제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

대용량 문서 집합에서 유사문서 탐색 시스템의 성능 개선

박선영*, 조환규*
*부산대학교 정보컴퓨터공학부
e-mail:parksy@pusan.ac.kr

Performance Improvement on Similar Texts Searching System for Massive Document Repository

Sun-Young Park*, Hwan-Gue Cho*
*Dept of Computer Science and Engineering, Pusan National University

요 약

최근 발생한 수많은 표절 논란으로 인해 많은 유사 문서 탐색 시스템이 개발되어 사용되고 있다. 많은 시스템 중 내용기반 유사문서 탐색 시스템인 DeVAC은 대용량 문서 1:1간의 비교에서 빠른 성능을 보여주지만 수천~수만 개의 문서 집합에 대해서는 적절한 성능을 보여주지 못한다. 이를 해결하기 위해 전역 사전(Global Dictionary)을 이용한 전처리 방법이 고안되어 적용되었다. 이 전처리 방법을 통해 비교해야 할 문서쌍이 줄어들고 전체 시스템의 성능을 향상시킬 수 있다는 것은 밝혀졌으나, 전처리를 위해 발생하는 추가 비용에 대한 계측이 이루어지지 않았을 뿐 아니라 문서 쌍이 얼마나 감소하는지 측정한 실험에서도 언어 처리용 실험적 데이터(말뭉치)에 대한 실험이 대부분을 차지하였기 때문에 실제 데이터에 대해 어떤 성능을 보일지 정확히 예측할 수 없었다. 본 논문에서는 전체 시스템에서 전처리를 위해 필요한 모든 추가 비용을 측정하고, 데이터를 1.5Gb, 6263개의 문서로 이루어진 실존하는 문서 집합으로 구성하여 성능 향상 정도를 측정함으로써 실제 데이터에 대한 전처리 신뢰도를 예측하였다. 실험 결과 전처리 후 찾아낸 유사한 문서 쌍을 전처리를 하지 않을 경우의 80~89.3% 정도로 유지하면서 검사 시간을 기존의 10.8%~15.4% 수준으로 대폭 감소시킬 수 있었다.

1. 서론

최근 학계와 예술계의 잦은 표절 논란으로 인해 연구 윤리, 창작 윤리 등이 크게 부각되고 있다[1]. 특히 전자 문서의 경우 상대적으로 표절하기가 용이하면서도 표절한 구간과 그렇지 않은 구간을 적절히 뒤섞어 놓을 경우 육안으로는 이를 찾아내기 어렵다는 특성 때문에 문서 간 표절 검사 시스템의 연구가 활발하게 이루어지고 있는 추세이다. 또한 인터넷의 발달로 인해 일반적인 사용자가 접근 가능한 문서의 수가 많아지고 문서의 크기도 대형화됨에 따라 대용량 및 다량의 문서 집합에서의 유사도 탐색 시간 역시 시스템의 중요한 성능 인자 중 하나가 되었다. 본 논문에서는 전역 사전을 이용한 전처리 시스템을 간단히 소개하고 이를 모듈화하여 이미 개발된 문서 유사도 탐색 시스템에 적용한 후, 전처리 모델에서 설정 가능한 여러 변수의 값을 조절하면서 실존하는 거대 문서 집합 간 유사도를 측정하여 계산 시간 및 탐색의 정확도를 비교하였다.

2. 관련 연구

유사 문서 탐색 시스템의 경우 문서 내의 단어 빈도를 측정하여 유사 문서를 탐색하는 Attribute Counting[2] 방법과 문서의 구조를 분석한 후 토큰 스트링의 유사성을 계산하는 Structured Metric 방법이 널리 사용[3]된다. 두 방법의 장점을 결합한 하이브리드 방법[4]도 존재하는데, 이 방법을 사용한 대표적인 예 중 하나가 내용 기반 유사 문서 탐색 시스템인 DeVAC(Document eVolution Analysis Center)[5]이다. 이 시스템은 문맥적 의미를 이용한 방법[6]을 완전히 배제하고 Attribute Counting의 하나인 fingerprint[7] 방식과 Structured Metric 방법을 효과적으로 결합한 방법을 사용함으로써 1:1 문서 탐색의 경우 10만 단어 이상의 문서에서도 매우 빠른 성능과 탐색 정확도를 보여준다. 다만 대용량 문서 집합의 경우 발생 가능한 모든 쌍에 대하여 검사를 수행하기 때문에 집합 내 문서의 수가 많아질수록 계산 시간이 기하급수적으로 증가한다는 문제가 있다. 이러한 문제를 해결하기 위하여 대용량 문서 집합에서 '전역 사전을 이용한 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계[8]'와 같은 연구가 진행되었으며, 이 연구에서는 유사 문서 탐색을 진행하는 과정에서 발생하는 단어 조각의 사전(Dictionary, DIC)을 종합한 전역 사전(Global Dictionary, GDIC)을 생성한 후 이를 이용하여 비교할 문서 쌍의 개수를 줄임으로써 탐색 시간을 줄일 수 있다는 것을 보여주었다. 여기에는 네 가지 환경 변수가 사용되는데, 불용어를 걸러낼 때 걸

러낼 비율을 결정하는 비율인 T_{ratio}, 유사 문서 쌍 후보로 등록하기 위한 최소의 공통 키 등장 횟수의 기준을 나타내는 N_{match}, 후보로 등록되기 위한 두 문서 간의 공통 키 등장 횟수 합계 기준을 나타내는 S_{match}, 후보로 등록하기 위한 공통 키 등장 비율의 기준을 나타내는 C_{ratio} 등으로, 이러한 변수를 효과적으로 설정(T_{ratio} : 0.002 ~ 0.005, N_{match} : 2 ~ 4, S_{match} : 7 ~ 12, C_{ratio} : 0.01 ~ 0.10)하면 검사할 문서 쌍의 개수를 90% 이상 줄일 수 있다는 것을 확인하였다.

3. 전처리 모듈의 한계

이전 전처리 모듈의 연구에서는 비교할 문서 쌍이 90%이상 줄어듦으로써, 전처리를 위해 발생되는 추가 비용을 고려해도 전체 시스템 성능이 향상된다는 것까지는 확인하였으나 실제로 전역 사전을 생성하는데 소모되는 시간과 전처리에 필요한 시간 등은 정확히 측정되지 않았다. 또한 성능 탐색을 위한 실험에 비슷한 형식으로 작성된 실제 데이터가 아니라 실험용 대용량 말뭉치 데이터를 사용함으로써 실재하는 유사 문서 탐색 성능의 향상 및 신뢰도 유지 여부를 완전하게 검증하지는 못했다. 즉, 전처리 모듈을 적용함으로써 전체 탐색 시스템의 성능이 향상됨을 보이기 위해서는 비슷한 성질(형태)를 갖는 실제 대용량 문서 집합을 사용하여 실험을 진행함으로써 전처리 시스템의 성능을 검증할 필요가 있다. 또한 전역 사전 생성 시간과 전처리에 필요한 계산 시간을 고려하여 전체 시스템의 성능 향상 수준을 정확히 측정하여야 한다. 실제 데이터에 대해 전처리 과정에서 탐색 결과가 최대한 보존되는 범위 내에서 가장 강력한 성능을 낼 수 있는 환경 변수를 찾아낼 필요가 있다.

4. 실험

4.1. 실험 데이터

실험에 사용된 데이터는 1999~2009년의 정부 정책 연구와 관련한 보고용 문서 6808건 중 10 ~ 120,000 개의 어절로 이루어진 6263건의 문서를 추출하여 사용하였다. 문서 집합의 총 용량은 1.52GB로, 문서 하나의 평균 크기는 250KB이다.

4.2. 실험 환경

실험은 Intel Xeon 서버 머신으로 진행하였으며, 사양은 <표 1>과 같다.

<표 1> 실험에 사용한 서버 머신의 사양 표

구분	성능
CPU	Intel Xeon 2000MHz
RAM	4096 MB
Graphic Card	Geforce GT 6600
HDD	300GB × 2

4.3. 실험 방법

실험은 크게 전처리 수행 없이 유사 문서 탐색을 진행하는 경우와 전처리 수행 후 유사 문서 탐색을 진행하는 경우로 나뉜다. 또한 각 경우에 대하여 연산 시간을 측정하는 실험과 유사도 500 이상의 구간이 존재하는 문서 쌍을 몇 개나 찾아냈는지 측정하는 실험으로 나뉜다. 여기에서 유사도란, 두 문서 사이에서 특정 구간이 얼마나 유사한지를 나타내는 점수이다. 일반적으로 두 문서 사이에서 100~150 이상의 구간이 존재한다면, 사람이 직접 해당 구간을 보고 표절 여부를 판단해볼 필요가 있다. 만약 유사도 500 이상의 구간이 존재한다면, 해당 문서의 두 저자가 각자 자신의 생각대로 해당 문구를 삽입하였음에도 그 구간이 우연히 일치할 가능성은 거의 없다고 볼 수 있다. 따라서 유사도 500 이상의 구간이 존재하는 문서 쌍의 개수가 전처리 유무에 따라 어떻게 달라지는지 측정한다면 전처리로 인한 유사 문서의 누락이 얼마나 발생하는지 판단할 수 있다.

우선 전처리 과정 없이 유사 문서 탐색을 수행할 경우 (6263×(6263-1))/2=19,609,453 회의 비교를 수행하여야 하는데, 이렇게 측정할 경우 시간이 너무 오래 걸리므로 데이터를 분할하여 측정한 후 전체 데이터에 대한 추정치를 계산하였다. 전처리를 하지 않을 경우 전체 시간은 문서의 메모리 적재 시간과 검사 시간으로 구성되는데, 6263건에 대한 메모리 적재 시간을 10회 측정하여 평균을 구하고, 검사 시간은 샘플링을 통해 전체 데이터에 해당하는 계산 시간을 추정하였다. 즉 전체 데이터의 10%에 해당하는 626건의 데이터를 random하게 잘라낸 후 이를 가지고 실험을 진행하면 195,625번의 비교를 수행하게 되므로, 데이터를 10등분하여 각각 2회씩 실험한 후 평균값의 100.24배 (195,625 × 100.24 = 19609450)를 곱하여 6263건의 실제 검사 시간에 해당하는 값과, 유사도 500 이상의 문서 쌍의 개수를 예측하였다.

전처리 수행 후 검사 시간은 전처리 시스템을 설계[7]하는 과정에서 확인한 T_{ratio}, N_{match}, S_{match}, C_{ratio} 등의 환경 변수값의 적정 범위(T_{ratio} : 0.002 ~ 0.005, N_{match} : 2 ~ 4, S_{match} : 7 ~ 12, C_{ratio} : 0.01 ~ 0.10)를 사용하여, 추가로 전역 사전을 생성하는 데 걸리는 시간, 전처리 과정에 필요한 시간 등을 모두 측정한 후 전체 시스템에서 걸리는 시간과 유사문서 쌍의 개수를 측정하였다.

4.4. 실험 결과

전처리 수행 없이 문서를 10등분하여 실험한 결과는 <표 2>와 같다. 실험 결과 626 건에 대해 평균 1445.4초의 검사 시간이 소요되었다. 6263건에 대한 예상 시간을 구하기 위해 100.24를 곱하면 144886.9초이고, 메모리 적재 시간 151초를 더하면 최종적으로 전처리 없이 6263건에 대한 유사 문서 탐색을 수행했을 경우 예상되는 시간은 145037.9(40.3시간)이다. 또한 마찬가지 방법으로 유사도 500 이상의 유사문서 쌍의 개수를 1303개로 예측하였다.

<표 2> 전처리 수행 없이 626건으로 이루어진 각 그룹에 대한 검사를 수행한 결과. 626건당 평균 1445.4sec 정도가 소요되며, 6263건으로 환산할 경우 대략 40.2시간이 소요될 것으로 추정됨. 유사 문서의 개수는 유사도가 500점 이상으로 측정된 문서쌍의 개수이며 동일 sample에 대한 검사이므로 1, 2차의 개수가 반드시 같아야 함.

그룹	소요 시간(sec)		유사 문서 쌍(개)	
	1차	2차	1차	2차
1	1469	1485	10	10
2	1359	1328	12	12
3	1501	1464	16	16
4	1425	1399	12	12
5	1463	1417	13	13
6	1484	1443	15	15
7	1487	1437	14	14
8	1390	1347	9	9
9	1635	1534	12	12
10	1461	1380	17	17
합계	14674	14234	130	130
전체 평균	1445.4		13	
6263건에 대한 예측 값	144886.9		1303	

전처리 수행 시 필요한 전체 탐색 시간은 전역 사전 생성시간, 메모리 적재 시간, 필터링 시간, 검사 시간의 합으로 이루어진다. 이중 메모리 적재 시간은 평균 155 초로 전처리를 하지 않은 경우와 큰 차이가 없다. 전역 사전 생성 시간은 12473.2초 정도로, 3.4시간에 해당한다.

<표 3> 전처리 수행 결과 - 전처리 및 검사 시간. 단위는 초. 전역 사전 생성시간(12473.2초)과 메모리 적재 시간(151초)이 더해진 최종 탐색 시간이다. N_{match}/S_{match}와 C_{ratio}가 증가할수록 검사 시간은 뚜렷하게 감소하며, T_{ratio}는 0.005의 경우를 제외한 구간에서는 불규칙적인 시간 분포를 보인다. 결과적으로 환경 변수에 따라 차이는 있으나 전처리 없이 검사 수행 시 예측 값인 40.2시간에 비해 3.7시간(13990초)~6.2시간(22329초) 정도로 빠른 시간 내에 탐색이 완료되었다.

T_{ratio}				0.003	0.004	0.005
C_{ratio}	0.2	N/S match	2/7	17515	16754	22329
			3/10	16355	16082	19801
			4/13	15064	15293	14939
	0.5	N/S match	2/7	16151	16129	19724
			3/10	15251	15029	18712
			4/13	14407	14636	14309
	0.8	N/S match	2/7	15291	15682	18679
			3/10	14328	14744	16364
			4/13	13990	14169	14344

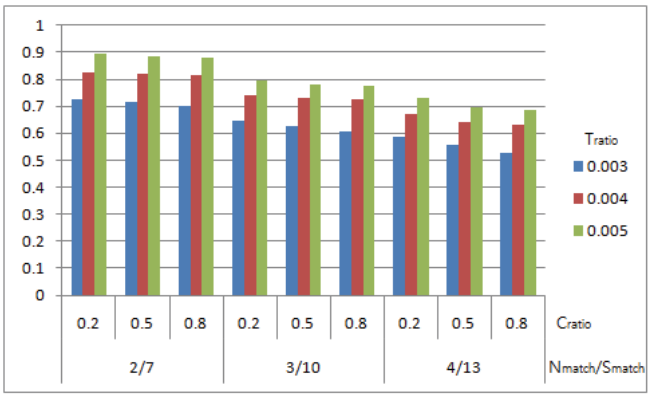
일단 전역 사전을 생성한 후에는 해당 문서 집합에 대해서는 이 사전을 재사용할 수 있지만 각 경우에 대해 전역 사전을 새로 만든다고 가정하고 각 실험의 측정값에 3.4시간을 더하였다. 각 환경 변수의 값을 달리하면서 필터링 시간과 검사 시간을 측정한 후 전역 사전 생성 시간과 메모리 적재 시간을 더한 결과는 <표 3>과 같다.

마찬가지 방법으로 유사도 500 이상의 유사문서 쌍의 개수를 측정한 결과는 <표 4>와 같다. 이상 두 실험에서 T_{ratio}의 실험 범위 0.002 ~ 0.005는 이전에 2만 개의 실험용 말뭉치 데이터로 실험했을 경우의 데이터인데, 이번 실험에서 T_{ratio}가 0.002인 경우, 50% 이하의 낮은 전처리 신뢰도를 보여 실험 결과에서 제외하였다. T_{ratio}는 문서 집합에서 문서의 총 개수에 큰 영향을 받는 변수인 만큼, T_{ratio}를 설정할 때에는 문서의 개수와 연동하여 책정하는 방법을 보완할 필요가 있다.

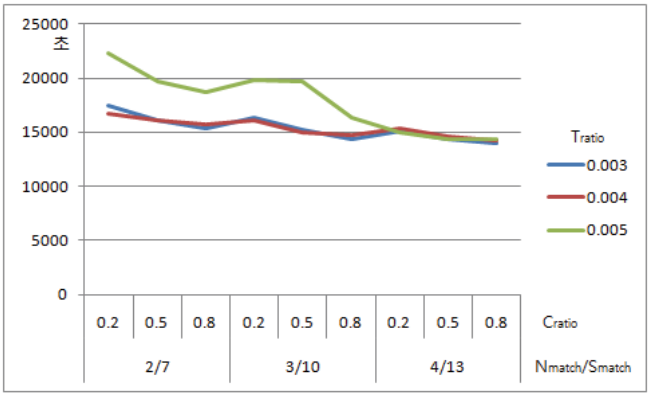
<표 4> 전처리 수행 결과 - 검사 후 유사도 500 이상으로 유사한 문서 쌍의 개수.N_{match}/S_{match}와 C_{ratio}가 감소할수록, T_{ratio}가 증가할수록 찾아낸 유사한 문서 쌍의 개수가 증가한다.

T_{ratio}				0.003	0.004	0.005
C_{ratio}	0.2	N/S match	2/7	948	1077	1164
			3/10	842	966	1037
			4/13	764	874	949
	0.5	N/S match	2/7	931	1067	1151
			3/10	818	950	1020
			4/13	725	837	909
	0.8	N/S match	2/7	914	1063	1147
			3/10	789	944	1013
			4/13	686	820	894

<표 4>의 결과와 6263건 전체에 대한 예측값인 1303을 비교하여 얻은 Sensitivity 그래프는 (그림 1)과 같으며, 각 환경변수와 연산 시간의 그래프는 (그림 2)와 같다. 두 그래프를 통해 각 환경 변수와 Sensitivity는 서로 밀접한 상관관계에 있으며, Sensitivity와 연산에 필요한 시간 사이에도 상관관계가 존재한다는 것을 알 수 있다.



(그림 1) 전처리 수행 결과 - 각 환경 변수에 따른 검사 후 유사도 500 이상의 문서에 대한 Sensitivity. 이 값이 1일 경우 전처리로 인해 누락된 유사 문서 쌍이 없다는 것을 의미하고, 0일 경우 모든 유사 문서쌍이 누락되었다는 것을 의미함. N_{match}/S_{match}와 C_{ratio}가 감소할수록, T_{ratio}가 증가할수록 Sensitivity가 증가한다.



(그림 2) 전처리 수행 결과 - 각 환경 변수에 따른 전체 탐색 시간(초). Sensitivity와 환경 변수의 상관관계에 비해 탐색 시간과 환경 변수 간의 상관관계가 명확하지는 않으나, N_{match}/S_{match}와 C_{ratio}가 감소할수록, T_{ratio}가 어느 정도 증가하는 양상을 보인다.

5. 결론 및 추후 연구

본 논문에서는 유사 문서 탐색 시스템의 대용량 문서 집합에 대한 성능을 개선하기 위하여, 전역 사전을 이용한 전처리 모델을 시스템에 적용하여 대용량 문서 전처리 과정의 성능 이득을 측정하였다. 이 과정에서 실험용 데이터가 아닌 실존하는 문서 집합을 사용함으로써 실험 결과의 신뢰도를 높였으며, 전처리 과정에서 발생하는 모든 추가 비용을 고려한 후 발생하는 실질적인 계산 시간의 이득을 측정하였다. 실험 결과 6263건의 실제 데이터에 대하여 Sensitivity를 80%로 유지할 경우 기존 연산시간의 10.8%, Sensitivity를 89.3%로 유지할 경우 기존 연산시간의 15.4% 정도의 시간만으로 검사를 완료함으로써 전체 탐색 시간이 큰 폭으로 줄어들었다는 것을 보였다. 다만 예전 실험에서 말뭉치 데이터에 대하여 Sensitivity가 100%를 유지했던 것에 비해 같은 수준의 환경 변수 값을 입력했음에도 불구하고 Sensitivity가 50%수준까지 떨어진 것에 대한 분석이 필요할 것으로 보인다. 특히 앞서 언급했듯이 문서 개수에 따른 T_{ratio} 책정 방법에 대한 연구가 반드시 필요하다고 생각된다. 또한 전체 검사 시간에서 전역 사전 생성 시간의 비율이 56.5% ~ 94.3% 정도로 매우 높은 비율을 보인 만큼, 이 시간을 줄일 수 있다면 시스템 전체의 성능을 더욱 향상시킬 수 있을 것으로 생각된다. 따라서 추후에는 다양한 실제 문서 집합에 대해 Sensitivity를 최대한 높이면서 성능을 향상시킬 수 있는 환경 변수 값에 대한 연구와 전역 사전의 생성 시간 감소를 위한 연구를 진행할 계획이다.

참고문헌

[1] 남형두, "표절과 저작권 침해," 창작과 권리 2009년 봄호(제54호), pp. 32-68, 세창출판사, 2009.

[2] 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.

[3] 류창건, 김형준, 조환규, 한글 말뭉치를 이용한 한글 표절 탐색 모델 개발, 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제14권, 제2호, pp. 231-235, 2008.

[4] 김형준, 조환규, 정렬을 이용한 내용기반 문서탐색 시스템의 전처리 과정 개선, 2008 한국컴퓨터종합학술대회 논문집, 제35호, 제1권(C), pp. 354-358, 2008.

[5] 류창건, 김형준, 박수현, 조환규, DeVAC(Document eVolution Analyzing Center), http://devac.cs.pusan.ac.kr/

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

[7] S. Schleimer, D. S. Wikerson, 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.

[8] 박선영, 김지훈, 김선영, 김형준, 조환규, 대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계, 한국정보과학회 2009 가을 학술발표논문집 제36권 제2호(A), pp. 76-77, 2009.