



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0084923
(43) 공개일자 2013년07월26일

(51) 국제특허분류(Int. Cl.)
G06F 17/20 (2006.01) G06F 17/30 (2006.01)
(21) 출원번호 10-2012-0005881
(22) 출원일자 2012년01월18일
심사청구일자 2012년01월18일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
이상근
서울특별시 동대문구 제기동 153 안암골벽산아파트 101-203
소슬기
서울특별시 관악구 봉천동 1152-11 금강아미움 106동 301호
(74) 대리인
특허법인엠에이피에스

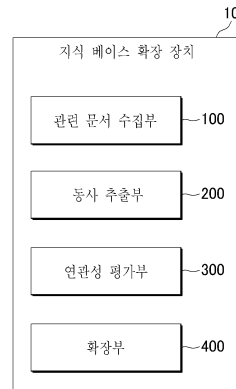
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 지식 베이스 확장 장치 및 방법

(57) 요약

본 발명은 계층 구조의 주제 분류 디렉토리의 각 카테고리별로 상기 카테고리별과 연관된 데이터를 수집하여 관련 문서를 생성하는 관련 문서 수집부; 상기 관련 문서에서 후보 동사를 추출하여 정제하는 동사 추출부; 상기 추출되어 정제된 후보 동사들 중에서 상기 카테고리별의 연관성이 높은 동사를 하나 이상 선별하는 연관성 평가부; 및 상기 선별된 동사를 상기 주제 분류 디렉토리 데이터에 통합하는 확장부를 포함하는 지식 베이스 확장 장치를 제공한다.

대표도 - 도1



특허청구의 범위

청구항 1

지식 베이스 확장 장치에 있어서,

계층 구조의 주제 분류 디렉토리의 각 카테고리별로 상기 카테고리와의 연관된 데이터를 수집하여 관련 문서를 생성하는 관련 문서 수집부;

상기 관련 문서에서 후보 동사를 추출하여 정제하는 동사 추출부;

상기 추출되어 정제된 후보 동사들 중에서 상기 카테고리와의 연관성이 높은 동사를 하나 이상 선별하는 연관성 평가부; 및

상기 선별된 동사를 상기 주제 분류 디렉토리 데이터에 통합하는 확장부를 포함하는 지식 베이스 확장 장치.

청구항 2

제 1 항에 있어서,

상기 지식 베이스 확장 장치는 상기 확장된 주제 분류 디렉토리를 활용하여 개인화 서비스를 제공하는 활용부를 더 포함하는 지식 베이스 확장 장치.

청구항 3

제 1 항에 있어서,

상기 주제 분류 디렉토리는 공개된 디렉토리 데이터(ODP: Open Directory Project)인 지식 베이스 확장 장치.

청구항 4

제 1 항에 있어서,

상기 관련 문서 수집부는 검색 엔진이 사용자의 질의에 대해 반환한 검색 결과 페이지에서 웹 페이지들의 제목 및 요약 데이터를 수집하거나, 검색 로그에서 웹 페이지들의 제목 및 요약 데이터를 수집하거나, 또는 공개된 디렉토리 데이터에 분류되어 있는 웹 페이지들의 제목 및 설명 데이터를 수집하는 지식 베이스 확장 장치.

청구항 5

제 1 항에 있어서,

상기 동사 추출부는 상기 관련 문서에서 상기 카테고리명을 포함하는 문장을 구성하는 단어 중 동사만 추출한 후, 상기 동사를 어근으로만 구분되도록 정제하여 상기 후보 동사를 추출하며, 상기 후보 동사 중 기지정된 불용어를 제거하는 지식 베이스 확장 장치.

청구항 6

제 1 항에 있어서,

상기 연관성 평가부는 상기 관련 문서에서 상기 후보 동사의 발생 빈도가 높을수록, 상기 카테고리명을 포함하는 문장에서 상기 후보 동사와 상기 카테고리간 거리가 짧을수록 높은 연관성 점수를 부여하는 지식 베이스 확장 장치.

청구항 7

제 1 항에 있어서,

상기 연관성 평가부는

문서 내에서 동사의 위치 값을 사용하여 후보 동사와 카테고리명간 문장내 평균 거리를 도출하고,

동사와 카테고리명 사이의 거리값, 및 동사의 발생 빈도를 사용하여 후보 동사의 문서내 정규화된 발생 빈도를

도출한 후,

상기 두 값을 기초로 상기 후보 동사와 상기 카테고리명간의 연관성을 계산하는 지식 베이스 확장 장치.

청구항 8

지식 베이스 확장 방법에 있어서,

(a) 계층 구조의 주제 분류 디렉토리의 각 카테고리별로 상기 카테고리명과 연관된 데이터를 수집하여 관련 문서를 생성하는 단계;

(b) 상기 관련 문서에서 후보 동사를 추출하여 정제하는 단계;

(c) 상기 추출되어 정제된 후보 동사들 중에서 상기 카테고리와의 연관성이 높은 동사를 하나 이상 선별하는 단계; 및

(d) 상기 선별된 동사를 상기 주제 분류 디렉토리 데이터에 통합하는 단계;를 포함하되,

상기 주제 분류 디렉토리는 공개된 디렉토리 데이터(ODP: Open Directory Project)인 지식 베이스 확장 방법.

청구항 9

제 8 항에 있어서,

상기 (a) 단계는

상기 공개된 디렉토리 데이터에 분류되어 있는 웹 페이지들의 제목 및 설명 데이터를 수집하는 단계;

검색 엔진이 사용자의 질의에 대해 반환한 검색 결과 페이지에서 웹 페이지들의 제목 및 요약 데이터를 수집하는 단계; 및

검색 로그에서 웹 페이지들의 제목 및 요약 데이터를 수집하는 단계;를 포함하는 지식 베이스 확장 방법.

청구항 10

제 8 항에 있어서,

상기 (b) 단계는

상기 관련 문서에서 상기 카테고리명을 포함하는 문장을 구성하는 단어 중 동사만 추출한 후, 상기 동사를 어근으로만 구분되도록 정제하여 상기 후보 동사를 추출하며, 상기 후보 동사 중 기지정된 불용어를 제거하는 지식 베이스 확장 방법.

청구항 11

제 8 항에 있어서,

상기 (c) 단계는

상기 연관성 평가부는

문서 내에서 동사의 위치 값을 사용하여 후보 동사와 카테고리명간 문장내 평균 거리를 도출하고,

동사와 카테고리명 사이의 거리값, 및 동사의 발생 빈도를 사용하여 후보 동사의 문서내 정규화된 발생 빈도를 도출한 후,

상기 두 값을 기초로 상기 후보 동사와 상기 카테고리명간의 연관성을 계산하는 지식 베이스 확장 방법.

명세서

기술분야

본 발명은 지식 베이스 확장 장치 및 방법에 관한 것이다.

배경기술

[0001]

- [0002] 최근 사용자별로 맞춤형 서비스를 제공해주는 개인화 서비스가 인기를 끌고 있다. 개인화 정보 서비스는 동일한 상황에서도 서비스를 사용하는 각 사용자에게 적합한 서비스를 제공하기 때문에 개인화가 적용되지 않은 일반 서비스에 비해 사용자의 만족도가 높고 서비스의 품질 또한 높게 평가 받는다.
- [0003] 예를 들어, 개인화된 뉴스 서비스의 경우 사용자가 요청하기 전에 사용자가 관심을 가질 만한 뉴스들을 추천해 주어 사용자의 만족도를 높일 수 있는 서비스를 제공한다. 또 다른 예로, 개인화 검색 엔진은 관심사가 다른 두 사용자가 동일한 질의를 요청한 경우, 각 사용자에게 다른 결과를 반환해준다.
- [0004] 대부분의 경우 사용자의 질의는 짧고 모호한 경우가 많기 때문에 사용자의 의도를 정확하게 나타내기 어렵다. 예컨대, '카페'라는 질의는 '온라인 커뮤니티'를 의미하는 동시에 '커피와 음료를 판매하는 장소'의 의미를 갖고 있다. 그러므로 개인화 정보 서비스에서 가장 중요한 것은 사용자의 의도를 추론하는 것이다.
- [0005] 이를 위해 사용자의 의도를 추론하기 위하여 사용자의 관심사를 공개된 디렉토리 데이터(ODP: Open Directory Project)와 같이, 계층화된 카테고리 표현하는 방법이 연구되었다. 오픈 소스인 ODP는 웹 사이트들을 주제별로 분류하기 위하여 만들어진 계층 구조의 웹 디렉토리로서, 100만개 이상의 디렉토리 및 490만개 이상의 웹 사이트를 트리 형태로 관리하고 있다. ODP는 규모와 품질 측면에서 사용자의 관심사를 나타내기에 유용한 소스라고 할 수 있다.
- [0006] 사용자의 의도를 ODP의 카테고리로 나타내는 것은 사용자 의도를 명사의 형태로 표현하는 것을 의미한다. 이는 사용자의 의도를 주제별로 표현하기에 적합한 방법이지만 사용자의 의도를 더욱 명확하게 나타내고자 하는 경우에는 충분하지 않다. 예를 들어, 사용자의 의도를 명사 형태인 '카페'라는 ODP 카테고리로 표현한 경우, 개인화 검색 엔진은 사용자의 의도가 '온라인 커뮤니티'가 아니라 '커피와 음료를 판매하는 장소'를 의미함을 추론하여 그에 맞는 결과를 반환해줄 수는 있지만, 사용자가 원하는 것이 약속 장소를 정하기 위해 적절한 카페를 찾기 위한 것인지, 창업을 하기 위해 적절한 프랜차이즈 브랜드를 물색하기 위한 것인지, 또는 아르바이트를 하기 위해 채용 공고를 낸 카페가 있는지 검색하기 위한 것인지는 추론할 수 없다는 한계가 있다.
- [0007] 따라서 개인화 서비스를 좀더 효과적으로 제공하기 위해 ODP와 같은 지식 베이스를 확장할 필요가 있다.
- [0008] 개인화 서비스와 관련하여 한국등록특허 제10-0341396호("계층 단어를 이용한 3차원 클러스터링 생성 시스템 및 그 방법")에는 유사도 기반의 클러스터링 기법에 계층 단어 개념을 접목한 구성이 개시되어 있다.
- [0009] 또한, 한국공개특허 제10-2011-0052821호("주제 분류 모듈 및 그를 이용한 문맥 광고 시스템")는 ODP를 활용하여 문맥 광고를 제공하는 구성이 개시되어 있다.

발명의 내용

해결하려는 과제

- [0010] 본 발명은 전술한 좀더 나은 개인화 서비스를 제공하는 문제를 해결하기 위한 것으로서, 그 목적은 사용자의 관심사를 좀더 효과적으로 표현할 수 있도록 주제 분류 디렉토리를 확장하는 지식 베이스 확장 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

- [0011] 상기와 같은 목적을 달성하기 위한 본 발명의 제 1 측면에 따른 지식 베이스 확장 장치는, 계층 구조의 주제 분류 디렉토리의 각 카테고리별로 상기 카테고리 및 연관된 데이터를 수집하여 관련 문서를 생성하는 관련 문서 수집부; 상기 관련 문서에서 후보 동사를 추출하여 정제하는 동사 추출부; 상기 추출되어 정제된 후보 동사들 중에서 상기 카테고리와의 연관성이 높은 동사를 하나 이상 선별하는 연관성 평가부; 및 상기 선별된 동사를 상기 주제 분류 디렉토리 데이터에 통합하는 확장부를 포함하는 것을 특징으로 한다.
- [0012] 상기와 같은 목적을 달성하기 위한 본 발명의 제 2 측면에 따른 지식 베이스 확장 방법은, (a) 계층 구조의 주제 분류 디렉토리의 각 카테고리별로 상기 카테고리 및 연관된 데이터를 수집하여 관련 문서를 생성하는 단계; (b) 상기 관련 문서에서 후보 동사를 추출하여 정제하는 단계; (c) 상기 추출되어 정제된 후보 동사들 중에서 상기 카테고리와의 연관성이 높은 동사를 하나 이상 선별하는 단계; 및 (d) 상기 선별된 동사를 상기 주제 분류 디렉토리 데이터에 통합하는 단계;를 포함하되, 상기 주제 분류 디렉토리는 공개된 디렉토리 데이터(ODP: Open Directory Project)인 것을 특징으로 한다.

발명의 효과

- [0013] 본 발명은 지식 베이스 확장 장치 및 방법에 있어, 사용자의 관심사를 좀더 효과적으로 표현할 수 있도록 주제 분류 디렉토리를 확장하는 효과를 얻는다.
- [0014] 구체적으로, 사용자의 관심사를 명사와 연관 동사의 조합으로 표현할 수 있다.
- [0015] 따라서 사용자 관심사 추론 성능이 향상되므로 좀더 향상된 개인화 서비스를 제공할 수 있다.
- [0016] 또한 다양한 외부 소스를 활용하여 효과적으로 연관 동사를 추출하므로 추출되어 통합된 연관 동사의 정확도가 높다.

도면의 간단한 설명

- [0017] 도 1은 본 발명에 따른 지식 베이스 확장 장치의 구조를 도시함.
- 도 2은 ODP의 실시예를 도시함.
- 도 3는 본 발명에 따른 확장된 ODP의 실시예를 도시함.
- 도 4는 본 발명에 따른 지식 베이스 확장 방법의 흐름을 도시함.
- 도 5은 본 발명에 따른 관련 문서 수집 단계의 흐름을 도시함.
- 도 6은 본 발명에 따른 후보 동사 추출 단계의 흐름을 도시함.
- 도 7은 본 발명에 따른 연관성 높은 동사 선별 단계의 흐름을 도시함.

발명을 실시하기 위한 구체적인 내용

- [0018] 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.
- [0019] 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.
- [0020] 도 1은 본 발명에 따른 지식 베이스 확장 장치(10)를 나타낸 블록도이다.
- [0021] 본 발명에 따른 지식 베이스 확장 장치(10)는 전술한 바와 같이 향상된 개인화 서비스를 제공하기 위해 지식 베이스를 확장한다. 지식 베이스는 각 사용자의 관심사를 주제별로 분류해놓은 것으로, 계층 구조를 갖는 주제 분류 디렉토리의 형태로 표현된다. 명사의 형태로 표현되어 있는 지식 베이스를 연관 동사까지 포함하도록 확장하여, 사용자의 관심사를 좀더 효과적으로 표현하는 것이 본 발명에 따른 지식 베이스 확장 장치(10)의 목적이다.
- [0022] 이를 위해 본 발명에 따른 지식 베이스 확장 장치(10)는 주제 분류 디렉토리의 각 카테고리별로 연관된 데이터를 수집하여 관련 문서를 생성하는 관련 문서 수집부(100), 관련 문서에서 후보 동사를 추출하여 정제하는 동사 추출부(200), 추출되어 정제된 후보 동사들 중에서 카테고리와의 연관성이 높은 동사를 하나 이상 선별하는 연관성 평가부(300), 선별된 동사를 주제 분류 디렉토리 데이터에 통합하는 확장부(400)를 포함한다.
- [0023] 또한 일실시예에서 지식 베이스 확장 장치(10)는 확장된 주제 분류 디렉토리를 활용하여 개인화 서비스를 제공하는 활용부(미도시)를 더 포함할 수 있다. 또는 지식 베이스 확장 장치가 개인화 서비스 장치의 사용자 의도 추론 모듈 내에 포함되어 사용되는 것도 가능하다.
- [0024] 바람직한 실시예에서, 주제 분류 디렉토리는 공개된 디렉토리 데이터(ODP: Open Directory Project)를 사용한다. ODP는 사용자가 자발적으로 참여하여 월드 와이드 웹 사이트들의 링크를 계층형 온톨로지 형태로 수집하여 놓은 개방형 디렉토리이다. 원래의 도메인 명이 `directory.mozilla.org`이었기 때문에 Dmoz라고도 불린다. 현재 ODP 사이트의 주소는 `http://www.dmoz.org`이다.

- [0025] 도 2에는 ODP의 실시예가 도시되어 있다.
- [0026] 본 실시예에서 사용자는 Computer>Company>Apple Inc. 순으로 ODP 디렉토리를 탐색하였다. 'apple'이라는 질의는 이렇게 '애플이라는 이름의 IT 회사'로 해석될 수도 있고, '사과라는 과일'을 뜻할 수도 있다. 도 2은 이러한 사용자의 관심사가 ODP 등의 주제 분류 디렉토리로 어떻게 표현되는지 도시하고 있다. 본 실시예에서는 사용자가 Computer>Company>Apple Inc. 순으로 ODP 디렉토리를 탐색하였지만, 또다른 사용자는 Shopping>Food>Fruit>Apple 순으로 ODP 디렉토리를 탐색할 수도 있을 것이다.
- [0027] 이때, 전술한 바와 같이 개인화 검색 엔진은 이러한 주제 분류 디렉토리를 이용하여, 사용자가 단순히 'apple'이라는 질의만을 입력하였어도, 각 사용자에게 맞춰 어떤 사용자에게는 회사인 애플사에 대한 정보가 상위에 표시된 검색 결과를, 또다른 사용자에게는 과일인 사과에 대한 정보가 상위에 표시된 검색 결과를 반환해줄 수 있다. 이러한 사용자의 의도 추론에는 유사도 계산이 사용된다. 예를 들어 도 2에 도시된 사용자에게 대한 유사도 점수는 Computer>Company>Apple Inc. 쪽이 0.86으로 Shopping>Food>Fruit>Apple 쪽의 0.2보다 높으므로, 이 사용자에게 대해서는 Apple Inc.가 상위에 표시된 검색 결과를 반환할 것이다.
- [0028] 도 3는 본 발명에 따른 확장된 ODP의 실시예를 도시하고 있다.
- [0029] 도 2에 도시된 사용자의 의도를 좀더 구체적으로 들여다보면, 이 사용자가 'apple'이라는 질의를 입력한 이유는 아이튠즈 등 콘텐츠를 다운로드하고 싶어서일 수도, 아이폰 등 제품을 구매하고 싶어서일 수도, 또는 CEO인 스티브 잡스에 대한 정보를 찾고 싶어서일 수도 있다. 각 경우에 대해 'download', 'buy', 'search'를 'Apple Inc.'에 대한 연관 동사로 추출할 수 있다.
- [0030] 연관 동사까지 유사도 계산을 확장하면 사용자의 의도를 좀더 정확하게 추론하여 좀더 향상된 개인화 서비스를 제공할 수 있다. 도 3에 도시된 사용자의 경우, (apple, download)의 유사도 점수는 0.67, (apple, buy)의 유사도 점수는 0.78, (apple, search)의 유사도 점수는 0.23이다. 따라서 이 사용자는 애플의 제품 구매에 관심이 있음을 알 수 있다. 따라서 개인화 검색 엔진은 애플사에 대한 정보 중에서도 제품에 대한 정보를 상위에 표시해줄 수 있다.
- [0031] 도 4는 본 발명에 따른 지식 베이스 확장 방법의 흐름을 도시하고 있다.
- [0032] 계층 구조의 주제 분류 디렉토리의 각 카테고리별로 연관된 데이터를 수집하여 관련 문서를 생성한다(S510). 생성된 관련 문서에서 후보 동사를 추출하여 정제하고(S520), 추출되어 정제된 후보 동사들 중에서 카테고리와의 연관성이 높은 동사를 하나 이상 선별한 후(S530), 선별된 동사를 주제 분류 디렉토리 데이터에 통합한다(S540).
- [0033] 도 5을 통해 관련 문서 생성 단계(S510)에 대해 더 자세히 살펴보자.
- [0034] 본 발명에 따른 지식 베이스 확장 방법은 다양한 외부 소스를 활용하여 카테고리별 연관된 동사를 많이 포함할 만한 문서를 생성한다. 즉, 여러 소스로부터 주어진 카테고리에 연관된 데이터를 수집하여 하나의 문서로 통합한다. 중요한 것은 카테고리와 연관성이 높은 동사를 포함하고 있을 만한 데이터를 찾는 것이므로, 검색 결과 페이지, ODP에 분류된 웹 페이지, 검색 로그와 같은 다양한 소스를 활용한다.
- [0035] 관련 문서를 생성하기 위해 본 발명에 따른 지식 베이스 확장 방법이 활용하는 정보 원천의 첫번째 예는 ODP에 분류되어 있는 웹 페이지들의 제목 및 설명 데이터이다(S610). ODP에 분류된 웹 페이지들은 각 카테고리별 연관된 페이지이기 때문에 연관된 동사를 많이 포함하고 있으므로 좋은 정보 수집 원천이다. 이때 하위에 분류된 웹 페이지의 개수가 적은 경우를 대비하여 상위 카테고리들에 분류된 웹 페이지들을 함께 활용할 수도 있다.
- [0036] 관련 문서를 생성하기 위해 본 발명에 따른 지식 베이스 확장 방법이 활용하는 정보 원천의 두번째 예는 카테고리를 질의로 사용하였을 때 검색 엔진이 반환한 검색 결과 페이지에 포함되어 있는 웹 페이지들의 제목 및 요약 데이터이다(S620). 검색 결과 페이지는 검색 엔진에서 질의에 대하여 반환한 웹 페이지들의 목록으로, 어떤 질의에 대해서도 적절하고 많은 데이터를 반환하기 때문에, 카테고리를 질의로 사용한다면 연관된 동사를 많이 포함한 웹 페이지 정보를 수집할 수 있다.
- [0037] 관련 문서를 생성하기 위해 본 발명에 따른 지식 베이스 확장 방법이 활용하는 정보 원천의 세번째 예는 검색 로그에 포함되어 있는 웹 페이지들의 제목 및 요약 데이터이다(S630). 검색 로그는 사용자가 검색할 때 사용하는 질의나 클릭한 URL, 클릭한 시간 등과 같은 웹 사용자들의 사용 패턴을 기록한 정보로, 사용자들이 자주 클릭한 웹 페이지 정보들을 포함하고 이러한 웹 페이지들은 주어진 질의와 연관된 많은 동사들을 포함하기 때문에

검색 로그는 유용한 외부 소스이다.

[0038] 구체적으로, 본 발명에 따른 지식 베이스 확장 방법은 검색 로그에서 효과적으로 정보를 추출하기 위해, 디렉토리 및 유사도가 높은 질의들에 대하여 클릭된 웹 페이지들의 제목과 URL, 요약 정보를 이용하여 임시 문서들을 생성한 후, 주어진 ODP 카테고리에 대하여 임시 문서와의 유사도를 측정하여 연관된 질의의 순위를 정하고, 상위에 랭크된 질의들에 대한 임시 문서를 통합하여 하나의 문서를 생성한다.

[0039] 정보 수집이 완료되면, 수집한 정보를 하나의 관련 문서로 통합한다(S640). 이때, 전술한 정보 수집 단계들(S610, S620, S630)을 수행하는 순서는 중요하지 않으며, 본 발명에 따른 지식 베이스 확장 방법이 관련 문서를 생성하기 위해 활용하는 정보 원천도 전술한 세가지 예에 한정되지 않는다.

[0040] 도 6은 생성된 관련 문서에서 후보 동사를 추출하는 단계(S520)를 도시하고 있다.

[0041] 먼저, 생성된 관련 문서에서 카테고리명을 포함하는 문장을 추출한 후(S710), 각 문장을 단어 단위의 토큰으로 분할한다(S720). 예를 들어, 관련 문서에 카테고리명 'apple'을 포함하는 'Apple designs and creates iPod and iTunes'라는 문장이 있다면, 'Apple', 'designs', 'and', 'creates', 'iPod', 'and', 'iTunes'라는 토큰들이 추출된다.

[0042] 다음, 각 토큰을 태깅하고 품사가 동사로 태깅된 토큰만 남긴다(S730). 이를 위한 도구로 본 발명에 따른 지식 베이스 확장 방법은 Stanford Natural Language Processing Tool을 사용할 수 있다. 예를 들어, 전술한 예에서 'designs'와 'creates'라는 토큰만 남는다.

[0043] 다음, 각 토큰의 어근만 남겨 다양한 형태의 동일한 동사들을 하나의 동사로 인지한다(S740). 이를 위한 도구로 본 발명에 따른 지식 베이스 확장 방법은 Porter stemmer를 사용할 수 있다. 예를 들어, 'designs'와 'designing'은 어근 'design'만 남는다.

[0044] 다음, 추출된 동사들 중에서 기지정된 불용어를 제거한다(S750). 불용어는 카테고리에 상관없이 자주 등장하는 동사로 'is', 'are'와 같이 연관 동사를 추출하는 데에 의미가 없는 동사를 의미한다.

[0045] 도 7을 통해 연관성 높은 동사 선별 단계(S530)에 대해 더 자세히 살펴보자.

[0046] 이 단계에서는 후보 동사 중 지식 베이스 확장에 사용될 동사를 선정하기 위해 연관성 점수가 계산된다. 연관성 점수란 후보 동사가 주어진 카테고리 및 얼마나 연관되어 있는지를 의미한다. 문서 내에서 동사의 발생 빈도가 높을수록, 문장 내에서 동사와 카테고리명간 거리가 짧을수록 높은 점수를 부여된다.

[0047] 따라서 먼저 후보 동사와 카테고리명간 문장내 평균 거리를 산출하고(S810), 후보 동사의 문서 내 정규화된 발생 빈도를 산출한 후(S820), 이 두 가지 정보를 바탕으로 후보 동사의 연관성을 산출한다(S830).

[0048] 구체적으로 살펴보면, 후보 동사와 카테고리간 연관성 점수는 다음 수식들을 사용하여 계산된다.

수학식 1

$$avd(v_i, c) = \frac{1}{vf(v_i)} \sum_{l \in loc(v_i)} dist(l, c)$$

수학식 2

$$nvf(v_i) = \frac{vf(v_i)}{\sum_{w \in V} vf(w)}$$

수학식 3

$$RelScore(v_i, c) = \frac{nvf(v_i)}{avd(v_i, c)}$$

[0051]

[0052] 여기에서 v_i 는 i 번째 동사이며, c 는 카테고리명이다. $nvf(v_i)$ 는 후보 동사의 문서내 정규화된 발생 빈도, $avd(v_i, c)$ 는 후보 동사와 카테고리명간 문장내 평균 거리를 의미한다.

[0053] $vf(v_i)$ 는 동사의 발생 빈도를 의미하며, V 는 문서에서 추출된 동사의 전체 집합을 의미한다. 즉, $nvf(v_i)$ 는 동사의 발생 빈도를 문서 내에서 발생한 모든 동사의 발생 빈도의 합으로 정규화한 값을 의미한다.

[0054] 또한, $loc(v_i)$ 는 동사 v_i 의 위치 집합을 나타낸다. 이는 문서 내에서 동일한 동사가 여러번 등장하기 때문에 각 위치값을 얻기 위한 함수이다. $dist(l, c)$ 는 동사와 디렉토리명 사이의 거리값을 얻기 위한 함수이다.

[0055] 모든 후보 동사에 대해 연관성 점수가 산출되면, 그중 연관성 점수가 높은 후보 동사를 일정 개수 선별하여, 해당 카테고리의 연관 동사로 지정한다(S840).

[0056] 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.

[0057] 본 발명의 범위는 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

[0058] 10: 지식 베이스 확장장치

100: 관련 문서 수집부

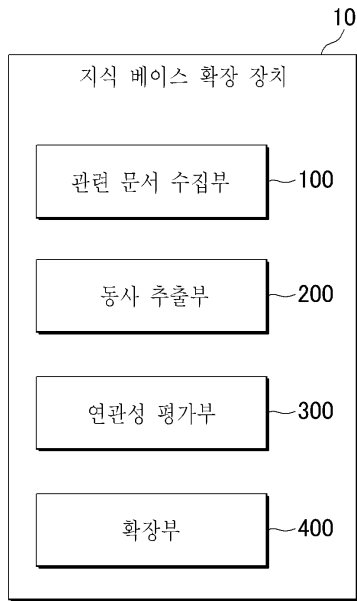
200: 동사 추출부

300: 연관성 평가부

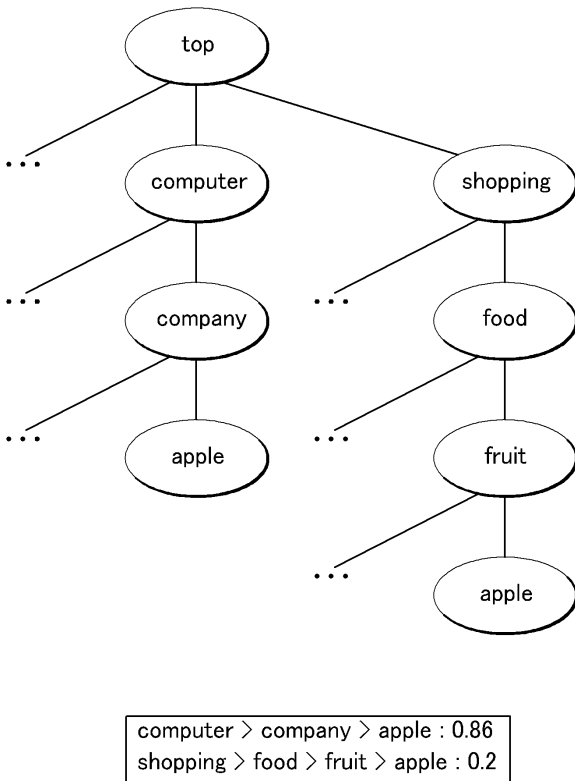
400: 확장부

도면

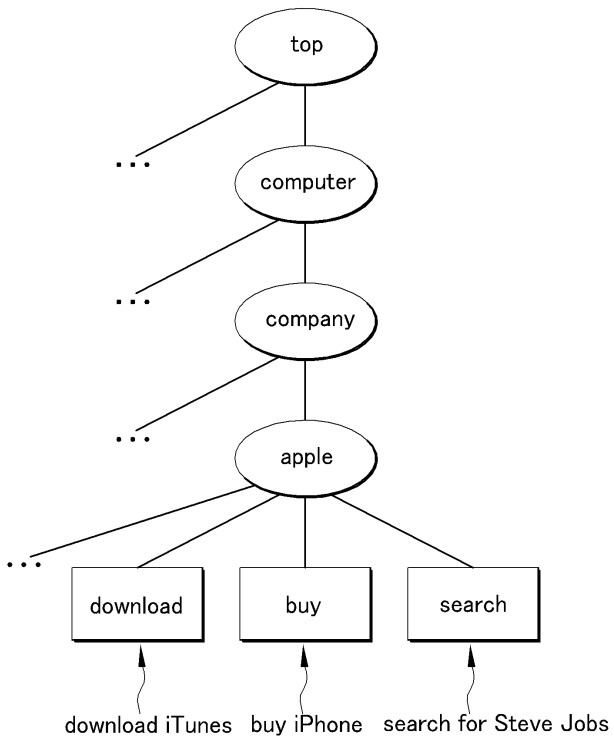
도면1



도면2

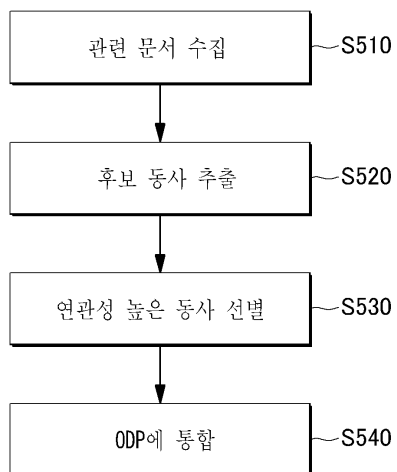


도면3



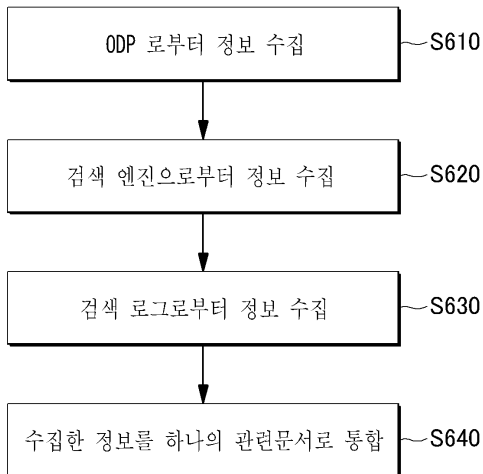
computer > company > apple, download : 0.67
 computer > company > apple, buy : 0.78
 computer > company > apple, search : 0.23

도면4



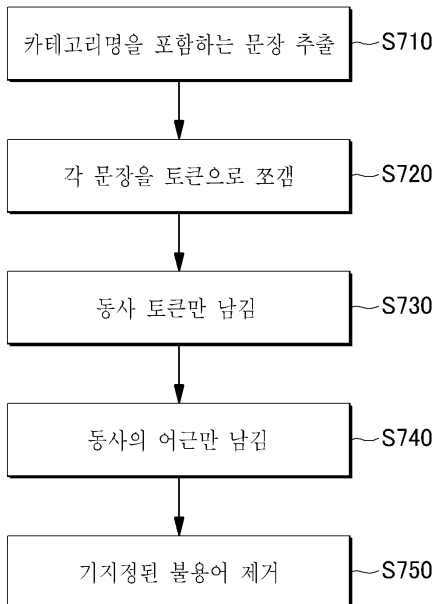
도면5

S510



도면6

S520



도면7

S530

