



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0090354
(43) 공개일자 2011년08월10일

(51) Int. Cl.

H04B 1/40 (2006.01) G06F 17/16 (2006.01)

(21) 출원번호 10-2010-0010078

(22) 출원일자 2010년02월03일

심사청구일자 2010년02월03일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

이상근

서울특별시 동대문구 제기동 153번지 안암골 벽산 아파트 105동 1001호

하중우

서울특별시 성북구 안암동1가 149-1번지 404호

이정현

서울특별시 영등포구 대림3동 현대1차아파트 102동 1905호

(74) 대리인

특허법인 엠에이피에스

전체 청구항 수 : 총 10 항

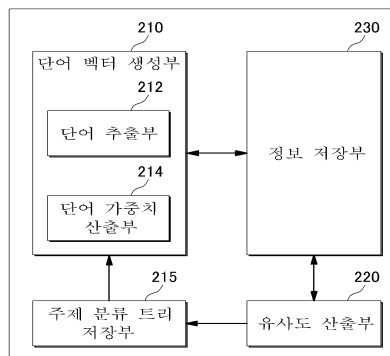
(54) 사용자 관심 주제를 추출하는 휴대용 통신 단말기 및 그 방법

(57) 요약

사용자 관심 주제를 추출하는 휴대용 통신 단말기는 휴대용 통신 단말기에 저장된 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성하는 단어 벡터 생성부와, 하나 이상의 학습 데이터를 포함하고, 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리를 저장하는 주제 분류 트리 저장부 및 단어 벡터와 주제 분류 트리에 포함된 각 노드별 학습 데이터간의 유사도를 산출하는 유사도 산출부를 포함하되, 유사도 산출부는 유사도에 따라 주제 분류 트리 중에서 사용자의 관심 주제를 대표할 노드명을 추출한다.

대표도 - 도9

200



특허청구의 범위

청구항 1

사용자 관심 주제를 추출하는 휴대용 통신 단말기에 있어서,

상기 휴대용 통신 단말기에 저장된 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성하는 단어 벡터 생성부,

하나 이상의 학습 데이터를 포함하고, 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리를 저장하는 주제 분류 트리 저장부 및

상기 단어 벡터와 상기 주제 분류 트리에 포함된 각 노드별 학습 데이터간의 유사도를 산출하는 유사도 산출부를 포함하되,

상기 유사도 산출부는 상기 유사도에 따라 상기 주제 분류 트리 중에서 사용자의 관심 주제를 대표할 노드명을 추출하는 것인 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 2

제 1 항에 있어서,

상기 단어 벡터 생성부는,

상기 텍스트 데이터로부터 단어를 추출하는 단어 추출부 및

상기 텍스트 데이터 내에서 사용되는 각 단어의 사용 빈도수, 및 상기 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출하는 단어 가중치 산출부를 포함하는 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 3

제 1 항에 있어서,

상기 유사도 산출부는,

상기 단어 벡터와 상기 주제 분류 트리의 각 노드에 포함된 학습 데이터간의 유사도를 산출하되,

상기 주제 분류 트리의 각 노드에 있어서 동일한 상위 노드에 속한 하위 노드들에 대해서는 각각의 유사도 값을 모두 합산하여 상기 상위 노드의 유사도 값으로 산출하는 것인 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 4

제 1 항에 있어서,

상기 유사도 산출부는,

상기 주제 분류 트리 중에서 상기 유사도가 가장 높은 노드명을 상기 사용자의 관심 주제로 추출하는 것인 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 5

제 1 항에 있어서,

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

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

상기 학습 데이터를 상기 주제 분류 트리에 포함된 디렉토리에 대응시키는 분류부를 더 포함하는 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 6

제 1 항 내지 제 5 항 중 하나의 항에 있어서,

상기 텍스트 데이터는,

상기 휴대용 통신 단말기에서 생성되는 문자 메시지, 파일명, 이메일, 모바일 웹 사용 히스토리 중 하나 이상에서 추출된 텍스트 데이터인 것인 사용자 관심 주제를 추출하는 휴대용 통신 단말기.

청구항 7

사용자 관심 주제 추출 방법에 있어서,

단어 추출부가 휴대용 통신 단말기에 저장된 텍스트 데이터로부터 단어를 추출하는 단계,

단어 가중치 산출부가 상기 텍스트 데이터 내에서 사용되는 각 단어의 사용 빈도수, 및 상기 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출하는 단계,

단어 벡터 생성부가 상기 단어 가중치를 기초로 상기 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성하는 단계 및

유사도 산출부가 하나 이상의 학습 데이터를 포함하고 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리에 포함된 각 노드별 학습 데이터와 상기 단어 벡터간의 유사도를 산출하는 단계를 포함하되,

상기 유사도 산출부는 산출된 상기 유사도에 따라 상기 주제 분류 트리 중에서 사용자의 관심 주제를 대표할 노드명을 추출하는 것인 사용자 관심 주제 추출 방법.

청구항 8

제 7 항에 있어서,

상기 유사도 산출부는,

상기 주제 분류 트리의 각 노드에 있어서 동일한 상위 노드에 속한 하위 노드들에 대해서는 각각의 유사도 값을 모두 합산하여 상기 상위 노드의 유사도 값으로 산출하는 것인 사용자 관심 주제 추출 방법.

청구항 9

제 7 항에 있어서,

상기 유사도 산출부는,

상기 주제 분류 트리 중에서 상기 유사도가 가장 높은 노드명을 상기 사용자의 관심 주제로 추출하는 것인 사용자 관심 주제 추출 방법.

청구항 10

제 7 항에 있어서,

상기 텍스트 데이터를 추출하는 단계에 앞서서, 상기 주제 분류 트리를 생성하는 단계를 포함하되,

상기 주제 분류 트리를 생성하는 단계는,

공개 디렉토리 데이터 수집부가 공개되어 있는 각종 디렉토리 및 각 디렉토리에 포함되는 웹 페이지 정보를 수집하는 단계,

주제 분류 트리 생성부가 상기 수집된 디렉토리 데이터를 가공하여 상기 주제 분류 트리를 생성하는 단계,

학습 데이터 생성부가 상기 수집된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 데이터 정보에 기초하여 각 디렉토리를 대표하는 상기 학습 데이터를 생성하는 단계 및

분류부가 디렉토리별 상기 학습 데이터를 상기 주제 분류 트리에 포함된 디렉토리에 대응시키는 단계를 포함하는 사용자 관심 주제 추출 방법.

명세서

기술 분야

- [0001] 본 발명은 사용자 관심 주제를 추출하는 휴대용 통신 단말기 및 그 방법에 관한 것으로서, 보다 상세하게는, 사용자 관심 주제를 추출하는 사용자 관심 주제를 추출하는 휴대용 통신 단말기 및 그 방법에 관한 것이다.

배경 기술

- [0002] 최근 휴대용 통신 단말기의 인터넷 활성화와 휴대용 통신 단말기의 개인화된 특성에 따라 휴대용 통신 단말기 환경에서 뉴스 추천, 맞춤형 광고 등 개인화된 정보 서비스를 제공하기 위한 기술이 개발되고 있다. 이러한 개인화된 정보 서비스를 제공하기 위해서는 사용자의 관심 사항을 파악하는 것이 필수적인 요소이다.
- [0003] 사용자의 관심 사항을 파악하기 위하여 일반적으로 활용되는 방법은 사용자의 관심 사항을 직접적으로 입력받는 것이다. 예를 들어, 사용자에게 관심정보 리스트를 제공하는 GUI를 통하여 사용자의 관심 사항을 획득하거나, 사용자가 관심 단어를 직접 지정할 수 있다.
- [0004] 그러나, 이러한 종래 방법은 사용자가 휴대용 통신 단말기와 같이 입력이 불편한 환경에서 적극적인 호응이 없을 경우 해당 사용자의 관심 사항을 파악하기가 어려운 문제점 있다.
- [0005] 또한, 사용자의 관심 사항을 자동으로 파악하기 위해, 사용자 단말기에 설치된 소프트웨어를 통하여 사용자의 검색 키워드를 자동으로 추출하고, 이를 사용자 관심 사항으로 간주하는 방법이 있다.
- [0006] 그러나, 이러한 종래 기술은 사용자의 검색 키워드를 모두 서버에 전송하기 때문에 전송 과정에서의 보안 문제가 발생할 수 있다. 그리고, 주로 사용자의 검색 키워드를 대상으로 하기 때문에 사용자 관심 사항을 파악하는 데에 있어서 추출 범위 및 추출 대상에 있어서 한계가 있다.

발명의 내용

해결하려는 과제

- [0007] 본 발명의 일부 실시예는 휴대용 통신 단말기에서 생성되는 데이터에 포함된 텍스트 데이터를 이용하여 사용자 관심 주제를 파악함으로써, 사용자가 명시적으로 자신의 관심 주제에 대한 데이터를 입력하지 않아도 보다 정확한 사용자 관심 주제를 추출할 수 있는 사용자 관심 주제를 추출하는 휴대용 통신 단말기를 제공하는 데에 그 목적이 있다.
- [0008] 또한, 본 발명의 일부 실시예는 공개된 디렉토리 데이터를 가공하여 생성된 주제 분류 트리의 노드명으로 사용자 관심 주제를 출력하므로, 사용자가 관심 주제에 대한 보안이 가능한 사용자 관심 주제를 추출하는 휴대용 통신 단말기를 제공하는 데에 그 목적이 있다.
- [0009] 또한, 본 발명의 일부 실시예는 주제 분류 트리의 특성상 사용자 관심 주제에 대한 상위 개념의 파악이 가능한 사용자 관심 주제를 추출하는 휴대용 통신 단말기를 제공하는 데에 그 목적이 있다.

과제의 해결 수단

- [0010] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 실시예에 따른 사용자 관심 주제를 추출하는 휴대용 통신 단말기는 휴대용 통신 단말기에 저장된 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성하는 단어 벡터 생성부와, 하나 이상의 학습 데이터를 포함하고, 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리를 저장하는 주제 분류 트리 저장부 및 단어 벡터와 주제 분류 트리에 포함된 각 노드별 학습 데이터간의 유사도를 산출하는 유사도 산출부를 포함하되, 유사도 산출부는 유사도에 따라 주제 분류 트리 중에서 사용자의 관심 주제를 대표할 노드명을 추출한다.
- [0011] 또한, 본 발명의 일 실시예에 따른 사용자 관심 주제 추출 방법은 단어 추출부가 휴대용 통신 단말기에 저장된 텍스트 데이터로부터 단어를 추출하는 단계와, 단어 가중치 산출부가 텍스트 데이터 내에서 사용되는 각 단어의 사용 빈도수, 및 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출하는 단계와, 단어 벡터 생성부가 단어 가중치를 기초로 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성하는 단계 및 유사도 산출부가 하나 이상의 학습 데이터를 포함하고 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리에 포함된 각 노드별 학습 데이터와 단어 벡터간의 유사도를 산출하는 단계

를 포함하되, 유사도 산출부는 산출된 유사도에 따라 주제 분류 트리 중에서 사용자의 관심 주제를 대표할 노드명을 추출한다.

발명의 효과

- [0012] 전술한 본 발명의 과제 해결 수단에 의하면, 휴대용 통신 단말기에서 생성되는 데이터에 포함된 텍스트 데이터를 이용하여 사용자 관심 주제를 파악함으로써, 사용자가 명시적으로 자신의 관심 주제에 대한 데이터를 입력하지 않아도 보다 정확한 사용자 관심 주제를 추출할 수 있는 장점이 있다.
- [0013] 또한, 전술한 본 발명의 과제 해결 수단에 의하면, 공개된 디렉토리 데이터를 가공하여 생성된 주제 분류 트리의 노드명으로 사용자 관심 주제를 출력하므로, 사용자 관심 주제에 대한 보안이 가능한 장점이 있다.
- [0014] 그리고, 주제 분류 트리의 특성상 사용자 관심 주제에 대한 상위 개념의 파악이 가능한 장점이 있다.

도면의 간단한 설명

- [0015] 도 1은 본 발명의 일 실시예에 따른 주제 분류 트리 장치의 블록도이다.
- 도 2는 본 발명의 일 실시예에 따른 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- 도 4는 본 발명의 일 실시예에 따라 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- 도 5는 본 발명의 일 실시예에 따른 오픈 디렉토리 프로젝트의 디렉토리 구조를 이용하여 생성된 주제 분류 트리의 일부를 나타낸다.
- 도 6은 본 발명의 일 실시예에 따른 학습 데이터 생성 단계를 수식화한 도면이다.
- 도 7은 본 발명의 일 실시예에 따른 각 노드에 생성된 학습 데이터에 대한 개념도이다.
- 도 8는 본 발명의 일 실시예에 따른 분류부의 구성을 수식화한 도면이다.
- 도 9는 본 발명의 일 실시예에 따른 사용자 관심 주제를 추출하는 휴대용 통신 단말기의 블록도이다.
- 도 10은 본 발명의 일 실시예에 따른 입력되는 데이터에서 텍스트 데이터를 추출하여 XML 형태로 저장하는 개념도이다.
- 도 11은 본 발명의 일 실시예에 따른 텍스트 데이터를 분석하여 단어 벡터를 생성하는 개념도이다.
- 도 12는 본 발명의 일 실시예에 따른 텍스트 데이터를 분석하여 유사도에 따라 사용자 관심 주제에 해당하는 노드명을 추출하는 개념도이다.
- 도 13은 본 발명의 일 실시예에 따른 주제 분류 트리를 생성하는 순서도이다.
- 도 14는 본 발명의 일 실시예에 따른 주제 분류 트리에서 사용자 관심 주제에 해당하는 노드명을 추출하는 순서도이다.

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] 도 1은 본 발명의 일 실시예에 따른 주제 분류 트리 장치의 블록도이다.
- [0019] 도 1에 도시된 바와 같이, 주제 분류 트리 장치(100)는 공개 디렉토리 데이터 수집부(110), 주제 분류 트리 생성부(120), 학습 데이터 생성부(130), 및 분류부(140)를 포함한다. 주제 분류 트리 장치(100)는 후술될 사용자

관심 주제를 추출하는 휴대용 통신 단말기(200)와 연동되어 동작하거나, 사용자 관심 주제를 추출하는 휴대용 통신 단말기(200)의 구성요소들을 포함할 수 있다. 사용자 관심 주제를 추출하는 휴대용 통신 단말기(200)에 대해서는 도 9에서 설명하기로 한다.

- [0020] 먼저 공개 디렉토리 데이터 수집부(110)는 공개되어 있는 각종 디렉토리 및 각 디렉토리에 포함되는 웹 페이지 정보를 수집한다. 예를 들면, 공개 디렉토리 데이터 수집부(110)는 오픈 디렉토리 프로젝트(ODP, Open Directory Project, <http://www.dmoz.org/>)의 데이터 또는 각종 포털 사이트에 게시된 공개 디렉토리 데이터를 마이닝한다.
- [0021] 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.
- [0022] 주제 분류 트리 생성부(120)는 공개된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성한다. 여기서 주제 분류 트리 생성부(120)는 실세계의 개념 및 주제를 나타내는 디렉토리 이외에도 단순한 사전식 분류를 위한 디렉토리 등이 있기 때문에, 분류 트리 생성시 이를 제거하여 주제 분류 트리를 생성할 수 있다.
- [0023] 또한, 주제 분류 트리 생성부(120)는 사용자의 관심 사항으로 적당한 개념 및 주제들을 나타내는 디렉토리를 선별하여 주제 분류 트리를 생성할 수 있다. 여기서, 본 발명의 실시예에서는 주제 분류 트리 생성을 위해 하기의 네 가지 필터링 단계를 거칠 수 있다.
- [0024] 첫 번째 단계에서는, 공개된 디렉토리 데이터의 전체 디렉토리 중에서 주제를 나타내지 않는 디렉토리가 제거된다. 예를 들어, 최상위 디렉토리 중 다른 디렉토리나 중복이 되고 또한 다양한 언어로 구성되는 "World", "Regional" 디렉토리가 제거될 수 있다.
- [0025] 여기서 주제 분류 트리의 대상 언어에 따라 "World" 디렉토리의 자식 디렉토리가 선별되어 적용될 수 있다. 또한, 단순히 알파벳(A, B, C, ...)이나 숫자(1, 2, 3, ...)를 디렉토리 이름으로 사용하는 중간 디렉토리들이 제거되고, 그 서브 트리가 상위 디렉토리에 연결될 수 있다. 또한, 특정 주제를 나타내지 않고 각 디렉토리마다 자주 반복되는 리프 디렉토리들이 제거될 수 있다.
- [0026] 두 번째 단계에서는, 공개된 디렉토리 데이터에 포함된 디렉토리들 중 임계값보다 작은 개수의 웹 페이지를 포함하는 디렉토리가 제거될 수 있다.
- [0027] 도 2는 본 발명의 일 실시예에 따른 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0028] 이어서 설명하자면, 도 2에 도시된 바와 같이, 주제 분류 트리 생성부(120)는 각 디렉토리(Di)에 포함된 웹 페이지의 개수(Pi)가 임계값(Pur1) 보다 적은 경우, 해당 디렉토리를 제거할 수 있다.
- [0029] 세 번째 단계에서는, 일반성과 구체성을 적절한 수준에서 보장하는 주제를 선별하기 위해, 주제 분류 트리 생성부(120)는 트리의 모든 리프 노드까지 이르는 각각의 경로에서 주어진 파라미터(Pdep) 비율로 하위노드들을 제거할 수 있다.
- [0030] 도 3은 본 발명의 일 실시예에 따른 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0031] 도 3의 수식에 기재된 바와 같이, 주제 분류 트리 생성부(120)는 각 디렉토리의 깊이(dep(Di))에 대한 최하위 디렉토리의 깊이(maxdep(Di))의 비율이 임계값(Pdep) 보다 큰 경우 해당 디렉토리를 제거할 수 있다.
- [0032] 그리고 도 3에 도시된 예에서, 디렉토리(D2)의 깊이는 2 이고, 최하위 디렉토리(D3)의 깊이는 3 이 된다. 임계값(Pdep)이 0.8이라고 가정하면, $\text{dep}(D2)/\text{maxdep}(D2)=2/3$ 이고, 이는 임계값(Pdep)보다 작으므로, 해당 디렉토리(D2)는 유지시킨다. 그러나, $\text{dep}(D3)/\text{maxdep}(D3)=3/3$ 이고, 이는 임계값(Pdep)보다 크므로, 해당 디렉토리(D3)는 제거될 수 있다.
- [0033] 네 번째 단계에서는, 각 주제들 간의 충분한 차별성을 보장하여 사용자 관심 주제의 정확도를 향상시키기 위해, 각 디렉토리의 웹 페이지에 벡터 스페이스 모델이 적용될 수 있다.
- [0034] 도 4는 본 발명의 일 실시예에 따른 공개된 디렉토리 데이터를 가공하는 방법을 수식화한 도면이다.
- [0035] 수식에 기재된 바와 같이, 주제 분류 트리 생성부(120)는 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

자식 디렉토리를 제거할 수 있다. 이때, 코사인 유사도는 수식식 1을 기초로 산출될 수 있다.

수식식 1

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

[0036]

[0037]

이와 같이, 주제 분류 트리 생성부(120)는 상술된 네 가지 단계에 따라 디렉토리를 제거할 수 있으며, 남은 디렉토리를 주제 분류 트리의 각 노드로 변환할 수 있다.

[0038]

그리고, 주제 분류 트리 생성부(120)는 디렉토리 이름을 노드 이름으로 정하며, 디렉토리 간의 부모-자식 관계를 각 노드 사이의 간선으로 변환하여 주제 분류 트리를 생성할 수 있다. 이때, 디렉토리를 필터링하는 단계들 중 하나 이상이 선택적으로 적용될 수 있다.

[0039]

도 5는 본 발명의 일 실시예에 따른 오픈 디렉토리 프로젝트의 디렉토리 구조를 이용하여 생성된 주제 분류 트리의 일부를 나타낸다.

[0040]

오픈 디렉토리 프로젝트에 존재하는 디렉토리의 구조는 트리 형태로 이루어져 있으며, 전체 디렉토리의 수는 약 59만개, 최대 트리 깊이는 15레벨이며, 최상위 레벨의 디렉토리의 개수는 16개이다.

[0041]

도 5에 도시된 바와 같이, 이러한 디렉토리들은 각각 실세계의 하나의 개념 및 주제를 나타내기 때문에, 이를 임의의 사용자의 관심 사항들로 보고 주제 분류 트리가 생성될 수 있다.

[0042]

여기서 상술된 네 가지 단계를 활용하여, 주제 분류 트리 생성부(120)는 사용자의 관심 사항으로 적당한 개념 및 주제들을 나타내는 디렉토리를 선별하여 주제 분류 트리를 생성할 수 있다.

[0043]

다시 도 1을 참조하면, 학습 데이터 생성부(130)는 공개된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 데이터 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성한다.

[0044]

도 6은 본 발명의 일 실시예에 따른 학습 데이터 생성 단계를 수식화한 도면이다. 그리고 도 7은 본 발명의 일 실시예에 따른 각 노드에 생성된 학습 데이터에 대한 개념도이다.

[0045]

이어서 설명하자면, 학습 데이터 생성부(130)는 웹 페이지의 텍스트 데이터 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 산출할 수 있다.

[0046]

여기서 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 도 6에 도시된 바와 같다.

[0047]

그리고, 각 웹 페이지의 텍스트 데이터 정보는 웹 사이트 주소(URL), 제목(Title), 설명(Description)을 포함하므로, 학습 데이터 생성부(130)는 이를 기초로 단어 벡터로 변환할 수 있다. 이때 도 7에 도시된 바와 같이, 학습 데이터 생성부(130)는 제목과 설명을 포함하는 텍스트 데이터(702)를 하나의 문서로 구성하고, 각 문서들을 주제 분류 트리의 각 노드들에 대한 학습 데이터로 생성할 수 있다. 단어 벡터 변환 및 센트로이드 벡터 산출과정은 도 4의 수식에 따라 진행될 수 있다.

[0048]

다시 도 1을 참조하면, 분류부(140)는 생성된 디렉토리별 학습 데이터를 주제 분류 트리에 포함된 디렉토리에 대응시킨다. 여기서, 분류부(140)는 산출된 병합 센트로이드 벡터를 주제 분류 트리에 포함된 디렉토리에 대응시킬 수 있다.

[0049]

도 8는 본 발명의 일 실시예에 따른 분류부의 구성을 수식화한 도면이다.

[0050]

이어서 설명하자면, 상술된 분류부(140)는 분류할 웹 페이지의 텍스트 데이터를 벡터 스페이스 모델에서 단어 벡터(d)로 변환한다. 또한, 상기 생성된 주제 분류 트리의 각 노드를 나타내는 병합 센트로이드 벡터(mc)와 단어 벡터(d)들 사이의 코사인 유사도를 계산하여, 가장 높은 코사인 유사도를 가지는 노드(class(d))로 웹 페이지

지의 주제를 결정한다. 여기서 분류부(140)는 벡터 스페이스 모델에서의 로치오(Rocchio) 분류기를 포함할 수 있다.

- [0051] 한편, 상술된 주제 분류 트리에 포함된 노드별 학습 데이터와 후술될 휴대용 통신 단말기에서 생성된 데이터에 포함된 텍스트 데이터를 대표하는 단어 벡터간의 유사도를 통해 사용자의 관심 사항이 주제 분류 트리의 노드명의 형태로 추출될 수 있다. 이에 대해서 이하 도 9에서 보다 구체적으로 후술하기로 한다.
- [0052] 도 9는 본 발명의 일 실시예에 따른 사용자 관심 주제를 추출하는 휴대용 통신 단말기의 블록도이다.
- [0053] 사용자 관심 주제를 추출하는 휴대용 통신 단말기(200)는 단어 벡터 생성부(210), 주제 분류 트리 저장부(215), 유사도 산출부(220) 및 정보 저장부(230)를 포함한다. 사용자 관심 주제를 추출하는 휴대용 통신 단말기(200)는 상술된 도 1의 구성요소들을 포함하거나 주제 분류 트리 장치(100)와 연동되어 동작할 수 있다.
- [0054] 단어 벡터 생성부(210)는 휴대용 통신 단말기에 저장된 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성한다. 이를 위해 단어 벡터 생성부(210)는 단어 추출부(212) 및 단어 가중치 산출부(214)를 포함한다.
- [0055] 단어 추출부(212)는 텍스트 데이터로부터 단어를 추출한다. 여기서 텍스트 데이터는 휴대용 통신 단말기에서 생성되는 문자 메시지, 파일명, 이메일, 모바일 웹 사용 히스토리 중 하나 이상에서 추출된 텍스트 데이터일 수 있다. 즉, 단어 추출부(212)는 사용자가 휴대폰으로 문자 메시지를 보내거나, 이메일 작성, 사용자 파일 저장 및 웹 연결을 통한 검색을 수행할 때 생성되는 텍스트 데이터에서 단어를 추출할 수 있다.
- [0056] 도 10은 본 발명의 일 실시예에 따른 입력되는 데이터에서 텍스트 데이터를 추출하여 XML 형태로 저장하는 개념도이다.
- [0057] 이어서 설명하자면, 단어 추출부(212)는 휴대용 통신 단말기에서 생성되는 문자 메시지(1002), 파일명(1004), 이메일(1006), 모바일 웹 사용 히스토리(1008) 등의 텍스트 데이터에서 단어를 추출한다(S1001). 여기서, 텍스트 데이터에는 해당 텍스트 데이터에 관련된 정보들로서 메타 데이터가 포함될 수 있다. 구체적으로, 텍스트 데이터에 관련된 정보에는 해당 텍스트 데이터가 포함된 텍스트 데이터의 종류(예컨대 SMS, FILE, EMAIL 등), 텍스트 데이터 생성시간, 생성위치, 연관된 응용 프로그램 정보 등이 포함될 수 있다.
- [0058] 또한, 단어 추출부(212)는 특정 설정된 기간 동안 생성된 텍스트 데이터에 대해서 단어를 추출할 수 있으며, 텍스트 데이터에 관련된 정보들을 XML 파일 형태로 하기 정보 저장부(230)에 저장할 수 있다(S1011). 여기서, XML 데이터(1010)에는 각 텍스트 종류에 대한 리스트(1)들과, 텍스트 데이터에 대한 정보(3)들, 및 해당 텍스트 데이터가 생성된 시간에 대한 정보(2)들이 포함될 수 있다.
- [0059] 다시 도 9로 돌아가서, 단어 가중치 산출부(214)는 텍스트 데이터 내에서 사용되는 각 단어의 사용 빈도수, 및 상기 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출한다. 예컨대, 사용자가 주로 문자 메시지를 많이 사용할 경우, 다른 텍스트 데이터의 종류에 비해 문자 메시지 타입에 더 많은 가중치가 부여될 수 있다. 그리고, 최근에 생성된 데이터가 현재 사용자의 관심 사항을 더 잘 나타낼 수 있으므로, 해당 단어가 포함된 텍스트 데이터가 생성된 시간에 대한 가중치가 부여될 수 있다.
- [0060] 따라서, 단어 가중치 산출부(214)는 각 단어의 사용 빈도수, 및 해당 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출하고, 단어 벡터 생성부(210)는 이를 기초로 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성할 수 있다. 여기서, 단어 가중치 산출부(214)는 텍스트 데이터의 종류에 따라 추가로 가중치를 더 부여할 수도 있다.
- [0061] 도 11은 본 발명의 일 실시예에 따른 텍스트 데이터를 분석하여 단어 벡터를 생성하는 개념도이다.
- [0062] 이어서 설명하자면, 단어 가중치 산출부(214)가 상술된 XML 데이터(1010)에 대해 XML 파서를 통해 파싱(Parsing)을 수행하고, 그 결과를 메모리에 로딩할 수 있다(S1101, S1111). 그리고, 단어 가중치 산출부(214)가 메모리에 로딩된 각 텍스트 데이터 및 메타 데이터를 분석하면, 단어 벡터 생성부(210)는 분석된 정보를 이용하여 단어 가중치를 기초로 한 벡터 스페이스 모델에서의 단어 벡터(1110)를 생성할 수 있다(S1121).
- [0063] 이러한 단어 벡터(1110)는 기본 분류 단위로서, 텍스트 데이터를 이루는 개별적인 단어(Term)들을 차원으로 가지고, 각 단어마다 단어 가중치(Term Weight)를 값으로 가지는 벡터로 정의될 수 있다. 이때, 단어 가중치는 상술된 바와 같이, 단어 빈도수(Term Frequency)와 시간 가중치(Time Weight)의 연산으로 계산될 수 있다.
- [0064] 단어 빈도수는, 도 11에 도시된 바와 같이 추출된 텍스트 데이터에 포함된 개별적인 단어들(1102)의 사용 횟수

(1104)를 나타내며, 이는 추출된 텍스트 데이터에 자주 나타난 단어가 사용자의 관심 사항을 나타내는데 더 중요한 단어임을 반영하는 것이다. 이때, 해당 단어가 포함된 텍스트 데이터가 추출된 텍스트 데이터의 종류에 따라 해당 단어에 가중치가 부여될 수 있다.

- [0065] 그리고, 시간 가중치(1106)란, 각 단어가 포함된 텍스트 데이터의 메타데이터 중 해당 텍스트 데이터 생성 시간에 따른 가중치이다. 이때, 단어가 포함된 텍스트 데이터의 생성 시간이 단어 벡터(1110)가 계산되는 현재 시간에 얼마나 가까운지를 0~1의 범위로 나타낼 수 있다.
- [0066] 이는 최근에 생성된 텍스트 데이터에 포함된 단어일수록 사용자의 관심 사항을 나타내는데 더 중요한 단어임을 반영하기 위한 것이다. 이러한 단어 벡터(1110)는 텍스트 데이터의 종류 예컨대, 추출된 텍스트 데이터가 생성된 위치, 해당 텍스트 데이터와 연관된 응용프로그램의 타입에 따라 하나씩 생성될 수 있다.
- [0067] 구체적으로, 단어 벡터 생성부(210)는 SMS(Short Message Service)/MMS(Multimedia Messaging Service) 문자 메시지, 파일명, 이메일, 및 모바일 웹 사용 히스토리 등의 텍스트 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 개별적으로 생성할 수 있다. 이때, 단어 빈도수 및 시간 가중치를 값으로 포함하는 단어 가중치에 따라 텍스트 종류별로 각 텍스트 데이터를 대표하는 단어 벡터가 하나 이상 생성될 수 있다.
- [0068] 다시 도 9로 돌아가서, 주제 분류 트리 저장부(215)는 하나 이상의 학습 데이터를 포함하고, 주제를 나타내는 복수의 노드가 트리 구조로 연결된 주제 분류 트리를 저장한다. 주제 분류 트리를 생성하는 방법에 대해서는 도 1에서 상세히 설명되었으므로 이하 생략하도록 한다.
- [0069] 유사도 산출부(220)는 휴대용 통신 단말기에서 생성되는 데이터에 포함된 텍스트 데이터를 대표하는 상기 단어 벡터(1110)와 주제 분류 트리에 포함된 각 노드별 학습 데이터간의 유사도를 산출한다.
- [0070] 도 12는 본 발명의 일 실시예에 따른 텍스트 데이터를 분석하여 유사도에 따라 사용자 관심 주제에 해당하는 노드명을 추출하는 개념도이다.
- [0071] 이어서 설명하자면 도 12에 도시된 바와 같이, 유사도 산출부(220)는 분류 알고리즘을 이용하여 휴대용 통신 단말기에서 생성되는 데이터에 포함된 텍스트 데이터를 대표하는 단어 벡터(1110)를 주제 분류 트리(1200)의 각 노드에 대응시킨다(S1201). 여기서, 분류 알고리즘은 지지 벡터 머신(Support Vector Machine), 고유 베이저안(Naive Bayesian), K-NN(K-Nearest Neighbor) 등을 포함할 수 있다.
- [0072] 그리고, 유사도 산출부(220)는 단어 벡터(1110)와 해당 노드의 학습 데이터간의 유사도(1202)를 산출한다. 여기서, 유사도 산출부(220)는 단어 벡터(1110)와 각 노드별 학습 데이터간의 유사도를 산출하여 유사도(1202)가 높은 상위 N개의 노드명(1204)을 유사도(1202) 점수에 따라 추출할 수 있다(S1211). 그리고, 유사도 산출부(220)는 주제 분류 트리(1200)의 각 노드에 있어서 동일한 상위 노드에 속한 하위 노드들에 대해서는 각각의 유사도 값을 모두 합산하여 상위 노드의 최종 유사도 값으로 산출할 수 있다.
- [0073] 그리고, 유사도 산출부(220)는 주제 분류 트리(1200)에 있어서 동일한 계층에 속한 상위 노드들 중 최종적으로 산출된 유사도 값이 가장 높은 노드명을 사용자 관심 주제로 추출할 수 있다.
- [0074] 예컨대, 스포츠(Sports)-축구(Soccer)로 이어지는 주제 분류 트리(1200)에서 축구 노드의 하위 노드들에 대해서는 유사도에 대한 값을 모두 합산하여 축구 노드의 유사도로 산출될 수 있다. 따라서, 유사도 값이 축구(Soccer)가 가장 큰 값인 경우, Top/Sports/Soccer란 노드명이 해당 사용자 관심 주제로 추출될 수 있다.
- [0075] 또한, 구성에 따라 축구(Soccer), 야구(Baseball) 등의 동일한 계층에 대한 유사도를 모두 합산하여 그 상위 노드의 스포츠(Sports) 노드의 유사도로 최종 산출할 수 있다. 이 경우, 동일한 계층에 있는 노드들 중에서 스포츠(Sports) 노드의 최종 유사도 값이 가장 큰 값인 경우, Top/Sports 란 노드명이 해당 사용자 관심 주제로 추출될 수 있다. 이와 같이, 주제 분류 트리의 노드명으로 사용자 관심 주제가 추출되므로, 사용자 관심 주제에 대한 보안이 가능하다.
- [0076] 다시 도 9를 참조하면, 정보 저장부(230)는 본 발명을 수행하기 위한 프로그램, 알고리즘, 설정값, 트리 구조에 대한 정보, 각 노드별 유사도값, 각 노드별 학습 데이터 정보, 데이터 타입 정보, 단어 벡터에 대한 정보 등을 저장한다.
- [0077] 도 13은 본 발명의 일 실시예에 따른 주제 분류 트리를 생성하는 순서도이다.
- [0078] 먼저, 공개 디렉토리 데이터 수집부(110)가 공개되어 있는 각종 디렉토리 및 각 디렉토리에 포함되는 웹 페이지 정보를 수집한다(S1301). 예를 들면, 오픈 디렉토리 프로젝트(ODP)의 데이터나 또는 각종 포털 사이트에 게시된

공개 디렉토리 데이터가 마이닝된다.

- [0079] 다음으로, 주제 분류 트리 생성부(120)가 수집된 디렉토리 데이터를 가공하여 주제 분류 트리를 생성한다(S1311). 주제 분류 트리 생성을 위해 앞서 설명한 네 가지 필터링 방법 중 하나 이상이 선택적으로 사용될 수 있다.
- [0080] 다음으로, 학습 데이터 생성부(130)가 수집된 디렉토리 데이터 중 각 디렉토리에 포함되는 웹 사이트의 텍스트 데이터 정보에 기초하여 각 디렉토리를 대표하는 학습 데이터를 생성한다(S1321). 이를 위해, 학습 데이터 생성부(130)는 웹 페이지의 텍스트 데이터 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.
- [0081] 다음으로, 분류부(140)가 디렉토리별 학습 데이터를 주제 분류 트리에 포함된 디렉토리에 대응시킨다(S1331). 여기서, 분류부(140)는 산출된 병합 센트로이드 벡터를 주제 분류 트리에 포함된 디렉토리에 대응시킬 수 있다.
- [0082] 도 14는 본 발명의 일 실시예에 따른 주제 분류 트리에서 사용자 관심 주제에 해당하는 노드명을 추출하는 순서도이다.
- [0083] 먼저, 단어 추출부(212)가 휴대용 통신 단말기에 저장된 텍스트 데이터로부터 단어를 추출한다(S1401).
- [0084] 다음으로, 단어 가중치 산출부(214)가 텍스트 데이터 내에서 사용되는 각 단어의 사용 빈도수, 및 상기 단어가 포함된 텍스트 데이터의 생성 시간에 기초하여 단어 가중치를 산출한다(S1411).
- [0085] 다음으로, 단어 벡터 생성부(210)가 단어 가중치를 기초로 텍스트 데이터의 종류에 따라 각 텍스트 데이터를 대표하는 단어 벡터를 생성한다 (S1421).
- [0086] 다음으로, 유사도 산출부(220)가 단어 벡터를 주제 분류 트리의 노드에 대응시키고(S1431), 단어 벡터와 주제 분류 트리에 포함된 각 노드별 학습 데이터간의 유사도를 산출한다(S1441).
- [0087] 그리고, 유사도 산출부(220)가 유사도에 따라 주제 분류 트리 중에서 사용자 관심 주제를 대표할 노드명을 추출한다(S1451).
- [0088] 여기서, 유사도 산출부(220)는 주제 분류 트리 중에서 유사도가 가장 높은 노드명을 사용자 관심 주제로 추출할 수 있다. 또한, 유사도 산출부(220)는 주제 분류 트리의 각 노드에 있어서 동일한 노드에 대해서는 유사도에 대한 값을 모두 합산하여 최종 산출값이 가장 높은 최상위 노드명을 사용자 관심 주제로 추출할 수 있다.
- [0089] 한편, 상술된 도 9의 주제 분류 트리 저장부(215) 및 정보 저장부(230)는 캐쉬, ROM(Read Only Memory), PROM(Programmable ROM), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM) 및 플래쉬 메모리(Flash memory)와 같은 비휘발성 메모리 소자 또는 RAM(Random Access Memory)과 같은 휘발성 메모리 소자 또는 하드디스크 드라이브(HDD, Hard Disk Drive), CD-ROM과 같은 저장 매체 중 적어도 하나로 구현될 수 있으나 이에 한정되지는 않는다
- [0090] 또한, 도 1 및 도 9에서 도시된 각각의 구성요소는 일종의 '모듈'로 구성될 수 있다. 상기 '모듈'은 소프트웨어 또는 Field Programmable Gate Array(FPGA) 또는 주문형 반도체(ASIC, Application Specific Integrated Circuit)과 같은 하드웨어 구성요소를 의미하며, 모듈은 어떤 역할들을 수행한다. 그렇지만 모듈은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. 모듈은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 실행시키도록 구성될 수도 있다. 구성요소들과 모듈들에서 제공되는 기능은 더 작은 수의 구성요소들 및 모듈들로 결합되거나 추가적인 구성요소들과 모듈들로 더 분리될 수 있다.
- [0091] 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.
- [0092] 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

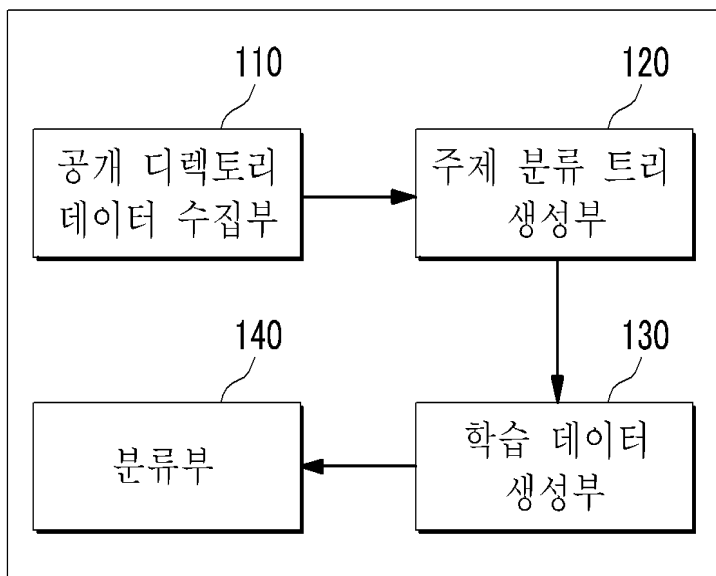
[0093]

110: 공개 디렉토리 데이터 수집부	120: 주제 분류 트리 생성부
130: 학습 데이터 생성부	140: 분류부
210: 단어 벡터 생성부	220: 유사도 산출부
212: 단어 추출부	214: 단어 가중치 산출부
215: 주제 분류 트리 저장부	220: 유사도 산출부
230: 정보 저장부	

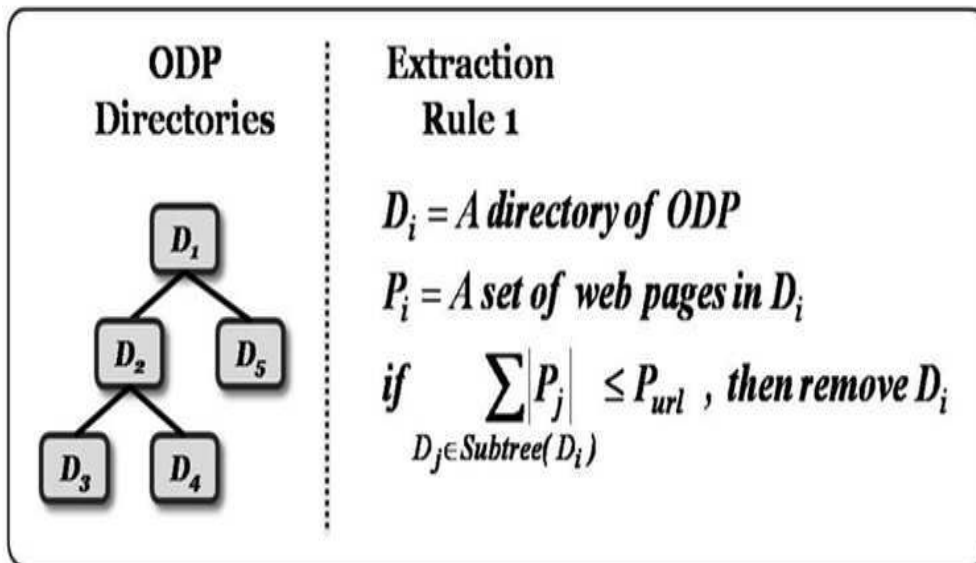
도면

도면1

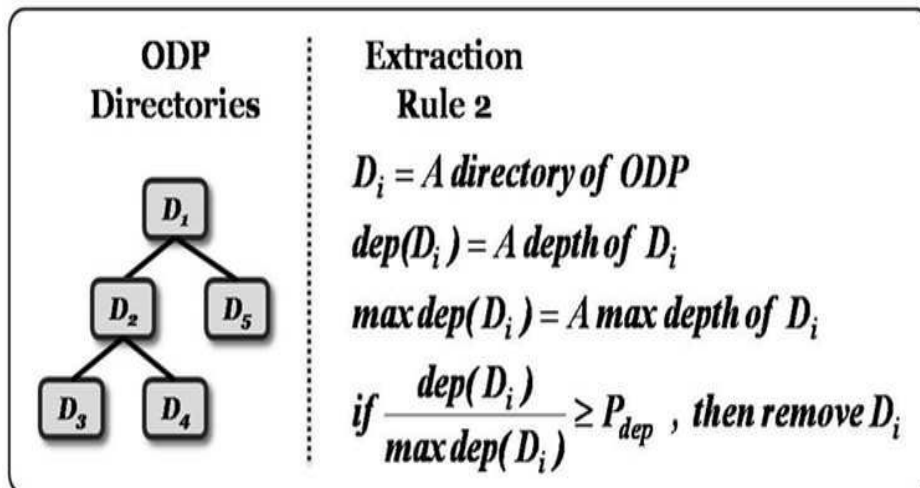
100



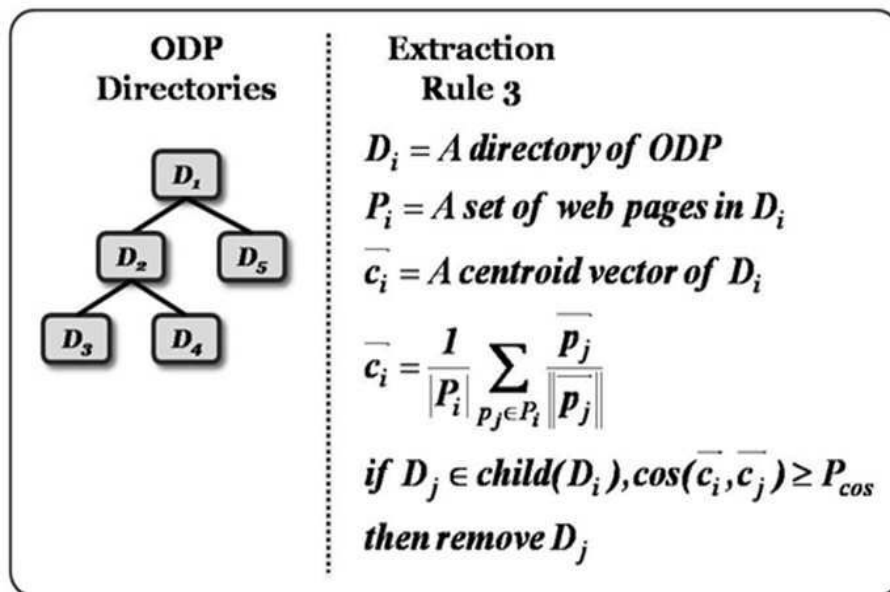
도면2



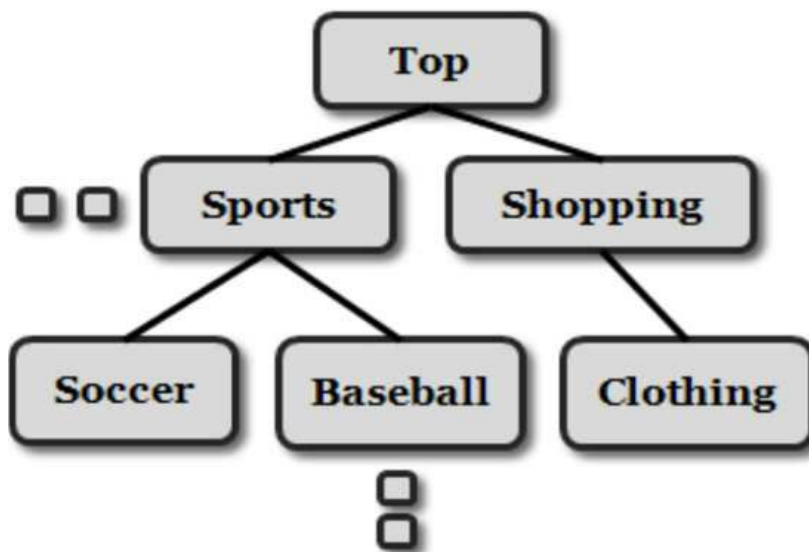
도면3



도면4



도면5

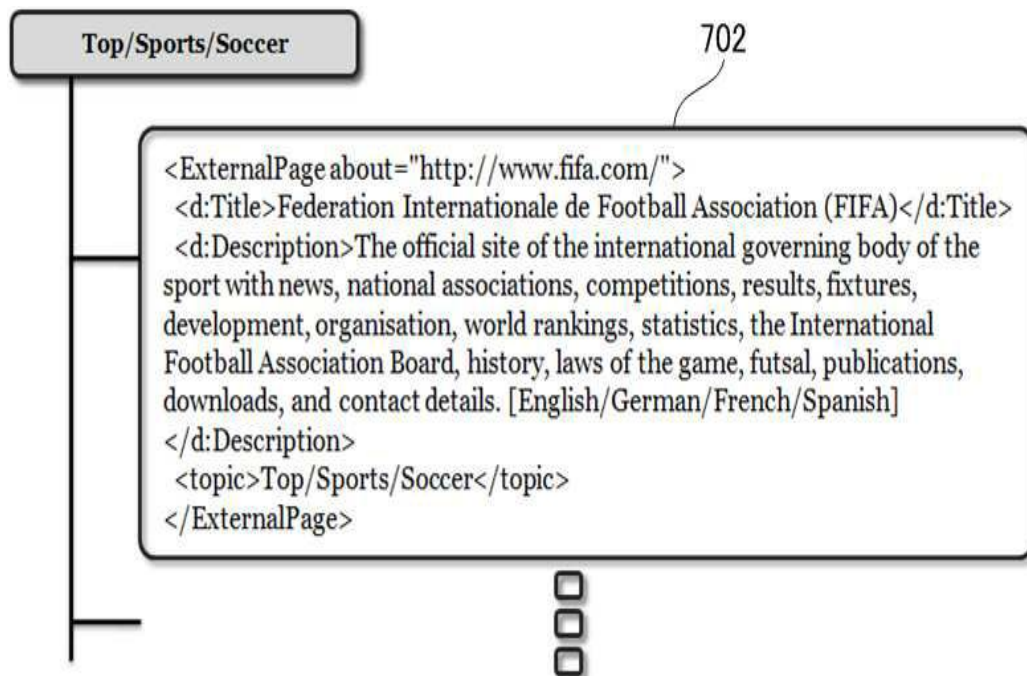


도면6

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

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

도면7

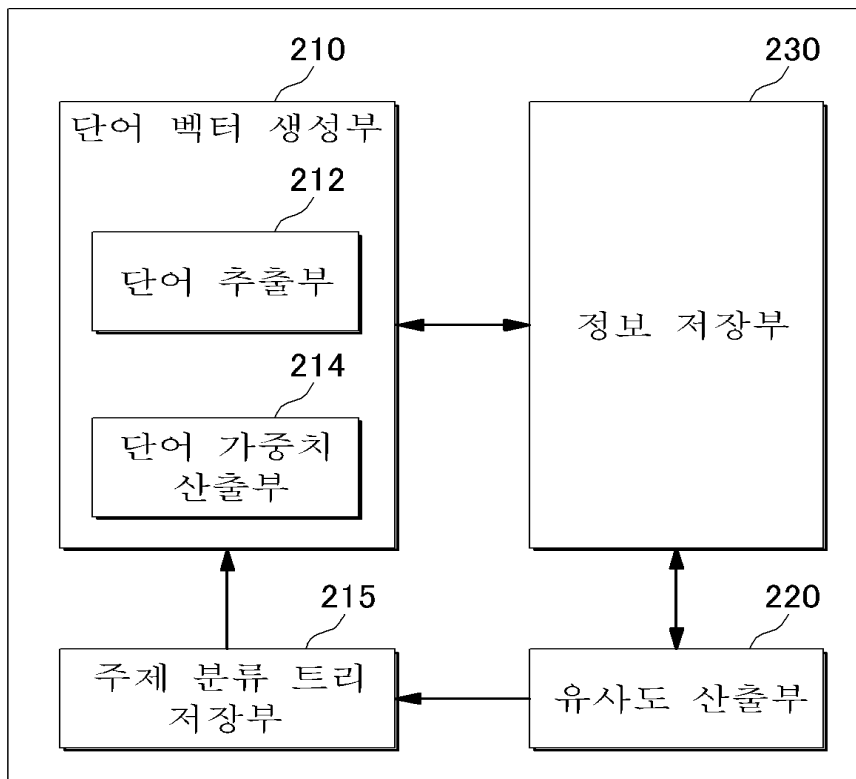


도면8

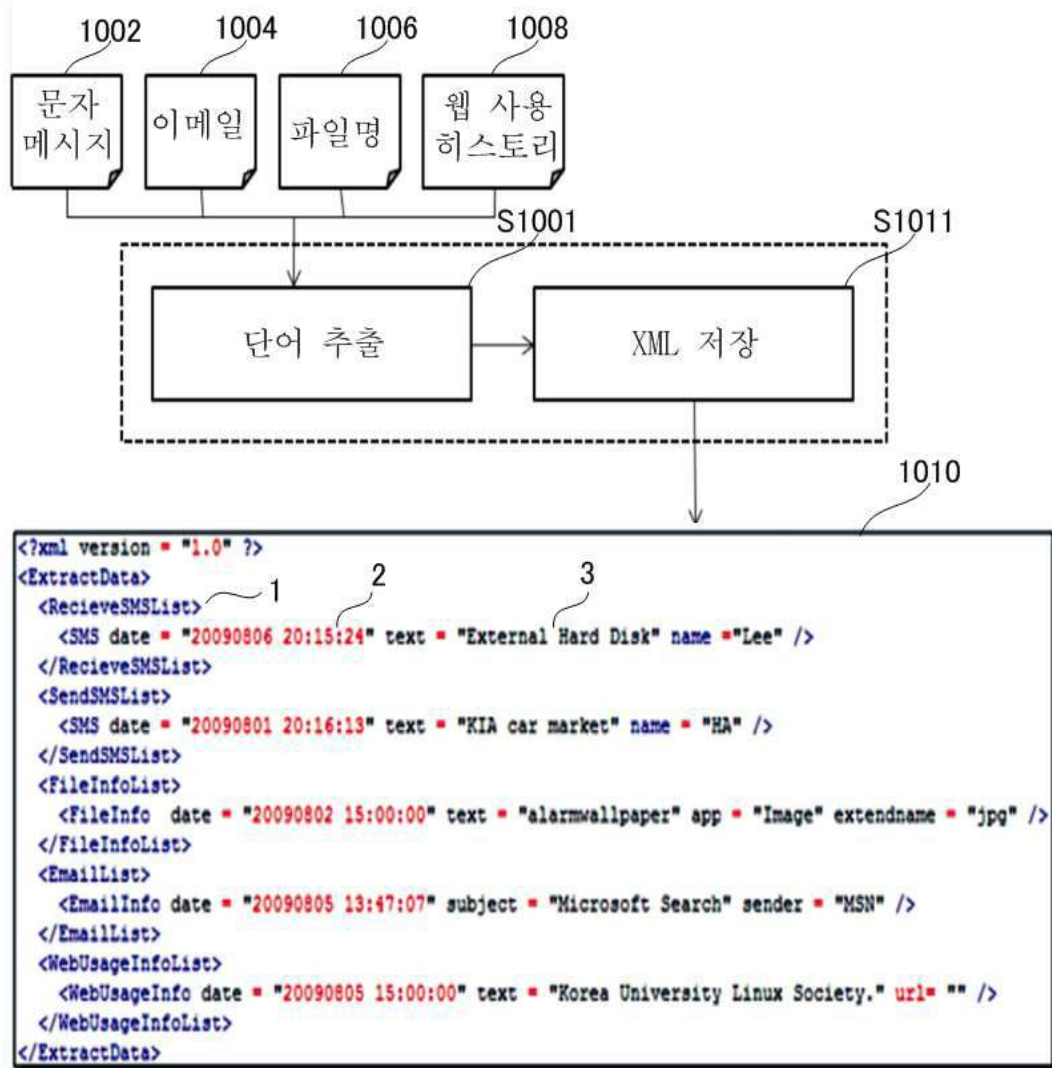
$$\begin{aligned}
 & \mathbf{d} = \text{A input page vector or ad vector} \\
 & \mathbf{C} = \text{A set of class in taxonomy} \\
 & \overline{\mathbf{mc}}_i = \text{A merge centroid vector of class } c_i \\
 & \text{class}(\mathbf{d}) = \arg \max_{c_i \in \mathbf{C}} \frac{\overline{\mathbf{mc}}_i}{\|\overline{\mathbf{mc}}_i\|} \cdot \frac{\mathbf{d}}{\|\mathbf{d}\|}
 \end{aligned}$$

도면9

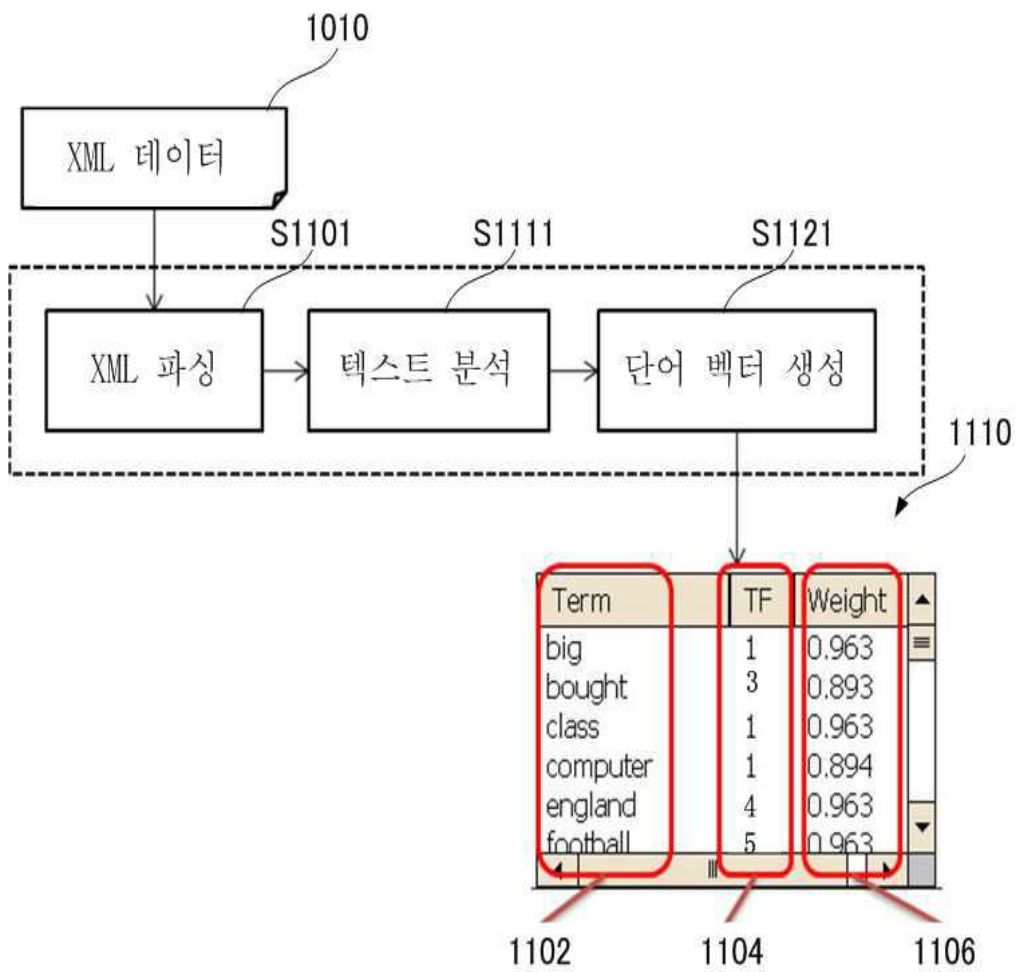
200



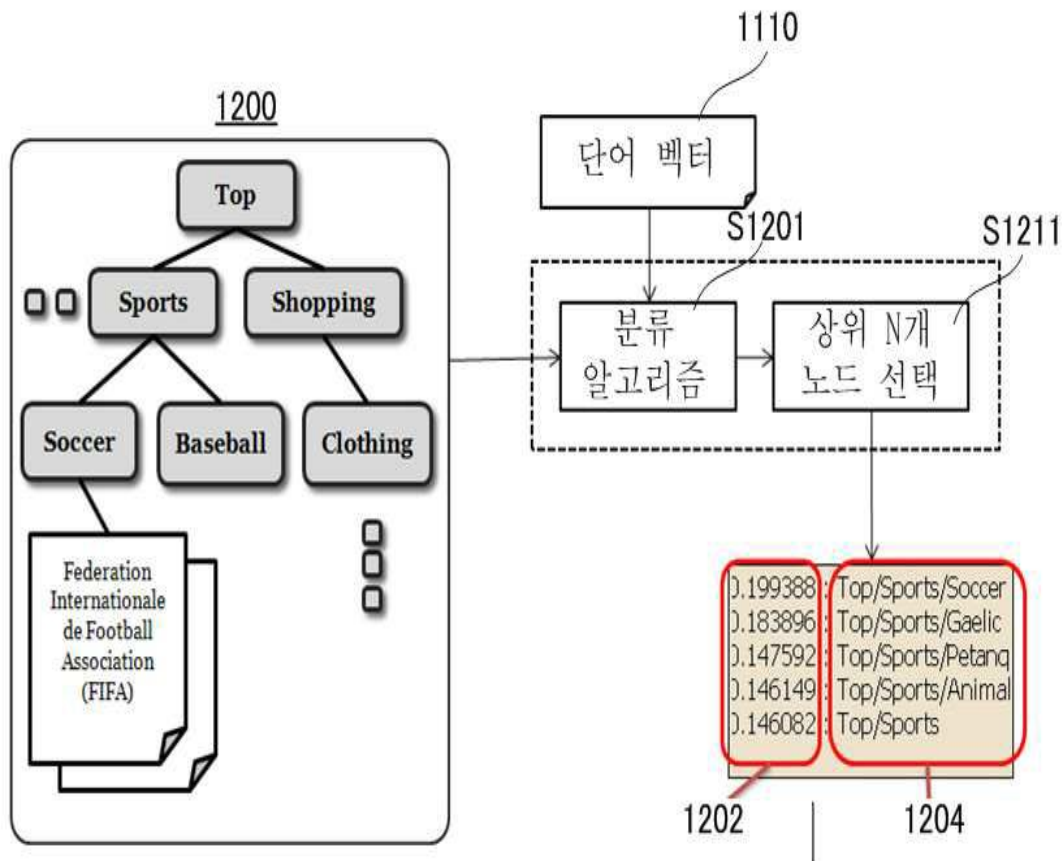
도면10



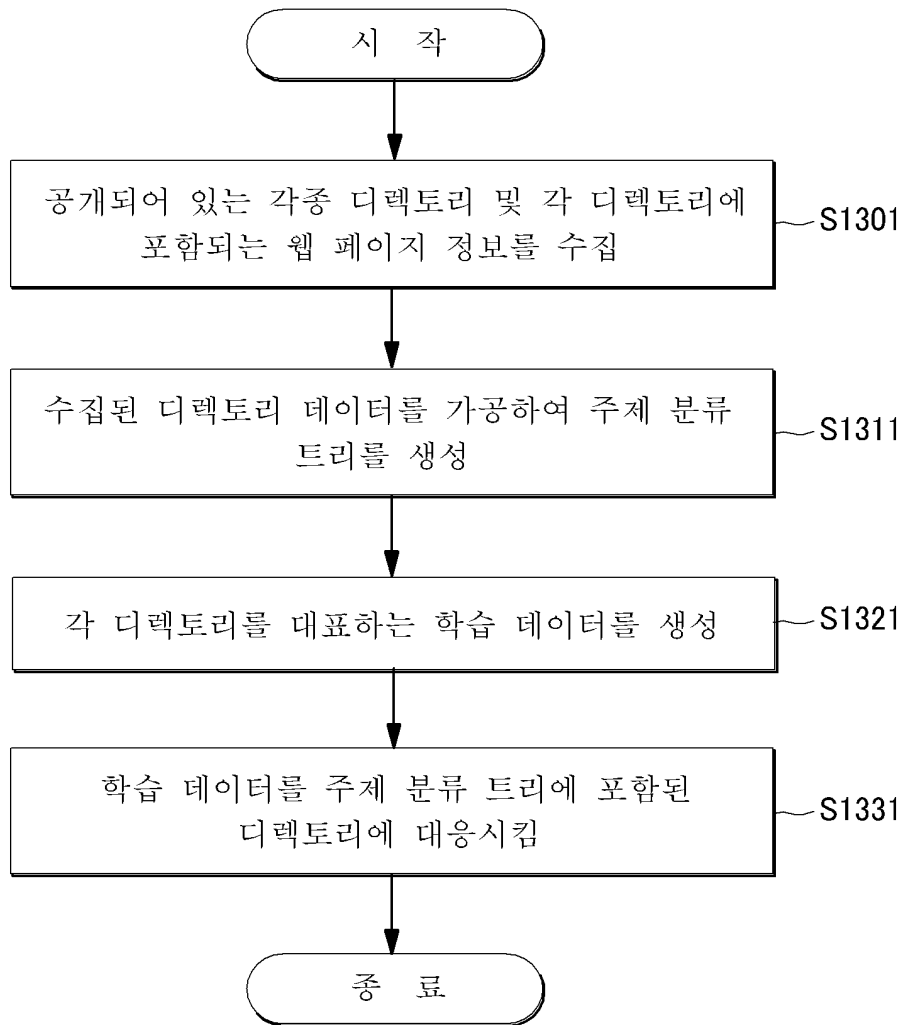
도면11



도면12



도면13



도면14

