일반화된 정렬 가중치 모델을 이용한 내용 기반 문서 검색 시스템

류창건°, 조환규
부산대학교 컴퓨터공학과
ckryu@pearl.cs.pusan.ac.kr, hgcho@pusan.ac.kr

Document-Content Retrieval System using General Alignment Weight Model

Chang-Keon Ryu°, Hwan-Gue Cho
Dept. of Computer Engineer, Pusan National University

이전 연구에서 제안했던 한글 표절 탐색 시스템(DEVAC)은 유전체의 상동성을 찾아내는 방법을 응용하여 사용하고 있다[1,2]. 문서를 표절할 경우 문장이 변해가는 과정이 유전체가 진화해나가는 모습과 유사하다는 아이디어를 기반으로 어절을 유전자로 문장의 순서를 유전자의 서열로 각각 대응시켜 유전자의 상동성을 검색하는 알고리즘을 적용시켰다. 이 유전자 상동성 검색 알고리즘으로 BLAST 방법을 사용하여 많은 대용량의 문서들 간의 표절 탐색을 빠른 시간 내에 정확하게 할 수 있게 된다[3]. BLAST 방법은 유전자에 삽입과 삭제가 발생할 때에도 그 상동성을 정확하게 탐색해주는 특징을 가지고 있으며, 이를 응용하여 만든 DEVAC 시스템도 표절 탐색에 이와 같은 장점을 가지고 있다[4]. BLAST 방법으로 문장 사이의 유사도를 측정하기 위해서는 갭 비용 방법을 사용하고 있으며, DEVAC 시스템은 단위 갭 벌점 모델을 사용하고 있다[5]. 본 논문은 이 갭 비용 방법을 더 나은 방법으로 변경하여 표절 탐색 시스템에 적용하고자 한다. 그리고 이 방법의 Specificity값과 Sensitivity값을 이전 방법의 해당 값들과 비교하여 시스템 성능의 향상을 증명하고자 한다. 또한 실험을 통해 구한 최적 파라미터 값을 이용하여, 실제 데이터에서 어파인 갭 벌점 모델이 어떤 영역에서 기능이 향상되는지를 보이고자 한다.

기존에 DEVAC 시스템이 사용하고 있는 갭 비용 방법은 단위 갭 벌점 모델로서 갭의 길이에 비례하여 벌점이 증가하는 방식이다. 이에 비해 새로 제안하고자 하는 방식은 어파인 갭 벌점 모델 방식으로 갭의 시작 부분에 비해 이후 갭에 대해서는 낮은 벌점을 부여하여 좀 더 노이즈에 강하지만 계산량은 크게 증가하지 않아 실용적인 알고리즘으로 평가받고 있다.

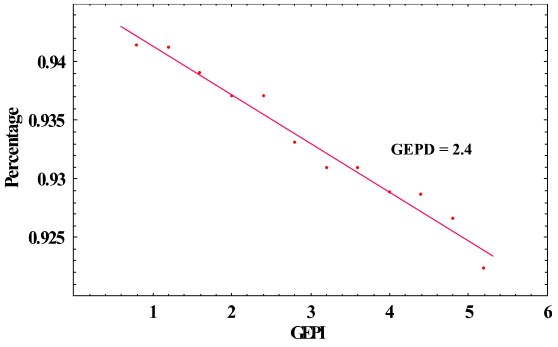

그림 1. 파라미터 값에 따른 specificity

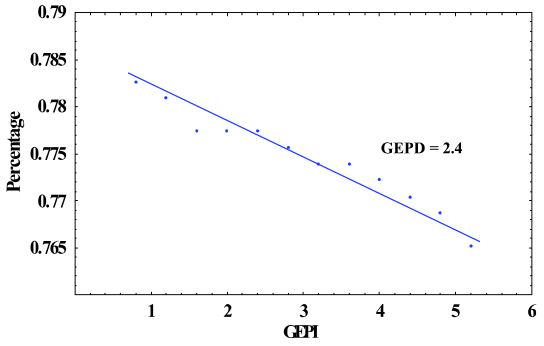

그림 2. 파라미터 값에 따른 sensitivity

생물정보학에서 사용하는 서열 정렬 방법을 한글 문서에 적용하기 위해서 기존 모델을 수정하여 시스템에 적용하였으며, 가장 좋은 성능을 얻기 위해 최적 파라미터 값을 구하였다.그림 1, 2는 어파인 갭 벌점 모델의 파라미터 값에 따른 specificity와 sensitivity를 나타내고 있다. GEPI(Gap Extend Penalty Insertion)는 두 번째 이후의 삽입 갭 벌점 값이며 GEPD(Gap Extend Penalty Deletion)는 두 번째 이후의 삭제 갭 벌점 값이다. 이 두 벌점 값은 어파인 갭 벌점 모델에서 새로 생긴 파라미터 값으로 두 값이 증가하여 5값에 가깝게 될 경우, 기존 모델과 동일하게 표절 탐색이 수행된다.이 실험을 통해 최적 파라미터 값을 찾았으며, 동시에 기존 시스템과 성능을 비교할 수 있었다.그림 1, 2를 살펴 보면 어파인 갭 벌점 모델을 적용할 경우, specificity와 sensitivity값 모두 증가하는 것을 알 수 있다.

표 1 단위 갭 벌점 모델 방식과 어파인 갭 벌점 모델 방식에서 찾아낸 유사 영역을 구성하는 단어 비율

	단위 갭	어파인 갭
유사 단어	68,998(66%)	69,855(55%)
다른 단어	12,082(12%)	19,075(15%)
삽입 단어	11,324(10%)	21,009(16%)
삭제 단어	12,072(12%)	17,302(14%)
합 계	104,476(100%)	127,241(100%)

표절 탐색의 성능을 알기 위해서는 실제 표절이 일어난 다량의 문서가 필요하다.이를 위해 다양한 원본 문서와 이를 순차적으로 표절한 문서들로 구성된 제작 표절 데이터를 제작하였다.690개의 표절 데이터를 제작하였으며, 이 데이터를 이용하여 표절 탐색을 수행하여 표1과 같은 결과를 얻었다. 단위 갭 벌점 모델 방식으로 시스템이 찾은 유사 영역은 총104,476단어였으며, 유사한 단어는 68,998 개였으며, 나머지 35,478개의 단어는 인위적으로 변경된 부분이었다.이와 같이 표절이 일어날 때엔 삽입과 삭제, 변형이 중요한 수단이며 이를 정확히 구별하여 찾는 것이 탐색 시스템의 성능에 많은 영향을 준다는 것을 알 수 있다. 만약 제작 표절 데이터에서 유사한 단어의 수로만 문서의 유사도를 계산하게 된다면 실제 유사도 값의 66%정도의 수치가 표시될 것이다.어파인 갭 벌점 모델은 이와 같이 표절 영역을 찾는데 중요한 요소인 삽입 삭제, 변형되는 단어들을 더 잘 찾기 위해 개선된 방법으로 일반 갭 방식과 동일한 실험을 한 결과 약 22%정도의 유사 영역을 더 찾았다. 유사 단어는 857개를 더 찾았으며, 인위적으로 변형이 일어난 부분은 21908개의 단어를 더 찾았다. 이는 삽입과 삭제, 변형을 통해 찾지 못했던 유사 단어를 추가적으로 더 찾은 것이다.표 1의 실험 결과를 통해 어파인 갭 벌점 모델이 삽입과 삭제, 변형에 더욱 강건하며 표절 탐색 성능도 더욱 뛰어나다는 것을 알 수 있다.

본 논문은 이전 연구에서 제안했던 한글 표절 탐색 시스템에 새로운 유사도 측정 방법을 적용하여 표절 탐색 성능이 향상되는 것을 증명하였으며, 실제 데이터를 이용하여 더 나아지는 기능이 무엇인지를 구체적으로 설명하였다.단위 갭 벌점 모델은 삽입이나 삭제되는 어절의 크기가 작을 경우에만 삽입이나 삭제된 문장 전체를 찾아 주지만, 클 경우에는 삽입이나 삭제된 문장에 의해 표절된 문장이 두 영역으로 나눠져 탐색된다. 이에 비해 어파인 갭 벌점 모델은 삽입이나 삭제된 어절의 크기에 관계없이 이 같은 문장을 찾아준다. 이를 다르게 본다면 단위 갭 벌점 모델은 지역적으로 표절을 찾는다고 할 수 있을 것이고, 어파인 갭 벌점 모델은 전체적으로 표절을 찾는다고 할 수 있다.어파인 갭 벌점 모델의 가장 큰 장점은 이와 같은 많은 장점을 가지면서 표절 탐색 시간이 단위 갭 벌점 모델과 거의 비슷하다는 것이다.

참고문헌

[1] Chang-Keon Ryu, Hyong-Jun Kim, Seung-Hyun Ji, Gyun Woo, and Hwan-Gue Cho. Detecting and tracing plagiarized documents by reconstruction plagiarism-evolution tree. Computer and International Technology 2008, July. 2008.

[2] Chang-Keon Ryu, Hyong-Jun Kim, and Hwan-Gue Cho. 한글 말뭉치를 이용한 한글 표절 탐색 모델 개발, 추계 정보과학회, Oct. 2007

[3] Bilu, Y., Agarwal, P. K., and Kolodny, R. 2006. Faster Algorithms for Optimal Multiple Sequence Alignment Based on Pairwise Comparisons. IEEE/ACM Trans. Comput. Biol. Bioinformatics 3, 4, 408-422. Oct. 2006.

[4] Cameron, M., Williams, H. E., and Cannane, A. Improved Gapped Alignment in BLAST. IEEE/ACM Trans. Comput. Biol. Bioinformatics 1, 3, 116-129. Jul. 2004.

[5] Michael Cameron, Hugh E. Williams, and Adam Cannane. Improved gapped alignment in blast. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 1(3):116-129, 2004.