분석 결과를 보여준다. 검색된 기사의 벡터로는 원본 기사의 벡터와 동일한 벡터를 사용한다. 따라서 검색된 기사에서 추출된 색인어 중 원본기사에 존재하지 않는 색인어의 경우에는 벡터에 포함되지 않는다. 그리고 검색된 기사에서 추출된 색인어가 원본 기사 색인어 집합에 포함되어 있으면 그 빈도를 검색된 기사 벡터에 삽입하고, 그렇지 않다면 그 위치에 0을 넣는다. <표 5>는 검색된 기사를 벡터로 변환한 결과를 보여준다.

이렇게 만들어진 원본기사의 벡터와 검색기사의 벡터 사이의 유사도는 다음 수식 (1) [11]과 같이 계산한다.

$$similarity(x,y) = \frac{x \cdot y}{|x| \times |y|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i^2}} \quad (1)$$

유사도를 계산하면 0과 1사이의 값이 나오게 되는데, 1에 가까울수록 유사도가 높은 것이다. (그림 16)은 하나의 원본 기사와 검색된 여러 개의 기사 각각의 유사도 계산 결과이다.

Result

URL	Match Count	score
http://www.munhwa.com/news/v...	3	0.898464391987807
http://news.naver.com/main/read....	5	0.957910519029552
http://mbn.mk.co.kr/news/newsRe...	2	0.747003642100158
http://news.naver.com/main/read....	2	0.713956940770475
http://www.hankyung.com/news/...	3	0.681647359698574
http://news.naver.com/main/read....	2	0.546816411427981
http://newslink.media.daum.net/ne...	2	NaN
http://newslink.media.daum.net/ne...	2	0.713956940770475
http://newslink.media.daum.net/ne...	2	0.756757425264248
http://newslink.media.daum.net/ne...	2	0.916142703413501
http://newslink.media.daum.net/ne...	5	0.971904642744971
http://www.fnnews.com/view?ra...	2	0.737374045088152
http://newslink.media.daum.net/ne...	2	0.901742039463297

(그림 16) 유사도 계산 결과

〈표 3〉 원본 기사의 형태소 분석 결과

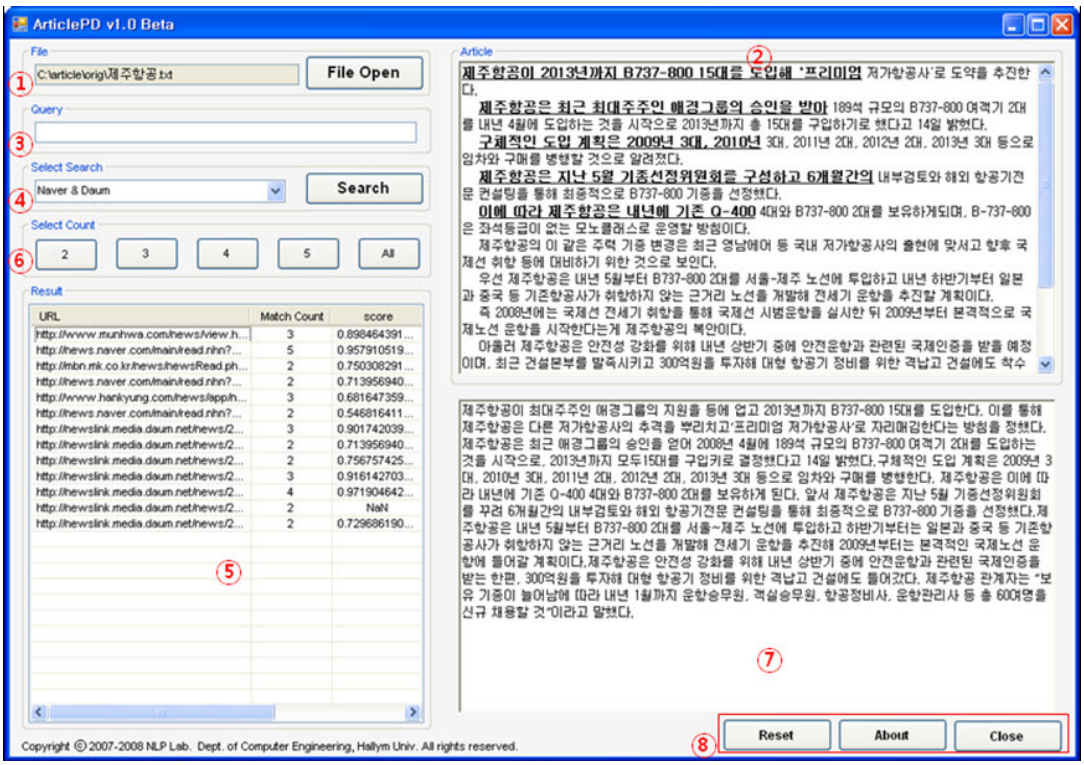
색인어	규모	그룹	근거리	급변	기	기존	기종	내년	내부	노선
빈도수	1	1	1	1	3	2	5	6	1	3

〈표 4〉 검색 기사의 형태소 분석 결과

색인어	규모	그룹	근거리	기	기존	기종	내년	내부	노선	대형
빈도수	1	2	1	2	2	3	4	1	3	1

〈표 5〉 검색 기사 벡터변환 결과

색인어	규모	그룹	근거리	급변	기	기존	기종	내년	내부	노선
빈도수	1	2	1	0	2	2	3	4	1	3



(그림 17) 시스템 UI

이 그림은 3.2 에서의 OpenAPI 검색결과에 유사도 계산 결과를 포함하여 나타낸 것이다. 세 번째 열에 유사도가 표시되는데, 중앙에 있는 'NaN'은 해당 기사의 내용이 너무 길어 형태소 분석기에서 색인어를 추출하지 못했다는 표시이다. 본 연구에서 사용한 형태소 분석기는 공개 S/W인데, 입력되는 텍스트의 길이를 제한한다. 따라서 'NaN' 표시로 나타나는 경우는 허용된 길이를 초과하는 텍스트를 말한다. 그러나 이 경우, 검색된 기사가 원본기사보다 훨씬 긴 내용을 가지게 되고, 이러한 기사는 직관적으로 표절기사일 확률이 매우 적기 때문에 문제되지 않는다고 보아 추가적인 처리는 하지 않았다.

3.4 구현

본 시스템은 C#으로 작성되었으며 사용 운영체제는 Windows XP SP2이다. 그리고 본 시스템은 지속적으로 인터넷 검색을 이용하기 때문에 인터넷이 항상 연결되어 있어야 한다.

(그림 17)은 본 시스템의 UI를 보여준다. ①은 원본 기사의 텍스트 파일을 입력하는 부분이다. 텍스트 파일의 종류는 확장자가 TXT인 파일만 가능하다. 파일을 열게 되면 ②에 원본기사의 내용을 출력한다. ③은 검색할 문장(Query) 하나를 입력하는 부분이다. 이것은 수동으로 한 문장을 검색할 때 사용한다. ④는 어떤 검색 OpenAPI를 사용할 것인지 선택하는 부분이다. 선택은 네이버, 다음 또는 모두를 선택할 수 있다. 선택한 후 Search버튼을 누르면 ②의 원본기사 내용에서 선택된 다섯 문장이 빨간색으로 나타나고 검색하여 탐지된 결과는 ⑤의 리스트박스에 보여준다. 결과 목록에는 탐지된 URL과 그 URL의 중복 횟수 그리고 유사도 점수를 보여준다. ⑥의 버튼은 결과에서 중복횟수 별로 보고자 할 때 사용한다. ⑦은 ⑤의 결과에서 URL을 더블클릭 하면 그 해당 기사의 추출된 본문을 보여준다. 마지막으로 ⑧은 모든 창을 초기화 하는 Reset 버튼, 프로그램을 닫는 Close 버튼, 프로그램 정보를 보여주는 About버튼 이다.

4. 실 험

본 시스템의 실험을 위한 원본 기사는 연합뉴스에서 발췌하여 사용하였다. 이 원본 기사는 다양한 분야에서 무작위로 선택한 것이고 개수는 총 30개 이다. 본 연구에서는 두 가지 실험을 하였는데, 첫 번째는 OpenAPI를 통해 검색한 기사 중 실제 표절 기사가 얼마나 되는지 네이버 검색, 다음 검색으로 각각 나누어 실험하였고, 두 번째는 0.01 단위로 Threshold를 변경하며 Recall과 Precision을 구하고 다시 F-Measure를 계산하여 최적의 Threshold를 정하고 이 값을 검증하는 실험을 하여 본 시스템의 성능을 측정하였다. 본 실험에서는 유사도가 threshold 이상으로 나오는 경우 표절로 판정하고자 하였다.

4.1 OepnAPI 검색을 통한 표절 가능 기사 탐지

<표 6>과 <표 7>은 네이버 검색을 선택하여 검색한 기사의 표절 기사의 개수 및 비율 이다.

표절 기사는 2장에서 설명한 세 개의 기사 표절 유형과 검색이 중복되는 횟수 별로 나누었다. <표 6>과 <표 7>에서 '동일 기사' 항목은 탐지된 기사가 연합 뉴스에서 나온 기사이거나, 기사의 출처가 연합뉴스라고 알리는 크래딧을 달아놓은 기사를 말한다. <표 6>을 보면, 30개의 원본 기사를 통해 시스템에서 탐지한 기사는 총 127개이다. 이 중 표절 기사와 동일 기사는 총 76개로 전체의 59.8%<표 7>이다. <표 7>은 중복 횟수 별 표절기사와 동일기사 모두의 비율이다. 2회 중복은 표절 기사와 동일 기사의 비율이 32.7%로 낮지만 중복 횟수가 증가할수록 표절 기사와 동일 기사의 비율이 높아지는 것을 볼 수 있다. 이는 중복 횟수가 높을수록 탐지된 기사가 표절이나 동일 기사일 가능성이 높다는 것을 알 수 있다.

<표 8>과 <표 9>는 다음 검색을 선택하여 검색한 기사의 표절 기사의 개수 및 비율이다.

탐지된 총106개의 기사 중 표절 기사와 동일 기사는 66개로 전체의 62.2%이다. 다음 검색 역시 네이버 검색 시와 비슷하게 절반 이상의 표절 기사를 탐지하는 것을 볼 수 있고 <표 9>에서와 같이 중복 횟수가 증가할수록 표절기사나 동일 기사의 비율이 높아지는 것을 볼 수 있다.

〈표 6〉 네이버 검색 결과 기사 개수

	표절유형 1	표절유형 2	표절유형 3	동일기사	표절이 아닌 기사	합계
2회		6	12		37	55
3회		3	10		12	25
4회	3	2	8	2	1	16
5회		2	4	24	1	31
합계	3	13	34	26	51	127

〈표 7〉 중복 횟수 별 표절기사+'동일기사' 비율(네이버)

	2회	3회	4회	5회	전체
표절기사 + 동일기사	32.7%	52.0%	93.8%	96.8%	59.8%

〈표 8〉 다음 검색 결과 기사 개수

	표절유형 1	표절유형 2	표절유형 3	동일기사	표절이 아닌 기사	합계
2회		5	8		30	43
3회		5	11		8	24
4회		3	3	6	2	14
5회			1	24		25
합계	0	13	23	30	40	106

〈표 9〉 중복 횟수 별 표절 기사+원본기사 비율(다음)

	2회	3회	4회	5회	전체
표절기사 + 동일기사	30.2%	66.6%	85.7%	100%	62.2%

4.2 Threshold 최적화를 통한 표절 탐지

두 번째 실험에서는 이전 실험에서 2회 이상의 URL 중복을 보여주는 검색 기사들을 대상으로 하여 이들 기사들과 원본 기사의 유사도를 계산하여 검색된 기사들의 표절 유무를 자동으로 판단하도록 하였다. 이 실험을 위하여, 총 30개의 기사를 대상으로 다음과 네이버로 검색하여 2회 이상 중복되어 검색된 기사들을 직접 사람이 확인하여 표절여부를 판별해 놓았다. 또한 0.01에서 1.00까지 유사도 Threshold 값을 0.01 단위로 변동시키며, 각각의 경우에 Recall과 Precision을 값을 다음 수식에 기반하여 구하고 그 두 값을 이용해 F-measure를 측정하였다. 수식 (2), (3), (4)는 Precision, Recall 및 F-Measure 값을 구하는 식을 Threshold 값과 연관시켜 설명한 것이다.

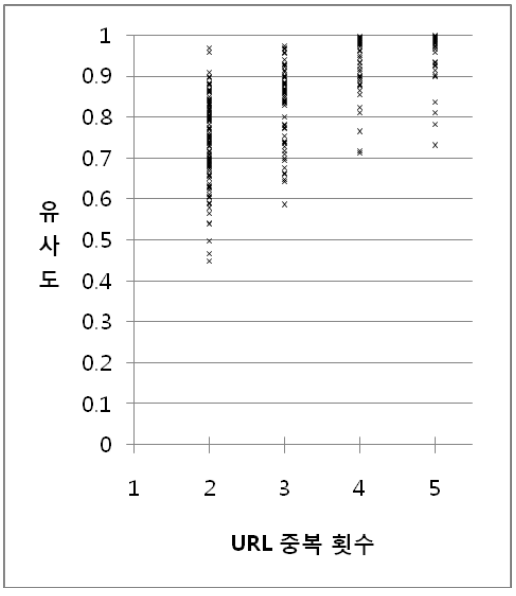
$$\text{Precision} = \frac{\text{Threshold 보다 높은 유사도를 보이는 검색된 기사 중 표절 또는 원본 기사의 수}}{\text{Theshold 보다 높은 유사도를 보이는 검색된 기사의 수}} \quad (2)$$

$$\text{Recall} = \frac{\text{Threshold 보다 높은 유사도를 보이는 검색된 기사 중 표절 또는 원본 기사의 수}}{\text{총 표절 또는 원본 기사의 수}} \quad (3)$$

$$\text{F} - \text{Measure} = \frac{2 \times \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

이렇게 구해진 각 threshold 별 f-measure 값 중에서 가장 큰 값을 찾고 그 때의 threshold 값을 기준threshold 값으로 삼아서 검색된 기사와 원본 기사와의 유사도가 기준 threshold 값보다 크면 해당 검색 기사를 표절 기사라 판명하도록 하였다.

먼저 (그림 18)은 URL의 중복 횟수 별 유사도의 분포를 보여주고 있다. 자연스러운 결과이겠지만 중복 횟수가 많을수록 검색 기사들의 유사도는 최상위에 몰려 있는 형태를 보여준다.



(그림 18) URL 중복 횟수 별 유사도 분포

그리고 (그림 19)는 위의 수식으로 구한 F-Measure 와 그에 따른 Threshold 값을 구한 결과이다. 여기서 굵은 점선은 recall을 나타내고, 얇은 점선은 precision을 그리고 마지막 실선은 f-measure 값을 나타낸다. 이 실험에서는 전체 30개의 원본 기사를 대상으로 Threshold 값을 0부터 1까지 0.01 단위로 증가시키며 모든 경우에 있어 recall, precision, 그리고 f-measure 값을 구하였다.

(그림 19)에서 보는 바와 같이 최대 F-Measure 값은 0.8685이고 그에 따른 Threshold 값은 0.71이 나왔다. 따라서 본 시스템에서는 향후 표절 탐지에서 유사도가 0.71 이상의 값을 가지는 기사들을 표절 기사라고 판정할 것이다.

이 방법론의 실 세계에서의 성능을 검증하기 위해 본 연구에서는 k-fold cross-validation 을 이용해 학습 샘플 (trained sample) 로 threshold 값을 구하고 이 값을 표절 여부의 기준으로 삼았을 때의 시험 샘플 (test sample) 의 f-measure 값을 구하여 그 성능을 평가하였다. 이 실험에서는 총 30개의 기사를 무작위로 10개씩 3개의 군으로 나누었다. 즉 K=3 이다. 먼저 하나의 군에서 최대 f-Measure를 보이는 threshold를 구하고 나머지 두 군을 합한 군에서 앞서 구한 threshold로 자동 채점을 시행하여 f-Measure 값을 구한다. 이러한 방법으로 나머지 두 번의 f-Measure를 구하고 세 값의 평균을 구한다. <표 10>은 이 실험 결과를 나타낸 것이다.

실험 결과 본 방법론을 통하여 표절을 탐지하게 된다면 약 0.84 정도의 f-measure 값을 보여줄 수 있었다. 각 실험에 따라서 recall과 precision의 분포는 약간씩 차이를 보이고 있는데, 이는 전체 기사가 30개에 불과하기 때문에 발생하는 차이라 보인다. 즉 각 기사 별로 적으면 5개 미만의 기사가 검색될 수도 있고, 많으면 30개 가까운 기사가 검색되어 표절 후보로 선정되게 되는데, 이러한 후보 기사의 차이가 성능의 차이를 유발한다. 따라서 보다 더 많은 기사를 가지고 실험한다면 이 차이는 더욱 줄어들 것이다. 그러나 f-measure 값은 세 실험의 경우 모두 비슷한 모습을 보여주고 있기 때문에, 샘플에 따른 전체적인 성능의 차이는 미미하다고 볼 수 있다.

다음 <표 11>에서 <표 14>까지는 각 중복 횟수 별 성능 평가 결과를 보여주고 있다. URL 중복 횟수가 높아지면, <표 5>에서 <표 8>까지의 데이터가 보여주듯이 원본 또는 표절 기사가 검색될 확률이 점차로 높아지고, 또한 동시에

〈표 10〉 3-fold cross-validation 을 이용한 평균 F-Measure

	1	2	3	
Threshold	0.76	0.82	0.7	Average F-Measure 0.8429
Recall	0.8766	0.7548	0.9591	
Precision	0.8282	0.9285	0.7540	
F-Measure	0.8517	0.8327	0.8443	

〈표 11〉 URL 2회 중복

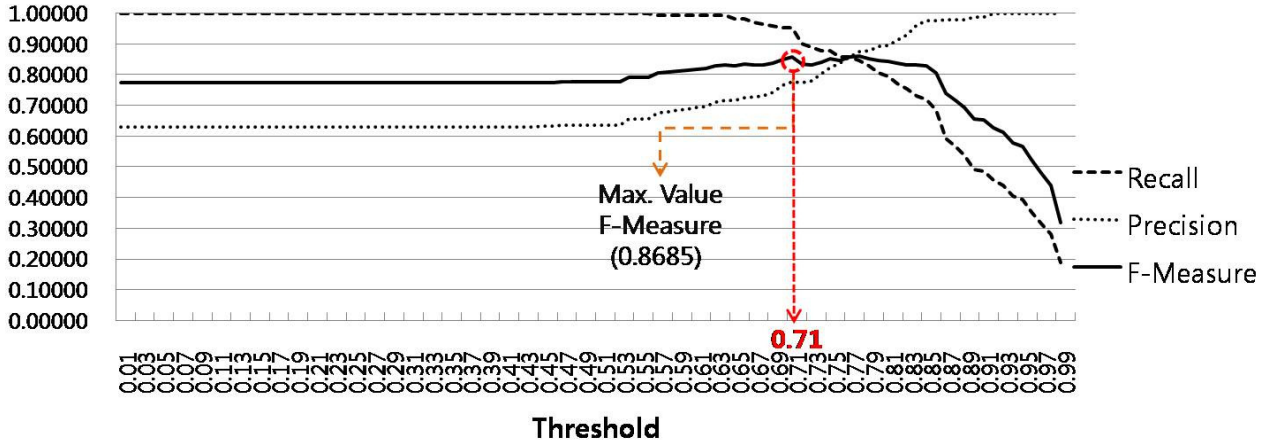
	1	2	3	
Threshold	0.82	0.75	0.7	Average F-Measure 0.6348
Recall	0.55	0.6923	0.9705	
Precision	0.8461	0.4864	0.5076	
F-Measure	0.6666	0.5714	0.6666	

〈표 12〉 URL 3회 중복

	1	2	3	
Threshold	0.01	0.84	0.01	Average F-Measure 0.8113
Recall	1	0.6071	1	
Precision	0.7895	0.9444	0.6842	
F-Measure	0.8823	0.7391	0.8125	

〈표 13〉 URL 4회 중복

	1	2	3	
Threshold	0.01	0.01	0.01	Average F-Measure 0.9863
Recall	1	1	1	
Precision	0.96	0.96	1	
F-Measure	0.9795	0.9795	1	



(그림 19) Threshold에 따른 Recall, Precision 및 F-Measure

〈표 14〉 URL 5회 중복

	1	2	3	
Threshold	0.01	0.01	0.01	Average F-Measure 1
Recall	1	1	1	
Precision	1	1	1	
F-Measure	1	1	1	

(그림 18)에서 보여지듯이 유사도 역시 1에 가까운 값들을 가지는 경향이 매우 강해진다. 따라서 URL 중복 횟수가 증가할수록 전체적인 성능은 더욱 좋은 모습을 보여주고 있는 것이다. 특히 중복 횟수가 증가하면서 매우 낮은 threshold 값으로도 더욱 좋은 성능을 보여주는데, 이 역시 위의 이유에서 비롯된 것이라 볼 수 있다.

5. 결론 및 향후 연구방향

본 논문에서는 기사의 표절을 효율적으로 탐지하는 웹 검색을 활용한 2단계 기사 표절 탐지 시스템을 제안하였다. 먼저 웹 검색 업체에서 제공하는 OpenAPI를 활용하여 표절이 의심되는 기사들을 웹에서 검색한다. 그리고 검색된 기사들과 원본 기사와의 유사도를 코사인 계산법을 활용하여 추정하여 최적화된 threshold 값 이상의 유사도를 보이는 기사들을 표절 기사로 판명하였다. 본 연구에서는 약 0.84 정도의 f-measure 값을 보여주었다.

향후 연구로는 다음과 같이 몇 가지가 고려되고 있다. 첫째, 기사 표절뿐만 아니라 블로그, 카페, 게시판 등의 문서의 표절도 탐지하고자 한다. 이러한 게시물들은 신문 기사와는 달리 문서들의 형식이 매우 다양하고 심지어는 본문이 원본 파일에서 나타나지 않는 경우도 있기 때문에 생각보다는 매우 까다롭다. 따라서 이러한 경우를 위해서는 고수준의 크롤링 기법이 개발되어야 할 것이다. 둘째, 멀티 스래딩 등을 통한 병렬처리 기법을 적용해야 한다. 본 연구에서는 5개의 질의어를 검색어로 선정하여 검색 작업을 하는데, 검색 작업이 많은 시간을 요구하기 때문에 전체적으로 많은 시간이 소모된다. 따라서 이 부분을 병렬화 시킨다면 전체 속도가 매우 크게 향상될 것이다. 마지막으로 OpenAPI를 활용하여 보다 다양한 응용에 접목시킬 것이다. 웹 2.0 시대를 맞이하여 보다 다양한 OpenAPI가 제공될 것이며, 이러한 다양한 API를 활용한다면 많은 자연언어처리 및 정보처리의 문제를 해결할 수 있을 것이다.

감사의 글

이 논문은 2006년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2006-331-D00534).

참 고 문 헌

[1] 손정우, 박성배, 이상조, 박세영, “Parse Tree Kernel을 이용한 소스코드 표절 검출”, 한국컴퓨터종합학술대회 논문집, Vol.33, No.1(B), 2006.
[2] 김영철, 황석찬, 최재영, “프로그램 유사도 평가 알고리즘”, 인터넷정보학회논문지, 제6권 제1호, pp.51-64, 2005.
[3] 김영철, “문서와 프로그래밍 언어의 표절 검사 기술에 관한 연구”, 한국경영교육학회 학술저널, 제48집, pp.25-43, 2007.
[4] Stefan Gruner, Stuart Naven, “Tool support for plagiarism detection in text documents”, Proceedings of the 2005 ACM symposium on Applied computing, DE, pp.776-781, 2005.
[5] 김지수, “OMUCS와 서열 정렬 기법을 이용한 영어 텍스트 표절 탐색 시스템의 설계 및 구현”, 중앙대학교 석사학위논문, 2005.
[6] 장정호, 김유섭, 장병탁, “헬름홀츠머신 학습 기반의 의미 커널을 이용한 문서 유사도 측정”, 한국정보과학회 학술발표 논문집, 제30권 제1호(B), pp.440-442, 2003.
[7] 전명재, “대용량 한글 문서를 위한 표절 검색 시스템 개발”, 부산대학교 석사학위논문, 2005.
[8] 류창건, 김형준, 박병준, 최혜정, 조환규, “한글 말뭉치를 이용한 한글 표절 탐색 모델 개발”, 한국정보과학회 학술발표 논문집, 제34권 제2호(A), pp.58-59, 2007.
[9] 천승환, 김미영, 이귀상, “유사 어절 트리와 비 색인어 기반의 문서 표절 유사도 분류 방법”, 컴퓨터산업교육학회 논문지, Vol.3, No.8, pp.1039-1048, 2002.
[10] 김혜숙, 박상철, 김수형, “단어/단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정”, 정보과학회논문지 : 소프트웨어 및 응용, 제31권 제12호, pp.1660-1671, 2004.
[11] Van Rijsbergen, C.J., Information Retrieval, 2nd Edition, London:Butterworths, 1979.

조 정 현
e-mail : showcjh@hallym.ac.kr
2008년 한림대학교 컴퓨터공학과(학사)
2008년~현 재 한림대학교 컴퓨터공학과 석사과정
관심분야 : 자연언어처리, 정보검색 등

정 현 기

e-mail : mayapple@hallym.ac.kr
2007년 한림대학교 컴퓨터공학과(학사)
2007년~현 재 한림대학교 컴퓨터공학과 석사과정
관심분야 : 자연언어처리, 기계번역 등

김 유 섭

e-mail : yskim01@hallym.ac.kr
1992년 서강대학교 전자계산학과(학사)
1994년 서울대학교 컴퓨터공학과(공학석사)
2000년 서울대학교 컴퓨터공학과(공학박사)
2008년~현 재 Visiting Scholar, University of Colorado at Boulder
2002년~현 재 한림대학교 컴퓨터공학과 부교수
관심분야 : 전산금융, 자연언어처리, 기계학습, e-learning 등