제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

유사 어절 트리를 이용한 표절 문서의 Clustering 방법

천승환*, 김미영**, 이귀상*
*전남대학교 전산학과
**담양대학 인터넷 IT 공학부
e-mail:shcheon@gyosuclub.com

Clustering Method Of Plagiarism Document To Use Similarity Syntagma Tree

Seung-Hwan Cheon*, Mi-Young Kim**, Guee-Sang Lee*
*Dept of Computer Science, Chonnam National University
**Dept of Information Technology , Damyang College

요 약

인터넷과 컴퓨터를 이용한 학생들의 과제물을 평가하는데 있어 표절의 용이성으로 인해 정확히 판별하는 것은 매우 어렵고 번거로운 일이다. 특히 동일한 주제에 대해서 작성되는 경우가 많으므로 독자적으로 작성된 문서와 표절되어진 문서를 판별하기가 쉽지 않다. 이것은 클러스터링 하고자 하는 문서들에서 주요 단어들 즉, 색인어들의 출현 빈도를 추출한 뒤 이를 이용하여 가장 적합한 Clustering을 찾는 기존의 정보 검색 방법들과는 전혀 다른 문제이다. 본 논문에서는 과제물의 평가에 지침을 제공할 수 있도록 유사 어절 트리를 이용한 표절 유사도에 따른 Cluster들을 생성하는 방법에 대해 제안한다.

1. 서론

신속한 정보화가 진행되고 컴퓨터의 많은 이용으로 요즘 학생들은 과제물을 내줘도 쉽게 해결한다. 컴퓨터를 이용해서 남의 것을 일부 또는 전부를 표절할 수 있기 때문이다. 인터넷상에 과제물을 모아놓은 사이트가 많아서 쉽게 이용할 수 있을 뿐 만 아니라 그 조회수가 무척 높은 실정에 있다. 이러한 현실에서 표절에 대한 대안이 있다면 양질의 문서 및 자기 힘으로 독창적인 과제물을 제출하는 학생들이 훨씬 많아져서 건전한 학습 진행에 많은 도움이 될 것이라고 생각된다. 그러나 여러 가지 문서의 분류나 검색에 관한 기존의 방법들은 이미 분류되어 있는 문서들(training set)에 대하여 클러스터링 하고자 하는 문서들로부터 주요 단어들 즉, 색인어 들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 클러스터링을 찾는 것이다. 역설적으로 검색의 용이함과 문서편집의 편리함으로 인해 그 표절 여부를 판별하는 것은 더욱 어렵고 번거로운 일이다.

일반적으로 문서의 자동 클러스터링 방법에는 통계적인 방법과 지식기반의 방법이 있고 그리고 이들 방법의 상호 보완적인 특징을 이용한 복합적인 방법이 있다.[1,2,3,4] 통계적인 방법은 문서들로부터 단어들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 Clustering을 찾는 방법으로 단순한 방법이지만 가장 기본적으로 사용되고 있다. 지식기반 방법은 키워드 집합을 이용해 Clustering 하고자 하는 문서가 그 키워드 집합을 내포하는 정도에 따라 Clustering을 정하는 방법이다. 이러한 문서의 Clustering 방법[8]은 단어 색인어를 기반으로 유사도에 따라 Clustering 한 문서 검색의 활용에 한정적으로 활용되고 있지만 표절 여부를 판별하는데는 그대로 적용하지 못한다 할 수 있다. 특히 학생들의 과제물의 경우, 다른 도서와 문서들과는 달리 동일한 주제에 대해서 작성되는 특성과 과제물의 평가에 지침이 되는 표절 여부의 판별을 목적으로 한 특성으로 인해 일반적인 유사도 Clustering 방법과는 전혀 다른 문제라고 볼수 있다.

본 논문에서는 문서의 빠른 탐색을 위해 유사 어절 트리를 제안하고 어절 색인어를 생성하여 Clustering

하는 방법에 대해 알아보고 이를 기반으로 문서의 표절 여부를 판단할 수 있는 방법을 제안한다.

2. 관련연구

2.1. 문서의 Clustering 방법

1) 통계적인 문서 Clustering

사람에 의해 이미 분류되어 있는 문서들(training set)로부터 각 Clustering 카테고리에 나타나는 단어들의 출현빈도에 대한 정보를 추출하고, Clustering 하고자 하는 문서들로부터 주요 단어들과 단어들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 Clustering을 찾거나 각 Clustering에 대하여 포함 여부를 판단하는 방법이다. 여기에 속하는 대표적인 방법으로 벡터 유사도에 의한 Clustering 방법이 있다.[4,5,6] 이 방법은 Clustering 하려는 문서와 Clustering 대상 카테고리들을 색인어들의 벡터로 구성하고, 두 벡터 사이의 유사한 정도를 비교하여 유사도가 가장 높은 Clustering 카테고리로 문서를 분류하는 방법이다.

2) 지식기반 문서의 Clustering

지식기반 문서의 Clustering 방법에 하나인 키워드 집합에 의한 방법[3,4]은 Clustering의 단서가 되는 단어들의 집합으로 패턴을 정의하는 것이다. 이를 키워드 집합이라고 한다. 문장 내에 단어의 순서에는 상관없이, 한 문장 속의 단어들과 주어진 키워드 집합의 단어들 간에 일치하는 단어의 수가 어느 정적 수준 이상 존재하게 되면 그 문장은 주어진 키워드 집합으로 표현된 문장과 같은 내용을 갖는다고 간주하여 문서를 Clustering 한다.

3) 개념 기반의 문서분류 방법

대부분의 문서분류는 문서에 나타난 용어를 기반으로 한다. 그러나 개념적인 문서 분류를 하려면 문서의 표현 언어인 자연어를 분석해야 한다. 자연어 분석은 기술 수준의 정도에 따라 분석 결과가 색인어 추출 정도에서 문서의 뜻을 파악하는 정도까지 다양하다. 문장을 해석하는 방법으로는 시소러스를 이용하는데 개념을 상하위 분류하여 용어가 가지고 있는 개념의 획득을 일관성 있게 해준다.[4]

2.2 문서들의 Clustering 도출 방법

1) 유사도 행렬에 의한 Clustering[8]

문서와 색인어의 관계를 벡터로 표현하여 문서-색인어 행렬을 작성하고 문서와 문서간의 유사도를 측정하여 문서-문서 유사계수행렬을 형성한다. 그리고 유사계수의 기준치(threshold)를 정하고 기준치가 넘는 문서들끼리 모아 클러스터를 형성하고 각 클러스터를 대표하는 센트로이드 벡터(centroid vector)를 산출한다. 이에 대한 방법으로 싱글링크(single-link)식 기법이 있는데 이는 계층적 클러스터를 형성하는 기법으로 우선 유사계수 행렬로부터 일단의 클러스터를 형성한 뒤 이 클러스터를 다시 합쳐서 더 큰 크기의 클러스터를 형성한다. 이렇게 반복함으로써 클러스터의 계층이 형성된다. 싱글링크 기법은 문서간의 그림[1]의 (a) Dissimilarity matrix과 같이 상이계수행렬을 작성하고, 그림[1]의 (b), (c), (d)와 같이 클러스터수준(cluster level)을 달리하여 각 수준에서 클러스터를 형성한다. 결과적으로 클러스터의 계층이 생산되며 클러스터의 수준은 기준치(threshold)의 값으로 결정된다.

2	.4			
3	.4	.2		
4	.3	.3	.3	
5	.1	.4	.4	.1
	1	2	3	4

(a) Dissimilarity matrix

2	0			
3	0	0		
4	0	0	0	
5	1	0	0	1
	1	2	3	4

(b) Threshold = .1

2	0			
3	0	1		
4	0	0	0	
5	1	0	0	1
	1	2	3	4

(c) Threshold = .2

2	0			
3	0	1		
4	1	1	1	
5	1	0	0	1
	1	2	3	4

(d) Threshold = .3

[그림 1] Binary matrix

2) 자기 발견적 Clustering[7,8]

문서를 클러스터에 재배치함으로써 초기의 클러스터를 점차 정렬해 가며, 이 결과 최종의 클러스터들이 형성된다. 자기 발견적 Clustering은 클러스터의 센트로이드와 문서간의 유사도 측정에 기초하므로 클러스터 센트로이드의 형성이 선행되어야 한다. 클러스터 센트로이드는 벡터형태로 표현되며, 클러스터 내의 가장 중심적인 문서의 문서벡터가 센트로이드가 된다. 일반적으로는 클러스터에 속하는 문서들의 평균벡터를 산출하여 센트로이드로 삼는다.

$$\text{센트로이드용어}k = \frac{1}{m}\sum_{i=1}^{m} \text{용어}k_i$$

용어 k_i는 문헌 i가 갖는 용어 k의 가중치이고 클러스터는 m개의 문헌으로 구성된다.

3. 제안 방법

과제물의 표절 판별 Clustering은 문서들에서 주요 단어들 즉, 색인어들의 출현 빈도를 추출한 뒤 이를 이용하여 가장 적합한 Clustering을 찾는 기존의 일반적인 정보 검색 방법들과는 전혀 다른 문제이다. 본 논문에서는 먼저 학생들이 제출한 과제물들이 갖는 특징을 파악하고 과제물의 평가에 지침을 제공할 수 있도록

표절 유사도를 측정하기 위하여 유사 어절 트리를 구성하여 특히, 표절과정을 거쳐서 작성된 과제물들에 대해서 표절 판별을 목적으로 하는 작업에 적용될 수 있는 방법을 제안한다.

3.1. 과제물 문서의 특징

인터넷을 통한 자료 획득과 컴퓨터를 이용한 문서의 편집의 용이성으로 인해 과제물과 같은 문서는 다음과 같은 유형과 특징을 갖고 있으며, 이로 인해 유사하나 표절 과정이 없이 작성된 것과 유사하지 않으나 표절된 문서들에 대해 반드시 구분은 해야 되지만 동일한 주제에 대해서 작성된 경우가 대부분 이라 할 수 있기 때문에 그 판별은 더욱 어려움이 있다.

유형	특징
부분적 블록(문단) 복제	여러 문서를 원본으로 하여 각각 필요한 부분들만 발췌하여 새로이 작성된 문서
출현빈도가 높은 단어의 변환	중요 단어들을 유사의미를 갖는 단어나 전체적인 의미를 훼손하지 않는 범위 내에서 교체하여 작성된 문서
원문요약	잘 작성된 문서를 원문으로 하여 주요 내용은 훼손되지 않도록 하는 범위에서 부가적인 내용들을 삭제하여 작성된 내용
블록(문단)의 순서 변경	전체적인 문서의 각 문단이나 블록들의 순서를 일부 바꾸어서 원문과 상관없이 보이도록 작성된 문서
전체적인 복제	원문을 그대로 도용하여 작성된 문서

3.2 유사 어절 트리 구조

문서를 Clustering 하는데 있어 기존의 방법들에서 사용되어진 키워드가 되는 단어들을 추출하기 위해서 형태소 분석기를 거친 후 색인어 추출을 하였다. 그 예는 다음과 같다.

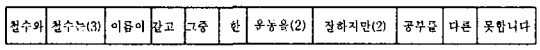
문서 A : 철수와 철수는 이름이 같고 그중 한 철수는 운동을 잘하지만 다른 철수는 공부를 잘하지만 운동을 못합니다.
문서 B : 영희와 영희는 이름이 같고 그중 한 영희는 운동을 잘하지만 다른 영희는 공부를 잘하지만 운동을 못합니다.

문서A의 형태소 분석 결과 : 철수(4), 이름, 운동(2), 공부
문서 B의 형태소 분석 결과 : 영희(4), 이름, 운동(2), 공부

일반적인 문서 분류 방법에서는 문서 A와 문서 B의 색인어로 각각 "철수"와 "영희" 등이 되므로 동일 카테고리에 속한다고 볼 수 없고 유사도 또한 낮다고 판별할 수 있는 허점이 있다. 그러나 두 문서가 표절되었다고 볼 수 있다.

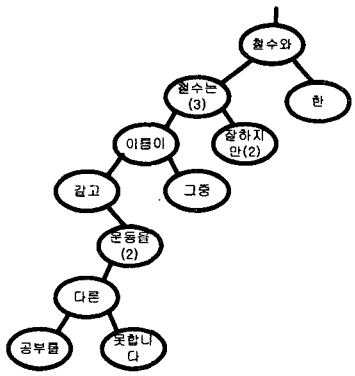
형태소 분석 과정을 포함하는 방법은 정보 검색 분야의 Clustering에 적합할 뿐 문서가 어느 정도 표절된 것인가에 대한 Clustering은 단어 기반의 색인어 보다는 어절 기반의 색인어을 생성하는 방법이 더 적합하다고 판단된다.

어절기반의 색인어를 찾기 위해 다음 [그림2]와 같이 Linked List를 이용하게 되면 문서 A 의 예로는 33회의 노드 탐색이 필요하여 수행시간의 Cost가 높다.



[그림 2] 문서A의 어절 링크

이를 개선하는 방법으로 다음 [그림 3]와 같이 문서 단위를 어절단위의 트리로 구성하되 문서내 어절의 순서에 따라 하위 레벨의 노드에는 상위 노드와 유사한 어절을 이루도록 한다. 결국 [그림2]에 비해서 탐색시간이 23회로 줄어들게 된어 수행시간의 Cost가 상대적으로 작다. 이는 동일한 어절의 빠른 탐색을 가능하게 하고 문서의 내용이 많을수록 트리를 구성하고 동일 어절의 출현 빈도수를 효과적으로 구하는데 효과적인 구조인 것이다.



[그림 3] 유사 어절 트리 구조

3.3 표절 유사도 클러스터의 생성

표절 판별을 목적으로 하는 문서의 Clustering 방법에서는 [그림3]의 유사 어절 트리를 이용하여 [그림5]에서와 같이 InOrder 인덱스를 구성하고 [그림6]와 같이 어절 빈도순 색인어 구조로 변환한다. 이는 비슷한 어절들의 클러스터를 생성한 구조를 얻어 유사 Clustering을 도출하는 응용에 유용하다 하겠다.

먼저 형태소 분석과정을 거친 [그림4]과 같이 생성된 단어 기반의 색인어들로 클러스터를 형성한 후 클러스터의 이외의 문서를 축출한다. 여기서 보면 동일한 주제 하에 작성된 문서들에서는 오히려 색인어 범주에 드는 단어들이 의미가 없게 되기 때문에 표절 판별을 위해 제외시킨다. 또한 출현빈도가 낮은 어절도 같은 개념으로 표절 정도의 측정에는 관

계될수 있으나 전체적인 문서의 주제의 표절 판별에는 관련성이 미미하다고 볼 수 있기 때문에 제외 시킨다.

철수	운동

[그림4] 단어 기반의 색인어

공부법	다른	못합니다	5고	운동을(2)	이럽이	그중	철수는(3)	잘하지만(2)	철수는	한

[그림5] 유사 어절트리의 InOrder 인덱스

철수는(3)	운동을(2)	잘하지만(2)	공부법	다른	못합니다	같고	이럽이	그중	철수와	한

[그림6] 어절 빈도순 인덱스

잘하지만	다른	못합니다	같고	이럽이	그중	철수와	한

[그림7] 비색인어 어절 벡터

그리고 각각 문서의 대표 클러스터 유사어절 트리를 생성하고 [그림6]와 같이 어절 빈도순 인덱스를 생성한다. [그림4]에서 색인어 와 어절 빈도가 낮은 어절 벡터를 제외 시킨 후 [그림7]과 같이 비색인어 어절 벡터를 생성한다. 각 문서의 비색인어 어절 벡터를 비교한 후 [그림1]과 같은 형식의 Dissimilarity matrix를 생성하고 유사도의 정도에 따른 다양한 산출을 위해서 다양한 Threshold을 주어 그에 따른 Binary matrix를 얻어 Binary matrix를 이용한 문서의 표절 유사도의 도식화를 만든다. 이는 유사한 문서들에 대해서 할 수 있는 정보를 가지고 있다. 결국 각각의 문서를 비색인어 어절 벡터에서 빈도수가 높은 색인어와 빈도수가 낮은 색인어 벡터를 제외한 후 유사계수 행렬을 구해 유사계수의 Theshold를 정하고 기준치가 넘는 문서들끼리 클러스터를 형성한다.

3.4 제안 알고리즘

문서 A를 예로 3.3절에서 기술한 내용을 다음과 같은 단계로 요약할 수 있다.

단계1 : 문서들의 형태소 분석 후 색인어 추출
단계2 : 색인어를 이용 Clustering 과정 수행
대표 클러스터의 이외의 문서 축출
단계3 : 대표 클러스터 문서 각각 유사 어절트리 생성
단계4 : 각 어절 트리에서 빈도순 인덱스 데이터 생성
단계5 : (단계1)의 색인어를 포함하는 어절 벡터와 빈도가 낮은 어절 벡터 삭제 후 비색인어 벡터 생성
단계6 : 각 문서의 비색인어 벡터를 비교하여 Dissimilarity Matrix 생성
단계7 : Threshold값에 따른 Binary Matrix 생성
단계8 : Binary Matrix를 이용한 문서의 표절 유사도의 도식화

4. 결론 및 향후과제

본 논문은 복제 및 표절의 판별을 효과적으로 하기 위해 유사 문서들간의 Clustering 판별 방법에 관하여 살펴보았다. 그리고 어절 트리 구조를 기반으로 Clustering을 적용하여 표절 유사도를 판별하는 방법에 대해 제안하였다. 결론적으로 어절 트리를 이용함으로써 예로든 문서 A의 경우 69% 탐색시간을 줄일 수 있음을 보였고 [그림7]과 같이 비색인어 어절 벡터를 이용할 경우 형태소 분석에서 얻은 색인어 벡터를 이용하는 방법에 비해 표절의 판별에 유리함을 보였다.

이는 제한적으로 문서의 복제나 표절의 판별에 이용된다면 양질의 문서를 판별하는데 부분적으로 기여할 수 있을 것이고 결과적으로 학생들의 과제물 표절 및 복제를 예방하고 합당하게 평가하는데 적용될 수 있을 것이다.

향후에는 이를 적용한 시스템을 설계 및 구현하여 그 결과를 검증할 것이며, 양질의 문서의 정도를 평가할 수 있는 개념기반의 Clustering 방법과 복합하여 적용시켜야 할 것이다.

5. 참고문헌

[1] Lewis.D. "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task", SIGIR'92.

[2] M. Blosseville. G. Hebrail, M. Monteil, and N. Penot., "Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together", SIGIR'92, 1992.

[3] K. M. Wong and Y. Y. Yao, "A statistical Similarity Measure, "In Proc. Intl. Conf. on Reasearch and Development in Information Retrieval. ACM SIGIR, pp. 3-12, 1987.

[4] 김준태, "지식기반 자연어처리를 이용한 문서의 자동분류와 지능형 색인에 관한 연구", 한국과학재단 연구결과보고서, 1998.4

[5] 최동시, 정경택, "주제와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현", 한국정보과학회 가을학술발표대회논문집, 22권 2호,1995.

[6] 최봉진, 김용성, 김순기, "2단계 필터링을 이용한 문서 선별 및 순위", 한국 정보처리학회 학술발표논문집(B), pp.315-317, 1999.

[7] 정영미, 이재윤, "지식 분류의 자동화를 위한 Clustering 모형연구", 정보관리학회지, pp.203-230, 2001

[8] 정영미, 정보검색론, 구미무역 출판사