정보검색 기법을 이용한 부분 문서 재사용 고속 탐색

배원식O 조명래 차정원
국립창원대학교 컴퓨터공학과
wonsigi529@changwon.ac.kr, joaudfo@changwon.ac.kr, jcha@changwon.ac.kr

High Speed Local Text Reuse Detection using IR Approach

Won-Sik BaeO Myung-Rae Jo Jeong-Won Cha
Department of Computer Engineering, Changwon National University

요 약

인터넷의 발달로 지식의 재사용이 폭발적으로 증가하였다. 이는 지식의 확산이라는 측면에서는 바람직하지만 지식의 도용이라는 문제점을 안고 있다. 따라서 문서의 전부나 일부분을 재사용한 것인지를 판단하고자하는 요구가 증가하고 있다. 본 논문에서는 정보검색 기술을 이용하여 문서에서 부분 문서 재사용 및 표절을 탐색하는 방법을 제안한다. 본 논문에서는 대용량 문서의 고속 탐색을 위해서 원본 문서와 대상 문서를 색인하여 검색에 이용한다. 또한 한글의 언어적 특성을 맞게 어순 변경 비교, 기능어 생략 비교, 갭(gap) 비교 등의 다양한 처리 조건을 제공하여 문서 재사용을 탐색할 수 있다. 실험을 통해서 기존의 시스템보다 정확하게 고속으로 문서 재사용 탐색이 가능함을 보였다. 특히 비교 문서가 증가하더라도 비교 시간이 급격하게 증가하지 않으며, 정보검색 기법을 사용하는 경우 취약하다고 알려져 있는 부분 문서 재사용 탐색에도 견고하며, 처리 조건에 따라 유연하게 문서 재사용 탐색이 가능하다.

1. 서 론

문서 재사용(Text Reuse)은 문서와 문서 사이에서 서로 유사한 부분을 찾아내는 문제를 다루는 분야이다. 영어권에서는 이미 오래전부터 연구가 활발하게 진행되어 왔으나 국내에서는 아직 많은 연구가 필요한 분야라 할 수 있다. 인터넷의 활발한 보급으로 다른 사람의 글을 쉽게 접할 수 있으며, 너무나도 손쉽게 그 글을 복사하여 사용할 수 있는 환경이 마련되었다. 사람들은 그 글을 아무 거리낌 없이 마치 자신의 글인 것처럼 사용하고 있다. 이처럼 타인이 작성한 인터넷 문서를 허락 없이 사용하는 것에서부터 학생들의 리포트 표절, 최근 사회적으로 큰 문제가 되고 있는 논문 표절이나 중복 논문 게제 등의 문제가 많이 발생함에 따라 문서 재사용 연구의 필요성이 대두되고 있다.

한글은 영어와는 다른 언어적 특성을 가지고 있는데, 그 차이점은 다음과 같이 정리할 수 있다. 첫 번째로 한글은 영어에 비해 어순이 자유롭다. 영어는 어순이 정형화되어 있지만, 한글은 술어만 문장의 마지막에 있으면 나머지 문장 성분은 순서에 상관없이 자유롭게 사용할 수 있다. 예를 들어 영어로 “I go to home.”이라는 문장은 한글로는 “나는 집에 간다.”, “집에 나는 간다.” 등 여러 형태로 표현할 수 있다. 두 번째로 한글은 조사와 어미 같은 기능어가 발달해 있다. 한글은 영어와 달리 같은 표현일지라도 조사와 어미만 바꾸어 다양한 형태로 표현할 수 있다. 세 번째로 단어(Word) 구분상의 차이이다. 영어는 띄어쓰기 단위로 간단하게 단어를 구분할 수 있다. 그러나 한글은 영어와 달리 띄어쓰기 단위로 구분하면 어절만 구분되고, 영어의 단어와 같은 개념인 형태소는 어절로부터 다시 분리해내야 한다. 그리고 이것은 형태소 분석기나 품사 태거를 이용해야 분리가 가능하다. 형태소 분석을 피하기 위하여 음절단위 n-gram과 같은 언어 모델(Language Model)을 사용하는 경우가 많은데, 이 방법 역시 한글보다는 영어에 적합한 방법으로 한글의 언어적 특성을 잘 반영하지는 못한다.

본 논문에서는 한국어 품사 태거를 이용하여 형태소 구분 문제를 해결하였으며, 다양한 처리 조건을 통한 유연한 문서 재사용 탐색이 가능하도록 하였다. 그리고 정보검색에서 사용되는 방법론을 이용하여 대용량 문서를 기존연구들보다 고속으로 처리 가능하도록 하였다. 또한 인터넷에서 수집한 스포츠 신문 기사를 이용한 실험을 통하여 제안 시스템의 속도와 성능을 측정하여 제안 시스템의 가능성을 시험하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 연구들을 소개하고, 3장에서는 본 연구의 제안 시스템에 대하여 자세히 설명한다. 4장에서는 제안 시스템의 성능 측정을 위한 실험에 관한 내용을 실었으며, 5장에서는 결론을 다룬다.

2. 기존 연구

문서 재사용을 탐색하기 위한 접근법은 크게 두 종류가 존재한다. 첫 번째 방법은 문서 지문법(Document Fingerprinting)을 이용하는 방법이다. 이 방법은 문서에서 나타난 단어나 일정 구간을 해싱하여 얻어진 해시값을 비교하는 방법으로 복사 탐색(Copy Detection) 방법으로 많은 연구가 이루어진 방법이다. Shivakumar와 Garcia-Molina[1, 2]와 Broder[3]는 자신의 연구에서 효율적인 문서 지문법 방법들을 소개하고 있다. 문서의 모든 지문(Fingerprints)을 다루는 것이 가장 정확한 방법이지만 그렇게 하기에는 너무 계산 비용이 많이 들기 때문에 선택적으로 지문을 다루고자하는 다양한 알고리즘이 제안되었다[4, 5, 6, 7, 8]. 문서 지문법은 다시 오버랩(Overlap) 방법과 비오버랩(Non-overlap) 방법으로 나눌

수 있는데, 오버랩 방법에서는 Schleimer와 Wilerson, Aiken이 제안한 n-gram과 슬라이딩 윈도우(Sliding Window) 방법을 이용한 Winnowing 알고리즘[8]이 좋은 성능을 보이고 있으며, 비오버랩 방법에서는 Brin과 Davis, Garcia-Molina가 제안한 Hash-breaking 알고리즘[7]과 서장원과 Bruce Croft가 Hash-breaking 알고리즘을 수정하여 제안한 DCT Fingerprinting 알고리즘[4]이 좋은 성능을 보이고 있다. 이러한 방법들은 문서의 수와 길이에 따라 탐색 시간에 많은 영향을 받는 문서 지문법의 단점을 보완하는데 큰 몫을 하였다. 그러나 어순이 바뀌거나 단락의 순서가 바뀌는 경우에는 잘 대응하지 못하는 단점이 여전히 존재하고 있다. 두 번째 방법은 정보검색(Information Retrieval) 기법을 이용하는 방법이다. 용어 빈도(TF, Term Frequency), 문서 빈도(DF, Document Frequency) 등과 같이 문서에서 얻을 수 있는 정보들을 이용하여 문서 간의 유사도를 계산하여 문서 재사용을 탐색하는 방법이다. 이 방법은 문서에 나타난 용어를 색인하여 검색하는 방식을 취하므로 검색 시간이 문서의 길이에 크게 영향을 받지 않아 대용량 문서를 빠른 시간 내에 처리할 수 있는 특징을 갖는다. 그렇지만 기본적인 정보검색 기법은 문서 재사용에 맞게 설계되어있지 않으므로 부분 문서 재사용 탐지가 용이하지 않다. 따라서 정보검색 기법을 사용하려면 이 부분을 보완할 필요가 있다.

2.1 기존 시스템의 성능

류창건 등이 [9]에서 제안한 DEVAC 시스템은 세종 말뭉치[10]를 이용하여 수행시간을 측정하였다. DEVAC 시스템은 어절 증가량의 제곱에 비례하여 증가하는 수행시간을 불용어 제거를 통해 선형적으로 증가되도록 하였다. 그러나 n-gram과 BLAST 알고리즘을 기반으로 동작하기 때문에 기본적으로 대용량 문서를 고속으로 처리하는 방법으로는 적합하지 않다.

최성원 등은 [11]에서 주변 문장 유사도를 이용하여 문서 재사용을 측정하는 시스템을 제안하였다. [11]에서 학생들이 작성한 리포트 26개와 인터넷에서 같은 주제로 수집한 15개의 문서를 대상으로 성능을 측정하였다. [11]의 시스템과 Greedy String Tiling 알고리즘을 사용한 시스템의 성능을 다음 [표 1]과 같이 정리하였다. 단, 실험 문서가 서로 다르므로 제안 시스템과의 직접적인 성능 비교는 다소 어려울 것으로 보인다.

표 78. 기존 시스템의 성능

시스템	정확도	재현률	F1
[11] 시스템	99.7%	92.1%	95.75%
Greedy String Tiling	97.6%	90.5%	93.92%

3. 정보검색 기반의 부분 문서 재사용 검색 시스템

제안 시스템의 구조도는 [그림 1]과 같다.

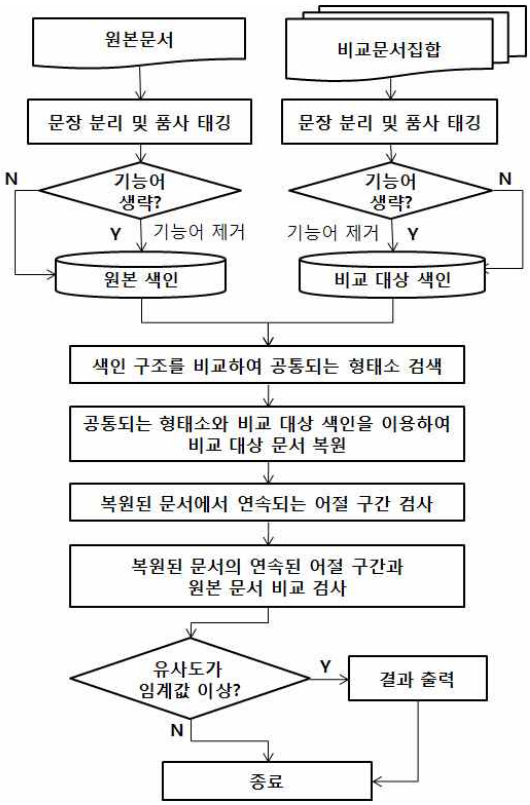


그림 85. 제안 시스템의 구조도.

3.1 전처리(Preprocessing)

전처리 과정에서는 문장 분리와 품사 태깅(Part of Speech Tagging)을 수행한다. 문서로부터 문장을 분리하고 한국어 품사 태거(1)NUGUNA Korean PoS-Tagger)를 이용하여 품사 태깅을 수행하여 형태소의 문자열과 품사 태그와 같은 어휘 정보(Lexical Information)와 위치 정보(Positional Information)를 얻는다. 제안 시스템에서는 기능어를 고려하거나 제외할 수 있는데, 기능어를 제외하는 경우에는 품사 태깅 결과에서 기능어에 해당되는 형태소 어휘 정보와 위치 정보를 배제시킨다. 기능어 제외는 영어에서 불용어(Stop-words) 제거와 같은 의미를 갖는다.

3.2 색인(Indexing)

색인 과정에서는 전처리 과정에서부터 얻은 형태소의 어휘 정보와 위치 정보를 색인한다. 제안 시스템에서는 하나의 원본 문서와 비교 대상이 되는 하나 이상의 문서를 각각 색인하여 두 개의 색인을 만든다. 두 색인을 구분하기 위해 원본 문서를 색인한 것을 '원본 색인', 비교할 문서를 색인한 것을 '비교 대상 색인'이라고 명명하기로 한다. 그리고 형태소 외에 문서의 식별자(이름), 총 어절 수, 총 형태소와 같은 문서 정보와 문장 내 총 어절 수, 총 형태소 수와 같은 문장 정보도 추가로 색인한다.

1) http://nlp.changwon.ac.kr/@page/demo/?kg_tagger

3.3 부분 문서 재사용 탐색(Local Text Reuse Detection)

부분 문서 재사용 탐색 과정에서는 먼저 색인 과정에서 만들어진 원본 색인과 비교 대상 색인을 비교하여 공통되는 형태소를 찾는다. 그리고 공통되는 형태소의 색인된 위치 정보를 이용하여 비교 대상 문서들을 복원한다. 복원된 문서의 형태소를 문서에서 나타난 형태소의 위치순으로 정렬하여 연속되는 어절 구간을 찾고, 그 연속된 구간 내의 형태소들을 원본 문서에서 나타난 문장별로 구분한다. 구분된 형태소들을 정렬하여 원본 문서와 비교 대상 문서에서 유사 구간을 찾아내는데, 유사 구간의 성립 조건은 어순 처리조건과 기능어 처리조건에 따라 달라진다. 어순을 고려하는 경우에는 비교 문서의 연속된 어절의 어순이 원본 문서와 일치하여야 유사 구간으로 검색되므로 표절과 같이 엄격하게 유사 구간을 검사할 필요가 있을 때 사용할 수 있다. 그리고 어순을 고려하지 않는 어순을 고려하지 않는 경우에는 단순히 원본 문서에서 같은 문장에서 내에서만 나타나면 유사 구간으로 간주하기 때문에 어순을 고려하는 경우보다 좀 더 유연하게 유사 구간을 찾을 수 있다는 특징이 있다. 또, 기능어를 고려하는 경우에는 어절이 완전히 일치되는 경우에만 유사 구간으로 간주하고, 기능어를 제외하는 경우에는 내용어만 같으면 기능어를 달리 쓰는 경우도 유사 구간으로 찾을 수 있다.

3.4 유사도 측정 및 탐색 결과 출력

부분 문서 재사용 탐색과정에서 찾은 원본 문서와 비교 대상 문서의 유사 구간의 정보를 이용하여 식 (1)과 같이 유사도를 측정한다.

$$sim(o,t) = \frac{|o \cap t|}{|o| + |t|} \qquad (1)$$

위 식에서 o는 원본 문서, t는 비교 대상 문서이며, |o|는 원본 문서의 크기를 말한다. 유사도가 임계값 이상이면 유사한 문서 쌍의 유사 구간을 HTML 문서 형태로 출력한다. 또한 부분 문서 재사용 탐색 결과를 보이기 위해 일정 길이 이상의 문자열이 동일하면 유사도와 관계없이 모두 출력하는 것도 가능하다.

4. 실 험

4.1 실험 문서 집합

문서 재사용 탐색 실험을 위해서는 동일한 주제를 바탕으로 기술된 문서 집합이 필요하다. 축구와 야구 관련 기사는 동일한 경기 내용과 결과를 바탕으로 작성되며, 연합뉴스의 기사를 기본으로 작성되는 경우가 많으므로 문서 재사용 탐색을 위한 실험 문서로 적합하다고 판단하였다. 따라서 본 연구에서는 4개 스포츠 신문사에 9월 1일부터 9월 11일 사이에 개제된 축구와 야구 관련 기사 1,952개를 수집하여 실험 문서로 사용하였다. 야구 관련 기사가 1,063개, 축구 관련 기사가 889개이며, 신문사별 기사의 분포는 [표 2]와 같다.

표 79. 신문사별 기사의 분포

신문사	기사 분량	
	야구 기사	축구 기사
스포츠서울	520 문서	524 문서
스포츠조선	299 문서	220 문서
스포츠동아	168 문서	107 문서
스포츠칸	76 문서	38 문서

4.2 실험 환경

문서 재사용 탐색 시스템의 수행 속도는 하드웨어 성능에 영향을 받는다. 본 연구에서는 서버급 컴퓨터와 일반 데스크탑 컴퓨터로 수행 속도를 측정하였다. 이후 서버급 컴퓨터는 '서버'라 명명하고, 일반 데스크탑 컴퓨터는 'PC'라 명명하기로 한다.

서버는 쿼드코어(Quad Core) 프로세서를 사용하며, PC는 단일코어(Single Core) 프로세세를 사용하는데, 제안 시스템은 다중 코어 환경에서 동작하도록 작성되지 않았으므로 코어의 수는 시스템의 수행속도에 영향을 주지 않는다. 하드웨어 중에서 시스템의 수행속도에 직접적인 영향을 미치는 요인이 될 수 있는 것은 CPU 자체의 계산 속도나 클럭 수, 메인 메모리와 하드디스크의 처리 속도, 운영체제 등이 있다. 서버와 PC의 하드웨어 사양은 다음 [표 3]과 같다.

표 80. 서버와 PC의 하드웨어 사양

	서버	PC
운영체제	리눅스(Linux)	Windows XP
CPU 클럭 수	2.0 GHz	2.0 GHz
메인 메모리	DDR2 2GB	DDR2 2GB
하드디스크	SCSI 레이드	S-ATA2

4.3 문서 수에 따른 색인 속도

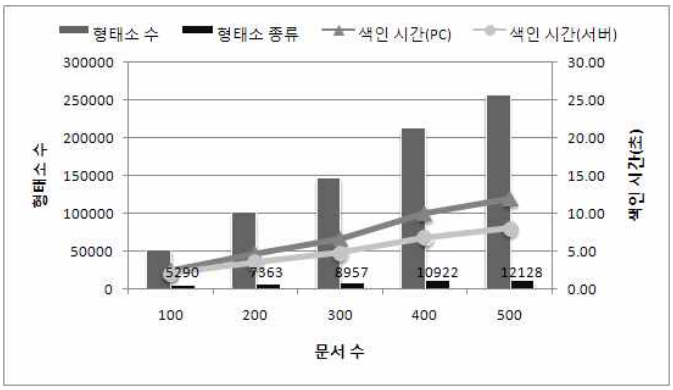


그림 86. 문서 수에 따른 색인 속도

[그림 2]에서 막대그래프는 형태소 수를 나타내고, 꺽은 선 그래프는 색인 시간을 나타낸다. 실험은 기능어를 포함한 모든 형태소를 색인하도록 진행하였다. 하드웨어

의 성능에 따라 절대적인 시간의 차이는 있으나, 문서 수가 증가함에 따라 색인 시간도 선형적으로 증가하는 것을 알 수 있다. 또한 서버가 PC보다 색인 시간이 더 빠른데, 이는 색인에 하드디스크의 접근이 많이 필요하므로, 레이드 환경이 구축된 서버가 PC보다 하드디스크의 처리 속도가 빠르기 때문에 나온 결과라 할 수 있다.

4.4 문서 수에 따른 비교 속도

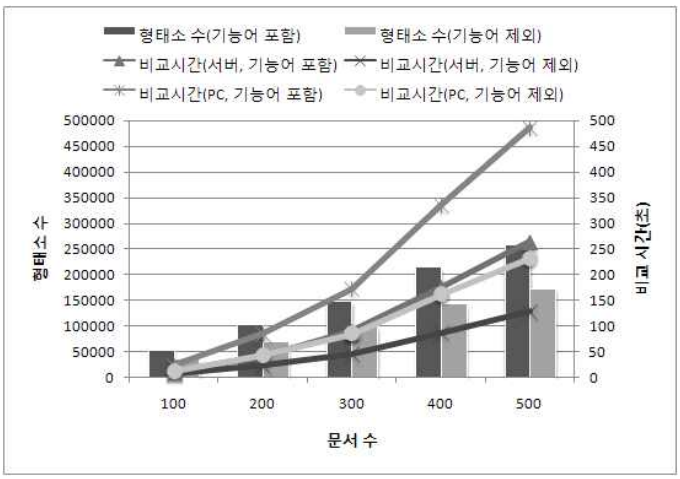


그림 87. 문서 수에 따른 비교 속도

문서 수와 기능어 처리 조건에 따른 비교 속도는 [그림 3]과 같다. 막대그래프는 형태소 수를 나타내고, 꺽은선 그래프는 비교 시간을 나타낸다. 실험은 최소 연속 어절 수를 6어절 이상, 어순을 고려하며, 기능어를 포함하여 진행하였다. 비교 시간 측정은 일정 수의 문서를 서로 교차 비교하였을 때의 시간을 측정한 것으로 문서의 총 비교횟수는 문서 수의 제곱 정도가 된다. 비교 시간도 하드웨어의 성능에 따라 절대적인 시간의 차이는 보이지만, 문서의 비교 횟수에 비례하여 선형적으로 증가한다는 것을 알 수 있다. 또한 기능어를 제외하게 되면 비교되는 형태소 수가 줄어들어 비교 시간이 40% 이상 단축된다는 것을 알 수 있다. 그리고 서버의 비교 속도가 PC보다 약 두 배 정도 빠른 것을 알 수 있다.

4.5 처리 조건에 따른 비교 속도

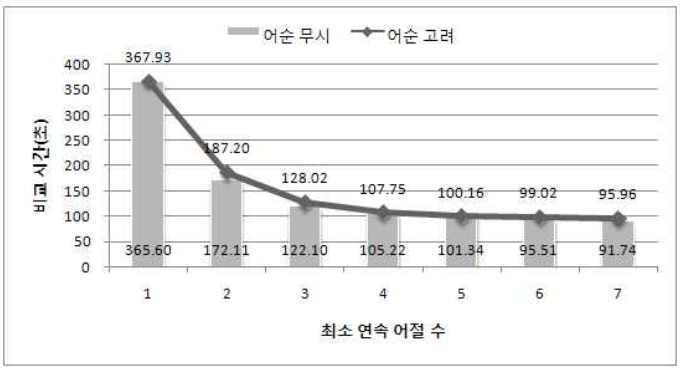


그림 88. 처리 조건에 따른 비교 속도

[그림 4]는 서버 환경에서 최소 연속 어절 수와 어순 고려/무시 처리 조건에 따른 비교 시간을 나타내는 그림이다. 최소 연속 어절 수가 낮으면 그만큼 비교하는 어절 수가 많아지므로 비교 속도가 증가함을 알 수 있다. 그리고 어순을 고려하는 경우가 어순을 무시하는 경우보다 대체적으로 비교 속도가 빠른 것을 알 수 있다. 이는 어순을 고려하면 원본 문서와 비교 대상 문서간의 어순을 검사하는 부분이 추가로 비교시간에 영향을 주는 것을 알 수 있다.

4.6 성능 평가 방법

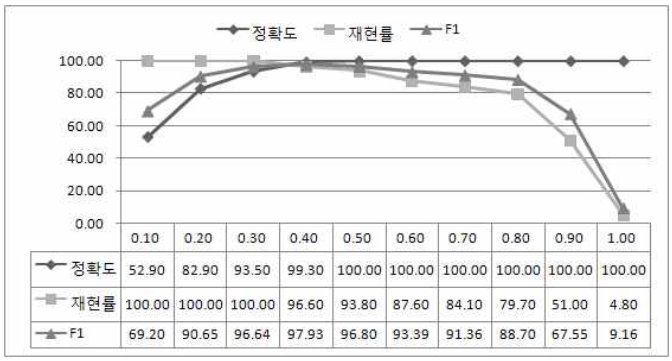
성능 평가 방법으로는 정보검색 시스템의 성능 평가에 사용되는 정확도(Precision)과 재현률(Recall)을 문서 재사용에 맞게 수정하여 사용하고, 정확도와 재현률을 하나의 값으로 표현하는 F_1 성능 평가 방법(F_1-measure)를 사용한다.

$$\text{정확도} = \frac{\text{제대로 판정된 표절 문서쌍의 수}}{\text{시스템이 판정한 표절 문서쌍의 수}}$$

$$\text{재현률} = \frac{\text{시스템이 제대로 판정한 표절 문서쌍의 수}}{\text{실제 표절 문서쌍의 수}}$$

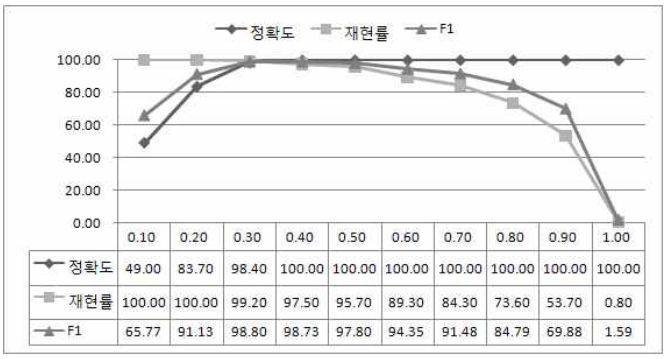
4.7 전체 문서 재사용 탐색

전체 문서 재사용 탐색의 성능 측정에 앞서 우선 축구 관련 기사와 야구 관련 기사 중에서 유사한 문서를 미리 분류하여 실제 표절 문서쌍이라고 가정하였다. 그리고 제안 시스템을 이용하여 같은 방식으로 표절 문서쌍의 유사도를 판정하였을 때, 나온 결과로 성능을 측정하였다. 실험은 최소 연속 어절 수를 6어절 이상, 어순을 고려하며, 기능어를 포함하도록 설정하여 진행하였다. [그림 5]는 야구 기사, [그림 6]은 축구 기사에 대해 유사도 임계값을 {0.1, 0.2, ..., 1.0}로 조정하여 측정한 성능을 정리한 것이다.



	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
정확도	52.90	82.90	93.50	99.30	100.00	100.00	100.00	100.00	100.00	100.00
재현률	100.00	100.00	100.00	96.60	93.80	87.60	84.10	79.70	51.00	4.80
F1	69.20	90.65	96.64	97.93	96.80	93.39	91.36	88.70	67.55	9.16

그림 89. 야구 기사에 대한 성능



	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
정확도	49.00	83.70	98.40	100.00	100.00	100.00	100.00	100.00	100.00	100.00
재현률	100.00	100.00	99.20	97.50	95.70	89.30	84.30	73.60	53.70	0.80
F1	65.77	91.13	98.80	98.73	97.80	94.35	91.48	84.79	69.88	1.59

그림 90. 축구 기사에 대한 성능

축구와 야구 문서집합 모두 임계값이 낮을수록 재현률은 높아지고, 정확도는 낮아지며, 임계값이 높을수록 그 반대 결과를 나타나는 것을 확인할 수 있다. 제안 시스템은 야구 문서 집합에서는 임계값이 0.4일 때, 97.93%의 최고 성능을 보였으며, 축구 문서 집합에서는 임계값이 0.3일 때, 98.80%의 최고 성능을 보였다. 오류를 확인해 본 결과, 오류는 실제 유사도가 낮은 문서 쌍에서 많이 발생하였다. 그러므로 문서 집합의 크기가 성능에 영향을 미치는 요인이 될 수 있다. 따라서 야구 문서 집합의 문서가 축구 문서 집합보다 약 20% 정도 많은 것이 성능에 영향을 미친 것으로 판단할 수 있다.

4.8 부분 문서 재사용 탐색

일반적으로 정보검색 기법으로 문서 재사용 탐색을 하면 부분 문서 재사용 탐색에 취약하다고 알려져 있다. 제안 시스템은 부분 문서 재사용 탐색에도 견고함을 보이는데 다음 [그림 7, 8, 9]와 같이 두 문서의 비교 결과를 통해 부분 문서 재사용 탐색에 대한 견고함을 확인해 보고자 한다.

[0] 유로2008 챔피언 스페인이 석 달 연속 세계 1위 자리를 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.	[0] 한편, 세계 1위 자리는 스페인이 석 달 연속 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.

그림 91. 어순과 기능어를 고려하였을 때의 부분 문서 탐색 결과

[0] 유로2008 챔피언 스페인이 석 달 연속 세계 1위 자리를 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.	[0] 한편, 세계 1위 자리는 스페인이 석 달 연속 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.

그림 92. 어순을 고려하지 않고 기능어를 고려하였을 때의 부분 문서 탐색 결과

[0] 유로2008 챔피언 스페인이 석 달 연속 세계 1위 자리를 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.	[0] 한편, 세계 1위 자리는 스페인이 석 달 연속 지켰고, 이탈리아가 독일을 끌어내리며 다시 2위로 올라섰다.

그림 93. 어순과 기능어를 고려하지 않았을 때의 부분 문서 탐색 결과

비교에 사용된 두 문서는 서로 내용이 같은데, 어순이 뒤바뀐 부분과 조사가 바뀌어 기술된 부분이 있다. 왼쪽 문서가 원본 문서이고 오른쪽 문서가 비교 문서이다. 실험은 제안 시스템의 처리 조건에 따라 3가지로 나누어 진행하였다. 첫 번째 실험은 어순과 기능어를 고려한 경우로 [그림 7]을 보면 문서에서 어순과 형태소가 완전히 일치되는 부분이 붉게 표시된 것을 볼 수 있다. 두 번째 실험은 어순을 고려하지 않고 기능어를 고려한 경우로 [그림 8]을 보면 첫 번째 실험에서는 잡아내지 못하던 "스페인이 석 달 연속"이라는 어순이 바뀐 구간을 잡아내는 것을 볼 수 있다. 세 번째 실험은 어순과 기능어를 모두 고려하지 않는 경우로 [그림 9]를 보면 두 번째 실험 결과에 추가로 원본 문서의 "세계 1위 자리를"과 비교 문서의 "세계 1위 자리는"이 서로 다른 조사를 사용하였음에도 불구하고 같은 구간으로 잡힌 것을 볼 수 있다. 실험 결과에서 알 수 있듯이 제안된 시스템은 정보검색 기법을 사용하면서도 부분 문서 재사용 탐색에 취약하지 않으며, 처리 조건에 따라 유연하게 유사한 구간을 찾아낼 수 있다는 것을 확인할 수 있다.

4.9 토의

본 절에서는 유사도 측정 방법에 대한 내용을 언급하고자 한다. 본 연구에서는 3.4 절에서 기술한 식 (1)과 같이 문서 간의 유사도를 계산한다. 이 유사도는 전체 문서 재사용에 대한 유사도이므로 부분 문서 재사용의 판단 지표로 사용하기에는 무리가 따른다. 왜냐하면 부분적으로 문서를 재사용하는 경우에는 유사도 값은 두 문서의 길이에 영향을 받기 때문에 부분적으로 문서를 재사용하더라도 전체 유사도 값이 낮게 계산되어 부분 문서 재사용의 판단 대상에서 제외되는 경우가 발생할 수 있기 때문이다. 따라서 본 연구에서는 유사도 외에 부분 문서 재사용의 판단 지표로 두 문서에서 연속되는 어절 수를 사용한다. 이 연속되는 어절 수에 임계값을 설정하여 임계값 이상인 경우에는 유사도가 낮더라도 부분 문서 재사용의 판단에서 제외되지 않도록 하였다. 물론 단순하게 연속되는 어절 수를 사용하여 부분 문서 재사용을 판단하는 것에는 한계가 있기 때문에 보다 나은 부분 문서 재사용의 판단 지표로 사용하기 위한 방법에 관한 연구가 필요하다.

5. 결 론

한글의 언어적 특성을 문서 재사용 탐색 시스템에 반영하기 위하여 본 연구에서는 한국어 품사 태거를 사용하였으며, 고속으로 문서 재사용 탐색을 위하여 정보검색 분야의 색인 구조를 사용하였다. 실험 결과를 통해 색인 구조로 인하여 비교 문서 수가 증가하더라도 비교 시간이 급격하게 증가하지 않았으며, 기존 시스템보다 고속으로 문서 재사용 탐색이 가능하다는 것을 확인하였다. 또한 어순과 기능어 문제와 같은 한글의 언어적 특성을 반영하기 위한 처리 조건을 통해 기존 시스템보다 유연하게 문서 재사용 탐색이 가능하다는 것을 확인할 수 있었다.

참고문헌

[1] N. Shivakumar and H. García-Molina, SCAM: A copy detection mechanism for digital documents, In Proceedings of the 2nd Annual International Conference on the Theory and Practice of Digital Libraries, 1995
[2] N. Shivakumar and H. García-Molina, Finding near-replicas of documents on the web, In International Workshop on the World Wide Web and Databases, 1999
[3] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, Syntactic clustering of the web. In Proceedings of the 6th International World Wide Web Conference, pp. 393-404, 1997
[4] Jangwon Seo, W. Bruce Croft, Local Text Reuse Detection, In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 571-578, 2008
[5] U. Manber, Finding similar files in a large file system, In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 517-523, 2005
[6] N. Heintze, Scalable document fingerprinting, In Proceedings of USENIX Workshop on Electronic Commerce, 1996
[7] S. Brin, J. Davis, and H. García-Molina, Copy detection mechanisms for digital documents, In Proceedings of the 1995 ACM SIGMOD International Conference on Management of data, pp. 398-409, 1995
[8] S. Schleimer, D. S. Wilkerson, and A. Aiken, Winnowing: local algorithms for document fingerprinting, In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 76-85, 2003
[9] 류창건, 김형준, 박병준, 최혜정, 조환규, 한글 말뭉치를 이용한 한글 표절 탐색 모델 개발, 추계 정보과학회, 2007
[10] 21세기 세종계획. http://www.sejong.or.kr/
[11] 최성원, 김상범, 임해창, 주변 문장의 유사도를 이용한 문서 재사용 측정 모델, 제17회 한글 및 한국어 정보처리 학술대회 논문집, pp.179-183, 2005