대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계

(An Efficient Preprocessing System for Searching Similar Texts among Massive Document Repository)

박 선 영 † 김 지 훈 ††
(Sun-Young Park) (Jihun Kim)
김 선 영 † 김 형 준 †
(SeonYeong Kim) (HyungJoon Kim)
조 환 규 †††
(Hwan-Gue Cho)

요 약 최근 문서 표절이 사회적 이슈가 되면서 문서간 유사도를 검사하는 시스템의 필요성이 대두되었다. 이에 따라 문서 유사도 검사 시스템에서의 중요한 요소인 검사 속도와 정확도를 충족시키기 위한 연구가 진행되고 있다. 본 논문에서는 유사 문서 탐색 시스템에서의 성능을 향상시키기 위해 전역 사전이라는 모델을 사용한 전처리 방법을 제시한다. 전역 사전이란 탐색 대상 문서군에서 사용된 모든 단어의 정보를 포함한 것으로, 유사한 문서가 어느 문서인지 빠르게 파악하는 데에 사용한다. 시스템에서 이 모델을 적용하는 방법에 대해 기술하고, 실험을 통해 각 방법의 전처리 성능을 분석하여 최적화된 문서 전처리 방법을 찾아낸다. 결과적으로 검사 대상 문서가 20,000건 이상인 경우에도 검사 대상 문서의 개수를 50개 이하로 획기적으로 줄여서 전체 시스템의 성능을 크게 향상시킬 수 있다는 것을 알 수 있었다.

키워드 : 표절, 유사 문서, 전처리, 전역 사전

Abstract Since the paper plagiarism has become one of important social issues, it is necessary to develop system for measuring the similarity between papers. The speed and accuracy of the system are very important features. So many researchers are studying the features. In this paper, we propose a preprocessing method using 'Global Dictionary' model to enhance performance of the system. The global dictionary includes information of all words in the document repository. The system uses the model to find similar papers with low computing time. Finally our experiment showed that a set of more than 20,000 documents could be reduced to about 50 documents drastically by our filtering techniques, which proves the excellence of our system.

Key words : Plagiarism, Similar Document, Global Dictionary, Preprocessing

1. 서 론

최근 들어 유명인사의 학력 위조, 학계의 논문 표절 등이 사회적 이슈로 대두됨에 따라, 표절에 대한 관심도 급증하고 있다[1]. 특히 그 중에서도 논문, 기사 등 한글 문서의 표절이 논란이 되고 있는데, 특허청에서는 표절을 "다른 사람의 저작물의 전부나 일부를 그대로, 또는 그 형태나 내용에 다소 변경을 기하여 자신의 것으로 제공 또는 제시하는 행위"라고 정의하고 있다[2]. 이러한 문제를 사회가 반영하듯, 문서 간 표절여부, 즉 유사성을 검사하는 시스템의 연구 개발이 상당히 보편화되어 있는 상황이다. 따라서 현재의 유사 문서 탐색 시스템에서는 유사 문서 판정의 정확도는 물론 다량의 문서에 대한 탐색 속도 또한 중요한 척도가 되고 있다. 일반적으로 유사 문서 탐색 속도는 대상 문서의 개수에 좌우되므로 모든 검사 대상에 대해 검사를 수행하는 것은 비효율적이다. 유사 문서 검사를 하기 전에 전처리 과정을 통하여 유사 가능성이 높은 문서만을 추출한다면 유사 문서 탐색 속도의 향상을 기대할 수 있다. 본 논문에서는 유사 문서 추출 과정을 통해 유사 문서 탐색에 소요되는 시간을 줄이고자, 전역 사전을 이용한 모델을 사용하여 대용량 문서 집합에서의 유사 문서 검사의 전처리 방법을 제시한다.

· 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음
· 이 논문은 제36회 추계학술발표회에서 '대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 부산대학교 정보컴퓨터공학부
parksy@pusan.ac.kr
s.y.kim@pusan.ac.kr
hjkim@pusan.ac.kr
†† 학생회원 : POSTECH 컴퓨터공학과
jihun735@postech.ac.kr
††† 종신회원 : 부산대학교 정보컴퓨터공학부 교수
hgcho@pusan.ac.kr
논문접수 : 2010년 1월 4일
심사완료 : 2010년 2월 18일

2. 관련 연구

현재 사용되는 문서 표절 탐색 방법에는 문장 구조를 이용한 방법과 문맥적 의미를 이용한 방법이 있는데, 이 중 문맥적 의미를 이용한 방법은 구현상의 어려움을 이유로 거의 사용되지 않는다[3]. 문서 표절 탐색 방법에는 Attribute counting 방법과 Structure metric 방법 두 가지로 분류할 수 있다. Attribute counting 방법은 문서 내의 단어 빈도수를 측정하고, 이것을 토대로 문서를 비교함으로써 유사 문서 탐색을 수행한다[4]. 따라서 문서의 길이가 탐색 속도에 큰 영향을 주지 않고, 탐색 속도가 빠른 장점이 있다. 반면 정확도가 낮고 어느 부분이 유사한지를 검출할 수 없다는 한계도 존재한다. Structure metric 방법은 문서의 구조를 분석한 후 토큰 스트링(Token String)의 유사성을 계산하여 유사 문서 탐색을 하는 방법이다. 이 방법은 문서의 길이에 비례하여 유사 문서 탐색 시간이 증가하지만, 부분적으로 비슷한 구문을 탐지하려는 경우에 유리하다. 현재 대다수의 유사 문서 탐색 시스템이 이러한 방법들을 사용하고 있으며, 이에 대한 연구도 진행되고 있다.

앞에서 언급한 두 유사 문서 탐색 방법의 장점을 결합하여 1:1 문서 간 유사도를 검사할 때 성능을 향상시킬 수 있다. Attribute Counting 방법 중 fingerprinting 방법[5]을 이용한 전처리를 통하여 비교 대상 문서의 수를 줄인 후 Structure metric 방법을 사용한 유사 문서 탐색을 수행하여 유사 문서 탐색의 정확도를 유지하면서 수행 시간을 단축할 수 있다. 이와 관련하여 한글 표절 탐색 모델 개발 연구[6,7], 문서 탐색 시스템의 전처리 과정 개선에 대한 연구[8]가 진행되고 있다. 이러한 방법들은 유사 문서 검사 전에 Attribute counting 방법을 이용한 전처리 과정을 통하여 유사 문서 탐색 대상 문서 개수를 줄여 유사 문서 탐색 속도를 향상시키고, 그 이후에 Structure metric 방법을 이용한 유사 문서 탐색을 수행하여 뛰어난 유사 문서 탐색 정확도를 보여준다. 주목할 것은 Attribute counting 방법을 적용하는 과정에서 성능을 더욱 향상시킬 수 있는 여지가 있다는 것이다. 중요한 것은 정확도(sensitivity, specificity)를 최대로 유지하면서 유사 문서 탐색에 걸리는 시간을 최소화하여 성능을 최대로 끌어올리는 것이다.

3. 대량 문서 집합 전처리 모델

3.1 대량 문서 집합 전처리 설계

전처리를 위해서 각각의 문서에 대해 개별적으로 단어 사전(DIC)이 필요하다. 사전(DIC)이란, 어떠한 문서를 3음절 단위로 나누어 이를 키(key)로 정하고, 이 키가 문서의 몇 번째 어절에 포함되는지 List 형태로 저장한 자료구조를 말한다. 이 개별 단어 사전은 두 문서 간의 단어 빈도를 조사하여 단어가 문서의 어느 위치에 존재하는지를 파악하는 데에 사용되며, 본 논문에서 활용한 시스템에서는 fingerprint 방법의 탐색을 위해 사용하는 사전을 그대로 이용하므로 별도의 계산 비용이 발생하지 않는다.

전역 사전(Global Dictionary, GDIC)은 유사 문서 탐색 대상으로 저장된 문서들의 사전을 종합하여 만들 수 있다. 입력된 전체 문서군을 하나의 문서로 보고 키를 저장하며 각 키에 대한 index를 만들어서 해당 index가 몇 번째 문서에 몇 번 사용되었는지에 대한 데이터를 모두 저장한다. 전역 사전은 대량 문서 집합 전처리 모델을 위해 적용되는 세 가지 방법 모두에 사용된다. 이 전역 사전을 생성하기 위해서는 부하가 존재하는데, 이는 개별 사전을 생성할 때 전역 사전에도 정보를 동시에 추가함으로써 최소화할 수 있다. 전체 검사 문서집합에 대한 전역 사전을 확보하면 불용어를 처리한 후, 일치 단어 수를 검사하는 방법, 일치하는 키 비율을 검사하는 방법으로 전처리를 수행하게 된다.

3.2 Specificity와 Sensitivity

유사 문서 필터링 결과에서 중요한 척도는 Specificity와 Sensitivity이다. Specificity는 시스템이 필터링한 후 문서의 총 개수 중 실제 유사 문서 수의 비율로, 높을수록 적은 개수의 후보 중 정답이 많이 포함된 상태이므로 짧은 시간 내에 효율적으로 탐색을 수행할 준비가 되었다는 뜻이다. Sensitivity는 필터링 후 결과에 포함된 유사 문서의 개수를 전체 문서 중 유사 문서의 수로 나눈 것으로, Sensitivity가 100%가 아니라면 필터링 후 정답이 누락된 것을 의미한다. 따라서 필터링을 수행할 때에는 Sensitivity를 최우선으로 고려하여야 하고, Specificity는 높을수록 시스템 성능이 향상되므로 이 또한 고려 대상이 된다.

Specificity와 Sensitivity는 다음과 같이 계산된다.

$$Specificity = \frac{\text{필터링 결과 중 유사문서의 수}}{\text{필터링 결과 후보문서의 총 개수}}$$

$$Sensitivity = \frac{\text{필터링 결과에 포함된 유사문서의 개수}}{\text{전체문서 중 유사문서의 수}}$$

전처리 이후의 시스템의 검색 정확도를 유지하기 위해, 본 논문에서 다룰 전처리 방법은 sensitivity를 100%로 유지하는 범위 내에서 specificity를 최대한 높이는 것을 목표로 한다.

3.3 대량 문서 집합 전처리 수행 방법

유사 의심 문서를 추출하는 데에는 3가지 방법을 적용하는데, 첫 번째는 불용어 처리이다. 우선 원본 문서의 DIC에 있는 키 중 GDIC에서 일정 비율 이상 사용된 키를 골라내어 불용어로 처리한다. 이 방법이 필요한 이유는, 많이 등장하는 단어가 반드시 해당 문서와 연관

되는 단어라고 할 수 없기 때문이다. 전역 사전을 통하여 조사한 결과 가장 많이 등장하는 단어는 '그', '이', '것이다' 등으로, 대부분의 문서에서 사용하는 단어이다. 이러한 단어를 걸러내기 위해 일정 비율 이상 사용된 키는 불용어로 처리해야 하며, 해당 키가 전체 문서 중 몇 개의 문서에서 사용되었는지를 나타내는 비율을 용어비율이라 정의한다.

용어 비율이 Threshold T_{ratio} 이상일 경우 해당 키를 불용어로 간주하고 사용하지 않는다. 표 1은 문서의 특정 키에 대한 용어 비율이 T_{ratio}보다 높은 키를 모두 불용어로 처리하여 제거할 경우 전처리 정확도를 나타낸 것이다.

표 1 T_{ratio}에 따른 전처리 결과와 정확도. T_{ratio}가 작을수록 전처리 후 문서 개수가 감소하나, 0.002 미만의 경우 정확도가 떨어진다.

T_{ratio}	문서의 수	정확도(%)
1.0000	20000	100.00
0.2000	719	100.00
0.1000	407	100.00
0.0100	53	100.00
0.0050	31	100.00
0.0040	24	100.00
0.0030	16	100.00
0.0020	10	100.00
0.0010	6	88.00
0.0005	1	12.00

두 번째로 불용어를 제외한 키 중 대상 문서와 GDIC에서 각각 일정 횟수 이상 동시에 등장하는 키가 존재해야만 전처리 결과에 포함시키는 방법을 사용하였다. 각각의 문서에서 특정 키가 나타나는 횟수가 N_{match} 이상이면서 동시에 두 문서를 합쳤을 때 해당 키가 나타나는 횟수가 S_{match} 이상 등장하면 대상 문서를 전처리 결과에 포함시킨다. N_{match}, S_{match}를 변화시키면서 전처리를 수행한 결과는 표 2와 같다. 유사 문서 추출 비율이 100%가 아닌 것은 전처리 과정에서 누락된 유사 문서가 존재한다는 것이다. 용어 비율이 0.001을 초과하고 동시에 유사 문서를 전처리 결과에 모두 포함시키는 문서 개수가 최소인 N_{match}/S_{match} 값은 4/12이다. 전체 실험 값 중 2/7~4/12의 범위 내에서는 양호한 전처리 결과가 나타난다.

세 번째로 대상 문서의 DIC와 GDIC 간 불용어를 제외한 키가 같은 경우를 카운트하여 두 문서 중 전체 키의 개수가 적은 쪽의 키 개수로 나눈 값이 일정 비율 이상이면 전처리 결과에 포함시키는 방법을 사용한다. 이 비율이 Common ratio C_{ratio}이상이면 해당 문서는 전처리 결과에 포함된다. 표 3은 C_{ratio}에 따른 전처리 성능을 나타낸 것이다.

표 2 N_{match}, S_{match}에 따른 전처리 후 문서의 개수. N_{match}, S_{match}가 클수록 전처리 후 문서의 개수가 감소한다. 괄호() 안은 Sensitivity를 나타낸 것으로 N_{match}, S_{match}가 너무 크면 유사 문서임에도 불구하고 전처리 결과에 포함되지 않는 경우가 발생한다.

T_{ratio} \ N_{match}/S_{match}	1/4	2/7	3/10	4/12	5/15
0.001	10.6 (100)	6.2 (90)	5.4 (90)	4.4 (90)	3.8 (68)
0.005	90.0 (100)	24.4 (100)	16.4 (100)	13.0 (100)	9.0 (88)
0.010	183.4 (100)	38.2 (100)	20.0 (100)	15.4 (100)	10.2 (100)
0.020	351.6 (100)	60.2 (100)	26.4 (100)	18.0 (100)	12.6 (100)
0.030	497.8 (100)	95.6 (100)	44.4 (100)	30.0 (100)	23.0 (100)
0.050	994.6 (100)	226.0 (100)	86.0 (100)	49.0 (100)	24.8 (100)
0.100	1584.6 (100)	286.0 (100)	92.2 (100)	49.4 (100)	24.8 (100)
0.200	2952.8 (100)	393.0 (100)	130.4 (100)	74.8 (100)	44.0 (100)

표 3 C_{ratio}에 따른 전처리 후 문서 개수. C_{ratio}가 높을수록 전처리 후 문서 개수가 감소한다.

C_{ratio} \ T_{ratio}	0.001	0.002	0.003	0.004	0.005	0.010
0.005	40.2	108.6	205.2	301.0	422.8	1095.0
0.010	13.4	28.2	49.2	64.4	90.2	240.4
0.020	7.8	12.0	21.0	22.6	26.8	46.6
0.030	6.6	10.0	14.6	15.4	16.0	23.4
0.040	6.6	8.8	12.2	13.4	14.0	17.4
0.050	6.4	6.8	11.2	11.8	12.0	14.6
0.100	5.8	6.0	7.2	8.4	9.4	10.4
0.200	5.4	5.8	6.0	6.0	6.0	6.2

결과를 분석해보면 대체로 0.030~0.100 정도의 값에서 높은 수준의 전처리 결과가 나왔음을 알 수 있다.

4. 실험

4.1 실험 데이터

유사 문서 탐색의 성능 향상을 위한 전처리 과정의 성능을 측정하기 위해, 비교 대상 문서 군으로 사용한 실험 데이터는 두 종류이며, 첫 번째는 '21세기 세종 계획'에서 제공하는 '연구/교육용 1,000만 어절 현대국어균형 말뭉치'[9]를 어절 단위의 파일로 분할하여 사용하였다.

일반적인 블로그 또는 뉴스의 길이와 비슷하게 하기

표 4 실험 데이터 - 말뭉치 자료의 장르별 구성 비

구어/문어 구분	매체 구분	상세 장르 구분
문어 (90%)	신문 (20%)	사설/칼럼 (30%)
		정치/사회/경제/외시/북한/종합 (30%)
		문화/매체/생활/과학 (30%)
		스포츠 (5%)
		기타 (5%)
	잡지 (10%)	
	책, 정보 (35%)	총류 (15%)
		교육자료 (10%)
		체험기술 (15%)
		인문 (20%)
		사회 (15%)
		자연 (10%)
		예술/취미/생활 (15%)
	책, 상상 (20%)	장편 (50%)
		중·단편 (40%)
		동화 (10%)
	기타 (5%)	
구어 (10%)	순구어 (5%)	
	준구어 (5%)	

표 5 실험 기사 200개의 범주 및 내용 구성

범주	문서 수	평균 어절 수
정치	80개	716.6
경제	21개	852.3
문화/생활	28개	612.7
IT/과학	11개	771.3
국제	8개	658.2
스포츠	11개	547.8
사회	17개	882.3
연예	24개	623.4

위하여 100줄 단위로 잘라 200~700 어절 사이의 글로 분할하여 파일로 저장하였다. 19,867개의 파일이 생성되었고, 이를 테스트 데이터로 사용하였다. 말뭉치 자료의 구성비는 표 4와 같다.

두 번째는 2009년 8월 15일부터 9월 10일까지의 신문 기사이다. 이 기사는 Naver, Daum, Google News에서 제공하는 신문사의 기사 내용을 정리하여 사용하였고, 총 200개로 구성되었다. 자료의 범주 및 개수, 어절 수는 표 5와 같다.

탐색 대상으로 사용할 문서는 앞서 소개한 실제 기사 중 5개의 문서에 대해 유사 문서에 해당하는 여러 종류의 조작을 가하여 생성하였다. 내용의 순서를 바꾼 경우, 일부의 내용만 포함한 경우, 내용 일부를 삭제한 후 필요 없는 내용을 첨가한 경우, 문서의 일부만 저장한 경우 등으로 5개 기사에 대하여 각각 5개씩의 변종을 생성하여 총 25개의 변종을 실험 데이터에 포함하였다. 시스템에서 전처리를 거친 후에도 이 데이터가 100% 포함되었을 경우 시스템의 정확도가 유지되었다고 볼 수 있다.

4.2 전처리 변수 환경 설정

전처리 과정에 있어서 중요한 변수는 T_{ratio}, N_{match}, S_{match}, C_{ratio} 등이다. 각 변수에 대해 3절에서 언급된 테스트 결과, 정확도를 유지하면서도 향상된 성능을 보여줄 수 있는 결과에 대한 변수 값의 범위를 알 수 있다. T_{ratio}의 경우 0.002~0.005, N_{match} 2~4, S_{match} 7~12, C_{ratio}의 경우 0.01~0.10이다.

4.3 실험 결과

5개의 문서에 대해 각각 유사 문서 5개씩 총 25개의 유사 문서가 포함된 20,000개의 실험 데이터에 대하여 유사 의심 문서 추출 검사를 수행했을 때, 각 변수에 대한 전처리 결과의 평균값은 표 6과 같으며, 전처리 후 문서 개수는 그림 1, 전처리 후 Specitifity 변화는 그림 2와 같다.

표 6 T_{ratio}, N_{match}, S_{match}, C_{ratio}에 따른 전처리 결과. Sensitivity는 100%를 유지하는 변수 범위에서, T_{ratio}가 작을수록, N_{match}, S_{match}가 클수록, C_{ratio} 가 클수록 Specificity가 상승한다.

C_{ratio}	T_{ratio} / N_{match}/S_{match}	0.002	0.003	0.004	0.005
0.03	2/7	12.6	23.0	25.8	27.2
	3/10	10.8	17.6	19.2	19.8
	4/12	10.4	16.8	17.4	18.0
0.05	2/7	9.6	21.8	24.2	25.2
	3/10	7.8	16.2	17.6	17.6
	4/12	7.2	15.0	15.4	15.4
0.10	2/7	8.8	21.6	23.8	25.0
	3/10	7.0	15.8	16.8	17.2
	4/12	6.4	14.0	14.0	14.6

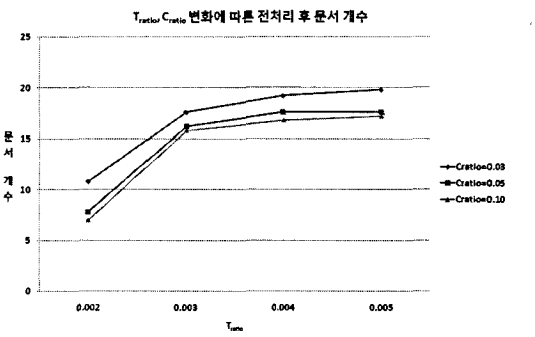

그림 1 C_{ratio}, T_{ratio} 변화에 따른 전처리 후 문서 개수. N_{match}=3, S_{match}=10으로 고정된 상태에서 실험한 결과 T_{ratio} 값이 커질 때, C_{ratio} 값이 작아질 때 전처리 후 문서 개수가 증가한다.

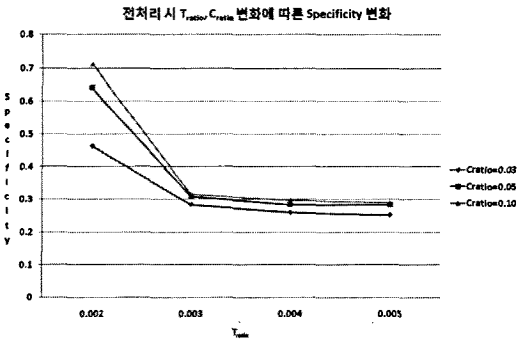

그림 2 C_{ratio}, T_{ratio} 변화에 따른 Specificity 변화. N_{match}=3, S_{match}=10로 고정된 상태에서 실험한 결과 T_{ratio} 값이 작아질 때, C_{ratio} 값이 커질 때 Specificity가 증가한다.

위의 표와 그래프에서 문서 당 찾아내야 할 유사 문서 5개는 모든 실험에서 5개로 100% 찾아내었기 때문에 시스템의 정확도는 유지하였고 볼 수 있다. 찾아낸 문서의 개수는 변수에 따라 다르나 20,000개 중 평균 15.77개 정도로, 검사 횟수를 20,000회에서 최대 30회 이하로 줄일 수 있다는 사실을 알 수 있다. 검사할 문서 집합의 종류와 특성에 따른 적절한 전처리 변수 값을 찾아낸다면 높은 정확도를 그대로 유지하면서 시스템의 성능을 크게 향상시킬 수 있다.

5. 결론 및 추후 연구

본 논문에서는 전처리 과정에서 N_{match}와 S_{match}를 이용한 문서 필터링 방법과 C_{ratio}를 이용한 방법을 통한 성능 개선을 시도하였다. 위의 두 가지 방법을 이용한 실험을 통해 두 방법의 성능을 살펴보았고, 두 방법을 동시에 적용한 모델을 제시하고 이 모델의 성능을 측정하였다.

실험 결과 20,000개의 문서 집합에 대한 전역 사전이 구축된 경우 시스템의 정확도는 유지하면서 문서 간 검사 횟수를 20,000회에서 30회 이하로 감소시킬 수 있었다. 현재 시스템에서는 전처리 필터링 기법을 위하여 DB에 구현된 GDIC를 위한 index table을 구축한다. 그러나 개별 사전(DIC)에서 GDIC를 생성하는 과정에서 발생하는 추가적인 비용을 최소화하기 위한 index table 최적화 방법, 가령 문서의 개수와 크기에 따라 생성되는 index table을 적절한 크기로 생성하는 자료구조와 알고리즘을 구현한다면 시스템 성능을 향상시키는 데 도움이 될 것이다.

추후 '21세기 세종 계획'에서 제공하는 '연구/교육용 1,000만 어절 현대국어 균형 말뭉치'를 이용하여 대용량 문서 군집에서의 문서 간 유사도 분포를 면밀히 조사하여 유사도가 Gumbel 모델을 따르는지의 여부와, 이러한 특성을 바탕으로 문서 간의 특성 유사도가 출현하는 확률에 대하여 연구할 계획이다.

참 고 문 헌

[1] H. D. Nam, "Plagiarism and Copyright Infringement," *Creation and Right Spring 2009*, vol.54, pp.32-36, Sechang, 2009. (in Korean)

[2] Korean Intellectual Property Office, http://www.kipo.go.kr/ (in Korean)

[3] S. M. Eissen, and B. Stein, "Intrinsic plagiarism detection," *Lecture Notes in Computer Science*, vol.3936, pp.565-569, Springer, 2006.

[4] J. L. Donaldson, A. Lancaster, and P. H. Sposato, A plagiarism detection system, *In Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education*, pp.21-25, 1981.

[5] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing : local algorithms for document fingerprinting," SIGMOD '03: *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pp.76-85, ACM, 2003.

[6] C. K. Ryu, H. J. Kim and H. G. Cho, "Developing of Text Plagiarism Detection Model using Korean Corpus Data," *Journal of KIISE : Computing Practices and Letters*, vol.14, no.2, pp.231-235, 2008. (in Korean)

[7] C. K. Ryu, H. J. Kim, S. H. Park and H. G. Cho, DeVAC(Document eVolution Analysis Center), http://devac.cs.pusan.ac.kr:8080/ (in Korean)

[8] H. J. Kim and H. G. Cho, "Improving Preprocessing step for Document retrieval system based on String Alignment," *Proc. of the KIISE Korea Computer Congress 2008*, vol.35, no.1(C), pp.248-251, 2008. (in Korean)

[9] 21th Century Sejong Project, http://www.sejong.or.kr/ (in Korean)