정렬을 이용한 내용기반 문서탐색 시스템의 전처리 과정 개선

김형준°, 조환규
부산대학교 컴퓨터공학과
hjkim83@pearl.cs.pusan.ac.kr, hgcho@pusan.ac.kr

Improving Preprocessing step for Document retrieval system based on String Alignment

Hyong-jun Kim°, Hwan-Gue Cho
부산대학교 컴퓨터공학과
Dept. of Computer Engineer, Pusan National University

요 약

문서 표절이 사회적으로 이슈가 됨에 따라 표절 문서를 판별할 수 있는 시스템의 필요성이 대두되었다. 문서 표절 검사 시스템에서 가장 중요한 이슈는 성능과 속도인데 이 두 가지를 모두 만족시키기 위해서는 표절을 상세하게 검사하기 전에 표절 의심 문서에 대한 비교 문서군이 크기를 최적화하여 표절 검사 범위를 최대한 작게 만들어야 한다. 비교 문서군의 크기를 최적화하기 위해서는 표절 의심 문서와 상관이 없는 문서를 필터링 하는 작업이 필요하다. 이 논문에서는 문서를 빠르게 필터링 하기 위해서 웹 문서 검색에 사용되는 Inverted Index을 이용하여 적당한 시간 안에 비교 문서군의 크기를 최적화 하는 방법들을 알아보고 각각의 방법들의 성능을 비교 분석하는 방법을 제시하며 그 방법들을 바탕으로 성능을 분석하여 최적화된 문서 필터링 방법을 알아본다.

1. 서 론

최근 한글 문서의 표절이 사회적으로 이슈가 되고 있다. 그로 인해 표절 문서를 판별하거나 표절을 사전에 차단하기 위해서 다른 유사한 문서를 빠르게 검색해주는 시스템의 필요성이 대두되고 있다. 이런 문서 표절 검사 시스템에서 중요한 이슈로는 성능과 속도가 있다. 그러나 문서 표절을 자세히 하여 성능을 향상시키기 위해서는 표절 검색 속도가 떨어지며 표절 검색 속도를 빠르게 하면 성능을 보장할 수 없게 된다. 표절 검사는 입력되는 문서 군들의 크기에 따라 처리해야할 양이 좌우되기 때문에 표절 검사를 수행하기 전에 표절 검사 문서 군들의 크기를 줄이면 표절 검색에 있어 속도와 성능을 모두 향상 시킬 수 있다. 표절 검사 문서 군들의 크기를 최적화하기 위해서는 표절 검사 문서 군에서 내용면에서 상관이 없는 문서들을 필터링하여 문서들의 개수를 줄여야 한다. 문서 내용을 이용한 빠르고 정확한 문서 필터링 기법은 우선 주어진 질의에 대하여 빠르고 비교적 정확한 결과를 돌려주어야 한다. 또한 차후에 표절 비교 문서군에 문서 추가가 용이하여야 한다. 이런 문서 필터링 기법은 이미 웹 문서 검색에서 사용되고 있다. 웹 문서 검색은 사전에 문서들을 Inverted Index를 이용하여 구조화한 다음 질의어가 받게 되면 Inverted Index를 빠르게 탐색하여 질의어가 포함되어 있거나 유사한 내용이 존재하는 웹 문서들을 돌려주게 된다. 또한 해당 문서의 Inverted Index만 만들면 문서 추가가 가능하기 때문에 차후 문서군 업데이트에도 용이하다.

이 논문에서는 Inverted Index를 이용하는 문서 필터링 방법들을 알아보고 각각의 방법들의 성능을 비교하기 위하여 성능 비교 방법을 제시한다. 그리고 그 방법들을 바탕으로 Inverted Index를 이용한 문서 필터링 방법들의 성능을 분석하여 표절 검색 문서군의 개수를 최적화 하는 문서 필터링 방법을 알아본다.

2. 관련 연구

표절 검사를 하는 방법에는 크게 대상에 따라 문서 표절 검사와 프로그램 소스 코드 표절 검사가 있다. 프로그램 소스 코드는 문서에 쓰이는 자연어에 비해 형식에 제약이 있고 의미를 나타내는 단어의 개수가 크지 않기 때문에 많은 연구가 진행되어 있다. 그에 비해 문서 표절 검사 방식은 자연언어의 복잡성으로 말미암아 아직 많은 연구가 진행되고 있다. 문서 표절 검사 방식에는 크게 문장 구조를 이용한 표절 검사 방식과 문맥적 의미를 이용한 표절 검사 방식이 존재한다. 이중에서 문맥적 의미를 이용한 표절 검사 방식은 문장의 구조나 단어의 순서에 상관없이 문장의 뜻을 이용하여 표절 검색을 하기 때문에 가장 이상적인 표절 검사 방식이기는 하나 자연어의 의미를 파악하는 작업이 어려우며 그 의미를 파악하는 시스템도 구현하기 어렵기

때문에 문맥적 의미를 이용한 표절 탐색 시스템 구현은 현재로서는 어려울 것으로 주장되어지고 있다[1].

문장 구조를 이용하여 표절 검사를 수행하는 방법은 현재 문서 표절 검사 방식에서 가장 활발히 연구되고 있는 분야이다. 표절이 일어나는 대부분의 경우가 뜻을 파악하여 표절하기 보다는 문장에서 몇 개의 단어만 수정하여 그대로 사용하는 경우가 많기 때문에 단어의 앞뒤 관계를 이용하는 문법적 탐색 기법은 상당한 표절 탐색 성능을 보이고 있다. 문장 구조를 이용한 표절 탐색 기법에는 크게 Attribute counting 방식과 Structure metric 방식이 존재한다. Attribute counting 방식은 문서에서 자주 사용되는 단어들 간의 유사성이나 빈도수를 검사하여 표절 정도를 측정하는 방법이다. 단어들 간의 유사성이나 빈도수에 초점을 맞추기 때문에 표절 탐색 시간이 문서의 길이에 크게 영향을 받지 않으며 문서의 단락 순서 또한 성능에 영향을 미치지 않는다. 이 방식을 이용하여 표절을 탐색하는 시스템으로는 CloneChecker[2], SCAM[3]이 있다. Structure metric 방식은 단어의 정확한 일치가 아닌 토큰 스트링(Token String)의 유사성을 계산하여 표절 탐색을 하는 방법이다. 표절 탐색 시간이 문서의 길이에 영향을 받으며 단락 순서 또한 표절 탐색 성능에 영향을 미치지만 지역 탐색에 유리하다. 이 방식을 현재 채용하고 있는 시스템들은 Plague[4], YAP[5], SIM[6]등이 있다. 또한 두 방식을 모두 이용하는 시스템들도 있는데 한글 표절 탐색 시스템인 DEVAC의 경우 2번의 표절 탐색을 수행하는데 처음에는 빠른 검색을 위해 Attribute Counting 방식중 하나인 'fingerprint[7]'을 이용하며 심층 검색을 위해 Structure metric 방식을 사용한다. 즉, 문서의 길이에 크게 영향을 받지 않는 Attribute Counting 방식을 이용하여 표절 검색 문서 군의 크기를 줄인 다음 Structure metric 방식으로 지역 탐색을 통해 자세한 표절 탐색을 수행한다.

3. 제안된 방법

3.1 DEVAC 예비 검사

DEVAC 예비 검사 기법은 현재 개발 중인 한글 표절 탐색 시스템인 DEVAC에서 심층 검사를 수행하기 전, 표절 탐색의 대상이 되는 문서 군을 제한하기 위해 사용되는 방법으로 Inverted Index를 이용한 fingerprint방식으로 문서들을 탐색하여 기준 문서와 대조 문서의 유사도를 계산한다. 기준 문서 Q에 대해 대조 문서 D의 점수 $Score(D,Q)$는 다음과 같다.

$$Score(D,Q)=\sum_{i=1}^{n} f(q_i,D)$$

q_i는 문서 Q에 포함되어 있는 단어이며 $f(q_i,D)$는 문서 D에서 단어 q_i가 등장하는 빈도이다. 위의 Ranking Function을 이용하여 임계값 미만의 Score를 가지는 문서 D는 심층 검사에서 제외함으로서 검색 대상을 축소하여 성능 향상을 도모한다. 하지만 현재 임계값에 대한 명확한 모델은 수립되어 있지 않다. 다음 그림은 DEVAC 예비 검사 기법의 Specifity와 Sensitivity그래프 이다.

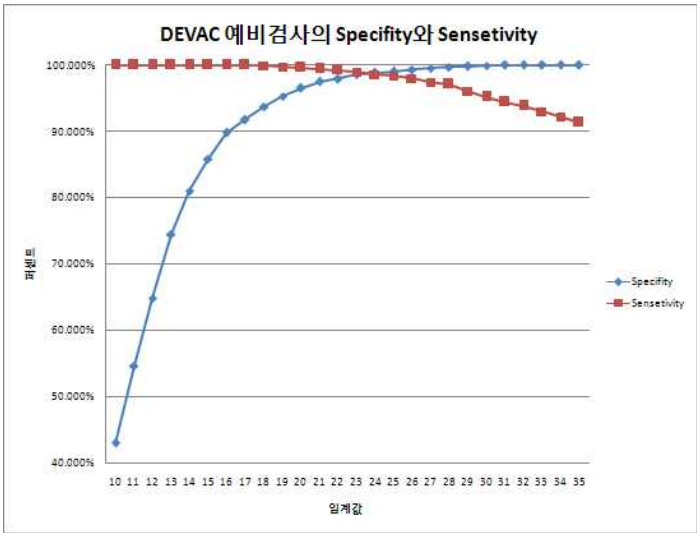

그림 1 . DEVAC 예비검사의 Specifity와 Sensetivity

위의 그래프는 DEVAC 예비검사 방법의 Score의 임계값에 따른 Specifity와 Sensitivity를 나타낸 그래프이다. Specifity는 찾은 답 중에 얼마나 정확하게 답들을 찾았는지를 나타내는 척도로 위의 그래프에서 Specifity가 증가할수록 관련 없는 문서들은 필터링 됨을 나타낸다. Sensitivity는 실제 답을 얼마나 놓치지 않았느냐를 나타내는 척도로 Sensitivity가 떨어질수록 문서 셋에서 서로 연관이 있지만 임계값보다 낮아서 필터링 되었음을 나타낸다. 위의 그래프를 이용하여 최적화된 임계값을 정할 수 있다.

3.2 Okapi BM25

Okapi BM25[8]는 검색 엔진에서 주어진 질의문에 대해 일치하는 문서들의 순위를 매기는데 사용되는 Ranking function이다. Okapi BM25는 각 문서의 단어들의 등장 빈도를 이용하여 확률적 모델을 적용하여 점수를 계산하게 된다. 단어 $q_1,\ldots,q_n$ 를 포함하고 있는 질의 문서 Q에 대한 문서 D의 Okapi BM25 점수는 다음과 같다.

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i)\frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{avgdl})}$$

여기서 $f(q_i,D)$는 문서 D에서 단어 q_i가 등장하는 빈도를 나타낸다. $|D|$는 문서 D의 단어 개수를 의미하며 $avgdl$은 비교 대상 문서군의 평균 단어 개수 이다. k_1과 b는 자유 파라미터로서 보통 $k_1=2.0$, $b=0.75$의 값을 사용한다. $IDF(q_i)$는 단어 q_i의 역 문서 빈도(inverse document frequency)로서 다음과 같이 계산된다.

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

위의 식에서 N은 비교 대상 문서군의 문서 개수이며 $n(q_i)$는 단어 q_i를 포함하는 문서의 개수이다. Okapi BM25의 Score값을 계산하여 임계값 이하의 문서를 필터링 하여 심층 검사에서 제외함으로서 검색 대상의 축소를 통한 성능 향사의 효과를 얻을 수 있다. Okapi BM25의 Specifity와

Sensetivity그래프는 다음과 같다.

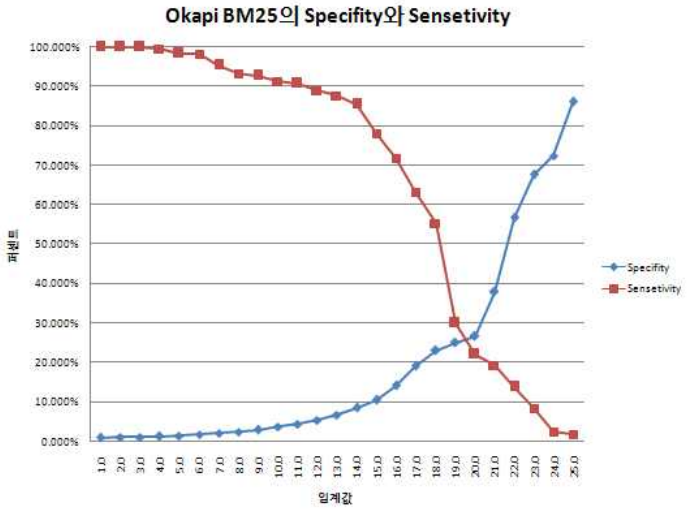

그림 2. Okapi BM25의 Specifity와 Sensetivity

위의 그래프는 Okapi BM25의 Score의 임계값에 따른 Specifity와 Sensetivity를 나타낸 그래프이다. 임계값이 증가함에 따라 Specifity는 증가하고 Sensetivity는 감소하는 것을 확인할 수 있다. DEVAC 예비검사 방법에 비해 Specifity와 Sensetivity는 크게 떨어지는 것으로 나타나지만 문서 필터링 성능에서 Specifity와 Sensetivity는 크게 중요하지 않고 단지 임계값을 설정하기 위해서 사용되어 진다.

4. 실 험

4.1 실험을 위한 데이터

표절 검색 문서군의 개수를 최적화 하는 문서 필터링 방법들의 성능을 확인하기 위하여 사용한 테스트 데이터는 다음과 같다.

표 1. 테스트 데이터 정보

	제작된 표절문서셋
문서의 갯수	690
문서의 쌍	115
문서당 평균 단어 갯수	72.78
최대 단어 수	155
최소 단어 수	36

제작된 표절 셋은 표절 검사 성능을 테스트하기 위해서 의도적으로 만든 표절 셋이다. 각 문서 쌍은 6개의 문서로 이루어져 있으며 문서 쌍은 총 115개로 되어 있으며 총 문서의 개수는 690개 이다. 각각의 문서는 평균 72개의 단어로 이루어져 있으며 최대 단어 수는 155개이며 최소 단어 수는 36개이다.

4.2 Score값 분포도

Okapi BM25의 임계값 별 Score 분포도는 다음과 같다.

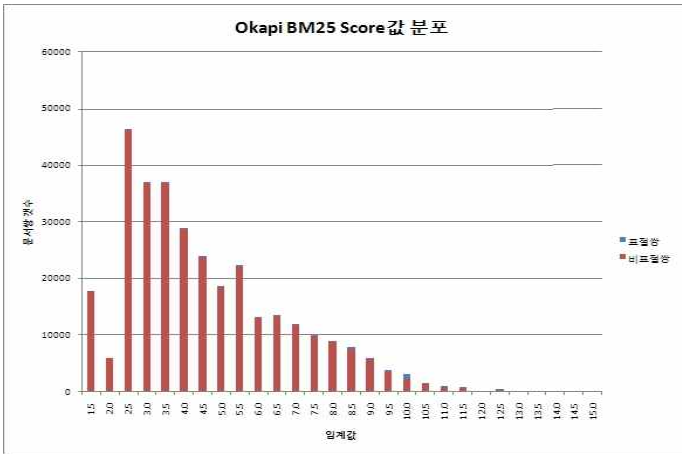

그림 3. Okapi BM25의 Score값 분포도

그래프의 x축은 임계값을 나타내며 y축은 문서 쌍의 Score값이 x축의 임계값 영역에 존재하는 문서 쌍의 개수이다. 빨간색 막대 그래프는 표절 상관관계가 없는 문서 쌍을 나타내며 파란색은 서로 표절 상관관계에 있는 문서 쌍을 나타낸다. 위의 그래프를 보면 표절 상관관계가 없는 문서 쌍들은 낮은 임계값 영역에 분포함을 알 수 있다.

다음 그래프는 DEVAC 예비 검사 방식의 Score 분포도이다.

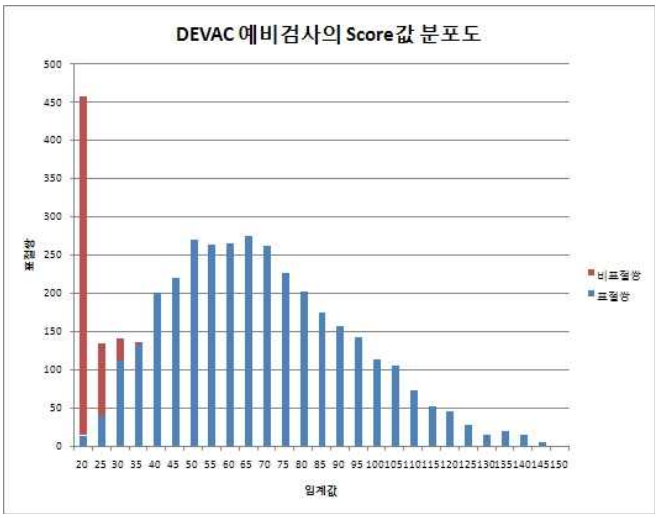

그림 4. DEVAC 예비 검사의 Score값 분포도

DEVAC 예비 검사의 Score값 분포는 위의 그래프와 같다. 낮은 임계값 영역에 비 표절 문서 쌍들이 존재하고 있으며 임계값을 조정하여 문서 필터링을 수행 할 수 있음을 알 수 있다.

4.3 임계값 별 문서군 크기 분포

임계값 별로 문서를 필터링 했을 경우 표절 비교 문서군의 크기가 어떻게 변화하는지를 알아보면 문서의 필터링을 위한 임계값 설정에 도움이 된다. 다음은 Okapi BM25의 임계값 별 문서군 분포 그래프이다.

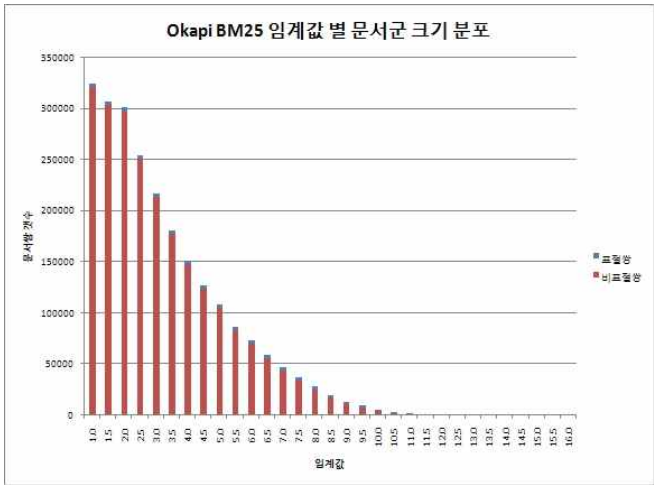

그림 5. Okapi BM25의 임계값 별 문서군 크기

위의 그래프는 Okapi BM25에서 임계값에 따라 문서군의 크기가 어떻게 변화하는지를 나타내고 있다. 임계값이 증가함에 따라 문서군의 크기가 감소하는 것을 확인 할 수 있다. 특히 비 표절 쌍의 크기가 눈에 띄게 감소하는 것을 확인 할 수 있는데 이는 임계값을 조정함에 따라 표절 심층 검사에서 필요하지 않은 문서 군들이 필터링 되어 성능 향상이 가능함을 보여주고 있다.

다음은 DEVAC 예비 검사에서의 임계값 별 문서군 크기 분포 그래프이다.

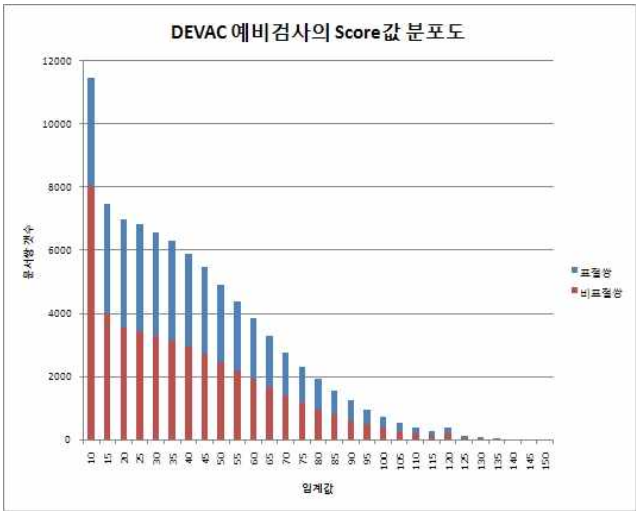

그림 6. DEVAC 예비검사의 Score값 분포도

위의 그래프는 DEVAC 예비검사에서 임계값에 따라 문서군의 크기가 어떻게 변화하는 지를 나타내고 있다. 임계값이 증가함에 따라 비 표절 쌍이 눈에 띄게 줄어드는 것을 확인 할 수 있다.

5. 결 론

최근 한글 문서 표절이 사회적으로 이슈가 되면서 표절을 탐색해 내거나 사전에 표절을 차단하기 위한 시스템의 필요성이 높아지고 있다. 표절 탐색 시스템은 성능과 속도 두 가지 요소를 모두 만족해야 하는데, 입력 데이터의 크기에 따라 성능에 영향을 받는 표절 탐색 시스템에서는 표절 대조 문서군의 크기를 제한하여 표절 탐색 시스템의 성능을 향상시키는 방법이 필요하게 되었다. 빠른 응답 속도와 비교적 정확한 결과를 가지기 위한 방법으로 웹검색에서 사용되는 Inverted Index방식이 제시되었고 이를 이용하는 두 방법의 성능을 살펴 보았다.

이 논문에서는 Inverted Index를 사용하는 Okapi BM25와 DEVAC시스템의 예비검사 방법을 알아보고 이 방법들의 성능을 측정해 보았다. 또한 그 결과를 분석하여 Document filtering에 있어서 적당한 임계값을 찾아서 표절 탐색 성능이 얼마나 향상되는지를 확인해 보았다.

참고 문헌

[1] Sven Meyer zu Eissen and Benno Stein, Intrinsic plagiarism detection, Lecture Notes in Computer Science, Vol 3936, page 565-569. Springer, 2006

[2] CloneChecker: A Software Plagiarism Detector. http://ropas.snu.ac.kr/n/clonechecker

[3] Narayanan Shivakumar and Hector Garcia-Molina, SCAM: A copy detection mechanism for digital documents. The 2nd International Conference on Theory and Practice of Digital Libraries, 1995

[4] Geoff Whale, Plague user mannual(release1.2), Department of Computer Science, University of New South Wales, 1989

[5] Wise, YAP3: Improved detection of similarities in computer program and other texts, SIGCSEB: SIGCSE Bulletin (ACM Special Interest Group on Computer Science Education), 28, 1996

[6] David Gitchell and Nicholas Tran, SIM: a utility for detecting similarity in computer programs. SIGCSE '99: The proceedings of the thirtieth SIGCSE technical symposium on Computer science education, pages 266-270, ACM Press, 1999

[7] Saul Schleimer, Daniel S. Wilkerson and Alex Aiken, Winnowing : local algorithms for document fingerprinting, SIGMOD '03: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pages 76-85, ACM, 2003

[8] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu and Mike Gatford, Okapi at TREC-3, Proceedings of the Third Text REtriveal Conference(TREC 1994), Gaithersburg, USA, Novemver 1994