(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년06월01일
(11) 등록번호 10-1150063
(24) 등록일자 2012년05월18일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(21) 출원번호 10-2005-0079872
(22) 출원일자 2005년08월30일
심사청구일자 2010년08월23일
(65) 공개번호 10-2006-0050800
(43) 공개일자 2006년05월19일
(30) 우선권주장
10/969,567 2004년10월20일 미국(US)
(56) 선행기술조사문헌
JP2000311176 A
JP2001209646 A
JP2004252911 A

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
앤더슨, 블레이크 이.
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
베어, 프레드릭 에이치 주니어
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
(뒷면에 계속)
(74) 대리인
제일특허법인

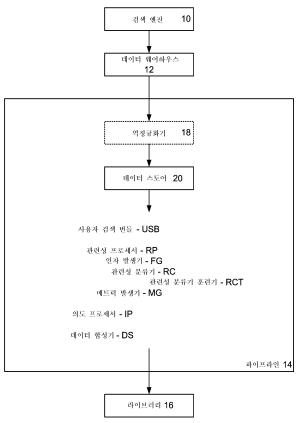
전체 청구항 수 : 총 18 항 심사관 : 석상문

(54) 발명의 명칭 **검색 시스템 등으로부터의 동작 데이터 및 기타 데이터를분석하는 시스템 및 방법**

(57) 요 약

시스템은 검색 엔진으로부터의 데이터를 분석한다. 사용자 검색 번들러는 사용자 검색을 분석하고 유사한 사용자 검색들을 사용자 검색 번들로 그룹화하며, 의도 프로세서는 사용자 검색 번들에 기초하여 의도를 생성한다. 인자 발생기는 사용자 검색 및 관련 정보를 고려하여 인자를 생성하고, 여기서 각각의 인자는 일련의 검색 결과로부터의 특정 결과에 관한 것이다. 관련성 분류기는 인자를 수신하고 이에 기초하여 각 결과에 대한 판정을 생성하는 동작을 한다. 메트릭 발생기는 인자 및 판정에 기초하여 메트릭을 생성하고, 데이터 합성기는 추출된 데이터를 데이터베이스로 포맷화한다.

대 표 도 - 도2



(72) 발명자

핑거, 제임스 씨

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

마스맨, 제니퍼 제이.

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

카나와트, 쿨딥

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

미들랜드, 마크 비.

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

말로렙싸이, 폴 엠

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

시미즈, 다께시

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

화이트, 토마스 디.

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

장, 잉

미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내

특허청구의 범위

청구항 1

검색 엔진으로부터의 데이터를 분석하는 시스템으로서,

상기 검색 엔진은 요청측 사용자로부터 수신된 쿼리 문자열(Query string)에 기초하여 사용자 검색(user search)을 생성하고, 상기 사용자 검색은 상기 쿼리 문자열 및 상기 검색 결과를 포함하며, 상기 검색 결과는 적어도 하나의 결과를 포함하고, 각각의 결과는 상기 쿼리 문자열과 관련있는 콘텐츠의 특정 아이템을 참조하며,

상기 시스템은,

이러한 사용자 검색 중 유사한 것을 찾아내기 위해 사용자 검색을 분석하고 이러한 유사한 사용자 검색을 사용자 검색 번들로 그룹화하는 사용자 검색 번들러(User Search Bundler, USB),

상기 USB로부터의 사용자 검색 번들에 기초하여 의도(intent)를 생성하는 의도 프로세서(Intent Processor, IP) - 각각의 의도는 서로 관련된 하나 이상의 세션의 그룹이고, 각 세션은 상기 쿼리 문자열에 관련된 각각의 검색 문자열에 기초하여 생성되는 일련의 관련된 사용자 검색들을 포함함 -,

사용자 검색 및 관련 정보에 기초하여 인자를 생성하는 인자 발생기(Factor Generator, FG) - 각각의 인자는 일련의 검색 결과로부터의 특정 결과에 관한 것이고, 각각의 인자는 하나 이상의 이벤트에 관련되어 있으며 각각의 이벤트는 질의하는 사용자가 수행한 행위에 관련한 정보임 -,

각각의 결과에 대해 상기 FG에 의해 발생된 상기 인자를 수신하고 상기 인자에 기초하여 상기 결과에 대한 판정(Judgement)을 생성하는 관련성 분류기(Relevance Classifier, RC) - 상기 판정은 사용자가 상기 검색 결과로부터 상기 결과에 액세스하기로 결정할 때 사용자가 상기 결과를 어떻게 판정했는지의 결정을 나타내는 것임-,

상기 FG에 의해 발생된 상기 인자들 및 상기 RC에 의해 생성된 상기 판정에 기초하여 메트릭을 생성하는 메트릭 발생기(Metric Generator, MG) - 각각의 메트릭은 결과, 사용자 검색, 또는 세션과 관련한 척도임 -, 및

상기 USB, IP, FG, RC 및 MG에 의해 발생된 데이터를 추출하고, 상기 추출된 데이터를 하나 이상의 데이터베이스로 포맷화하며, 상기 데이터베이스를 라이브러리에 저장하는 데이터 합성기(data synthesizer, DS)를 포함하는, 시스템.

청구항 2

제1항에 있어서, 상기 검색 엔진은 상기 쿼리 문자열, 상기 검색 결과 및 상기 관련 정보를 데이터 웨어하우스에 정규화된 형태로 저장하고,

상기 시스템은 상기 데이터 웨어하우스로부터 정규화된 데이터를 복원하고(retrieve), 상기 정규화된 데이터를 역정규화하고 역정규화된 데이터를 데이터 스토어에 저장하는 역정규화기(denormalizer)를 더 포함하는, 시스템.

청구항 3

제1항에 있어서, 상기 USB는 쿼리 문자열의 유사성 및 검색 결과의 유사성 중 적어도 하나에 대한 상기 사용자 검색을 분석하는, 시스템.

청구항 4

제1항에 있어서, 각각의 이벤트는 상기 사용자가 특정의 결과를 선택하는 것과 상기 특정의 결과에 대한 뷰(view)를 닫는 것 중 적어도 하나를 수행한 시각을 포함하고,

상기 FG는 사용자가 결과를 본 시간 길이를 나타낸 "체류 시간(Dwell time)" 인자를 계산하며,

상기 체류 시간 인자는 상기 사용자가 상기 결과를 선택한 때와 상기 결과에 대한 뷰를 닫은 때 사이의 시간차에 기초하고 이 각각의 때는 대응하는 타임 스탬프드 이벤트(time-stamped Event)로 표현되는, 시스템.

청구항 5

제1항에 있어서, 상기 RC는 "수락" 판정, "탐색" 판정, 및 "거부" 판정 중 적어도 하나를 포함하는 판정, 및 상기 판정이 옳을 가능성에 대한 확신을 나타내는 대응하는 값을 생성하는, 시스템.

청구항 6

제1항에 있어서, 상기 FG로부터 명시적인 판정 인자를 수신하고 이에 기초하여 상기 RC를 발생하는 관련성 분류기 훈련기(Relevance Classifier Trainer, RCT) -각각의 명시적인 판정 인자는 대응하는 결과에 관한 상기 사용자로부터의 명시적인 피드백을 나타냄-를 더 포함하고,

상기 RCT는 상기 명시적인 판정 인자로부터 어떤 인자가 어느 판정을 암시하는지를 학습하고 그에 기초하여 RC를 생성하는, 시스템.

청구항 7

제1항에 있어서, 상기 MG는 결과와 관련하여,

어떻게 사용자가 상기 결과를 정렬한 것으로 판정되는지에 관한 위치 메트릭(Position Metric),

상기 결과가 상기 검색 결과 내에서 어떻게 배치되어 있는가에 관한 관련성 위치 메트릭(Relevance Positioin Metric), 및

상기 위치 메트릭 및 상기 관련성 위치 메트릭에 기초하여, 상기 결과가 자신이 있어야만 하는 곳으로부터 얼마나 '멀리' 있는지에 관한 오정렬된 결과 메트릭(Mis-ranked Result Metric) 중 적어도 하나를 생성하는, 시스템.

청구항 8

제1항에 있어서, 상기 IP는 검토된 사용자 검색 번들에 기초하여 세션들에 걸친 공통 결과 및 세션들에 걸친 공통 질의어를 찾아냄으로써 세션들 간의 관계값(relationship value)을 결정하고, 이러한 공통 결과가 발견된 경우 공통성의 세기(Strength of Commonality)를 확인하며,

이러한 공통성의 세기는 2개의 세션이 공통 목적을 가짐으로써 서로 관련될 가능성이 얼마나 되는지를 나타내고, 상기 IP는 결정된 임계값(determined threshold)을 넘는 공통성의 세기를 갖는 세션 쌍들을 그룹화하여 상기 의도를 생성하는, 시스템.

청구항 9

제1항에 있어서, 상기 DS는 상기 추출된 데이터를 관계 데이터베이스로 포맷화하는, 시스템.

청구항 10

검색 엔진으로부터의 데이터를 분석하는 방법으로서,

상기 검색 엔진은 요청측 사용자로부터 수신된 쿼리 문자열에 기초하여 사용자 검색을 생성하고, 상기 사용자 검색은 상기 쿼리 문자열 및 상기 검색 결과를 포함하며, 상기 검색 결과는 적어도 하나의 결과를 포함하고, 각각의 결과는 상기 쿼리 문자열과 관련있는 콘텐츠의 특정 아이템을 참조하며,

상기 방법은,

이러한 사용자 검색 중 유사한 것을 찾아내기 위해 사용자 검색을 분석하고 이러한 유사한 사용자 검색을 사용자 검색 번들로 그룹화하는 단계,

사용자 검색 번들러(USB)로부터의 상기 사용자 검색 번들에 기초하여 의도를 생성하는 단계 - 각각의 의도는 서로 관련된 하나 이상의 세션의 그룹이고, 각 세션은 상기 쿼리 문자열에 관련된 각각의 검색 문자열에 기초하여 생성되는 일련의 관련된 사용자 검색들을 포함함 -,

사용자 검색 및 관련 정보에 기초하여 인자를 생성하는 단계 - 각각의 인자는 일련의 검색 결과로부터의 특정 결과에 관한 것이고 각각의 인자는 하나 이상의 이벤트에 관련되어 있으며 각각의 이벤트는 질의하는 사용자가 수행한 행위에 관련한 정보임 -,

관련성 분류기(Relevance Classifier, RC)에서, 각각의 결과에 대해 발생된 상기 인자를 수신하고 상기 인자에 기초하여 상기 결과에 대한 판정을 생성하는 단계 - 상기 판정은 사용자가 상기 검색 결과로부터 상기 결

과에 액세스하기로 결정할 때 사용자가 상기 결과를 어떻게 판정했는지의 결정을 나타냄 -,

상기 인자 및 상기 판정에 기초하여 메트릭을 생성하는 단계 - 각각의 메트릭은 결과, 사용자 검색, 또는 세션과 관련한 척도임 -, 및

상기 사용자 검색 번들, 상기 의도, 상기 인자, 상기 판정, 및 상기 메트릭을 포함하는 데이터를 추출하고, 상기 추출된 데이터를 하나 이상의 데이터베이스로 포맷화하며, 상기 데이터베이스를 라이브러리에 저장하는 단계를 포함하는, 방법.

청구항 11

제10항에 있어서,

상기 쿼리 문자열, 상기 검색 결과 및 상기 관련 정보를 저장하는 데이터 웨어하우스로부터 정규화된 데이터를 복원하고, 상기 정규화된 데이터를 역정규화하여, 역정규화된 데이터를 데이터 스토어에 저장하는 단계를 더 포함하는, 방법.

청구항 12

제10항에 있어서, 쿼리 문자열의 유사성 및 검색 결과의 유사성 중 적어도 하나에 대한 상기 사용자 검색을 분석하는 단계를 포함하는, 방법.

청구항 13

제10항에 있어서, 각각의 이벤트는 상기 사용자가 특정의 결과를 선택하는 것과 상기 특정의 결과에 대한 뷰(view)를 닫는 것 중 적어도 하나를 수행한 시각을 포함하고,

상기 인자를 선택하는 단계는, 사용자가 결과를 본 시간 길이를 나타내는 "체류 시간" 인자를 계산하는 단계를 포함하며, 상기 체류 시간 인자는 상기 사용자가 상기 결과를 선택한 때와 상기 결과에 대한 뷰를 닫은 때 사이의 시간차에 기초하고 이 각각의 때는 대응하는 타임스탬프드 이벤트(time-stamped Event)로 표현되는, 방법.

청구항 14

제10항에 있어서, 상기 판정을 생성하는 단계는, "수락" 판정, "탐색" 판정, 및 "거부" 판정 중 적어도 하나를 포함하는 판정, 및 상기 판정이 옳을 가능성에 대한 확신을 나타내는 대응하는 값을 생성하는 단계를 포함하는, 방법.

청구항 15

제10항에 있어서, 명시적인 판정 인자를 수신하고 이에 기초하여 상기 관련성 분류기를 발생하는 단계를 더 포함하고, 상기 관련성 분류기는 각각의 결과에 대해 생성된 상기 인자를 수신하고 이에 기초하여 상기 결과에 대한 상기 판정을 생성하는 동작을 하며, 각각의 명시적인 판정 인자는 상기 결과에 관한 상기 사용자로부터의 명시적인 피드백을 나타내며 그에 따라 이러한 명시적인 판정 인자에 기초하여 어떤 인자가 어느 판정을 암시하는지가 학습될 수 있는, 방법.

청구항 16

제10항에 있어서, 상기 메트릭을 생성하는 단계는, 결과와 관련하여,

어떻게 사용자가 상기 결과를 정렬한 것으로 판정되는지에 관한 위치 메트릭,

상기 결과가 상기 검색 결과 내에서 어떻게 배치되어 있는가에 관한 관련성 위치 메트릭, 및

상기 위치 메트릭 및 상기 관련성 위치 메트릭에 기초하여, 상기 결과가 자신이 있어야만 하는 곳으로부터 얼마나 '멀리' 있는지에 관한 오정렬된 결과 메트릭 중 적어도 하나를 생성하는 단계를 포함하는, 방법.

청구항 17

제10항에 있어서, 상기 의도를 생성하는 단계는

검토된 사용자 검색 번들에 기초하여 세션들에 걸친 공통 결과 및 세션들에 걸친 공통 질의어를 찾아냄으로써

세션들 간의 관계값을 결정하는 단계;

이러한 공통 결과가 발견된 경우 공통성의 세기를 확인하는 단계; 및

결정된 임계값을 넘는 공통성의 세기를 갖는 세션 쌍들을 그룹화하여 상기 의도를 생성하는 단계를 포함하고,

상기 공통성의 세기는 2개의 세션이 공통 목적을 가짐으로써 서로 관련될 가능성이 얼마나 되는지를 나타내는, 방법.

청구항 18

제10항에 있어서, 상기 데이터베이스는 관계 데이터베이스인, 방법.

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

[0010] 본 발명은 검색 요청에 응답하는 시스템 등의 시스템의 동작 동안에 컴파일된 데이터를 비롯한 데이터를 분석하는 시스템 및 방법에 관한 것이다. 보다 상세하게는, 본 발명은 응답을 향상시키는 것 및 시스템 성능을 유지하는 것을 비롯한 목적들을 위해 비교적 대량의 데이터를 분석하는 이러한 시스템 및 방법에 관한 것이다.

[0011] 일반적인 검색 엔진 등의 시스템과 관련하여, 이 시스템에 액세스하는 사용자는 아마도 부울 연산자와 함께 하나 이상의 검색어를 포함하는 검색 문자열 등을 입력함으로써 검색을 요청한다. 이에 응답하여, 검색 엔진은 검색 문자열에 기초하여 하나 이상의 데이터베이스를 검색하고, 그에 기초하여 일련의 검색 결과를 발생하며, 이러한 검색 결과를 아마도 사용자가 검토할 수 있는 정보로의 링크 또는 정보 페이지의 형태로 요청측 사용자에게 반환한다. 특히 전자의 경우, 사용자는 특정의 검색 결과와 관련한 콘텐츠를 검토하기 위해 링크들 중 하나 이상에 액세스할 수 있고, 검색 결과의 하나 이상의 링크와 연관된 콘텐츠가 사용자에게 만족스러운 경우, 이러한 사용자는 일반적으로 계속하여 이러한 만족스러운 콘텐츠를 적당하다고 생각되는 임의의 방식으로 이용하게 된다.

[0012] 그렇지만, 그 대신에 검색 결과의 콘텐츠 중 어느 것도 적어도 이러한 사용자의 관점에서 볼 때 요청된 검색을 충족시키지 못한다는 점에서 그 검색 결과가 사용자에게 만족스럽지 못할 경우도 있을 수 있다. 이러한 경우, 사용자는 새로운 검색 문자열 또는 이전에 입력된 검색 문자열의 수정을 입력하고 이러한 새로운 또는 수정된 검색 문자열에 기초한 검색 엔진으로부터의 검색 결과를 검토하기로 결정할 수도 있다. 잘 알고 있는 바와 같이, 이러한 프로세스는 사용자가 만족스러운 검색 결과를 찾거나 또는 포기할 때까지 검색 세션의 형태로 몇번씩 반복될 수 있다.

[0013] 일반적으로, 전술한 검색 엔진 등의 고품질 시스템에서, 검색 문자열에 기재된 것같은 사용자로부터의 각 쿼리는 그 쿼리에 응답하는 콘텐츠를 나타내는 검색 결과에 정확하게 매핑되어야만 한다. 이러한 목표는 양호한 검색 경험을 제공하는 데 필수적이며, 실제로 이러한 목표를 달성한다는 것은 새로운 검색 세션으로 검색 엔진에 되돌아오는 행복하고 만족한 사용자와 그 대신에 다른 검색 엔진을 방문하는 화가 나고 불만족한 사용자 간의 차이를 말할 수 있다.

[0014] 이러한 목표를 달성하기 위해, 시스템 자체가 응답을 개선시키고 또 시스템 성능을 유지하기 위해 조정 또는 "튜닝"되어야만 한다는 것을 알았다. 잘 알고 있는 바와 같이, 이러한 조정은 전적으로 그래야만 하는 것은 아니지만 주로 시스템의 동작 동안에 컴파일된 동작 데이터 및 기타 데이터에 기초하여 행해질 수 있다. 그렇지만, 시스템이 특히 대규모이거나 많은 트래픽을 갖거나 하는 경우, 분석에 이용가능한 이러한 데이터의 양은 특히 대량이며 따라서 다루기 힘드고 그렇지 않으면 그를 가지고 작업하기가 어려울 수 있다. 따라서, 대량의 데이터, 특히 검색 시스템 등으로부터의 대량의 데이터를 분석하는 시스템 및 방법이 필요하다.

발명이 이루고자 하는 기술적 과제

[0015] 전술한 필요성은 검색 엔진으로부터의 데이터를 분석하는 시스템이 제공되는 본 발명에 의해 적어도 부분적으

로 만족된다. 검색 엔진은 요청측 사용자로부터 수신된 쿼리 문자열에 기초하여 일련의 검색 결과를 발생하며, 여기서 쿼리 문자열 및 검색 결과가 공동으로 사용자 검색을 구성한다. 검색 결과는 적어도 하나의 결과를 포함하며, 각각의 결과는 쿼리 문자열과 관련있는 것으로 생각되는 콘텐츠의 특정 아이템을 참조한다. 일련의 관련 사용자 검색이 세션을 이루며, 검색 엔진은 각각의 사용자 검색 및 관련 정보를 저장한다.

[0016] 이 시스템에서, 사용자 검색 번들러(User Search Bundler, USB)는 이러한 사용자 검색 중 유사한 것을 찾아내기 위해 사용자 검색을 분석하고 이러한 유사한 사용자 검색을 사용자 검색 번들로 그룹화하며, 의도 프로세서(Intent Processor, IP)는 USB로부터의 사용자 검색 번들에 기초하여 의도(intent)를 생성한다. 각각의 의도는 서로 관련된 것으로 생각되는 하나 이상의 세션의 그룹이다.

[0017] 인자 발생기(Factor Generator, FG)는 사용자 검색 및 관련 정보를 고려하여 인자를 생성하며, 여기서 각각의 인자는 일련의 검색 결과로부터의 특정 결과에 관한 것이다. 각각의 인자는 하나 이상의 이벤트에 관련되어 있으며, 여기서 각각의 이벤트는 질의하는 사용자가 수행한 행위에 관련한 정보이다. 관련성 분류기(Relevance Classifier, RC)는 각각의 결과에 대해 FG에 의해 발생된 인자를 수신하고 그에 기초하여 그 결과에 대한 판정(Judgement)을 생성하는 동작을 하며, 여기서 판정이란 사용자가 검색 결과로부터 결과에 액세스하기로 결정할 때 결과를 어떻게 판단했는지의 결정을 말한다. 메트릭 발생기(Metric Generator, MG)는 FG에 의해 발생된 인자 및 RC에 의해 생성된 판정에 기초하여 메트릭을 생성하며, 여기서 각각의 메트릭은 결과, 사용자 검색, 또는 세션과 관련한 척도이다. 마지막으로, 데이터 합성기(data synthesizer, DS)는 USB, IP, FG, RC 및 MG에 의해 발생된 데이터를 추출하고, 추출된 데이터를 하나 이상의 데이터베이스로 포맷화하며, 이 데이터베이스를 라이브러리에 저장하고, 그에 따라 데이터는 피드백을 제공하거나 리포트를 생성하기 위해 검토되고 통합될 수 있다.

발명의 구성 및 작용

[0018] 이상의 요약은 물론 본 발명의 실시예들에 대한 이하의 상세한 설명은 첨부 도면을 참조하여 읽어가면 보다 잘 이해될 것이다. 본 발명을 설명하기 위해, 도면에는 현재 양호한 실시예들에 도시되어 있다. 그렇지만, 잘 알고 있는 바와 같이, 본 발명은 도면에 도시된 구체적인 구성 및 수단에 한정되지 않는다.

[0019] 도 1 및 이하의 기재는 본 발명 및/또는 그의 일부분이 구현될 수 있는 적당한 컴퓨팅 환경에 대한 간략한 일반적인 설명을 제공하기 위한 것이다. 본 발명이 클라이언트 워크스테이션 또는 서버 등의 컴퓨터에 의해 실행되는 프로그램 모듈 등의 컴퓨터 판독가능 명령어의 일반적 관점에서 기술되고 있지만, 꼭 그럴 필요가 있는 것은 아니다. 일반적으로, 프로그램 모듈은 특정 태스크를 수행하거나 특정 추상 데이터 유형을 구현하는 루틴, 프로그램, 객체, 컴포넌트, 데이터 구조 등을 포함한다. 게다가, 본 발명 및/또는 그의 일부분이 핸드헬드 디바이스, 멀티프로세서 시스템, 마이크로프로세서 기반 또는 프로그램가능 가전 제품, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터 등을 비롯한 다른 컴퓨터 시스템 구성에서 실시될 수 있음을 잘 알 것이다. 본 발명은 또한 태스크들이 통신 네트워크를 통해 연결되어 있는 원격 프로세싱 장치에 의해 수행되는 분산 컴퓨팅 환경에서도 실시될 수 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 로컬 및 원격 메모리 저장 장치 둘다에 위치될 수 있다.

[0020] 도 1에 도시된 바와 같이, 전형적인 범용 컴퓨팅 시스템은 프로세싱 유닛(121), 시스템 메모리(122), 및 시스템 메모리를 비롯한 여러가지 시스템 컴포넌트를 프로세싱 유닛(121)에 연결시키는 시스템 버스(123)를 포함하는 종래의 퍼스널 컴퓨터(120) 등을 포함한다. 시스템 버스(123)는 메모리 버스 또는 메모리 콘트롤러, 주변 버스, 및 다양한 버스 아키텍처 중 임의의 것을 사용하는 로컬 버스를 비롯한 몇가지 유형의 버스 구조 중 임의의 것일 수 있다. 시스템 메모리는 판독 전용 메모리(ROM)(124) 및 랜덤 액세스 메모리(RAM)(125)를 포함한다. 시동 중과 같은 때에 퍼스널 컴퓨터(120) 내의 구성요소들 간의 정보의 전달을 돕는 기본적인 루틴을 포함하는 기본 입/출력 시스템(126)(BIOS)은 ROM(124)에 저장되어 있다.

[0021] 퍼스널 컴퓨터(120)는 하드 디스크(도시 생략)로부터 판독하고 그에 기록하기 위한 하드 디스크 드라이브(127), 분리형 자기 디스크(129)로부터 판독하거나 그에 기록하기 위한 자기 디스크 드라이브(128), 및 CD-ROM 또는 기타 광학 매체 등의 분리형 광학 디스크(131)로부터 판독하거나 그에 기록하기 위한 광학 디스크 드라이브(130)를 더 포함한다. 하드 디스크 드라이브(127), 자기 디스크 드라이브(128) 및 광학 디스크 드라이브(130)는 각각 하드 디스크 드라이브 인터페이스(132), 자기 디스크 드라이브 인터페이스(133), 및 광학 드라이브 인터페이스(134)에 의해 시스템 버스(123)에 연결된다. 이들 드라이브 및 그의 관련 컴퓨터 판독가능 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 및 퍼스널 컴퓨터(120)의 기타 데이터의 비휘발성 저장을 제공한다.

[0022] 본 명세서에 기술된 전형적인 환경이 하드 디스크, 분리형 자기 디스크(129) 및 분리형 광학 디스크(131)를 이용하고 있지만, 컴퓨터에 의해 액세스가능한 데이터를 저장할 수 있는 다른 유형의 컴퓨터 판독가능 매체도 역시 전형적인 오퍼레이팅 환경에서 사용될 수 있음을 잘 알 것이다. 이러한 다른 유형의 매체로는 자기 카세트, 플래쉬 메모리 카드, DVD(digital video disk), 베르누이 카트리지, 랜덤 액세스 메모리(RAM), 판독 전용 메모리(ROM) 등이 있다.

[0023] 오퍼레이팅 시스템(135), 하나 이상의 애플리케이션 프로그램(136), 기타 프로그램 모듈(137), 및 프로그램 데이터(138)를 비롯한 다수의 프로그램 모듈이 하드 디스크, 자기 디스크(129), 광학 디스크(131), ROM(124) 또는 RAM(125) 상에 저장될 수 있다. 사용자는 키보드(140) 및 포인팅 디바이스(142) 등의 입력 장치를 통해 퍼스널 컴퓨터(120)에 명령 및 정보를 입력할 수 있다. 다른 입력 장치(도시 생략)는 마이크로폰, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 및 다른 입력 장치는 종종 시스템 버스에 연결되어 있는 직렬 포트 인터페이스(146)를 통해 프로세싱 유닛(121)에 연결되지만, 병렬 포트, 게임 포트, 또는 USB(유니버설 직렬 버스) 등의 다른 인터페이스에 의해 연결될 수 있다. 모니터(147) 또는 다른 유형의 디스플레이 장치도 역시 비디오 어댑터(148) 등의 인터페이스를 통해 시스템 버스(123)에 연결된다. 모니터(147) 이외에, 퍼스널 컴퓨터는 일반적으로 스피커 및 프린터 등의 다른 주변 출력 장치(도시 생략)를 포함한다. 도 1의 전형적인 시스템은 또한 호스트 어댑터(155), SCSI(Small Computer System Interface) 버스(156), 및 SCSI 버스(156)에 연결된 외부 저장 장치(162)를 포함한다.

[0024] 퍼스널 컴퓨터(120)는 원격 컴퓨터(149) 등의 하나 이상의 원격 컴퓨터로의 논리적 연결을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(149)는 다른 퍼스널 컴퓨터, 서버, 라우터, 네트워크 PC, 피어 장치 또는 다른 통상의 네트워크 노드일 수 있으며, 일반적으로 퍼스널 컴퓨터(120)와 관련하여 전술한 구성 요소들 중 다수 또는 모두를 포함하지만, 도 1에서는 단지 메모리 저장 장치(150)만이 예시되어 있다. 도 1에 예시된 논리적 연결은 근거리 통신망(LAN)(151) 및 원거리 통신망(WAN)(152)을 포함한다. 이러한 네트워킹 환경은 사무실, 기업 규모의 컴퓨터 네트워크, 인트라넷 및 인터넷에서 통상적인 것이다.

[0025] LAN 네트워킹 환경에서 사용될 때, 퍼스널 컴퓨터(120)는 네트워크 인터페이스 또는 어댑터(153)를 통해 LAN(151)에 연결된다. WAN 네트워킹 환경에서 사용될 때, 퍼스널 컴퓨터(120)는 일반적으로 모뎀(154) 또는 인터넷 등의 원거리 통신망(153)을 통해 통신을 설정하는 다른 수단을 포함한다. 내장형 또는 외장형일 수 있는 모뎀(154)은 직렬 포트 인터페이스(146)를 통해 시스템 버스(123)에 연결된다. 네트워크화된 환경에서, 퍼스널 컴퓨터(120) 또는 그의 일부분과 관련하여 도시된 프로그램 모듈은 원격 메모리 저장 장치에 저장될 수 있다. 도시된 네트워크 연결은 전형적인 것이고 컴퓨터들 간에 통신 링크를 설정하는 다른 수단이 사용될 수 있음을 잘 알 것이다.

[0026] 데이터 분석 파이프라인

[0027] 전문 용어와 관련하여 예비 단계로서, 본 발명과 관련하여 이용될 수 있는 검색 엔진 등의 검색 엔진과 관련하여, 이 검색 엔진에 액세스하는 각 사용자는 아마도 부울 연산자와 함께 하나 이상의 검색어를 갖는 검색 문자열을 포함하는 쿼리를 입력함으로써 검색을 요청한다는 것을 잘 알 것이다. 이에 응답하여, 검색 엔진은 검색 문자열에 기초하여 일련의 검색 결과를 발생하고(이러한 검색 결과가 실제로 이용가능한 것으로 가정할 경우) 이러한 검색 결과를 요청측 사용자에게 반환한다. 반환된 검색 결과는 검색 요청과 관련있는 것으로 생각되는 콘텐츠 또는 결과의 특정 아이템을 포함할 수 있지만, 각각의 특정 결과는 그 대신에 검색 결과 내의 대응하는 링크를 통해 액세스될 경우가 더 많다.

[0028] 특히 반환된 검색 결과가 만족스럽지 못한 경우, 사용자는 새로운 검색 문자열이나 이전에 입력한 검색 문자열의 수정을 갖는 다른 쿼리 문자열을 입력하고 그에 따라 이 다른 쿼리에 기초하여 검색 엔진으로부터의 다른 일련의 검색 결과를 발생할 수 있다. 그러면, 일련의 관련 쿼리는 전체 검색 세션을 구성하며, 바람직하게는 이러한 세션은 사용자가 검색 결과에서 만족스러운 결과를 찾을 때 종료된다.

[0029] 사용자가 쿼리 문자열을 입력하고 검색 결과를 검토하는 동안, 검색 엔진 또는 관련 엔티티는 이러한 세션과 관련된 데이터를 식별 및 저장할 수 있고 종종은 그렇게 한다. 특히, 세션을 식별하는 것 이외에, 검색 엔진 또는 관련 엔티티(이후, "검색 분석기(search analyzer)"라 함)는 세션의 각각의 쿼리 문자열, 및 그 중에서도 특히 각각의 쿼리 문자열에 대해 반환된 검색 결과를 식별 및 저장할 수 있다. 게다가, 검색 분석기는 검색 결과의 각각의 반환된 결과의 각각의 링크에 대해 사용자가 그에 따라 결과에 액세스했는지 여부 및 그 중에서도 특히 사용자가 이러한 결과를 검토하는 데 얼마 동안 걸렸는지를 식별 및 저장할 수 있다. 따라서 잘

알고 있는 바와 같이, 검색 분석기 또는 다른 엔티티는 이러한 정보를 가지고서 사용자가 세션으로부터 반환된 검색 결과의 각 세트에 대해 얼마나 만족하는지, 즉 "흡족해하는지"의 정량적이 아닌 정성적인 척도를 개발할 수 있다.

[0030] 상기 기능을 수행하는 검색 분석기가 공지되어 있거나 당업자에게는 명백하며 따라서 상세히 기술하지 않는다는 것에 유의하기 바란다. 따라서, 본 발명과 관련하여 임의의 적절한 검색 분석기가 이용될 수 있다.

[0031] 세션을 구성하는 쿼리 문자열 그룹들의 일례로서, 이하의 쿼리 문자열 "자동차", "포드", "포드 엣셀", "저렴한 휴가" 및 "런던 여행 비용"을 생각해보자. 각각의 쿼리 문자열이 사용자가 검색을 할 때 입력한 실제 텍스트라는 것을 잘 알 것이다. 게다가, 각각의 쿼리 문자열은 결과에 링크되어 있는 일련의 검색 결과를 생성하고, 각 링크가 선택되었는지 여부, 결과를 보는 데 걸린 체류 시간(dwell time), 결과와 관련하여 취한 스크롤 및 다른 행동, 그리고 다른 유사한 사용자 거동 등의 검색 분석기 관련 데이터에 의해 검색 결과와 연관될 수 있다.

[0032] 상기 정보 모두에 기초해 볼 때, 사용자가 2가지 서로 다른 의도를 가진 것으로 보여지기 때문에, 처음 3개의 쿼리 문자열(즉, "자동차", "포드", "포드 엣셀")이 제1 세션의 부분이고, 마지막 2개의 쿼리 문자열(즉, "저렴한 휴가", "런던 여행 비용")이 제2 세션의 부분임을 알 것이다. 즉, 세션이 단일 사용자에 의해 시간상 연속적으로 행해진 동일한 의도 또는 목적을 갖는 쿼리 문자열 및 그의 대응하는 검색 결과의 그룹(즉, 사용자 검색)이라는 것을 잘 알 것이다.

[0033] 이제 도 2를 참조하면, 본 발명이 검색 요청자로부터의 요청된 검색을 수행하고 적절한 검색 결과로 그에 응답하는 검색 엔진(10) 등의 시스템을 가정하고 있음을 알 수 있다. 이러한 검색 엔진(10)은 물론 본 발명의 정신 및 범위를 벗어나지 않는 임의의 적절한 검색 엔진일 수 있다. 이러한 검색 엔진(10)의 동작은 일반적으로 공지되어 있거나 당업자에게는 명백한 것이며, 따라서 달리 언급하지 않는 한 여기에 상세히 기술할 필요는 없다. 보다 일반적으로, 이 시스템은 다시 말하면 본 발명의 정신 및 범위를 벗어나지 않는 검색 엔진(10) 이외의 시스템일 수 있다.

[0034] 도 2에서 알 수 있는 바와 같이, 데이터 웨어하우스(12)가 시스템/검색 엔진(10)과 관련하여 유지되고 있는 것으로 가정하고 있다. 잘 알고 있는 바와 같이, 데이터 웨어하우스(12)는 아마도 검색 분석기에 의해 생성된 상기 데이터(이에 한정되는 것은 아님)를 비롯하여 검색 엔진(10)의 동작과 관련한 데이터를 저장한다. 이러한 데이터는 본 발명의 정신 및 범위를 벗어나지 않고 데이터 웨어하우스(12)에 의해 저장된 것이거나 임의의 다른 적절한 데이터일 수도 있으며, 데이터 웨어하우스(12) 및 그의 구성도 마찬가지로 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 데이터 웨어하우스 및 구성일 수 있다.

[0035] 잘 알고 있는 바와 같이, 데이터 웨어하우스(12) 내의 데이터는 무한정으로 증가하거나 또는 주기적으로 만료 및 삭제될 수 있다. 따라서, 본 발명에서는, 파이프라인(14)이 이러한 데이터를 분석하는 데 이용되며, 이하에 보다 상세히 기술하는 바와 같이, 데이터의 전부 또는 데이터의 슬라이딩 윈도우를 비롯하여 데이터 웨어하우스(12) 내의 데이터의 일부분을 분석할 수 있다. 잘 알고 있는 바와 같이, 파이프라인(14)에 의해 수행되는 이러한 데이터 분석은 몇가지 컴포넌트를 포함할 수 있으며, 그 각각은 웨어하우스로부터 데이터의 일부분을 가져와 그에 대해 동작한다. 본 발명의 일 실시예에서, 파이프라인(14)의 출력은 라이브러리(16)에 배치되며, 출력 데이터는 관계 데이터베이스, 다차원 테이블 등등과 같은 형태로 조직화된다. 따라서, 적절한 콘트롤을 사용하여, 웹 브라우저 등에 있는 사용자는 이러한 조직화된 데이터를 보고 아마도 그 데이터를 추가로 분석할 수 있다.

[0036] 잘 알고 있는 바와 같이, 파이프라인(14)에 의해 수행되는 프로세스는 검색 엔진(10)이 어떻게 사용되고 있는지에 관한 통계를 제공한다. 게다가, 본 발명의 일 실시예에서, 이러한 수행된 프로세스는 검색 엔진(10)의 사용자가 이 검색 엔진(10)이 어떻게 이용되고 있는지의 패턴을 식별하는 데 사용되며, 이러한 패턴이 진정으로 가치있는 정보를 제공할 수 있는 한 특히 그렇다.

[0037] 웨어하우스(12) 내의 데이터가 고도로 정규화될 수 있고, 이러한 데이터의 정규화가 새로운 데이터를 효율적인 방식으로 부가될 수 있게 해준다는 것을 알 수 있는 한 특히 그러하다는 것에 유의한다. 그렇지만, 이러한 정규화된 데이터는 분석 이전에 역정규화(de-normalize)되어야만 하며, 따라서, 데이터 파이프라인(14)은 그의 입력으로서 역정규화기(de-normalizer)(18)를 포함할 수 있다. 이러한 역정규화기(18)는 임의의 적절한 역정규화기일 수 있으며, 임의의 적절한 방식으로 동작할 수 있다. 이러한 역정규화기(18)의 동작은 공지되어 있거나 당업자에게는 명백하며, 따라서 여기에서 상세히 설명할 필요는 없다.

[0038] 본 발명의 파이프라인(14)에서 사용되는 역정규화기(18)는 웨어하우스(12) 내의 데이터를 역정규화하고, 또한 나중의 처리를 위해 역정규화된 데이터의 배치(batch)를 생성할 수 있다. 일괄 처리(batching)는 임의의 적절한 기준에 따라 수행될 수 있다. 예를 들어, 배치(batch)는 데이터의 이전의 배치(batch) 이래로 도착한 모든 데이터일 수 있으며, 여기서 배치(batch)는 역정규화기(18)가 매일 한번씩 동작되는 경우 하루의 데이터일 수 있다. 이와 유사하게, 역정규화기(18)는 매주 한번씩 동작될 수 있지만, 일주일의 매 시간마다 배치(batch)를 발생하도록 설정될 수 있다.

[0039] 알고 있는 바와 같이, 역정규화기(18)에 의해 출력된 배치(batch)는 파이프라인(14)의 다른 컴포넌트들에 의한 추가의 처리를 위해 데이터 스토어(20)에 저장된다. 이러한 데이터 스토어(20) 및 그 안에 있는 배치(batch)의 구성은 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 스토어 및 구성일 수 있다. 도 2에 도시한 바와 같이, 데이터 스토어(20) 내의 데이터의 배치(batch)를 이용하는 컴포넌트들은 사용자 검색 번들(USB), 관련성 프로세서(Relevance Processor, RP), 의도 프로세서(IP), 데이터 합성기(DS) 및 기타 등등을 포함할 수 있다.

[0040] 데이터 파이프라인(14)의 사용자 검색 번들러(USB)는 사용자 검색을 분석하는데, 각각의 사용자 검색은 사용자가 검색 엔진(10)에 대해 쿼리를 호출하는 것을 말하며, 그 중에서도 특히 쿼리 문자열 및 반환된 검색 결과 둘다에 관한 정보를 포함한다. 2명의 사용자가 동일한 쿼리 문자열을 호출하면 2개의 사용자 검색을 생성하고, 단일 사용자가 동일한 쿼리 문자열을 두번 호출해도 역시 2개의 사용자 검색이 얻어진다. 본 발명의 일 실시예에서, USB는 "비슷하게 보이는(look alike)" 따라서 유사한 사용자 검색들을 찾으려고 하며, 이러한 유사한 사용자 검색들을 사용자 검색 번들로 그룹화한다. USB는 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 방식으로 동작할 수 있다. USB를 동작시키는 방법은 공지되어 있거나 당업자에게 자명한 것이며, 따라서 여기에 상세히 기술할 필요는 없다.

[0041] 예를 들어, USB는 쿼리 문자열을 분석하는 "Look Alike" 알고리즘을 사용할 수 있다. 예를 들어, 2명의 사용자가 각각 "Wilkes-Barre, PA"에 대한 검색을 호출하는 경우, 이들 쿼리 문자열은 비슷하게 보인다(look alike). 보다 흥미로운 것은, 한명의 사용자가 "duck"에 대해 검색하고 또한명의 사용자가 "ducks"에 대해 검색하는 경우(하나는 단수, 하나는 복수), 알고리즘이 스테밍(stemming), 대소문자 정규화(case normalization), 및 기타 유사한 쿼리 압축 기술을 사용하는 한, 이들 2개의 쿼리 문자열은 아주 "비슷하게 보인다(look alike)".

[0042] 그렇지만, USB는 유사성이 있는지 쿼리 문자열을 분석하는 것에 한정되지 않는다. 그 대신에, 본 발명의 일 실시예에서, USB는 또한 반환된 검색 결과 분석 알고리즘을 이용할 수 있다. 예를 들어, USB는 반환된 검색 결과 분석 알고리즘에 기초하여 대응하는 쿼리 문자열이 전혀 유사하지 않음에도 2개의 사용자 검색이 유사하고 그 각각이 유사한 일련의 검색 결과를 생성하는 것으로 결론내릴 수 있다. 이에 부가하여 또는 다른 대안으로서, USB는 콘텐츠 만족도 분석 알고리즘에 기초하여 각각의 사용자 검색이 사용자가 만족스러운 것으로 판정한 특정의 콘텐츠 또는 결과를 생성함에도, 다시 말하면 대응하는 쿼리 문자열이 전혀 유사하지 않음에도 2개의 사용자 검색이 유사하다고 결론내릴 수 있다. 이와 마찬가지로, USB는 콘텐츠 불만족도 분석 알고리즘에 기초하여 각각의 사용자 검색이 사용자가 불만족스러운 것으로 판정한 특정의 결과를 생성함에도 2개의 사용자 검색이 유사한 것으로 결론내릴 수 있다.

[0043] USB의 출력은 번들화된 사용자 검색의 데이터베이스로서 라이브러리(16)에 저장될 수 있다. 다른 대안에서, 이러한 출력은 파이프라인(14)의 다른 컴포넌트들이 이용가능하도록 데이터 스토어(20)에 다시 저장될 수 있다.

[0044] 본 발명의 관련성 프로세서(RP)는 인자 발생기(FG), 관련성 분류기(RC), 및 메트릭 발생기(MG)로 세분될 수 있다. 그렇지만, 이러한 RP의 세분된 구성요소들은 그 대신에 RP 아래에 그룹화되어 있지 않고 본 발명의 파이프라인(14)에 개별적으로 존재할 수 있음에 유의해야 한다.

[0045] 인자 발생기(FG)는 사용자 검색과 관련한 정보를 고려하여 인자를 생성하며, 여기서 이러한 인자는 관련성 분류기(RC)에 입력으로서 인가된다. 각각의 인자는 일련의 검색 결과로부터의 특정의 결과에 대한 것이다. 인자들은 종종 몇가지 이벤트를 고려하여 생성되지만 항상 그러한 것은 아니며, 여기서 각각의 이벤트는 일반적으로 질의하는 사용자가 수행한 행동에 관한 정보이다. 예를 들어, 특정의 결과를 선택하는 것은 이벤트일 수 있으며, 그 결과가 디스플레이되는 것, 닫히는 것, 인쇄되는 것, 특정의 리스트에 부가되는 것, 다시열기(re-open)되는 것, 및 기타 등등도 이벤트일 수 있다. 이벤트는 또한 명시적인 사용자 피드백이 제공되는 경우 이를 포함할 수 있다.

[0046] 본 발명의 일 실시예에서, FG는 이벤트에 기초하여 값들을 계산한다. 예를 들어, FG는 "체류 시간(Dwell Time)"을, 사용자가 결과를 본 시간 길이를 나타내는 인자로서 계산할 수 있고, 사용자가 결과를 보기 시작한 때와 종료한 때[이 각각은 대응하는 타임스탬프드 이벤트(timestamped Event)로 나타내어짐] 사이의 시간차에 기초하여 그와같이 할 수도 있다. 그렇지만, 다른 적절한 타임스탬프드 이벤트에 의해 나타내어지는 바와 같이 사용자가 결과를 보는 것을 중단했을지 여부를 비롯하여, 체류 시간 인자를 발생할 때 다른 이벤트들도 역시 FG에 의해 고려될 수 있음에 유의해야 한다.

[0047] FG에 의해 발생되는 또하나의 인자는 사용자가 즐겨찾기 리스트 등의 특정의 리스트에 결과를 부가했는지 여부에 대한 인자일 수 있다. 이러한 경우, 이러한 "즐겨찾기" 인자는 TRUE 값으로 설정될 수 있다. 이제 잘 알 수 있는 바와 같이, 이벤트에 기초한 여러 타입의 인자들은 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 인자일 수 있다. 이러한 인자들은 공지되어 있거나 당업자에게는 자명한 것이며, 따라서 여기에서 상세히 설명할 필요가 없다.

[0048] 인자들은 또한 이벤트 이외의 다른 것들로부터 나올 수 있음에 유의해야 한다. 예를 들어, 인자는 결과를 작성한 사용자의 식별자일 수 있다. 따라서, 인자는 결과의 프로퍼티일 수 있다. 다른 인자들은 쿼리 문자열의 로케일(예를 들어, 미국 영어, 캐나다 영어, 브라질 스페인어)을 포함할 수 있고, 사용자 검색 내의 정보로부터 도출될 수도 있다. 보다 일반적으로 말하면, 본 발명의 정신 및 범위를 벗어나지 않고 인자는 이용가능한 임의의 정보로부터 FG에 의해 또는 다른 컴포넌트에 의해 생성될 수 있다.

[0049] 인자는 FG에 의해 발생되며, 아마도 다른 경우에는 적절한 데이터베이스 내의 라이브러리(16)에 저장될 수 있다. 다른 대안에서, 이러한 인자는 파이프라인(14)의 다른 컴포넌트들이 이용가능하도록 다시 데이터 스토어(20)에 저장될 수 있거나 이러한 컴포넌트들로 직접 전달될 수 있다.

[0050] 관련성 프로세서(RP)의 관련성 분류기(RC)는 FG에 의해 또한 아마도 다른 곳에서 발생된 인자들을 수신한다. RC는 그에의 입력으로서 수신된 인자들에 기초하여 동작하는 기계 발생 결정 트리(machine-generated decision tree)이다. RP에 의해 개시될 때, RC는 데이터 스토어(20) 또는 다른 곳으로부터 결과에 대한 인자를 판독하고 그 결과에 대한 판정을 생성한다. 일반적으로, 이러한 판정은 일련의 검색 결과로부터의 결과에 액세스하려고 결정할 때 사용자가 그 동일 결과를 어떻게 판단했는지의 결정이다.

[0051] 판정은 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 판정 시스템에 따라 표현될 수 있다. 예를 들어, 판정은 숫자 또는 문자 등급 점수(grade score)일 수 있고, "수락(Accept)"(즉, 사용자가 결과에 만족함), "탐색(Explore)"(즉, 사용자가 결과에 만족하지도 불만족하지도 않음), 및 "거부(Reject)"(즉, 사용자가 결과에 불만족함) 중 하나 또는 기타 등등일 수 있다. 게다가, 특정 결과에 대한 판정은 또한 RC에 의해 결정되는 바와 같이 판정이 옳을 가능성이 얼마나 되는지에 대한 확신(confidence)을 나타내는 값을 포함할 수 있다.

[0052] RC에 의해 및 아마도 다른 곳에서 발생되는 판정은 적절한 데이터베이스 내의 라이브러리(16)에 저장될 수 있다. 다른 대안에서, 이러한 인자들은 파이프라인(14)의 다른 컴포넌트들이 이용가능하도록 데이터 스토어(20)에 다시 저장될 수 있거나 이러한 컴포넌트들로 직접 전달될 수 있다.

[0053] RC의 결정 트리를 기계 발생하기 위해, RP가 명시적인 판정 인자를 갖는 이러한 결과를 고려할 수 있다. 명시적인 판정이란 결과에 관하여 사용자로부터의 명시적인 피드백을 나타내는 일종의 인자이다. 명시적인 판정 인자를 갖는 각각의 결과를 취하고 그 결과에 대한 다른 이용가능한 인자들을 분석함으로써, RP 또는 다른 곳의 관련성 분류기 훈련기(Relevance Classifier Trainer, RCT)는 어떤 인자가 어느 판정을 암시하는지를 "학습"할 수 있고 그에 기초하여 RC를 구축할 수 있다. 인자들이 무엇을 암시하는지를 학습하고 그로부터 RC를 구축하는 것은 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 방식으로 행해질 수 있다. 그와 같이 하는 것은 공지되어 있거나 당업자에게 자명한 것이며, 따라서 여기에 상세히 설명할 필요는 없다.

[0054] RP의 메트릭 발생기(MG)는 메트릭을 생성하고, 여기서 메트릭은 결과, 사용자 검색, 세션 또는 기타 등등에 대한 척도이다. 일반적으로, MG는 FG에 의해 및 아마도 다른 곳에서 발생된 인자들, RC에 의해 생성된 판정, 및 이용가능한 다른 관련 정보에 기초하여 이러한 메트릭을 생성한다. 이러한 메트릭을 생성하는 것은 본 발명의 정신 및 범위를 벗어나지 않고 임의의 적절한 방식으로 행해질 수 있다. 그와 같이 하는 것은 공지되어 있거나 이하에 기술하는 정보로부터 당업자에게는 자명한 것이며, 따라서 여기에 상세히 기술할 필요는 없다.

[0055] MG는 결과에 대해 이하의 메트릭을 생성할 수 있다.

[0056] - 수락, 탐색, 거부. 각각은 결과에 대한 판정으로부터 도출되며, 확신을 포함할 수 있다.

[0057] - 오정렬된 결과(Mis-ranked Result). 결과가 검색 결과 내에서 어떻게 배치되어 있는가 및 어떻게 사용자가 결과를 정렬한 것으로 판정되는가에 기초하여 결과가 있어야만 하는 곳으로부터 얼마나 "멀리" 있는지의 척도.

[0058] - 위치. 어떻게 사용자가 결과를 정렬한 것으로 판정되는가.

[0059] - 관련성 위치. 결과가 검색 결과 내에서 어떻게 배치되어 있는가.

[0060] MG는 사용자 검색에 대해 이하의 메트릭을 생성할 수 있다.

[0061] - 결과 세트 정렬 점수(Result Set Ranking Score). 각 결과에 대해 생성된 오정렬된 결과 메트릭과 유사하지만, 사용자 검색의 모든 결과에 대한 것임. 이러한 값은 결과 세트가 있어야 하는 곳으로부터 얼마나 멀리 떨어져 있는가를 반영하려고 한다.

[0062] - 요약된 수락, 탐색, 거부. 사용자 검색에서의 각각의 결과에 대해 각각 모든 수락, 탐색 및 거부 메트릭에 대한 요약.

[0063] MG는 세션에 대해 이하의 메트릭을 생성할 수 있다.

[0064] - 콘텐츠 양. 세션이 사용자가 무엇을 검색하고 있는지에 대해 이용가능한 정보가 없음을 나타내는지 여부.

[0065] - 의도 결정. 세션이 사용자가 무엇을 검색하고 있는지를 결정할 수 없음을 나타내는지 여부.

[0066] MG에 의해 또한 아마도 다른 곳에서 생성된 메트릭은 적절한 데이터베이스 내의 라이브러리(16)에 저장될 수 있다. 다른 대안에서, 이러한 메트릭은 파이프라인(14)의 다른 컴포넌트들이 이용가능하도록 하기 위해 데이터 스토어(20)에 다시 저장될 수 있거나 이러한 컴포넌트들로 직접 전달될 수 있다.

[0067] 파이프라인(14)의 의도 프로세서(IP)는 의도를 생성하며, 여기서 사용자가 각 경우에서 동일한 검색 결과를 찾고 있기 때문에, 각각의 의도는 서로 관련되어 있는 것으로 생각되는 하나 이상의 세션들의 그룹이다. 즉, 의도는 공통 목적을 공유하는 세션들의 그룹이다.

[0068] IP는 각각의 세션, 각각의 세션의 각각의 사용자 검색, 및 각각의 사용자 검색의 각각의 결과를 고려함으로써 의도를 생성한다. 본 발명의 일 실시예에서, 공통 결과를 갖는 세션들이 관련되어 있을 가능성이 있음을 염두에 두면, 수락의 판정을 갖는 결과들만이 조사된다. 그렇지만, 본 발명의 정신 및 범위를 벗어나지 않고 다른 판단을 갖는 결과들도 역시 이용될 수 있다.

[0069] 일반적으로, IP는 검토된 사용자 검색 번들에 기초하여 세션들에 걸친 공통 질의어 및 세션들에 걸친 공통 결과를 찾아냄으로써 세션들 간의 관계값(relationship value)을 결정하고, 발견된 경우 공통성의 세기(Strength of Commonality)를 확인한다. 이러한 세기는 2개의 세션이 공통 목적을 가짐으로써 서로 관련될 가능성이 얼마나 되는지를 나타낸다. 이어서, 어떤 정해진 임계값을 넘는 세기를 갖는 세션쌍은 IP에 의해 의도로 번들화될 수 있다.

[0070] IP에 의해 또한 아마도 다른 곳에서 생성된 의도는 적절한 데이터베이스 내의 라이브러리(16)에 저장될 수 있다. 다른 대안에서, 이러한 의도는 파이프라인(14)의 다른 컴포넌트들이 이용가능하도록 하기 위해 데이터 스토어(20)에 다시 저장될 수 있거나 이러한 컴포넌트들로 직접 전달될 수 있다.

[0071] 마지막으로, 지금까지 생성된 모든 데이터를 가지고서, 파이프라인(14)은 데이터 스토어(20), 라이브러리(16) 또는 다른 곳로부터 이러한 데이터를 추출하고 이러한 데이터를 라이브러리(20)에 저장되어야 하는 하나 이상의 데이터베이스로 포맷화하는 데이터 합성기(DS)를 포함한다. 이러한 포맷화는 특히 서로 다른 컴포넌트가 데이터를 서로 다른 테이블, 데이터베이스 또는 기타 등등에 기록한 경우에 필요하다. 예를 들어, USB는 각각의 사용자 검색에 관한 데이터를 사용자 검색 번들화된 테이블에 부가했을 수 있고, MG는 데이터를 메트릭 테이블에 부가했을 수 있다. 따라서, DS는 이러한 테이블 및 기타의 것들을 적절한 경우 의미있는 형태로 결합하며, 그로부터 이러한 데이터가 검색되거나, 통합되거나, 제공되거나 기타 등등이 행해진다.

[0072] 유의할 점은 DS가 적절한 경우 데이터의 일부분만을 포맷화하도록 프로그램될 수 있다는 것이다. 예를 들어, DS는 저장된 데이터의 일부가 유용하지 않은 것으로 생각되는 경우 이러한 데이터를 누락시킬 수 있다. 다른 대안에서, DS는 동일한 데이터를 다수의 테이블로 복사하거나 데이터를 다수의 테이블로 분할할 수 있다. 보다 일반적으로 말하면, 본 발명의 정신 및 범위를 벗어나지 않고 DS는 파이프라인(14)으로부터의 데이터를 임의의 적절한 포맷으로 포맷화하도록 프로그램될 수 있다.

[0073] 본 발명의 일 실시예에서, DS는 데이터를 SQL 데이터베이스 등과 같은 관계 데이터베이스로 포맷화한다. 그 자체로서, 데이터는 사실 테이블(fact table)이 중앙에 있고 차원 테이블(dimension table)이 이를 둘러싸고 있는 "별" 형태로 나타내어질 수 있다. 잘 알고 있는 바와 같이, 그와 같이 함으로써, 사실 테이블 및 차원 테이블은 검색 엔진(10)의 관리자 등이 여러가지 조건들에 기초하여 데이터를 효과적으로 통합할 수 있게 해주는 OLAP 큐브 등과 같은 데이터 큐브(data cube)로 구축될 수 있다. 예를 들어, 이러한 관리자 등은 어떤 데이터 범위에 존재하는 모든 세션에 대한 의도 결정 메트릭의 평균을 구하도록 큐브에 요청하고 그를 의도별로 통합할 수 있다. 이와 마찬가지로, 큐브에 대해 리포트가 실행될 수 있고, 이어서 그로부터의 결과가 이러한 관리자 등에 전달될 수 있다. 물론, 이러한 관리자 등은 이러한 큐브에 대해 임의의 다른 적절한 쿼리를 배치하거나 이러한 큐브에 대해 실행되는 임의의 다른 리포트를 수신할 수 있다.

발명의 효과

[0074] 본 발명은 검색 엔진(10) 또는 기타를 포함하는 임의의 적절한 시스템으로부터의 데이터를 분석하는 것과 관련하여 실시될 수 있다. 이제 잘 알고 있는 바와 같이, 여기 기술된 본 발명에서 시스템을 표현하는 데이터는, 응답을 개선하기 위해 시스템을 조정 또는 "튜닝"하는 것, 시스템 성능을 유지하는 것, 및 다른 방식으로 시스템이 만족스럽게 동작하도록 보장하는 것을 비롯한 어떤 목적을 위해서도 분석될 수 있다.

[0075] 본 발명과 관련하여 수행되는 프로세스를 실시하는 데 필요한 프로그래밍은 비교적 간단하며 관련 프로그래밍 당업자에게는 자명할 것이다. 따라서, 이러한 프로그래밍은 여기에 첨부하지 않는다. 본 발명의 정신 및 범위를 벗어나지 않고 본 발명을 실시하는 데 임의의 특정 프로그래밍이 이용될 수 있다.

[0076] 상기 설명에서, 본 발명이 대량의 데이터, 특히 검색 엔진(10) 등으로부터의 대량의 데이터를 분석하는 새롭고 유용한 장치를 포함한다는 것을 알 수 있다. 이러한 장치는 검색 엔진(10)이 특히 대규모이고 많은 트래픽 등을 가지며, 분석에 이용가능한 이러한 데이터의 양이 특히 많고 따라서 다루기 힘들고 또 어려운 때에 특히 유용하다.

[0077] 본 발명의 발명 개념을 벗어나지 않고 상기 기술한 실시예들에 여러 변경이 행해질 수 있음을 잘 알 것이다. 일반적으로 말하면, 따라서 본 발명이 개시된 특정 실시예들에 한정되지 않고 첨부된 청구항들에 의해 정의된 본 발명의 정신 및 범위 내의 여러 수정들을 포함한다는 것을 잘 알 것이다.

도면의 간단한 설명

[0001] 도 1은 본 발명의 양태 및/또는 그의 일부분이 포함될 수 있는 범용 컴퓨터 시스템을 나타내는 블록도.

[0002] 도 2는 본 발명의 실시예들에 따라 검색 엔진 등의 대규모 시스템으로부터의 데이터를 분석하는 데이터 분석 파이프라인을 도시하는 블록도.

[0003] <도면의 주요 부분에 대한 부호의 설명>

[0004] 10: 검색 엔진

[0005] 12: 데이터 웨어하우스

[0006] 14: 파이프라인

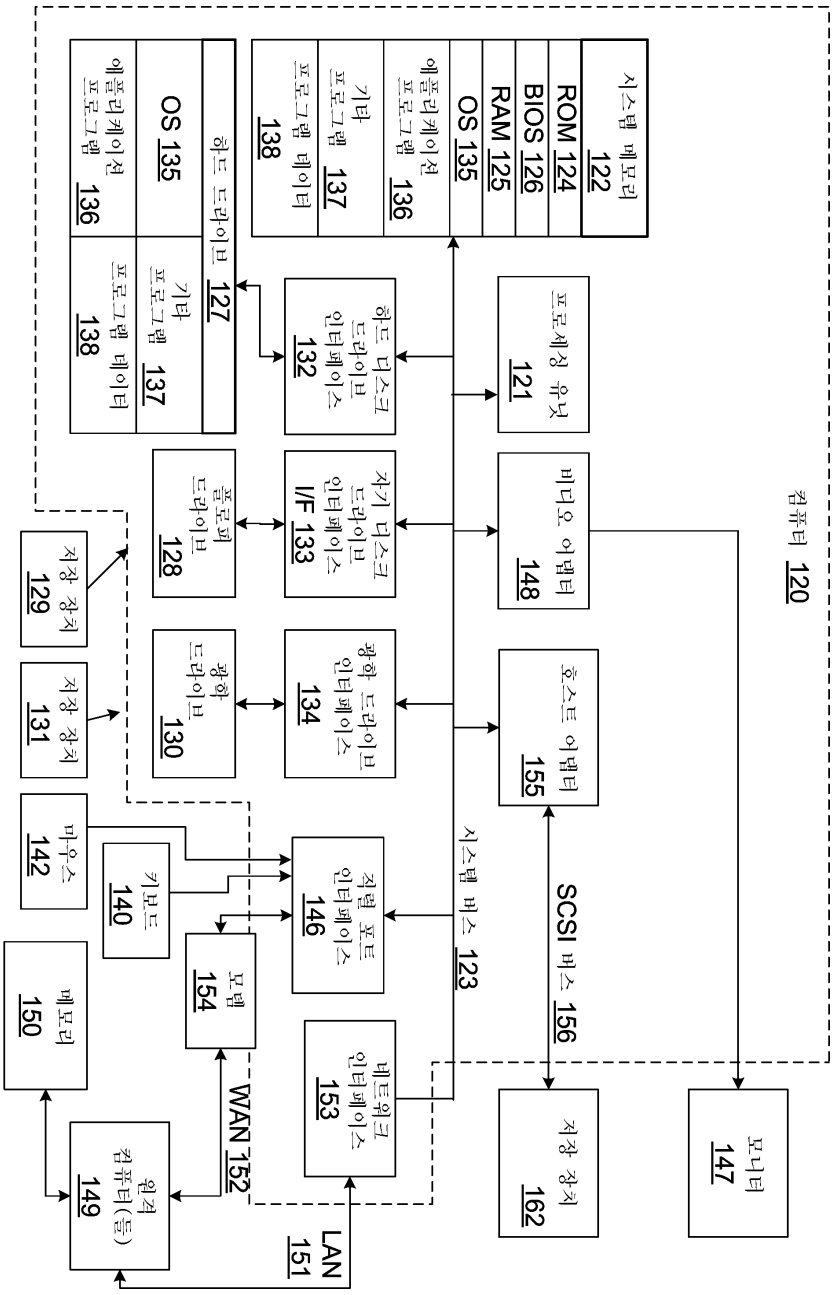
[0007] 16: 라이브러리

[0008] 18: 역 정규화기

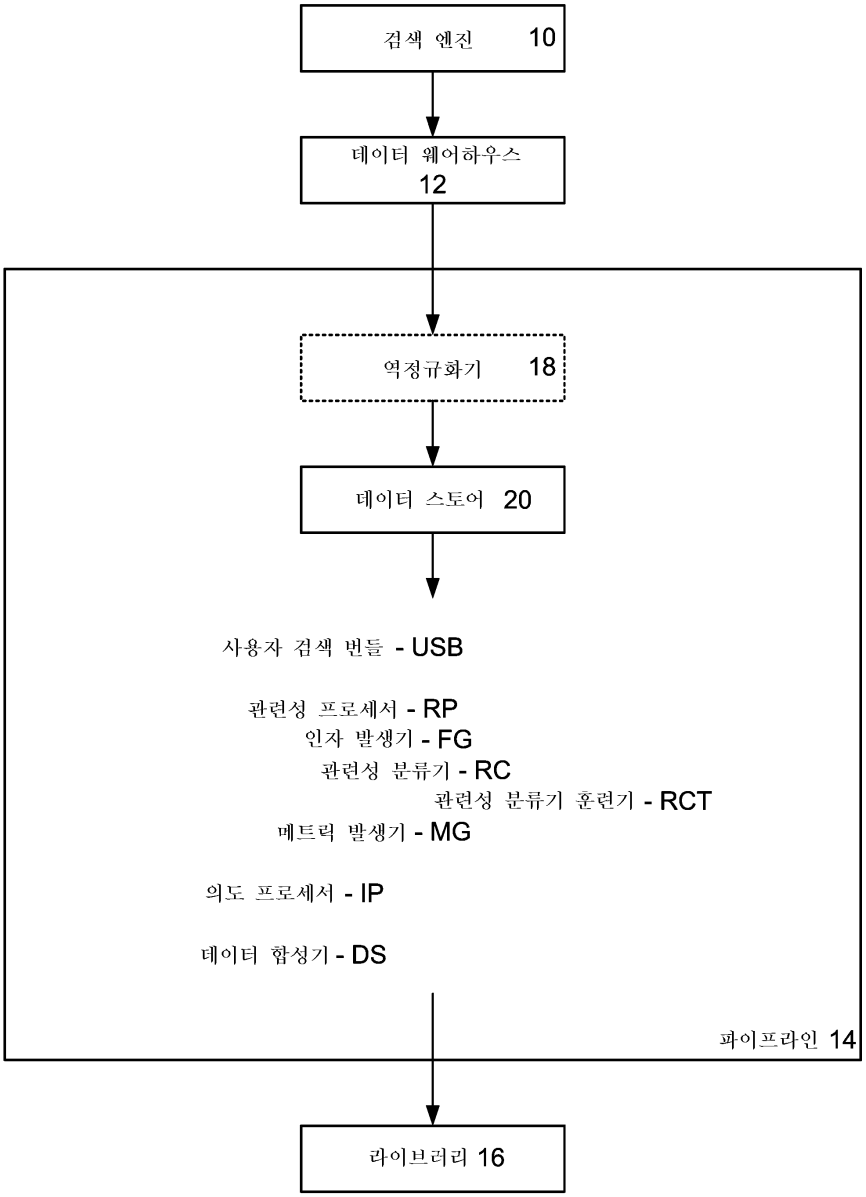
[0009] 20: 데이터 스토어

도면

도면1


컴퓨터 120
시스템 메모리 122
ROM 124
BIOS 126
RAM 125
OS 135
애플리케이션 프로그램 136
기타 프로그램 137
프로그램 데이터 138
하드 드라이브 127
OS 135
애플리케이션 프로그램 136
기타 프로그램 137
프로그램 데이터 138
하드 디스크 드라이브 인터페이스 132
프로세싱 유닛 121
비디오 어댑터 148
자기 디스크 드라이브 인터페이스 I/F 133
플로피 드라이브 128
저장 장치 129
광학 드라이브 인터페이스 134
광학 드라이브 130
저장 장치 131
호스트 어댑터 155
시스템 버스 123
SCSI 버스 156
저장 장치 162
모니터 147
직렬 포트 인터페이스 146
마우스 142
키보드 140
모뎀 154
네트워크 인터페이스 153
메모리 150
WAN 152
원격 컴퓨터(들) 149
LAN 151

도면2


검색 엔진 10
데이터 웨어하우스 12
역정규화기 18
데이터 스토어 20
사용자 검색 번들 - USB
관련성 프로세서 - RP
인자 발생기 - FG
관련성 분류기 - RC
관련성 분류기 훈련기 - RCT
메트릭 발생기 - MG
의도 프로세서 - IP
데이터 합성기 - DS
파이프라인 14
라이브러리 16