DEVAC 전처리 모듈에서의 환경 변수에 따른 시스템 성능 분석

System Performance Analysis of Preprocessing Module of DEVAC for Finding Similar Texts by Environment Parameters

박선영
Park Sun-Young
부산대학교 컴퓨터공학과
parksy@pusan.ac.kr

ABSTRACT

본 보고서에서는 DeVAC 전처리 모듈을 사용하여 전체 시스템의 성능을 최대로 향상시키기 위하여, 전처리에 사용되는 최적의 환경 변수를 찾아내기 위한 실험을 실시하여 이를 분석하였다. 국민일보에서 제공한 총 용량 1.5GB 분량의 정부 보고서 6263 건의 문서를 활용하여 전처리 수행 여부와 환경 변수의 변화에 따른 최적의 수행 조건을 측정하였다. 실험 결과 전처리 후 찾아낸 유사문서의 개수를 80% $\sim$ 89.3% 정도로 유지하면서 검사 시간을 기존의 10.8% $\sim$ 15.4% 수준으로 감소시킬 수 있었으나, 2 만 건의 말뭉치 데이터로 실험을 진행하였을 때보다 전처리 신뢰도가 낮아졌다. 이는 T_{ratio} 를 선택할 때 문서 집합 내 문서의 개수를 고려하지 않았기 때문으로 판단된다. 따라서 최적의 T_{ratio} 와 입력 문서의 개수 간의 상관관계에 대한 분석이 필요하다는 결론을 얻었다.

KEYWORDS Plagiarism, Similar Document, DeVAC, Preprocessing

1 서 론

최근 정치 문화 예술계 인사의 잦은 표절 의혹이 이슈가 되고 있다. 표절이란 다른 사람의 저작물의 전부나 일부를 그대로 또는 그 형태나 내용에 다소 변경을 가하여 자신의 것으로 제공 또는 제시하는 행위를 의미한다.[1] 표절 문제에 대한 사회적 관심이 높아지면서, 2010 년 3 월경 저작권 위원회에서 내부에 표절 위원회를 구성하여 표절 문제에 대한 대응에 나섰다.[2] 전자문서의 경우 상대적으로 표절하기가 용이하면서 표절한 구간과 그렇지 않은 구간을 적절히 뒤섞을 경우 육안으로는 이를 찾아내기 힘들다는 특성을 가지고 있다. 이 때문에 문서 유사도 탐색 시스템의 연구가 활발하게 이루어지고 있다. 인터넷과 검색 엔진의 급격한 발달로 인하여 일반 사용자가 접근 가능한 문서의 수가 기하급수적으로 늘어나고, 문서의 크기도 대형화됨에 따라 대용량 및 다량의 문서 집합에서 유사도 탐색 성능 역시 시스템의 중요한 성능 인자 중 하나가 되었다. 본 논문에서는 이전의 연구 중 전역 사전을 이용한 전처리 시스템을 간단히 소개하고, 이 시스템에서 설정 가능한 환경 변수의 값을 조절하면서 실존하는 거대 문서 집합간 유사도를 측정하는 실험을 통해 환경 변수와 탐색 시스템 성능 간의 상관관계를 밝힘과 동시에 최적의 환경 변수를 찾아내려 한다.

2 관련 연구

유사 문서를 탐색하는 방법에는 여러 가지가 있다. 가장 최근의 연구 결과를 살펴보면 통신 기사의 경우 원 저자권 개념에 기반한 뉴스 기사 표절 판정을 위한 프레임 워크[3]에 관한 연구가 진행되고 있다. 또한 내용 기반 유사 문서 탐색 시스템인 DeVAC[4] 의 경우 유사도 탐색을 위해 Attribute Counting 방법[?]의 일종인 Fingerprint[5] 방법과 Structured Metric 방법[6]의 장점을 적절히 취하여 용량이 큰 문서 간의 비교 성능을 개선하였으며, 최근 전역 사전(Global DICtionary, GDIC)을 이용한 전처리 모듈[7]을 적용하여 대용량 문서 집합에 대한 성능을 개선하였다. 전처리 성능을 검증하기 위하여 말뭉치 데이터를 사용하여 2만 건의 문서 집합을 생성한 후 이를 통해 실험함으로써 유사 문서쌍의 누락 없이 전처리를 할 수 있다는 것을 보였다. 전처리 시스템에는 네 가지 환경 변수가 존재한다. 불용어를 걸러낼 때 걸러낼 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 $\sim$ 0.005, N_{match} : 2 $\sim$ 4, S_{match} : 7 $\sim$ 12, C_{ratio} : 0.01 $\sim$ 0.10)하면 검사할 문서 쌍의 개수를 80% 이상 줄일 수 있다는 것을 확인하였다.

3 전처리 시스템 연구의 개선할 점

즉 이전 연구에서는 비교할 문서 쌍의 개수가 줄어듦으로써 전처리를 위해 발생되는 추가 비용을 고려하더라도 전체 시스템의 성능은 향상된다는 것을 증명하였다. 하지만 몇 가지 문제점과 함께 개선해야 할 점들이 있는데, 다음과 같다.

a. 전역 사전을 생성하는 데에 소모되는 시간, 메모리 적재 시간, 그리고 전처리 시간 등이 고려되지 않았기 때문에 전체 시스템에서 어느 정도의 성능 향상이 이루어지는지 확인하지 못했다.

b. 실험용으로 수집된 말뭉치 데이터를 사용하여 실험함으로써 실존하는 데이터에 대한 연산 시간의 감소에 대해 검증하지 못했다.

c. 실험용 데이터를 사용함으로 인해 실존하는 데이터에 대한 검사 신뢰도가 유지되는지 확인하지 못했다.

따라서 이 연구에서는 전체 시스템에서 전처리 모듈이 어느 정도의 성능 향상을 보여주는지를 실험을 통해 검증하고, 실존하는 데이터에 대해 환경 변수값을 조절하면서 실제 계산에 소요되는 시간과 신뢰도를 측정함으로써 실험 데이터가 아닌 실제 값에 대한 최적의 환경 변수값을 찾아낼 필요가 있다.

4 실 험

4.1 실험 데이터 및 환경

실험에 사용된 데이터는 1999 $\sim$ 2009년의 정부 정책 연구와 관련한 보고용 문서 6808건 중 10 $\sim$ 120,000 개의 어절로 이루어진 6263건의 문서를 추출하여 사용하였다. 문서 집합의 총 용량은 1.52GB

로, 문서 하나의 평균 크기는 250KB 이다. 또한 실험은 Intel Xeon 서버 머신으로 진행하였으며, 사양은 표 1 과 같다.

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

표 1. 실험에 사용한 서버 머신의 사양 표

4.2 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

	소요시간(sec)		유사문서쌍(개)	
그룹	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
전체평균	1445.40		13.00	
6263 건에 대한 예측값	144886.90		1303	

표 2. 전처리 수행 없이 626 건으로 이루어진 각 그룹에 대한 검사를 수행한 결과. 626 건당 평균 1445.40 초 정도가 소요되며, 6263 건으로 환산할 경우 대략 40.2 시간이 소요될 것으로 추정됨. 유사문서의 개수는 유사도가 500 점 이상으로 측정된 문서쌍의 개수를 나타낸 것으로, 동일 sample 에 대한 검사이므로 1,2 차 개수가 반드시 같아야 함

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 $(6263 \cdot (6263-1))/2 = 19,609,453$ 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 $(195,625 \cdot 100.24 = 19609450)$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 환경변수값의 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.

T_{ratio}				0.003	0.004	0.005
C_{ratio}	0.2	N_{match} / S_{match}	2 / 7	17515	16754	22329
			3 / 10	16335	16082	19801
			4 / 13	15064	15293	14939
	0.5	N_{match} / S_{match}	2 / 7	16151	16129	19724
			3 / 10	15251	15029	18712
			4 / 13	14407	14636	14309
	0.8	N_{match} / S_{match}	2 / 7	15291	15682	18679
			3 / 10	14328	14744	16364
			4 / 13	13990	14169	14344

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_{match} / S_{match}	2 / 7	948	1077	1164
			3 / 10	842	966	1037
			4 / 13	764	874	949
	0.5	N_{match} / S_{match}	2 / 7	931	1067	1151
			3 / 10	818	950	1020
			4 / 13	725	837	909
	0.8	N_{match} / S_{match}	2 / 7	914	1063	1147
			3 / 10	789	944	1013
			4 / 13	686	820	894

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.

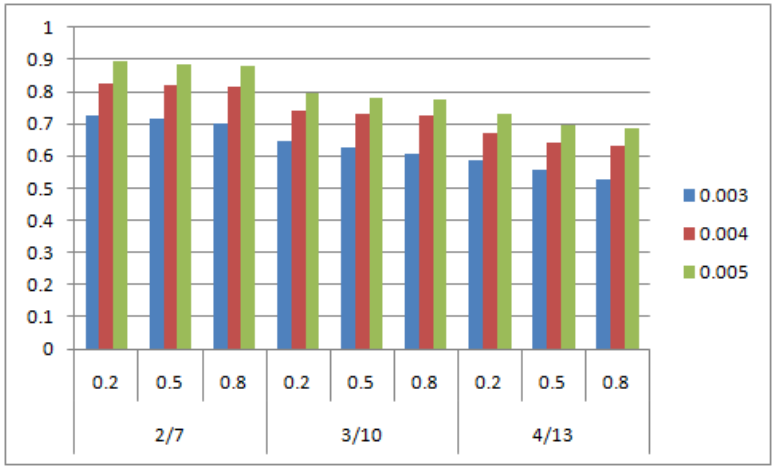

그림 1. 전처리 수행 결과 - 각 환경 변수에 따른 검사 후 유사도 500 이상의 문서에 대한 Sensitivity. 이 값이 1 일 경우 전처리로 인해 누락된 유사 문서 쌍이 없다는 것을 의미하고, 0 일 경우 모든 유사 문서쌍이 누락되었다는 것을 의미함. x 축의 윗 부분이 C_{ratio}, 아랫 부분이 N_{match}, S_{match} 이며, N_{match}, S_{match} 와 C_{ratio} 가 감소할수록, T_{ratio} 가 증가할수록 Sensitivity 가 증가한다. Sensitivity 와 연산 시간은 대체로 비례한다.

4.3 실험 결과

전처리 수행 없이 문서를 10 등분하여 실험한 결과는 표 2 와 같다. 실험 결과 626 건에 대해 평균 1445.4 초의 검사 시간이 소요되었다. 6263 건에 대한 예상 시간을 구하기 위해 100.24 를 곱하면 144886.9 초이고, 메모리 적재 시간 151 초를 더하면 최종적으로 전처리 없이 6263 건에 대한 유사 문서 탐색을 수행했을 경우 예상되는 시간은 145037.90(40.3 시간) 이다. 또한 마찬가지 방법으로 유사도

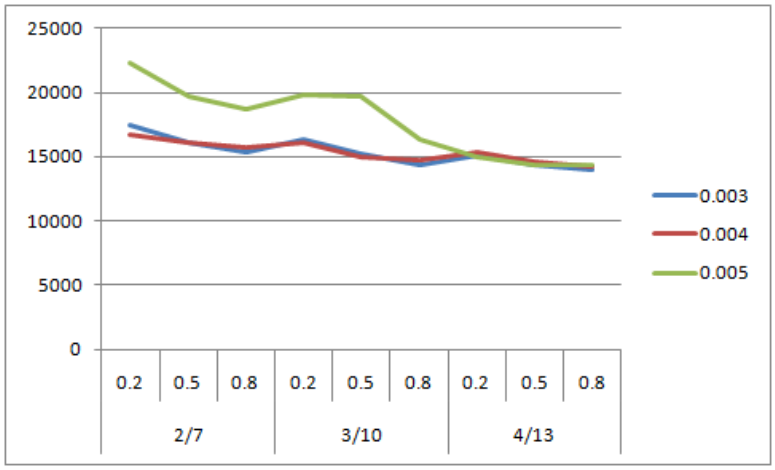

그림 2. 전처리 수행 결과 - 각 환경 변수에 따른 전체 탐색 시간(초). Sensitivity와 환경 변수의 상관관계에 비해 탐색 시간과 환경 변수 간의 상관관계가 명확하지는 않으나, N_{match}, S_{match}와 C_{ratio}가 감소할수록, T_{ratio}가 어느 정도 증가하는 양상을 보인다.

500 이상의 유사문서 쌍의 개수를 1303개로 예측하였다. 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 12473.20초 정도로, 3.4시간에 해당한다.

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 3과 같다. 마찬가지 방법으로 유사도 500 이상의 유사문서 쌍의 개수를 측정한 결과는 표 4와 같다. 이상 두 실험에서 Tratio의 실험 범위 0.002 ~ 0.005는 이전에 2만 개의 실험용 말뭉치 데이터로 실험했을 경우의 데이터인데, 이번 실험에서는 T_{ratio}가 0.002인 경우, 50% 이하의 낮은 Sensitivity를 보여 실험 결과에서 제외하였다. T_{ratio}는 문서 집합에서 문서의 총 개수에 큰 영향을 받는 변수인 만큼, Tratio를 설정할 때에는 문서의 개수와 연동하여 결정하는 방법을 보완할 필요가 있다고 판단된다.

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

5 결론

5.1 모델 적용 후 시스템 향성 정도

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 6263 건의 실제 데이터에 대하여 Sensitivity 를 80% 로 유지할 경우 기존 연산 시간의 10.8%, Sensitivity 를 89.3% 로 유지할 경우 기존 연산시간의 15.4% 정도의 시간만으로 검사를 완료함으로써 전체 탐색 시간이 큰 폭으로 줄어들었다는 것을 보였다.

5.2 추후 연구

예전 실험에서 말뭉치 데이터에 대하여 Sensitivity 가 100% 를 유지했던 것에 비해, 실제 데이터에 대해서 같은 수준의 환경 변수 값을 입력했음에도 불구하고 일부 환경 변수에 대해 Sensitivity 가 50% 수준까지 떨어진 것에 대한 분석이 필요하다. 이전 말뭉치 데이터의 경우 집합 내의 문서 개수가 2 만개에 달하기 때문에 전체 문서 중 T_{ratio} 가 0.003 정도의 비율의 문서라고 해도 60 개 정도가 되나, 문서의 개수가 6 천개로 줄어든 상황에서 0.003 의 비율은 18 개의 문서에 해당하므로, 불용어로 처리되지 않아야 할 일부 단어가 불용어로 처리되면서 Sensitivity 가 급격히 하락한 것으로 보인다. 따라서 문서 개수에 따라 유동적으로 T_{ratio} 를 책정하는 정책이 필요하며, 이에 대한 연구가 반드시 필요하다고 판단된다. 또한 전체 검사 시간에서 전역 사전 생성 시간의 비율이 56.5% $\sim$ 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 크기의 실제 문서 집합에 대해 Sensitivity 를 최대한 높이면서 성능을 향상시킬 수 있는 환경 변수 값에 대한 연구와 전역 사전의 생성 시간 감소를 위한 연구를 진행할 계획이다.

참고 문헌

1. “저작권 위원회,” http://www.copyright.or.kr/, 2010.
2. “특허청,” http://www.kipo.go.kr/, 2010.
3. 김정민, 정현숙, 이종영, 강남준, “뉴스 기사 표절 판정을 위한 시스템 프레임워크,” in *Proc. of the KIIT*, 2009, pp. 228–233.
4. 류창건, 김형준, 조환규, “한글 말뭉치를 이용한 한글 표절 탐색 모델 개발,” in *Proc. of the KIISE*, 2008, vol. 14, pp. 231–235.
5. S. Schleimer, D. S. Wikerson, and A. Aiken, “Winnowing : local algorithms for document fingerprinting,” in *Proc. of the ACM SIGMOD international conference on Management of data*. 2003, pp. 76–85, ACM.
6. A. Apostolico, “The myriad virtues of subword trees,” *Combinatorial Algorithms on Words*, vol. 37, no. 3, pp. 85–96, 1985.

7. 박선영, 김지훈, 김선영, 김형준, 조환규, "대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계," in *Proc. of the KIISE*, 2009, vol. 36, pp. 76–77.

8. J. L. Donaldson, A. Lancaster, and P.H. Sposato, "A plagiarism detection system," in *Proc. of the Twelfth SIGCSE Technical Symposium on Computer Science Education*, 1981, pp. 21–25.