



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0052821
(43) 공개일자 2011년05월19일

(51) Int. Cl.

G06F 17/00 (2006.01) G06Q 30/00 (2006.01)

(21) 출원번호 10-2009-0109501

(22) 출원일자 2009년11월13일

심사청구일자 2009년11월13일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

이상근

서울특별시 동대문구 제기동 153 안암골 벽산아파트 101동 203호

이정현

서울특별시 영등포구 대림3동 현대1차아파트 102동 1905호

하중우

서울특별시 성북구 안암동1가 149-1번지 404호

(74) 대리인

특허법인 엠에이피에스

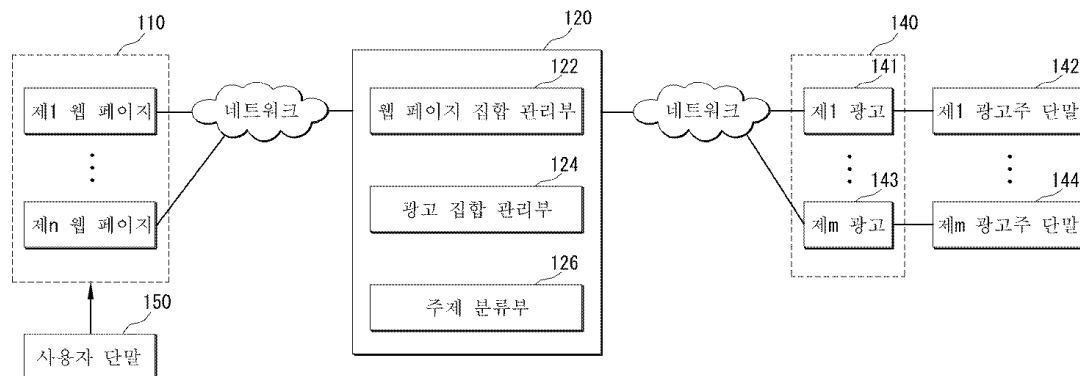
전체 청구항 수 : 총 14 항

(54) 주제 분류 모듈 및 그를 이용한 문맥 광고 시스템

(57) 요약

공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성하는 주제 분류 트리 생성부, 상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성하는 학습 데이터 생성부 및 상기 학습 데이터 생성부를 통해 생성된 디렉토리별 학습 데이터를 상기 주제 분류 트리 생성부를 통해 생성된 상기 주제 분류 트리에 포함된 디렉토리에 대응시킨 분류부를 포함하되, 상기 분류부는 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 주제 분류 모듈을 제공한다.

대표도



특허청구의 범위

청구항 1

주제 분류 모듈에 있어서,

공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성하는 주제 분류 트리 생성부,

상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성하는 학습 데이터 생성부 및

상기 학습 데이터 생성부를 통해 생성된 디렉토리별 학습 데이터를 상기 주제 분류 트리 생성부를 통해 생성된 상기 주제 분류 트리에 포함된 디렉토리에 대응시킨 분류부를 포함하되,

상기 분류부는 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 주제 분류 모듈.

청구항 2

제 1 항에 있어서,

상기 주제 분류 트리 생성부는 상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 주제를 나타내지 않는 디렉토리를 제거하는 것인 주제 분류 모듈.

청구항 3

제 1 항에 있어서,

상기 주제 분류 트리 생성부는 상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 임계값보다 작은 개수의 웹 페이지를 포함하는 디렉토리를 제거하는 것인 주제 분류 모듈.

청구항 4

제 1 항에 있어서,

상기 주제 분류 트리 생성부는 상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 각 디렉토리의 깊이에 대한 최하위 디렉토리의 깊이의 비율이 임계값보다 큰 경우 해당 디렉토리를 제거하는 것인 주제 분류 모듈.

청구항 5

제 1 항에 있어서,

상기 주제 분류 트리 생성부는 상기 공개된 디렉토리 데이터에 포함된 디렉토리에 대하여, 각 디렉토리의 웹 페이지에 대한 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하고,

부모 디렉토리와 자식 디렉토리의 센트로이드 벡터 사이의 코사인 유사도를 산출하여, 그 값이 임계값 이상인 경우에는 상기 자식 디렉토리를 제거하는 것인 주제 분류 모듈.

청구항 6

제 1 항에 있어서,

상기 학습 데이터 생성부는 상기 웹 페이지의 텍스트 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하고

최하위 디렉토리부터 자식 디렉토리의 센트로이드 벡터와 부모 디렉토리의 센트로이드 벡터간의 벡터 합을 구하여, 각 부모 디렉토리의 병합 센트로이드 벡터를 산출하는 것인 주제 분류 모듈.

청구항 7

제 6 항에 있어서,

상기 분류부는 상기 산출된 병합 센트로이드 벡터를 추출하여 상기 주제 분류 트리에 포함된 디렉토리에 대응시키는 것인 주제 분류 모듈.

청구항 8

제 7 항에 있어서,

상기 분류부는 상기 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 센트로이드 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 주제 분류 모듈.

청구항 9

문맥 광고 시스템에 있어서,

광고가 게재될 하나 이상의 웹 페이지를 관리하는 웹 페이지 집합 관리부,

웹 페이지에 게재될 하나 이상의 광고를 관리하는 광고 집합 관리부 및

상기 웹 페이지의 주제와 상기 광고의 주제의 유사도에 기초하여 상기 광고와 상기 웹 페이지를 매칭시켜, 상기 광고가 상기 웹 페이지 상에서 실행되도록 하는 주제 분류 모듈을 포함하되,

상기 주제 분류 모듈은 상기 제 1 항 내지 제 8 항 중 어느 한 항에 해당하는 주제 분류 모듈인 것인 문맥 광고 시스템.

청구항 10

웹 페이지 또는 광고의 주제별 분류 방법에 있어서,

공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성하는 단계,

상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성하는 단계,

상기 디렉토리별 학습 데이터를 상기 주제 분류 트리에 포함된 디렉토리에 대응시켜 분류부를 생성하는 단계 및

상기 생성된 분류부를 이용하여 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 단계

를 포함하는 웹 페이지 또는 광고의 주제별 분류 방법.

청구항 11

제 10 항에 있어서,

상기 주제 분류 트리를 생성하는 단계는

상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 주제를 나타내지 않는 디렉토리를 제거하는 제 1 단계,

상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 임계값보다 작은 개수의 웹 페이지를 포함하는 디렉토리를 제거하는 제 2 단계,

상기 공개된 디렉토리 데이터에 포함된 디렉토리들 중 각 디렉토리의 깊이에 대한 최하위 디렉토리의 깊이의 비율이 임계값보다 큰 경우 해당 디렉토리를 제거하는 제 3 단계 및

상기 공개된 디렉토리 데이터에 포함된 디렉토리에 대하여, 각 디렉토리의 웹 페이지에 대한 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하고, 부모 디렉토리와 자식 디렉토리의 센트로이드 벡터 사이의 코사인 유사도를 산출하여, 그 값이 임계값 이상인 경우에는 상기 자식 디렉토리를 제거하는 제 4 단계 중 하나 이상의 단계를 포함하는 것인 웹 페이지 또는 광고의 주제별 분류 방법.

청구항 12

제 10 항에 있어서,

상기 학습 데이터를 생성하는 단계는,

상기 웹 페이지의 텍스트 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하는 단계 및

최하위 디렉토리부터 자식 디렉토리의 센트로이드 벡터와 부모 디렉토리의 센트로이드 벡터간의 벡터 합을 구하여, 각 부모 디렉토리의 병합 센트로이드 벡터를 산출하는 단계

를 포함하는 것인 웹 페이지 또는 광고의 주제별 분류 방법.

청구항 13

제 12 항에 있어서,

상기 분류부를 생성하는 단계는,

상기 산출된 병합 센트로이드 벡터를 추출하여 상기 주제 분류 트리에 포함된 디렉토리에 대응시키는 것인 웹 페이지 또는 광고의 주제별 분류 방법.

청구항 14

제 13 항에 있어서,

상기 주제를 결정하는 단계는,

상기 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 센트로이드 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 웹 페이지 또는 광고의 주제별 분류 방법.

명 세 서

발명의 상세한 설명

기술 분야

본 발명은 주제 분류 모듈 및 그를 이용한 문맥 광고 시스템에 관한 것이다.

[0001]

배정 기술

- [0002] 최근 들어 인터넷을 이용한 광고의 비중이 날로 커짐에 따라, 웹 광고는 인터넷 비즈니스의 큰 수익원으로서 자리매김하고 있다. 이러한 웹 광고는 스폰서 광고와 문맥 광고의 두 가지의 형태로 나누어진다. 스폰서 광고는 웹 검색 엔진을 포함하는 포털 사이트에서 주로 이용되는 것으로, 질의에 대한 검색 결과와 함께 질의와 연관된 텍스트 형태의 광고를 배치하는 형태로 광고를 수행한다. 반면에, 문맥 광고는 웹 페이지에서 웹 페이지의 내용과 연관된 광고를 배치하는 형태로 수행된다. 검색 결과에 광고를 배치하는 스폰서 광고 보다 일반 웹 페이지에 광고를 배치하는 문맥 광고의 광고 노출 횟수가 상대적으로 많기 때문에 문맥 광고에 대한 관심이 증가하는 추세이다.
- [0003] 문맥 광고에 참여하는 구성원은 광고 중개자, 광고주(Advertiser), 웹 페이지 제작자(Publisher), 웹 사용자(User)로 구분될 수 있다. 광고 중개자는 광고주의 광고를 수집하고, 웹 페이지 제작자의 웹 페이지에 적절한 광고를 배치하여, 웹 사용자가 해당 웹 페이지를 방문하였을 때, 광고 클릭을 유도하게 된다. 광고의 클릭이 발생하면, 광고주는 사전에 채택한 클릭 당 비용(Pay Per Click)을 광고 중개자에게 지불하고, 광고 중개자는 이 수익을 클릭된 광고가 게재된 웹 페이지 제작자와 공유한다.
- [0004] 이러한 문맥 광고의 환경에서는 웹 페이지의 내용과 광고 내용의 의미적 연관성이 가장 중요하다. 의미적 연관성이 높아질수록 웹 사용자의 광고 클릭율이 높아지고, 이에 광고 중개자와 웹 페이지 제작자의 수익이 향상된다. 또한, 웹 사용자는 자신이 원하는 정보를 획득할 수 있는 추가적인 기회를 제공 받으며, 이에 따라 광고주는 더 높은 광고 효과를 기대할 수 있다.
- [0005] 웹 페이지의 내용과 광고 내용의 의미적 연관성을 높이기 위한 연구들이 계속되어 왔다. 초기에 제안된 방법들은 벡터 스페이스 모델을 이용한 웹 페이지와 광고의 키워드 매칭을 활용하였으나, 이 매칭 방법에서는 단어들의 다의성으로 인해 의미적 매칭의 정확도가 낮았다. 이 문제를 해결하기 위해, 웹 페이지와 광고의 주제를 분류하기 위한 주제 분류 트리 및 분류기를 이용하여 웹 페이지와 광고의 의미적 매칭을 하는 방법이 제안되었으며, 기존의 키워드 매칭과 의미적 매칭의 조합이 정확도를 향상시킬 수 있었다. 그러나, 여기서 사용된 주제 분류 트리 및 분류기의 생성에는 많은 사람들의 노력이 필요하기 때문에 큰 생성 비용이 요구 된다. 주제 분류 트리란, 모든 웹 페이지와 광고의 주제들을 트리 형태로 구성한 것으로서, 각 주제들은 "is-a" 관계를 가지고 연결되는데, 모든 주제들을 정의하고 그들의 관계를 연결하기 위해선 각 분야의 다수의 전문가들의 장기간에 걸친 노력이 필요하다. 분류기란, 생성된 주제 분류 트리의 주제들 중 웹 페이지와 광고의 내용에 가장 적합한 주제를 자동으로 결정해주는 것으로서, 정확한 분류기를 만들기 위해선 사람이 손수 정의해 놓은 많은 양의 학습 데이터가 필요하다.
- [0006] 또한, 문맥 광고의 웹 페이지와 광고의 의미적 매칭을 위해 생성된 주제 분류 트리 및 분류기는 다음과 같은 특성을 만족해야 한다. 첫째, 주제 분류 트리는 웹 페이지와 광고에 대한 모든 범위의 주제를 포함하는 크기여야 한다. 그러나 너무 많은 주제를 포함하면 분류기의 정확도가 하락할 수 있기 때문에 높은 정확도를 유지할 수 있는 적절한 크기여야 한다. 둘째, 주제 분류 트리의 각 주제들은 구체성을 지녀야 한다. 너무 상세하면 분류기의 정확도가 하락할 수 있고, 너무 일반적이면 웹 페이지와 광고의 의미적 매칭의 정확도가 하락할 수 있기 때문에 적절한 구체성을 지녀야 한다. 셋째, 주제 분류 트리의 각 주제들은 서로 간의 차별성을 가져야 한다. 주제 분류 트리내의 주제들이 차별성이 없다면, 분류기의 정확도가 하락할 수 있기 때문에 서로 간의 적절한 차별성을 지녀야 한다. 이러한 특성들을 만족시키고 생성 과정에서 사람의 노력을 최소화하는 주제 분류 트리 및 분류기의 생성 방법이 요구된다.

발명의 내용

해결 하고자하는 과제

- [0007] 본 발명의 일 실시예는 공개된 디렉토리 데이터를 적절히 가공하여 생성한 주제 분류 모듈 및 그를 이용한 문맥 광고 시스템을 제공한다.

과제 해결수단

- [0008] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 제 1 측면은 공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성하는 주제 분류 트리 생성부, 상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성하는 학습 데이터 생성부 및 상기 학습 데이터 생성부를 통해 생성된 디렉토리별 학습 데이터를 상기 주제 분류 트리 생성부를 통해 생성된 상기 주제 분류 트리에 포함된 디렉토리에 대응시킨 분류부를 포함하되, 상기 분류부는 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 주제 분류 모듈을 제공한다.
- [0009] 또한, 본 발명의 제 2 측면은 광고가 게재될 하나 이상의 웹 페이지를 관리하는 웹 페이지 집합 관리부, 웹 페이지에 게재될 하나 이상의 광고를 관리하는 광고 집합 관리부 및 상기 웹 페이지의 주제와 상기 광고의 주제의 유사도에 기초하여 상기 광고와 상기 웹 페이지를 매칭시켜, 상기 광고가 상기 웹 페이지 상에서 실행되도록 하는 주제 분류 모듈을 포함하는 문맥 광고 시스템을 제공한다.
- [0010] 또한, 본 발명의 제 3 측면은 공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성하는 단계, 상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성하는 단계, 상기 디렉토리별 학습 데이터를 상기 주제 분류 트리에 포함된 디렉토리에 대응시켜 분류부를 생성하는 단계 및 상기 생성된 분류부를 이용하여 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 상기 분류부의 디렉토리를 대표하는 벡터 간의 유사도를 산출하여 상기 웹 페이지 또는 광고의 주제를 결정하는 단계를 포함하는 웹 페이지 또는 광고의 주제별 분류 방법을 제공한다.

효 과

- [0011] 전술한 본 발명의 과제 해결 수단에 의하면, 공개된 디렉토리 데이터를 이용하여 분류부를 생성하므로, 분류부 생성에 소요되는 비용을 최소화할 수 있다. 또한, 공개된 디렉토리 데이터는 방대한 디렉토리 및 웹 페이지를 포함하고 있어, 이를 이용하여 분류부를 생성하는 경우 웹 페이지 또는 광고의 주제 분류 성능을 획기적으로 개선할 수 있다.

발명의 실시를 위한 구체적인 내용

- [0012] 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.
- [0013] 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.
- [0014] 도 1은 본 발명에 적용되는 문맥 광고 시스템을 도시한 도면이다.
- [0015] 본 발명의 문맥 광고 시스템(120)은 웹 페이지 제작자에 의하여 제공되는 웹 페이지 집합(110)에 포함되는 임의의 웹 페이지와 광고주에 의하여 제공되는 광고 집합(140)에 포함되는 임의의 광고를 매칭시키고, 매칭된 광고가 해당 웹 페이지에 노출되도록 한다.
- [0016] 웹 페이지 집합(110)에 포함되는 웹 페이지는 업무용 웹 사이트뿐만 아니라 비업무용 웹 사이트들 통해 접속되는 모든 웹 페이지가 포함될 수 있다.
- [0017] 광고 집합(140)에 포함되는 광고(141, 143)는 광고주 단말(142, 144) 등을 통해, 문맥 광고 시스템(120)에 전달된다. 이와 같은 광고는 상기 웹 페이지에 하나 이상 게재될 수 있으며, 다른 광고 매칭 시스템에 의하여 게재된 광고와 함께 게재될 수 도 있다.
- [0018] 사용자들은 사용자 단말(150)을 통해 웹 페이지에 접속하며, 웹 페이지에 게시된 광고를 클릭함에 따라 광고에 노출되게 된다.

- [0019] 이때, 웹 페이지 집합(110)의 각 웹 페이지, 문맥 광고 시스템(120), 광고 집합(140)의 각 광고 또는 사용자 단말(150)은 네트워크를 통해 접속되며, 네트워크는 근거리 통신망(Local Area Network; LAN), 광역 통신망(Wide Area Network; WAN) 또는 부가가치 통신망(Value Added Network; VAN) 등과 같은 유선 네트워크나 이동 통신망(mobile radio communication network) 또는 위성 통신망 등과 같은 모든 종류의 무선 네트워크로 구현될 수 있다.
- [0020] 또한, 광고주 단말(142, 144), 사용자 단말(150)은 컴퓨터나 휴대용 단말기로 구현될 수 있다. 여기서, 컴퓨터는 예를 들어, 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop) 등을 포함하고, 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, PCS(Personal Communication System), GSM(Global System for Mobile communications), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access)-2000, W-CDMA(W-Code Division Multiple Access), Wibro(Wireless Broadband Internet) 단말 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치를 포함할 수 있다.
- [0021] 다음으로, 문맥 광고 시스템(120)은 광고가 게재될 하나 이상의 웹 페이지를 관리하는 웹 페이지 집합 관리부(122), 웹 페이지에 게재될 하나 이상의 광고를 관리하는 광고 집합 관리부(124), 광고 또는 웹 페이지의 주제를 분류하여, 광고 집합에 포함되는 광고를 웹 페이지에 매칭시키거나 웹 페이지 집합에 포함되는 웹 페이지를 광고에 매칭시키는 주제 분류부(126)를 포함한다.
- [0022] 웹 페이지 집합 관리부(122)는 웹 사이트의 운영자 등에 의하여 광고 게재 요청이 접수된 웹 페이지에 대한 정보를 관리한다. 예를 들면, 웹 페이지의 주소, 웹 페이지 상에서 광고가 게재될 위치, 웹 페이지의 내용 등에 대한 정보를 관리한다. 특히, 본 발명에서는 웹 페이지와 광고의 의미적인 관련성에 기초하여 유사도를 판단하므로, 웹 페이지에 포함되는 텍스트 정보를 각 웹 페이지 별로 관리한다.
- [0023] 광고 집합 관리부(124)는 광고주 등에 의하여 광고 게재 요청이 접수된 광고에 대한 정보를 관리한다. 예를 들면, 클릭당 광고료, 광고 기간, 광고의 내용 등에 대한 정보를 관리한다. 특히, 본 발명에서는 웹 페이지와 광고의 의미적인 관련성에 기초하여 유사도를 판단하므로, 광고에 포함되는 텍스트 정보를 각 광고 별로 관리한다.
- [0024] 주제 분류부(126)는 웹 페이지와 광고의 의미적인 관련성에 기초하여 웹 페이지 또는 광고의 주제를 분류하고, 적절한 광고를 각 웹 페이지에 매칭시켜 광고가 노출되도록 한다. 도면을 통해 광고 매칭부(126)의 상세 구성을 살펴보기로 한다.
- [0025] 도 2는 본 발명의 일 실시예에 따른 주제 분류 모듈(장치)의 상세 구성을 도시한 도면이다.
- [0026] 주제 분류 모듈(126)은 공개 디렉토리 데이터 수집부(130), 주제 분류 트리 생성부(132), 학습 데이터 생성부(134), 분류부(136)를 포함한다. 참고로, 본 발명의 실시예에 따른 도 1 및 도 2에 도시된 구성 요소들은 소프트웨어 또는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)와 같은 하드웨어 구성 요소를 의미하며, 소정의 역할들을 수행한다.
- [0027] 그렇지만 '구성 요소들'은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, 각 구성 요소는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다.
- [0028] 따라서, 일 예로서 구성 요소는 소프트웨어 구성 요소들, 객체지향 소프트웨어 구성 요소들, 클래스 구성 요소들 및 태스크 구성 요소들과 같은 구성 요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.
- [0029] 구성 요소들과 해당 구성 요소들 안에서 제공되는 기능은 더 작은 수의 구성 요소들로 결합되거나 추가적인 구성 요소들로 더 분리될 수 있다.
- [0030] 공개 디렉토리 데이터 수집부(130)는 공개되어 있는 각종 디렉토리 및 각 디렉토리에 포함되는 웹 페이지 정보를 수집한다. 예를 들면, 오픈 디렉토리 프로젝트(ODP, Open Directory Project, <http://www.dmoz.org/>)의 데이터 또는 각종 포털 사이트에 게시된 공개 디렉토리 데이터를 마이닝한다. 오픈 디렉토리 프로젝트란, 인간의 손에 의해 편집되는 가장 포괄적인 웹 디렉토리로서, 세계규모의 거대한 자원 편집 커뮤니티에 의해 구축 및 관리되고 있다. 오픈 디렉토리 프로젝트에서 공개하는 데이터에는 웹 사이트를 분류하기 위한 디렉토리의 구조와

각 디렉토리마다 사람에 의해 분류된 웹 사이트들의 리스트가 포함된다.

- [0031] 도 3은 본 발명의 일 실시예에 따른 공개 디렉토리로서 사용되는 오픈 디렉토리 프로젝트의 디렉토리 구조를 나타내며, 도 4는 오픈 디렉토리 프로젝트의 디렉토리에 포함되는 웹 사이트에 대한 정보를 나타낸다.
- [0032] 도 3에 도시된 바와 같이, 디렉토리 구조는 트리 형태로 이루어져 있으며, 전체 디렉토리의 수는 약 59만개, 최대 트리 깊이는 15레벨이며, 최상위 레벨의 디렉토리의 개수는 16개이다. 또한, 도 4에 도시된 바와 같이, 사람의 손에 의해 분류된 웹 사이트들의 정보는 웹 사이트 주소(URL), 제목(Title), 설명(Description)을 포함한다. 전체 디렉토리에 분류된 웹 사이트들의 개수는 약 460만개에 이른다. 이와 같이, 공개된 디렉토리 데이터에 포함된 디렉토리 구조 및 각 디렉토리에 포함된 웹 사이트들의 정보를 이용하여 주제 분류 모듈을 생성한다.
- [0033] 다시 도 2를 참조하면, 주제 분류 트리 생성부(132)는 상기 공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성한다. 본 발명에서는 주제 분류 트리 생성을 위해 네 가지 필터링 단계를 거칠 수 있다.
- [0034] 첫 번째 단계에서는, 공개된 디렉토리 데이터의 전체 디렉토리 중에서 주제를 나타내지 않는 디렉토리를 제거한다. 예를 들어, 최상위 디렉토리 중 "World", "Regional" 디렉토리를 제거한다. 이 두 디렉토리는 다른 디렉토리와 중복이 되고, 또한 다양한 언어로 구성되어 있기 때문에 제거한다. 여기서 주제 분류 트리의 대상 언어에 따라 "World" 디렉토리의 자식 디렉토리를 선별하여 적용할 수 있다. 또한, 단순히 알파벳(A, B, C, ...)이나 숫자(1, 2, 3, ...)를 디렉토리 이름으로 사용하는 중간 디렉토리들을 제거하고, 그 서브 트리를 상위 디렉토리에 연결한다. 또한, 특정 주제를 나타내지 않고 각 디렉토리마다 자주 반복되는 리프 디렉토리들을 제거한다.
- [0035] 두 번째 단계에서는, 공개된 디렉토리 데이터에 포함된 디렉토리들 중 임계값보다 작은 개수의 웹 페이지를 포함하는 디렉토리를 제거한다.
- [0036] 도 5는 본 발명의 일 실시예에 따라 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0037] 도시된 바와 같이, 각 디렉토리(D_i)에 포함된 웹 페이지의 개수(P_i)가 임계값(P_{url}) 보다 적은 경우, 해당 디렉토리를 제거한다.
- [0038] 세 번째 단계에서는, 일반성과 구체성을 적정한 수준에서 보장하는 주제를 선별하기 위해, 트리의 모든 리프 노드까지 이르는 각각의 경로에서 주어진 파라미터(P_{dep}) 비율로 하위노드들을 제거한다.
- [0039] 도 6은 본 발명의 일 실시예에 따라 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0040] 수식에 기재된 바와 같이, 각 디렉토리의 깊이($dep(D_i)$)에 대한 최하위 디렉토리의 깊이($maxdep(D_i)$)의 비율이 임계값(P_{dep}) 보다 큰 경우 해당 디렉토리를 제거한다. 도시된 예에서, 디렉토리(D_2)의 깊이는 2 이고, 최하위 디렉토리(D_3)의 깊이는 3 이 된다. 임계값(P_{dep})이 0.8이라고 가정하면, $dep(D_2)/maxdep(D_2)=2/3$ 이고, 이는 임계값(P_{dep})보다 작으므로, 해당 디렉토리(D_2)는 유지시킨다. 그러나, $dep(D_3)/maxdep(D_3)=3/3$ 이고, 이는 임계값(P_{dep})보다 크므로, 해당 디렉토리(D_3)는 제거한다.
- [0041] 네 번째 단계에서는, 각 주제들 간의 충분한 차별성을 보장하여 분류기의 정확도를 향상시키기 위해, 각 디렉토리의 웹 페이지에 벡터 스페이스 모델을 적용한다.
- [0042] 도 7은 본 발명의 일 실시예에 따라 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0043] 수식에 기재된 바와 같이, 각 디렉토리의 웹 페이지에 대한 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터(centroid vector)를 산출하고, 부모 디렉토리와 자식 디렉토리의 센트로이드 벡터 사이의 코사인 유사도를 산출하여, 그 값이 임계값 이상인 경우에는 상기 자식 디렉토리를 제거한다. 이때, 코사인 유사도는 수학적 1을 기초로 산출된다.
- [0044] [수학식 1]

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

[0045]

[0046] 이와 같이, 상기 네 가지 단계에 따라 디렉토리가 제거되고 남은 디렉토리를 주제 분류 트리의 각 노드로 변환

하고, 디렉토리 이름은 노드 이름으로 정하며, 디렉토리 간의 부모-자식 관계를 각 노드 사이의 간선으로 변환하여 주제 분류 트리를 생성할 수 있다. 이때, 상기 디렉토리를 필터링하는 단계들 중 하나 이상을 선택적으로 적용할 수 있다.

[0047] 다시 도 2를 참조하면, 학습 데이터 생성부(134)는 상기 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성한다.

[0048] 도 8은 본 발명의 일 실시예에 따른 학습 데이터 생성 단계를 수식화한 도면이다.

[0049] 학습 데이터 생성부(134)는 웹 페이지의 텍스트 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하고, 최하위 디렉토리부터 자식 디렉토리의 센트로이드 벡터와 부모 디렉토리의 센트로이드 벡터간의 벡터 합을 구하여, 각 부모 디렉토리의 병합 센트로이드 벡터(merge centroid vector)를 산출한다. 앞서 설명한 바와 같이, 각 웹 페이지의 텍스트 정보는 웹 사이트 주소(URL), 제목(Title), 설명(Description)을 포함하므로, 이를 기초로 단어 벡터로 변환한다. 한편, 단어 벡터 변환 및 센트로이드 벡터 산출과정은 도 7의 수식에 따라 진행된다.

[0050] 병합 센트로이드 벡터를 산출하는 이유는 부모 디렉토리가 자식 디렉토리의 개념을 포함하도록 하기 위함이다. 이를 위해, 가장 하위의 디렉토리부터 시작하여 자식 디렉토리들의 센트로이드 벡터와 자신의 센트로이드 벡터간의 벡터 합을 구하여 자신의 병합 센트로이드 벡터를 계산한다. 상세 알고리즘은 도시된 바와 같다.

[0051] 다시 도 2를 참조하면, 분류부(136)는 학습 데이터 생성부(134)를 통해 생성된 디렉토리별 학습 데이터를 주제 분류 트리 생성부(132)를 통해 생성된 상기 주제 분류 트리에 포함된 디렉토리에 대응시켜 생성된다.

[0052] 도 9는 본 발명의 일 실시예에 따른 분류부의 구성을 수식화한 도면이다.

[0053] 상기 분류부(136)는 분류할 웹 페이지 또는 광고의 텍스트를 벡터 스페이스 모델에서 단어 벡터(d)로 변환한다. 또한, 상기 생성된 주제 분류 트리의 각 노드를 나타내는 병합 센트로이드 벡터(mc)와 상기 단어 벡터(d)들 사이의 코사인 유사도를 계산하여, 가장 높은 코사인 유사도를 가지는 노드(class(d))로 웹 페이지 또는 광고의 주제를 결정한다. 한편, 분류부(136)는 벡터 스페이스 모델에서의 로치오(Rocchio) 분류기를 기반으로 생성될 수 있다.

[0054] 도 10은 본 발명의 일 실시예에 따른 웹 페이지 또는 광고의 주제 분류 방법을 도시한 순서도이다.

[0055] 먼저, 공개되어 있는 각종 디렉토리 및 각 디렉토리에 포함되는 웹 페이지 정보를 수집한다(S1010). 예를 들면, 오픈 디렉토리 프로젝트(ODP)의 데이터나 또는 각종 포털 사이트에 게시된 공개 디렉토리 데이터를 마이닝한다.

[0056] 다음으로, 상기 수집된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성한다(S1012). 주제 분류 트리 생성을 위해 앞서 설명한 네 가지 필터링 방법 중 하나 이상을 선택적으로 사용한다.

[0057] 다음으로, 상기 수집된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성한다(S1014). 이를 위해, 웹 페이지의 텍스트 정보를 단어 벡터로 변환한 후 각 단어 벡터를 기초로 각 디렉토리에 대한 센트로이드 벡터를 산출하고, 최하위 디렉토리부터 자식 디렉토리의 센트로이드 벡터와 부모 디렉토리의 센트로이드 벡터간의 벡터 합을 구하여, 각 부모 디렉토리의 병합 센트로이드 벡터를 산출한다.

[0058] 다음으로, 상기 디렉토리별 학습 데이터를 상기 주제 분류 트리에 포함된 디렉토리에 대응시켜 분류부를 생성한다(S1016). 상기 산출된 병합 센트로이드 벡터를 주제 분류 트리에 포함된 디렉토리에 대응시킴으로써, 분류부를 생성한다.

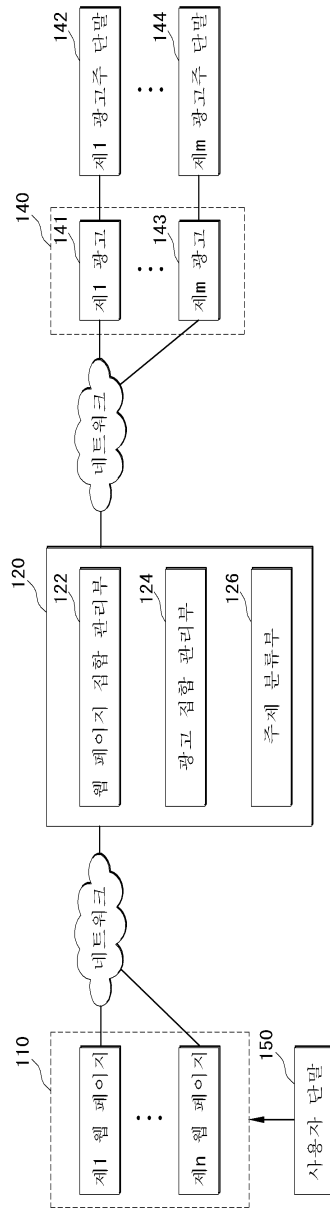
[0059] 다음으로, 생성된 분류부를 이용하여 분류 대상 웹 페이지 또는 분류 대상 광고의 텍스트를 대표하는 단어 벡터와 분류부의 디렉토리를 대표하는 단어 벡터 간의 유사도를 산출하여 웹 페이지 또는 광고의 주제를 결정한다(S1018).

[0060] 이와 같이, 본 발명에서는 공개된 디렉토리 데이터를 이용하여 분류부를 생성하므로, 분류부 생성에 소요되는 비용을 최소화할 수 있다. 또한, 공개된 디렉토리 데이터는 방대한 디렉토리 및 웹 페이지를 포함하고 있어, 이를 이용하여 분류부를 생성하는 경우 웹 페이지 또는 광고의 주제 분류 성능을 획기적으로 개선할 수 있다.

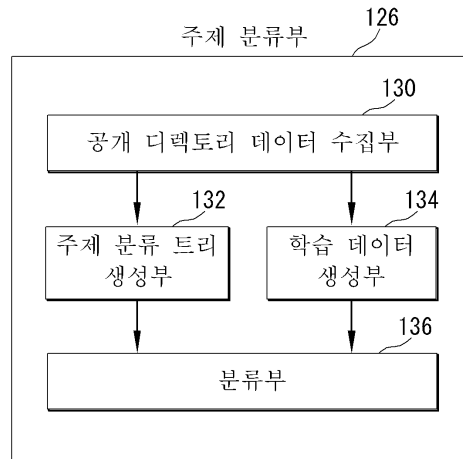
[0061] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함

도면

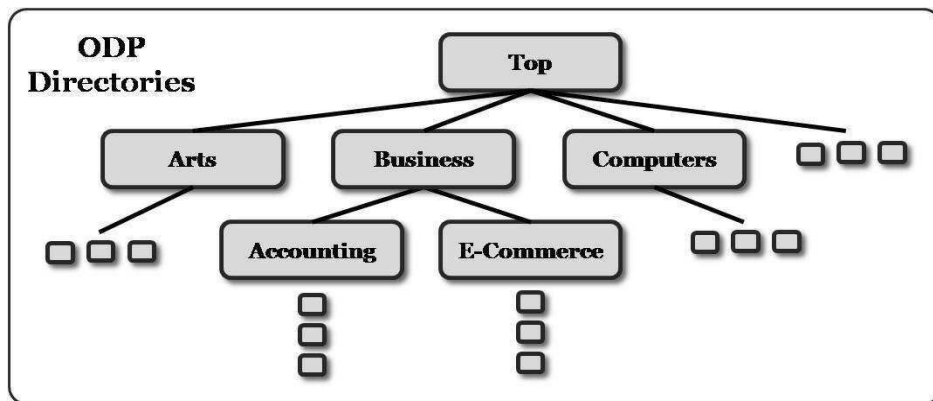
도면1



도면2



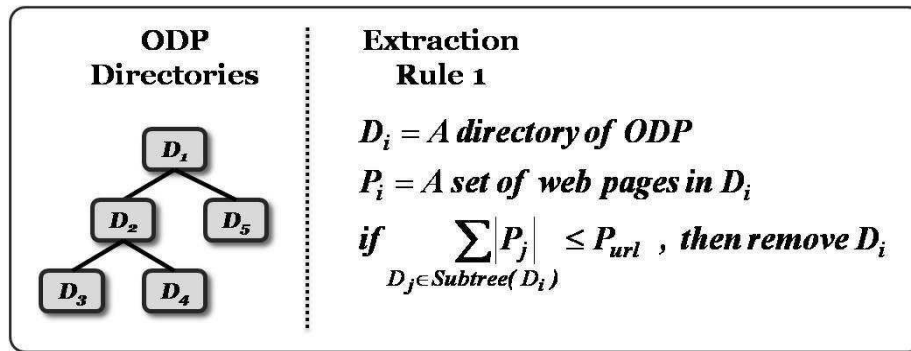
도면3



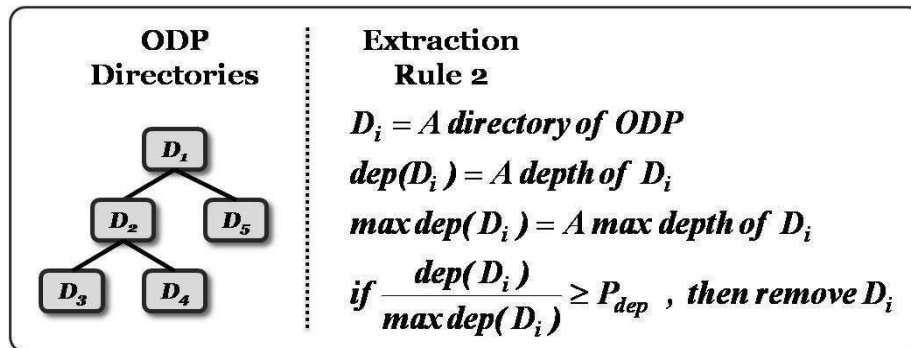
도면4



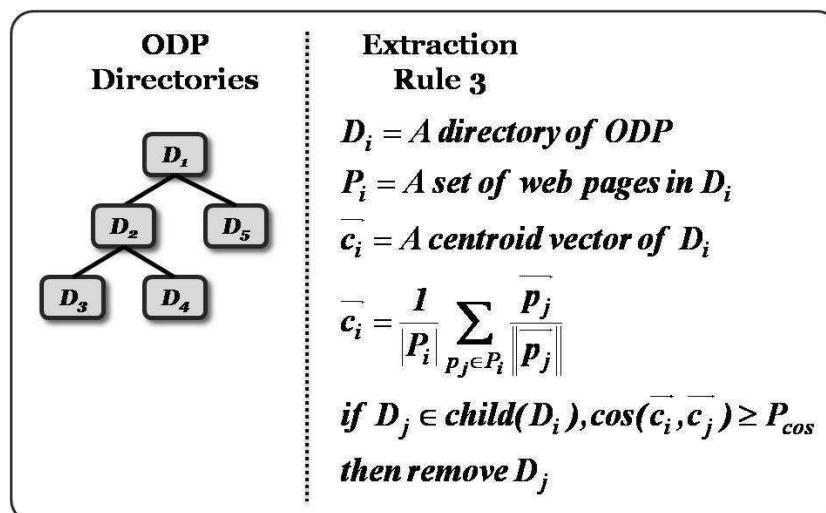
도면5



도면6



도면7



도면8

Input : root directory D_{root} in ODP
Output : merge centroid vectors for each directory in ODP
Algorithm : merge_centroid(D_i)

- 1: $\vec{c}_i \leftarrow \text{centroid}(D_i)$
- 2: if ($|child(D_i)| = 0$)
- 3: $\vec{mc}_i \leftarrow \vec{c}_i$
- 4: return $\vec{c}_i$
- 5: else
- 6: $\vec{mc}_i \leftarrow \vec{c}_i / \|\vec{c}_i\|$
- 7: for($c_j \in child(D_i)$)
- 8: $\vec{mc}_j \leftarrow \text{merge_centroid}(D_j)$
- 9: $\vec{mc}_i \leftarrow \vec{mc}_i + \vec{mc}_j / \|\vec{mc}_j\|$
- 10: return $\vec{mc}_i / 1 + |child(D_i)|$

도면9

d = A input page vector or ad vector
 C = A set of class in taxonomy
 $\vec{mc}_i$ = A merge centroid vector of class c_i

$$\text{class}(d) = \arg \max_{c_i \in C} \frac{\vec{mc}_i}{\|\vec{mc}_i\|} \cdot \frac{\vec{d}}{\|\vec{d}\|}$$

도면10

