(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2011년06월23일
(11) 등록번호 10-1043523
(24) 등록일자 2011년06월15일

(51) Int. Cl.
G06F 9/00 (2006.01)
(21) 출원번호 10-2005-0031597
(22) 출원일자 2005년04월15일
심사청구일자 2010년04월07일
(65) 공개번호 10-2006-0045782
(43) 공개일자 2006년05월17일
(30) 우선권주장
10/826,159 2004년04월15일 미국(US)
(56) 선행기술조사문헌
JP2003316819 A
KR1020030069640 A
논문
전체 청구항 수 : 총 21 항

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
장, 벤유
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
정, 후아-준
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
(뒷면에 계속)
(74) 대리인
주성민, 이중희, 백만기

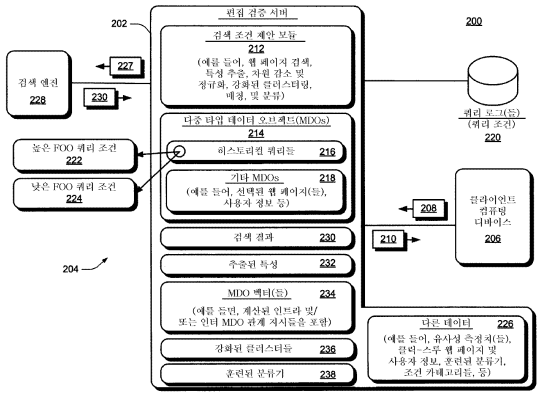
심사관 : 복진요

(54) 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된클러스터링

(57) 요 약

관련된 용어 제안을 위한 시스템 및 방법에 개시된다. 일 양상에서, 2개 이상의 다중 타입 데이터 오브젝트들중 각각의 오브젝트들간의 관계는 식별된다. 다중 타입 데이터 오브젝트들중 개별 오브젝트는 제1 타입의 적어도 하나의 오브젝트와, 제1 타입과는 구별되는 제2 타입의 적어도 하나의 오브젝트를 포함한다. 다중 타입 데이터 오브젝트는 개별 관계를 고려하여 반복적으로 클러스팅되어 강화된 클러스터를 생성한다.

대 표 도



(72) 발명자

리, 리

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

나즘, 타레크

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

마, 웨이-잉

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

리, 잉

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

첸, 젱

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

특허청구의 범위

청구항 1

명령어들을 실행하는 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 수행되는 컴퓨터-구현된 방법으로서,

적어도 하나의 제1 타입의 오브젝트 및 상기 제1 타입과는 다른 적어도 하나의 제2 타입의 오브젝트를 포함하는 다중타입 데이터 오브젝트(multi-type data object)들 사이의 관계들을 식별하는 단계 - 상기 관계들은 상기 다중타입 데이터 오브젝트들 사이의 계층간(inter-layer) 관계들 또는 계층내(intra-layer) 관계들 중 적어도 하나임 - ;

강화된 클러스터(cluster)들을 생성하기 위해, 상기 관계들에 기초하여 상기 다중타입 데이터 오브젝트들을 반복적으로 클러스터링하는 단계;

상기 강화된 클러스터들을 사용하여 비드 용어(bid term)와 관련된 제안된 검색 용어들의 리스트를 생성하는 단계 - 상기 검색 용어들은 사용자로부터 상기 비드 용어를 수신하는 것에 응답하여 생성됨 - ;

상기 제안된 검색 용어들의 리스트를 네트워크 인터페이스를 통해 사용자에게 전송하는 단계; 및

상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들의 중요도를,

$$\begin{cases} a(X) = \beta L_X{}^T h(X) + (1-\beta) L_{XY} i(Y) \\ h(X) = \beta L_X a(X) + (1-\beta) L_{XY} i(Y) \\ i(X) = a(X) + h(X) \\ \\ a(Y) = \gamma L_Y{}^T h(Y) + (1-\gamma) L_{YX} i(X) \\ h(Y) = \gamma L_Y a(Y) + (1-\gamma) L_{YX} i(X) \\ i(Y) = a(Y) + h(Y) \end{cases}$$

에 기초하여, 오브젝트 타입 내에서 및 서로 다른 오브젝트 타입들 사이에서 상호 강화하는 단계

를 포함하고,

$X=\{x_1,x_2,...x_m\}$ 및 $Y=\{y_1,y_2,...y_n\}$은 방향성을 고려하는 경우 관계들 R_X, R_Y, R_{XY} 및 R_{YX}를 갖는 이질성(heterogeneous) 오브젝트 타입의 각각의 오브젝트 집합들을 나타내며,

L_X 및 L_Y는 각각 집합 X 및 Y 내의 관계들을 식별하는 링크들의 인접 매트릭스들을 나타내고, L_{XY} 및 L_{YX}는 X 내의 오브젝트들로부터 Y 내의 오브젝트들로의 관계들을 식별하는 링크들의 인접 매트릭스들을 나타내고, a(X) 및 h(X)는 각각 X 내의 노드들의 오쏘리티(authority) 스코어 및 허브(hub) 스코어이고, a(Y) 및 h(Y)는 각각 Y 내의 노드들의 오쏘리티 스코어들 및 허브 스코어들을 나타내고, i(X) 및 i(Y)는 각각 X 및 Y 내의 노드들의 중요도를 나타내고, β 및 γ는 서로 다른 관계들로부터 유도된 링크들의 영향(influence)을 조정하기 위한 가중치 파라미터(weight parameter)들인 컴퓨터-구현된 방법.

청구항 2

제1항에 있어서,

상기 계층간 관계들은, 컨텐트(content) 관련 정보, 관련 주제에 대한 사용자의 관심, 및 관련 웹페이지에 대한 사용자의 관심 중 적어도 하나를 포함하는 컴퓨터-구현된 방법.

청구항 3

제1항에 있어서,

상기 계층내 관계들은, 쿼리 정제(query refinement), 추천된 웹페이지, 및 각각의 사용자들 사이의 관계 중 적어도 하나를 포함하는 컴퓨터-구현된 방법.

청구항 4

제1항에 있어서,

상기 다중타입 데이터 오브젝트들 각각은, 선택된 웹페이지 타입 및 사용자 정보 타입 중 적어도 하나에 관련되는 컴퓨터-구현된 방법.

청구항 5

제1항에 있어서,

상기 계층내 관계들은 제1 가중 스킴(weighting scheme)을 포함하고,

상기 계층내 관계들은 상기 다중타입 데이터 오브젝트들 중 관련된 오브젝트들에 대한 중요도를 나타내기 위해, 상기 제1 가중 스킴과 다른 제2 가중 스킴을 포함하는 컴퓨터-구현된 방법.

청구항 6

제1항에 있어서,

상기 식별하는 단계 및 상기 반복적으로 클러스터링하는 단계는 상기 제안된 검색 용어들에 대해 수행되는 컴퓨터-구현된 방법.

청구항 7

제1항에 있어서,

상기 반복적으로 클러스터링하는 단계는, 제1 반복의 클러스터링 결과들을 상기 다중타입 데이터 오브젝트들 중 모든 관련 데이터 오브젝트들에 전파시키는 단계를 포함하고,

상기 관련 데이터 오브젝트들 중 적어도 2개의 오브젝트들은 이질성 데이터 타입이며,

상기 전파는 강화된 클러스터링 동작들의 제2 반복에서 상기 다중타입 데이터 오브젝트 중 각각의 오브젝트들의 클러스터링을 강화하는 데 사용되는 컴퓨터-구현된 방법.

청구항 8

제1항에 있어서,

상기 반복적으로 클러스터링하는 단계는, 상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들 사이의 유사성(similarity)을 판정하는 단계를 포함하고,

상기 유사성은 상기 계층내 관계들과 상기 계층간 관계들 사이의 유사성들과, 오브젝트내(intra-object)와 오브젝트간(inter-object) 컨텐트 유사성 중 적어도 하나의 함수인 컴퓨터-구현된 방법.

청구항 9

제1항에 있어서,

상기 반복적으로 클러스터링하는 단계는,

상기 다중타입 데이터 오브젝트들 중 관련된 오브젝트들을 머지(merge)하여 상기 관련된 오브젝트들의 특성 공간 차원(feature space dimensionality)을 줄이는 단계를 포함하는 컴퓨터-구현된 방법.

청구항 10

제1항에 있어서,

상기 비드 용어를 상기 강화된 클러스터들 내의 오브젝트들의 특성 공간과 비교하여 상기 제안된 검색 용어들을 식별하는 단계

를 더 포함하는 컴퓨터-구현된 방법.

청구항 11

프로세서에 의해 실행가능한 컴퓨터-실행가능 명령어들을 포함하는 컴퓨터-판독가능 저장 매체로서, 상기 컴퓨

터-실행가능 명령어들은,

다중타입 데이터 오브젝트들 사이의 계층간 및 계층내 관계들 중 적어도 하나를 식별하는 단계 - 상기 다중타입 데이터 오브젝트들은 적어도 하나의 제1 타입의 오브젝트 및 상기 제1 타입과는 다른 적어도 하나의 제2 타입의 오브젝트를 포함함 - ;

강화된 클러스터들을 생성하기 위해, 상기 관계들 중 적어도 하나의 관계에 의해 상기 다중타입 데이터 오브젝트들을 반복적으로 클러스터링하는 단계;

상기 강화된 클러스터들을 사용하여 비드 용어와 관련된 제안된 검색 용어들의 리스트를 생성하는 단계 - 상기 제안된 검색 용어들은 사용자로부터 상기 비드 용어를 수신하는 것에 응답하여 생성됨 - ; 및

상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들의 중요도를,

$$\begin{cases} a(X) = \beta L_X{}^T h(X) + (1-\beta) L_{XY} i(Y) \\ h(X) = \beta L_X a(X) + (1-\beta) L_{XY} i(Y) \\ i(X) = a(X) + h(X) \\ \\ a(Y) = \gamma L_Y{}^T h(Y) + (1-\gamma) L_{YX} i(X) \\ h(Y) = \gamma L_Y a(Y) + (1-\gamma) L_{YX} i(X) \\ i(Y) = a(Y) + h(Y) \end{cases}$$

에 기초하여, 오브젝트 타입 내에서 및 서로 다른 오브젝트 타입들 사이에서 상호 강화하는 단계

를 수행하고,

$X=\{x_1,x_2,...x_m\}$ 및 $Y=\{y_1,y_2,...y_n\}$은 방향성을 고려하는 경우 관계들 R_X, R_Y, R_{XY} 및 R_{YX}를 갖는 이질성 오브젝트 타입의 각각의 오브젝트 집합들을 나타내며,

L_X 및 L_Y는 각각 집합 X 및 Y 내의 관계들을 식별하는 링크들의 인접 매트릭스들을 나타내고, L_{XY} 및 L_{YX}는 X 내의 오브젝트들로부터 Y 내의 오브젝트들로의 관계들을 식별하는 링크들의 인접 매트릭스들을 나타내고, a(X) 및 h(X)는 각각 X 내의 노드들의 오쏘리티 스코어 및 허브 스코어이고, a(Y) 및 h(Y)는 각각 Y 내의 노드들의 오쏘리티 스코어들 및 허브 스코어들을 나타내고, i(X) 및 i(Y)는 각각 X 및 Y 내의 노드들의 중요도를 나타내고, β 및 γ는 서로 다른 관계들로부터 유도된 링크들의 영향을 조정하기 위한 가중치 파라미터들인 컴퓨터-판독가능 저장 매체.

청구항 12

제11항에 있어서,

상기 계층간 관계들은, 컨텐트 관련 정보, 관련 주제에 대한 사용자의 관심, 및 관련 웹페이지에 대한 사용자의 관심 중 적어도 하나를 포함하는 컴퓨터-판독가능 저장 매체.

청구항 13

제11항에 있어서,

상기 계층내 관계들은, 쿼리 정제, 추천된 웹페이지, 및 각각의 사용자들 사이의 관계 중 적어도 하나를 포함하는 컴퓨터-판독가능 저장 매체.

청구항 14

제11항에 있어서,

상기 다중타입 데이터 오브젝트들 각각은, 검색 쿼리 데이터 오브젝트 타입, 선택된 웹페이지 타입 및 사용자 정보 타입 중 적어도 하나에 관련되는 컴퓨터-판독가능 저장 매체.

청구항 15

제11항에 있어서,

식별된 관계들 중 적어도 하나는 상기 다중타입 데이터 오브젝트들 중 관련된 오브젝트들에 대한 중요도를 나타내기 위해 가중되는 컴퓨터-판독가능 저장 매체.

청구항 16

제11항에 있어서,

상기 식별하는 단계 및 상기 반복적으로 클러스터링하는 단계는 상기 제안된 검색 용어들에 대해 수행되는 컴퓨터-판독가능 저장 매체.

청구항 17

제11항에 있어서,

상기 반복적으로 클러스터링하는 단계는, 제1 반복의 클러스터링 결과들을 상기 다중타입 데이터 오브젝트들 중 모든 관련 데이터 오브젝트들에 전파시키는 단계를 포함하고,

상기 관련 데이터 오브젝트들 중 적어도 2개의 오브젝트들은 이질성 데이터 타입이며,

상기 전파는 강화된 클러스터링 동작들의 제2 반복에서 상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들의 클러스터링을 강화하는데 사용되는 컴퓨터-판독가능 저장 매체.

청구항 18

제11항에 있어서,

상기 반복적으로 클러스터링하는 단계는, 상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들 사이의 유사성을 판정하는 단계를 포함하고,

상기 유사성은 식별된 관계들 중 적어도 하나 사이의 유사성들과, 오브젝트 컨텐트 유사성 중 적어도 하나의 함수인 컴퓨터-판독가능 저장 매체.

청구항 19

제11항에 있어서,

상기 반복적으로 클러스터링하는 단계는, 상기 다중타입 데이터 오브젝트들 중 관련된 오브젝트들을 머지하여 상기 관련된 오브젝트들의 특성 공간 차원(feature space dimensionality)을 줄이는 단계를 포함하는 컴퓨터-판독가능 저장 매체.

청구항 20

제11항에 있어서,

상기 명령어들은, 상기 다중타입 데이터 오브젝트들 중 각각의 오브젝트들의 중요도를 오브젝트 타입 내에서 및 서로 다른 오브젝트 타입들 사이에서 상호 강화하는 단계를 수행하기 위한 명령어들을 더 포함하는 컴퓨터-판독가능 저장 매체.

청구항 21

제11항에 있어서,

상기 명령어들은, 상기 비드 용어를 상기 강화된 클러스터들 내의 오브젝트들의 특성 공간과 비교하여 상기 제안된 검색 용어들을 식별하는 단계를 수행하기 위한 명령어들을 더 포함하는 컴퓨터-판독가능 저장 매체.

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

[0030] 본 출원은, 다음과 같은 특허 출원들과 연관되며, 그 각각은 본 출원의 양수인에게 일반 양도되고, 참조로 본 원에 포함된다.

[0031] ● "Object Clustering Using Inter-Layer Links"라는 제목으로 2003년 5월 1일에 출원된 미국 특허 출원 번호 10/427,548; 및

[0032] ● "Related Term Suggestion for Multi-Sense Query"라는 제목으로 2004년 4월 15일에 출원된 미국 특허 출원(출원 번호는 배정될 예정임).

[0033] 본 발명은 데이터 마이닝(data mining)에 관한 것이며, 보다 구체적으로는 검색어 제안을 위한 시스템 및 방법들을 개선하기 위한 이종 오브젝트들의 클러스터링에 관한 것이다.

[0034] 키워드 또는 문구는, WWW(World Wide Web) 상의 관련 웹 페이지/사이트를 검색할 때, 웹 서퍼(Web surfer)에 의해 검색 엔진에 제공되는 단어 또는 용어 세트이다. 검색 엔진들은 페이지/사이트 상에 나타나는 키워드들 및 키워드 문구들에 기초하여 웹 사이트의 연관성을 결정한다. 상당 부분의 웹 사이트 트래픽이 검색 엔진의 사용으로부터 초래되기 때문에, 웹 사이트 프로모터(promoter)들은, 적절한 키워드/문구 선택이 원하는 사이트 노출을 얻기 위해 사이트 트래픽을 증가시키는 데 중요하다는 것을 안다. 검색 엔진 결과 최적화를 위해 웹 사이트에 연관된 키워드들을 식별하는 기술들은, 예를 들어 사람에 의한 웹 사이트 컨텐트의 평가를 포함하고 연관된 키워드(들)을 식별하고자 한다. 이러한 평가는 키워드 인기도 툴(keyword popularity tool)의 사용을 포함할 수 있다. 그러한 툴들은, 얼마나 많은 사람이 특정 키워드 또는 키워드를 포함한 문구를 검색 엔진에 제공했는지를 결정한다. 웹 사이트에 연관되고 검색 쿼리들을 생성하는데 보다 자주 사용될 것으로 판단된 키워드들이, 웹 사이트에 대해 검색 엔진 결과 최적화를 위해 일반적으로 선택된다.

[0035] 웹 사이트의 검색 엔진 결과 최적화를 위한 키워드 세트를 식별한 후, 프로모터는 웹 사이트를 (다른 웹 사이트 검색 엔진 결과들의 표시된 위치들과 비교하여) 검색 엔진의 결과들 내의 보다 높은 위치로 전진시키기를 원할 수 있다. 이를 위해, 키워드(들)과 연관된 프로모터의 목록들에 웹 서퍼가 클릭할 때마다 프로모터가 얼마를 지불할 것인지를 나타내도록 프로모터가 키워드(들)을 입찰한다. 즉, 키워드 입찰은 클릭-당-지불(pay-per-click) 입찰이다. 키워드 입찰량이 동일한 키워드에 대한 다른 입찰에 비해 보다 클수록, 검색 엔진이, 연관된 웹 사이트를 키워드에 기초한 검색 결과들 내에 보다 높게(중요성에 있어서 보다 두드러지게) 표시할 것이다.

[0036] 웹 사이트 컨텐트와 연관된 입찰 용어(들)을 식별하는 종래의 시스템 및 기술들은, 통상적으로 클러스터링 알고리즘을 사용하여, 동일한 클러스터로부터의 오브젝트들은 유사하고 서로 다른 클러스터들로부터의 오브젝트들은 상이하다는 방식으로 오브젝트 세트를 그룹들 또는 클러스터들로 분할한다. 그러한 클러스터링 접근법들은, 클러스터될 데이터 오브젝트들이 독립적이고 동일한 클래스이고, 특징/속성 값들의 고정 길이 벡터에 의해 종종 모델링된다고 가정한다. 데이터 마이닝 검색의 최근 격동에 있어, 이러한 고전적인 문제가 큰 데이터베이스 배경에서 다시 검토되어왔다. 그러나, 웹 마이닝 및 협업 필터링(collaborative filtering)과 같은 소정의 떠오르는 애플리케이션들이 그러한 가정에 도전장을 던지더라도, 클러스터될 데이터 오브젝트들의 동질성(homogeneity)은 여전히 기본적인 가정이 되고 있다. 그러한 애플리케이션들에서, 데이터 오브젝트들은 서로 다른 유형들이고 고도로 서로 관련되어 있다. 불행하게도, 이종 오브젝트 유형들에 걸쳐 분산된 오브젝트들이 고도로 서로 관련되어 있더라도, 종래의 클러스터링 동작들은 통상적으로, 서로 다른 오브젝트 유형들의 임의의 서로 관련된 측면에 대한 고려없이, 각각의 오브젝트 유형들을 개별적으로 클러스터링한다.

발명이 이루고자 하는 기술적 과제

[0037] 이러한 것에 대한 한가지 원인은, 서로 다른 유형의 데이터 오브젝트들 간의 관계들이 종종 성기고 식별하기 어렵기 때문이다. 다른 원인은, 오브젝트 각각에 부착된 정적인 고정 길이 값 벡터 - 상기 벡터는 오브젝트 속성들 및 서로 다른 유형의 관련된 오브젝트의 속성들 모두를 나타냄 - 를 갖는 임의의 그러한 관계들의 표현이 오브젝트 속성/특징 벡터들을 매우 높은 차원(dimensionality)(특징 공간)으로 생성하기 때문이다. 그러한 높은 차원은, 데이터가 특징 공간 내에서 서로 멀리 떨어져 있기 때문에 바람직하지 않고, 효율적인 모델들이 작은 영역들 내의 그와 같이 드문 양의 데이터로 충분히 훈련될 수 없다.

[0038] 따라서, 이종 데이터 오브젝트들 간의 관계의 측면에서 관련 오브젝트들(예를 들면 용어들)을 식별하고 그룹화하는 보다 나은 클러스터링 기술들이 유용할 것이다. 이러한 클러스터링 기술들이 사용되어, 예를 들면, 검색 엔진 최적화 및 용어 입찰을 위해 용어(들)을 식별하는 시스템 및 방법들을 제공함으로써, 관련 용어(들)을 실질적으로 보다 높은 확률로 식별할 수 있다.

발명의 구성 및 작용

[0039] <발명의 요약>

[0040] 관련어 제안을 위한 시스템 및 방법들이 설명된다. 한 측면에서, 2개 이상의 다중 유형 데이터 오브젝트들의 각각의 오브젝트들 간에 계층간(inter-layer) 및/또는 계층내(intra-layer) 관계들이 식별된다. 다중 유형 데이터 오브젝트들의 각각의 오브젝트들은, 제1 유형의 적어도 하나의 오브젝트 및 제1 유형과는 상이한 제2 유형의 적어도 하나의 오브젝트를 포함한다. 다중 유형 데이터 오브젝트들은 관계들 각각의 측면에서 반복적으로 클러스터링되어 강화된 클러스터들을 생성한다.

[0041] 도면들에서, 구성요소의 참조 번호의 맨좌측 숫자는 그 구성요소가 처음 나타난 특정 도면을 나타낸다.

[0042] **개관**

[0043] 도 1은 서로 관계가 있는 이질적인 오브젝트 데이터 타입들의 예시적인 프레임워크(100)를 도시한다. 프레임워크(100)는 이질적인 데이터 오브젝트들/노드들의 다층들(02) 및 관련 층간(interlayer) 및 층간 데이터 오브젝트 링크들/관계들을 포함한다. 각 층(102-1 내지 102-N)은 동일한 타입(동질)의 데이터 오브젝트들 또는 노드들의 개별적인 세트를 포함한다. 즉, 노드 세트 P는 각각이 동일한 데이터 타입인 하나 이상의 데이터 오브젝트들(p_1 내지 p_j)을 포함하고, 노드 세트 U는 각각이 동일한 데이터 타입인 하나 이상의 데이터 오브젝트들(u_1 내지 u_k)를 포함한다. 따라서, 상이한 개별적인 층들(102)에 있는 데이터 오브젝트들의 타입들은 서로에 대해 이질적이다.

[0044] 이러한 구현에 있어서, 예를 들어:

[0045] · 층(102-1)은 검색 쿼리 데이터 오브젝트(들)/노드(들)(p_1 내지 p_j)를 포함한다. 검색 쿼리 오브젝트들은 후술하는 바와 같이 쿼리 용어(들)를 포함하며 쿼리 로그로부터 도출된 히스토리 쿼리들의 개별적인 쿼리들을 나타낸다.

[0046] · 층(102-1)은 도출된 웹페이지 층이고, 웹페이지 데이터 오브젝트(들)/노드(들)(u_1 내지 u_k)를 포함한다.

[0047] · 층(102-3)은 도출된 사용자층이고, 사용자 정보 오브젝트(들)/노드들(w_1 내지 w_m)을 포함한다.

[0048] · 층(102-N)은 개별적인 상이한 오브젝트 타입들(x_1 내지 x_o)를 포함하는 임의 수의 층들(102)이 존재할 수 있음을 예시하기 위해 도시된다.

[0049] 한쌍의 데이터 오브젝트들 사이에 연장하는 라인들/링크들은 개별적인 데이터 오브젝트들 사이에 존재하도록 결정된 개별적인 도출 관계들을 나타낸다. 클러스터링에 관한 특정 실시예들에서, 라인들/링크들은 "에지들(edges)"로 지칭된다. 일반화된 용어인 라인 또는 링크는 본 개시물에서는 링크들, 에지들, 또는 오브젝트들 사이의 관계를 기술하는 또 다른 오브젝트에 대한 하나의 오브젝트의 임의의 커넥터를 기술하기 위해 사용된다. (데이터 오브젝트들 사이의 관계를 나타내는 화살촉들에 의해 제공된 바와 같은) 링크 방향은 참가 오브젝트들의 속성 함수로서 어느 한쪽 방향으로 지향될 수 있다. 링크들은 예시적으로 여겨지고 범위가 한정되지 않는다. 프레임워크(100)에 의해 도시된 바와 같은 웹 환경에서의 특정 링크들은 한 방향으로 보다 적절하게 지향될 수 있고, 화살촉의 방향은 일반적으로 후술된 강화 클러스터링 동작들에 영향을 주지 않을 것이다.

[0050] 오브젝트 쌍들 사이의 링크들은 층간 또는 층간 링크들로서 분류될 수 있다. 층간 링크는 동일한 타입의 상이

한 오브젝트들 사이에서 식별된 관계를 나타낸다. 따라서 층간 링크들(104)은 동일층(102) 내에서 오브젝트들을 접속시킨다. 예를 들어, 개별적인 쌍의 데이터 오브젝트들 사이의 실선(104)은 층간 링크를 나타낸다. 이러한 예에서, 층간 링크는 웹페이지 오브젝트 u_2로부터 다른 웹페이지 오브젝트 u_3으로 연장하고, 상이한 웹페이지들 사이의 관계(들)를 나타낸다.

[0051] 층간 링크는 상이한 타입들의 데이터 오브젝트들 사이의 관계들을 기술한다. 층간 링크들은 한쌍의 이질적인 오브젝트들의 개별적인 오브젝트들 사이로 연장하기 때문에, 참가중인 데이터 오브젝트들의 쌍의 각각은 상이한 개별적인 데이터 오브젝트/노드 세트층(102) 상에 도시된다. 도 1에 도시된 바와 같이, 실선이 아닌 오브젝트들의 쌍을 연결하는 임의의 라인은 층간 링크이다. 예를 들어, 링크(106)는 제1의 오브젝트들의 쌍으로부터 제2의 오브젝트들의 쌍으로의 레퍼런스(예를 들어, 하이퍼링크)를 나타내고, 링크/라인(108)은 제1의 오브젝트들의 쌍으로부터 제2의 오브젝트들의 쌍으로 공유/참조된 발행물(예를 들어, 내용)를 나타내고, 링크/라인(110)은 제1의 오브젝트들의 쌍으로부터 제2의 오브젝트들의 쌍으로의 브라우즈 링크를 나타낸다. 다른 예에서, 링크는 사용자 오브젝트 w_4로부터 검색 쿼리 오브젝트 p_5 및 웹페이지 오브젝트 u_5로 연장하고, 사용자마다 관련성을 선택한 웹페이지를 리턴시키는 쿼리를 사용자가 제출하는 것를 나타낼 수 있다.

[0052] 도 1의 예에서, 그리고 내부 및 층간 링크들의 개별적인 링크들에 의해 도시된 바와 같이, 상이한 오브젝트들의 타입들(p, u, w, ...)은 관련된다. 예를 들어, (오브젝트들 w에 의해 표시된) 사용자는 쿼리들(오브젝트들 p)을 발행하고; 사용자는 발행된 쿼리들의 수신에 응답하여 검색 엔진에 의해 리턴된 웹페이지들(오브젝트들 u)을 브라우징하고; 각 검색 쿼리(오브젝트 p)는 하나 이상의 개별적인 웹페이지들(오브젝트들 u)을 참조한다. 이 관점에서, 웹 사용자 정보가 클러스터링되면, 사용자가 브라우징하고 있는 웹페이지(들) 및 개별적인 웹페이지(들)를 얻기 위해 사용되는 쿼리들은 보다 많은 유사성을 가져야 하고 클러스터링 프로세스에서 함께 클러스터링되는 경향이 있다. 마찬가지로, 웹페이지들을 클러스터링할 때, 웹페이지들이 사용자에 의해 사용되고 개별적인 검색 쿼리들에 의해 참조되는 방식을 고려해야 한다. 이를 해결하기 위해, 후술되는 바와 같이, 강화된 클러스터링 알고리즘은 이러한 이질적인 데이터 오브젝트들을 데이터 오브젝트들의 개별적인 오브젝트들 사이의 도출 관계들의 함수로서 클러스터링한다.

[0053] 본 개시물의 일 양태는 내재 상호 관계에 기초하며, 클러스터링되는 오브젝트들에는 다른 오브젝트들에 대한 링크들이 제공된다. 각 오브젝트에 접속하는 링크들의 특정 링크들(및 그 링크들이 접속하는 오브젝트들)은 그들의 관련성을 그 오브젝트에 반영하기 위해 상이한 중요성으로 가중화될 수 있다. 예를 들어, 클러스터링되는 타입과 동일한 타입의 오브젝트들에는 상이한 타입의 오브젝트들보다 큰 중요성이 제공될 수 있다. 본 개시물은 상이한 오브젝트들 또는 상이한 타입의 오브젝트들에 할당될 수 있는 중요성 레벨들을 가변시키는 메카니즘을 제공한다. 상이한 레벨의 중요성을 상이한 오브젝트들(또는 상이한 타입의 오브젝트들)에 할당하는 것은 여기서 중요성에 의한 클러스터링이라 지칭된다. 상이한 오브젝트들의 중요성 레벨들을 가변시키는 것에 의해 종종 클러스터링 결과들 및 효율성이 향상된다. 검색 용어 제안들을 위한 다중 타입 데이터 오브젝트들의 강화된 클러스터링에 대한 상기 및 다른 양태들은 이제 설명된다.

[0054] 웹사이트와 관련되고, 최종 사용자들에 의해 검색 쿼리들을 생성할 때 보다 종종 사용되도록 결정되는 용어(들)/키워드(들)은 일반적으로 웹사이트들에 대하여 검색 엔진 결과 최적화를 위해 웹사이트 프로모터들/광고자들에 의해 선택된다. 이것을 염두해 두고, 이하에 개시된 시스템들 및 방법들은, 본 구현에서는 검색 용어 제안인 작업과 상호 관계되도록 결정된 다중 타입 데이터 오브젝트들을 도출한다. 이러한 다중 타입 데이터 오브젝트들은 검색 엔진에 대한 히스토리 쿼리들, 특정 히스토리 검색 쿼리에 응답하여 사용자에 의해 선택된 웹페이지들의 세트, 사용자에게 고유한 정보(예를 들어, 사용자의 웹사이트 액세스 정보, 검색 쿼리를 발생시키는데 사용되는 기계의 IP 어드레스 등), 및/또는 관련 데이터 오브젝트들의 타입들을 제출함으로써 얻어지는 결과들로부터 도출된 의미 콘텍스트(예를 들어, 텍스트, URL, 결과 타이틀, 및 각 결과에 대한 짧은 기술 등)에 의해 향상되는 도출된 히스토리 검색 쿼리들의 용어(들)를 포함한다.

[0055] 이들 다중 타입 데이터 오브젝트들 사이의 유사도는 식별되고 가중된 컨텐트 유사도와 계산된 오브젝트간(inter-object) 및 오브젝트내(intra-object) 관계 유사도들의 선형 조합으로서 결정된다. 오브젝트간 및 오브젝트내 타입 관계들로부터 도출된 링크 구조를 분석함으로써 데이터 오브젝트들에 상이한 가중들이 할당된다. 따라서, 데이타 오브젝트들의 각각의 것들 사이의 유사도는 그들 자신의 속성의 유사도 뿐만 아니라 그들의 관계의 유사도를 포함한다.

[0056] 계산된 다중 타입 오브젝트 관계들의 관점에서, 강화된 클러스터링 알고리즘은 각각의 오브젝트의 식별된 오브젝트간 및 오브젝트내 관계 속성들의 함수로서 다중 타입 데이터 오브젝트들을 반복적으로 클러스터링한다. 이

러한 구현에 있어서, 수정된 다이렉트-k-평균(direct-k-means) 클러스터링 알고리즘이 사용되어 클러스터 내의 오브젝트들의 가중 합을 사용함으로써 클러스터 중심값들을 결정한다. 이것은 그들의 각각의 관계 속성들을 업데이트함으로써 모든 관련된 데이터 오브젝트들에 클러스터링 결과들을 전달하는 반복적인 프로세스이다. 즉, 한가지 유형의 오브젝트의 클러스터링 결과들은 새로운 특성 공간을 형성하며, 이것은 그외의 관련되었으나 상이한 유형의 오브젝트들로 투영되고 전달된다. 그 다음에 이 업데이트된 특성 공간을 이용하여 관련된 유형의 오브젝트들에 대한 클러스터링을 수행한다. 이 반복적인 강화 프로세스는 각각의 오브젝트 타입들에 대해 실행되어 실질적으로 관련된 클러스터 노드들을 병합하여 특성 공간 차원을 감소시키고, 모든 다중 타입 오브젝트들에 걸친 클러스터링 결과들이 수렴될 때까지 계속된다. 이것은 다중 타입 데이터 오브젝트들과 실질적으로 매우 관련된 강화된 클러스터들을 발생시킨다.

[0057] 최종 사용자로부터 용어를 수신하는 것에 응답하여, 시스템 및 방법은 용어(들)을 용어/쿼리(query) 오브젝트 유형에 기초한 강화된 클러스터들의 용어(들) 중 각각의 것들과 비교한다. 강화된 용어 클러스터들은 문맥적으로 서로 관련되어 있는 용어(들)을 포함하기 때문에, 제출된 비드(bid)를 클러스터들 내의 용어들과 비교하는 경우에, 용어 문장은 임의의 다수의 관련된 컨텍스트들 또는 "의식(senses)"들의 관점에서 평가된다. 또한, 각각의 강화된 용어 클러스터가 매우 관련된 다중 타입 오브젝트들의 세트들로부터 도출되기 때문에, 알고리즘은 순수한 컨텐트 기반의 방법의 결함을 극복할 수 있다, 즉, 쿼리 용어들 사이의 의미론적인 관계들을 효과적으로 강화시키고, 용어 컨텍스트의 노이즈의 영향을 억제할 수 있다. 수신된 용어를 강화된 클러스터들의 오브젝트들의 특성 공간들과 비교하는 것에 응답하여, 하나 이상의 검색어 제안들이 식별된다. 이 검색어 제안들은 최종 사용자로 전달된다.

[0058] **예시의 시스템**

[0059] 비록 요구되지는 않았지만, 본 발명은 퍼스널 컴퓨터에 의해 실행되는 컴퓨터 실행가능한 명령어들(프로그램 모듈들)의 일반적인 문맥으로 서술된다. 프로그램 모듈들은 특정 태스크들을 수행하거나 또는 특정 추상 데이터 타입을 구현하는 루틴, 프로그램, 오브젝트, 컴포넌트, 데이터 구조 등을 일반적으로 포함한다. 시스템 및 방법들이 전술한 문맥으로 서술되지만, 이하에 기술되는 작용들 및 동작들도 하드웨어에서 구현될 수 있다.

[0060] 도 2는 검색어 제안을 위한 다중 타입 데이터 오브젝트들의 강화된 클러스터링을 위한 예시적인 시스템(200)을 도시한다. 본 구현에 있어서, 시스템(200)은 네트워크를 통해 클라이언트 컴퓨팅 디바이스(206)에 연결된 EVS(editorial verification server)(202)를 포함한다. 수신 용어(들)(208)에 응답하여, 예를 들면, 클라이언트 컴퓨팅 디바이스(206) 또는 EVS(202) 상에서 실행하는 다른 애플리케이션(도시되지 않음)으로부터, EVS(202)는 검색어 리스트(210)를 발생시켜 클라이언트 컴퓨팅 디바이스(206)로 전달하여, 최종 사용자가 용어(들)에 대해 실제적으로 비딩하기 전에 용어(들)(208)에 의미론적으로 및/또는 문맥적으로 관련된 용어들의 세트를 평가할 수 있게 한다. 네트워크(204)는 사무실, 기업형 컴퓨터 네트워크, 인트라넷, 및 인터넷에서 일반적인 것들인 LAN(local area network) 및 일반적인 WAN(wide area network) 통신 환경들의 임의의 조합을 포함할 수 있다. 시스템(200)이 클라이언트 컴퓨팅 디바이스(206)를 포함하는 경우에, 클라이언트 컴퓨팅 디바이스는 퍼스널 컴퓨터, 랩탑, 서버, 이동 컴퓨팅 디바이스(예를 들어, 셀룰러 폰, PDA(personal digital assistant), 휴대용 컴퓨터) 등과 같은 임의의 타입의 컴퓨팅 디바이스이다.

[0061] EVS(202)는 다수의 프로그램 모듈들을 포함하여 제안어 리스트(210)를 발생시킨다. 컴퓨터 프로그램 모듈들은, 예를 들어, STS(search term suggestion) 모듈(212)을 포함한다. 본 구현에서, 논의 및 예시의 설명을 목적으로, STS 모듈(212)은 이력적 쿼리어 마이닝(historical query term mining), 웹 페이지 검색, 특성 추출, 특성 공간 차원 감소 및 정규화, 다중 타입 데이터 오브젝트들의 강화된 클러스터링, 사용자 비드 용어(들)을 검색어 제안을 수행하기 위해 강화된 클러스터들의 컨텐츠와 매칭, 및 용어 분류와 같은 다수의 기능들을 수행하는 것으로서 기술된다. 이 동작들의 각각의 것들은 STS 모듈(212)과 통신하여 하나 이상의 그외의 컴퓨터 프로그램 모듈들(도시되지 않음)에 의해 수행될 수 있다.

[0062] 채굴된 히스토리 검색 쿼리 d의 의미론 구문(Semantic Context)에 의한 향상

[0063] 이러한 구현에서, 준비된 태스크에 상호 관련되도록 결정된 다중 타입의 데이터 오브젝트(MDO)(214)를 채굴하는 STS 모듈(212)은 검색 조건(용어)을 제안한다. 이러한 다중 타입 데이터 오브젝트(214)는 히스토리 쿼리들(216) 각각을 검색 엔진에 제출함으로써 구해진 검색 결과들로부터 채굴된 의미론 구문(예를들어, 텍스트, URL, 결과 타이틀, 및 각 결과의 간단한 설명 등)을 갖는 STS 모듈(212)에 의해 향상될 채굴된 히스토리 검색 쿼리(216) 조건(용어)과, 특정 히스토리 검색 쿼리에 응답하여 사용자에 의해 선택된 웹 페이지 세트, 사용자에 특정한 정보(예를들어, 사용자의 웹 사이트 액세스 정보, 검색 쿼리 생성을 위해 사용된 머신의 IP 어드레스 등),

및/또는 관련된 다중 타입 데이터 오브젝트들과 같은 "다른 MDO"(218)를 포함한다.

[0064] 특히, STS 모듈(212)은 쿼리 로그(220)로부터 한 세트의 히스토리 쿼리(216)를 복원한다. 히스토리 쿼리(216)는 하나 이상의 사용자들에 의해 이전에 검색 엔진으로 제출된 검색 쿼리 조건(용어)을 포함한다. STS 모듈(212)은 높은 발생 빈도(high frequency of occurence)의 검색 조건(용어)(222)과 비교적 낮은 발생 빈도의 검색 조건(용어)(224)을 식별하기 위한 발생 빈도의 함수로서 히스토리 쿼리(216)를 평가한다. 이러한 구현에서, 구성가능한 트레시홀드값이 히스토리 쿼리가 비교적 높은 빈도로 발생한 것인지 낮은 빈도로 발생한 것인지 여부를 판정하기 위해 사용된다. 예를들어, 히스토리 쿼리(216)에서 적어도 트레시홀드배수 발생한 검색 쿼리 조건(용어)은 높은 발생 빈도를 갖는 것이라고 한다. 유사하게, 히스토리 쿼리(216)에서 트레시홀드배수 이하로 발생하는 검색 쿼리 조건은 낮은 발생 빈도를 갖는 것이라고 한다. 설명할 목적으로, 이러한 트레시홀드 값을 "다른 데이터"(226)의 각 부분으로서 도시한다.

[0065] STS 모듈(212)은 각 쿼리를 검색 엔진(228)에 하나씩(검색 쿼리(227)) 제출함으로써 높은 발생 빈도 쿼리 조건(222)의 의미론/구문 의미를 채굴한다. 검색 쿼리(227) 수신에 응답하여, 검색 엔진(228)은 검색 결과(230)에 랭크된 리스트(번호가 구성가능한)를 STS 모듈(212)로 입력한다. 랭크된 리스트는 URL, 결과 리스트, 및 간단한 설명 및/또는 제출된 검색 쿼리(227)에 관련된 쿼리 조건의 배경(상황)을 포함한다. 랭크된 리스트는 검색 결과(230)에 기억된다. 이러한 검색 결과 복원은 각 검색 쿼리(227)에 대해 행해진다.

[0066] STS 모듈(212)은 URL, 결과 타이틀 및 간단한 설명 및/또는 각 복원된 검색 결과(230)로부터 각 쿼리 조건(222)에 해당하는 쿼리 조건의 배경을 추출하기 위해 웹 페이지 하이퍼텍스트 생성 언어(HTML)를 해석한다. URL, 결과 타이틀 및 간단한 설명 및/또는 쿼리 조건의 배경, 및 검색 쿼리(227)는 추출된 특성(232)의 각 기록에서 STS 모듈(212)에 의해 저장된 복원 검색 결과(230)를 구하기 위해 사용된다.

[0067] 높은 발생 빈도 쿼리 조건(222)에 대한 검색 결과(230)를 해석한 다음, STS 모듈(212)은 추출된 특성(232)에 대해 텍스트 전처리 동작을 행하여 추출된 특성들로부터의 언어적 토큰(토큰화하는)을 개별 키워드로 발생한다. 토큰의 크기를 감소시키기 위해, STS 모듈(212)은 (예를들면, "the", "a", "is" 등의) 단어를 제거하고 키워드를 표준화하기 위해 예를들어 공지된 포터 스테밍 알고리즘(Porter stemming algorithm)을 사용하여 공통의 접미사를 제거한다. STS 모듈(212)은 결과의 추출된 특성(232)을 다중 타입 데이터 오브젝트(MDO) 벡터(234)에 기초한 하나 이상의 조건으로 배열한다.

[0068] 다중 타입 데이터 오브젝트(MDO) 벡터(234)에 기초한 각 조건은 빈도를 나타내는 조건과 반전된 문서 빈도(TFIDF) 기록에 기초한 크기를 갖는다. i^{th} 벡터에 대한 웨이트와 j^{th} 키워드는 다음과 같이 계산된다.

[0069] $$w_{ij} = TF_{ij} x \log(N / DF_j)$$

[0070] 여기서, TF_{ij}는 빈도의 조건을 나타내고(i^{th} 기록에서 키워드 j의 발생 횟수), N은 쿼리 조건의 전체 수, 그리고 DF_j는 키워드 j를 포함하는 기록 수를 나타낸다.

[0071] 각 쿼리 조건의 벡터 표시가 주어지면, 한 쌍의 조건 간의 관계를 판정하기 위해 코사인 함수가 사용된다(벡터들이 정상화되었음을 상기하기 바람).

[0072] $$sim(q_j, q_k) = \sum_{i=1}^{d} w_{ij} \cdot w_{ik}$$

[0073] 따라서, 두개 조건들 간의 거리(유사성 판정)가

[0074] $$dist(q_j, q_k) = 1 - sim(q_j, q_k)$$

[0075] 으로서 정해진다.

[0076] 이러한 유사성 판정은 "다른 데이터"(226)의 각 부분으로서 도시된다. 예시적인 이러한 유사성 값들은 이하에서 설명된 표 1의 예시적으로 제안된 조건 리스트(210)에 도시된다.

[0077] 사용자가 선택한 웹 페이지(들) 및 사용자 정보 마이닝(Mining User Selected Web Page(s) and User Information)

[0078] 웹 사이트의 검색 엔진 결과 최적하에 대해 실질적으로 가장 관련있는 세트의 용어(들)을 식별하기 위해(검색 용어 제안), STS 모듈(212)은 히스토리컬 쿼리(216)와는 다른/이질적적인 다중-타입 데이터 오브젝트들(214)을 마이닝한다. 논의의 편의상, 이들 마이닝된 오브젝트들은 "기타 MDOs(other MDOs)"로 표현된다. 일 실시예에서, "기타 MDOs"(218)는, 예를 들어, 최종 사용자가 선택한 웹 페이지 및/또는 사용자 특정 정보를 포함하고, 여기서 사용자는 검색 엔진(228)에 히스토리컬 쿼리(216)를 제출하는 것과 관련된 사용자다. 최종 사용자가 선택한 웹 페이지들은 스파스(sparse)이거나 또는 스파스가 아닐 수 있고, 여기서 스파스는, 예를 들어, 히스토리컬 쿼리(216) 당 2개 내지 3개의 웹 페이지를 평균화한다. STS 모듈(212)은 쿼리 로그(들)(220) 또는 기타 데이터 소스들로부터 사용자 특정 정보를 추출한다. 사용자 특정 정보는, 예를 들어, 히스토리컬 쿼리(216), GUID 및/또는 웹 사이트 액세스 정보(예를 들어, 마이크로소프트사의 ".net Passport information" 등)를 각각 제출하는데 사용되는 머신의 인터넷 프로토콜(IP) 어드레스를 포함한다.

[0079] 강화된 다중-타입 데이터 오브젝트 클러스터링(Reinforced Multi-Type Data Object Clustering)

[0080] STS 모듈(212)은 클러스터링 분석을 위해 다중-타입 상호 관련된 데이터 오브젝트들[MDOs(214)]간 관계를 모두 탐색한다. 다중-타입 데이터 오브젝트(214)는 n개의 서로 다른 타입의 오브젝트들 X_1, X_2, ..., X_n[예를 들어, 히스토리컬 쿼리(216) 및 "기타 MDOs"(218) 등]을 포함한다. 각 타입의 데이터 오브젝트 Xi는 특징들의 세트 Fi로 묘사된다. 동일한 타입 내에서의 데이터 오브젝트들은 인트라-타입 관계식 Ri ⊆ Xi ×Xi로 상호 관련된다. 이러한 관계식으로부터 구별하기 위해, Fi는 데이터 오브젝트들의 컨텐츠 특징(content feature)으로서 참조된다. 특정 오브젝트 x ∈ Xi에 대하여, x.Fi를 사용하여 그 컨텐츠 특징들을 나타내고, x.Ri ⊆ Xi와 x.Rij ⊆ Xj를 사용하여 Xi 및 Xj 각각에서 관련되는 오브젝트들을 표시한다. 다중-타입 상호 관련된 데이터 오브젝트들의 클러스터링 문제점은 각 타입의 오브젝트들 Xi를 Ki의 클러스터로 구획화하여 각 클러스터의 데이터 오브젝트들이 매우 유사하게 하고, 다른 클러스터들로부터의 오브젝트들이 유사하지 않게 하도록 하는 것이다.

[0081] 다중-타입 데이터 오브젝트들(214) 중 하나의 오브젝트가 다중-타입 데이터 오브젝트들(214)에서의 다른 오브젝트(들)과 컨텐츠 특성들 및 관계들 양자 모두를 갖는다고 고려하면, 이들 2개의 오브젝트들간 유사성은 이하의 식에 따라 결정된다:

수학식 1

[0082] $$S = \alpha \bullet s_f + \beta \bullet s_{intra} + \gamma \bullet s_{inter}$$

[0083] 여기서, s_f는 컨텐츠 유사성이고, s_{intra}와 s_{inter}는 각각 인트라-타입 유사성과 인터-타입 유사성이며, α, β, γ는 상이한 유사성에 대한 가중치로서, α+ β+ γ = 1이다.

[0084] 식 (1)로부터, 2개 오브젝트들간 유사성은 컨텐츠 유사성과 관계 유사성들의 선형 조합이다. α, β 및 γ에 서로 다른 값들을 할당하므로써, STS 모듈(212)은 전체 유사성에서 서로 다른 유사성들의 가중치를 조절/구성할 수 있다. 예를 들어, α= 1이고, β= γ= 0이면, 컨텐츠 특징들간 유사성이 고려된다. β= 0로 설정하므로써, STS 모듈(212)은 인트라-타입 유사성의 효과들을 회피한다.

[0085] 식 (1)에서의 유사성은 다른 함수들을 사용하여 정의될 수 있고, 일반적으로는 오브젝트들의 타입과 애플리케이션들에 의해 결정된다. 예를 들어, 2개의 웹-페이지들간 컨텐츠 유사성은 그들의 컨텐츠로부터 유도되는 2개의 키워드 벡터들의 코사인 함수 x ∈ X, yi ∈ x.Ry로서 정의될 수 있다.

[0086] 특정 오브젝트의 관계 특징은 그 엔트리가 자신과 관련된 오브젝트들에 대응하는 MDO 벡터(234)에 의해 표현된다. 일 실시예에서, 각각의 엔트리는 관계의 가중치에 대응하는 숫자 값이다. 예를 들어, 2개의 오브젝트 타입 X = {x_1, x_2, L_{xm}} 및 Y = {y_1, y_2, L_{yn}}이 주어지면, 벡터의 인터-타입 관계는 $V_x = [v_1, v_2, L, v_n]^T$이고, 여기서 $v_i \neq 0$이고, 그렇지 않으면 $v_i = 0$이다. 그리고, X에서의 2개의 오브젝트들간 인터-타입 관계 R_{xy}에 대한 유사성 $S_{inter-XY}$는 또한 2개 벡터들의 코사인 함수로서 정의될 수 있다.

[0087] Xi에서의 오브젝트들이 여러 데이터 오브젝트 타입들과 인터-타입 관계들을 가지면, 최종 인터-타입 관계 유사성은 모든 인터-타입 유사성들의 선형 조합일 수 있다.

[0088] 정의된 유사 함수들(similarity funtions)를 가지고, STS 모듈(212)은 기록 쿼리들(historical queries)(216)와 "기타 MDO들" 사이의 층내 관계/링크(intralayer relationships/link) 및 층간 링크(interlayer links)를

식별한다. 클러스터링(clustering)시의 층간 링크의 이용은 한가지 타입의 오브젝트의 클러스터링이 다른 타입의 오브젝트에 의해서 영향을 받을 수 있음을 인식한다. 예컨대, 웹 페이지 오브젝트들의 클러스터링은 사용자 오브젝트 구성, 상태 및 특성에 의해서 영향을 받을 수 있을 것이다. 따라서, 이들 얻어진 층내 및 층간 관계는 아래에 기술되는 바와 같이 상호관련된 데이터 오브젝트들의 클러스터 질을 향상시키는 데에 이용된다. 얻어진 층간 및 층내 데이터 오브젝트 관계는 각 오브젝트의 각각의 MDO 벡터(234)내에 저장된다.

[0089] 일 실시예에서 식별된 층간 링크/관계는, 예컨대 아래의 것 중 하나 이상의 사항을 나타낸다.

[0090] · 예컨대, 기록 쿼리(216) 및 이에 상응하는 (클릭을 통한)사용자 선택 웹 페이지내의 링크와 같은 컨텐츠 관련 정보.

[0091] · 예컨대, 기록 쿼리(216)와 사용자 특정 정보내의 링크에 의해서 결정되는 관련 주제에 대한 사용자 관심.

[0092] · 예컨대, 사용자 특정 정보와 선택된 웹 페이지 사이의 링크를 통해서 결정되는 선택된 웹 페이지에 대한 사용자 관심.

[0093] 일 실시예에서 식별된 층간 링크/관계(동일한 데이터 타입의 오브젝트들 사이의 관계)는, 예컨대 아래의 것 중 하나 이상의 사항을 나타낸다.

[0094] · 아래에 보다 상세히 기술되는 것과 같은 쿼리 내의 링크.

[0095] · 사용자 선택 웹 페이지내의 하이퍼링크에 지시되어 나타나는 추천 웹 페이지(들).

[0096] · 예컨대, 각각의 사용자들 사이에 식별된 관계/링크에 의해서 나타나는 인간 관계들. 일 실시예에서, 이러한 타입의 관계 정보는 사용자 프로파일의 유사성에 의해서 계산되어 얻어진다. 예컨대, 사용자 프로파일은 인구통계학, 기하학적 위치, 관심 등을 포함한다. 일 실시예에서, 사용자 프로파일은 각각의 사용자에 의해서 제공되는 웹 사이트 액세스 정보에 의해서 액세스된다.

[0097] 쿼리 내의 링크에 관하여, 쿼리 내의 링크에 의해서 지시되는 층내 관계들은 초기 기록 쿼리(216) 및/또는 (또한 각 기록 쿼리(216)에 의해서 나타나는)후속 쿼리 정제(refinement) 사이의 링크를 나타낸다. 일 실시예에서, 이러한 정보는 쿼리 로그(220)로부터 검색된 클릭을 통한 웹 페이지 정보로부터 얻어진다. 보다 구체적으로, 초기 검색 쿼리 결과가 만족스럽지 못한 것으로 판정한 경우에, 사용자가 초기 쿼리가 제출된 시간으로부터 구성가능한 시간 내에 하나 이상의 정제된 쿼리를 검색 엔진(228)에 제출할 것이 추정된다. 구성가능한 시간은 쿼리 세션을 나타낸다. 이러한 하나 이상의 검색 쿼리 용어 정제 후에, 사용자는 만족할 만한 검색 결과를 얻을 수 있을 것이다. 예컨대, 사용자가 제품 지원 웹 사이트를 방문하여 "cookie"라는 초기 쿼리를 제출하는 경우를 고려하자. 검색 결과자 만족스럽지 못한 것으로 판정한 경우에(예컨대, 너무 광범위한 경우), 사용자는 쿼리 용어를 "enable cookie"로 변경/정제 하여 보다 만족스러운 검색 결과를 얻을 수 있을 것이다.

[0098] 일 실시예에서, STS 모듈(212)은 하나 이상의 쿼리 로그(220)의 부분을 분할함으로써 쿼리 내의 링크를 식별한다. 각각의 쿼리 세션은 초기 쿼리, 하나이상의 쿼리 정제 및 가능하게는 하나 이상의 웹 페이지 클릭을 통한 지시를 포함할 수 있을 것이다. 초기 쿼리 및 하나 이상의 관련 쿼리 정제를 카테고리화하기 위하여, STS 모듈(212)은 각 쿼리 세션의 쿼리들 사이의 용어 유사성을 계산한다. 유사성의 임계 기준을 하나 이상 만족하는 검색 쿼리들은 선택되어 쿼리 및 이에 상응하는 쿼리 정제 내의 링크를 생성한다. 일 실시예에서 쿼리 유사성은, 예컨대 상기된 단락에 기술된 예시적인 연산을 이용하여 판정된다.

[0099] MDO 벡터들(234) 중 대응하는 것들에 모델링된 관계 특징으로서 다중 타입 데이터 오브젝트들(214) 사이의 관계를 맵핑한 후, 각 타입의 데이터 오브젝트들은 종래의 클러스터링 기술로(즉, 본 명세서에 개시된 강화 클러스터링 동작을 사용하지 않고) 개별적으로 클러스터링될 수 있다. 그러나 처음에는 데이터 오브젝트의 개별적 클러스터링이 가능한 것으로 보이지만, 이 기술은 실질적으로 한계가 있으며, 문제가 있다. 그에 대한 하나의 이유는 관계에 대한 특징 벡터의 크기가 매우 커서 오브젝트들의 수가 매우 커지기 때문이다. 그리고, 관련 오브젝트들의 정확한 매칭에 기초하여 관계 특징들 상에 정의된 유사성은 논-제로 엔티티들의 희소성을 갖는다. 또 다른 이유는 종래의 클러스터링 기술은 데이터 오브젝트들 사이의 관계가 데이터 오브젝트들에 할당된 특징들에 완전히 반영될 수 없고 클러스터링 프로세스 자체 동안에만 발견될 수 있다는 것을 고려하고 있지 않다는 것이다. 즉, 기존의 클러스터링 기술은 클러스터링 동작들이 후속 분석/클러스터링 동작들에서 데이터를 강화하는데 유용한 구조화된 정보를 제공할 수 있다는 점을 고려하고 있지 않다.

[0100] STS 모듈(212)은 적어도 한 데이터 오브젝트 타입의 클러스터링 결과를 그와 관련된 모든 데이터 오브젝트 타입에 이들 각각의 관계 특징들을 갱신하도록 전파함으로써 종래의 클러스터링 기술의 상기 문제/한계를 해결한다.

즉, STS 모듈(212)은 강화된 클러스터들(236)의 컨텐츠에 기초하여 표시된 데이터 오브젝트 관계(들)를 개별 다중 타입 데이터 오브젝트들(214)로 집합시킴으로써 강화된 클러스터들(236)을 생성했다. 예를 들어, 2개의 지원 노드가 클러스터링에 이어 존재하는 경우, 이 가장 가까운 2개의 지원 노드는 예를 들어 이 두 지원 노드의 벡터 값들을 평균함으로써 병합될 수 있다. 이러한 병합은 개별 노드들이 고려되어야 할 노드 수를 줄이도록 결합되는 것을 허용한다. 이로써, MDO 벡터(234)의 치수가 감소된다. 이어서, STS 모듈(212)은 MDO 벡터(234)를 클러스터링한다. 이 프로세스는 모든 오브젝트 타입들에서의 클러스터링 결과가 수렴할 때까지 반복적으로 수행된다.

[0101] 반복 클러스터링 투영 기술은 동일 타입의 오브젝트를 각각 포함하는 개별 층들에 배열된 개별 타입의 오브젝트들로부터 클러스터링 정보를 얻는 것에 의존한다. 노드 정보는 링크 정보와 함께 클러스터링이 수렴할 때까지 클러스터링된 결과들을 반복적으로 투영하고 전파하는 데 이용된다(클러스터링 알고리즘은 계층들 사이에 제공된다). 즉, 각 타입의 상이한 종류의 노드들 및 링크들은 클러스터링에 사용될 수 있는 구조적인 정보를 얻기 위해 검사된다. 예를 들어, 구조적인 정보는 상이한 데이터 오브젝트들을 연결하는 링크들의 타입을 고려하여(예를 들어, 링크가 계층간 링크인지 또는 계층 내의 링크인지에 따라) 얻어질 수 있다. 한 타입의 오브젝트의 결과들을 다른 타입의 오브젝트의 클러스터링 결과들에 반복적으로 클러스터링함으로써 데이터 희소성과 관련된 문제를 줄일 수 있다. 이러한 반복적인 투영에 의해 한 계층 클러스터링에서의 유사성 측정치가 다른 타입의 클러스터들의 개별 그룹들 대신에 클러스터들에 상에서 계산된다.

[0102] 예를 들어, 2개의 오브젝트 타입 X={x1,x2,L x_m} 및 Y={y1,y2,L y_n}에 비추어 프로세스를 설명한다. STS 모듈(212)은 먼저 임의의 종래 클러스터링 방법을 이용하여 Y 내의 오브젝트들을 {C_1, C_2, ..., Ck}로 표시되는 k 클러스터들로 클러스터링한다. x∈X의 관계 특징 벡터를 포함하는 MDO 벡터(234)가 Y 내의 하나의 오브젝트에 대응하는 각각의 성분을 가진 $Vx=[v_1, v_2,L ,v_n]^T$로서 초기에 정의된다는 것을 상기하자. Y 내의 클러스터들과 관련하여, Y의 한 클러스터에 대응하는 각각의 성분을 가진 $Vx'=[v_1',v_2',L ,v_x']^T$로 Vx를 대체하며, x.Ry∩Ci≠Φ이면, v_i'는 0이 아니다. v_i'의 수치 값은 |x.Ry∩Ci|로 설정될 수 있는데, 이 값은 오브젝트 x에서 클러스터 Ci 내의 오브젝트들로의 관계들의 수, 또는 관련 오브젝트들의 중요도와 같은 다른 값들을 나타낸다(오브젝트 중요도는 후술한다). 이어서, X 내의 오브젝트의 클러스터링은 새로운 상호 타입 관계 특징(inter-type relationship feature)에 기초한다. 프로세스는 수렴할 때까지 계층간 관계에 의해 한 타입의 클러스터링 결과들을 다른 타입으로 반복적으로 투영함으로써 계속된다.

[0103] 상술한 강화 클러스터링 알고리즘의 장점은, 클러스터링이 컨텐츠로부터의 데이터 분포를 반영할 뿐만 아니라 다른 데이터 타입과의 관계도 반영하게 된다는 점이다. 이는, 데이터 스파스니스(sparseness) 문제를 어느 정도 해결할 수도 있다. 고정된 특징 스페이스에 대한 유사성을 정의하는 현존하는 클러스터링 접근법과 대조하면, 다중-타입 데이터 오브젝트의 강화 클러스터링을 위한 상술한 시스템 및 방법은, 클러스터링 프로세스 동안에 2개 오브젝트 간의 유사성을 갱신하여, 새로 발견된 관계 특징 스페이스를 채택한다. 또한, 실시에 있어서, 임의의 전형적인 클러스터링 알고리즘을, 클러스터링 성능을 강화하는 제안된 프레임워크에 임베디드할 수 있다.

[0104] <u>오브젝트의 링크 분석 및 중요성</u>

[0105] 몇몇 데이터 오브젝트 및 애플리케이션의 경우, 동일한 타입 내 다중-타입 데이터 오브젝트(214)는, 클러스터링 프로세스 내에서 상이한 중요성을 가질 수도 있다. 통상적인 예로는, 어떤 웹 페이지가 믿을 만한 페이지일 때 더욱 중요한 웹 페이지/사용자 클러스터링, 및 몇몇 사용자가 아이템의 소속성 결정에 더욱 믿을 만해야 하는 공동 필터링을 위한 아이템/사용자 클러스터링 등을 포함한다. 오브젝트를 노드로 간주하고, 오브젝트 간의 관계를 링크로 간주한다면, HITS 알고리즘과 같은 종래의 링크 분석 방법을 사용하여, 각 데이터 오브젝트의 고유한 값을 계산한다. 그러나, 다중 타입의 데이터 오브젝트가 포함되는 경우, 상이한 타입의 오브젝트의 중요성이 비교될 수 없기 때문에, 이 방법은 도움이 되지 않을 것이다.

[0106] 이러한 문제점을 처리하기 위하여, 다중-타입 데이터 오브젝트의 강화 클러스터링을 위한 상술한 시스템 및 방법은, HITS 알고리즘을 다음과 같이 확장한다. 타입 내 오브젝트 중요성의 공통 강화뿐만 아니라 타입 간의 공통 강화도 고려한다. 각각의 노드가 *hub* 스코어 및 *authority* 스코어에 할당된다.

[0107] 설명의 간소화를 위하여, 2개 타입의 상호 관계를 갖는 오브젝트를 포함하는 경우를 예로서 사용하여, 제안된 알고리즘을 설명한다. 방향성을 고려하여, 2개 타입의 오브젝트 X={x_1,x_2,L x_m}, Y={y_1,y_2,L y_n} 및 R_X, R_Y, R_{XY}

및 R_{YX}의 관계가 주어진다. 인접한 매트릭스는 링크 정보를 표현하기 위하여 사용한다. L_X 및 L_Y는, 집합 X 및 Y 내 링크 구조의 인접한 매트릭스를 각각 의미한다. L_{XY} 및 L_{YX}는, X 내 오브젝트로부터 Y 내 오브젝트까지의 링크의 인접 매트릭스를 의미한다. 예를 들어, 노드 x_i부터 노드 y_i까지 하나의 링크가 존재하면, $L_{XY}(i,j)=1$이다.

[0108] 2개 레벨의 계산이 존재하는데, 하나는, 동일한 타입으로부터의 오브젝트의 *hub* 값 및 *authority* 값이, 인트라-타입의 관계에 의해 서로 강화되는 것이며, 다른 하나는, 상이한 타입의 노드의 *importance*가 인터-타입의 관계에 의해 서로 강화되는 것이다. 이러한 접근법에서의 계산은 다음과 같이 기록된다.

수학식 2

$$\begin{cases} a(X) = \beta L_X^T h(X) + (1-\beta) L_{XY} i(Y) \\ h(X) = \beta L_X a(X) + (1-\beta) L_{XY} i(Y) \\ i(X) = a(X) + h(X) \\ \\ a(Y) = \gamma L_Y^T h(Y) + (1-\gamma) L_{YX} i(X) \\ h(Y) = \gamma L_Y a(Y) + (1-\gamma) L_{YX} i(X) \\ i(Y) = a(Y) + h(Y) \end{cases}$$

[0109]

[0110] 여기서, a(X) 및 h(X)는 각각 X 내 노드의 *authority* 스코어 및 *hub* 스코어이다. 유사하게, a(Y) 및 h(Y)는 각각 Y 내 노드의 *authority* 스코어 및 *hub* 스코어이며, i(X) 및 i(Y)는 X 및 Y 내 노드의 *importance*를 각각 의미한다. β 및 γ는, 상이한 관계로부터 도출된 링크의 영향을 조정하는 가중치 파라미터이다.

[0111] 계산의 시작 시, 모든 벡터 a(X), h(X), a(Y) 및 h(Y)는 1로 초기화된다. *hub* 스코어 및 *authority* 스코어는, 각각의 반복 시에 수학식 2를 사용하여 갱신한다. 각각의 반복의 종료 시, 벡터는 그 다음 반복 계산을 위하여 정규화된다. 이러한 알고리즘은 정규화되고 균일한 중요성을 각각의 오브젝트 타입에 제공하며, 다른 타입의 관련 오브젝트의 중요성을 인터-타입의 관계를 통해 고려함으로써 좀더 타당한 결과를 얻는다.

[0112] 오브젝트의 중요성 스코어가 주어진 경우, 상술한 강화 클러스터링 프로세스는 오브젝트의 중요성을 반영하도록 변경된다. 이러한 실시에 있어서, k-평균 클러스터링 알고리즘은 가중화된 k-평균 알로리즘으로 변경된다. 즉, 클러스터 센트로이드(centroid)를 계산할 때, 클러스터 멤버의 가중화된 합을 새로운 센트로이드로서 사용하여, 클러스터는 그러한 중요한 오브젝트에 바이어스된다.

[0113] 상술한 관점에서, STS 모듈(212)은 인터 및 인트라 타입의 관계 모두에 기초한 다중-타입 데이터 오브젝트의 중요성을 다중-타입 데이터 오브젝트(214) 중에서 차별화한다. 이러한 중요성은 클러스터링 프로세스에 포함된다.

[0114] 비드 용어(term)의 예시적인 프로세싱

[0115] 엔드-사용자(예컨대, 광고주, 웹 사이트 프로모터 등)로부터 용어(208)를 수신한 다음, STS 모듈(212)은 용어(208)를, 강화 용어 클러스터(236) 내 용어/구의 각각의 용어와 비교한다. 강화 용어 클러스터(236)은, 서로 문맥적으로 관련있을 뿐만 아니라, 상호 관계로부터 웹 페이지 및 사용자에게서 도출되어 서로 의미적으로도 관련있는 용어를 포함하기 때문에, 용어(208)는 복수의 관련있는 역사적인 문맥, 또는 "센스"의 관점에서 평가된다.

[0116] 일 실시예에서, STS 모듈(212)이, 용어(들)(208)가 보강된 클러스터(236)로부터의 용어(들)와 매칭한다고 결정하면, 검색 용어 제시 모듈(212)은 보강된 클러스터(236)로부터 제시되는 용어 리스트(210)를 생성한다. 이 실시예에서, 매칭은 정확한 것일수도 있고, 단수/복수 형태, 철자오류, 구두점, 등과 같은 작은 수의 변동들을 포함하는 것일 수도 있다. 반환된 리스트는 FOO와 신뢰값의 조합에 의해 정렬된다.

[0117] 일 실시예에서, 용어(들)가 클러스터로부터의 용어와 매칭한다면, 클러스터는 제시된 용어 리스트의 말단 사용자에게 반환된다. 제시된 용어 리스트(210)는, 용어(들)에 대해 의미론적으로 및/또는 문맥적으로 관련되어 결정된 용어/문구, 각 용어(들) 대 용어(들)에 대한 유사성 측정치(신뢰값), 및 각 용어(들)의 발생 빈도(FOO)를 포함한다. 반환된 리스트(210)는 FOO와 신뢰값의 조합에 의해 정렬된다.

[0118] STS 모듈(212)이, 용어(들)(208)이 다수의 보강된 용어 클러스터들(236)의 용어들과 매칭된다고 결정한다면, 검색 용어 제시 모듈(212)은, 보강된 용어 클러스터들(236) 중 다수의 보강된 클러스터의 용어들로부터 다수의 제시된 용어 리스트(210)를 생성한다. 이 리스트는 클러스터의 크기에 의해 정렬되고, 각 리스트내의 용어는 FOO와 신뢰값의 조합에 따라서 정렬된다.

[0119] 매칭되는 클러스터가 식별되지 않는다면, 쿼리 용어는, 저 FOO의 쿼리 용어로부터 생성된 확장 클러스터에 대해서 더 매칭된다. 일 실시예에서, 저 FOO의 쿼리 용어는, 역사적 쿼리 로그 용어가 높은 빈도로 발생함으로써 생성되는 보강된 용어 클러스터(236)에 대해 분류자(예를 들어, K-최근접 이웃 분류자)를 훈련함으로써 클러스터화된다. 낮은 빈도로 발생하는 것으로 결정된 역사적 쿼리 용어가 검색 엔진에 하나씩 제출된다. 그 다음, 반환된 검색 결과 중에 선택된 것(예를 들어, 처음 탑-랭크된 웹 페이지, 및/또는 기타 등등)으로부터 특성들이 추출된다. 추출된 특성들은 표준화되어, 저 FOO의 쿼리 용어를 나타내는데 사용된다. 그 다음, 쿼리 용어는 기존의 클러스터들로 분류되어, 훈련된 분류자에 기초해서 확장된 클러스터를 생성한다. 다음으로, 말단 사용자가 제출한 용어(들)는 이러한 확장된 클러스터들의 관점에서 평가되어, 제시된 용어 리스트를 식별하고 말단 사용자에게 반환된다.

[0120] 저 FOO 용어의 분류

[0121] 높은 발생 빈도의(FOO) 쿼리 용어들(222)로부터 생성된 보강된 용어 클러스터들(236)이 말단 사용자 입력 용어(들)(208)에 대해 동일한 용어를 포함하지 않는다면, STS 모듈(212)은, 높은 발생 빈도의(FOO) 쿼리 로그 용어(222)로부터 생성된 보강된 용어 클러스터들(236)로부터 훈련된 분류자(238)를 생성한다. 보강된 용어 클러스터들(236)의 용어는, 분류 동작에 적합한 벡터 공간 모델에서 상응하는 키워드 벡터를 이미 포함하고 있다. 추가적으로, 스톱-단어 삭제 및 단어 제거(접미사 삭제)는 (어느 클러스터(236)에 기초하는지에 따라서) 용어 벡터(234)의 차원성(dimensionality)을 감소시킨다. 일 실시예에서, 추가의 차원성 축소 기술, 예를 들어, 특성 선택 또는 재-파라미터화 기술이 사용될 수 있다.

[0122] 이 실시예에서, 계급-불명의 쿼리 용어(222)를 분류하기 위해서, STS 모듈(212)은, 상응하는 특성 벡터에 따라서, 계급-공지된 모든 쿼리 용어들(222)에서 k와 가장 유사한 이웃들을 찾도록 k-최근접 이웃 분류자 알고리즘을 사용하고, 새로운 쿼리 용어의 계급을 예상하기 위해서, 이웃들의 가중치가 부가된 대부분의 계급 라벨들을 사용한다. 여기에서 보강된 용어 클러스터들(236)에 이미 존재하는 각 쿼리 용어에는, 상응하는 클러스터의 라벨과 동일한 라벨이 할당되지만, 각각의 보강된 클러스터(236)는 단순한 시퀀스 넘버에 의해서 표시된다. 이들 이웃들은 X에 대해 각 이웃들의 유사성을 이용해서 가중치가 부가되고, 여기에서 유사성은 두 벡터 사이의 유클리드 기하학적 거리 또는 코사인 값에 의해 측정된다. 코사인 유사성은 다음과 같다:

$$sim(X, D_j) = \frac{\sum_{t_i \in (x1\ D_j)} x_i \cdot d_{ij}}{\|X\|_2 \cdot \|D_j\|_2}$$

[0123]

[0124] 여기에서, X는 테스트 용어, 즉, 벡터로서 표현되는 분류될 쿼리 용어이고, D_j는 j번째 훈련 용어이고, t_i는 X와 D_j에 의해 공유되는 단어이고, x_i는 X에서 키워드 t_i의 가중치이고, d_{ij}는 D_j에서 키워드 t_i의 가중치이고, $\|X\|_2 = \sqrt{x_1^2 + x_2^2 + x_3^2}$ 는 X의 평균이고, $\|D_j\|_2$ 은 D_j의 평균이다. 따라서, 테스트 용어 X의 계급 라벨은 모든 이웃들의 가중치가 부가된 대부분의 계급 라벨이다:

$$label(X) = \arg\max_{l_i} \left(\sum_{\text{All } D_j \text{ where } lable(D_j)=l_i} sim(X, D_i) \right)$$

[0125]

[0126] 또 다른 실시예에서, 최근접-이웃 분류 기술 이외의, 여러가지 통계적 분류 및 기계 학습 기술(예를 들어, 복귀 모델, 베이스(Bayesian) 분류자, 결정 트리, 신경 네트워크, 및 지원 벡터 기계)이 훈련된 분류자(238)를 생성하는데 사용될 수 있다.

[0127] STS 모듈(212)은 (각각의 검색 쿼리(227)를 통해) 하나씩 낮은 발생 빈도(frequency of occurrence; FOO) 쿼리어(224)를 검색 엔진(228)에 제출한다. 특정 검색 쿼리(227)와 관계되는 검색 결과(들)(230)을 수신하는 것에 응답하고, 전술한 기술을 사용하여, STS 모듈(212)은 검색 결과(들)(230)에 의해 식별되는 하나 이상의 검색된

검색 결과(230)로부터 피쳐를 추출한다(추출된 피쳐(232)). 본 구현에서, 피쳐는 제1의 최상위 랭크된 검색 결과(들)(230)로부터 추출된다. 각각의 검색되고 구문해석된(parsed) 검색 결과(들)(230)에 대하여, STS 모듈(212)는 추출된 피쳐(232)의 각 레코드에 다음의 정보를 저장한다; URL, 결과 타이틀, 쿼리어의 간단한 설명 및/또는 컨텍스트, 및 검색된 검색 결과(들)(230)을 얻기 위해 사용되는 검색 쿼리(227). 다음에, STS 모듈(212)은 낮은 FOO 쿼리어(224)로부터 도출된 추출 피쳐(232)를 토큰화, 디멘젼 감축, 및 정규화한다. 다음에, STS 모듈(212)은 쿼리어를 용어 클러스터(236)의 개별 세트로 클러스터링한다. 이러한 클러스터링 동작은 (높은 FOO 쿼리어(222)로부터 생성된) 훈련된 분류기(238)을 사용하여 수행된다.

[0128] STS 모듈(212)은 (낮은 FOO 쿼리어(224)에 기초하여 생성된) 이들 확장 용어 클러스터의 관점에서 엔드-유저 제출 용어(들)(208)를 평가하여 하나 이상의 제안된 용어 리스트(210)를 식별하여 엔드-유저에게 리턴한다. 그러한 절차의 예가 단락 [0063] 및 [0066]에 전술되어 있고 다음의 섹션에서 설명된다.

[0129] 예시적 검색 용어 제안 리스트

[0130] 제안된 용어 리스트(210)는, 예를 들면, 용어(들)(208)에 관련되도록 결정된 용어(들), 각각의 용어(들) 대 용어(들)(208) 유사도 측정값(비밀값), 및 각각의 용어(들) 발생 빈도(FOO) - 히스토리 쿼리 로그에서의 빈도를 포함한다. 관련 용어를 식별하고, 유사성 측정값을 생성하고 FOO 값을 생성하는 기술은 전술하였다.

[0131] 표 1은 "메일"의 용어(들)(208)에 관련되도록 결정된 용어의 예시적 제안 용어 리스트(210)를 나타낸다. 용어(들)(208) 관련 용어는 본 예에서 제목이 "제안된 용어"인 컬럼 1에 도시되어 있다.

표 1

[0132] 비드(bid) 용어 "MAIL"용 예시적 제안 용어 리스트

제안 용어	유사도	빈도	<컨텍스트>
hotmail	0.246142	93161	온라인 이메일 관련
yahoo	0.0719463	165722	
mail.com	0.352664	1455	
yahoo mail	0.0720606	39376	
www.mail.com	0.35367	711	
email.com	0.484197	225	
www.hot	0.186565	1579	
www.msn.com	0.189117	1069	
mail.yahoo.com	0.0962268	4481	
free email	0.230611	1189	
www.aolmail.com	0.150844	654	
check mail	0.221989	66	
check email	0.184565	59	
msn passport	0.12222	55	
www.webmail.aol.com	0.0200538	108	
webmail.yahoo.com	0.08789	71	
free email account	0.0234481	65	
제안 용어	유사도	빈도	
mail	1	2191	전통적인 메일 관련
usps	0.205141	4316	
usps.com	0.173754	779	
united parcel service	0.120837	941	
postal rates	0.250423	76	
stamps	0.156702	202	
stamp collecting	0.143618	152	
state abbreviations	0.104614	300	
postal	0.185255	66	
postage	0.180112	55	
postage rates	0.172722	51	
usps zip codes	0.138821	78	
us postmaster	0.109844	58	

[0133] 표 1을 참조하면, 제안된 용어 리스트에 있는 용어들은 용어 유사도 값("유사도" 제목의 컬럼 1 참조) 및 발생 빈도 점수("빈도" 제목의 컬럼 3 참조)로 매핑된다. 부제가 "용어 클러스터링"로 붙은 하기에 설명되는 바와 같이 계산된 각각의 용어 유사도 값은 대응하는 제안 용어(컬럼 1)과 용어(들)(208) 간의 유사도 측정값을 제공하며, 본 예에서는 "mail"을 예로 들었다. 각각의 빈도값, 또는 점수는 제안된 용어가 히스토리 쿼리 로그에서 발생하는 횟수를 나타낸다. 제안된 용어 리스트는 용어 유사도의 함수로서 분류되고, 및/또는 발생 빈도는 사업 목표의 함수로서 계산한다.

[0134] 임의의 주어진 용어(들)(208)(예를 들면, 메일 등)는 비드 용어가 사용될 수 있는 싱글 이상의 컨텍스트를 가질 수 있다. 이것을 설명하기 위해, STS 모듈(212)은 어느 제안된 용어가 용어(들)(208)의 다수의 컨텍스트 중 어느 것에 대응하는지 제안된 용어 리스트(210)에 지시를 제공한다. 예를 들면, 표 1을 참조하면, "mail"의 용어(들)(208)는 두개의 (2) 컨텍스트: (1) 전통적인 오프라인 메일 및 (2) 온라인 이메일를 갖는다. 관련 용어의 각 리스트는 이들 두개의 비드 용어 컨텍스트에 대하여 도시되어 있다.

[0135] 또한, 임의의 용어(들)(208)에 대한 제한된 용어는 비드 용어의 동의어 이상일 수 있다. 예를 들면, 표 1을 참조하면, 제안된 용어 "usps"는 메일을 핸들링하는 조직에 대한 두문자어이지 비드 용어 "mail"에 대한 동의어는 아니다. 그러나, "usps" 또한 "mail" 비드 용어에 매우 관련이 있는 용어이고, 따라서 제안된 용어 리스트(210)에 도시되어 있다. 일 구현예에서, STS 모듈(212)은 관련 용어 R(예를 들면, "usps")과 목표 용어 T(예를 들면, "mail") 간의 관계를 다음의 관계 규칙의 함수로서 결정한다: itr(T) -> itr(R), 여기에서, "itr"은 "interested in"을 나타낸다. 유저(광고업자, 웹 사이트 프로모터 등)가 R에 관심이 있다면, 그 유저는 T에 또한 관심이 있다.

[0136] 예시적인 프로시저

[0137] 도 3은 검색 용어 제안용 다중타입 데이터 오브젝트의 강화된 클러스터링을 위한 예시적인 프로시저(300)를 나타낸다. 설명의 편의상, 이 프로시저의 동작은 도 2의 특징과 관련하여 설명한다 (모든 참조 번호는 먼저 컴포넌트가 도입되는 도면의 번호로 시작된다). 블록(302)에서, 검색 용어 제안(STS) 모듈(212; 도 2)은 쿼리 로그(220)로부터 히스토리 쿼리 용어(216)를 집합한다. STS 모듈(212)은 히스토리 쿼리 용어(216)를 발생 빈도 함수로서 조직한다. 블록(304)에서, STS 모듈(212)은 발생 쿼리 용어(222)의 고 빈도를 전송하여 엔진(228)을 검색하고 대응하는 검색 결과(230)를 수신한다. 블록(306)에서, STS 모듈(212)은 각 검색 결과(230)로부터 단편적 설명을 추출하고 단편적 설명들을 함께 병합하여 용어 기반 MDO 벡터(234)를 형성하게 된다. 각 용어 벡터는 발생 쿼리 용어(222)의 고 빈도 각각에 대하여 발생한다.

[0138] 블록(308)에서, STS 모듈(212)은, 예를 들어, 쿼리 로그(220), 웹사이트 사용자 액세스 정보 등으로부터 다른 MDO(218)를 찾아낸다. STS 모듈(212)은 각 MDO 벡터(234)를 발생하여 찾아낸 다른 MDO(218)의 특징 공간을 나타낸다. 블록(310)에서, STS 모듈(212)은 인트라 오브젝트 및 인터 오브젝트 관계/MDO(214)의 각각 간의 링크를 식별한다. 블록(312)에서, STS 모듈(212)은 자신의 각각의 MDO 벡터(234)에 기초하여 MDO(214)의 강화된 클러스터링을 수행하여 강화된 클러스터(236)를 발생한다. 블록(312)의 이종 데이터 오브젝트의 강화된 클러스터링의 상세는 도 5를 참조하여 후술된다. 프로시저(300)는 온 페이지 참조 "A"에 의해 도시된 바와 같이 도 4의 블록(402)에서 계속된다.

[0139] 도 4는 검색 용어 제안용 다중타입 데이터 오브젝트의 강하된 클러스터링을 위한 도 3의 예시적인 프로시저(300)의 연속을 나타낸다. 설명의 편의상, 이 프로시저의 동작은 도 2의 특징과 관련하여 설명한다. 블록(402)에서, 그리고 최종 사용자로부터 (도 2) 수신 용어(208)에 응답하여, STS 모듈(212)은, 용어(208)에 관련되며 이에 상당히 유사한 것으로 결정된 강화 용어 클러스터링(236)으로부터의 임의의 용어로부터 제안된 용어 리스트(210)를 발생한다. 상이한 오브젝트 타입들 간의 상호 관계를 이용하여 클러스터링을 개선한다. 블록(404)에서, STS 모듈(212)은, 키워드 클러스터(236)로부터의 임의의 용어가 용어(208)에 관련되는지/상당히 유사한지 여부를 결정한다. 만약 그렇다면, 프로시저는 블록(406)에서 계속되며, 여기서 STS 모듈(212)은 대응하는 제안된 용어 리스트(210)를 최종 사용자에게 전송한다. 그렇지 않다면, 프로시저는 온 페이지 참조 "B"에 의해 도시된 바와 같이 도 5의 블록(502)에서 계속된다.

[0140] 도 5는 검색 용어 제안용 다중타입 데이터 오브젝트의 강화된 클러스터링을 위한 도 3과 4의 예시적인 프로시저(300)의 연속을 나타낸다. 설명의 편의상, 이 프로시저의 상세는 도 2의 특징과 관련하여 설명한다 (모든 참조

번호는 먼저 구성 요소가 도입되는 도면의 번호로 시작된다). 블록(502)에서, STS 모듈(212)는 강화된 용어 클러스터(236)로부터 분류자(훈련된 분류자)를 발생하며, 이 때 그 강화된 용어 클러스터는 쿼리 용어(222) 발생의 고 빈도에 기초한다. 블록(504)에서, STS 모듈(212)은 쿼리 용어(224) 발생의 저 빈도를 하나씩 전송하여 엔진(228)을 검색하고 대응하는 검색 결과(230)를 수신한다. 블록(506)에서, STS 모듈(212)은 검색 결과(230)로부터 단편적 설명(추출된 특징(232))을 추출하고, 이로부터 용어 벡터(234)를 발생한다.

[0141] 블록(802)에서, STS 모듈(212)은 트레이닝된 분류자9238)에서 발생 질의 용어(224)의 저 빈도로부터 발생하는 용어 벡터(234)를 분류하여 발생 질의 용어(224)의 저 빈도에 기초하여 각각의 강화된 용어 클러스터(236)를 발생한다. 블록(510)에서, STS 모듈(212)은, 용어(들)(208)에 상당히 유사한 것으로 결정된 발생 질의 용어(224)의 저 빈도에 기초하여 강화된 용어 클러스터(236)로부터의 키워드/키 구로부터 제안된 용어 리스트(210)를 발생한다. 블록(512)에서, STS 모듈(212)은 제안된 용어 리스트(210)를 최종 사용자에게 전송한다.

[0142] 도 6은 도 3의 블록(312)의 강화된 클러스터링 동작의 예시적인 상세를 나타낸다. 설명의 편의상, 블록(310)의 동작은 도 1과 2의 특징과 관련하여 설명한다. STS 모듈(212)에 의해 구현되는 강화된 클러스터링 알고리즘으로의 입력은, 식별되고 가중된 인터 오브젝트 및 인트라 오브젝트 관계를 포함하여, 대응하는 노드의 컨텐츠 특징(fi 및 gj)을 포함하는 그래프(100)와 같은 다중층 프레임워크 그래프를 포함한다. 이 클러스터링 알고리즘의 출력은, 다중타입 데이터 오브젝트의 강화된 클러스터링을 반영하는 새로운 프레임워크 그래프(100)를 포함한다. 이 새로운 프레임워크 그래프의 소정의 구현예에서, 그래프(100) 차원성을 줄이도록 새로운 노드 위치로 변경되고 그리고/또는 다른 노드와 병합된 구(old) 노드 각각의 변동을 나타낼 수 있다.

[0143] 블록(602)에서, (각 클러스터링 반복 전에) 원 프레임워크 그래프가 입력된다. 블록(604)에서, 클러스터링되는 각 노드의 중요성은 수학식(2)을 이용하여 계산되거나 결정된다. 블록(606)에서, 클러스터링을 위해 임의층을 선택한다. 블록(608)에서, 선택된 층에서의 노드는 적절한 방식으로 (예를 들어, 컨텐츠 특징에 따라) 클러스터링되어 강화된 클러스터(236)를 발생한다. 소정의 구현예에서, 이 노드는 클러스터링을 개선하기 위해 (도시하지 않은) 원하는 필터링 알고리즘을 이용하여 필터링될 수 있다. 블록(610)에서, 각 클러스터의 노드는 하나의 노드로 병합된다. 예를 들어, 2개의 후보 노드가 필터링에 이어 존재하면, 예를 들어 그 2개 후보 노드의 벡터 값을 평균화함으로써 가장 근접하는 2개의 후보 노드를 병합할 수 있다. 이러한 병합에 의해 노드가 개별적으로 결합되어 고려해야 하는 노드 수를 줄일 수 있다. 이처럼, 병합 동작을 이용하여 중복 및 근사 중복의 발생을 줄일 수 있다. 블록(612)에서, 대응하는 링크가 병합(mergining in; 610)에 기초하여 갱신된다. 블록(614)에서, 클러스터링 알고리즘은 클러스터링을 위해 (임의의 선택한 층으로부터) 제2 층으로 전환된다. 블록(312)의 동작은 온 페이지 참조 "C"에 의해 도시된 바와 같이 도 7의 블록(702)에서 계속된다.

[0144] 도 6의 연산(operation)을 참조하면, 초기 클러스터링 패스(initial clustering pass)에서는, 컨텐츠의 특징(content features)들만이 이용된다. 대부분의 경우, 링크 특징(link feature)은 클러스터링에 이용되기에는 시작부에서 너무 빈약(sparse)하다. 후속하는 클러스터링 패스에서는 도 7을 참조하여 아래에서 설명하고 있는 바와 같이, 컨텐츠의 특징과 링크의 특징들은 조합되어 클러스터링을 효과적으로 강화시킨다. 컨텐츠의 특징과 링크의 특징들을 조합함으로써, 서로 상이한 값들로 가중치(weights)들이 특정되고, 그 결과치가 비교될 수 있어 정확도가 개선된 클러스터링이 제공될 수 있다.

[0145] 도 7은 도 3과 도 6의 블록 312의 강화된 클러스터링 연산의 전형적 연속성을 예시하고 있다. 블록 702에서, 제2 층의 노드들은 그 컨텐츠의 특징과 업데이트된 링크의 특징에 따라 클러스트화된다. 블록 704에서, 각각의 클러스터의 노드들은 하나의 노드로 병합된다. 블록 706에서, 다른층의 초기 링크 구조와 초기 노드들은 복원된다. 블록 708에서, 제2층의 각 클러스터의 노드들은 병합되며, 대응하는 링크들이 갱신된다. 블록 710에서, 이러한 반복적 클러스터링 처리는 컴퓨터 환경내에서 지속된다. 블록 702에서, 프레임 그래프의 교정 버전(revised version)이 출력된다.

[0146] 전형적인 운용환경

[0147] 도 8은 적절한 컴퓨팅 환경(800)의 예를 예시하고 있으며, 검색 용어의 제안을 위한 다중형 데이터 오브젝트들의 강화된 클러스터링을 위해 도 3 내지 도 6의 방법론과 도 2의 시스템이 완전히 혹은 부분적으로 구현될 수도 있다. 예시적인 컴퓨팅 환경(800)은 적절한 컴퓨팅 환경 중 하나의 예일 뿐이며, 본 명세서에서 설명되고 있는 시스템 및 방법론의 기능 혹은 사용 범위에 대해 어떠한 제한을 가하고자하는 의도는 없는 것이다. 컴퓨팅 환경(800)에서 예시되고 있는 요소(components)들 중 어느 하나 혹은 조합에 관하여 어떠한 의존성이나 요구사항(requirements)을 갖는 것으로 컴퓨팅 환경(800)이 해석되어서는 안될 것이다.

[0148] 본 명세서에서 설명하고 있는 방법과 시스템들은 수많은 기타 범용의 목적으로 혹은 특수한 목적의 컴퓨팅 시스템 환경 혹은 구성으로 조작 가능하다. 공지의 컴퓨팅 시스템, 환경 및/또는 사용에 적합한 구조들은 국한되는 것은 아니지만, 개인용 컴퓨터, 서버 컴퓨터, 다중프로세서 시스템, 마이크로프로세서기반 시스템, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 이들 시스템 혹은 장치 등을 포함하는 분산형 컴퓨팅 환경(distributed computing environment)을 포함하고 있다. 프레임 워크의 콤팩트 혹은 서브셋 버전들은 휴대용 컴퓨터 혹은 기타 컴퓨팅 장치 등과 같은 제한된 리소스의 클라이언트에서 구현될 수도 있다. 본 발명은 분산형 컴퓨팅 환경에서 실행되며, 이 경우, 태스크들은 통신망을 통해 링크되는 원격 처리 장치들에 의해 수행된다. 분산형 컴퓨팅 환경에서, 프로그램 모듈들은 로컬 및 원격 메모리 기억장치(storage devices) 모두에 위치할 수도 있다.

[0149] 도 8을 참조하면, 검색 용어 제안을 위한 다중형 데이터 오브젝트들의 강화된 클러스터링의 전형적인 시스템은 컴퓨터(810)의 형태로 범용 컴퓨팅 장치를 포함하고 있다. 다음에 설명되고 있는 컴퓨터(810)의 양상들은 클라이언트 컴퓨팅 장치 PSS 서버(202) (도 2) 및/또는 클라이언트 컴퓨팅 장치(206)의 전형적인 구현예이다. 컴퓨터(810)의 구성요소(components)는 국한되는 것은 아니지만, 처리유닛(들)(820), 시스템 메모리(830), 및 이 시스템 메모리를 포함한 각종 시스템 구성요소들을 처리부(820)에 연결하는 시스템 버스(821)를 포함할 수도 있다. 시스템 버스(821)는 메모리 버스 혹은 메모리 콘트롤러, 주변 버스, 그리고 다양한 버스 구조 중 어느 하나의 버스를 이용하는 로컬 버스를 포함한 여러 버스 구조 중의 어느 하나일 수도 있다. 한 예로서 그리고 제한하고자 하는 것은 아니지만, 그러한 구조들은 ISA(Industry Standard Architecture) 버스, MCA(Micro Channel Architecture) 버스, EISA(Enhanced) 버스, VESA(Video Electronics Standards Association) 로컬 버스, 그리고 중간 버스(Mezzanine bus)로도 알려져 있는 PCI(Peripheral Component Internet) 버스를 포함할 수도 있다.

[0150] 컴퓨터(810)는 통상 다양한 컴퓨터 판독가능 매체를 포함하고 있다. 컴퓨터 판독 가능 매체는 컴퓨터(810)에 의해 액세스될 수 있는 어떠한 이용가능한 매체도 가능하며, 휘발성 및 불휘발성 매체, 착탈식 및 비착탈식 매체 모두를 포함한다. 한 예로서 그리고 국한시키고자 하는 것은 아니지만, 컴퓨터 판독 가능 매체는 컴퓨터 스토리지 매체 그리고 통신 매체를 포함할 수도 있다. 컴퓨터 스토리지 매체는 컴퓨터 판독 가능 명령, 데이터 구조, 프로그램 모듈 혹은 기타 데이터 등과 같은 정보의 저장을 위한 어느 방법 혹은 기술로 구현되는 휘발성 및 비휘발성, 착탈식 및 비착탈식 매체를 포함하고 있다. 컴퓨터 스토리지 매체는 국한되는 것은 아니지만, RAM, ROM, EEPROM, 플래시 메모리 혹은 기타 메모리 기술, CD-ROM, DVD(Digital Versatile Disks) 혹은 기타 광디스크 스토리지, 자기 카셋트, 자기 테이프, 자기 디스크 스토리지 혹은 기타 자기 스토리지 장치, 혹은 원하는 정보를 저장하는데 사용될 수 있으며 컴퓨터에 의해 액세스될 수 있는 기타 어떠한 매체도 포함하는 것이다.

[0151] 통신 매체는 통상적으로 반송파 또는 기타 전송 메카니즘 등의 변조된 데이터 신호에 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈, 또는 다른 데이터를 구현하며, 임의의 정보 전달 매체를 포함한다. "변조된 데이터 신호"라는 용어는 신호 내에 정보를 인코딩하도록 설정되거나 변환된 특성을 하나 또는 그 이상을 갖는 신호를 의미한다. 예로서, 통신 매체는 유선 네트워크 또는 직접 유선 접속 등의 유선 매체와, 음향, RF, 적외선 및 기타 무선 매체 등의 무선 매체를 포함하지만, 이에 한정되지 않는다. 상술한 것들 중의의 임의의 조합이 컴퓨터 판독가능 매체의 범위 내에 포함되어야 한다.

[0152] 시스템 메모리(830)는 리드 온리 메모리(ROM)(831) 및 랜덤 액세스 메모리(RAM)(832)와 같은 휘발성 및/또는 비휘발성 메모리의 형식의 컴퓨터 저장 매체를 포함한다. 개시(스타트업) 동안과 같이, 컴퓨터(810) 내에서 구성소자 사이의 정보 전달을 돕는 기본 루틴을 포함하는 기본 입출력 시스템(833)(BIOS)은 통상적으로 ROM(831)에 저장된다. RAM(832)은 통상적으로 즉시 액세스가능하고 및/또는 프로세싱 유닛(820)에 의해 현재 동작되고 있는 데이터 및/또는 프로그램 모듈을 포함한다. 예를 들어, 도 8은 오퍼레이팅 시스템(834), 어플리케이션 프로그램(835), 다른 프로그램 모듈(836), 및 프로그램 데이터(838)를 나타내지만, 이에 한정되지 않는다. 일 실시예에서, 컴퓨터(810)는 PSS 서버(202)이다. 이 시나리오에서, 어플리케이션 프로그램(835)은 검색 용어 제안 모델(212)을 포함한다. 이 동일한 시나리오에서, 프로그램 데이터(838)는 다중타입 데이터 오브젝트(214), 검색 결과(230), 추출 특징(232), MDO 벡터(234), 강화 클러스터(236), 훈련된 분류기(trained classifier)(238), 및 다른 데이터(226)를 포함한다.

[0153] 또한 컴퓨터(810)는 다른 분리형/비분리형, 휘발설/비휘발성 컴퓨터 저장 매체를 포함한다. 예를 들어, 도 8은 비분리형 비휘발성 마그네틱 매체로부터 판독하거나 기록하는 하드 디스크 드라이브(841), 분리형 비휘발성 마그네틱 디스크(852)로부터 판독하고 기록하는 마그네틱 디스크 드라이브(851), 및 CD ROM 또는 다른 광 매체와 같이 분리형 비휘발성 광 디스크(856)로부터 판독하고 기록하는 광 디스크 드라이브(855)를 나타내고 있지만,

이에 한정되지 않는다. 오퍼레이팅 환경의 예로서 사용될 수 있는 다른 분리형/비분리형, 휘발성/비휘발성 컴퓨터 저장 매체는 마그네틱 테잎 카세트, 플래시 메모리 카드, 디지털 다목적 디스크, 디지털 비디오 테잎, 고체 상태 RAM, 고체 상태 ROM 등을 포함하지만, 이에 한정되지 않는다. 하드 디스크 드라이브(841)는 통상적으로 인터페이스(840)와 같은 비분리형 메모리 인터페이스를 통해 시스템 버스(821)에 접속되고, 마그네틱 디스크 드라이브(851) 및 광 디스크 드라이브(855)는 통상적으로 인터페이스(850)와 같은 분리형 메모리 인터페이스에 의해 시스템 버스(821)에 접속된다.

[0154] 상술하고 도 8에 나타낸 드라이브 및 그들의 관련 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 및 컴퓨터(810)에 대한 다른 데이터를 저장하도록 제공된다. 도 8에서, 예를 들어, 하드 디스크 드라이브(841)는 오퍼레이팅 시스템(844), 어플리케이션 프로그램(845), 다른 프로그램 모듈(846) 및 프로그램 데이터(848)를 저장하는 것으로 도시된다. 이들 컴포넌트는 오퍼레이팅 시스템(834), 어플리케이션 프로그램(835), 다른 프로그램 모듈(836), 및 프로그램 데이터(838)와 동일하거나 상이할 수 있다는 것에 주목하여야 한다. 오퍼레이팅 시스템(844), 어플리케이션 프로그램(845), 다른 프로그램 모듈(846), 및 프로그램 데이터(848)에는 최소한 다른 복사물을 나타내기 위하여 다른 참조번호를 부여한다.

[0155] 사용자는 일반적으로 마우스, 트랙볼, 또는 터치 패드라 불리우는 포인팅 장치(861) 및 키보드(862)와 같은 입력 장치를 통해 컴퓨터(810)에 명령 및 정보를 입력할 수 있다. (도시되지 않은) 기타 입력 장치는 마이크로폰, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 입력 장치 및 그외의 입력 장치는 시스템 버스(821)에 연결된 사용자 입력 인터페이스(860)를 통해 종종 프로세싱 유닛(820)에 접속되지만, 병렬 포트, 게임 포트 또는 유니버설 시리얼 포트(USB)와 같은 기타 인터페이스 및 버스 구조에 의해 접속될 수 있다.

[0156] 모니터(891) 또는 다른 유형의 디스플레이 장치는 또한 비디오 인터페이스(890) 등의 인터페이스를 통해 시스템 버스(821)에 접속된다. 모니터 이외에도, 컴퓨터는 또한 출력 주변 인터페이스(895)를 통해 접속될 수 있는 스피커(898) 및 프린터(896) 등의 기타 주변 출력 장치를 포함할 수 있다.

[0157] 컴퓨터(810)는 원격 컴퓨터(880)와 같은 하나 이상의 원격 컴퓨터로의 논리적 접속을 이용한 네트워크 환경에서 동작할 수 있다. 원격 컴퓨터(880)는 퍼스널 컴퓨터, 서버, 라우터, 네트워크 PC, 피어(peer) 장치, 또는 기타 공통 네트워크 노드일 수 있으며, 비록 도 8에는 메모리 저장 장치(881)만이 도시되어 있지만, 그 특별한 실시예의 기능으로서 컴퓨터(810)에 관하여 상술한 구성요소 중 다수 또는 모든 구성요소를 포함할 수 있다. 도 8에 도시된 논리적 접속은 근거리 통신망(LAN)(881) 및 원거리 통신망(WAN)(883)을 포함하지만, 그 외의 네트워크를 포함할 수도 있다. 이러한 네트워크 환경은 사무실, 기업 광역 컴퓨터 네트워크(enterprise-wide computer network), 인트라넷, 및 인터넷에서 일반적인 것이다.

[0158] LAN 네트워크 환경에서 사용된 경우, 컴퓨터(810)는 네트워크 인터페이스 또는 어댑터(880)를 통해 LAN(881)에 접속된다. WAN 네트워크 환경에서 사용된 경우, 컴퓨터(810)는 일반적으로 인터넷과 같은 WAN(883)을 통한 통신을 확립하기 위한 모뎀(882) 또는 기타 수단을 포함한다. 모뎀(882)은 내부 또는 외부에 있을 수 있으며, 사용자 입력 인터페이스(860) 또는 기타 적절한 메커니즘을 통해 시스템 버스(821)에 접속된다. 네트워크 환경에서, 컴퓨터(810)와 관련하여 도시된 프로그램 모듈 또는 그 일부는 원격 메모리 스토리지 장치에 저장될 수 있다. 이것에 국한되는 것은 아니지만, 예를 들어 도 8에서는 원격 애플리케이션 프로그램(885)이 메모리 장치(881)에 위치하는 것으로 예시되어 있다. 도시된 네트워크 접속은 예시를 위한 것이며 컴퓨터들 사이의 통신 링크를 확립하기 위한 다른 수단이 사용될 수도 있다.

[0159] **결론**

[0160] 검색 용어 제안을 위한 다중타입 데이터 오브젝트의 강화된 클러스터링을 위한 시스템 및 방법을 구조적 특징 및/또는 방법론적 동작 또는 작용에 특정한 언어로 설명하였지만, 첨부된 청구항에 정의된 구현예는 설명된 특정한 특징 또는 작용에 국한될 필요는 없다. 예를 들어, 다중타입 데이터 오브젝트의 강화된 클러스터링을 검색 용어 제안 애플리케이션에 관하여 설명하였지만, 다중타입 데이터 오브젝트의 강화된 클러스터링은 클러스터링을 이용하는 다수의 다른 타입의 애플리케이션에도 적용될 수 있다. 따라서, 상기한 특정한 특징 및 작용들은 청구된 주제(subject matter)를 구현하기 위한 예시적 형태로서 개시된 것이다.

발명의 효과

[0161] 본 발명에 따르면, 이종 데이터 오브젝트들 간의 관계의 측면에서 관련 오브젝트들(예를 들면 용어들)을 식별하고 그룹화하는 보다 나은 클러스터링 기술들을 제공한다. 이러한 클러스터링 기술들이 사용되어, 예를 들면,

검색 엔진 최적화 및 용어 입찰을 위해 용어(들)을 식별하는 시스템 및 방법들을 제공함으로써, 관련 용어(들)을 실질적으로 보다 높은 확률로 식별할 수 있다.

도면의 간단한 설명

[0001] 도 1은, 이종 데이터 오브젝트들/노드들의 다중 계층(102) 및 연관된 계층간 및 계층내 데이터 오브젝트 링크들/관계들을 포함하는, 다중 계층화된 프레임워크 그래프(100)를 나타낸다.

[0002] 도 2는, 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된 클러스터링(clustering)에 대한 예시적인 시스템을 나타낸다.

[0003] 도 3은, 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된 클러스터링에 대한 예시적인 절차를 나타낸다.

[0004] 도 4는, 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된 클러스터링에 대한 도 3의 예시적인 절차(300)의 연속 도면이다.

[0005] 도 5는, 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된 클러스터링에 대한 도 3 및 4의 예시적인 절차(300)의 연속 도면이다.

[0006] 도 6은, 도 3의 블럭(312)의 강화된 클러스터링 동작들의 예시적인 상세를 나타낸다.

[0007] 도 7은, 도 3 및 6의 블럭(312)의 강화된 클러스터링 동작들의 예시적인 연속 도면이다.

[0008] 도 8은, 검색어 제안을 위한 다중 유형 데이터 오브젝트들의 강화된 클러스터링을 위한 후속 설명된 시스템, 장치 및 방법들이 완전히 또는 부분적으로 구현될 수 있는 예시적인 적절한 컴퓨팅 환경을 나타낸다.

[0009] <도면의 주요부분에 관한 부호의 설명>

[0010] 102, 102-1, 102-2, 102-N: 층(노드 셋)

[0011] 106: 참조 링크

[0012] 108: 이슈 링크

[0013] 104: 층간 링크

[0014] 110: 브라우즈 링크

[0015] 228: 검색 엔진

[0016] 222: 높은 FOO 쿼리 조건(용어)

[0017] 224: 낮은 FOO 쿼리 조건(용어)

[0018] 202: 편집 검증 서버

[0019] 212: 검색 조건(용어) 제안 모듈

[0020] 214: 다중 타입 데이터 오브젝트(MDOs)

[0021] 216: 히스토리컬 쿼리들

[0022] 218: 기타 MDOs

[0023] 230: 검색 결과

[0024] 232: 추출된 특성

[0025] 234: MDO 벡터(들)

[0026] 236: 강화된 클러스터들

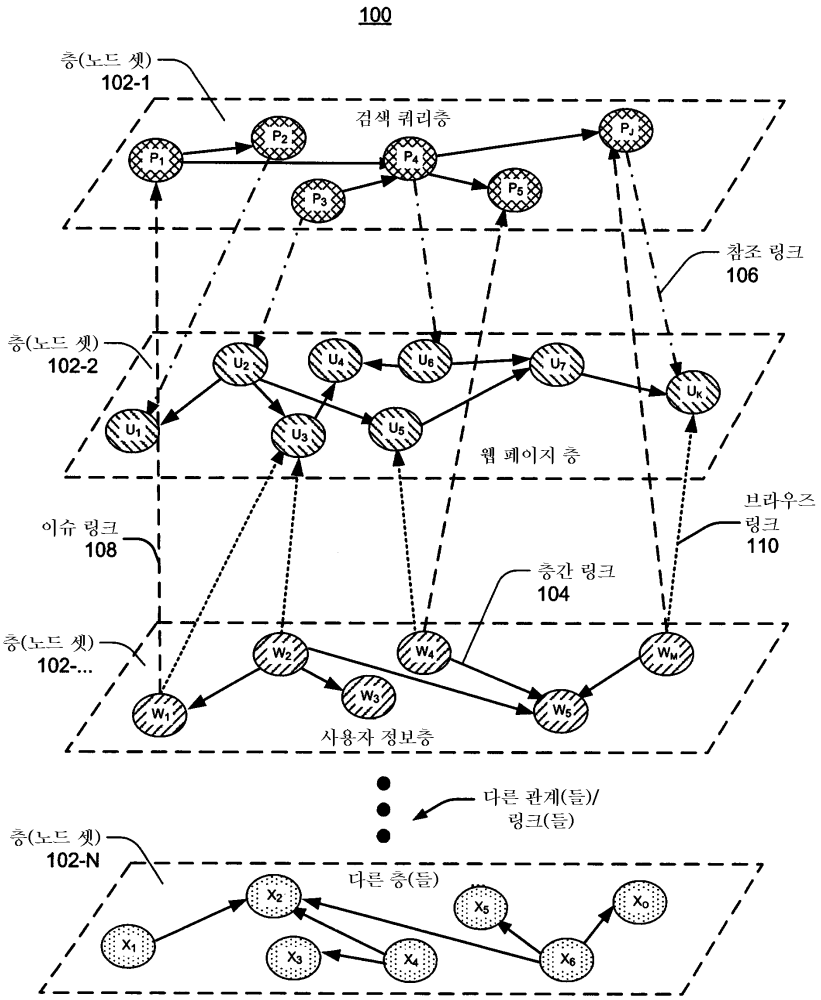
[0027] 238: 훈련된 분류기

[0028] 220: 쿼리 로그(들)(쿼리 조건)

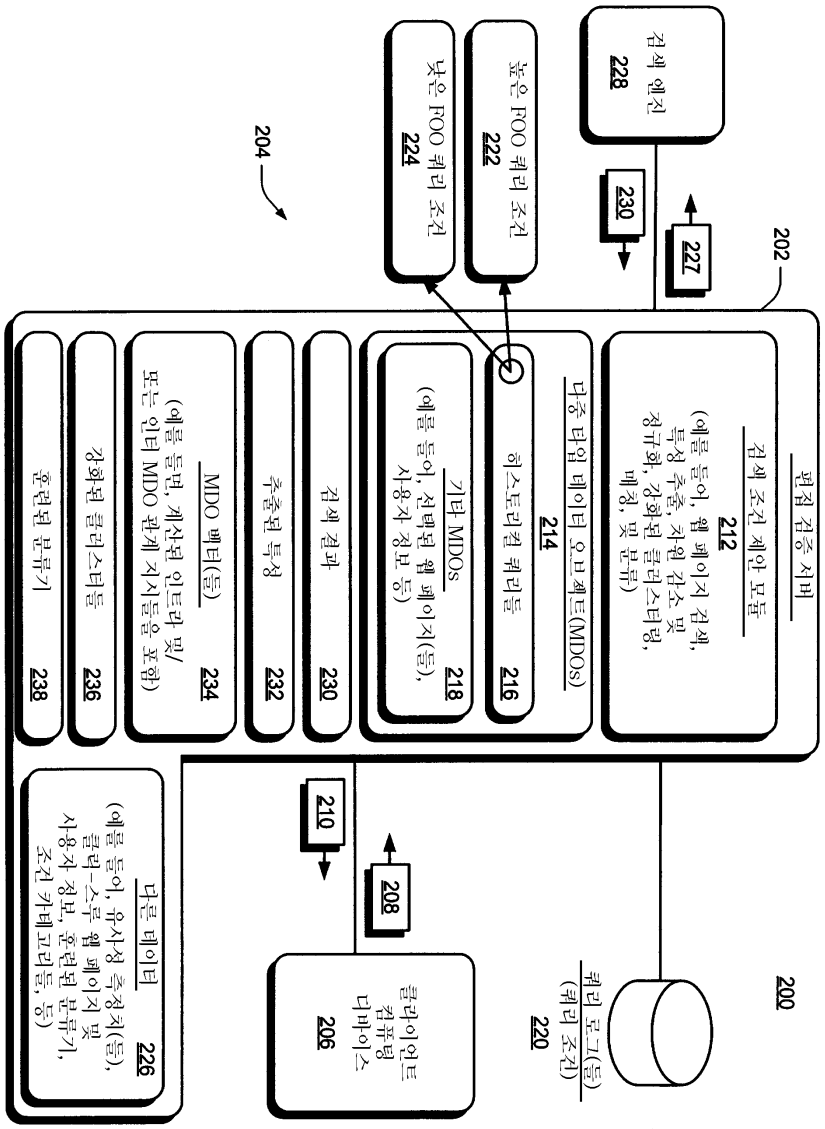
[0029] 206: 클라이언트 컴퓨팅 디바이스

도면

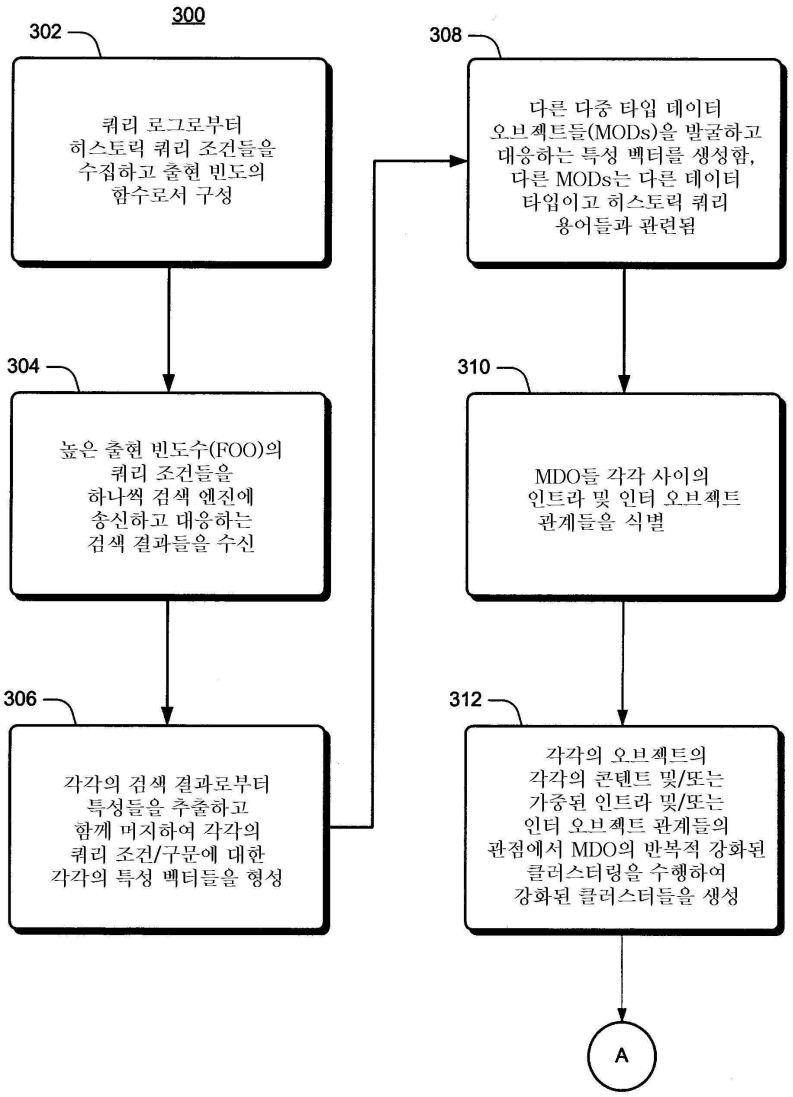
도면1



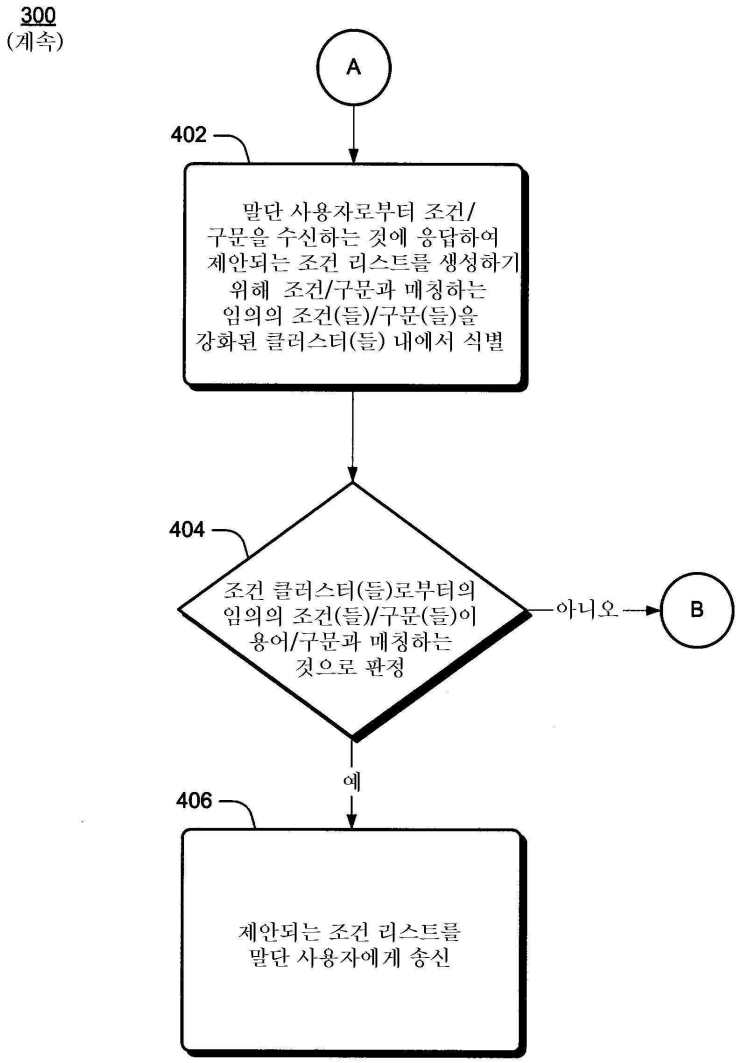
도면2


편집 검증 서버
202
검색 조건 제안 모듈
212
(예를 들어, 웹 페이지 검색, 특성 추출, 차원 감소 및 정규화, 강화된 클러스터링, 메칭, 및 분류)
다중 타입 데이터 오브젝트(MDOs)
214
히스토리컬 쿼리들 216
기타 MDOs 218
(예를 들어, 선택된 웹 페이지(들), 사용자 정보 등)
검색 결과 230
추출된 특성 232
MDO 벡터(들) 234
(예를 들면, 계산된 인트라 및/또는 인터 MDO 관계 지시들을 포함)
강화된 클러스터들 236
훈련된 분류기 238
다른 데이터 226
(예를 들어, 유사성 측정치(들), 클릭-스루 웹 페이지 및 사용자 정보, 훈련된 분류기, 조건 카테고리들, 등)
204
높은 FOO 쿼리 조건 222
낮은 FOO 쿼리 조건 224
검색 엔진 228
227
230
210
208
클라이언트 컴퓨팅 디바이스 206
쿼리 로그(들) (쿼리 조건) 220
200

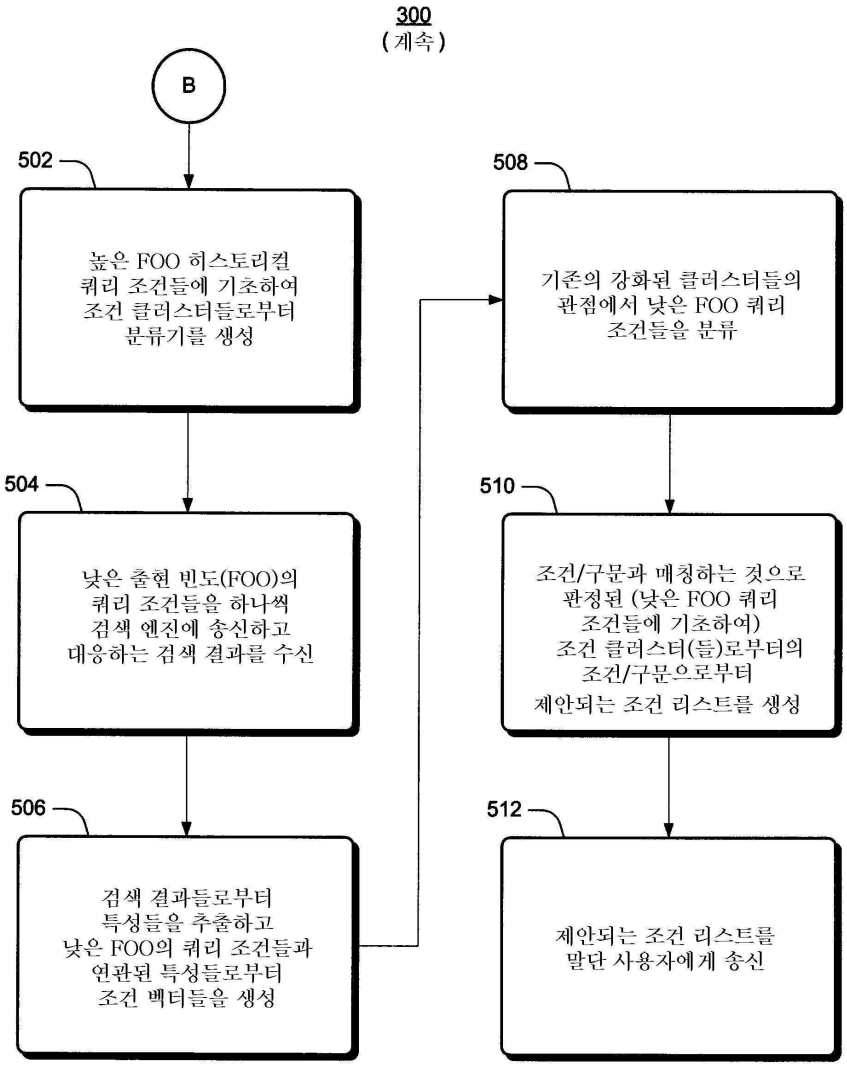
도면3


300
302
쿼리 로그로부터 히스토릭 쿼리 조건들을 수집하고 출현 빈도의 함수로서 구성
304
높은 출현 빈도수(FOO)의 쿼리 조건들을 하나씩 검색 엔진에 송신하고 대응하는 검색 결과들을 수신
306
각각의 검색 결과로부터 특성들을 추출하고 함께 머지하여 각각의 쿼리 조건/구문에 대한 각각의 특성 벡터들을 형성
308
다른 다중 타임 데이터 오브젝트들(MODs)을 발굴하고 대응하는 특성 벡터를 생성함, 다른 MODs는 다른 데이터 타임이고 히스토릭 쿼리 용어들과 관련됨
310
MDO들 각각 사이의 인트라 및 인터 오브젝트 관계들을 식별
312
각각의 오브젝트의 각각의 콘텐트 및/또는 가중된 인트라 및/또는 인터 오브젝트 관계들의 관점에서 MDO의 반복적 강화된 클러스터링을 수행하여 강화된 클러스터들을 생성
A

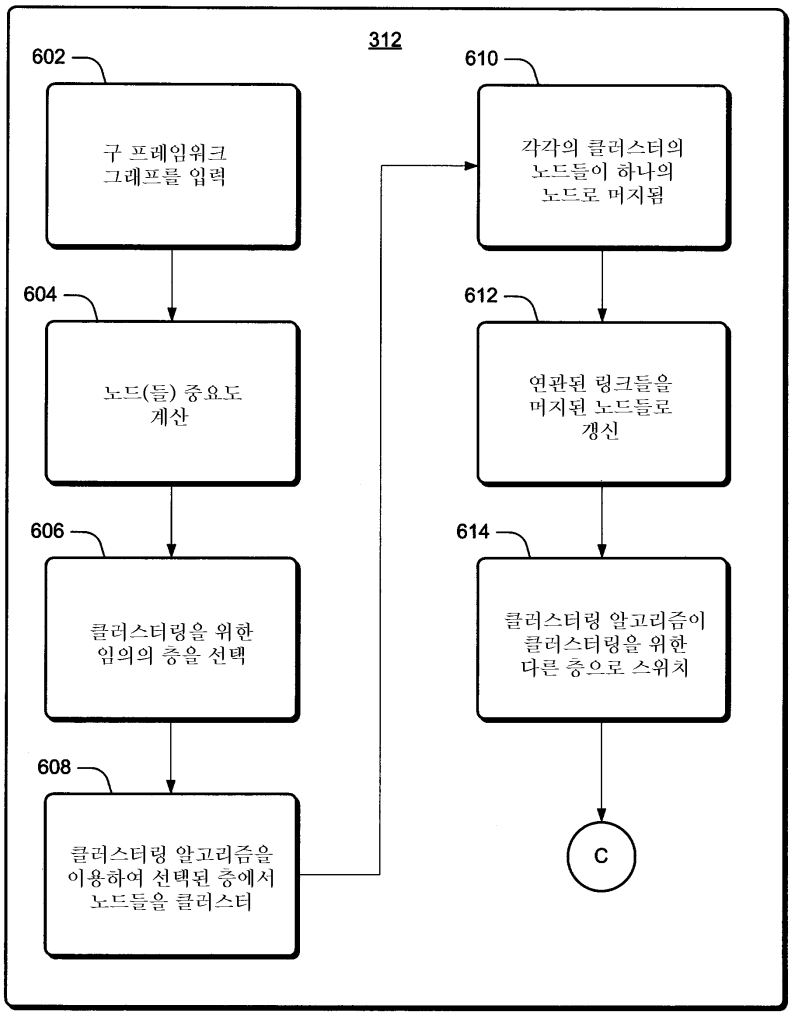
도면4


300
(계속)
A
402
말단 사용자로부터 조건/
구문을 수신하는 것에 응답하여
제안되는 조건 리스트를 생성하기
위해 조건/구문과 매칭하는
임의의 조건(들)/구문(들)을
강화된 클러스터(들) 내에서 식별
404
조건 클러스터(들)로부터의
임의의 조건(들)/구문(들)이
용어/구문과 매칭하는
것으로 판정
아니오
B
예
406
제안되는 조건 리스트를
말단 사용자에게 송신

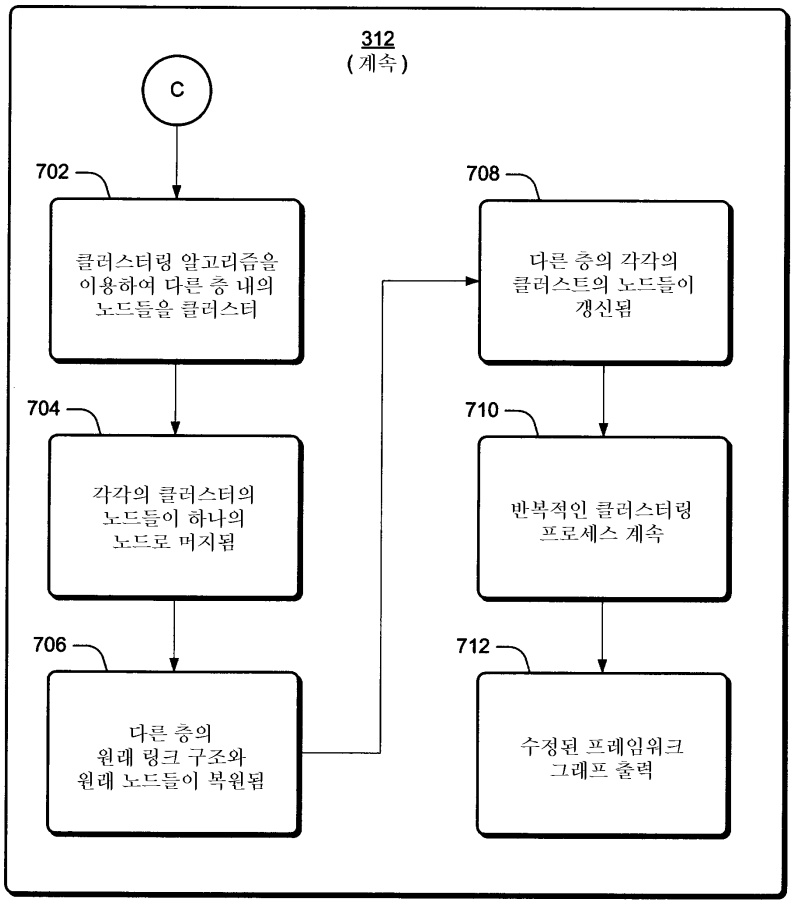
도면5


300
(계속)
B
502
높은 FOO 히스토리컬 쿼리 조건들에 기초하여 조건 클러스터들로부터 분류기를 생성
504
낮은 출현 빈도(FOO)의 쿼리 조건들을 하나씩 검색 엔진에 송신하고 대응하는 검색 결과를 수신
506
검색 결과들로부터 특성들을 추출하고 낮은 FOO의 쿼리 조건들과 연관된 특성들로부터 조건 벡터들을 생성
508
기존의 강화된 클러스터들의 관점에서 낮은 FOO 쿼리 조건들을 분류
510
조건/구문과 매칭하는 것으로 판정된 (낮은 FOO 쿼리 조건들에 기초하여) 조건 클러스터(들)로부터의 조건/구문으로부터 제안되는 조건 리스트를 생성
512
제안되는 조건 리스트를 말단 사용자에게 송신

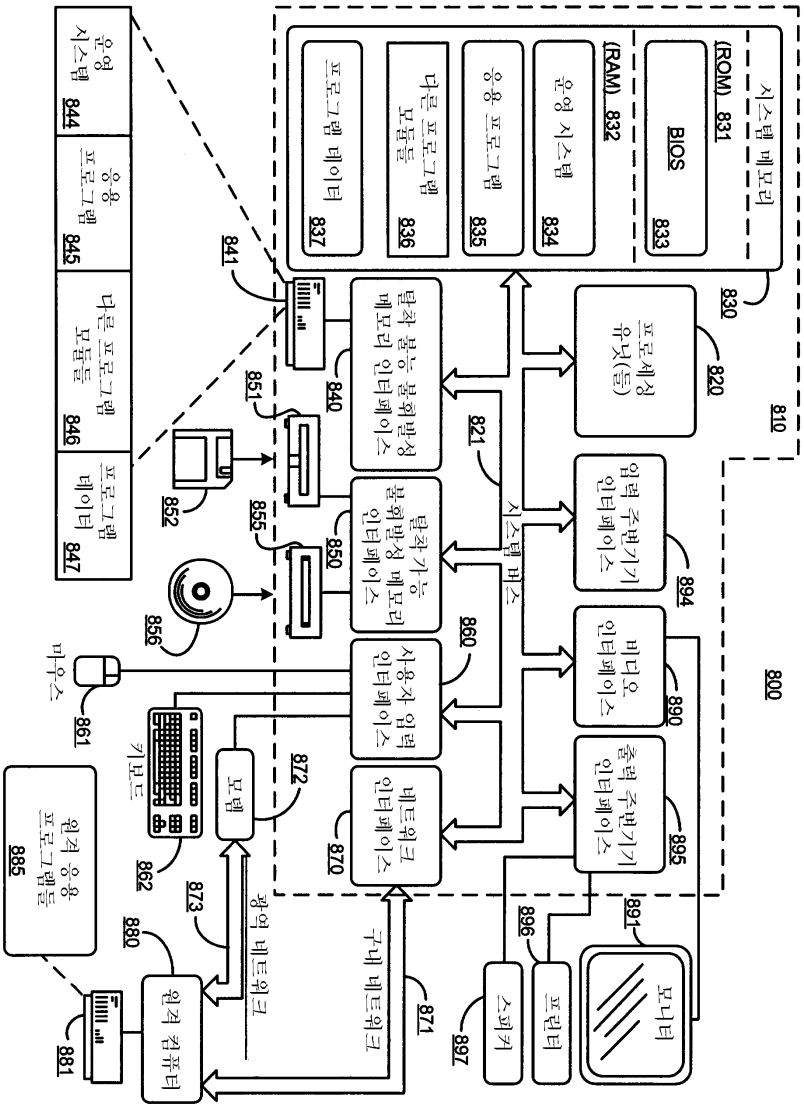
도면6


312
602
구 프레임워크 그래프를 입력
604
노드(들) 중요도 계산
606
클러스터링을 위한 임의의 층을 선택
608
클러스터링 알고리즘을 이용하여 선택된 층에서 노드들을 클러스터
610
각각의 클러스터의 노드들이 하나의 노드로 머지됨
612
연관된 링크들을 머지된 노드들로 갱신
614
클러스터링 알고리즘이 클러스터링을 위한 다른 층으로 스위치
C

도면7


312
(계속)
C
702
클러스터링 알고리즘을 이용하여 다른 층 내의 노드들을 클러스터
704
각각의 클러스터의 노드들이 하나의 노드로 머지됨
706
다른 층의 원래 링크 구조와 원래 노드들이 복원됨
708
다른 층의 각각의 클러스트의 노드들이 갱신됨
710
반복적인 클러스터링 프로세스 계속
712
수정된 프레임워크 그래프 출력

도면8


시스템 메모리
(ROM) 831
BIOS 833
(RAM) 832
운영 시스템 834
응용 프로그램 835
다른 프로그램 모듈들 836
프로그램 데이터 837
830
프로세싱 유닛(들) 820
810
800
821
시스템 버스
탈착 불능 불휘발성 메모리 인터페이스 840
841
851
852
탈착 가능 불휘발성 메모리 인터페이스 850
855
856
860
사용자 입력 인터페이스
861
마우스
키보드
862
872
모뎀
873
네트워크 인터페이스 870
871
구내 네트워크
광역 네트워크
880
원격 컴퓨터
881
원격 응용 프로그램들 885
입력 주변기기 인터페이스 894
비디오 인터페이스 890
출력 주변기기 인터페이스 895
896
891
스피커 897
프린터
모니터
운영 시스템 844
응용 프로그램 845
다른 프로그램 모듈들 846
프로그램 데이터 847