

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2019-0135129
(43) 공개일자 2019년12월06일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01) *G06F 17/27* (2006.01)
G06N 3/08 (2006.01)
 (52) CPC특허분류
G06F 16/3347 (2019.01)
G06F 16/93 (2019.01)
 (21) 출원번호 10-2018-0060185
 (22) 출원일자 2018년05월28일
 심사청구일자 2018년05월28일
 기술이전 희망 : 기술양도

(71) 출원인
 인천대학교 산학협력단
 인천광역시 연수구 아카데미로 119 (송도동)
 (72) 발명자
 전광길
 인천광역시 연수구 경원대로119번길 21, 114동
 904호(동춘동, 연수2차풍림아파트)
 (74) 대리인
 특허법인충정

전체 청구항 수 : 총 7 항

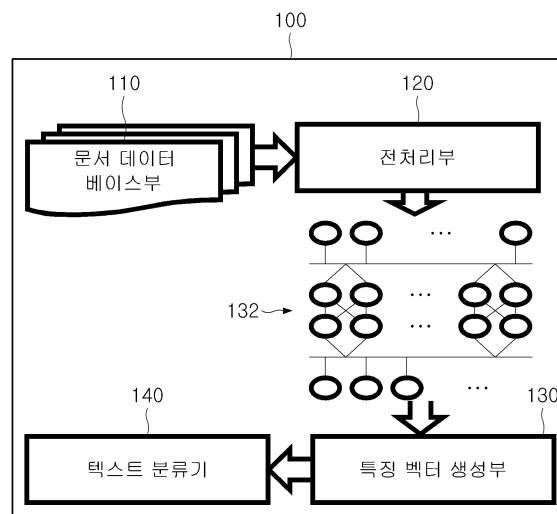
(54) 발명의 명칭 문서 구조와 딥러닝을 이용한 문서 분류 장치 및 방법

(57) 요약

문서 구조와 딥러닝을 이용한 문서 분류 장치 및 방법은 문서 분류 시 문서에 부여되는 가중치 적용 방식과, 이에 대응하는 복수의 분류기 중 분류 성능이 우수한 분류기와 가중치 적용 방식을 선택하여 텍스트를 분류한다.

문서 분류 시 문서에 부여되는 가중치 부여 방식과 이에 대응하는 지도 학습 기술을 F 측정값을 비교하고 가장 높은 측정값의 분류기를 선택하여 우수한 성능의 분류기를 적용할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06F 17/2755 (2013.01)

G06N 3/08 (2013.01)

명세서

청구범위

청구항 1

적어도 하나의 문서를 포함하는 문서 데이터베이스부;

상기 문서 데이터베이스부로부터 수신한 문서를 문자열로 변환하고, 전치사, 접속사, 대명사 등의 정보가 없는 단어를 제거하며, 동일한 개념적 의미를 가진 단어 그룹을 만드는 단어의 형태소 분석을 수행하는 전처리부;

벡터 공간 모델을 이용한 인덱싱 처리를 수행하여 문서에 포함된 단어에 가중치 부여 방식에 의해 가중치를 부여하여 특징 벡터들의 특징 세트를 구성하는 특징 벡터 생성부; 및

상기 특징 벡터 생성부로부터 구성된 특징 벡터들의 특징 세트를 수신하고, 상기 수신한 특징 세트를 결정 트리(Decision Tree), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine) 중 하나의 지도 학습 기술을 이용하여 텍스트를 분류하는 텍스트 분류기를 포함하며,

상기 가중치 부여 방식은 Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy 중 하나의 방식을 적용하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 장치.

청구항 2

제1항에 있어서,

상기 텍스트 분류기는 상기 결정 트리, 상기 나이브 베이즈, 상기 서포트 벡터 머신의 각각에 대하여 상기 5개의 가중치 부여 방식마다 공지된 F-Measure 방법에 적용하여 F 측정값을 각각 계산하고, 상기 계산된 F 측정값이 가장 높은 가중치 부여 방식을 상기 특징 벡터 생성부의 가중치 부여 방식으로 선택하고, 상기 가장 높은 가중치 부여 방식에 대응하는 지도 학습 기술이 텍스트 분류 기술로 적용되는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 장치.

청구항 3

제1항에 있어서,

상기 특징 벡터 생성부는 상기 전처리부로부터 전처리된 문서를 입력받아 문서를 적절한 클래스로 분류하기 위하여 Deep Belief Network(DBN) 알고리즘을 이용하여 학습 데이터와 특징 벡터들의 특징 세트를 구성하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 장치.

청구항 4

제2항에 있어서,

상기 F-Measure 방법은 하기의 수학적식 1, 수학적식 2, 수학적식 3에 의해 F 측정값을 계산하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 장치.

[수학적식 1]

$$P = \frac{\sum_{c=1}^N TP_c}{\sum_{c=1}^N (TP_c + FP_c)}, \quad \text{Micro-Averaging}$$

[수학식 2]

$$R = \frac{\sum_{c=1}^N TP_c}{\sum_{c=1}^N (TP_c + FN_c)}, \quad \text{Micro-Averaging}$$

[수학식 3]

$$F = \frac{2 \times P \times R}{(P + R)}.$$

여기서, N은 학습 기술의 성능을 평가하기로 결정한 문서 클래스의 수이고, TP는 각 문서 클래스에 대해 정확하게 식별되는 문서 개수이고, FP는 문서 개수가 각 문서 클래스에 대해 잘못 식별된 문서들의 개수이고, FN은 각각의 해당 문서 클래스에서 누락된 문서의 개수임.

청구항 5

문서 데이터베이스로부터 적어도 하나의 문서를 수신하는 단계;

상기 문서 데이터베이스로부터 수신한 문서를 문자열로 변환하고, 전치사, 접속사, 대명사 등의 정보가 없는 단어를 제거하며, 동일한 개념적 의미를 가진 단어 그룹을 만드는 단어의 형태소 분석을 수행하는 전처리 과정을 수행하는 단계;

벡터 공간 모델을 이용한 인덱싱 처리를 수행하여 문서에 포함된 단어에 가중치 부여 방식에 의해 가중치를 부여하여 특징 벡터들의 특징 세트를 구성하는 단계;

상기 구성된 특징 벡터들의 특징 세트를 수신하고, 상기 수신한 특징 세트를 결정 트리(Decision Tree), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine) 중 하나의 지도 학습 기술을 이용하여 텍스트를 분류하는 단계를 포함하며,

상기 가중치 부여 방식은 Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy 중 하나의 방식을 적용하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 방법.

청구항 6

제5항에 있어서,

상기 결정 트리, 상기 나이브 베이즈, 상기 서포트 벡터 머신의 각각에 대하여 상기 5개의 가중치 부여 방식마다 공지된 F-Measure 방법에 적용하여 F 측정값을 각각 계산하는 단계; 및

상기 계산된 F 측정값이 가장 높은 가중치 부여 방식과 이에 대응하는 지도 학습 기술이 텍스트 분류 기술로 적용하는 단계를 포함하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 방법.

청구항 7

제5항에 있어서,

상기 특징 벡터들의 특징 세트를 구성하는 단계는,

상기 전처리된 문서를 입력받아 문서를 적절한 클래스로 분류하기 위하여 Deep Belief Network(DBN) 알고리즘을 이용하여 학습 데이터와 특징 벡터들의 특징 세트를 구성하는 단계를 더 포함하는 것을 특징으로 하는 문서 구조와 딥러닝을 이용한 문서 분류 방법.

발명의 설명

기술 분야

본 발명은 문서 분류 장치에 관한 것으로서, 특히 문서 분류 시 문서에 부여되는 가중치 적용 방식과, 이에 대

[0001]

응하는 복수의 분류기 중 분류 성능이 우수한 분류기와 가중치 적용 방식을 선택하여 텍스트를 분류하는 문서 구조와 딥러닝을 이용한 문서 분류 장치 및 방법에 관한 것이다.

배경 기술

- [0002] 최근 빅 데이터와 데이터를 분류 및 예측하는 기계 학습 분야의 연구가 활발히 진행되고 있다. 특히, 패턴 분류 기술은 뉴스, 블로그, SNS 등의 문서 분류, 이슈 분류, 감성 분류에 활용되고 있다.
- [0003] 기계 학습 분야에서 간단한 이진 분류기들은 데이터가 오직 두 개의 클래스로 분류되므로 복잡한 실제 데이터를 표현하기에 효과적이지 않으므로 데이터를 다중 클래스 중 하나로 분류하는 다중 분류기(Multiclass Classifier)에 대한 연구가 활발히 진행되고 있다.
- [0004] 문서 분류기는 유용한 정보라고 판단되는 여러 특징과 해당 특징이 어떤 의미를 가지는가에 대한 의견을 나타내는 어휘 정보를 추출하는 특징 추출 단계를 포함한다. 일반적으로 문서 분류기는 모든 클래스의 중요도를 동일하게 보는데, 여러 분류기를 조합하거나 카테고리별로 다른 언어 모델과 분류기를 사용하여야 하기 때문에 그 구현 알고리즘이 너무 복잡한 문제가 있으며, 그에 따라 현실적으로 구현이 어려운 단점이 있다.

선행기술문헌

특허문헌

- [0005] (특허문헌 0001) 한국 등록특허번호 제10-1178068호

발명의 내용

해결하려는 과제

- [0006] 이와 같은 문제점을 해결하기 위하여, 본 발명은 문서 분류 시 문서에 부여되는 가중치 적용 방식과, 이에 대응하는 복수의 분류기 중 분류 성능이 우수한 분류기와 가중치 적용 방식을 선택하여 텍스트를 분류하는 문서 구조와 딥러닝을 이용한 문서 분류 장치 및 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

- [0007] 상기 목적을 달성하기 위한 본 발명의 특징에 따른 문서 구조와 딥러닝을 이용한 문서 분류 장치는,
- [0008] 적어도 하나의 문서를 포함하는 문서 데이터베이스부;
- [0009] 상기 문서 데이터베이스부로부터 수신한 문서를 문자열로 변환하고, 전치사, 접속사, 대명사 등의 정보가 없는 단어를 제거하며, 동일한 개념적 의미를 가진 단어 그룹을 만드는 단어의 형태소 분석을 수행하는 전처리부;
- [0010] 벡터 공간 모델을 이용한 인덱싱 처리를 수행하여 문서에 포함된 단어에 가중치 부여 방식에 의해 가중치를 부여하여 특징 벡터들의 특징 세트를 구성하는 특징 벡터 생성부; 및
- [0011] 상기 특징 벡터 생성부로부터 구성된 특징 벡터들의 특징 세트를 수신하고, 상기 수신한 특징 세트를 결정 트리(Decision Tree), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine) 중 하나의 지도 학습 기술을 이용하여 텍스트를 분류하는 텍스트 분류기를 포함하며,
- [0012] 상기 가중치 부여 방식은 Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy 중 하나의 방식을 적용하는 것을 특징으로 한다.
- [0014] 본 발명의 특징에 따른 문서 구조와 딥러닝을 이용한 문서 분류 방법은,
- [0015] 문서 데이터베이스부로부터 적어도 하나의 문서를 수신하는 단계;
- [0016] 상기 문서 데이터베이스부로부터 수신한 문서를 문자열로 변환하고, 전치사, 접속사, 대명사 등의 정보가 없는 단어를 제거하며, 동일한 개념적 의미를 가진 단어 그룹을 만드는 단어의 형태소 분석을 수행하는 전처리 과정을 수행하는 단계;

- [0017] 벡터 공간 모델을 이용한 인덱싱 처리를 수행하여 문서에 포함된 단어에 가중치 부여 방식에 의해 가중치를 부여하여 특징 벡터들의 특징 세트를 구성하는 단계;
- [0018] 상기 구성된 특징 벡터들의 특징 세트를 수신하고, 상기 수신한 특징 세트를 결정 트리(Decision Tree), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine) 중 하나의 지도 학습 기술을 이용하여 텍스트를 분류하는 단계를 포함하며,
- [0019] 상기 가중치 부여 방식은 Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy 중 하나의 방식을 적용하는 것을 특징으로 한다.

발명의 효과

- [0020] 전술한 구성에 의하여, 본 발명은 특정한 가중치 부여 방식과 지도 학습 기술을 통해 수집된 문서의 분류 성능을 향상시키는 효과가 있다.
- [0021] 본 발명은 문서 분류 시 문서에 부여되는 가중치 부여 방식과 이에 대응하는 지도 학습 기술을 F 측정값을 비교하고 가장 높은 측정값의 분류기를 선택하여 우수한 성능의 분류기를 적용할 수 있다.

도면의 간단한 설명

- [0022] 도 1은 본 발명의 실시예에 따른 문서 분류 시스템의 구성을 나타낸 도면이다.
- 도 2는 본 발명의 실시예에 따른 문서 구조와 딥러닝을 이용한 문서 분류 장치의 구성을 간략하게 나타낸 도면이다.
- 도 3 및 도 4는 본 발명의 제1 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이다.
- 도 5 및 도 6은 본 발명의 제2 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이다.
- 도 7 및 도 8은 본 발명의 제3 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0023] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.
- [0024] 도 1은 본 발명의 실시예에 따른 문서 분류 시스템의 구성을 나타낸 도면이며, 도 2는 본 발명의 실시예에 따른 문서 구조와 딥러닝을 이용한 문서 분류 장치의 구성을 간략하게 나타낸 도면이다.
- [0025] 본 발명의 실시예에 따른 문서 분류 시스템은 복수의 전자기기(20, 30, 40, 50) 및 문서 분류 장치(100)를 포함한다.
- [0026] 복수의 전자기기(20, 30, 40, 50)는 고정형 단말이거나 이동형 단말일 수 있다. 복수의 전자기기(20, 30, 40, 50)은 스마트폰(Smart Phone), 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC 등이 있다.
- [0027] 문서 분류 장치(100)은 무선 또는 유선 통신 방식을 이용하여 네트워크(10)를 통해 다른 전자기기들과 통신할 수 있다.
- [0028] 본 발명의 실시예에 따른 문서 구조와 딥러닝을 이용한 문서 분류 장치(100)는 문서 데이터베이스부(110), 전처리부(120), 특징 벡터 생성부(130) 및 텍스트 분류기(140)를 포함한다.
- [0029] 문서 데이터베이스부(110)는 적어도 하나의 클래스로 구성되고, 적어도 하나의 문서를 포함할 수 있으며, 통신부(미도시)를 통해 수신한 문서일 수 있다. 여기서, 문서는 논문, 이력서, 리뷰(Review), SNS 데이터 등 텍스트로 이루어진 문서, 문장을 모두 포함한다.
- [0030] 통신부(미도시)는 무선 통신 또는 유선 통신을 통해서 외부의 전자 장치에 연결될 수 있다.
- [0031] 전처리부(120)는 문서 데이터베이스부(110)로부터 수신한 문서를 문자열로 변환하고, 전치사, 접속사, 대명사

등의 정보가 없는 단어를 제거하며, 동일한 개념적 의미를 가진 단어 그룹을 만드는 단어의 형태소 분석을 수행한다.

[0032] 전처리부(120)는 형태소 분석기를 이용하여 공지의 형태소 분석 알고리즘을 이용하여 일련의 규칙을 통해 영어 단어를 어간으로 변환한다.

[0033] 전처리부(120)는 문서를 설명하기 위한 방법으로 벡터 공간 모델(Vector Space Model)을 이용한 인덱싱 처리를 수행하여 문서에 포함된 단어를 특징 벡터로 생성한다.

[0034] 특징 벡터 생성부(130)는 하기의 [수학식 1]과 같이, 문서에 포함된 단어에 가중치를 부여하여 특징 벡터를 생성한다.

수학식 1

$$D = W_{wd}$$

[0035]

여기서, d는 문서, w는 단어, W는 가중치, D는 행렬 벡터를 나타낸다.

[0036]

각 문서에 존재하는 단어에 대해 가중치를 부여하는 알고리즘은 TF(Term Frequency), Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy를 포함한다.

[0037]

TF는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값이고,

[0038]

TFIDF는 정보 검색과 텍스트 마이닝에서 이용하는 가중치로 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치를 나타내며, 단어 빈도와 역문서 빈도의 곱으로 나타낸다.

[0039]

특징 벡터 생성부(130)는 TF를 이용하여 특징 벡터를 정수 벡터로 변환할 수 있다.

[0040]

특징 벡터들의 사이즈는 DBN 알고리즘에 대한 전체 조건 입력과 동일하다.

[0041]

특징 벡터 생성부(130)는 전처리부(120)로부터 전처리된 문서를 입력받아 문서를 적절한 클래스로 분류하기 위하여 Deep Belief Network(이하, 'DBN'라 칭함) 알고리즘을 이용하여 학습 데이터와 특징 벡터들의 특징 세트를 구성한다.

[0042]

DBN 알고리즘은 Restricted Boltzmann Machine(이하 'RBM'이라 칭함)이 여러 층으로 쌓여져 있는 형태로 구성되어 소정의 학습 훈련을 하기 위한 관측값이 입력되는 입력 레이어와 상기 입력 레이어의 각각의 노드들과 연결되는 각각의 노드들로 구성되는 상위 레이어와, 상기 상위 레이어와 연결되는 제1 은닉 레이어 및 상기 제1 은닉 레이어와 연결되는 제2 은닉 레이어로 구성될 수 있다.

[0043]

RBM은 볼 수 있는 층(Visible Layer)와 숨겨진 층(Hidden Layer)의 두 층으로 이루어지며 하나의 층을 구성하는 노드들과 다른 층을 구성하는 노드들은 각각의 가중치(W)에 의하여 연결 관계가 결정되고 소정의 학습 규칙에 따라 원래의 입력 데이터가 확률적으로 복구될 때까지 각각의 히든 노드값 및 각각의 가중치를 업데이트함으로써 학습을 하게 된다.

[0044]

DBN의 목적은 다차원 신경망을 통해 주어진 입력 데이터(즉, 특징 벡터)로부터 학습하고, 높은 확률로 입력 데이터의 내용을 재구성한다.

[0045]

DBN은 하나의 입력 레이어, N개의 히든 레이어, 여러 개의 노드들의 세트를 갖는 하나의 출력 레이어를 포함하며, 히든 레이어 및 노드의 수는 사용자에게 의해 조정될 수 있다.

[0046]

하기의 [수학식 2]는 입력 레이어와 히든 레이어의 결합 확률 분포를 나타낸다.

[0047]

수학식 2

$$P(v, h^1, \dots, h^l) = P(v|h^1) \left(\prod_{k=1}^l P(h^k|h^{k+1}) \right)$$

[0048]

[0049] 여기서, V 는 입력 레이어로부터 특징 벡터이고, l 은 히든 레이어들의 수이고, h^k 는 k 번째 레이어의 특징 벡터를 나타낸다.

[0050] 인접한 두 레이어(k 와 $k+1$)에 대한 조건부 분포는 $P(h^k|h^{k+1})$ 라고 하며, 제한된 볼츠만 머신(Boltzmann Machines, RBM)을 통해 계산된다.

[0051] 노드의 경우, DBN은 현재 노드에서 상위 레벨의 노드의 확률을 학습한다. 역방향 유효성 검사는 DBN에 의해 수행되어 레이어 전체의 노드 간에 가중치를 조정하여 입력 특징 세트를 재구성한다.

[0052] 텍스트 분류기(140)는 특징 벡터 생성부(130)로부터 구성된 특징 벡터들의 특징 세트를 수신하고, 수신한 특징 세트를 결정 트리(Decision Tree), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(Support Vector Machine)의 지도 학습 기술을 이용하여 텍스트를 분류한다.

[0053] 나이브 베이즈는 베이즈 정리에 기초한 확률론적 분류기로서 모든 특징이 서로 독립적이라는 가정하에 작동하고 조건부 확률을 서로 곱하여 확률 점수를 계산한다.

[0054] 패턴 p 의 패턴 클래스 $L(p)$ 는 문서의 Problem Definition의 설명에 나타나는 단어와 어떤 관계가 있다고 가정한다. 베이지안 공식의 [수학식 3]은 문서의 설명에 사용된 단어가 주어진 패턴 클래스의 확률을 산출한다.

수학식 3

$$p(L_i|t_1, \dots, t_{m_i}) = \frac{p(t_1, \dots, t_{m_i}|L_k) p(L_k)}{\sum_{l \in L} p(t_1 \dots t_{m_i}|l) p(l)}.$$

[0055]

[0056] 여기서, 사전 확률 $p(l)$ 은 설계 문제가 알려진 단어 앞에 클래스 레이블 $l \in L$ 에 속할 확률을 나타낸다. 클래스 L_k 가 주어지면 단어의 조건부 확률(패턴 및 문서의 설명에 포함)은 하기의 [수학식 4]를 나타낸다.

수학식 4

$$p(t_1, \dots, t_{m_i}|L_k) = \prod_{j=1}^{m_i} p(t_j|L_k).$$

[0057]

[0058] 여기서, 단어 $p(t_j|L_k)$ 의 확률 추정치는 훈련 세트의 문서에서 빈도에 의존하는 나이브 베이 분류기의 학습을 하기 위해 필요하다.

[0060] 의사 결정 트리 기반 알고리즘은 분류 결정을 내리기 위해 일련의 규칙을 생성하고, 정보 엔트로피의 개념을 사용하고 훈련 세트로부터 결정을 내린다.

[0061] 문서의 클래스를 예측하기 위해서 단어 t_i 는 레이블이 지정된 문서의 학습 세트 T 에서 선택된다.

[0062] 그 다음에, 훈련 세트 T 는 두 개의 부분 집합 T_+ (단어 t_i 를 가진 문서)와 T_- (단어 t_i 가없는 문서)로 분할된다.

동일한 단계가 반복되어 T+ 및 T-에 적용된다. 재귀 프로세스는 모든 문서가 동일한 클래스 Lk에 속하면 중지된다.

[0064] 서포트 벡터 머신은 의사 결정 화면을 찾고 가장 가까운 데이터 요소를 통해 여백을 결정하는 가장 효과적인 텍스트 분류 알고리즘 중 하나이다. 일반적으로 SVM 알고리즘은 훈련 세트의 문서를 양수에 대해 $y = +1$ 과 음수에 대해 $y = -1$ 로 구분하여 분류할 수 있다.

[0065] 이어서, 하기의 [수학식 5]는 입력 벡터의 집합에 대해 $y = 0$ 에 하이퍼 평면을 정의하는데 사용된다. 문서 d에 대한 각각의 입력 벡터는 하기의 [수학식 6]에 도시된 바와 같이 단어들의 카운트로서 표현된다.

수학식 5

$$\vec{t_d} = t_{d1}, \dots, t_{dM}.$$

수학식 6

$$y = f(\vec{t_d}) = b_0 + \sum_{j=1}^M b_j t_{dj}.$$

[0068] 서포트 벡터 머신은 양의 클래스 $\vec{t_d}$ 를 사용하여 새로운 문서를 분류하고, $f(\vec{t_d}) > 0$ 가 아니라면 음의 클래스를 사용하여 새로운 문서를 분류한다.

[0069] 텍스트 분류기(140)는 DT, NB, SVM의 각각에 대하여 Binary, TFIDF(Term Frequency Inverse Document Frequency), TFC(Term Frequency Collection), LTC(Length Term Collection), Entropy의 가중치 부여 방식마다 공지된 F-Measure 방법에 적용하여 F 측정값을 각각 계산한다.

[0070] 텍스트 분류기(140)는 계산된 F 측정값이 가장 높은 가중치 부여 방식을 특징 벡터 생성부(130)의 가중치 부여 방식으로 선택하고, 가장 높은 가중치 부여 방식에 대응하는 지도 학습 기술이 텍스트를 분류하는 방법으로 적용된다.

[0072] 도 3 및 도 4는 본 발명의 제1 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이고, 도 5 및 도 6은 본 발명의 제2 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이며, 도 7 및 도 8은 본 발명의 제3 실시예에 따른 텍스트 분류기에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 나타낸 도면이다.

[0073] 하기에서는 텍스트 분류기(140)에서 가중치 부여 방식과 지도 학습 기술을 선택하는 과정을 일례로 들어 설명한다.

[0074] 텍스트 분류기(140)에서 텍스트를 분류할 문서의 예시는 다음과 같다.

[0075] (1) Gang-of-Four(GoF) 문서 컬렉션

[0076] GoF 문서 컬렉션에는 Creational, Structural 및 Behavioral이라는 세 그룹으로 나누어진 23개의 객체 지향 문서가 포함되어 있다. GoF 문서 컬렉션은 중지 단어와 형태소 분석 단어를 제거한 후 23개의 모든 문서 중 1465개의 반복되지 않는 단어가 포함된다.

[0078] (2) 더글라스 문서 컬렉션

[0079] 더글라스(Douglass) 문서 컬렉션에는 동시성, 안전성 및 신뢰성, 배포, 메모리 및 자원이라는 5가지 범주로 나뉘어진 34개의 실시간 시스템 관련 문서가 포함되어 있다. 더글라스 문서 컬렉션은 중지 단어와 형태소 분석 단어를 제거한 후 34개 문서 중 1271개의 반복되지 않는 단어가 포함되어 있다.

[0081] (3) 보안 문서 컬렉션

[0082] 보안 문서 컬렉션에는 SACA(시스템 액세스 및 제어 아키텍처), ACM(액세스 제어 모델), IA(식별 및 인증), OSAC(운영 체제 액세스 제어), SIA(보안 인터넷 응용 프로그램), FA(방화벽 아키텍처), ESRM(기업 보안 및 리스크 관리) 및 회계(Accounting) 등이 있다. 이 사례 연구에는 중지 단어와 형태소 분석 단어를 제거한 후 34개의 문서 중 1230개의 반복되지 않는 단어가 포함된다.

[0084] 텍스트 분류 관점에서 정확도(Precision)와 재현율(Recall)은 하기의 [수학식 7]과 [수학식 8]와 같이 마이크로 평균 방정식을 이용하여 추정할 수 있다.

수학식 7

$$P = \frac{\sum_{c=1}^N TP_c}{\sum_{c=1}^N (TP_c + FP_c)}, \quad \text{Micro-Averaging}$$

[0085]

수학식 8

$$R = \frac{\sum_{c=1}^N TP_c}{\sum_{c=1}^N (TP_c + FN_c)}, \quad \text{Micro-Averaging}$$

[0086]

수학식 9

$$F = \frac{2 \times P \times R}{(P + R)}.$$

[0087]

[0088] 각각의 지도 학습 기술(DT, NB, SVM)은 최상의 가중치 부여 방식(이진, TFIDF, LTC, TFC 및 엔트로피)을 선택하기 위해서 각각의 가중치 부여 방식에 대한 결과가 F 측정 방법(F-Measure)을 이용하여 F 측정값으로 계산된다.

[0089] 결과적으로 가장 높은 F 측정값을 갖는 가중치 부여 방식이 해당 지도 학습 기술에 가장 적합하게 선택된다.

[0090] [수학식 7] 내지 [수학식 9]에서 N은 학습 기술의 성능을 평가하기로 결정한 문서 클래스의 수이다. 예를 들면, GoF 문서 수집의 경우 N은 3이다. TP는 각 문서 클래스에 대해 정확하게 식별되는 문서 개수이고, FP는 문서 개수가 각 문서 클래스에 대해 잘못 식별된 문서들의 개수이고, FN은 각각의 해당 문서 클래스에서 누락된 문서의 개수이다.

[0091] P, R 및 F의 값은 [수학식 7] 내지 [수학식 9]를 사용하여 계산된다.

[0092] 이어서, [수학식 9]는 F 측정값이 가장 높은 가중치 부여 방식을 사용하여 DT, NB, SVM의 지도 학습 기술의 성과를 평가하는데 사용할 수 있다.

- [0094] GoF 문서 컬렉션은 전처리 과정을 거치고 특징 벡터 생성부(130)를 통해 특징 벡터들의 특징 세트를 구성한다.
- [0095] 텍스트 분류기(140)는 F 측정 방법(수학식 3, 수학식 4, 수학식 5)에 따라 지도 학습 기술(DT, NB, SVM)에 대한 가장 높은 F 측정값을 가진 가중치 부여 방식(이진, TFIDF, LTC, TFC 및 엔트로피)을 계산한다.
- [0096] 도 3에 도시된 바와 같이, TFC 및 TFIDF의 가중치 부여 방식을 사용하여 NB, SVM 및 DT의 성능이 향상되는 것을 확인할 수 있다.
- [0097] 본 발명은 NB(F 측정값 = 0.76) 및 SVM(F 측정값 = 0.78)과 비교할 때 가중치 부여 방식 TFIDF를 가진 DT(F 측정값 = 0.80)가 우수한 성능의 분류기로 관찰되었다.
- [0098] 따라서, 특징 벡터 생성부(130)는 TFIDF를 가중치 부여 방식으로 사용하고, 텍스트 분류기(140)는 DT를 지도 학습 기술로 적용하여 텍스트를 분류한다.
- [0099] 도 4는 가중치 부여 방식(TFIDF)과 지도 학습 기술(DT)이 적용된 텍스트 분류 기술로 X축이 n개의 순위가 매겨진 상위 n개의 특징 벡터의 개수이고, Y축이 해당 F 측정값이다.
- [0100] 본 발명은 특정한 가중치 부여 방식(TFIDF)과 지도 학습 기술(DT)을 통해 구성된 특징 벡터가 분류 성능에 중요한 영향을 미치는 것을 관찰할 수 있다.
- [0101] DT의 분류 성능은 순위가 매겨진 상위 10개의 특징 벡터들이 2.5%에서 11.25%(F 측정 기준) 범위에서 평균적으로 향상되었다.
- [0102] 본 발명은 순위가 결정된 상위 10개의 특징 벡터들이 DT의 분류 성능이 12.5%(F 측정 방법) 향상되었다.
- [0104] 더글러스 문서 컬렉션은 전처리 과정을 거치고 특징 벡터 생성부(130)를 통해 특징 벡터들의 특징 세트를 구성한다.
- [0105] 텍스트 분류기(140)는 F 측정 방법(수학식 3, 수학식 4, 수학식 5)에 따라 지도 학습 기술(DT, NB, SVM)에 대한 가장 높은 F 측정값을 가진 가중치 부여 방식(이진, TFIDF, LTC, TFC 및 엔트로피)을 계산한다.
- [0106] 도 5에 도시된 바와 같이, TFC 및 TFIDF의 가중치 부여 방식을 사용하여 NB, SVM 및 DT의 성능이 향상되는 것을 확인할 수 있다.
- [0107] 본 발명은 NB(F 측정값 = 0.79) 및 DT(F 측정값 = 0.80)과 비교할 때 가중치 부여 방식 TFIDF를 가진 SVM(F 측정값 = 0.83)이 우수한 성능의 분류기로 관찰되었다.
- [0108] 따라서, 특징 벡터 생성부(130)는 TFIDF를 가중치 부여 방식으로 사용하고, 텍스트 분류기(140)는 SVM을 지도 학습 기술로 적용하여 텍스트를 분류한다.
- [0109] 도 6은 가중치 부여 방식(TFIDF)과 지도 학습 기술(SVM)이 적용된 텍스트 분류 기술로 X축이 n개의 순위가 매겨진 상위 n개의 특징 벡터의 개수이고, Y축이 해당 F 측정값이다.
- [0110] 본 발명은 특정한 가중치 부여 방식(TFIDF)과 지도 학습 기술(SVM)을 통해 구성된 특징 벡터가 분류 성능에 중요한 영향을 미치는 것을 관찰할 수 있다.
- [0111] SVM의 분류 성능은 순위가 매겨진 상위 10개의 특징 벡터들이 2.40%에서 10.84%(F 측정 기준) 범위에서 평균적으로 향상되었다.
- [0112] 본 발명은 순위가 결정된 상위 10개의 특징 벡터들이 SVM의 분류 성능이 10.84% 향상되었다.
- [0114] 보안 문서 컬렉션은 전처리 과정을 거치고 특징 벡터 생성부(130)를 통해 특징 벡터들의 특징 세트를 구성한다.
- [0115] 텍스트 분류기(140)는 F 측정 방법(수학식 3, 수학식 4, 수학식 5)에 따라 지도 학습 기술(DT, NB, SVM)에 대한 가장 높은 F 측정값을 가진 가중치 부여 방식(이진, TFIDF, LTC, TFC 및 엔트로피)을 계산한다.
- [0116] 도 7에 도시된 바와 같이, LTC 및 TFC의 가중치 부여 방식을 사용하여 NB, SVM 및 DT의 성능이 향상되는 것을 확인할 수 있다.

- [0117] 본 발명은 NB(F 측정값 = 0.79) 및 DT(F 측정값 = 0.75)와 비교할 때 가중치 부여 방식 TFC를 가진 SVM(F 측정값 = 0.82)이 우수한 성능의 분류기로 관찰되었다.
- [0118] 따라서, 특징 벡터 생성부(130)는 TFC를 가중치 부여 방식으로 사용하고, 텍스트 분류기(140)는 SVM을 지도 학습 기술로 적용하여 텍스트를 분류한다.
- [0119] 도 8은 가중치 부여 방식(TFC)과 지도 학습 기술(SVM)이 적용된 텍스트 분류 기술로 X축이 n개의 순위가 매겨진 상위 n개의 특징 벡터의 개수이고, Y축이 해당 F 측정값이다.
- [0120] 본 발명은 특정한 가중치 부여 방식(TFC)과 지도 학습 기술(SVM)을 통해 구성된 특징 벡터가 분류 성능에 중요한 영향을 미치는 것을 관찰할 수 있다.
- [0121] SVM의 분류 성능은 순위가 매겨진 상위 10개의 특징 벡터들이 4.80%에서 13.41%(F 측정 기준) 범위에서 평균적으로 향상되었다.
- [0122] 본 발명은 순위가 결정된 상위 10개의 특징 벡터들이 SVM의 분류 성능이 13.41% 향상되었다.

표 1

Performance evaluation of classifiers on different design pattern catalogs.

Pattern catalog (i.e. Dataset)	Classifiers	Best weighting method	F1-criteria
GoF Pattern Catalog	DT	TFIDF	<u>0.80</u>
	NB	TFC	0.76
	SVM	TFC	0.78
Douglass Pattern Catalog	DT	TFIDF	0.80
	NB	TFC	0.79
	SVM	TFIDF	<u>0.83</u>
Security Pattern Catalog	DT	TFC	0.75
	NB	LTC	0.79
	SVM	TFC	<u>0.82</u>

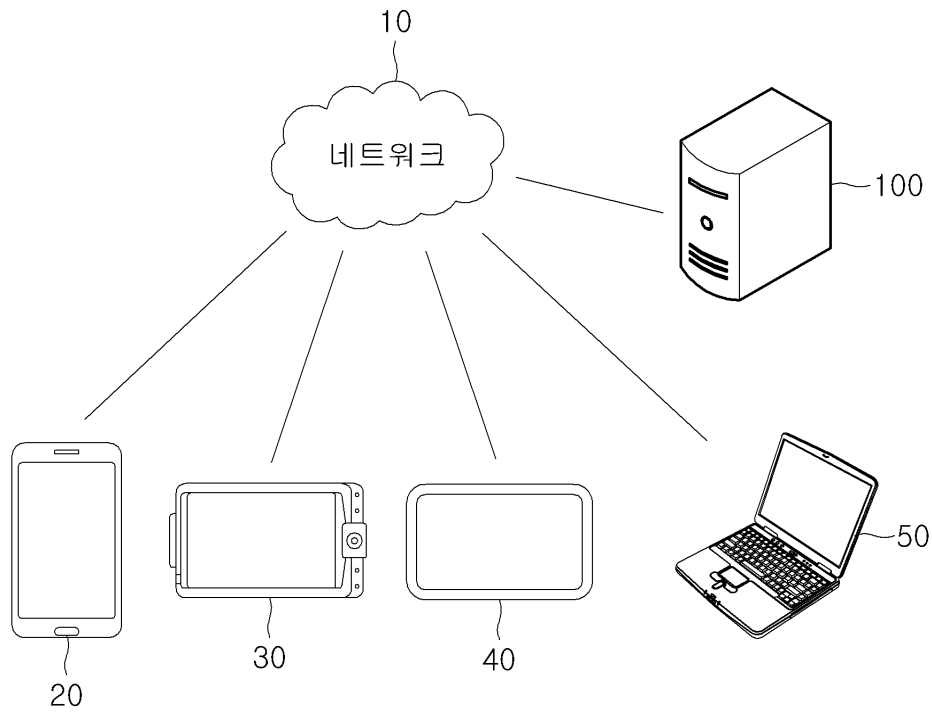
- [0124]
- [0125] 전술한 [표 1]에 도시된 바와 같이, 더글라스 문서 컬렉션 및 보안 문서 컬렉션의 경우, 분류기가 SVM이 더 분류 성능이 우수하고, Gang-of-Four(GoF) 문서 컬렉션 및 더글라스 문서 컬렉션의 경우, 가중치 적용 방식 TFIDF가 다른 가중치 적용 방식보다 성능이 우수하다.
- [0126] 이상에서 본 발명의 실시예는 장치 및/또는 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하기 위한 프로그램, 그 프로그램이 기록된 기록 매체 등을 통해 구현될 수도 있으며, 이러한 구현은 앞서 설명한 실시예의 기재로부터 본 발명이 속하는 기술분야의 전문가라면 쉽게 구현할 수 있는 것이다.
- [0127] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

부호의 설명

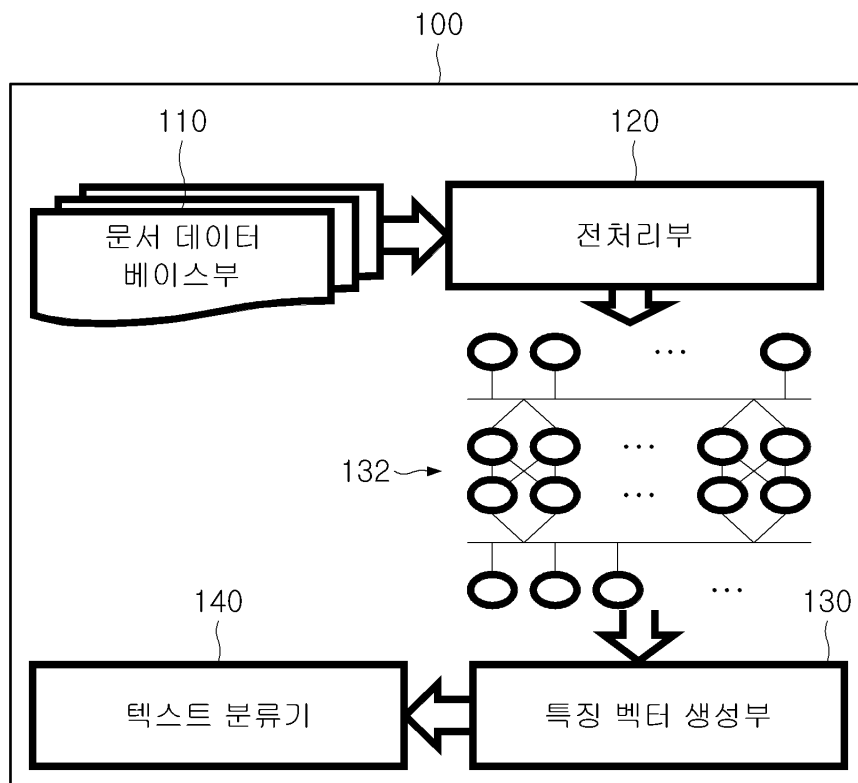
- [0128] 100: 문서 분류 장치
- 110: 문서 데이터베이스부
- 120: 전처리부
- 130: 특징 벡터 생성부
- 140: 텍스트 분류기

도면

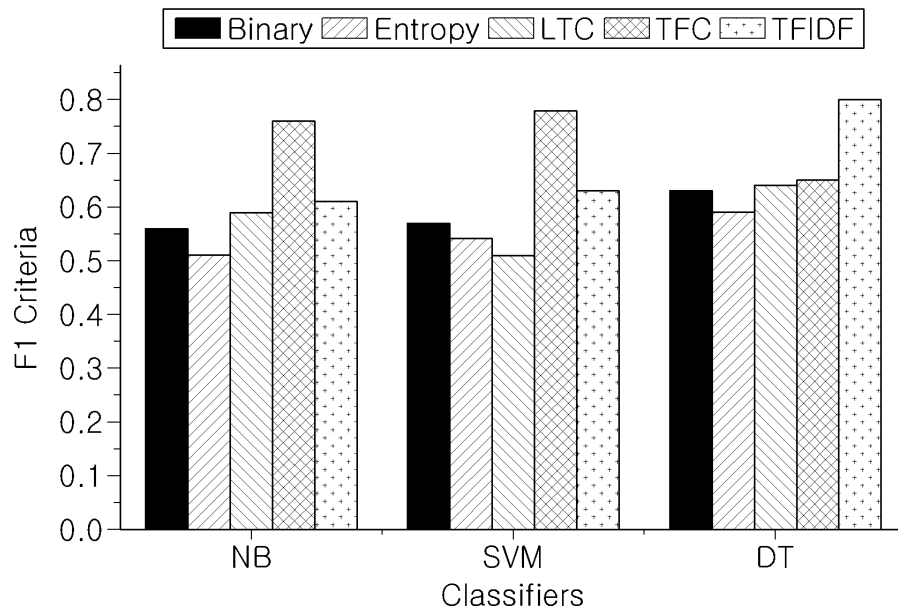
도면1



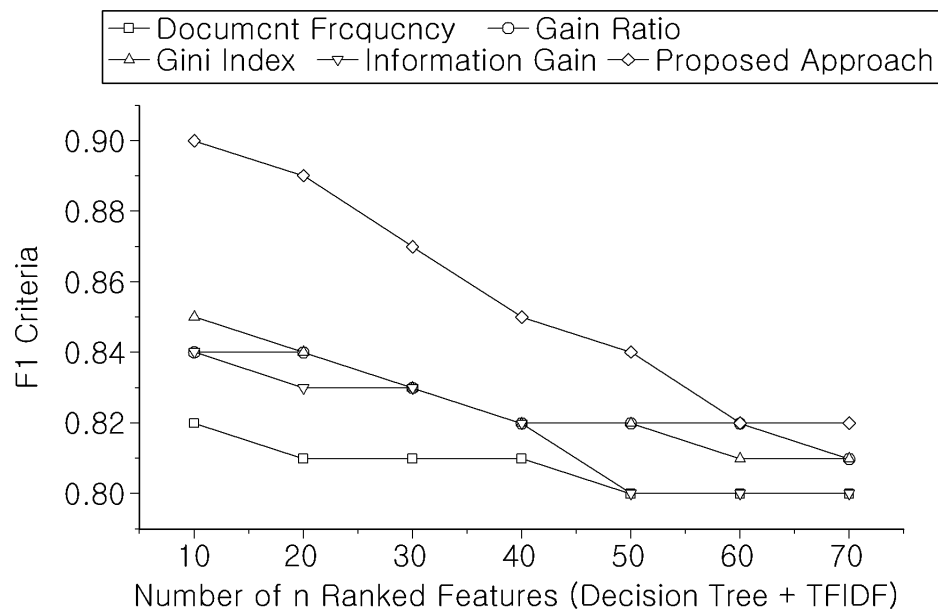
도면2



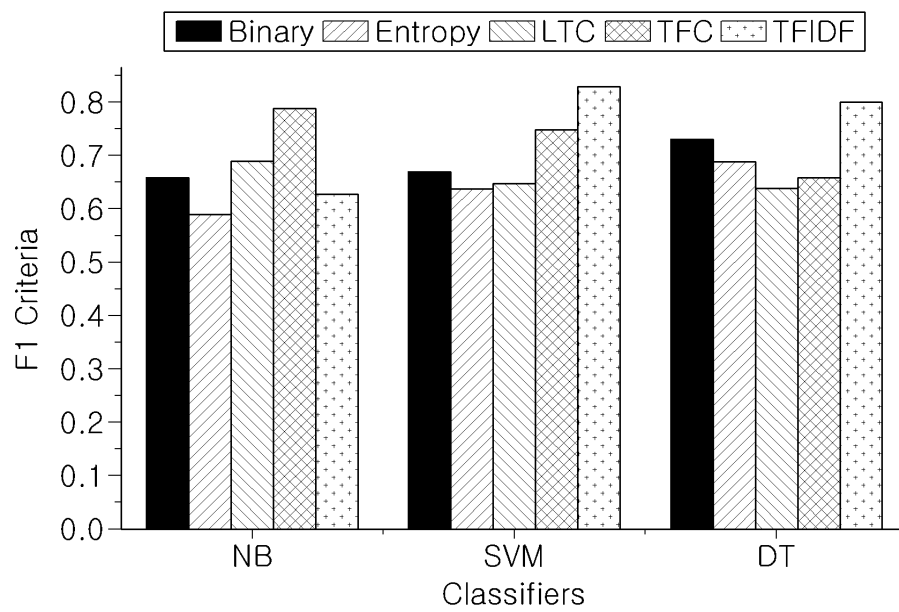
도면3



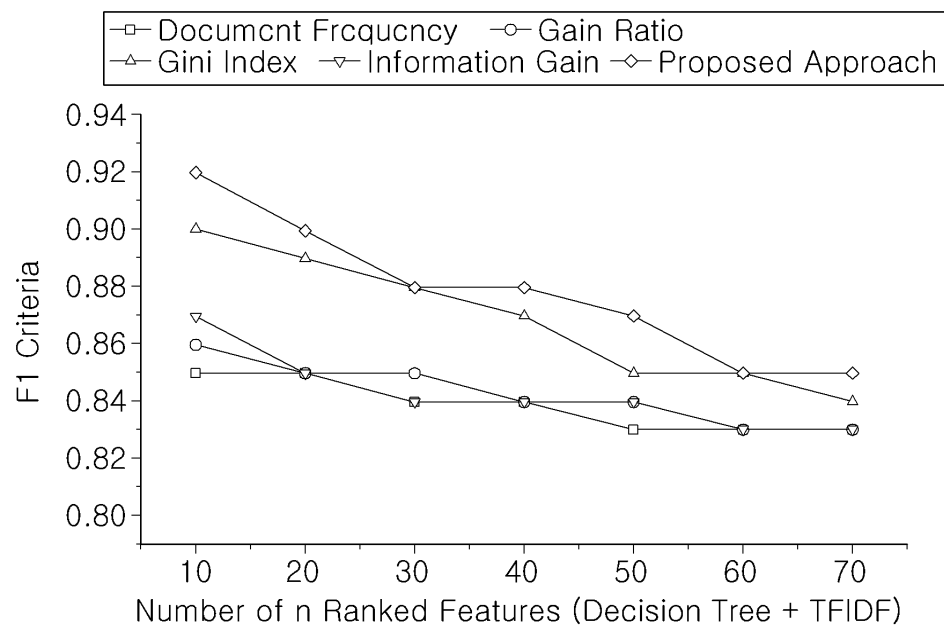
도면4



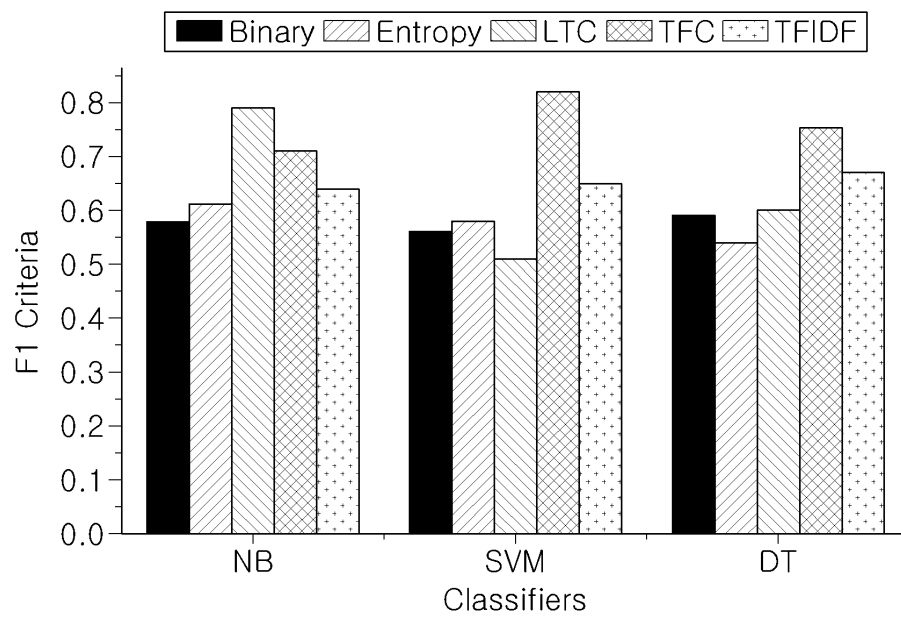
도면5



도면6



도면7



도면8

