



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0066994

(43) 공개일자 2015년06월17일

(51) 국제특허분류(Int. Cl.)

G06F 17/00 (2006.01) G06F 17/30 (2006.01)

(21) 출원번호 10-2014-0070694

(22) 출원일자 2014년06월11일

심사청구일자 2014년06월11일

(30) 우선권주장

1020130152166 2013년12월09일 대한민국(KR)

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

이상근

서울특별시 강남구 선릉로 221, 205동 1104호 (도곡동, 도곡테크아파트)

하중우

서울특별시 성북구 개운사1길 54, 502호 (안암동5가)

이정현

서울특별시 영등포구 도신로4길 14, 102동 1905호 (대림동, 현대1차아파트)

(74) 대리인

특허법인엠에이피에스

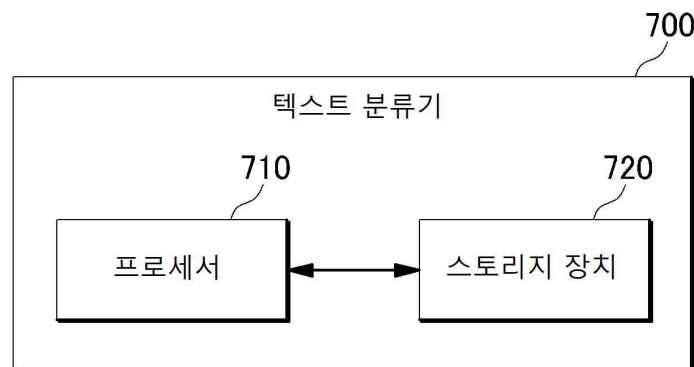
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 조상 후손 카테고리를 활용한 오픈 디렉터리 프로젝트 기반 텍스트 분류 방법 및 장치

(57) 요약

본 발명은 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에 있어서, 오픈 디렉터리 프로젝트의 각 카테고리에 저장된 텍스트 집합을 특징 벡터로 변환하는 단계; 카테고리의 특징 벡터를 이용하여 카테고리의 조상 병합 특징 벡터를 생성하는 단계; 카테고리의 특징 벡터를 이용하여 카테고리의 후손 병합 특징 벡터를 생성하는 단계; 및 카테고리의 특징 벡터, 조상 병합 특징 벡터 및 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하는 단계를 포함하고, 오픈 디렉터리 프로젝트의 구조는 미리 정의되어 있는 것을 특징으로 한다.

대표도 - 도8



명세서

청구범위

청구항 1

오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에 있어서,
상기 오픈 디렉터리 프로젝트의 각 카테고리에 저장된 텍스트 집합을 특징 벡터로 변환하는 단계;
상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 조상 병합 특징 벡터를 생성하는 단계;
상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 후손 병합 특징 벡터를 생성하는 단계; 및
상기 카테고리의 특징 벡터, 상기 조상 병합 특징 벡터 및 상기 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하는 단계;를 포함하고,
상기 오픈 디렉터리 프로젝트의 구조는 미리 정의되어 있는,
텍스트 분류기 생성 방법.

청구항 2

제 1항에 있어서,
상기 텍스트 집합을 특징 벡터로 변환하는 단계는,
상기 오픈 디렉터리 프로젝트의 상기 카테고리에 속하는 상기 텍스트 집합에서 단어를 추출하는 단계;
상기 카테고리에 속하는 텍스트별로 단어 발생 빈도 벡터를 생성하는 단계; 및
상기 카테고리에 속하는 텍스트에 대한 단어 발생 빈도를 이용하여 상기 특징 벡터를 생성하는 단계;를 포함하는,
텍스트 분류기 생성 방법.

청구항 3

제2항에 있어서,
상기 특징 벡터를 생성하는 단계는, 상기 카테고리에 속한 단어 각각의 평균 발생 빈도를 이용하여 상기 특징 벡터를 생성하는 것을 포함하는,
텍스트 분류기 생성 방법.

청구항 4

제1항에 있어서,
상기 조상 병합 특징 벡터를 생성하는 단계는,
상기 카테고리가 최상위 카테고리인 경우, 상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 조상 병합 특징 벡터를 생성하고,
상기 카테고리가 최상위 카테고리가 아닌 경우, 상기 카테고리의 특징 벡터와 상기 카테고리의 부모 카테고리의 조상 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 조상 병합 특징 벡터를 생성하되, 상기 카테고리의 특징 벡터의 가중치를 상기 부모 카테고리의 조상 병합 특징 벡터의 가중치보다 높게 설정하는,

텍스트 분류기 생성 방법.

청구항 5

제 1항에 있어서,

상기 후손 병합 특징 벡터를 생성하는 단계는,

상기 카테고리가 최말단 카테고리인 경우, 상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 후손 병합 특징 벡터를 생성하고,

상기 카테고리가 최말단 카테고리가 아닌 경우, 상기 카테고리의 특징 벡터와 상기 카테고리의 자식 카테고리의 평균 후손 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 후손 병합 특징 벡터를 생성하되, 상기 카테고리의 특징 벡터의 가중치를 상기 자식 카테고리의 후손 병합 특징 벡터 평균의 가중치보다 높거나 같게 설정하는,

텍스트 분류기 생성 방법.

청구항 6

제 1항에 있어서,

상기 조상 후손 병합 특징 벡터를 생성하는 단계는 상기 조상 병합 특징 벡터와 상기 후손 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 조상 후손 병합 특징 벡터를 생성하되,

상기 조상 병합 특징 벡터의 가중치와 상기 후손 병합 특징 벡터의 가중치를 서로 상이하게 설정하는,

텍스트 분류기 생성 방법.

청구항 7

오픈 디렉터리 프로젝트 환경에서 텍스트 분류기를 이용한 텍스트 분류 방법에 있어서,

상기 신규 텍스트를 특징 벡터로 변환하는 단계; 및

상기 특징 벡터와 상기 텍스트 분류기를 이용하여 상기 신규 텍스트의 카테고리를 분류하는 단계;를 포함하되,

상기 텍스트 분류기는 상기 오픈 프로젝트 디렉터리의 각 카테고리에 대한 특징 벡터, 상기 카테고리의 조상 병합 특징 벡터 및 상기 카테고리의 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하여 구축된 것인,

텍스트 분류 방법.

청구항 8

제7항에 있어서,

상기 카테고리 분류 단계는,

상기 신규 텍스트의 특징 벡터와 상기 오픈 프로젝트 디렉터리에 포함된 각 카테고리의 조상 후손 병합 특징 벡터에 대하여 유사도를 계산하는 단계;

상기 유사도가 가장 높은 카테고리를 선택하는 단계; 및

상기 단계에서 선택한 카테고리를 상기 신규 텍스트의 카테고리로 분류하는 단계;를 포함하는,

텍스트 분류 방법.

청구항 9

텍스트 분류기에 있어서,

오픈 디렉터리 프로젝트의 카테고리에 기반하여 신규 텍스트의 카테고리를 분류하는 텍스트 분류 프로그램이 저장된 스토리지 장치, 및

상기 스토리지 장치에 저장된 상기 텍스트 분류 프로그램을 실행하는 프로세서를 포함하되,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라, 미리 정의된 오픈 디렉터리 프로젝트 구조에 기반하여 상기 카테고리의 특징 벡터, 상기 카테고리의 조상 병합 특징 벡터 및 상기 카테고리의 후손 병합 특징 벡터를 이용하여 조상 후손 특징 벡터를 생성하고, 상기 조상 후손 병합 특징 벡터를 이용하여 텍스트 분류기를 생성하며, 상기 텍스트 분류기를 이용하여 상기 신규 텍스트의 카테고리를 분류하는,

텍스트 분류기.

청구항 10

제9항에 있어서,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라,

상기 오픈 디렉터리 프로젝트에 포함된 텍스트 집합에서 단어를 추출하고, 상기 카테고리에 속한 단어에 대한 발생 빈도를 이용하여 상기 카테고리의 특징 벡터를 생성하는,

텍스트 분류기.

청구항 11

제10항에 있어서,

상기 카테고리의 특징 벡터는 상기 카테고리에 속한 단어에 대한 각각의 평균 발생 빈도를 이용하여 생성되는 것인,

텍스트 분류기.

청구항 12

제9항에 있어서,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라,

상기 카테고리가 최상위 카테고리일 경우, 상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 조상 병합 특징 벡터를 생성하고,

상기 카테고리가 최상위 카테고리가 아닌 경우, 상기 카테고리의 특징 벡터와 상기 카테고리의 조상 카테고리의 조상 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 조상 병합 특징 벡터를 생성하되, 상기 카테고리의 특징 벡터의 가중치를 상기 부모 카테고리의 조상 병합 특징 벡터의 가중치보다 높게 설정하는,

텍스트 분류기.

청구항 13

제9항에 있어서,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라,

상기 카테고리가 최말단 카테고리일 경우, 상기 카테고리의 특징 벡터를 이용하여 상기 카테고리의 후손 병합 특징 벡터를 생성하고,

상기 카테고리가 최말단 카테고리가 아닌 경우, 상기 카테고리의 특징 벡터와 상기 카테고리의 자식 카테고리의 평균 후손 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 후손 병합 특징 벡터를 생성하되, 상기 카테고리의 특징 벡터의 가중치를 상기 자식 카테고리의 평균 후손 병합 특징 벡터의 가중치보다 높게 설정하는,

텍스트 분류기.

청구항 14

제 9항에 있어서,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라,

상기 조상 병합 특징 벡터와 상기 후손 병합 특징 벡터에 대한 가중합을 이용하여 상기 카테고리의 조상 후손 병합 특징 벡터를 생성하되, 상기 조상 병합 특징 벡터의 가중치와 상기 후손 병합 특징 벡터의 가중치를 상이하게 설정하는,

텍스트 분류기.

청구항 15

제9항에 있어서,

상기 프로세서는 상기 텍스트 분류 프로그램의 실행에 따라,

상기 신규 텍스트의 카테고리 분류를 위하여 상기 신규 텍스트의 특징 벡터와 상기 오픈 디렉터리 프로젝트의 각 카테고리의 조상 후손 병합 특징 벡터에 대하여 유사도를 계산한 후, 상기 유사도가 가장 높은 카테고리를 상기 신규 텍스트의 카테고리로 분류하는,

텍스트 분류기.

발명의 설명

기술 분야

[0001] 본 발명은 조상 후손 카테고리를 활용한 오픈 디렉터리 프로젝트 기반 텍스트 분류 방법 및 장치에 관한 것이다.

배경 기술

[0002] 오픈 디렉터리 프로젝트(open directory project)는 일종의 웹상의 텍스트 데이터베이스이다. 오픈 디렉터리 프로젝트는 월드 와이드 웹(world wide web)상의 웹 문서 등의 텍스트를 계층적 온톨로지(hierarchical ontology)를 이용하여 카테고리(category)로 세분화하고, 트리(tree) 자료 구조(data structure)를 이용하여 분류하였다.

[0003] 오픈 디렉터리 프로젝트의 카테고리에 기반하여 신규 텍스트의 카테고리를 분류하기 위한 종래기술은 나이브 베이즈(Naive Bays), k-근접 이웃 방법(k-nearest neighbor), 서포트 벡터 머신(support vector machine) 등 기계 학습 알고리즘(machine learning algorithm) 기반 기존 텍스트 분류(text classification) 기법을 활용하는 것이다. 그러나 이러한 종래 기술을 이용하는 경우, 오픈 디렉터리 프로젝트의 카테고리가 증가함에 따라, 텍

스트 분류기를 생성할 때 목표(target)가 되는 카테고리에 포함된 학습 데이터가 부족하고, 이 학습 데이터를 이용하여 학습된 텍스트 분류기의 분류 정확도가 낮아진다는 문제점이 있다.

[0004]

한국 등록특허공보 제1158750호(발명의 명칭: 문서분류장치 및 그것의 문서분류방법)는 카테고리별로 이미 분류된 학습 데이터에 대하여 기계 학습 알고리즘을 이용하여 공통 패턴을 추출하고, 신규 텍스트가 입력되면 공통 패턴과의 유사도(similarity)를 측정하여 카테고리를 분류하는 방법을 개시하고 있다. 이 발명은 각 카테고리에 대한 유사도 계산 시, 유사도가 가장 높은 카테고리의 유사도가 미리 설정된 값에 수렴할 경우에는 해당 카테고리로 분류하고, 수렴하지 않으면 계산된 유사도를 공통 패턴에 포함된 학습 데이터의 가중치로 고려하여 다시 학습을 수행한 후 카테고리를 분류한다. 이 발명은 카테고리별로 유사도를 반영하여 반복적으로 텍스트 분류기를 학습하므로 텍스트 분류 정확도가 향상될 수 있다. 그러나 공통 패턴 추출 시 기계 학습을 이용하므로, 학습 데이터가 부족한 경우에 대한 정확도가 저하되는 문제를 해결하지 못한다.

[0005]

한국 등록특허공보 제1095069호(발명의 명칭: 사용자 관심 주제를 추출하는 휴대용 통신 단말기 및 그 방법)는 후손 카테고리의 학습데이터를 이용하여 학습 데이터 부족 문제를 해결하는 병합 센트로이드(merge-centroid) 텍스트 분류기를 개시하고 있다. 병합 센트로이드는 텍스트 분류기 생성 시 대상이 되는 임의의 목표 카테고리의 자식 카테고리를 포함한 모든 후손 카테고리에 텍스트를 학습 데이터로 이용하여 텍스트 분류기를 생성된다. 이 방법은 텍스트 분류기에 목표 카테고리 및 목표 카테고리의 후손 카테고리의 가중치를 동일하게 반영된다. 그러므로 후손 카테고리에 포함된 데이터가 많을 경우, 목표 카테고리의 정보보다 후손 카테고리의 정보가 잘 반영되어, 목표 카테고리에 대한 분류 정확도가 낮아질 수 있다.

발명의 내용

해결하려는 과제

[0006]

본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 본 실시예는 오픈 디렉터리 프로젝트의 조상 카테고리 정보와 후손 카테고리 정보를 병합한 조상 후손 카테고리 정보를 생성하고, 조상 후손 카테고리 정보를 활용한 오픈 디렉터리 프로젝트 기반 텍스트 분류 방법 및 장치를 제공하는데 그 목적이 있다.

[0007]

또한, 본 실시예는 조상 후손 카테고리 정보를 이용한 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법 및 텍스트 분류기의 분류 방법을 제공하는데 그 목적이 있다.

[0008]

다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0009]

본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에 있어서, 오픈 디렉터리 프로젝트의 각 카테고리에 저장된 텍스트 집합을 특징 벡터로 변환하는 단계; 카테고리의 특징 벡터를 이용하여 카테고리의 조상 병합 특징 벡터를 생성하는 단계; 카테고리의 특징 벡터를 이용하여 카테고리의 후손 병합 특징 벡터를 생성하는 단계; 및 카테고리의 특징 벡터, 조상 병합 특징 벡터 및 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하는 단계를 포함하고, 오픈 디렉터리 프로젝트의 구조는 미리 정의되어 있는 텍스트 분류기 생성 방법을 포함한다.

[0010]

본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 환경에서 텍스트 분류기를 이용한 텍스트 분류 방법에 있어서, 신규 텍스트를 특징 벡터로 변환하는 단계; 및 특징 벡터와 텍스트 분류기를 이용하여 신규 텍스트의 카테고리를 분류하는 단계를 포함하되, 텍스트 분류기는 오픈 프로젝트 디렉터리의 각 카테고리에 대한 특징 벡터, 카테고리의 조상 병합 특징 벡터 및 카테고리의 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하여 구축된 것인, 텍스트 분류 방법을 포함한다.

[0011]

본 발명의 일 실시예에 따른 텍스트 분류기에 있어서, 오픈 디렉터리 프로젝트의 카테고리에 기반하여 신규 텍스트의 카테고리를 분류하는 텍스트 분류 프로그램이 저장된 스토리지 장치, 및 스토리지 장치에 저장된 텍스트 분류 프로그램을 실행하는 프로세서를 포함하되, 프로세서는 텍스트 분류 프로그램의 실행에 따라, 미리 정의된 오픈 디렉터리 프로젝트 구조에 기반하여 카테고리의 특징 벡터, 카테고리의 조상 병합 특징 벡터 및 카테고리

의 후손 병합 특징 벡터를 이용하여 조상 후손 특징 벡터를 생성하고, 조상 후손 병합 특징 벡터를 이용하여 텍스트 분류기를 생성하며, 텍스트 분류기를 이용하여 신규 텍스트의 카테고리를 분류하는 텍스트 분류기를 포함한다.

발명의 효과

- [0012] 전술한 과제 해결 수단 중 어느 하나에 의하면, 조상 후손 카테고리를 활용한 오픈 디렉터리 프로젝트 기반 텍스트 분류 방법 및 장치를 발명함에 인하여, 관련 사업 전반에서 다양한 과급 효과를 기대할 수 있다.
- [0013] 본 발명은 텍스트 분류기의 생성을 위하여 선택된 카테고리의 정보뿐만 아니라, 선택된 카테고리의 조상 카테고리 및 후손 카테고리의 정보를 활용할 수 있다. 이렇게 조상 카테고리 및 후손 카테고리를 정보를 활용하여 선택된 카테고리의 정보가 부족하여 텍스트 분류기의 정확도가 저하되는 문제를 해결할 수 있다.
- [0014] 본 발명은 텍스트 분류기 생성을 위하여, 조상 카테고리 또는 후손 카테고리의 가중치를 선택된 카테고리의 가중치에 비하여 낮게 설정할 수 있다. 이렇게 선택된 카테고리의 정보를 조상 카테고리 또는 후손 카테고리보다 중요하게 고려하여 텍스트 분류기를 생성하여, 텍스트 분류기의 정확도를 향상할 수 있다.

도면의 간단한 설명

- [0015] 도 1은 본 발명에 사용되는 오픈 디렉터리 프로젝트의 구조에 대한 예시도이다.
- 도 2는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법을 도시한 순서도이다.
- 도 3은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리별 특징 벡터 생성 방법을 도시한 순서도이다.
- 도 4는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 조상 병합 특징 벡터 생성 방법을 도시한 순서도이다.
- 도 5는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 후손 병합 특징 벡터 생성 방법을 도시한 순서도이다.
- 도 6은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 조상 후손 병합 특징 벡터 생성 방법을 도시한 순서도이다.
- 도 7은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기의 신규 텍스트 분류 방법을 도시한 순서도이다.
- 도 8은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기를 개략적으로 도시한 구성도이다.
- 도 9는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 과정을 설명하기 위한 오픈 디렉터리 프로젝트 예시도이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0017] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 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.

- [0018] 또한, 도면을 참고하여 설명함에 있어서, 동일한 명칭으로 나타난 구성일지라도 도면에 따라 도면 번호가 달라질 수 있고, 도면 번호는 설명의 편의를 위해 기재된 것에 불과하고 해당 도면 번호에 의해 각 구성의 개념, 특징, 기능 또는 효과가 제한 해석되는 것은 아니다.
- [0019] 먼저, 오픈 디렉터리 프로젝트의 구조를 설명한다.
- [0020] 도 1은 본 발명에 사용되는 오픈 디렉터리 프로젝트의 구조에 대한 예시도이다.
- [0021] 본 발명에서 오픈 디렉터리 프로젝트는 도 1과 같이 카테고리를 노드(node)로 가지는 트리 구조로 표현될 수 있다. 트리 구조는 그래프(graph) 구조의 일종으로 순환 구조가 없고, 서로 다른 두 노드를 잇는 길이 하나뿐인 자료 구조이다.
- [0022] 하나 이상의 노드를 포함하는 트리 구조는 하나의 최상위 노드와 하나 이상의 최하위 노드를 가질 수 있다. 최상위 노드는 루트 노드(root node)라고도 하며, 최하위 노드는 리프 노드(leaf node)라고 한다. 도 1에서 최상위 노드는 카테고리 C_1 이 되고, 최하위 노드는 C_4 , C_5 , C_6 및 C_7 이다.
- [0023] 트리 구조에서 임의 노드의 부모 노드(parent node)는 임의 노드와 상위에 연결된 노드를 의미한다. 만약, 임의 노드가 최상위 노드인 경우에는 부모 노드가 없을 수 있다. 최상위 노드가 아닌 임의 노드는 하나의 부모 노드를 가진다. 자식 노드(child node)는 임의 노드와 하위에 연결된 노드를 의미한다. 만약, 임의 노드가 최하위 노드인 경우에는 자식 노드가 없을 수 있다. 최하위 노드가 아닌 임의 노드는 하나 이상의 자식 노드를 가진다. 예를 들어, 도 1의 카테고리 C_2 의 부모 노드는 카테고리 C_1 이며, 자식 노드는 카테고리 C_4 와 카테고리 C_5 이다.
- [0024] 트리 구조에서 임의 노드의 조상 노드(ancestor node)는 임의 노드의 부모 노드부터 최상위 노드까지 상위로 연결된 모든 노드를 의미한다. 예를 들어, 도 1의 카테고리 C_5 의 조상 노드는 카테고리 C_2 와 카테고리 C_1 이다. 임의 노드의 후손 노드(descendant node)는 임의 노드의 자식 노드부터 최하위 노드까지 하위로 연결된 모든 노드를 의미한다. 예를 들어, 도 1의 카테고리 C_1 의 후손 노드는 카테고리 C_2 , 카테고리 C_3 , 카테고리 C_4 , 카테고리 C_5 , 카테고리 C_6 및 카테고리 C_7 이다.
- [0025] 다음으로, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법을 설명한다.
- [0026] 도 2는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법을 도시한 순서도이다.
- [0027] 본 발명의 일 실시예는 오픈 디렉터리 프로젝트 기반 텍스트 분류기 생성 방법에 있어서, 미리 정의된 오픈 디렉터리 프로젝트의 각 카테고리에 저장된 텍스트 집합을 특징 벡터(feature vector)로 변환할 수 있다(S100). 다음으로 텍스트 분류기는 각 카테고리의 특징 벡터를 이용하여 조상 병합 특징 벡터(merge-centroid ancestor feature vector)를 생성하고(S110), 동일한 각 카테고리의 특징 벡터를 이용하여 후손 병합 특징 벡터(merge-centroid descendant feature vector)를 생성할 수 있다(S120). 이렇게 생성된 조상 병합 특징 벡터와 후손 병합 특징 벡터를 이용하여 각 카테고리에 조상 후손 병합 특징 벡터(merge-centroid ancestor-descendant feature vector)를 생성할 수 있으며(S130), 이 조상 후손 병합 특징 벡터를 이용하여 텍스트 분류기를 생성할 수 있다.
- [0028] 각 카테고리에 저장된 텍스트 집합을 특징 벡터로 변환하기 위하여(S100), 오픈 디렉터리 프로젝트에 속하는 모든 텍스트에 대한 단어 집합을 생성하고, 단어 집합을 이용하여 각 카테고리의 특징 벡터를 생성할 수 있다. 특징 벡터로 변환하는 방법(S100)에 대한 구체적인 방법은 다음에 도 3을 이용하여 설명한다.
- [0029] 조상 병합 특징 벡터를 생성하기 위하여(S110), 카테고리의 특징 벡터와 카테고리의 부모 카테고리의 조상 병합 특징 벡터를 이용할 수 있다. 이때, 부모 카테고리의 조상 병합 특징 벡터를 이용하므로, 해당 카테고리의 모든 조상 카테고리 정보를 이용하여 텍스트 분류기를 생성할 수 있게 된다. 조상 병합 특징 벡터 생성 방법(S110)에 대한 구체적인 방법은 다음에 도 4를 이용하여 설명한다.
- [0030] 후손 병합 특징 벡터를 생성하기 위하여(S120), 카테고리의 특징 벡터와 카테고리의 자식 카테고리의 후손 병합 특징 벡터를 이용할 수 있다. 이때, 자식 카테고리의 후손 병합 특징 벡터를 이용하므로, 해당 카테고리의 모든 후손 카테고리 정보를 이용하여 텍스트 분류기를 생성할 수 있게 된다. 후손 병합 특징 벡터 생성 방법(S120)에 대한 구체적인 방법은 다음에 도 5를 이용하여 설명한다.

[0031] 조상 후손 병합 특징 벡터를 생성하기 위하여(S130), 카테고리의 조상 병합 특징 벡터와 카테고리의 후손 병합 특징 벡터를 모두 이용할 수 있다. 이렇게, 텍스트 분류기를 생성하기 위하여 조상 병합 특징 벡터와 후손 병합 특징 벡터를 모두 이용하므로, 해당 카테고리의 학습 데이터뿐 아니라 조상 카테고리나 후손 카테고리로부터 학습 데이터를 확장하는 효과를 볼 수 있다. 조상 후손 병합 특징 벡터 생성 방법(S130)에 대한 구체적인 방법은 다음에 도 6을 이용하여 설명한다.

[0032] 다음으로, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기의 카테고리별 특징 벡터 생성 방법을 설명한다.

[0033] 도 3은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리별 특징 벡터 생성 방법을 도시한 순서도이다.

[0034] 본 발명의 오픈 디렉터리 프로젝트 기반 텍스트 분류기 생성 방법은 텍스트 집합을 특징 벡터로 변환하기 위하여, 오픈 디렉터리 프로젝트의 카테고리에 속하는 텍스트 집합에서 단어를 추출하는 단계(S200), 카테고리에 속하는 텍스트별로 단어 발생 빈도 벡터를 생성하는 단계(S210), 카테고리에 속하는 텍스트에 대한 단어 발생 빈도를 이용하여 특징 벡터를 생성하는 단계(S220)를 포함할 수 있다.

[0035] 본 발명의 텍스트 분류기 생성 방법은 오픈 디렉터리 프로젝트의 카테고리에 속하는 모든 텍스트에 대하여 단어를 추출하고 단어 집합을 생성할 수 있다. 이때, 단어 집합의 생성을 위하여 형태소 분석기(morphological analyzer) 등을 이용할 수 있다. 형태소 분석기는 어휘 사전을 이용하여 모든 텍스트의 형태소를 분석하는 것이다. 이렇게 생성된 오픈 디렉터리 프로젝트의 단어 집합을 특징 벡터 추출 시 활용된다.

[0036] 오픈 디렉터리 프로젝트의 단어 집합을 생성한 후, 이 단어 집합을 기반으로 카테고리별 단어 발생 빈도 벡터를 생성할 수 있다(S220). 예를 들어, 오픈 디렉터리 프로젝트에서 추출된 단어 집합이 $\{w_1, w_2, w_3, w_4, w_5\}$ 이고 임의의 카테고리에서 속하는 텍스트가 2개이며, 2개의 텍스트에서 단어 w_3 가 3번 단어 w_4 가 2번 추출되면, 해당 카테고리의 단어 발생 빈도 벡터는 [0, 0, 3, 2, 0]이 된다.

[0037] 본 발명의 텍스트 분류기 생성 방법은 카테고리에 속한 단어 각각의 평균 발생 빈도를 이용하여 특징 벡터를 생성하는(S230) 단계를 포함할 수 있다. 이때, 카테고리에 대한 특징 벡터를 생성을 위하여 수학적 식 1을 사용할 수 있다. 수학적 식 1에서 c_i 는 카테고리 i 를 의미하며, D_{c_i} 는 카테고리 i 에 속한 텍스트 집합, $\vec{d}$ 는 각 텍스트 집합에 포함된 단어의 발생 빈도 벡터, $|D_{c_i}|$ 는 카테고리 i 에 속하는 텍스트 집합의 크기를 의미한다. 수학적 식 1에서처럼 카테고리 i 에 대한 특징 벡터 $\vec{\mu}_i$ 는 카테고리 i 에 속하는 단어들의 평균 발생 빈도를 이용하여 생성될 수 있다. 예를 들어, 2개의 텍스트가 있는 카테고리의 단어 발생 빈도 벡터가 [0, 0, 3, 2, 0]이라면, 평균 발생 빈도 벡터는 [0, 0, 1.5, 1, 0]이므로, 이 카테고리의 특징 벡터는 [0, 0, 1.5, 1, 0]이 된다.

수학적 식 1

$$\vec{\mu}_i = \frac{1}{|D_{c_i}|} \sum_{d \in D_{c_i}} \vec{d}$$

[0038]

[0039] 다음은, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반 텍스트 분류기의 조상 병합 특징 벡터를 생성하는 방법을 설명한다.

[0040] 도 4는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 조상 병합 특징 벡터 생성 방법을 도시한 순서도이다.

[0041] 오픈 디렉터리 프로젝트에서 텍스트 분류기를 생성할 카테고리가 선택되면(S300), 선택된 카테고리의 특징 벡터

를 가져올 수 있다(S310). 다음으로 선택된 카테고리가 최상위 카테고리인지 판단하여(S320), 만약, 선택된 카테고리가 최상위 카테고리인 경우, 선택된 카테고리의 특징 벡터를 이용하여 조상 병합 특징 벡터로 생성할 수 있다(S321). 만약 선택된 카테고리가 최상위 카테고리가 아닌 경우, 선택된 카테고리의 부모 카테고리를 선택하고(S330), 부모 카테고리의 조상 병합 특징 벡터를 가져올 수 있다(S340). 선택된 카테고리의 특징 벡터와 선택된 카테고리의 부모 카테고리의 조상 병합 특징 벡터의 가중합을 이용하여, 선택된 카테고리의 조상 병합 특징 벡터를 생성할 수 있다(S350). 이때, 선택된 카테고리의 특징 벡터의 가중치를 선택된 카테고리의 부모 카테고리의 조상 병합 특징 벡터의 가중치보다 높게 설정할 수 있다.

[0042]

본 발명에서는 선택된 카테고리의 조상 병합 특징 벡터를 생성하기(S350) 위하여, 부모 카테고리의 조상 병합 특징 벡터를 이용하므로, 선택된 카테고리의 정보뿐만 아니라, 선택된 카테고리의 모든 조상 카테고리의 정보를 이용할 수 있다. 이때, 선택된 카테고리의 특징 벡터에 대한 가중치를 부모 카테고리의 조상 병합 특징 벡터의 가중치보다 높게 설정하는 것은 텍스트 분류기를 생성할 때 선택된 카테고리의 특징 벡터의 중요도를 높이기 위해서이다. 이렇게 현재 텍스트 분류기를 생성하는 카테고리의 정보를 조상 카테고리의 정보보다 중요하게 고려하여 현재 카테고리를 분류하는 텍스트 분류기의 정확도를 높일 수 있다.

[0043]

선택된 카테고리가 최상위 카테고리가 아닌 경우, 선택된 카테고리에 대한 조상 병합 특징 벡터를 생성하기 위

하여(S350), 아래의 수학식 2를 사용할 수 있다. 수학식 2에서 $\vec{\mu}_i$ 는 카테고리 c_i 의 특징 벡터이며,

$\|\vec{\mu}_i\|$ 는 특징 벡터의 크기를 나타낸다. 또한, $parent(c_i)$ 는 카테고리 c_i 의 부모 카테고리를 나타내고, $|parent(c_i)|$ 는 부모 카테고리의 개수를 나타낸다. 카테고리 c_i 의 부모 카테고리 중 하나인 c_j 의

조상 병합 특징 벡터는 $\vec{mca}_j$ 이며, $\|\vec{mca}_j\|$ 는 조상 병합 특징 벡터의 크기를 나타낸다.

수학식 2

$$\vec{mca}_i = \frac{1}{2} \left(\frac{\vec{\mu}_i}{\|\vec{\mu}_i\|} + \frac{1}{1 + |parent(c_i)|} \sum_{c_j \in parent(c_i)} \frac{\vec{mca}_j}{\|\vec{mca}_j\|} \right)$$

[0044]

[0045]

수학식 2에서는 카테고리 c_i 특징 벡터의 가중치를 1, 조상 병합 특징 벡터의 가중치는 1로 설정할 수 있다. 오픈 디렉터리 프로젝트는 트리 자료구조이므로 최상위 카테고리가 아닌 경우 부모 카테고리의 개수는 언제나 하나만 존재하고, 부모 카테고리의 크기를 나타내는 는 언제나 1이 된다. 그러므로 수학식 2에서 조상 병합 특징 벡터의 가중치는 0.5가 되어 항상 카테고리 c_i 의 가중치에 비하여 작게 설정된다.

[0046]

다음은, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반 텍스트 분류기의 후손 병합 특징 벡터를 생성하는 방법을 설명한다.

[0047]

도5은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 후손 병합 특징 벡터 생성 방법을 도시한 순서도이다.

[0048]

후손 병합 특징 벡터를 생성하기 위하여 오픈 디렉터리 프로젝트의 카테고리가 선택되면(S400), 선택된 카테고리의 특징 벡터를 가져올 수 있다(S410). 최하위 카테고리 여부를 판단하여(S420), 만약 선택된 카테고리의 특징 벡터가 최하위 카테고리인 경우, 선택된 카테고리의 특징 벡터를 후손 병합 특징 벡터로 하여 후손 병합 특징 벡터를 생성할 수 있다(S421). 만약 선택된 카테고리의 특징 벡터가 최하위 카테고리가 아닌 경우, 카테고리의 특징 벡터와 카테고리의 자식 카테고리의 평균 후손 병합 특징 벡터에 대한 가중합을 이용하여 카테고리의 후손 병합 특징 벡터를 생성할 수 있다(S460).

[0049] 자식 카테고리의 평균 후손 병합 특징 벡터를 생성하기(S450) 위하여 선택된 카테고리의 복수개의 자식 카테고리를 선택하고(S430), 복수개의 선택된 자식 카테고리의 후손 병합 특징 벡터를 가져올 수 있다. 다음으로 복수개의 선택된 자식 카테고리의 후손 병합 특징 벡터의 평균을 구하여 평균 후손 병합 특징 벡터를 생성할 수 있다(S460).

[0050] 본 발명에서는 선택된 카테고리의 후손 병합 특징 벡터를 생성하기(S460) 위하여, 자식 카테고리의 후손 병합 특징 벡터를 이용하므로, 선택된 카테고리의 정보뿐만 아니라, 선택된 카테고리의 모든 후손 카테고리의 정보를 이용할 수 있다. 이때, 선택된 카테고리의 특징 벡터의 가중치를 자식 카테고리의 평균 후손 병합 특징 벡터의 가중치보다 크게 설정할 수 있다. 가중치를 크게 설정하는 것은 텍스트 분류기를 생성할 때, 선택된 카테고리의 특징 벡터의 중요도를 평균 후손 병합 특징 벡터보다 높게 설정하기 위해서이다. 이렇게 현재 텍스트 분류기를 생성하는 카테고리의 정보를 후손 카테고리의 정보보다 중요하게 고려하여 현재 카테고리를 분류하는 텍스트 분류기의 정확도를 높일 수 있다.

[0051] 본 발명은 후손 병합 특징 벡터를 생성하기 위하여(S460) 아래의 수학적 식 3을 사용할 수 있다. $child(c_i)$ 은 카테고리 c_i 의 자식 카테고리 집합을 나타내며, $|child(c_i)|$ 은 카테고리 c_i 의 자식 카테고리 집합의 크기를 나타낸다. 또한, $\vec{mcd}_j$ 는 카테고리 c_i 의 자식 카테고리 중 하나인 카테고리 c_j 의 후손 병합 특징 벡터를 나타내며, $\|\vec{mcd}_j\|$ 는 후손 병합 특징 벡터의 크기이다.

수학적 식 3

$$\vec{mcd}_i = \frac{1}{2} \left(\frac{\vec{\mu}_i}{\|\vec{\mu}_i\|} + \frac{1}{1 + |child(c_i)|} \sum_{c_j \in child(c_i)} \frac{\vec{mcd}_j}{\|\vec{mcd}_j\|} \right)$$

[0052]

[0053] 수학적 식 3에서는 조상 병합 특징 벡터를 생성하기 위하여 카테고리 c_i 의 가중치를 1로 설정하며, 자식 카테고리의 평균 후손 병합 특징 벡터의 가중치를 $\frac{1}{1 + |child(c_i)|}$ 로 설정할 수 있다. 오픈 디렉터리 프로젝트는 트리 자료구조이므로 최하위 카테고리가 아닌 경우 자식 카테고리 수는 1 이상의 자연수가 되며, 최하위 카테고리인 경우 자식 카테고리 수는 0이 된다. 즉, 자식 카테고리의 크기의 범위는 $0 \leq |child(c_i)| \leq \infty$ 이다. 그러므로 자식 카테고리의 평균 후손 병합 특징 벡터의 가중치는 0.5보다 작다. 즉, 수학적 식 3에서 평균 후손 특징 벡터의 가중치는 선택된 카테고리의 가중치보다 언제나 작다. 수학적 식 3에서는 이렇게 평균 후손 특징 벡터의 가중치가 선택된 카테고리의 가중치보다 언제나 작게 설정된다.

[0054] 도 6은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 방법에서 카테고리의 조상 후손 병합 특징 벡터 생성 방법을 도시한 순서도이다.

[0055] 선택된 카테고리의 조상 후손 병합 특징 벡터를 생성하는 것은 카테고리를 선택하고(S500), 선택된 카테고리의 조상 병합 특징 벡터와 후손 병합 특징 벡터를 가져올 수 있다(S510, S520). 조상 병합 특징 벡터와 후손 병합 특징 벡터에 대한 가중합을 이용하여 카테고리의 조상 후손 병합 특징 벡터를 생성할 수 있다(S530). 이때, 조상 병합 특징 벡터의 가중치와 후손 병합 특징 벡터의 가중치를 서로 상이하게 설정할 수 있다.

[0056] 아래의 수학적 식 4를 이용하여 조상 후손 병합 특징 벡터를 생성할 수 있다(S530). 수학적 식 4에서 $\vec{mcad}_i$ 는

카테고리 c_i 의 조상 후손 병합 특징 벡터이며, α 는 후손 병합 특징 벡터의 가중치이다.

수학식 4

$$\vec{mca}_i = \frac{1}{1+\alpha} \left(\frac{\vec{mca}_i}{\|\vec{mca}_i\|} + \alpha \frac{\vec{mcd}_i}{\|\vec{mcd}_i\|} \right)$$

[0057]

[0058]

조상 병합 특징 벡터와 후손 병합특징 벡터의 가중 평균으로 계산된다. 수학식 4에서는 조상 병합 특징 벡터의 가중치를 1로, 후손 병합 특징 벡터의 가중치를 α 로 설정한다. 즉, 후손 병합 특징 벡터의 상대적인 가중치인 α 만을 이용하여, 조상 후손 병합 특징 벡터에서 조상 병합 특징 벡터와 후손 병합 특징 벡터의 중요도를 조정할 수 있다.

[0059]

도 7은 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기의 신규 텍스트 분류 방법을 도시한 순서도이다.

[0060]

본 발명의 오픈 디렉터리 프로젝트 환경에서 텍스트 분류기를 이용한 텍스트 분류 방법은 신규 텍스트가 입력되면(S600), 신규 텍스트를 특징 벡터로 변환하고(S610), 특징 벡터와 텍스트 분류기를 이용하여 신규 텍스트의 카테고리를 분류할 수 있다(S640). 이때, 본 발명의 텍스트 분류기는 오픈 프로젝트 디렉터리의 각 카테고리에 대한 특징 벡터, 해당 카테고리의 조상 병합 특징 벡터 및 후손 병합 특징 벡터를 이용하여 조상 후손 병합 특징 벡터를 생성하여 구축될 수 있다.

[0061]

신규 텍스트의 카테고리를 분류하기(S640) 위하여 텍스트 분류기는 신규 텍스트의 특징 벡터와 오픈 프로젝트 디렉터리에 포함된 각 카테고리의 조상 후손 병합 특징 벡터에 대하여 유사도를 계산할 수 있다(S630). 본 발명은 유사도를 계산하는 방법으로 코사인(cosine) 유사도, 피어슨 상관계수(Pearson's correlation coefficient), 스피어만 상관계수(Spearman's correlation coefficient), 자카드 계수(jaccard coefficient) 및 유클리드 거리(Euclidean distance) 등을 이용하여 유사도를 계산하는 것 등을 모두 포함할 수 있다.

[0062]

텍스트 분류기는 각 카테고리에 대하여 유사도를 계산한 후, 유사도가 가장 높은 카테고리를 선택하고(S630), 신규 텍스트를 선택된 카테고리으로 분류하는 것(S640)을 포함할 수 있다.

[0063]

도 8은 본 발명의 일 실시예에 따르는 오픈 디렉터리 프로젝트 기반의 텍스트 분류기를 개략적으로 도시한 구성도이다.

[0064]

텍스트 분류기(700)는 오픈 디렉터리 프로젝트의 카테고리에 기반하여 신규 텍스트의 카테고리를 분류하는 텍스트 분류 프로그램이 저장된 스토리지 장치(720) 및 스토리지 장치(720)에 저장된 텍스트 분류 프로그램을 실행하는 프로세서(710)를 포함할 수 있다.

[0065]

프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 미리 정의된 오픈 디렉터리 프로젝트 구조에 기반하여 카테고리의 특징 벡터, 카테고리의 조상 병합 특징 벡터 및 카테고리의 후손 병합 특징 벡터를 이용하여 조상 후손 특징 벡터를 생성할 수 있다. 생성된 조상 후손 병합 특징 벡터를 이용하여 텍스트 분류기를 생성하고, 텍스트 분류기를 이용하여 신규 텍스트의 카테고리를 분류할 수 있다.

[0066]

프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 오픈 디렉터리 프로젝트에 포함된 텍스트 집합에서 단어를 추출하고, 카테고리에 속한 단어에 대한 발생 빈도를 이용하여, 카테고리의 특징 벡터를 생성할 수 있다. 이때, 카테고리의 특징 벡터는 카테고리에 속한 단어에 대한 각각의 평균 발생 빈도를 이용하여 생성될 수 있다.

[0067]

프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 카테고리가 최상위 카테고리일 경우, 카테고리의 특징 벡터를 이용하여 카테고리의 조상 병합 특징 벡터를 생성할 수 있다. 카테고리가 최상위 카테고리가 아닌

경우, 카테고리의 특징 벡터와 카테고리의 조상 카테고리의 조상 병합 특징 벡터에 대한 가중합을 이용하여 카테고리의 조상 병합 특징 벡터를 생성할 수 있다. 이때, 카테고리의 특징 벡터 가중치를 부모 카테고리의 조상 병합 특징 벡터 가중치보다 높게 설정할 수 있다. 프로세서(710)는 텍스트 분류기 생성을 위하여 부모 카테고리의 조상 병합 특징 벡터를 이용하므로, 해당 카테고리의 모든 조상 카테고리의 정보를 이용할 수 있다. 또한, 해당 카테고리의 가중치를 부모 카테고리의 조상 병합 특징 벡터 가중치보다 높게 설정하므로, 해당 카테고리의 중요도를 높게 설정하여 텍스트 분류기의 분류 정확도를 향상할 수 있다.

[0068] 프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 카테고리가 최말단 카테고리일 경우, 카테고리의 특징 벡터를 이용하여 카테고리의 후손 병합 특징 벡터를 생성하고, 카테고리가 최말단 카테고리가 아닌 경우, 카테고리의 특징 벡터와 카테고리의 자식 카테고리의 평균 후손 병합 특징 벡터에 대한 가중합을 이용하여 카테고리의 후손 병합 특징 벡터를 생성할 수 있다. 이때, 카테고리의 특징 벡터의 가중치를 자식 카테고리의 평균 후손 병합 특징 벡터의 가중치보다 높게 설정할 수 있다.

[0069] 프로세서(710)는 텍스트 분류기 생성을 위하여 자식 카테고리의 평균 후손 병합 특징 벡터를 이용하므로, 해당 카테고리의 모든 후손 카테고리의 정보를 이용할 수 있다. 또한, 해당 카테고리의 가중치를 자식 카테고리의 평균 후손 병합 특징 벡터 가중치보다 높게 설정하므로, 해당 카테고리의 중요도를 높게 설정하여 텍스트 분류기의 분류 정확도를 향상할 수 있다.

[0070] 프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 조상 병합 특징 벡터와 후손 병합 특징 벡터에 대한 가중합을 이용하여 해당 카테고리의 조상 후손 병합 특징 벡터를 생성할 수 있다. 이때, 조상 병합 특징 벡터 및 후손 병합 특징 벡터의 상대적인 중요도를 기반으로 조상 병합 특징 벡터의 가중치와 후손 병합 특징 벡터의 가중치를 상이하게 설정할 수 있다. 이렇게 생성된 조상 후손 병합 특징 벡터를 기반으로 텍스트 분류기를 생성하고, 스토리지 장치(720)에 저장할 수 있다.

[0071] 프로세서(710)는 텍스트 분류 프로그램의 실행에 따라, 신규 텍스트의 카테고리 분류를 위하여 신규 텍스트의 특징 벡터와 오픈 디렉터리 프로젝트의 각 카테고리의 조상 후손 병합 특징 벡터에 대하여 유사도를 계산한 후, 유사도가 가장 높은 카테고리를 신규 텍스트의 카테고리로 분류할 수 있다.

[0072] 다음으로, 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반 텍스트 분류기 생성 과정을 설명한다.

[0073] 도 9는 본 발명의 일 실시예에 따른 오픈 디렉터리 프로젝트 기반의 텍스트 분류기 생성 과정을 설명하기 위한 오픈 디렉터리 프로젝트 예시도이다.

[0074] 도 9와 같이 오픈 디렉터리 프로젝트의 카테고리별로 특징 벡터가 주어졌을 때, 본 발명의 일 실시예에 따른 카테고리 C₃의 특징 벡터는 [0, 0, 1, 0, 0, 0, 0], 조상 병합 특징 벡터는 [0.25, 0, 0.5, 0, 0, 0, 0]이다. 또한, 카테고리 C₃의 후손 병합 특징 벡터는 [0, 0, 0.5, 0, 0, 0.167, 0.167]이 된다. 이때, 조상 병합 특징 벡터의 가중치인 α 를 1로 설정하면 카테고리 C₃의 조상 후손 병합 특징 벡터는 [0.223, 0, 0.899, 0, 0, 0.15, 0.15]이 되며, 가중치인 α 를 2로 설정하면, [0.15, 0, 0.90, 0, 0, 0.20, 0.20]이 된다. 이렇게 가중치에 따라 조상 병합 특징 벡터와 후손 병합 특징 벡터의 상대적 가중치를 상이하게 설정할 수 있다. 또한, 조상 후손 병합 특징 벡터에는 C₃의 특징 벡터에 포함되지 않은 정보에 대한 가중치가 포함되어 있는 것을 확인할 수 있다.

[0075] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 메커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.

[0076] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0077]

전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0078]

본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

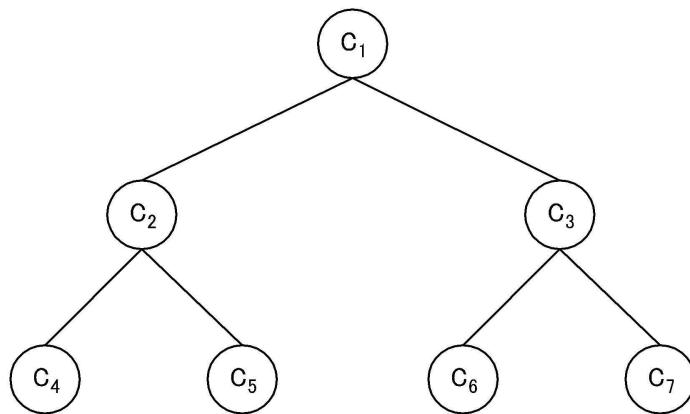
[0079]

700: 텍스트 분류기 710: 프로세서

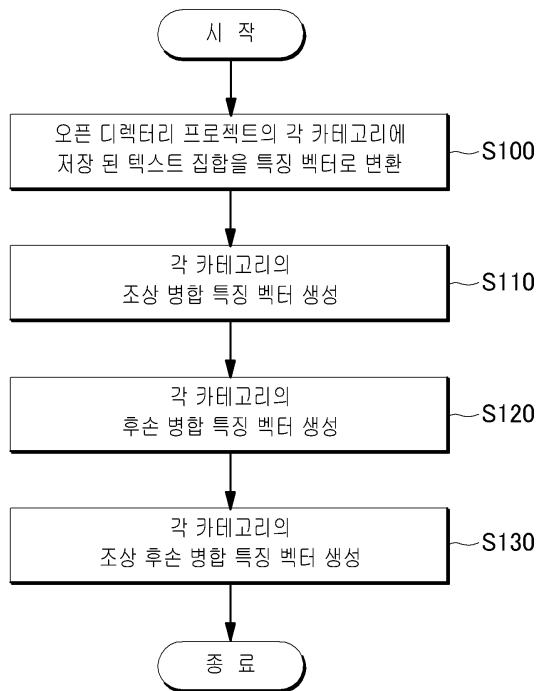
720: 스토리지 장치

도면

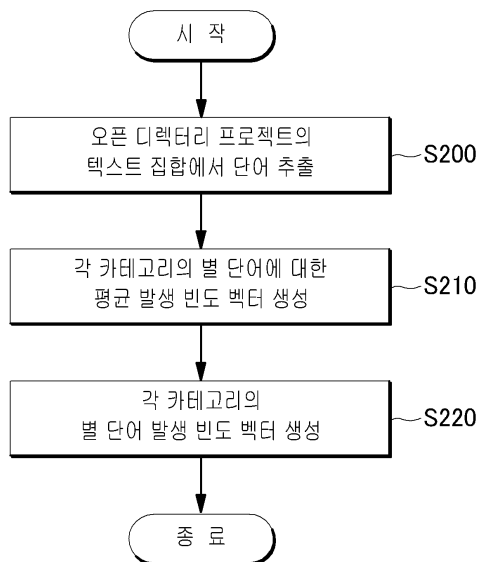
도면1



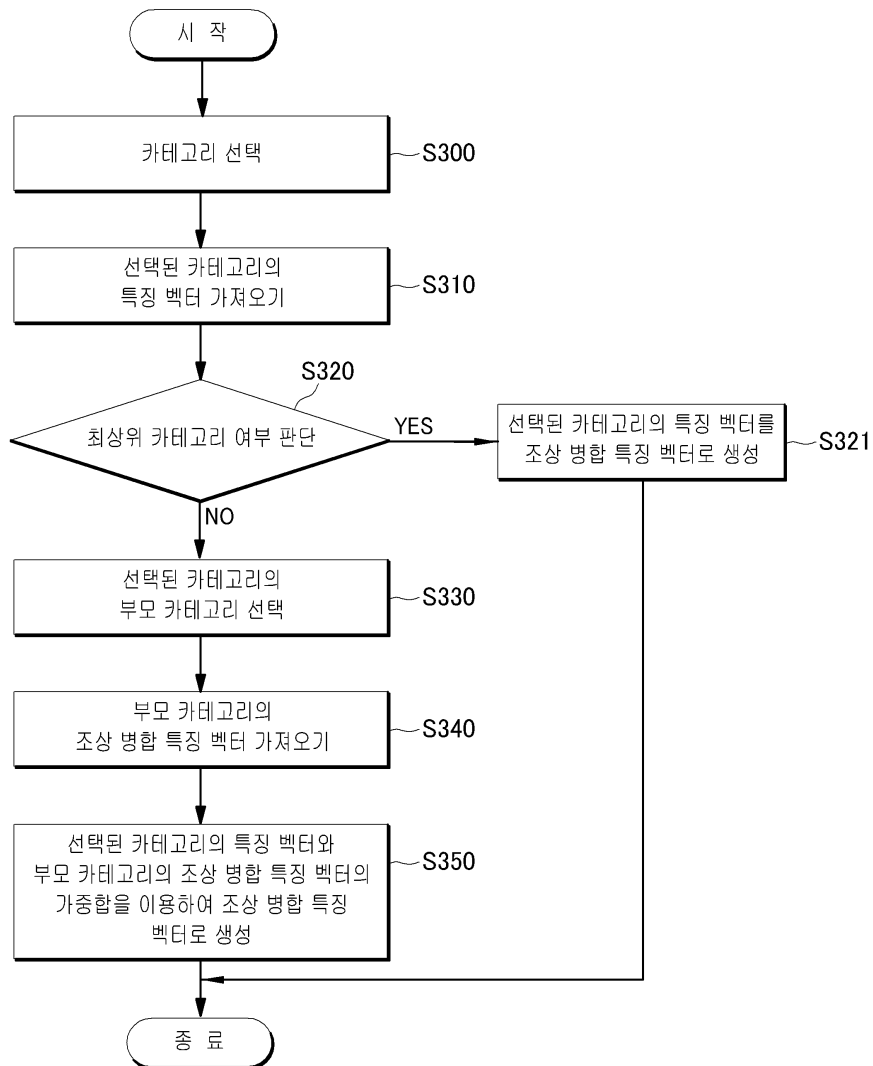
도면2



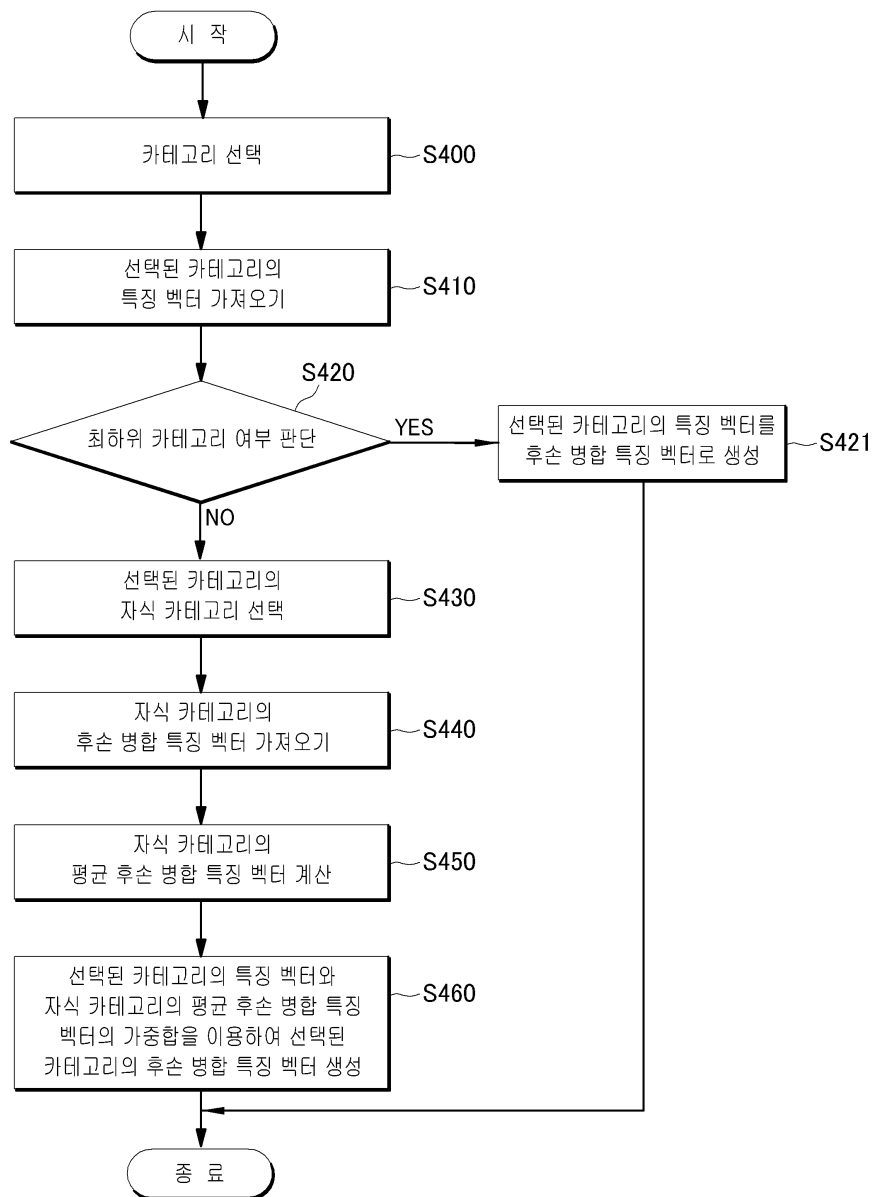
도면3



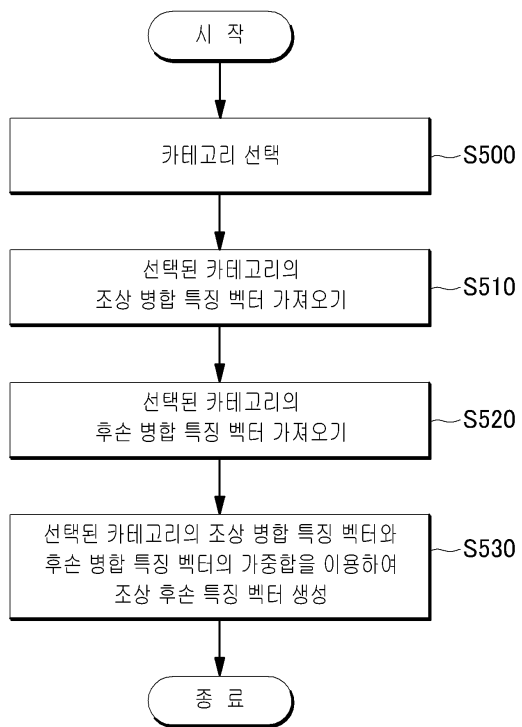
도면4



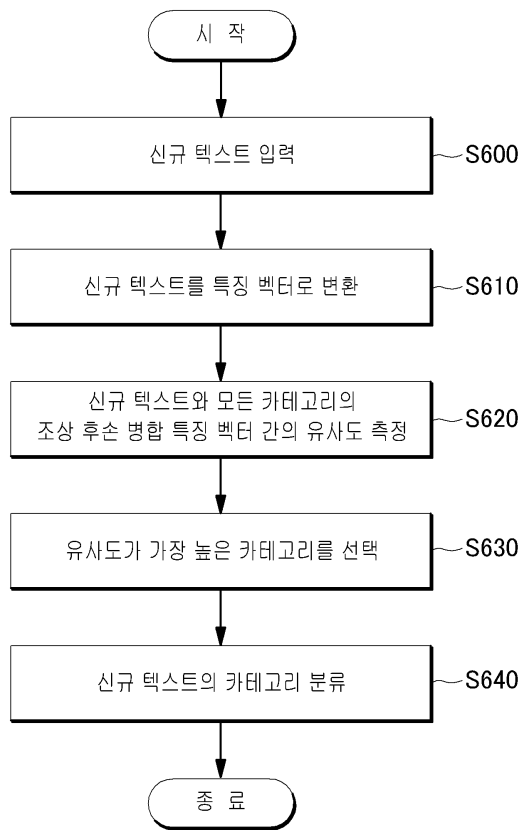
도면5



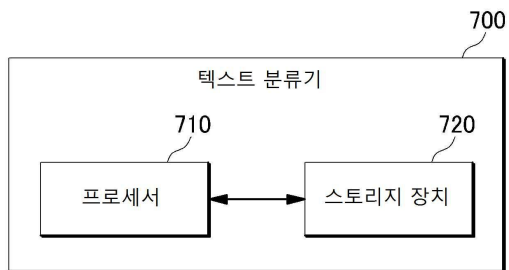
도면6



도면7



도면8



도면9

