論文
3-8-8

유사 어절 트리와 비 색인어 기반의 문서 표절 유사도 분류 방법 (The Classification Method of the Document Plagiarism Similarity based on Similar Syntagma Tree and Non-Index Term)

천 승 환* 김 미 영** 이 귀 상***
(Seung-Hwan Cheon) (Mi-Young Kim) (Guee-Sang Lee)

요 약

전자문서와 온라인으로 수신된 문서들은 표절 여부를 판별하기가 매우 어렵고 번거로운 일이다. 특히 학생들에게 부여된 과제물의 경우 동일한 주제에 대해서 작성되는 경우가 많으므로 독자적으로 작성된 문서와 표절되어진 문서를 판별하기가 쉽지 않다. 이것은 분류하고자 하는 문서들에서 주요 단어들 즉, 색인어들의 출현 빈도를 추출한 뒤 이를 이용하여 가장 적합한 카테고리를 찾는 기존의 방법들과는 전혀 다른 문제이다.

본 논문에서는 어절들의 -유사 어절 트리 구조와 색인어를 제외한 어절- 벡터를 기반으로 하여 비슷하게 작성된 문서들의 표절 판별을 목적으로 하는 작업에 적용될 수 있는 방법을 제안한다.

ABSTRACT

It is difficult and laborious to distinguish between the original and the plagiarism about the electrical documents or on-line received documents, specially student homeworks because in many case, the homeworks are written on the same subject. Existing methods are not appropriate to solve this problem, which find the most appropriate category using the expression frequency of index term in documents to be classified.

In this paper, a new classification method was proposed to distinguish between the original and the plagiarism about documents which were written similarly which is based on the syntagma vector - except the similar syntagma tree structure and non-index term.

1. 서론

신속한 정보화가 진행되면서 요즘 학생들은 어려운 과제물을 내줘도 쉽게 해결한다. 컴퓨터를 이용해서 남의 것을 일부 또는 전부를 표절할 수 있기 때문이다.

과제물 작성을 할 경우 거의 하루 전까지 내버려 두다가 당일 조금만 신경을 쓰면 짧은 시간 내에 쓰는 것은 문제가 없고 심지어 인터넷상에 과제물을 모아놓은 사이트가 많아서 쉽게 이용할 수 있을 뿐만 아니라 그 조회수가 무척 높은 실정에 있다. 이러한 현실에서 표절에 대한 대안이 있다면 양질의

* 정회원 : 전남대학교 전산학과 박사수료
** 정회원 : 담양대학 인터넷 IT 공학부 부교수
*** 정회원 : 전남대학교 전산학과 교수, 정보통신연구소

논문접수 : 2002. 7. 16.
심사완료 : 2002. 8. 12.

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 분류하고자 하는 문서들로부터 주요 단어들 즉, 색인어 들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 카테고리를 찾는 것이다. 역설적으로 검색의 용이함과 문서편집의 편리함으로 인해 그 표절 여부를 판별하는 것은 더욱 어렵고 번거로운 일이다.

일반적으로 문서의 자동 분류 방법에는 통계적인 방법과 지식기반의 방법이 있고 통계적인 방법의 대표적인 것으로는 분류하려는 문서와 분류 대상 카테고리들을 색인어 들의 벡터로 구성하고, 두 벡터 사이의 유사한 정도를 비교하여 유사도가 가장 높은 카테고리로 문서를 분류하는 벡터 유사도의 의한 분류 방법[1,2,4,5,6]이 있고 지식기반의 방법에는 분류의 단서가 되는 단어들의 집합으로 패턴을 정의하는 것으로 키워드 집합에 의한 방법이 있다.[7,8,9] 그리고 이들 방법의 상호 보완적인 특징을 이용한 복합적인 방법이 있는데 통계적 방법의 장점인 높은 분류율 과 지식기반 방법의 장점인 높은 정확도를 살려서 통계적 방법의 단점인 낮은 정확도와 지식기반 방법의 단점인 남은 분류율을 상호 보완할 수 있도록 통합하여야 한다. 지금까지 기술한 문서의 분류 방법은 용어를 기반으로 한 것들이다. 여기에 나아가 자연어를 분석하여 문서의 뜻을 파악하여 단순 용어가 아닌 개념을 기반으로 한 분류 방법이 있다.[2]

그러나 지금까지의 방법들은 문서의 유사도에 따른 자동분류를 기반으로 한 문서 검색의 활용에 제한적으로 활용되고 있는 실정으로 문서 분류의 또 하나의 분야인 표절 유사도 에는 그대로 적용하지 못한다. 형태소 분석과정을 거친 단어, 즉 색인어 들의 출현빈도나 문장의 의미만으로는 유사 여부는 파악할 수 있어도 표절 여부를 파악하는 것은 부적합하다고 볼 수 있다. 유사하지만 표절 과정이 없이 작성된 문서가 있고 유사하지 않지만 표절된 문서가 있을 수 있다. 한 문서가 어느 카테고리에 포함될 정도로 유사한 것인가를 아는 것과는 달리 전혀 다른 또 하나의 분야이자 방법을 적용해야 한다.

본 논문에서는 유사 어절 트리 구조를 이용하여 표절과정을 거쳐서 작성된 문서들에 대해서 표절 판별을 목적으로 하는 작업에 적용될 수 있는 색인어 절들을 오히려 제외한 차선의 어절 그룹을 기반으로 한 표절 유사도 분류 방법을 제안한다. 2장에서는 기존의 문서 자동 분류 방법들에 대해서 알아보고 3장에서는 이와는 방법적으로 유사하지만 그 최종 목적이 전혀 다른 표절 유사도 검사 방법에 대해서 알아보고 4장에서는 실험한 결과를 보이며 5장에서 결론을 맺는다.

2. 관련연구

2.1 문서의 자동분류 방법

(1) 통계적인 문서분류

사람에 의해 이미 분류되어 있는 문서들(training set)로부터 각 분류 카테고리에 나타나는 단어들의 출현빈도에 대한 정보를 추출하고, 분류하고자 하는 문서들로부터 주요 단어들과 단어들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 카테고리를 찾거나 각 카테고리에 대하여 포함 여부를 판단하는 방법이다. 여기에 속하는 대표적인 방법으로 벡터 유사도에 의한 분류 방법이 있다.[4,16] 이 방법은 분류하려는 문서와 분류 대상 카테고리들을 색인어들의 벡터로 구성하고, 두 벡터 사이의 유사한 정도를 비교하여 유사도가 가장 높은 카테고리로 문서를 분류하는 방법이다. 벡터 사이의 유사도는 두 벡터 사이의 각도를 계산하여 각도가 작은 경우가 높은 유사도를 갖도록 한다. 예를 들어 문서 D는 w_1, w_2, w_3, w_6의 색인어를 갖고, 카테고리 C_j는 w_1w_2, w_5, w_7, w_8의 색인어를 갖는다고 하면 벡터 D와 C_j는 D=(1,1,1,0,0,1,0,0), Cj=(1,1,0,0,0,1,1,1)이 된다. 두 벡터의 유사도는 다음과 같이 계산한다.

$$\text{similarity}(D,Cj) = \cos\theta = \frac{D * C_j}{|D||C_j|} \qquad \text{(식 1)}$$

여기에 정확도를 높이기 위하여 역문헌 빈도(IDF)를 이용한 방법이 있는데[9,16] 적은 수의 문서에 나

타난 색인어에 대해서 높은 가중치를 주는 것으로 색인어 W_i의 가중치는 다음과 같다

$$W_i = freq_{ij} * (\log(N)-\log(DF_i)+1) \quad (식\ 2)$$

$freq_{ij}$는 색인어 W_i의 문서 j에서의 빈도수, N은 총 문서의 개수, DF_i는 색인어 W_i를 포함하는 문서의 개수이다.

그러나 이 방법도 문제가 있는데 C_1, C_2를 분류 카테고리, D_1, D_2, D_3, D_4를 분류된 실험집단 문서, W_1, W_2를 색인어 라고 하고, 이들이 다음과 같이 분류되어 있다고 하자.

C_1 : D_1, D_3 W_1 : D_1, D_2에 나타남
C_2 : D_2, D_4 W_2 : D_1, D_3에 나타남

문서의 분류를 위해서는 W2가 W1보다 분류에 더 도움이 되지만 두 단어의 가중치가 같게 되어 그 특성을 반영하지 못하는 단점이 있다.

(2) 지식기반 문서의 분류

지식기반 문서의 분류 방법에 하나인 키워드 집합에 의한 방법[10,12]은 분류의 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 분류한다. 예를 들어 다음과 같은 하나의 키워드 집합은 아래의 두 문장에 모두 매치 되며 이들은 모두 같은 카테고리로 분류될 수 있다.[12,13,14]

키워드 집합 : {과속, 중앙선, 충돌}
문장 1 : "어젯밤 빗길을 과속으로 달리던 승용차가 중앙선을 넘어 앞에 오던 화물차와 정면 충돌 하였다."
문장 2 : "어젯밤에 일어난 고속도로 정면충돌 사건의 원인은 과속 주행하던 승용차의 중앙선 침범 때문인 것으로 밝혀졌다."

키워드 집합을 이용하는 방법은 문장내의 단어의 순서에 제약이 없으므로 더 많은 문장들이 매치 될 수 있는 반면, 그 만큼 다른 의미의 문장을 잘못 인식할 수 있는 가능성도 커지게 된다. 따라서 키워드 집합에 의한 패턴 표현에서는 구문 패턴에 의한 방법보다 평균적으로 더 많은 수의 단어들을 포함시킨다.

어떤 문서가 두개 이상의 서로 다른 카테고리를 나타내는 키워드 집합에 매치 되는 경우에는 매치된 키워드 집합간에 중요도를 고려하여 더 높은 중요도를 갖는 키워드 집합의 카테고리로 문서를 분류할 수 있다. 키워드 집합간에 중요도를 계산하는 방법에는 단어 집합이 가지고 있는 단어의 개수를 기준으로 하는 방법이 있는데 단어의 개수가 많은 키워드 집합일수록 중요하다고 볼 수 있는 이유는, 키워드 집합이 가지고 있는 단어가 개수가 많을수록 특정 카테고리를 더 정확하게 표현한다고 할 수 있으며 동시에 그 키워드 집합에 의한 매치가 오 분류일 확률이 상대적으로 낮기 때문이다.

(3) 복합적인 분류에 의한 방법

통계적 방법과 지식기반 방법은 상호 보완적인 특징을 가지고 있으므로 두 방법을 적절히 통합하면 분류 성능을 향상시킬 수 있다. 두 방법을 통합하는 방법은 여러 가지가 있을 수 있는데, 통합은 두 방법의 장점을 살피며 단점을 보완하는 방법이 되어야 한다. 즉, 통계적 방법의 장점인 높은 분류율과 지식기반 방법의 장점인 높은 정확도를 살려서 통계적 방법의 단점인 낮은 정확도와 지식기반 방법의 단점인 낮은 분류율을 상호보완 할 수 있도록 통합하여야 한다. 따라서 분류되지 않은 새로운 문서를 분류할 때, 패턴에 의한 분류 방법으로 1 차 분류를 시도한 후, 2 차로 통계적인 분류를 시도하면 모든 문서를 분류해 내면서 전체적인 분류 정확도를 높일 수 있다.[13,15]

2.2 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 색인

어 추출 정도에서 문서의 뜻을 파악하는 정도까지 다양하다. 문장을 해석하는 방법으로는 시소러스를 이용하는데 개념을 상하위 분류하여 용어가 가지고 있는 개념의 획득을 일관성 있게 해준다. 즉 개념 예를 들면 (물체)-(구체물)-(유생물)-(동물)-(인간), 이에 해당하는 단어는 (물건, 물체, 것)-(물체, 구체물, 물건, 개체)-(유생물)-(동물, 어류, 조류)-(사람, 철학자) 이다. 문서의 분류 영역에 시소러스를 사용한다면 단순 용어가 아닌 개념을 기반으로 한 문서분류를 할 수 있다. [13,14]

3. 제안 방법

먼저 표절된 문서의 유형들을 다양한 경우들에 대해서 분류를 해보고 그들 문서들이 갖는 특징을 파악 하는게 중요하다. 일반적으로 검색을 용이하게 하는 분류 방법과는 달리 표절의 정도를 도출하는데 본 논문의 목적이 있으므로 효과적으로 적용될 수 있는 방법으로 유사어절 트리를 이용한 문서의 구성 방법과 비 색인어 기반에서 문서의 표절 유사도 분류를 할 수 있는 방법에 대하여 제안한다.

3.1 표절된 문서의 유형

유 형	설 명
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 표절 판별을 위한 문서 분류의 특징

일반적으로 문서 분류 목적은 이미 분류되어 있는 문서들(training set)에 대하여 분류하고자 하는 문서들로부터 주요 단어들 즉 색인어 들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 카테고리를 찾는데 있다. 이러한 방법은 기본적으로 문서가 어느 정도 다른 문서들과 유사 하는가를 판별하는 것과 그 기법이 일치한다 할 수 있다.

그러나 형태소 분석과정을 거친 단어, 즉 색인어들의 출현빈도나 문장의 의미만으로는 유사 여부는 파악할 수 있어도 표절 여부를 파악하는 것은 부적합하다고 볼 수 있다. 유사하지만 표절 과정이 없이 작성된 문서가 있고 유사하지 않지만 표절된 문서가 있을 수 있다. 특히 동일한 주제에 대해서 작성된 문서들의 경우 독자적으로 작성된 문서와 표절되어진 문서를 판별하는 것은 기존의 문서 분류 방법과는 전혀 다른 문제이다.

특히 전자문서와 온라인으로 수신된 문서들에 대해서 표절 여부를 판별하는 것은 더욱 어렵고 번거로운 일이다. 그래서 본 논문에서는 표절과정을 거쳐서 작성된 문서들에 대해서 표절 판별을 목적으로 하는 작업에 적용될 수 있는 방법을 제안한다.

3.3 유사 어절 트리 구조

통계적인 문서 분류 방법들과 지식기반의 문서 분류 방법들에서 사용되어진 키워드가 되는 단어들을 추출하기 위해서 형태소 분석기를 거친 후 색인어 추출을 하는데 문서가 표절된 것인가, 아닌가는 문장에 대해서 적용할 수 있고 다시 어절을 분석하면 표절 여부에는 더 적합하다, 왜냐하면 단어들의 출현빈도는 문서의 카테고리별 분류에 적합하다면 단어보다는 조사를 포함한 어절 단위로 문장을 분석하는 것이 표절의 여부에 더 관계가 있다고 판단된다. 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.

문서 A의 형태소 분석 결과 : 철수(4), 이름, 운동(2), 공부

문서 A의 문서내의 어절 : 철수와, 철수는(3), 이름이, 같고 그 중, 한, 운동을(2), 잘하지만(2), 공부를, 다른, 못합니다.

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

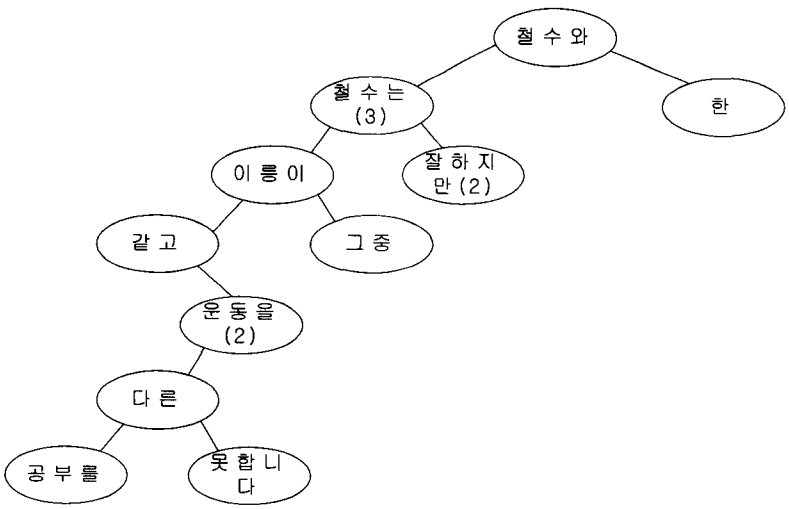
문서 B의 문서내의 어절 : 영희와, 영희는(3), 이름이, 같고 그 중, 한, 운동을(2), 잘지만(2), 공부를, 다른, 못합니다.

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. 이는 문서내의 어절들의 정보를 이용하면 유사한 문서로써 특히 표절의 정도가 매우 높다고 볼 수 있다.

본 논문에서는 이를 개선하는 방법으로 문서단위를 어절단위의 트리로 구성하되 하위 레벨의 노드에는 상위 노드와 유사한 어절을 이루도록 하여 동일한 어절의 빠른 검색을 가능하게 하였다. 문서내의 내용이 많을수록 트리를 구성하고 동일 어절의 출현 빈도수를 찾는데 유리한 구조인 것이다.

3.4 비 색인어절 기반의 유사도 판별

표절 판별을 목적으로 하는 문서의 분류 방법에서는 [그림 1] 의 어절의 출현 빈도에 따라 통계적으로 빈도수가 높은 어절을 문서의 유사도 분류, 즉 표절 판별의 벡터 값으로 추출 하기 보다는 각 문서들에 대한 어절 트리를 [그림 2](a)와 같이 InOrder 인덱스 구조로 변환한다. 이는 비슷한 어절들이 그룹 지어지는 구조를 얻어 다음 응용에 유용하다 하겠다.



[그림 1] 유사 어절 트리 구조

[Fig. 1] similar syntagma tree structure

공부를	다른	못합니다	같고	운동을(2)	이름이	그중	철수는(3)	잘하지만(2)	철수는	한

(a) InOrder 인덱스

(a) Inorder Index

철수는(3)	운동을(2)	잘하지만(2)	공부를	다른	못합니다	같고	이름이	그 중	철수와	한

(b) 빈도순 인덱스

(b) Freguency order index

[그림 2] 빈도순 인덱스

[Fig. 2] Freguency Index

결국 [그림2](b)와 같이 빈도순 인덱스 구조로 변환하여 가중치를 부여할 색인어절 구간 벡터를 추출한다.

현실적으로 문서 내에서 출현 빈도가 높은 어절은 동일 주제하의 문서들에서 표절 판별을 위한 분류의 색인어절로써는 부적합하다. 이는 3.1 절에서 기술한 바와 같이 표절된 문서의 유형 대부분과 동일한 주제 하에 작성된 문서들에서는 오히려 색인어 범주에 드는 단어들이 이미 의미가 없게 되고, 때문에 표절 판별을 위한 가중치로써는 0의 값을 부여하여 판별 벡터 구간에서 제외시킨다. 또한 출현 빈도가 낮은 어절도 같은 개념으로 표절 정도의 측정에는 관계될 수 있다고 할 수 있으나 전체적인 문서의 주제의 표절 판별에는 관련성이 미미하다고 볼 수 있기 때문에 제외시킨다.

출현빈도에 따른 빈도순 인덱스 리스트에서 상위 구간과 하위구간을 절삭한 구간의 선정 기준은 문서의 종류, 특징, 분류 의도, 분류의 소요시간 등에 따른 통계적 오차의 허용범위와 같은 의미를 지니므로 문서의 표절 판별 검사를 수행하는 시점에서 정할 수 있도록 하면 되겠다.

결국 문서내의 주요 색인어절들과 출현 빈도가 매우 낮은 어절들을 제외한 후 각 문서들에 대해 구간 벡터를 구하고 다음 식을 이용하여[16] 벡터 유사도 검사를 한다. 결과는 유사도의 정도에 따라 0-1 사이의 값을 얻게 된다. 값이 0 에 가까울수록 문서간 유사도가 높다.

$$f = 1 - \frac{2|X \cap Y|}{|X| + |Y|} = \frac{|X \Delta Y|}{|X| + |Y|} \qquad \text{(식3)}$$

3.5 유사 표절 문서들의 group 도출방법

유사도 검사를 수행하여 나온 결과 값으로 Similarity matrix 값을 구성한 다음 도식화된 도출결과를 표현하는데 필요한 Binary matrix를 구성하기 위해 문서들 간의 Dissimilarity matrix를 생성하고 유사도의 정도에 따른 다양한 산출을 위해서 다양한 threshold을 주어 그에 따른 Binary matrix를 얻는데 이는 유사한 문서들에 대해서 그룹화 할 수 있는 정보를 가지고 있다.

다음에 보이는 표절 문서들의 groupping를 위해서 백분율에 따라 threshold 값이 지정된다. threshold 값이 증가할수록 유사한 문서들을 더 정확히 분류할 수 있으며 유사도가 더 낮은 문서들을 찾아내게 된다.

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

[그림 3] Dissimilarity matrix

[Fig. 3] Dissimilarity matrix

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

Threshold = .1

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

Threshold = .2

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

Threshold = .

[그림 4] Binary matrix

[Fig. 4] Binary matrix

4. 실험 결과

아래 테이블들은 실제로 대학의 강의 수강생 25명의 1회분 연습문제 풀이 리포트 파일, A4 4-7 페이지 정도에 대해서 실험한 결과치를 도식화한 것이다. 리포트 문서들간의 유사도가 매우 높은 결과를 보이며 이는 표절의 정도가 높다는 것을 의미한다. <표 1>의 값들은 각 문서에 대해서 표절 가능성이 70% 이상인 유사한 다른 문서들과 그 유사도 값들을 보여주고 있는데 25개의 리포트 중에서 15개의 리포트가 최소 1에서 5개의 리포트와 유사함을 보여주고 있다.

다음 <표 2>는 표절 가능성이 75% 이상인 결과이며 이는 25개의 리포트 중에서 9개의 리포트가 최소 1에서 2개의 리포트와 유사함을 보여주고 있고, <표 3>은 표절 가능성이 80% 이상인 결과이며 이는 25개의 리포트 중에서 7개의 리포트가 최소 1에서 2개의 리포트와 유사함을 보여주고 있다. 표절 가능성 즉, 유사도를 높게 측정할 경우 점차적으로 문서들간의 classification 지어진 수가 적게 나타나는 것을 알 수 있다.

<표 1> 유사도 70% 이상의 분류

<Table 1> classification of similarity 70% upper

문서	유사 문서수	표절 문서와 유사도
A	5	B:89.9 C:81.2 F:74.2 H:73.4 D:72.5
B	4	A:89.9 C:75.8 D:70.6 H:70.2
C	4	A:81.2 B:75.8 D:73.2 F:71.0
D	3	C:73.2 A:72.5 B:70.6
E	2	I:94.7 G:70.4
F	2	A:74.2 C:71.0
G	2	I:72.7 E:70.4
H	2	A:73.4 B:70.2
I	2	E:94.7 G:72.7
J	1	K:77.2
K	1	J:77.2
L	1	O:85.0
M	1	N:72.1
N	1	M:72.1
O	1	L:85.0

<표 2> 유사도 75% 이상의 분류

<table 2> classification of similarity 75% upper

문서	유사 문서수	표절 문서와 유사도
A	2	B:89.9 C:81.2
B	2	A:89.9 C:75.8
C	2	A:81.2 B:75.8
D	1	F:77.2
E	1	F:94.7
F	1	A:77.2
G	1	I:85.0
H	1	E:94.7
I	1	G:85.0

<표 3> 유사도 80% 이상의 분류

<Table 3> classification of similarity 80% upper

문서	유사 문서수	표절 문서와 유사도
A	2	C:89.9 E:81.2
B	1	F:94.7
C	1	A:89.9
D	1	G:85.0
E	1	A:81.2
F	1	B:94.7
G	1	D:85.0

<표 4> 측정 유사도에 따른 class 수

<table 4> class counter of similarity

측정 유사도	문서 class수	class별 문서수
70%	15	1~5
75%	9	1~2
80%	7	1~2

5. 결론 및 향후과제

본 논문은 문서의 분류 방법들 중 하나에 속할 수 있는 복제, 표절의 방지를 위해서 문서의 유사도 측정을 효과적으로 하기 위한 데이터의 구조와 유사 문서들 간의 classification 판별 방법에 관하여 살펴보았다. 어절 트리 구조를 기반으로 표절 유사도를 판별하는 방법에 대해 설계 및 구현하여 그 결과가 효과적임을 보였다. 단, 이것은 제한적으로 문서의 복제나 표절의 판별에 이용될 수 있으며 양질의 문서를 판별하는데 부분적으로 기여할 뿐 이 결과만으로 양질의 문서만을 판별하는 방법에까지 적용되기에는 무리가 있다. 여기서의 목적은 학생들의 과제물 표절 및 복제를 예방하고 합당하게 측정하는데 적용될 수 있을 것이다.

온라인으로 제출된 학생들의 과제물의 경우, 그 특성이 동일한 주제에 관해 작성된 내용이 대부분일 것이라 본다면 어떤 과제물이 다른 과제물들과 차별적이고 복제되지 않았다고 하더라도 다른 과제물에 비해 너무 동떨어진 내용일 수도 있을 것이다. 이는 양질의 과제물이 아닐 수 있는데 현재의 방법은 다른 과제물들에 대해 너무 차별적으로 유사도가 작은 문서에 대해서는 별도의 수동적인 확인과정이 필요하다. 향후에는 이러한 부분을 판별할 수 있는 방법에 대해서 보완이 필요하다고 생각된다.

※ 참고문헌

[1] Fuhr. N. "Models for retrieval with probablistic indexing," information Processing and Management, 25(1), 1989.

[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] Larkey. L. and W. Croft, "Combling classifiers in text categorization", SIGIR'96, 1996.

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

[5] Masand. B. "Classifying News Stories using Memory Based Reasoning", SIGIR'92

[6] M E.. Maron, "Automatic Indexing: An Experimental Inquiry," Journal of the ACM, 8:404-417, 1961.

[7] P. Hayes and S. Weinstein, "CONSTRUE/TIS: A system for content-based indexing of a database of news stories", Second Annual Conference on Innovative Applecations of Artficial Intelligence, 1990.

[8] Hoch. R. "Using IR techniques for text classification in document analysis", SIGIR'94, 1994.

[9] Jacobs. P., "Using statistical methods to improve knowledge-based news catagoriztion", IEEE Expert, April, 1993.

[10] 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.

[11] 권오욱, 확률벡터와 메타범주를 이용한 최적 문서 범주화 모델, 석사학위 논문, 한국 과학기술원 전산학과, 1995.

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

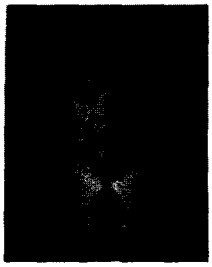
[13] 김준태, "지식기반 자연어처리를 이용한 문서의 자동분류와 지능형 색인에 관한 연구", 한국과학재단 연구결과보고서, 1998.4

[14] 강원석, 강현규, 김영섬, "개념기반 문서분류기 TAXON의 설계 및 구현", 한국정보과학회 가을 학술발표논문집, pp197-200, 1997.

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

[16] 조광제, 김준태, "역커테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류", 한국정보과학회 '97 봄 학술발표논문집 24권 1호, pp.507-510

천 승 환

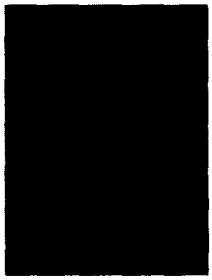


1993년 전남대 공대
화학공학과 졸업(학사)
1995년 전남대 대학원
전산통계학과 석사
1999년 전남대 대학원
전산통계학과 박사수료
1996년 ~현재
(주)도울정보기술 대표이사
관심분야 : 영상처리, 정보검색, 자연어처리, 인터넷 settop box
E-Mail :
shcheon@gyosuclub.com

이 귀 상

1980년 서울대 공대
전기공학과 졸업(학사)
1982년 서울대 대학원
전자계산기공학과 석사
1982년 금성통신 연구소
1991년 Pennsyivania 주립대학
전산학과 박사
1984년 ~현재 전남대
전산학과 교수
관심분야 : 멀티미디어통신, 영상처리 및 복원, 논리합성, VLSI/CAD
E-Mail : gslee@chonnam.ac.kr

김 미 영



1983년 전남대 계산통계학과
졸업(학사)
1985년 이화여대 대학원
전산통계학과 석사
1995년 전남대 대학원
전산통계학과 박사
1998년 ~현재 담양대학
인터넷 IT 공학부 부교수
관심분야 : VLSI/CAD, 정보검색, 자연어처리
E-Mail :
kimmee@
damyang.damyang.ac.kr