https://doi.org/10.5392/JKCA.2017.17.11.600

빅 데이터 환경에서 계층적 문서 유형 분류를 위한 클러스터링 기반 다중 SVM 모델

Multi-class Support Vector Machines Model Based Clustering for Hierarchical Document Categorization in Big Data Environment

김영수, 이병엽
배재대학교 사이버보안학과

Young Soo Kim(experkim@gmail.com), Byoung Yup Lee(bylee@pcu.ac.kr)

요약

최근 인터넷의 급격한 확장에 따른 정보의 양이 기하급수적으로 증가하고 있다. 그러나 실제 사용자에게 필요한 정보는 극히 일부분으로 사용자가 원하는 정보를 찾는데 까지는 부가적인 시간과 노력이 요구된다. 따라서 검색어로 검색된 문서에 대한 유사도 평가를 통한 계층적 유사 정보와 검색 우선순위에 대한 정보를 제공할 필요성이 있다. 이를 위해서 검색어를 구성하고 있는 키워드의 동시 발생 빈도를 고려한 검색 문서에 대한 유사도를 기반으로 문서 클러스터를 구성하고 SVM을 적용한 빅 데이터 기반 계층적 유형 분류 모델을 제안한다. 계층적 분류방법과 SVM 분류기의 결합은 문서의 계층이 기하급수적으로 늘어나는 웹 문서의 경우에 높은 성능을 얻을 수 있다. 제안된 모델은 정확하고 신속한 검색을 제공하는 정보검색시스템의 응용 모델로 활용될 수 있다.

■ 중심어 : | 빅 데이터 | 계층적 유형 | 문서 분류 | 클러스터링 | 다중 SVM | 유사도 |

Abstract

Recently data growth rates are growing exponentially according to the rapid expansion of internet. Since users need some of all the information, they carry a heavy workload for examination and discovery of the necessary contents. Therefore information retrieval must provide hierarchical class information and the priority of examination through the evaluation of similarity on query and documents. In this paper we propose an Multi-class support vector machines model based clustering for hierarchical document categorization that make semantic search possible considering the word co-occurrence measures. A combination of hierarchical document categorization and SVM classifier gives high performance for analytical classification of web documents that increase exponentially according to extension of document hierarchy. More information retrieval systems are expected to use our proposed model in their developments and can perform a accurate and rapid information retrieval service.

■ keyword : | Big Data | Hierarchical Class | Document Categorization | Clustering | Similarity |

* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1A2B1003678). 이 논문은 2017학년도 배재대학교 교내학술연구비 지원에 의하여 수행된 것임.

접수일자 : 2017년 08월 30일 심사완료일 : 2017년 09월 27일
수정일자 : 2017년 09월 27일 교신저자 : 이병엽, e-mail : bylee@pcu.ac.kr

Ⅰ.서 론

전형적인 정보 접근 모델은 색인어라고 불리는 키워드 집합을 사용하여 문서를 표현하고, 이들 색인어의 가중치를 상호 독립적이라고 가정하고 유사 문서를 분류하여 검색하고 있지만 실제 문서 내에서의 색인어 출현은 서로 무관하지 않기 때문에 이는 색인어들 간의 상대적인 중요도를 나타내지 못하고 유사도에 대한 정보를 제공하지 못한다. 그러나 최근 인터넷의 급격한 확장에 따른 정보의 양이 기하급수적으로 증가하고 있고 실제 사용자에게 필요한 정보는 극히 일부분으로 사용자가 원하는 정보를 찾는데 까지는 부가적인 시간과 노력을 더 요구한다. 따라서 검색어로 검색된 문서에 대한 유사도 평가를 통한 계층적 유사 정보와 검색 우선순위에 대한 정보를 제공할 필요성이 있다[1-5]. 이를 위하여 검색어로부터 분리된 각각의 키워드의 동시 발생 빈도를 고려하여 문서 클러스터를 구성하고 SVM을 적용하여 검색 문서를 계층적으로 분류하고 검색 우선 순위에 대한 정보를 제시한다. [그림 1]과 같은 연구 모델을 사용하여 빅 데이터 환경에서 클러스터링 모델 기반 SVM을 사용한 계층적 유형분류 모델을 제안한다.

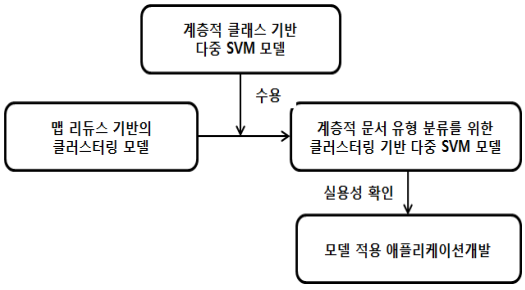

그림 1. 연구모델

본 논문은 다음과 같이 구성된다. 2절에서는 맵리듀스 기반의 클러스터링 모델을 분석하였고 3절에서는 계층적 클래스 기반 다중 SVM 분류 모델을 분석하였다. 4절에서는 빅 데이터 환경에서 계층적 문서 유형 분류를 위한 클러스터링 기반 선형 SVM 모델을 제안하였다. 5절에서는 결론과 시사점을 기술한다.

Ⅱ. 맵 리듀스 기반의 클러스터링 모델

2.1 맵 리듀스 모델

오늘날 빅 데이터를 활용하여 가치있는 정보를 실시간으로 추출하기 위해서 신속한 분석의 필요성이 요구되고 있다. 신속하게 데이터를 마이닝하고 고객에게 더 좋은 서비스를 제공하는 기업들은 효과적인 정책 수립과 경쟁우위 확보를 위해서 빅 데이터를 활용하고 있다. 빅 데이터의 신속한 분석기법 중의 하나는 구글에서 개발한 맵 리듀스로 분산 환경에서의 병렬 데이터 처리 기법이자 프로그래밍 모델이다. 구글의 맵 리듀스 기술을 구현한 프레임워크 중에서 아파치 하둡이 가장 주목을 받고 있다. 하둡의 주요 구성요소는 하둡 분산 파일시스템(HDFS)과 맵 리듀스(Map Reduce)이다[6][7]. HDFS는 메타 데이터를 네임노드라는 서버에 저장하고 데이터는 블록 단위로 분해하여 데이터 노드라는 서버에 저장하는데 사용되는 반면 맵 리듀스는 맵을 통해 키와 밸류라는 두개의 값을 쌍으로 가지고 있는 구조로 데이터를 관리하고 리듀스는 맵의 키를 기준으로 밸류를 더하거나 평균을 내는 등의 처리를 수행한다. 대규모 문서를 갖는 빅 데이터 환경에서 유사한 문서를 신속하고 발견하는 방법은 검색어와 문서간의 유사도 계산을 필요로 한다. 이를 위해서 [그림 2]의 (DocID(i), Term)과 같이 문서식별자와 질의어를 구성하는 키워드로 구성된 두개의 값을 쌍으로 가지는 형태로 맵을 구성하고 리듀스에서 유사도를 계산하여 문서식별자와 유사도로 구성된 두개의 값을 쌍으로 가지는 형태로 결과를 생성해서 유사문서를 제공한다.

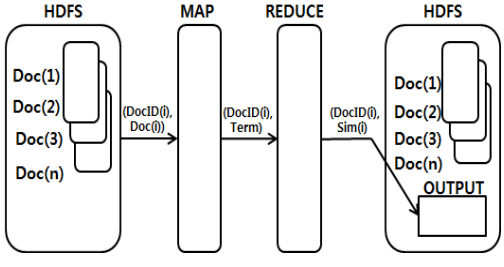

그림 2. 맵 리듀스 모델

2.2 클러스터링 모델

클러스터란 비슷한 특성을 가진 데이터들의 집단을 의미하고 클러스터 모델이란 데이터가 속해 있는 데이터 집단을 모르는 상태에서 유사한 혹은 동질의 데이터끼리 클러스터로 묶어 주는 비지도 학습 모델이다. 동질의 클러스터로 묶는 데이터 간의 유사성(similarity)은 거리(Distance)를 가지고 측정한다. 오늘날 블러그와 소셜 네트워크 그리고 웹 사이트에 의해서 폭발적으로 증가하는 온라인 문서를 효율적으로 관리하기 위해서는 자동화된 문서 범주화와 클러스터링 이 필요하다. 제안 모델에서 응용되는 클러스터링 모델은 [그림 3]과 같이 문서를 클러스터로 소속시켜서 분류하는 키워드의 빈도수를 기반으로 유사문서를 그룹으로 분류하는 K-Means 클러스터링 알고리즘이다. 검색어의 키워드 빈도수를 가중치로 사용하여 문서를 분류하는 기법으로서 주어진 키워드 집합에 따라 문서를 특정 카테고리로 분류하고 문서를 관련된 내용 별로 자동으로 구조화함으로써 사용자가 많은 양의 문서들을 좀 더 편리하게 접근할 수 있게 해준다[8-12]. 키워드 빈도수 기반 K-Means 클러스터링 모델은 검색어를 키워드로 분리한 후에 키워드의 개수만큼 클러스터 개수를 설정하고 문서를 스트링 배열로 분리한다. cluster[j]= argmax(freq(문서i,키워드j))은 문서에 포함된 키워드의 배열 첨자값을 배열로 저장한다는 의미이고 키워드의 배열 첨자값을 참조하여 클러스터로 묶인 문서를 출력한다.

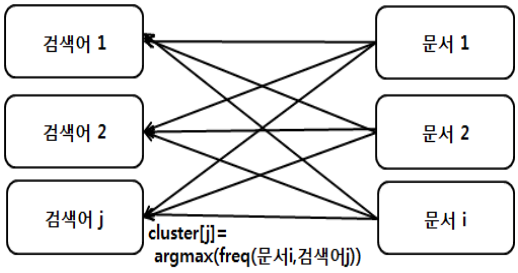

그림 3. K-means 클러스터링 모델

각 문서와 키워드의 비교를 통한 문서와 키워드간의 유사도를 측정하여 클러스터링을 수행하는 키워드 기반 K-means 클러스터링 모델에 대한 알고리즘은 알고리즘 1과 같다.

```
알고리즘 1.
클러스터 배열 초기화
for i=1 to n(문서의 개수)
for j=1 to n(검색어를 구성하는 키워드 개수)
  cluster[i][j]=0;
문서 클러스터 구성
for i=1 to n
cluster[i][j] =argmax(j = 1 to n) { freq(S[i], c[j]) }
```

Ⅲ. 계층적 클래스 기반 다중 SVM 분류 모델

3.1 범용적 SVM 모델

SVM(Support Vector Model)은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리로 분류할지 판단하는 분류 모델이다. 선형분류에 사용되는 SVM의 경우에는 지도학습 모델로서 주어진 이진 클래스의 데이터 집합으로부터 [그림 4]와 같이 선형 분류 모델 $X_1+2X_2-5.5=0$을 찾아서 이를 이용하여 분류를 수행한다. 삼각형과 원의 형태로 주어진 두 클래스에서 가장 가까운 좌표 (1,1)과 (2,3)이 서포트 벡터가 되고 이 두 서포트 벡터의 기울기가 1/2로 이를 벡터로 표현한 (1, 2)이 가중치 벡터가 된다. Decision Boundary는 support vector인 (1,1)과(2,3)의 중간지점(1.5,2)을 지나고 기울기가 1/2 이므로 1/2X1+X2+b=0 와 같은 방정식 모델로 표현된다. 따라서 x+2y+2b=0 식에 (1.5, 2)을 대입하면 1.5+4+2b=0 2b=-5.5 따라서 SVM 모델은 x+2y-5.5=0로 구축 된다. SVM은 주어진 데이터 점들이 두 개의 클래스 안에 각각 속해 있다고 가정했을 때, 새로운 데이터 점이 두 클래스 중 어느 곳에 속하는지 결정하는 것이 목표이다. 서포트 벡터는 선형 분류모델과 떨어진 거리가 가장 짧은 데이터 포인트가 바로 서포트 벡터이다[13][14]. 이 서포트 벡터를 이용 하면 새로운 데이터포인트를 분류할 때 전체 데이터포인트와의 거리를 계산하지 않고 서포트 벡터와 거리만 구하면

되므로 상당히 계산비용을 줄일 수 있다. 웹에서 수집한 평가의견이 긍정적인지 부정적인지를 제시해 주는 피드백 분석 시스템의 구현에 SVM 모델을 적용하여 평가의견을 분류하고 반영함으로써 사용자의 만족도를 개선하는데 활용된다[15][16].

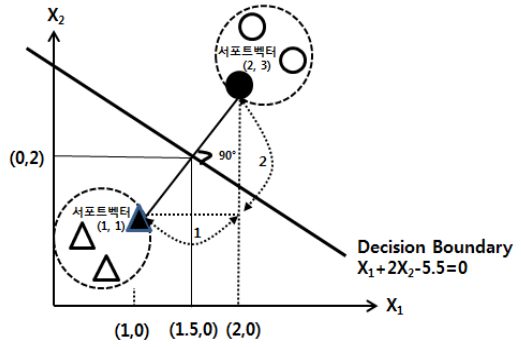

그림 4. 선형 SVM 모델

3.2 계층적 클래스 기반 다중 SVM 분류 모델

문서가 계층적 구조로 분류 될 때 단일의 문서는 하나 이상의 클래스에 속하게 되고, 상위 계층으로 올라갈수록 보다 폭넓은 범주에 속하게 된다. 일반적으로 문서 분류시 선형 구조의 클래스 분류보다 계층 구조로 클래스를 분류하는 것이 보다 좋은 성능을 보여주고 있다[17-20]. 계층적 분류 방법은 [그림 5]에서 알 수 있는 것처럼 루트 노드에서만 분류기가 존재하는 선형 분류 방법과는 달리 계층적 방법에서는 매 단계마다 분류기가 할당된다. 클래스간의 계층적 구조를 형성하기 위해서는 유사한 클래스간의 군집을 구성하는 클러스터링 모델을 활용하기도 한다. [그림 5](a)에서 보는 것처럼 다중 SVM 트리 분류모델은 각 노드의 SVM 분류기를 적용하기에 앞서 클러스터링 모델을 사용하여 상위 노드에서 분할된 데이터 그룹에 대해서 각각 2개의 서브 클러스터로 묶고 이를 긍정과 부정이라는 레이블을 할당하여 이진 SVM 분류기를 학습시키는 과정을 반복하여 [그림 5](b)와 같은 SVM 트리를 생성한다. [그림 5](c)는 계층적 트리에서 클래스 22222의 카테고리를 벡터의 형태로 표현하고 있다. 이들 벡터의 형태로 구축된 모델을 활용하여 검색어와 유사한 그룹에 소속된 문서를 신속하게 제공할 수 있게 된다. 루트 노드의 탑 레벨에서 수행되는 SVM1은 검색어에 대한 모든 문서의 유사도 1에서 5까지의 값을 {1,2,3}과 {4,5}의 2개의 서브 그룹으로 분류해서 학습을 하게 된다. 두 번째 레벨에서 수행되는 SVM2는 {1}과 {2,3}의 2개의 서브 그룹으로 분류해서 학습을 하게 된다. 하나의 클래스가 존재할 때 까지 이와 같은 과정을 반복 수행한다.

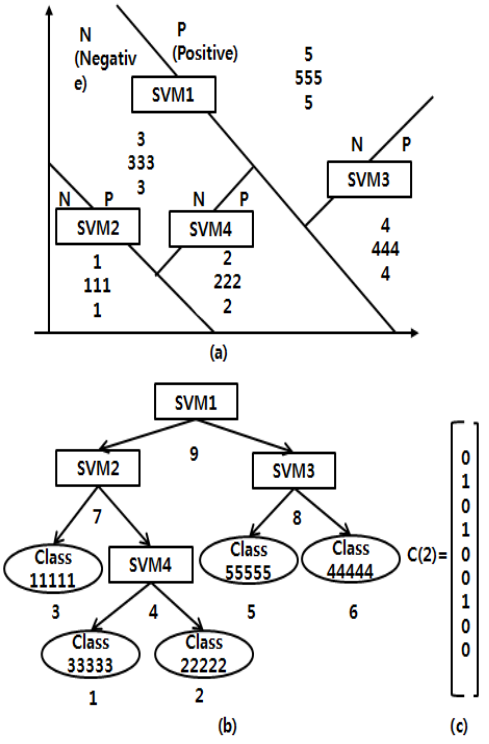

그림 5. 계층적 클래스 기반 다중 SVM 분류 모델

Ⅳ. 계층적 문서 유형 분류를 위한 클러스터링 기반 다중 SVM 모델

4.1 클러스터링 기반 다중 SVM 모델

문서의 검색 성능을 향상시키기 위해서 유사도 평가를 통한 문서 클러스터링과 계층적 분류와 같은 활동을 포함한다. K-means 클러스터링은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로 유클리드 거리를

계산하여 최소의 거리를 갖는 데이터를 클러스터로 묶고 클러스터의 중심점을 평균에 의해서 구한 다음에 이 중심점과 다른 데이터와의 유클리드를 계산하여 최소의 거리를 갖는 데이터와 클러스터를 구성한다. 이와 같은 과정을 반복하면 클러스터 테이블은 한 개의 클러스터로 구성되고 반복을 종료하게 된다. [표 1][표 2][표 3] 그리고 [표 4]는 3개의 점을 가지고 클러스터의 중심점을 구하여 클러스터와 거리행렬을 구성하는 과정을 보여주고 있다. 초기에는 하나의 점이 하나의 클러스터를 구성하므로 [표 1]과 같은 클러스터 테이블이 만들어진다.

표 1. 클러스터 테이블

Point(점)	X좌표	Y좌표
P1	1.5	1
P2	2	4
P3	4.5	3.5

클러스터간의 거리행렬은 [표 2]와 같이 유클리드 거리를 계산하여 구성한다. 두 점에 대한 유클리드거리는 $d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ 의 공식에 의하여 계산한다. P1과 P2의 거리는 9.25로 계산된다.

표 2. 거리 행렬

Point(점)	X좌표	Y좌표	P1	P2	P3
P1	1.5	1	0		
P2	2	4	9.25	0	
P3	4.5	3.5	12.25	6.5	0

위의 거리행렬에서 P2와 P3의 거리가 6.5로서 가장 가까우므로 즉 유사하므로 P21, P3)를 새로운 클러스터로 묶어준다. [표 3]과 같이 클러스터가 처음 하나의 점으로 구성된 3개의 클러스터에서 2개로 줄었다

새로 묶인 클러스터(P2, P3)의 중심(centroid)을 평균을 이용해서 구해보면 (P2,P3)/2=((2+4.5),(4+3.5))/2 =(6.5,7.5)/2=(6.5/2, 7.5/2)=(3.25, 3.75)가 된다. 따라서 새로운 클러스터 테이블2는 [표 3]과 같이 구성된다.

표 3. 클러스터 테이블

Point(점)	X좌표	Y좌표
P1	1.5	1
(P2,P3)	3.25	3.75

새로 구성된 위의 클러스터간의 유클리드 거리를 계산하여 구성된 거리행렬은 [표 4]와 같다.

표 4. 거리 행렬

Point(점)	X좌표	Y좌표	P1	(P2,P3)
P1	1.5	1	0	
(P2,P3)	3.25	3.75	9.25	0

거리행렬에서 P1과 (P2,P3)의 거리가 '9.25'로서 가장 가까우므로 즉 유사하므로 P1과 (P2,P3)를 새로운 클러스터로 묶어준다. 드디어 클러스터가 1개로 반복을 종료한다.

SVM은 기본적으로 선형 분류기이므로 가우시안 커널을 사용하는 범용적 다중 SVM 모델은 SVM 트리의 각 노드에 할당된 SVM 분류기는 두 개의 서브 클러스터를 결과로 제공한다. 따라서 n개의 유형 분류를 위해서는 n-1개의 분류기가 필요하다. 범용적 다중 SVM 분류기는 각 노드에서 마진이 가장 큰 두 개의 서브 클러스터를 확인하기 위해서 C^2_n의 비교 연산이 필요하다. 따라서 기존에 사용자가 질의한 다수 검색어에 대한 문서의 계층적 분류 모델을 통합해서 구축할 수 있도록 자식노드가 2로서 차수가 2인 이진 트리의 형태로 문서 유형을 계층적으로 분류하는 모델을 확장하여 자식노드가 n개로서 차수가 n인 일반 트리 구조로 문서 유형을 계층적으로 다중 분류함으로 동일한 검색어를 사용한 질의의 경우에 검색 속도를 개선하고 검색문서의 유사정보를 제공할 수 있는 모델을 제안한다.

[그림 6]과 같이 문서 유사도에 따른 데이터 셋에 대한 선형 SVM 모델식을 평가해서 서포트 벡터를 지나는 선형 모델식에 존재하는 클러스터를 제외한 클러스터는 최종 클래스로 분류한다. 이들 클러스터에 대해서 긍정과 부정에 대한 라벨을 부착한 후에 서포트 벡터를 지나는 선형 모델식에 존재하는 클러스터에 대해서 다

시 선형 SVM 모델식을 평가하고 서포트 벡터를 지나는 선형 모델식에 존재하는 클러스터를 제외한 클러스터를 최종 클래스로 분류한다. 이러한 평가 과정을 반복해서 계층적 문서 유형 분류를 위한 SVM 트리를 구성한다.

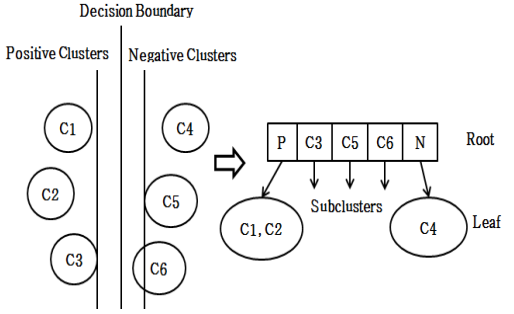

그림 6. 클러스터링 기반 다중 SVM 모델

기존의 SVM은 단일 분류기에 의해 이루어지는 이분류 기법으로 설계되었기 때문에 다분류 SVM을 구현하기 위한 다양한 접근법이 제안되었지만 크게 SVM을 복수개 사용하여 사후적으로 결합하여 예측을 수행하는 SVM 확장 접근법을 사용하거나 SVM을 다분류 모델에 적합하도록 변경하여 예측을 수행하는 SVM 변형 접근법을 사용한다. 제안 모델에서는 SVM 확장 접근법과 SVM 변형 접근법을 함께 사용하여 검색의 정확도를 높이고 검색에 소요되는 시간을 감소하여 처리 성능을 개선하였다. 가우시안 커널을 사용하는 범용적 다중 SVM은 식 (1)과 같이 학습용 입력벡터 X를 고차원의 특징공간으로 사상시킨 다음 두 집단 사이의 폭을 최대화시키는 분리 경계면을 찾는 모델이다. α_i는 각 학습표본 (x_i,y_i) 마다 주어지는 가중치 벡터이다.K(x(i), x)는 커널함수로 x라는 문서가 들어왔을 때 서포트 벡터 x(i)와의 거리를 계산하는 가우시안 커널로 식 (2)와 같이 커널의 폭을 제어하는 감마 ɣ와 유클리드거리인 $\| x(i)-x \|$를 사용한다.

$$y=b+\sum \alpha_i y_i K(x(i), x) \quad (1)$$

$$K(x(i), x)=\exp(-\gamma \| x(i)-x \|^2) \quad (2)$$

본 논문에서는 식 (3)과 같이 가우시안 커널 함수를 클러스터링 함수로 대치하여 지도학습 모델인 SVM모델을 비지도 학습 모델로 변형하여 라벨링을 자동화하여 분류를 수행하도록 하였다.

$$y=b+\sum \alpha_i y_i KMeans(x(i), x) \quad (3)$$

4.2 모델 검증

SVM 분류기와 클러스터 모델을 별도로 분리 구현하고 그 결과로서 리턴 되는 클러스터와 문서의 유사도를 SVM의 입력 데이터로 주어서 문서를 계층적으로 분류하도록 구현하였다. [그림 7]과 같이 검색어에 대한 질의 결과로 리턴되는 문서들에 대해서 키워드와 매칭되는 동시발생 빈도수를 고려한 유사도와 이와 대응되는 클러스터를 SVM의 입력으로 사용하여 SVM트리 형태로 모델을 구축하여 문서의 계층적 분류를 수행하였다.

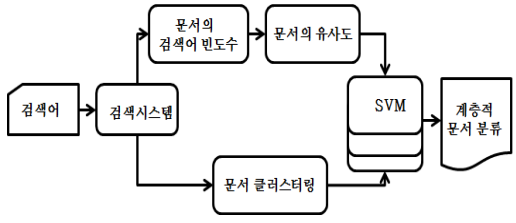

그림 7. 구현 모델

실용성 검증을 위하여 그림 8과 같이 가우시안 커널을 사용하는 SVM 모델과 K-Means clustering을 사용하는 SVM모델을 사용하여 실용성을 확인하였다

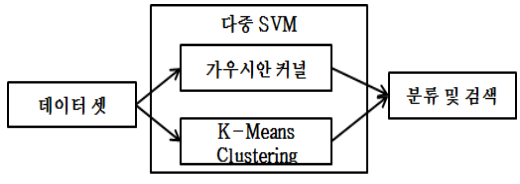

그림 8. 검증 모델

가우시안 커널을 사용하는 범용적 다중 SVM모델은 gamma와 cost에 대한 파라미터를 조정하여 분류 정확도와 검색 성능을 개선하는 반면에 K-Means 클러스터링을 사용하는 본 논문의 다중 SVM 모델은 클러스터의 개수와 초기 중심점의 설정이 분류 정확도와 검색

성능에 영향을 미치므로 이들 파라미터의 최적 조합을 찾아서 모델을 구축하여야 한다. 검색결과에 따른 문서를 [표 5]과 같이 구분하고 이를 이용한 검색효율의 검증을 위해서 식 (4)와 같은 재현율과 식 (5)와 같은 정확률을 사용하였다. 적합문서는 소속된 클러스터로 정확히 분류된 문서이고 부적합문서는 비소속 클러스터로 분류된 문서로 정의한다. 이때 재현율은 단일 클래스에 소속된 전체문서에서 실제 해당 클러스터로 분류된 검색문서의 개수로 적합문서를 검색하는 능력을 표시하며 정확률은 단일 클래스로 분류된 전체 검색문서에서 실제 해당 클러스터로 분류된 검색 문서의 개수로 검색된 문헌들의 적합도를 표시한다.

표 5. 검색결과에 따른 문서의 구분

적합문서	부적합문서	구분
검색된 적합문서(A)	검색된 부적합문서(B)	검색된 문서
미검색된 적합문서(C)	미검색된 부적합문서(D)	미검색된 문서

재현율=(A/A+C)*100 (4)

정확률=(A/A+B)*100 (5)

[그림 9]와 [그림 10]는 가우시안 커널을 사용하는 다중 SVM 모델과 클러스터링 기반 다중 SVM 모델의 분류의 정확도와 분류 및 검색에 소요되는 시간을 보여주고 있다. 본 논문에서 제안한 클러스터링 기반 다중 SVM 모델이 가우시안 커널을 사용하는 다중 SVM 분류 모델에 비해서 검색어를 통한 문서 검색의 횟수가 증가할수록 검색 속도가 우수한 반면 분류의 정확도는 떨어진다는 것을 확인할 수 있다.

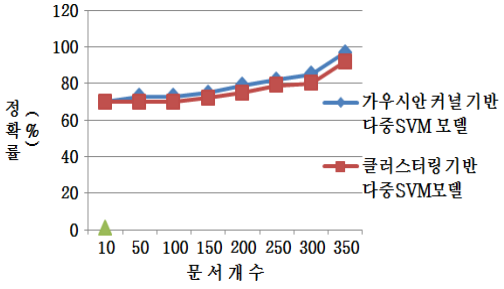

그림 9. 분류의 정확도

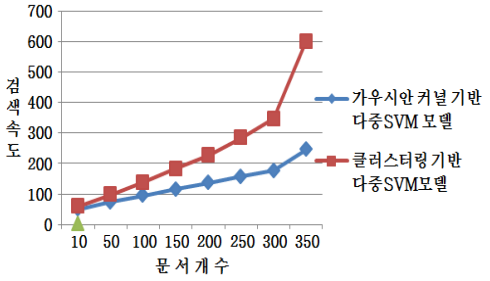

그림 10. 분류 및 검색 속도

V. 결 론

최근 인터넷의 급격한 확장에 따른 정보의 양이 기하급수적으로 증가하고 있다. 그러나 실제 사용자에게 필요한 정보는 극히 일부분으로 사용자가 원하는 정보를 찾는데 까지는 부가적인 시간과 노력이 요구된다. 따라서 검색어로 검색된 문서에 대한 유사도 평가를 통한 계층적 유사 정보와 검색 우선순위에 대한 정보를 제공할 필요성이 있다. 이의 해결책으로 검색어를 구성하고 있는 키워드의 동시 발생 빈도를 고려한 검색 문서에 대한 유사도를 기반으로 문서 클러스터를 구성하고 SVM을 적용한 빅 데이터 기반 계층적 유형 분류 모델을 제안한다. 계층적 분류방법과 SVM 분류기의 결합은 문서의 계층이 기하급수적으로 늘어나는 웹 문서의 경우에 높은 성능을 얻을 수 있다. 기존의 SVM은 단일 분류기에 의해 이루어지는 이분류 기법으로 설계되었기 때문에 다분류 SVM을 구현하기 위해서 다양한 접근법이 제안되었지만 일반적으로 SVM을 복수개 사용하여 사후적으로 결합하여 예측을 수행하는 SVM 확장 접근법을 사용하거나 SVM을 다분류 모델에 적합하도록 변경하여 예측을 수행하는 SVM 변형 접근법을 사용한다. 본 논문에서는 SVM 확장 접근법과 SVM 변형 접근법을 함께 사용하여 검색의 정확도를 높이고 검색에 소요되는 시간을 감소시켜 처리 성능을 개선하였다.

본 연구의 한계점으로는 실용성 검증에서 제안모델이 기존 모델과 비교하여 문서의 재현율과 정확률은 다소 떨어지는 반면 검색처리 성능은 더 좋은 결과를 보이지만 실용성 검증을 위해서 대화형 웹문서 대신에 일

괄처리 문서로 제한하고 있는 한계가 있다. 따라서 향후 대화형 웹 문서의 검색 결과에 대한 재현율과 정확율의 개선에 대한 연구를 지속할 예정이다.

제안 모델은 문서를 주제별로 계층을 만들어 디렉토리 형태로 분류하여 사용자들이 쉽게 찾을 수 있는 정보검색 서비스가 요구되는 웹 포탈 사이트와 정확하고 신속한 검색이 필요한 정보검색시스템의 응용 모델로 활용될 수 있다.

참 고 문 헌

[1] 김현주, 박소미, 박석, “확장된 질의 처리를 위해 경로간 의미적 유사도를 고려한 XML 문서 순위화 기법,” 한국정보과학회 학술발표논문집, Vol.36(1A), pp.8-13, 2009.

[2] 윤용욱, 이창기, 이근배, “지지 벡터 기계를 이용한 계층적 문서 분류,” 한국정보과학회 언어공학연구회 학술발표 논문집, pp.7-13, 10. 2003.

[3] 유재학, 김성윤, 이한성, 김명섭, 박대희, “계층적 다중 클래스 SVM을 이용한 인터넷 애플리케이션 트래픽 분류,” 한국컴퓨터종합학술대회 논문집, Vol.35, No.1(A), 2008.

[4] 김영수, 문형진, 조혜선, 김병익, 이진해, 이진우, 이병엽, “계층적침해자원기반의 침해사고 구성 및 유형 분석,” 한국콘텐츠학회논문지, 제16권, 제11호, pp.139-153, 2016.

[5] 김영수, “보안 인텔리전트 유형 분류를 위한 다중 프로파일링 앙상블 모델,” 한국콘텐츠학회논문지, Vol.17, No.3, pp.231-237, 2017(3).

[6] C. Chu, S. K. Kim, Y. Lin, Y. Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun, “Map-Reduce for Machine Learning on Multicore,” pp.281-288, NIPS 2006.

[7] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” OSDI'04: Sixth Symposium on Operating System Design and Implementation, Dec 2004.

[8] C. W. Hsu and C. J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, Vol.13, pp.415-425, 2002.

[9] S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases, In Proc. ACM SIGMOD Int. Conf. on Management of Data, Seatle, WA, 1998.

[10] Y. Kanza and Y. Sagiv, ″Flexible Queries over Semistructured Data,″ Proc. of 12th ACM SIGMODSIDACT-SIGART symposium on Principles of database systems, pp.40-51, 2001.

[11] K. N. Rao, T. V. Rao, and D. R. Lakshmi, “A Novel Class Imbalance Learning Method using Subset Filtering,” International Journal of Scientific and Engineering Research, Vol.3, No.9, pp.1-9, 2012.

[12] Xiping Liu, Changxuan Wan, and Lei Chen, “Returning Clustered Results for Keyword Search on XML Documents,” IEEE Transac-tions On Knowledge and Data Engineering, Vol.23, No.12, Dec. 2011.

[13] C. Cortes and V. Vapnik, “Support-vector network,” Machine Learning, Vol.20, pp.273-297, 1995.

[14] C. F. Tan, “Short Text Classification Based on LDA and SVM,” International Journal of Applied Mathematics and Statistics (IJAMS), Vol.51, No.22, pp.205-214, 2013.

[15] Lijuan Cai and Thomas Hofmann, “Implementation of Support Vector Machine Technique in Feedback Analysis System,” International Journal of Computer Applications, Vol.96, No.17, pp.24-27, Jun. 2014.

[16] J. Hernandez, L. E. Sucar, E. F. Morales, “Multidimensional hierarchical classification,” Expert Systems with Applications, Vol.41, No.17, pp.7671-7677, 2014.

[17] C. Freeman, D. Kulic, and O. Basir, "Joint feature selection and hierarchical classifier design," IEEE International Conference on Systems Man and Cybernetic, pp.1728-1734, 2011.

[18] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.170-178, 1997.

[19] Tao Li, Shenghuo Zho, and Mitsunori Orkhara, "Topic Hierarchy Generation via Linear Discriminant Projection," Proceedings of SIGIR 2003, the Twenty-Sixth Annual International ACM SIGIR Conference, pp.421-422, 2003.

[20] H. Yu, J. Han, and K. C. Chang, PEBL: Positive-example based learning for Web page classification using SVM. In Proc. 8th Int. Conf. Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.

저 자 소 개

김 영 수(Young Soo Kim) 정회원

- 2003년 8월 : 국민대학교정보관리학(정보관리학박사)
- 현재 : 충남 재활IT 융합 기술원 대표 컨설턴트
- 현재 : 배재대학교 사이버보안학과

<관심분야> : 빅 데이터서비스 보안, 정보 보안

이 병 엽(Byoung Yup Lee) 종신회원

- 1991년 2월 : 한국과학기술원 전산학과(공학사)
- 1993년 2월 : 한국과학기술원 전산학과(공학석사)
- 1997년 2월 : 한국과학기술원 경영정보공학(공학박사)
- 1993년 1월 ~ 2003년 2월 : 대우정보시스템 차장
- 2003년 3월 ~ 현재 : 배재대학교 사이버보안학과 교수

<관심분야> : XML, 지능정보시스템, 데이터베이스 시스템, 전자상거래학