



(51) 국제특허분류:

G06F 17/30 (2006.01)

(21) 국제출원번호:

PCT/KR2018/002834

(22) 국제출원일:

2018 년 3 월 9 일 (09.03.2018)

(25) 출원언어:

한국어

(26) 공개언어:

한국어

(30) 우선권정보:

10-2017-0044683 2017 년 4 월 6 일 (06.04.2017) KR

10-2017-0085316 2017 년 7 월 5 일 (05.07.2017) KR

(71) 출원인: 네이버 주식회사 (NAVER CORPORATION)

[KR/KR]; 13561 경기도 성남시 분당구 불정로 6 그린팩토리, Gyeonggi-do (KR).

(72) 발명자: 민혜진 (MIN, Hye-Jin); 13561 경기도 성남시 분당구 불정로 6 그린팩토리, Gyeonggi-do (KR). 김진홍 (KIM, Jinhong); 13561 경기도 성남시 분당구 불정로 6 그린팩토리, Gyeonggi-do (KR). 박찬훈 (PARK, Chan Hoon); 13561 경기도 성남시 분당구 불정로 6 그린팩토리, Gyeonggi-do (KR). 김광현 (KIM, Kwang Hyun); 13561 경기도 성남시 분당구 불정로 6 그린팩토리, Gyeonggi-do (KR).

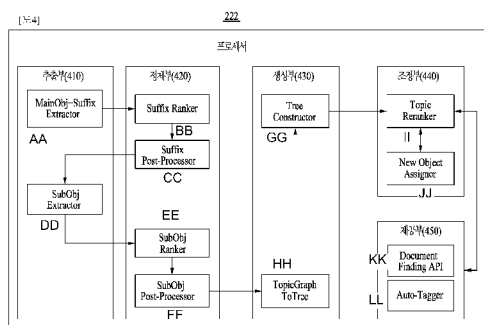
(74) 대리인: 양성보 (YANG, Sungbo); 06099 서울시 강남구 선릉로125길 14 삼성빌딩 2층, 피엔티특허법률사무소, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: AUTOMATIC EXTRACTION AND STRUCTURIZATION, BY SUBJECT, OF SUB-TOPIC OF QUERY

(54) 발명의 명칭: 주제별 질의의 서브토픽 자동 추출 및 구조화



222 ... Processor
410 ... Extracting unit
420 ... Refining unit
430 ... Generating unit
440 ... Adjusting unit
450 ... Providing unit
AA ... MainObj+Suffix Extractor
BB ... Suffix Ranker
CC ... Suffix Post-Processor
DD ... SubObj Extractor
EE ... SubObj Ranker
FF ... SubObj Post-Processor
GG ... Tree Constructor
HH ... TopicGraph ToTree
II ... Topic Reranker
JJ ... New Object Assigner
KK ... Document Finding API
LL ... Auto-Tagger

(57) Abstract: A technique for automatically extracting and structuring a sub-topic suitable for a query is disclosed. A topic structuring method can comprise the steps of: extracting, by subject, a sub-topic related to a subject; generating a topic tree for the sub-topic by using hierarchical information of the subject; and hierarchically providing the sub-topic as a search word related to a query according to a topic tree of a subject to which the query belongs, when the query for a search is provided.

(57) 요약서: 질의에 적합한 서브토픽을 자동 추출하고 구조화 하는 기술이 개시된다. 토픽 구조화 방법은, 주제 별로 상기 주제와 관련된 서브토픽을 추출하는 단계; 상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 단계; 및 검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로 제공하는 단계를 포함할 수 있다.

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 주제별 질의의 서브토픽 자동 추출 및 구조화 기술분야

- [1] 아래의 설명은 질의에 적합한 서브토픽을 자동 추출하고 구조화 하는 기술에 관한 것이다.

배경기술

- [2] 검색 시스템은 사용자가 입력한 질의의 검색 결과를 제공할 때 검색 조건에 매칭된 문서들 외에 사용자의 추가 탐색을 도와주는 여러 가지 기능을 함께 제공한다. 사용자의 추가 탐색을 도와주는 대표적인 것들로 연관 검색어, 연관 태그, 검색어 자동 완성 등이 있다. 이들은 단어쌍이 함께 나타나는 정도(Co-occurrence)를 기반으로 질의와 자주 등장하는 검색어나 태그를 파악하는 것을 기본으로 한다.
- [3] 예를 들어, 공개특허공보 제10-2012-0096806호(공개일 2012년 08월 31일)에는 사용자 단말의 위치 정보에 근거하여 추천 검색어를 선정하여 사용자에게 제공하는 검색어 추천 시스템 및 검색어 추천 방법이 개시되어 있다.
- [4] 한편, 쇼핑 검색 시스템에서는 쇼핑 의도가 있는 질의의 경우, 제품의 브랜드, 색상, 가격 등과 같은 계층 정보를 활용하여 상품의 탐색을 도와주는 기능을 제공하기도 한다.

발명의 상세한 설명

기술적 과제

- [5] 주제가 분명한 질의의 토픽을 제공하고 효율적인 정보 탐색을 위한 토픽의 구조화가 필요하다.
- [6] 특정 주제가 주어졌을 때 해당 주제에 적합한 서브토픽만을 추출하고 특정 주제에 대해 자동으로 구축된 계층 정보를 활용하여 사용자가 원하는 정도(세분화)에 따라 적절히 구조화하여 제시하는 방법을 제공한다.

과제 해결 수단

- [7] 컴퓨터로 구현되는 토픽 구조화 방법에 있어서, 주제 별로 상기 주제와 관련된 서브토픽을 추출하는 단계; 상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 단계; 및 검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로 제공하는 단계를 포함하는 토픽 구조화 방법을 제공한다.
- [8] 일 측면에 따르면, 상기 추출하는 단계는, 상기 주제를 결정하는 핵심 오브젝트와 관련된 단어들을 분석하여 상기 서브토픽을 추출할 수 있다.
- [9] 다른 측면에 따르면, 상기 서브토픽을 문서 출현 빈도와 검색 빈도 중 적어도 하나에 따라 필터링 하는 단계를 더 포함할 수 있다.

- [10] 또 다른 측면에 따르면, 상기 서브토픽을 유의어 또는 서브스트링(부분문자열) 관계에 따라 클러스터링 하여 각 클러스터의 대표를 선정하는 단계를 더 포함할 수 있다.
- [11] 또 다른 측면에 따르면, 상기 생성하는 단계는, 상기 서브토픽을 상기 계층 정보의 각 클래스 이름으로 레이블링(labeling) 하여 상기 토픽 트리를 생성할 수 있다.
- [12] 또 다른 측면에 따르면, 상기 생성하는 단계는, 상기 서브토픽에 대해 워드임베딩 데이터에서 유사 단어를 추출하는 단계; 상기 유사 단어를 유의어 또는 서브스트링(부분문자열) 관계에 따라 클러스터링 하는 단계; 및 상기 클러스터링 된 단어를 언어 분류학 상의 각 클래스에 매핑함으로써 레이블링 하는 단계를 포함할 수 있다.
- [13] 또 다른 측면에 따르면, 상기 토픽 트리의 너비(breadth)와 깊이(depth) 중 적어도 하나를 줄여 상기 토픽 트리에 대한 리밸런싱(rebalancing)을 수행하는 단계를 더 포함할 수 있다.
- [14] 또 다른 측면에 따르면, 상기 제공하는 단계는, 상기 질의와 상기 서브토픽의 연관성을 나타내는 주제 점수와, 상기 서브토픽에 대응되는 문서 수, 및 상기 질의에 대한 정답성 토픽 여부 중 적어도 하나의 조건에 따라 상기 서브토픽을 필터링 하는 단계를 포함할 수 있다.
- [15] 컴퓨터로 구현되는 검색 결과 제공 방법에 있어서, 검색을 위한 질의가 주어지면 상기 질의에 대응되는 검색 결과를 제공하는 단계; 상기 질의가 속한 주제의 계층 정보에 따라 상기 질의에 대한 연관 검색어로서 상기 주제와 관련된 서브토픽을 복수의 뎀스(depth)를 가진 계층형으로 제공하는 단계; 및 상기 서브토픽 중 적어도 하나의 검색어가 선택되는 경우 상기 선택된 검색어를 포함한 상기 질의에 대응되는 검색 결과를 제공하는 단계를 포함하는 검색 결과 제공 방법을 제공한다.
- [16] 컴퓨터 시스템과 결합되어 토픽 구조화 방법을 실행시키기 위해 컴퓨터 판독 가능한 기록 매체에 기록된 컴퓨터 프로그램에 있어서, 상기 토픽 구조화 방법은, 주제 별로 상기 주제와 관련된 서브토픽을 추출하는 단계; 상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 단계; 및 검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로 제공하는 단계를 포함하는, 컴퓨터 프로그램을 제공한다.
- [17] 컴퓨터로 구현되는 토픽 구조화 시스템에 있어서, 컴퓨터가 판독 가능한 명령을 실행하도록 구현되는 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 주제 별로 상기 주제와 관련된 서브토픽을 추출하는 추출부; 상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 생성부; 및 검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로 제공하는

제공부를 포함하는 토픽 구조화 시스템을 제공한다.

발명의 효과

- [18] 본 발명의 실시예들에 따르면, 특정 주제가 주어졌을 때 해당 주제에 적합한 서브토픽만을 추출하고 특정 주제에 적절한 계층 정보를 자동으로 구축한 후 사용자가 원하는 정도(세분화)에 따라 적절히 구조화하여 제시함으로써 주제에 적합하고 해당 질의와 관련된 세부적인 속성/소주제들을 사용자가 효율적으로 파악하고 실질적인 추가 탐색을 할 수 있도록 도와주는데 기여할 수 있다.

도면의 간단한 설명

- [19] 도 1은 본 발명의 일 실시예에 따른 네트워크 환경의 예를 도시한 도면이다.
- [20] 도 2는 본 발명의 일 실시예에 있어서, 전자 기기 및 서버의 내부 구성을 설명하기 위한 블록도이다.
- [21] 도 3은 본 발명의 일 실시예에 있어서 패턴화 된 질의를 계층화 하는 과정의 예시를 도시한 도면이다.
- [22] 도 4는 본 발명의 일 실시예에 따른 서버의 프로세서가 포함할 수 있는 구성요소의 예를 도시한 도면이다.
- [23] 도 5는 본 발명의 일 실시예에 따른 서버가 수행할 수 있는 방법의 예를 도시한 흐름도이다.
- [24] 도 6은 본 발명의 일 실시예에 따라 '팜' 및 '감자'이라는 질의에 대하여 서브토픽 후보를 필터링 및 그룹핑하는 과정의 일 예를 나타낸 것이다.
- [25] 도 7은 본 발명의 일 실시예에 있어서 계층 정보를 구축하는 과정의 일례를 설명하기 위한 순서도이다.
- [26] 도 8은 여행 주제에 대해서 클러스터링 및 언어 분류학을 이용하여 구축한 계층 정보의 일 예를 보여준다.
- [27] 도 9는 본 발명의 일 실시예에 따라 구축된 토픽 네트워크를 깊이가 2인 트리(2-depth 트리)로 변환하는 과정의 예시를 도시한 것이다.
- [28] 도 10은 본 발명의 일 실시예에 있어서 트리 리밸런싱 과정의 일례를 설명하기 위한 순서도이다.
- [29] 도 11 내지 도 12는 본 발명의 일 실시예에 따른 트리 리밸런싱 과정의 예시들을 나타내는 도면이다.
- [30] 도 13 내지 도 14는 본 발명의 일 실시예에 있어서 2 뎁스 토픽 구조가 반영된 검색 결과 화면의 예시를 도시한 도면이다.

발명의 실시를 위한 최선의 형태

- [31] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [32]
- [33] 본 발명의 실시예들은 질의에 적합한 서브토픽을 자동 추출하고 구조화 하는 기술에 관한 것이다.
- [34] 본 명세서에서 구체적으로 개시되는 것들을 포함하는 실시예들은 주제가

분명한 질의의 토픽을 제공하고 효율적인 정보 탐색을 위한 토픽의 구조화가 가능하고, 이를 통해 정확성, 효율성, 확장성, 비용 절감 등의 측면에 있어서 상당한 장점들을 달성한다.

- [35] 도 1은 본 발명의 일 실시예에 따른 네트워크 환경의 예를 도시한 도면이다. 도 1의 네트워크 환경은 복수의 전자 기기들(110, 120, 130, 140), 복수의 서버들(150, 160) 및 네트워크(170)를 포함하는 예를 나타내고 있다. 이러한 도 1은 발명의 설명을 위한 일례로 전자 기기의 수나 서버의 수가 도 1과 같이 한정되는 것은 아니다.
- [36] 복수의 전자 기기들(110, 120, 130, 140)은 컴퓨터 장치로 구현되는 고정형 단말이거나 이동형 단말일 수 있다. 복수의 전자 기기들(110, 120, 130, 140)의 예를 들면, 스마트폰(smart phone), 휴대폰, 태블릿 PC, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player) 등이 있다. 일례로 제1 전자 기기(110)는 무선 또는 유선 통신 방식을 이용하여 네트워크(170)를 통해 다른 전자 기기들(120, 130, 140) 및/또는 서버(150, 160)와 통신할 수 있다.
- [37] 통신 방식은 제한되지 않으며, 네트워크(170)가 포함할 수 있는 통신망(일례로, 이동통신망, 유선 인터넷, 무선 인터넷, 방송망)을 활용하는 통신 방식뿐만 아니라 기기들간의 근거리 무선 통신 역시 포함될 수 있다. 예를 들어, 네트워크(170)는, PAN(personal area network), LAN(local area network), CAN(campus area network), MAN(metropolitan area network), WAN(wide area network), BBN(broadband network), 인터넷 등의 네트워크 중 하나 이상의 임의의 네트워크를 포함할 수 있다. 또한, 네트워크(170)는 버스 네트워크, 스타 네트워크, 링 네트워크, 메쉬 네트워크, 스타-버스 네트워크, 트리 또는 계층적(hierarchical) 네트워크 등을 포함하는 네트워크 토폴로지 중 임의의 하나 이상을 포함할 수 있으나, 이에 제한되지 않는다.
- [38] 서버(150, 160) 각각은 복수의 전자 기기들(110, 120, 130, 140)과 네트워크(170)를 통해 통신하여 명령, 코드, 파일, 콘텐츠, 서비스 등을 제공하는 컴퓨터 장치 또는 복수의 컴퓨터 장치들로 구현될 수 있다.
- [39] 일례로, 서버(160)는 네트워크(170)를 통해 접속한 제1 전자 기기(110)로 어플리케이션의 설치를 위한 파일을 제공할 수 있다. 이 경우 제1 전자 기기(110)는 서버(160)로부터 제공된 파일을 이용하여 어플리케이션을 설치할 수 있다. 또한, 제1 전자 기기(110)가 포함하는 운영체제(Operating System, OS)나 적어도 하나의 프로그램(일례로 브라우저나 상기 설치된 어플리케이션)의 제어에 따라 서버(150)에 접속하여 서버(150)가 제공하는 서비스나 콘텐츠를 제공받을 수 있다. 예를 들어, 제1 전자 기기(110)가 어플리케이션의 제어에 따라 네트워크(170)를 통해 서비스 요청 메시지를 서버(150)로 전송하면, 서버(150)는 서비스 요청 메시지에 대응하는 코드를 제1 전자 기기(110)로 전송할 수 있고, 제1 전자 기기(110)는 어플리케이션의 제어에 따라 코드에 따른 화면을 구성하여

표시함으로써 사용자에게 콘텐츠를 제공할 수 있다.

- [40] 도 2는 본 발명의 일 실시예에 있어서, 전자 기기 및 서버의 내부 구성을 설명하기 위한 블록도이다. 도 2에서는 하나의 전자 기기에 대한 예로서 제1 전자 기기(110), 그리고 하나의 서버에 대한 예로서 서버(150)의 내부 구성을 설명한다. 다른 전자 기기들(120, 130, 140)이나 서버(160) 역시 동일한 또는 유사한 내부 구성을 가질 수 있다.

- [41] 제1 전자 기기(110)와 서버(150)는 메모리(211, 221), 프로세서(212, 222), 통신 모듈(213, 223) 그리고 입출력 인터페이스(214, 224)를 포함할 수 있다. 메모리(211, 221)는 컴퓨터에서 판독 가능한 기록 매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 또한, 메모리(211, 221)에는 운영체제나 적어도 하나의 프로그램 코드(일례로 제1 전자 기기(110)에 설치되어 구동되는 어플리케이션 등을 위한 코드)가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 메모리(211, 221)와는 별도의 컴퓨터에서 판독 가능한 기록 매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록 매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록 매체를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록 매체가 아닌 통신 모듈(213, 223)을 통해 메모리(211, 221)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 어플리케이션의 설치 파일을 배포하는 파일 배포 시스템(일례로 상술한 서버(160))이 네트워크(170)를 통해 제공하는 파일들에 의해 설치되는 프로그램(일례로 상술한 어플리케이션)에 기반하여 메모리(211, 221)에 로딩될 수 있다.

- [42] 프로세서(212, 222)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(211, 221) 또는 통신 모듈(213, 223)에 의해 프로세서(212, 222)로 제공될 수 있다. 예를 들어 프로세서(212, 222)는 메모리(211, 221)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

- [43] 통신 모듈(213, 223)은 네트워크(170)를 통해 제1 전자 기기(110)와 서버(150)가 서로 통신하기 위한 기능을 제공할 수 있으며, 다른 전자 기기(일례로 제2 전자 기기(120)) 또는 다른 서버(일례로 서버(160))와 통신하기 위한 기능을 제공할 수 있다. 일례로, 제1 전자 기기(110)의 프로세서(212)가 메모리(211)와 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청(일례로 검색 요청)이 통신 모듈(213)의 제어에 따라 네트워크(170)를 통해 서버(150)로 전달될 수 있다. 역으로, 서버(150)의 프로세서(222)의 제어에 따라 제공되는 제어 신호나 명령, 콘텐츠, 파일 등이 통신 모듈(223)과 네트워크(170)를 거쳐 제1 전자 기기(110)의 통신 모듈(213)을 통해 제1 전자 기기(110)로 수신될 수 있다. 예를 들어 통신 모듈(213)을 통해 수신된 서버(150)의 제어 신호나 명령 등은 프로세서(212)나

메모리(211)로 전달될 수 있고, 콘텐츠나 파일 등은 제1 전자 기기(110)가 더 포함할 수 있는 저장 매체로 저장될 수 있다.

[44] 입출력 인터페이스(214)는 입출력 장치(215)와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 입력 장치는 키보드 또는 마우스 등의 장치를, 그리고 출력 장치는 어플리케이션의 통신 세션을 표시하기 위한 디스플레이와 같은 장치를 포함할 수 있다. 다른 예로 입출력 인터페이스(214)는 터치스크린과 같이 입력과 출력을 위한 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수도 있다. 보다 구체적인 예로, 제1 전자 기기(110)의 프로세서(212)는 메모리(211)에 로딩된 컴퓨터 프로그램의 명령을 처리함에 있어서 서버(150)나 제2 전자 기기(120)가 제공하는 데이터를 이용하여 구성되는 서비스 화면이나 콘텐츠가 입출력 인터페이스(214)를 통해 디스플레이에 표시될 수 있다. 입출력 인터페이스(224) 또한 마찬가지로 서버(150)의 프로세서(222)가 메모리(221)에 로딩된 컴퓨터 프로그램의 명령을 처리함에 있어 서버(150)가 제공하는 데이터를 이용하여 구성되는 정보를 출력할 수 있다.

[45] 또한, 다른 실시예들에서 제1 전자 기기(110) 및 서버(150)는 도 2의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 제1 전자 기기(110)는 상술한 입출력 장치(215) 중 적어도 일부를 포함하도록 구현되거나 또는 트랜시버(transceiver), GPS(Global Positioning System) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수도 있다. 보다 구체적인 예로, 제1 전자 기기(110)가 스마트폰인 경우, 일반적으로 스마트폰이 포함하고 있는 가속도 센서나 자이로 센서, 카메라, 각종 물리적인 버튼, 터치패널을 이용한 버튼, 입출력 포트, 진동을 위한 진동기 등의 다양한 구성요소들이 제1 전자 기기(110)에 더 포함되도록 구현될 수 있음을 알 수 있다.

[46] 이하에서는 주제별 질의의 서브토픽 자동 추출 및 구조화를 위한 토픽 구조화 방법 및 토픽 구조화 시스템의 구체적인 실시예를 설명하기로 한다.

[47] 한번의 질의에 대해 최대한 다양한 검색 결과를 제공하고 효율적인 정보 탐색을 가능하게 하기 위해서는 토픽의 구조화(그룹화 및 계층화)가 필요하다.

[48] 검색 시스템이 사용자가 입력한 질의의 검색 결과를 제공할 때 검색 조건에 매칭된 문서들 외에 사용자의 추가 탐색을 도와주는 여러 가지 기능을 함께 제공한다. 대표적인 것으로 연관 검색어, 연관 태그, 검색어 자동 완성 등이 존재한다. 이러한 기능들은 다음과 같은 점에서 유용하다고 볼 수 있다.

[49] 첫째, 사용자가 알고자 하는 정보와 관련된 질의명을 명확히 알지 못할 경우 질의를 추천해주는 기능을 한다.

[50] 둘째, 해당 질의와 관련된 세부적인 속성/소주제들을 사용자가 간접적으로 파악할 수 있게 한다.

[51] 그러나, "연관"이라는 관계는 다음과 같은 한계점을 가진다.

[52] 첫째, "연관"이라는 관계가 모호하여 해당 질의와의 구체적인 관계(예:

상/하위개념, 유의어 또는 형제개념)를 알 수 가 없다. 따라서, 제공되는 검색어나 태그의 수가 많아지면 구조적으로 정렬하기 어렵고, 이러한 점 때문에 사용자 이용 관점에서 사용자에게 제공하는 수를 줄일 수 밖에 없다.

[53] 둘째, 질의가 중의성을 가지는 경우 각 의미에 따라 연관된 검색어나 태그가 정리되지 않은 상태로 제공되어 추가 탐색에 그다지 도움이 되지 않는다.

[54] 한편, 쇼핑 의도가 있는 질의의 경우 사용자가 원하는 상품의 탐색을 도와주기 위해 제공되는 계층 정보가 존재한다. 제품의 브랜드, 색상, 가격 등 그 계층이 매우 체계적이라 빠르고 효율적인 탐색에 도움이 되긴 하지만, 이러한 정보는 각 판매 업체들이 수동으로 입력한 정보들이라 확장성에 한계를 가지며, 또한 쇼핑 질의에 대해서만 잘 적용되는 한계점이 있다.

[55] 본 발명에서는 사용자의 추가 탐색을 도와주면서 해당 질의와 관련된 세부적인 속성/소주제들을 사용자가 효율적으로 파악할 수 있도록 하는 기능은 가지되, 위에서 서술한 한계점을 해결할 수 있는 서브토픽 자동 추출 및 구조화 기술을 제안한다.

[56] 본 발명에 따른 토픽 구조화 시스템의 핵심 내용은 다음과 같다.

[57] (1) 토픽 구조화 시스템은 주제 별 주요 질의를 "메인 오브젝트+서브토픽"으로 패턴화 한다. 이때, 메인 오브젝트는 주제를 결정하는 핵심 오브젝트를 의미하고, 서브토픽은 서브 오브젝트와 속성 중 적어도 하나를 포함한다. 서브 오브젝트는 주제를 구체화 하는 오브젝트를 의미하고, 속성은 suffix 또는 prefix 등으로 주제의 속성을 나타내는 단어를 의미한다.

[58] (2) 토픽 구조화 시스템은 서브 오브젝트와 속성들의 의미 관계에 따라 패턴화된 질의들을 계층화 한다. 도 3은 본 발명의 일 실시예에 있어서 패턴화 된 질의를 계층화 하는 과정의 예시를 도시한 도면이다. 도 3에 도시한 바와 같이, 특정 주제의 질의, 즉 메인 오브젝트(MainObj)를 기준으로 메인 오브젝트(MainObj)와 서브토픽들(SubObj, Suffix)을 계층화 할 수 있다.

[59] (3) 토픽 구조화 시스템은 계층화 된 질의와 서브토픽을 검색 결과(문서)와 함께 사용자에게 제공할 수 있다. 이때, 토픽 구조화 시스템은 사용자가 원하는 정도(세분화)에 따라 서브토픽을 적절히 구조화 하여 제시함으로써 주제에 적합하고 해당 질의와 관련된 세부적인 서브토픽들을 사용자가 효율적으로 파악하고 실질적인 추가 탐색을 도와주는데 기여할 수 있다.

[60] 도 4는 본 발명의 일 실시예에 따른 서버의 프로세서가 포함할 수 있는 구성요소의 예를 도시한 도면이고, 도 5는 본 발명의 일 실시예에 따른 서버가 수행할 수 있는 방법의 예를 도시한 흐름도이다.

[61] 도 4에 도시된 바와 같이 서버(150)의 프로세서(222)는 구성요소들로서 추출부(410), 정제부(420), 생성부(430), 조정부(440), 및 제공부(450)를 포함할 수 있다. 이러한 프로세서(222) 및 프로세서(222)의 구성요소들은 도 5의 방법이 포함하는 단계들(S510 내지 S550)을 수행하도록 서버(150)를 제어할 수 있다. 이때, 프로세서(222) 및 프로세서(222)의 구성요소들은 메모리(221)가 포함하는

운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다. 또한, 프로세서(222)의 구성요소들은 운영체제나 적어도 하나의 프로그램이 제공하는 제어 명령에 따라 프로세서(222)에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 예를 들어, 프로세서(222)가 상술한 제어 명령에 따라 메인 오브젝트와 서브토픽을 추출하는 기능적 표현으로서 추출부(410)가 사용될 수 있다.

- [62] 프로세서(222)의 구성요소들을 먼저 살펴보면 다음과 같다.
- [63] 추출부(410)의 MainObj+Suffix 추출 모듈은 주제 별로 메인 오브젝트 및 속성을 추출한다. 메인 오브젝트와 속성을 추출하기 위해서는 자연어 이해(Natural Language Understanding) 기술이 사용될 수 있다.
- [64] 추출부(410)의 SubObj 추출 모듈은 각 주제에 대해 메인 오브젝트+(속성)이 있을 때 서브 오브젝트를 추출한다. 서브 오브젝트를 추출하기 위해서는 여러 가지 통계 정보(예컨대, 클릭 수, 좋아요 수, 댓글 수, 작성자 수 등)나 사전(dictionary) 정보 등을 이용할 수 있다.
- [65] 정제부(420)의 Ranker (Suffix Ranker, SubObj Ranker) 모듈은 메인 오브젝트에 대하여 서브 오브젝트 및/또는 속성을 중요도 순으로 랭킹을 결정한다. 이때, 중요도를 결정하기 위해서 사용자로부터 클릭된 수, 좋아요 수, 댓글 수, 작성자 수 등 여러 가지 정보가 활용될 수 있다.
- [66] 정제부(420)의 Post-Processor (Post-Processor Ranker, Post-Processor Ranker) 모듈은 랭킹이 결정된 서브 오브젝트 및/또는 속성을 받아 유의어나 서브스트링(부분문자열) 관계에 따라 이를 클러스터링하고 각 클러스터의 대표값을 선정한다.
- [67] 생성부(430)의 TopicGraphToTree 모듈은 클러스터링 된 서브 오브젝트 및/또는 속성을 모아 질의나 문서에서 얼마나 같이 자주 나타나는지 관계 강도를 찾은 후 이를 네트워크(그래프) 구조로 만든 다음 트리(검색/클러스터 기반 트리)로 변환하게 된다. 구체적인 검색/클러스터 기반 트리에 대한 설명은 아래에서 기술하기로 한다.
- [68] 생성부(430)의 TreeConstructor 모듈은 사전 기반 트리와 검색/클러스터 기반 트리 구조를 통합하여 최종 토픽 트리(예를 들면, 2-depth 트리 구조)를 구성한다.
- [69] 조정부(440)의 Topic reranker 모듈은 토픽 트리에 대해 필터링 조건(일례로, 문서 수, 주제 적합도, 정당성 여부 등)에 따라 추가 필터링 한다.
- [70] 조정부(440)의 New Object Assigner 모듈은 메인 오브젝트와 관련된 새로운 아이টে를 추출하여 원 트리 구조에 할당한다. 만일 트리 구조에 맞지 않을 정도로 새로운 아이템이 많아지는 경우 처음부터 프로세스를 다시 진행하여 트리를 새로 구성한다.
- [71] 제공부(450)의 Document Finding API 모듈은 최종 토픽 트리를 바탕으로 질의를 구성하여 적합 문서를 추출한다. 이때, 필터링 기능도 포함될 수 있다.
- [72] 제공부(450)의 Auto-Tagger 모듈은 최종 토픽 트리를 바탕으로 토픽 태그를

구성하여 적합 문서에 태깅한다.

- [73] 상기한 구성요소를 포함하는 프로세서(222)를 통해 도 5의 방법이 포함하는 단계들(S510 내지 S550)을 수행할 수 있다.
- [74] 도 5에서, 단계(S510)에서 추출부(410)는 각 주제 별로 해당 주제를 결정하는 핵심 오브젝트인 메인 오브젝트와 해당 주제를 구체화 하는 서브토픽을 추출할 수 있다. 이때, 추출부(410)는 문서 상에 메인 오브젝트와 함께 자주 등장하는 단어들을 분석하거나 검색 시스템에서 메인 오브젝트와 함께 검색에 자주 이용되는 단어들을 분석하여 서브 오브젝트 및/또는 속성 후보들을 추출할 수 있다.
- [75] 단계(S520)에서 정제부(420)는 서브토픽을 문서 내 출현 빈도나 검색 빈도에 따라 필터링 한 후 단어 간 관계에 의한 그룹핑을 수행할 수 있다. 후보 필터링 과정에서는 서브 오브젝트 및/또는 속성 후보들 중 적어도 일부를 필터링 하여 사용할 수 있다. 구체적인 일례로, 정제부(420)는 문서 내 출현 빈도 및 사용자의 검색 빈도 중 적어도 하나에 따라 서브 오브젝트 및/또는 속성을 필터링할 수 있다. 또한, 출현 빈도를 특정 기간의 데이터 한정하여 필터링 할 수 있다. 이때, 필터링 방법은 주제의 특성에 따라 달라질 수 있는데, 예를 들어 시의성이 높은 주제라면 최근 일정 기간(예를 들어, 현재로부터 1주일 전까지의 기간) 데이터를 필터링 할 수 있다. 그리고, 정제부(420)는 후보 필터링 과정을 통해 선정된 서브 오브젝트 및/또는 속성 후보들을 유의어, 서브스트링 관계 등을 고려하여 그룹핑할 수 있으며, 그룹핑 후 각 그룹마다 대표를 선정할 수 있다. 대표를 선정하는 방법은 다양할 수 있는데, 일 실시예로서는 검색 빈도가 제일 높은 것을 대표로 선정할 수 있다. 다시 말해, 정제부(420)는 서브토픽을 중요도(예컨대, 문서 내 출현 빈도, 검색 빈도 등) 순으로 랭킹한 후 랭킹된 서브토픽을 유의어나 서브스트링 관계에 따라 클러스터링 하고 각 클러스터의 대표를 선정할 수 있다. 상기한 서브토픽의 추출 및 정제 과정은 후보 선정, 그룹핑/대표선정 과정으로 이루어진다. 도 6은 본 발명의 일 실시예에 따라 '팜' 및 '감자'이라는 질의에 대하여 후보를 필터링, 그룹핑 및 대표 토픽을 선정하는 과정의 일 예를 나타낸 것이다. 도 6에서 WTRIP 및 FOOD는 질의의 주제를 나타내는 분류 코드(카테고리 분류 코드)이고, 단어 옆의 숫자는 질의와 함께 연관되어 검색된 빈도수를 나타내는 것이다.
- [76] 다시 도 5에서, 단계(S530)에서 생성부(430)는 그룹핑 된 서브토픽에 대해 해당 주제의 계층 정보를 이용하여 토픽 트리를 생성할 수 있다. 생성부(430)는 각 그룹핑된 서브토픽을 계층 정보를 활용하여 계층의 각 클래스 이름에 맞도록 레이블링(labeling) 함으로써 토픽 트리를 생성할 수 있다. 주제에 따라 계층 정보가 아마 존재하는 경우와 계층 정보가 존재하지 않는 경우가 있다. 예를 들어, 다양한 종류의 콘텐츠를 포함하고 있는 데이터베이스로부터 구축된 사전 정보는 계층 정보에 유용하게 활용할 수 있는 정보의 하나이다. 이때, 생성부(430)는 사전 정보와 같이 기 존재하는 계층 정보가 있는 경우 이를

기반으로 토픽 트리를 생성할 수 있다. 예를 들어, 요리나 레시피 주제는 요리백과사전을 기반으로 한 풍부한 계층 정보가 존재한다. 반면에, 여행이나 쇼핑 주제는 계층 정보가 존재하지 않고, 시기에 따라서 다양한 서브토픽이 생성되거나 변경될 수 있다. 생성부(430)는 계층 정보가 존재하지 않는 경우 워드임베딩(word embedding) 기반의 클러스터링 기법과 언어 분류학(taxonomy)을 기초로 계층 정보를 구축하여 토픽 트리를 생성하는데 활용할 수 있다. 본 발명은 계층 정보가 존재하지 않는 경우에도 자동적으로 토픽을 계층화 할 수 있다는 점에 장점이 있다.

- [77] 도 7은 본 발명의 일 실시예에 있어서 워드임베딩 기반의 클러스터링 기법과 언어 분류학을 이용하여 계층 정보를 구축하는 과정의 일례를 설명하기 위한 순서도이다. 도 7을 참조하면, 생성부(430)는 서브토픽에 대해 워드임베딩 데이터에서 유사 단어를 추출하고(S701), 추출된 단어를 유의어나 서브스트링 관계에 따라 클러스터링 한 후(S702), 클러스터링 된 단어를 언어 분류학에 기초하여 레이블링 할 수 있다(S703). 도 8은 여행 주제에 대해서 클러스터링 및 언어 분류학을 이용하여 구축한 계층 정보의 일 예를 보여준다. 워드임베딩 기반의 클러스터링 과정(S702)에서는 주제별 문서(예컨대, 블로그 게시물 등)로 워드임베딩 데이터를 학습하고 이러한 학습 데이터로부터 클러스터링이 필요한 서브토픽의 워드벡터 값을 파악한 후 워드벡터 값을 기반으로 클러스터링을 한다. 이때, 클러스터링은 계층적 클러스터링(hierarchical clustering), K-평균 알고리즘(K-means algorithm), 밀도 클러스터링(density clustering) 등 여러 방법들을 활용 가능하다. 그리고, 언어 분류학 활용 과정(S703)에서는 클러스터링 된 결과를 언어 분류학 상의 각 클래스에 매핑함으로써 레이블링이 가능하다. 이때, 언어 분류학은 범용적이라 주제에 특화된 계층 정보와 비교할 때 불필요한 클래스들이 많이 존재한다. 따라서, 불필요한 클래스를 삭제하는 작업이 필요하며, 이는 이하 조정부(440)의 리밸런싱(rebalancing) 과정에서 설명하기로 한다.
- [78] 토픽 계층화 단계에서는 클러스터링 된 서브토픽을 모아서 질의나 문서에서 얼마나 같이 자주 나타나는지 관계 강도를 찾고 네트워크(그래프) 구조로 만든 후 토픽 트리로 변환하고(클러스터 기반 트리), 사전 기반으로 구축된 토픽 트리과 클러스터링 기반으로 구축된 토픽 트리를 통합하여 최종 트리 구조를 구성할 수 있다.
- [79] 다시 도 5에서, 단계(S540)에서 조정부(440)는 생성부(430)의 토픽 계층화 단계에서 구축된 토픽 트리에 대하여 사용자나 시스템의 목적에 따라 리밸런싱을 수행할 수 있다. 또한, 조정부(440)는 토픽 트리에 대해 주제 적합도, 검색 의도, 검색 결과의 양 등을 고려하여 적당히 가지치기(Pruning)를 수행할 수 있다.
- [80] 표 1은 본 발명의 일 실시예에 따른 토픽 네트워크의 정의를 나타낸 것이다.

[81] [표1]

- $G = (V, E)$
- $V =$

$$\{mainObj, Suffix\} - \{suffix | suffix \in mainObjSet\} + \{suffix | suffix \in seedSuffixSet\}$$
 - $mainObjSet = \{ \text{주제별 mainobj} \}$
 - $seedSuffixSet = \{ \text{주제별 주요 suffix} \}$
- $E = \{e(V_i, V_j)\}$
- $$e(V_i, V_j) = \frac{\sum_{SynoV_i} \sum_{SynoV_j} with(V_i, V_j) + with(mainObj, V_i, V_j) \times \alpha}{n(SynoV_i) \times n(SynoV_j)} \quad (1)$$
 - $SynoV_i = \{ \text{suffix 동의어/유의어 집합} \}$
 - $e(V_i, V_j) = with(\max(len(V_i), len(V_j))) \times \alpha \quad (2)$
 - V_i 와 V_j 가 서브스트링 관계

[82] 생성부(430)는 검색 빈도와 클러스터링을 이용하여 구축한 정보들을 이용하여 토픽 트리를 생성한다. 우선, 각 단어들을 노드(node)로, 간 단어들간의 연관 관계를 간선(edge)으로 나타내어 토픽 네트워크(G)를 구축한다. 본 발명의 일 실시예에 따른 토픽 네트워크(G)에서의 노드(V) 및 간선(E)는 표 1과 같이 정의될 수 있다. 이때, 생성부(430)는 토픽 네트워크를 검색 빈도를 고려하여 토픽 트리로 변경한다. 도 9는 본 발명의 일 실시예에 따라 구축된 토픽 네트워크를 깊이가 2인 트리(2-depth 트리)로 변환하는 과정의 예시를 도시한 것이다. 네트워크를 트리로 변경하는 방법은 다양한 알고리즘이 존재할 수 있으며, 예를 들어 가중치 그래프에서의 최소 신장 트리(minimum spanning tree) 구축 알고리즘 등을 응용할 수 있다. 이후, 조정부(440)는 언어 분류학을 기반으로 한 트리와 검색빈도/클러스터링을 기반으로 한 트리를 합쳐서 사용자나 시스템의 목적에 따라 리밸런싱을 수행할 수 있다. 도 10은 본 발명의 일 실시예에 있어서 트리 리밸런싱 과정의 일례를 설명하기 위한 순서도이다. 도 10을 참조하면, 조정부(440)는 클러스터링 기반 트리에서 리프 노드에 해당되는 클러스터를 사전 기반 트리의 해당 클래스에 삽입할 수 있다(S1001). 토픽 트리의 너비(breadth)와 깊이(depth)는 질의 별로 각각 다르며, 사용자가 활용하기에는 대체로 깊이 및 너비 모두 값이 큰 편이라 이를 줄이는 작업이 필요하다(S1002~S1003). 너비와 깊이를 줄이는 방법은 언어 분류학을 기반으로 한 트리와 검색빈도/클러스터링을 기반으로 한 트리를 결합하는 과정에서 이루어진다. 토픽 트리의 너비 값과 깊이 값은 시스템의 필요 사항에 따라서

다르게 설정할 수 있으며, 본 발명의 실시예에서는 깊이가 2인 것으로 가정한다(2 depth tree화). 그리고, 조정부(440)는 주제 적합도, 검색 의도, 검색 결과의 양 등을 고려하여 토픽 트리의 가지치기를 수행할 수 있다(S1004). 도 11은 트리 리밸런싱 과정 중 너비를 줄이기 위한 몇 가지 방법을 나타내는 것으로, 상향식(bottom-up) 노드 이동 및/또는 하향식(top-down) 노드 이동을 통해 토픽 트리의 너비를 줄일 수 있다. 또한, 도 12는 트리 리밸런싱 과정 중 깊이를 줄이기 위한 몇 가지 방법을 나타내는 것으로, 일부 노드를 자식 노드와 교체하는 방식을 통해 토픽 트리의 깊이를 줄일 수 있다.

- [83] 다시 도 5에서, 단계(S550)에서 제공부(450)는 질의가 속하는 주제의 토픽 트리를 이용하여 서브토픽을 해당 질의에 대응되는 검색 결과와 함께 제공할 수 있다. 이때, 제공부(450)는 질의에 대한 연관 검색어로서 서브토픽을 여러 조건에 따라 필터링 한 후 검색 결과와 함께 제공할 수 있다. 일례로, 제공부(450)는 질의의 주제 적합도에 따라 서브토픽을 필터링 할 수 있다. 검색을 위한 질의가 주어지면 해당 질의가 속하는 주제를 확인할 수 있고, 이때 질의가 여러 주제에 속하는 경우 주어진 주제에 적합하지 않은 토픽을 필터링 할 수 있다. 이를 위해, 질의와 서브토픽의 연관성을 나타내는 점수인 '질의+서브토픽'의 주제 점수를 활용할 수 있다. 주제 점수 파악 방법으로는 텍스트 분류(Text categorization) 알고리즘(예컨대, SVM(support vector machine), kNN(k-Nearest Neighbor), CNN(Convolutional Neural Networks) 등)을 이용할 수 있다. 다른 예로, 제공부(450)는 서브토픽에 대응되는 문서 수를 이용하여 서브토픽을 필터링 할 수 있다. 검색 결과에 포함된 문서 수가 일정 개수 이하로 적다면 유용성이 떨어지므로 해당 서브토픽을 제외시킬 수 있다. 또 다른 예로, 제공부(450)는 질의에 대한 정답성 토픽 여부에 따라 서브토픽을 필터링 할 수 있다. 검색 결과로 여러 문서를 제공하는 것보다 정답성 정보를 제공하는 것이 더 적절한 서브토픽의 경우(예를 들어, 팜 날씨와 같이 정답성 정보가 요구되는 경우) 질의에 대한 연관 검색어로 포함시킬 수 있다.

- [84] 제공부(450)는 사용자가 검색을 위해 입력한 질의에 대해 연관 검색어로서 해당 질의와 관련된 세부적인 서브토픽(서브 오브젝트 및/또는 속성)을 계층형으로 노출할 수 있다. 주제 별 토픽 트리는 일정 주기를 단위로 업데이트 될 수 있으며, 주제에 따라 해당 주제의 특성 등을 고려하여 토픽 트리의 업데이트 주기가 결정될 수 있다.

- [85] 도 13 내지 도 14는 본 발명의 일 실시예에 있어서 2 레벨의 토픽 구조가 반영된 검색 결과 화면의 예시를 도시한 도면이다. 예를 들어, 사용자가 입력한 질의에 대해 입력 질의에 대응되는 검색 결과를 제공하는 과정에서 입력 질의에 해당되는 주제의 계층 정보에 따라 입력 질의의 연관 검색어로서 depth1의 쿼리들과 depth2의 쿼리들을 제공할 수 있다. 도 13을 참조하면, 사용자가 검색창(1301)에 질의 '팜'을 입력한 경우 입력 질의와 대응되는 검색 결과와 함께, 해당 주제 '팜'의 계층 정보에 따라 입력 질의 '팜'의 연관 검색어로서 depth1의

쿼리들(1310)과 depth2의 쿼리들(1320)을 제공할 수 있다. 이때, 연관 검색어로 제공되는 쿼리 각각은 사용자에게 의해 선택 가능한 형태로 구성되며, 사용자가 선택한 쿼리가 검색창(1301)에 자동으로 추가 입력된다. 도 13에 도시한 바와 같이, 사용자가 초기 질의 '팜'의 연관 검색어로 제공되는 depth1의 쿼리들(1310) 중에서 '맛집'을 선택하는 경우 검색창(1301)에 '맛집'이 추가로 입력되고 '팜 맛집'을 질의로 한 depth1의 검색 결과(1302)가 노출될 수 있다. 다음으로, 사용자가 도 14에 도시한 바와 같이 depth2의 쿼리들(1320) 중에서 쿼리 '수제버거'를 선택한 경우 검색창(1301)에 '수제버거'가 추가로 입력되고 '팜 맛집 수제버거'를 질의로 한 depth2의 검색 결과(1402)가 노출될 수 있다.

[86] 이처럼 본 발명의 실시예들에 따르면, 본 발명에서는 검색결과(문서)를 계층 구조의 토픽들과 함께 제공하여 효율적인 추가탐색을 하게 도와주고 한번의 질의로 최대한 다양한 검색결과를 제공할 수 있다. 더 나아가, 계층화 된 토픽 구조를 검색 랭킹에 활용하는 것 또한 가능하다. 다시 말해, 서브오브젝트 및 속성이 포함된 문서는 상대적으로 고품질의 문서일 가능성이 있으므로 검색 랭킹 시 이러한 문서를 부스팅 하는데 활용할 수 있다.

[87] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[88] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 컴퓨터

저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

- [89] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 이때, 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수 개의 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 어플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

발명의 실시를 위한 형태

- [90] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.
- [91] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

청구범위

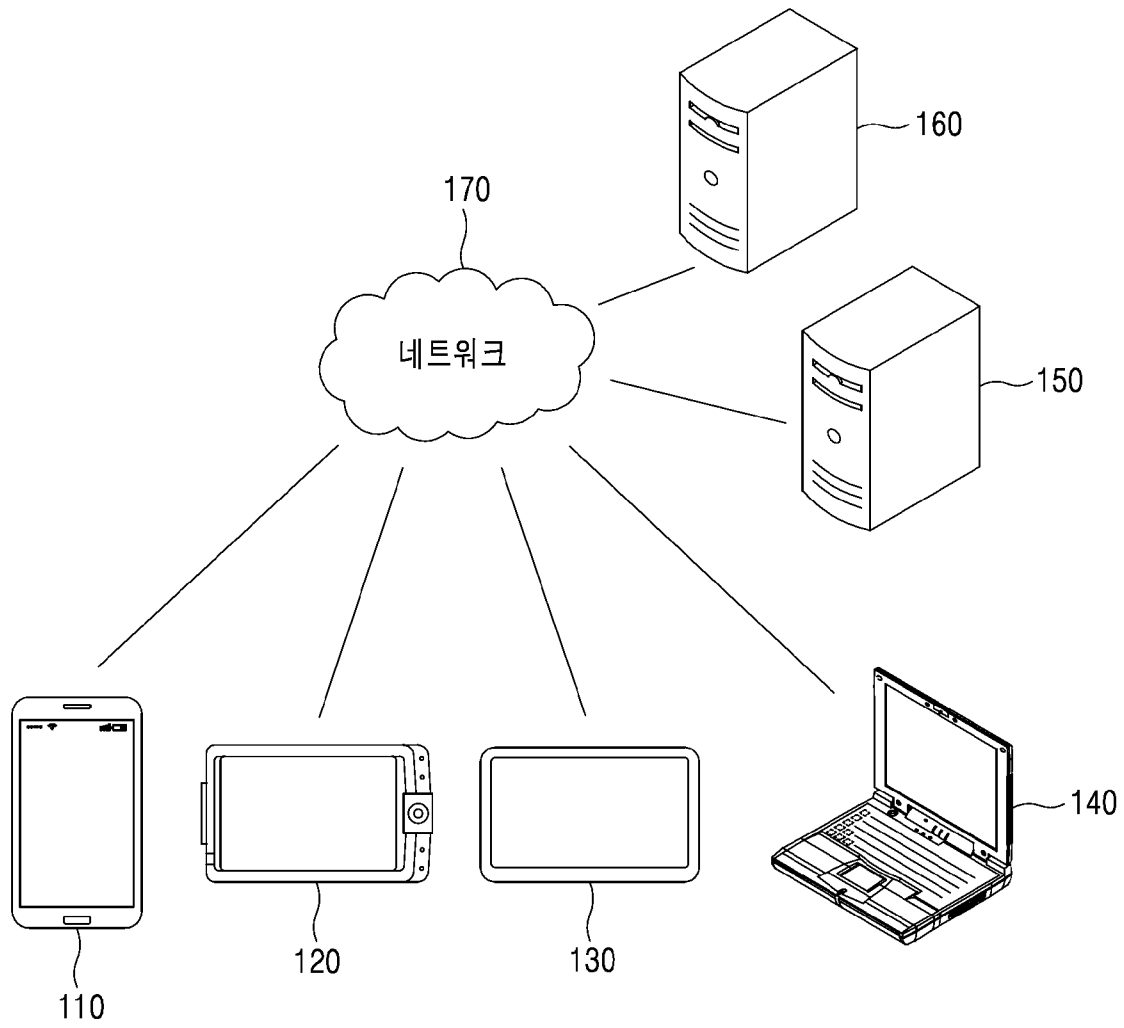
- [청구항 1] 컴퓨터로 구현되는 토픽 구조화 방법에 있어서,
주제 별로 상기 주제와 관련된 서브토픽을 추출하는 단계;
상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를
생성하는 단계; 및
검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라
상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로
제공하는 단계
를 포함하는 토픽 구조화 방법.
- [청구항 2] 제1항에 있어서,
상기 추출하는 단계는,
상기 주제를 결정하는 핵심 오브젝트와 관련된 단어들을 분석하여 상기
서브토픽을 추출하는 것
을 특징으로 하는 토픽 구조화 방법.
- [청구항 3] 제1항에 있어서,
상기 서브토픽을 문서 출현 빈도와 검색 빈도 중 적어도 하나에 따라
필터링 하는 단계
를 더 포함하는 토픽 구조화 방법.
- [청구항 4] 제1항에 있어서,
상기 서브토픽을 유의어 또는 서브스트링(부분문자열) 관계에 따라
클러스터링 하여 각 클러스터의 대표를 선정하는 단계
를 더 포함하는 토픽 구조화 방법.
- [청구항 5] 제1항에 있어서,
상기 생성하는 단계는,
상기 서브토픽을 상기 계층 정보의 각 클래스 이름으로 레이블링(labeling)
하여 상기 토픽 트리를 생성하는 것
을 특징으로 하는 토픽 구조화 방법.
- [청구항 6] 제1항에 있어서,
상기 생성하는 단계는,
상기 서브토픽에 대해 워드임베딩 데이터에서 유사 단어를 추출하는
단계;
상기 유사 단어를 유의어 또는 서브스트링(부분문자열) 관계에 따라
클러스터링 하는 단계; 및
상기 클러스터링 된 단어를 언어 분류학 상의 각 클래스에 매핑함으로써
레이블링 하는 단계
를 포함하는 토픽 구조화 방법.
- [청구항 7] 제1항에 있어서,

상기 토픽 트리의 너비(breadth)와 깊이(depth) 중 적어도 하나를 줄여 상기 토픽 트리에 대한 리밸런싱(rebalancing)을 수행하는 단계를 더 포함하는 토픽 구조화 방법.

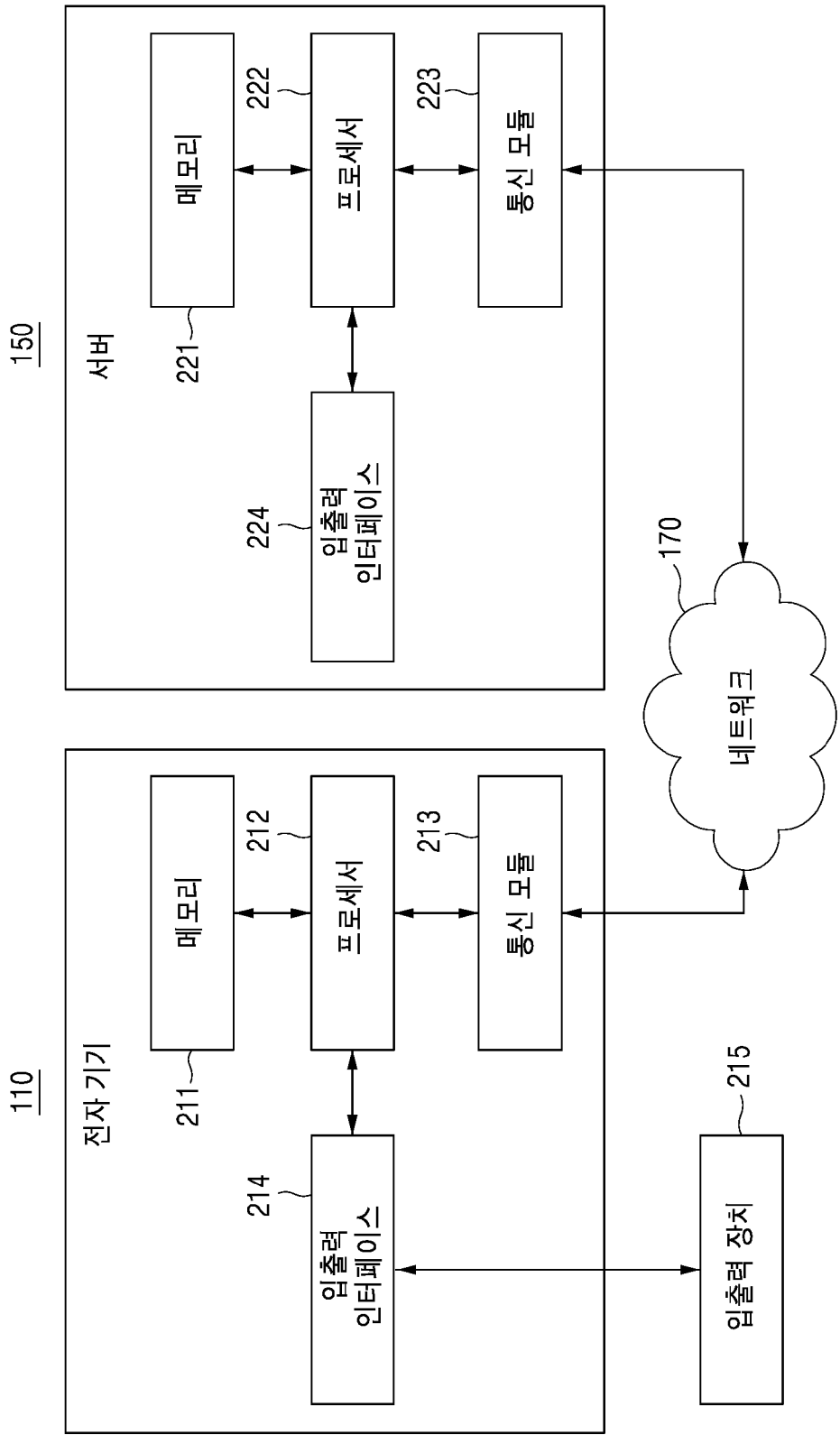
- [청구항 8] 제1항에 있어서,
상기 제공하는 단계는,
상기 질의와 상기 서브토픽의 연관성을 나타내는 주제 점수와, 상기 서브토픽에 대응되는 문서 수, 및 상기 질의에 대한 정답성 토픽 여부 중 적어도 하나의 조건에 따라 상기 서브토픽을 필터링 하는 단계를 포함하는 토픽 구조화 방법.
- [청구항 9] 컴퓨터로 구현되는 검색 결과 제공 방법에 있어서,
검색을 위한 질의가 주어지면 상기 질의에 대응되는 검색 결과를 제공하는 단계;
상기 질의가 속한 주제의 계층 정보에 따라 상기 질의에 대한 연관 검색어로서 상기 주제와 관련된 서브토픽을 복수의 템스(depth)를 가진 계층형으로 제공하는 단계; 및
상기 서브토픽 중 적어도 하나의 검색어가 선택되는 경우 상기 선택된 검색어를 포함한 상기 질의에 대응되는 검색 결과를 제공하는 단계를 포함하는 검색 결과 제공 방법.
- [청구항 10] 컴퓨터 시스템과 결합되어 토픽 구조화 방법을 실행시키기 위해 컴퓨터 판독 가능한 기록 매체에 기록된 컴퓨터 프로그램에 있어서,
상기 토픽 구조화 방법은,
주제 별로 상기 주제와 관련된 서브토픽을 추출하는 단계;
상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 단계; 및
검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로 제공하는 단계
를 포함하는, 컴퓨터 프로그램.
- [청구항 11] 컴퓨터로 구현되는 토픽 구조화 시스템에 있어서,
컴퓨터가 판독 가능한 명령을 실행하도록 구현되는 적어도 하나의 프로세서
를 포함하고,
상기 적어도 하나의 프로세서는,
주제 별로 상기 주제와 관련된 서브토픽을 추출하는 추출부;
상기 서브토픽에 대해 상기 주제의 계층 정보를 이용하여 토픽 트리를 생성하는 생성부; 및
검색을 위한 질의가 주어지면 상기 질의가 속한 주제의 토픽 트리에 따라 상기 질의에 대한 연관 검색어로서 상기 서브토픽을 계층형으로

제공하는 제공부
를 포함하는 토픽 구조화 시스템.

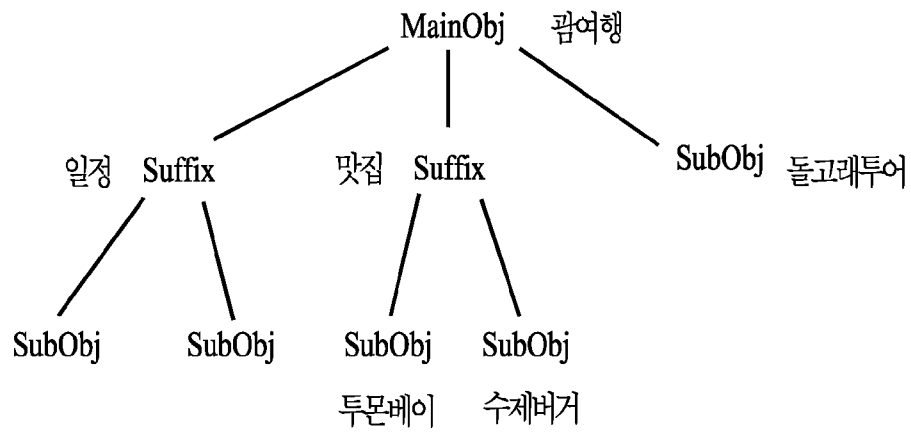
[도1]



[도2]



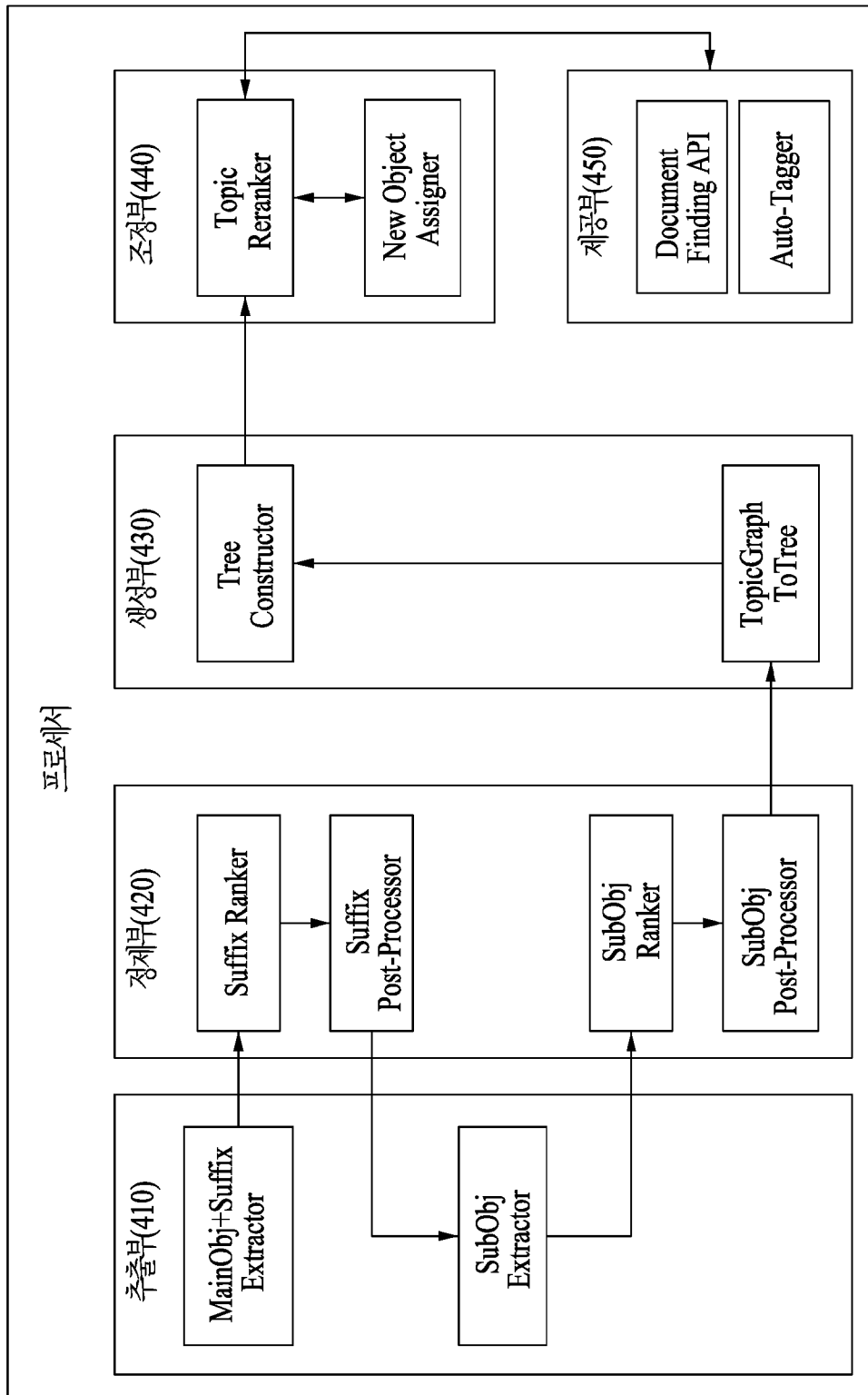
[도3]



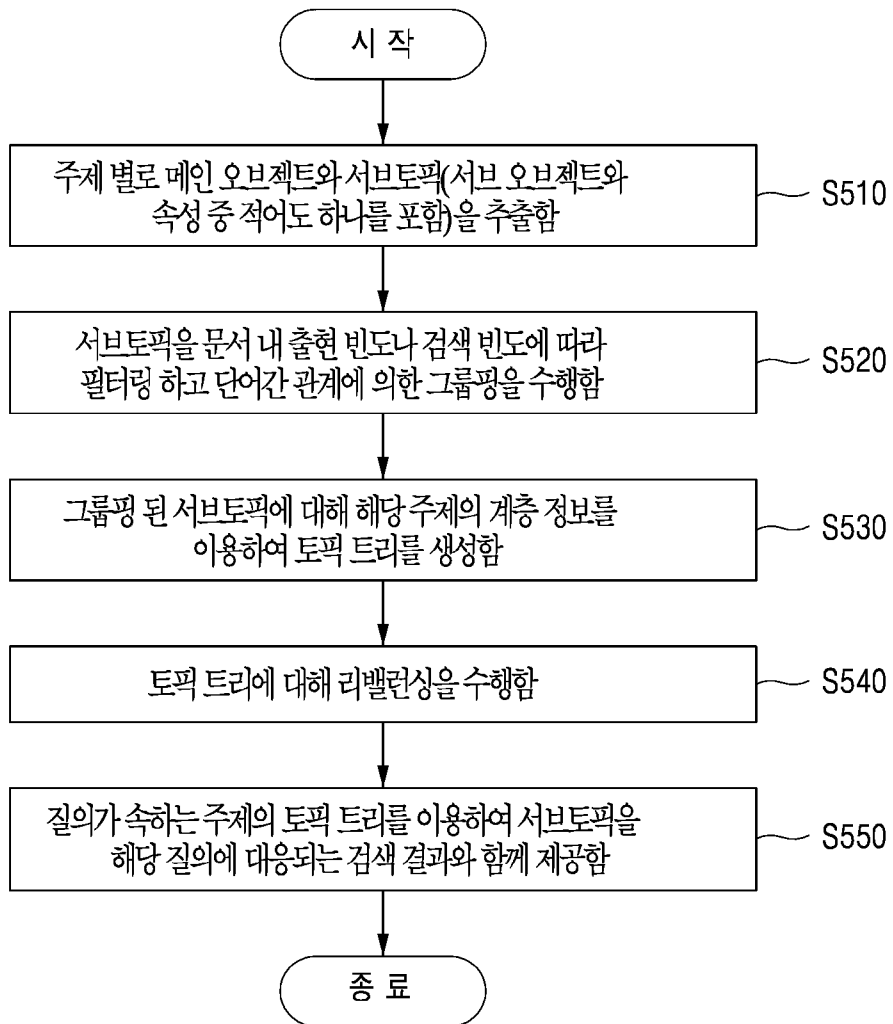
[도4]

222

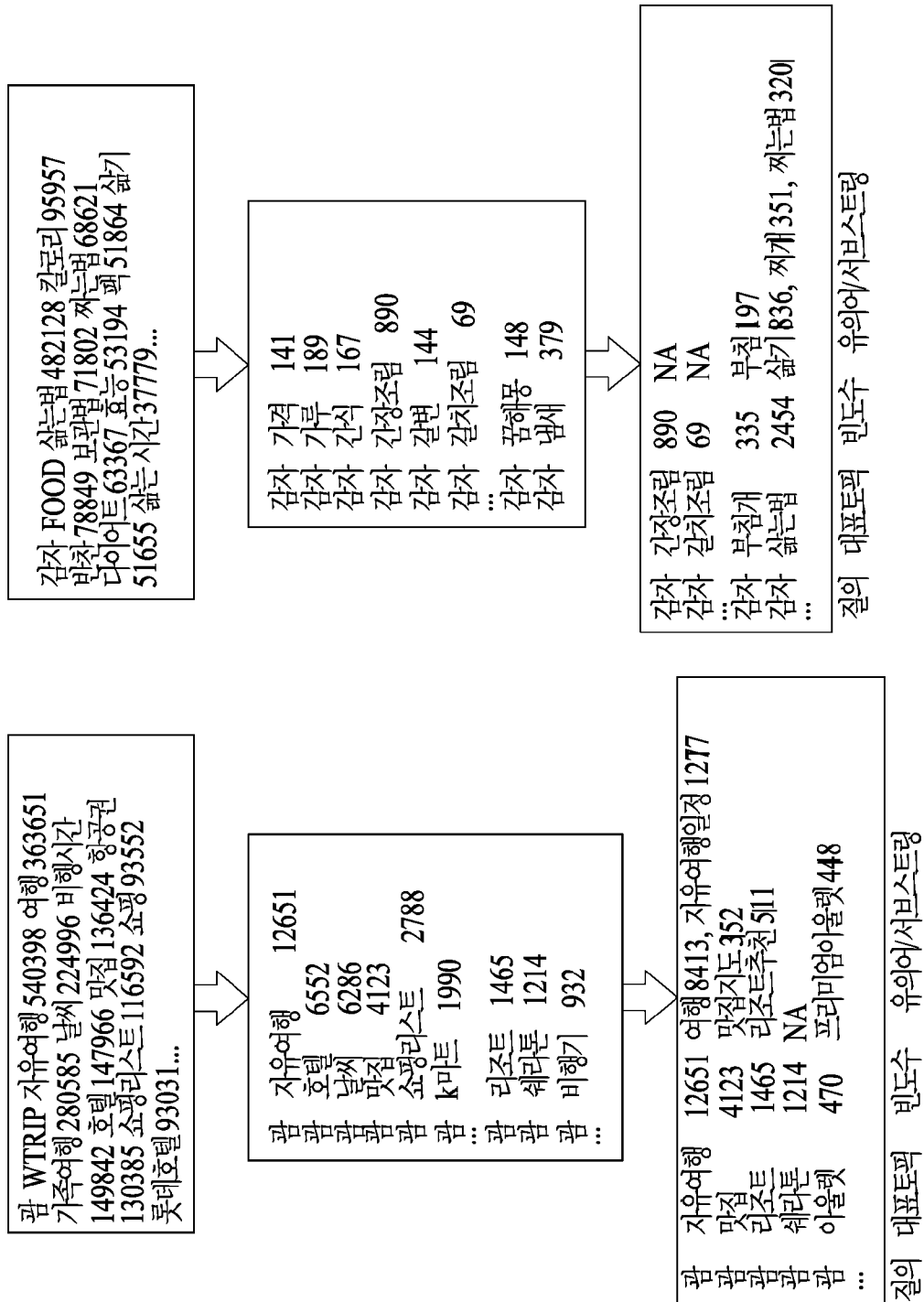
프로세서



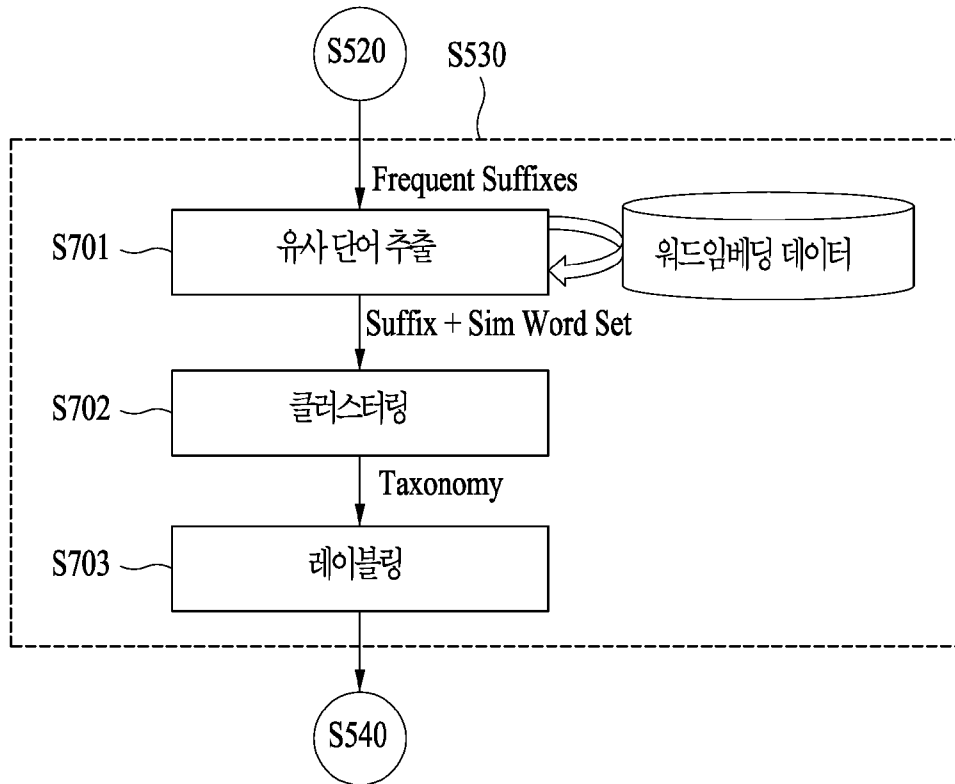
[도5]



[도6]



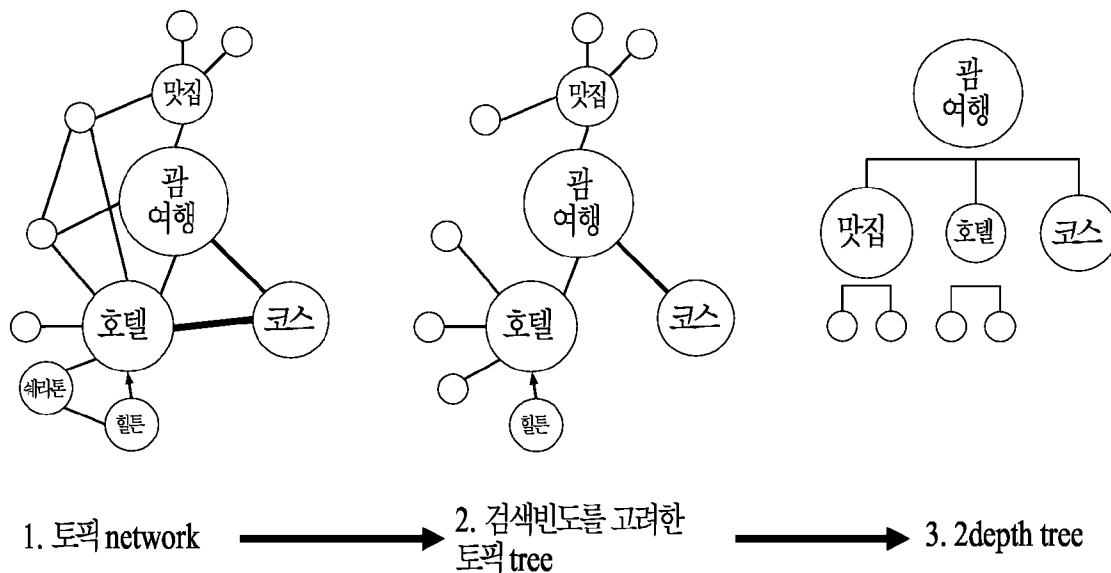
[도7]



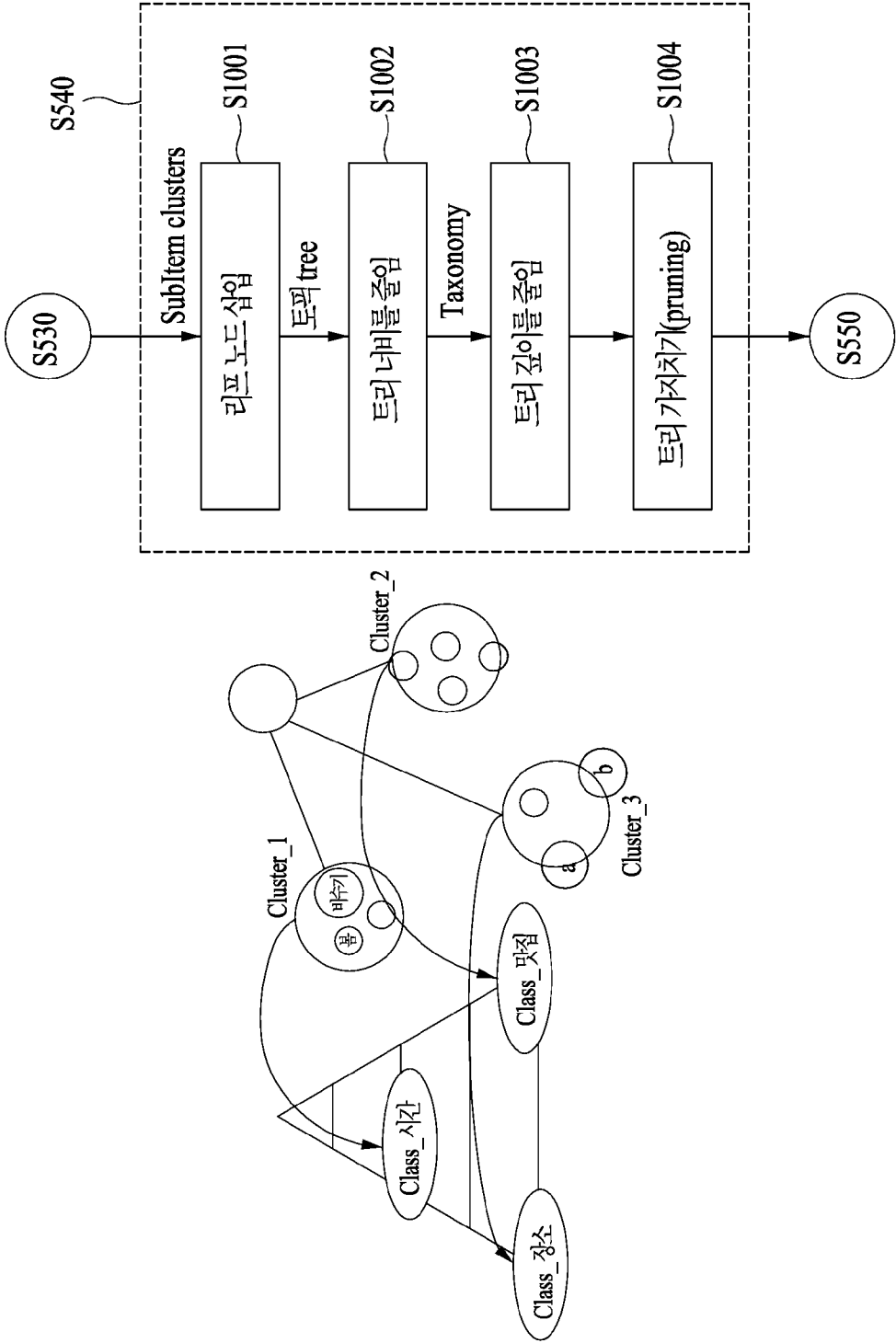
[도8]

cluster_31	cluster_7	계절	돈	육상교통기관	cluster_17	하루중단계
이올렛	쉐라톤	겨울	엔화	택시	모마	아침
버버리	인터컨티넨탈	바수기	환율	자동차	브로드웨이	밤
샤넬	힐튼	여름	환전	렌트		저녁
구찌	메리어트	성수기	폐소	차		
	콘래드		달러	자전거		
	노보텔		환전소	오토바이		
	웨스턴		화폐	스쿠터		
	히얏트		은행			

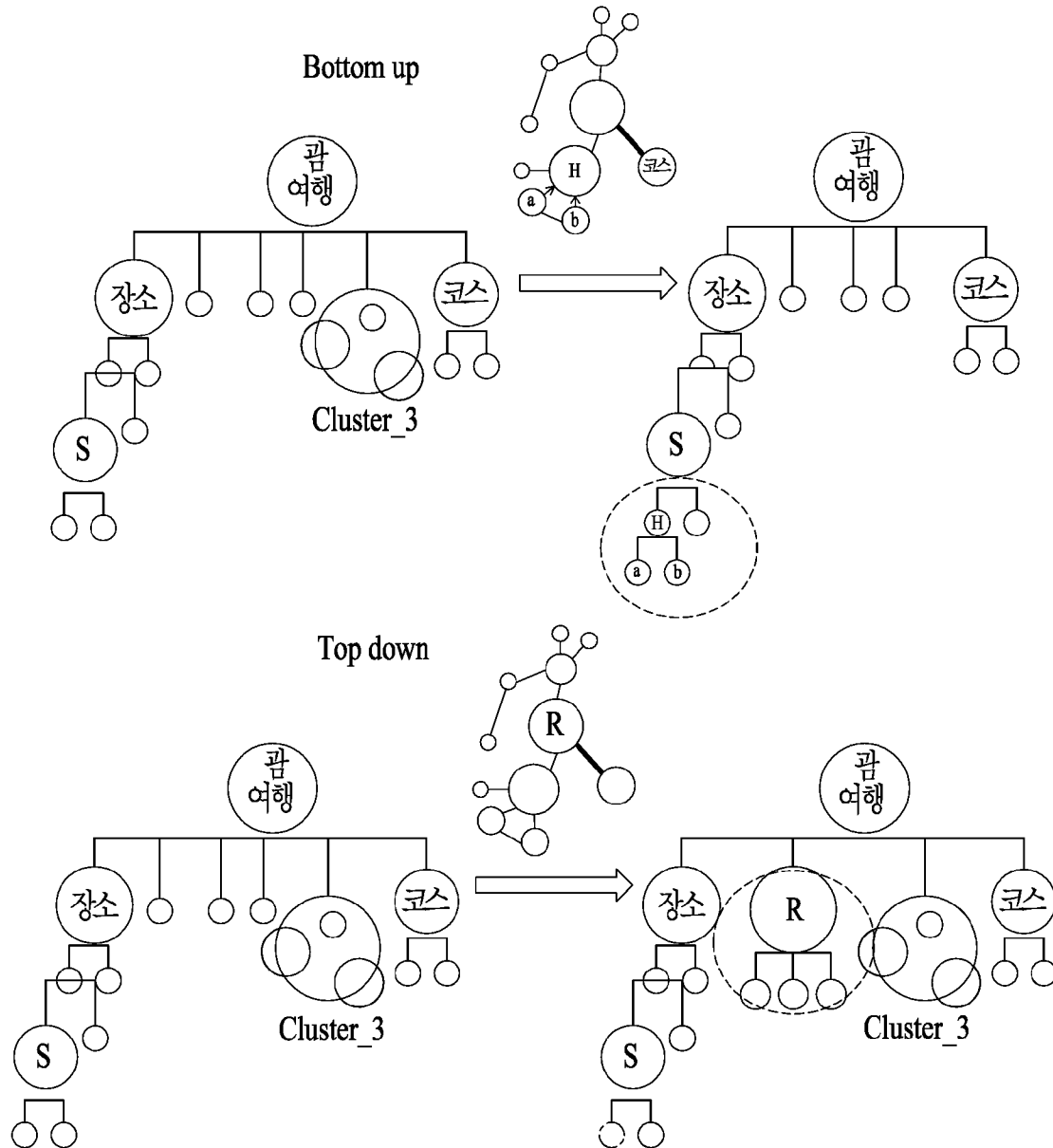
[도9]



[도10]

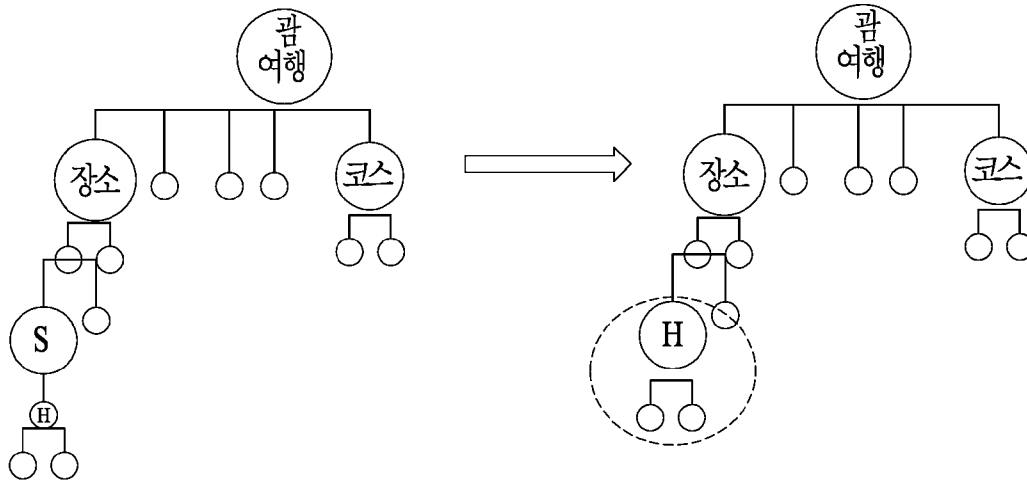


[도11]

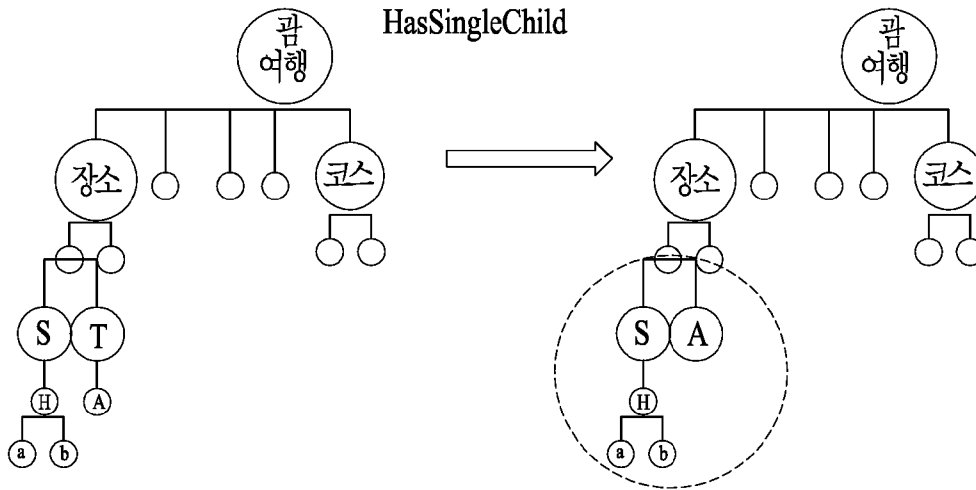


[도12]

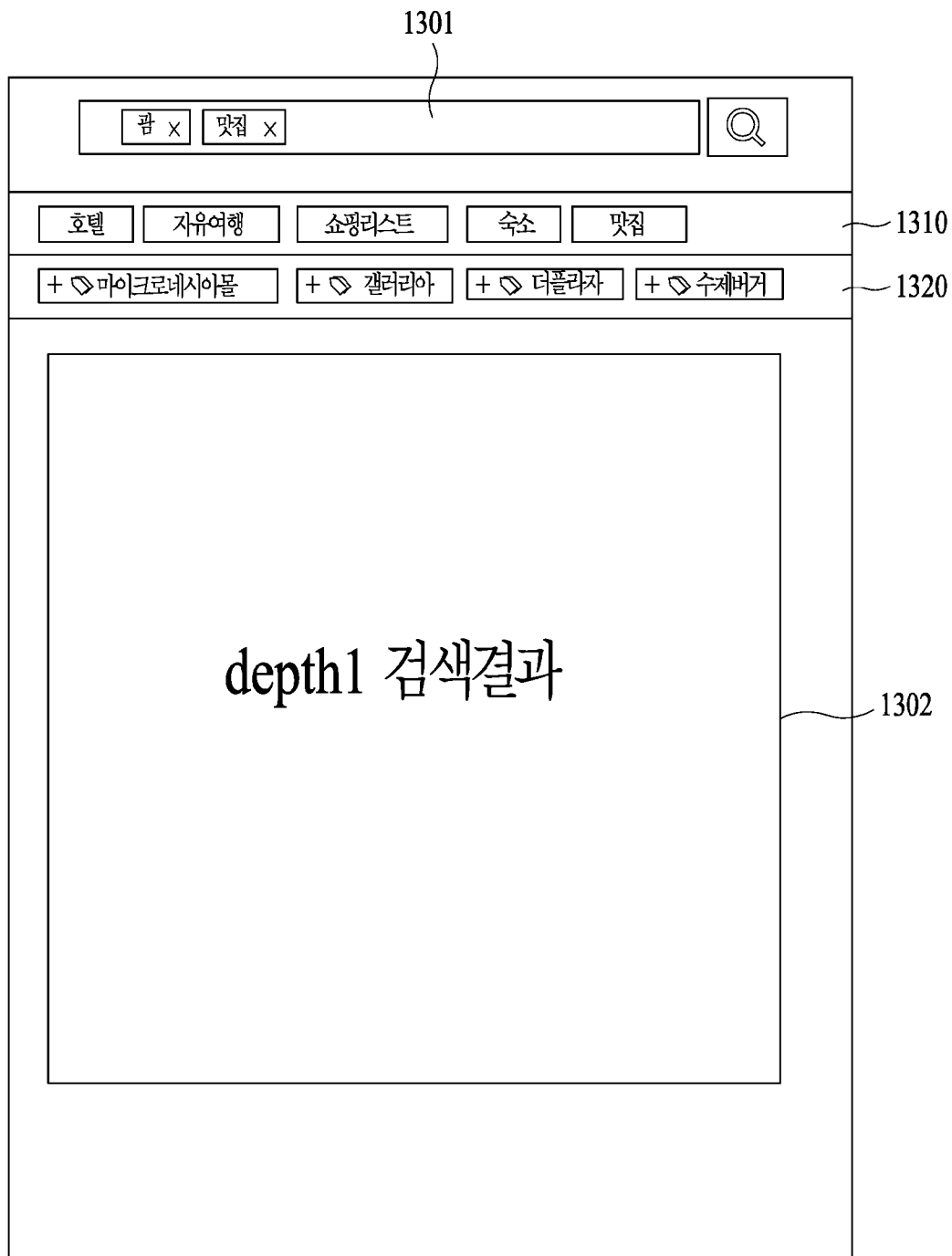
IsSingleChild



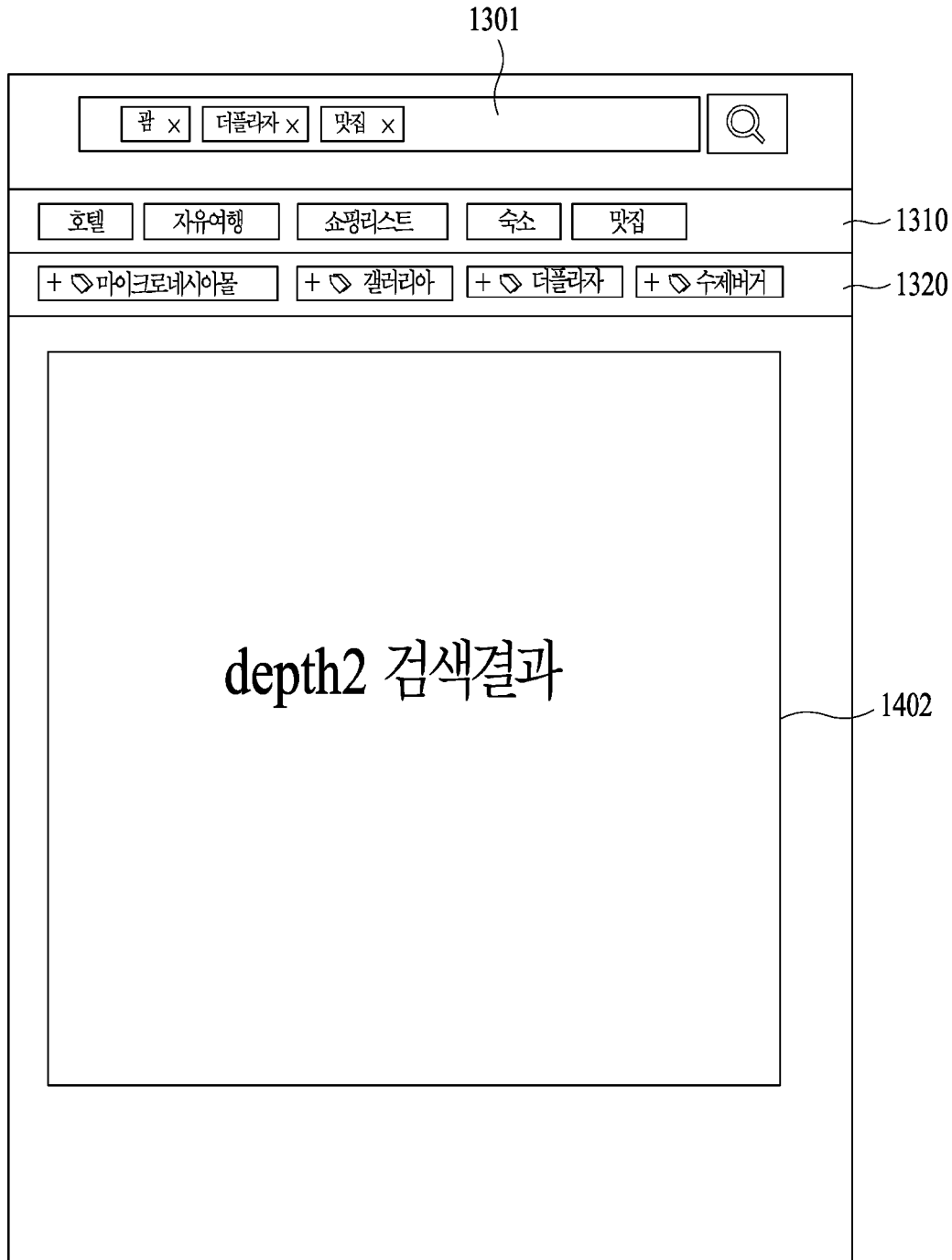
HasSingleChild



[도13]



[도 14]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2018/002834

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 17/30; G06F 3/048; G06Q 30/02; G06Q 20/10; H04B 1/40; G06F 17/40

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean Utility models and applications for Utility models: IPC as above

Japanese Utility models and applications for Utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: related search word, topic tree, hierarchical, subtopic, subject

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KR 10-2011-0099574 A (INDUSTRY-ACADEMIC COOPERATION FOUNDATION, YONSEI UNIVERSITY) 08 September 2011 See paragraphs [0036]-[0040], [0058]-[0061]; and figure 9.	1-11
Y	KR 10-2010-0080099 A (SALT LUX INC.) 08 July 2010 See paragraphs [0041], [0046], [0053]-[0054], [0085]; claims 1, 5, 7; and figures 4, 6.	1-11
Y	KR 10-1485940 B1 (NAVER CORPORATION) 27 January 2015 See paragraphs [0035], [0104]; and figure 8.	4,7
A	KR 10-0818357 B1 (WIPS CO., LTD.) 02 April 2008 See paragraphs [0044]-[0054]; and figures 1-3.	1-11
A	US 2017-0061485 A1 (EXCALIBUR IP, LLC.) 02 March 2017 See paragraphs [0052]-[0053]; and figures 8-10.	1-11



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

14 JUNE 2018 (14.06.2018)

Date of mailing of the international search report

14 JUNE 2018 (14.06.2018)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex-Daejeon, 189 Sconsa-ro, Daejeon 302-701,
Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2018/002834

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-2011-0099574 A	08/09/2011	KR 10-1101641 B1	02/01/2012
KR 10-2010-0080099 A	08/07/2010	KR 10-1059557 B1	26/08/2011
KR 10-1485940 B1	27/01/2015	JP 2015-041386 A	02/03/2015
		JP 5830592 B2	09/12/2015
		US 2015-0058309 A1	26/02/2015
KR 10-0818357 B1	02/04/2008	KR 10-2007-0109207 A	15/11/2007
US 2017-0061485 A1	02/03/2017	CN 103430176 A	04/12/2013
		CN 103430176 B	09/02/2018
		EP 2689354 A2	29/01/2014
		EP 2689354 A4	06/05/2015
		JP 2014-515846 A	03/07/2014
		JP 6231472 B2	15/11/2017
		TW 201243632 A	01/11/2012
		TW 201512867 A	01/04/2015
		TW 201631505 A	01/09/2016
		TW 201723896 A	01/07/2017
		TW 1479347 B	01/04/2015
		TW 1543009 B	21/07/2016
		TW 1585598 B	01/06/2017
		TW 1604325 B	01/11/2017
		US 2012-0246165 A1	27/09/2012
		US 9495461 B2	15/11/2016
		US 9892426 B2	13/02/2018
		WO 2012-128888 A2	27/09/2012
		WO 2012-128888 A3	14/03/2013

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06F 17/30(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06F 17/30; G06F 3/048; G06Q 30/02; G06Q 20/10; H04B 1/40; G06F 17/40

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 연관 검색어, 토픽 트리, 계층형, 서브토픽, 주제

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	KR 10-2011-0099574 A (연세대학교 산학협력단) 2011.09.08 단락 [0036]-[0040], [0058]-[0061]; 및 도면 9 참조.	1-11
Y	KR 10-2010-0080099 A (주식회사 솔트룩스) 2010.07.08 단락 [0041], [0046], [0053]-[0054], [0085]; 청구항 1, 5, 7; 및 도면 4, 6 참조.	1-11
Y	KR 10-1485940 B1 (네이버 주식회사) 2015.01.27 단락 [0035], [0104]; 및 도면 8 참조.	4,7
A	KR 10-0818357 B1 ((주)웹스) 2008.04.02 단락 [0044]-[0054]; 및 도면 1-3 참조.	1-11
A	US 2017-0061485 A1 (EXCALIBUR IP, LLC.) 2017.03.02 단락 [0052]-[0053]; 및 도면 8-10 참조.	1-11

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌

“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

“&” 동일한 대응특허문헌에 속하는 문헌

“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

국제조사의 실제 완료일

2018년 06월 14일 (14.06.2018)

국제조사보고서 발송일

2018년 06월 14일 (14.06.2018)

ISA/KR의 명칭 및 우편주소



대한민국 특허청
(35208) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)

팩스 번호 +82-42-481-8578

심사관

노지명

전화번호 +82-42-481-8528



국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2011-0099574 A	2011/09/08	KR 10-1101641 B1	2012/01/02
KR 10-2010-0080099 A	2010/07/08	KR 10-1059557 B1	2011/08/26
KR 10-1485940 B1	2015/01/27	JP 2015-041386 A	2015/03/02
		JP 5830592 B2	2015/12/09
		US 2015-0058309 A1	2015/02/26
KR 10-0818357 B1	2008/04/02	KR 10-2007-0109207 A	2007/11/15
US 2017-0061485 A1	2017/03/02	CN 103430176 A	2013/12/04
		CN 103430176 B	2018/02/09
		EP 2689354 A2	2014/01/29
		EP 2689354 A4	2015/05/06
		JP 2014-515846 A	2014/07/03
		JP 6231472 B2	2017/11/15
		TW 201243632 A	2012/11/01
		TW 201512867 A	2015/04/01
		TW 201631505 A	2016/09/01
		TW 201723896 A	2017/07/01
		TW I479347 B	2015/04/01
		TW I543009 B	2016/07/21
		TW I585598 B	2017/06/01
		TW I604325 B	2017/11/01
		US 2012-0246165 A1	2012/09/27
		US 9495461 B2	2016/11/15
		US 9892426 B2	2018/02/13
		WO 2012-128888 A2	2012/09/27
		WO 2012-128888 A3	2013/03/14