(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년03월29일
(11) 등록번호 10-1721338
(24) 등록일자 2017년03월23일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30864 (2013.01)
G06F 17/30646 (2013.01)
(21) 출원번호 10-2015-0169288
(22) 출원일자 2015년11월30일
심사청구일자 2015년11월30일
(65) 공개번호 10-2016-0149978
(43) 공개일자 2016년12월28일
(30) 우선권주장
201510342427.4 2015년06월18일 중국(CN)
(56) 선행기술조사문헌
KR1020150010740 A*
(뒷면에 계속)
전체 청구항 수 : 총 20 항

(73) 특허권자
바이두 온라인 네트웍 테크놀러지 (베이징) 캄파니 리미티드
중국 베이징 하이뎬 디스트릭트 샹디 10번가 10 바이두 캠퍼스 3층
(72) 발명자
팡, 가오린
중국, 베이징 100085, 샹디 10번 스트리트 하이뎬 디스트릭트, 넘버 10, 바이두 캠퍼스 3층
(74) 대리인
손민

심사관 : 박승철

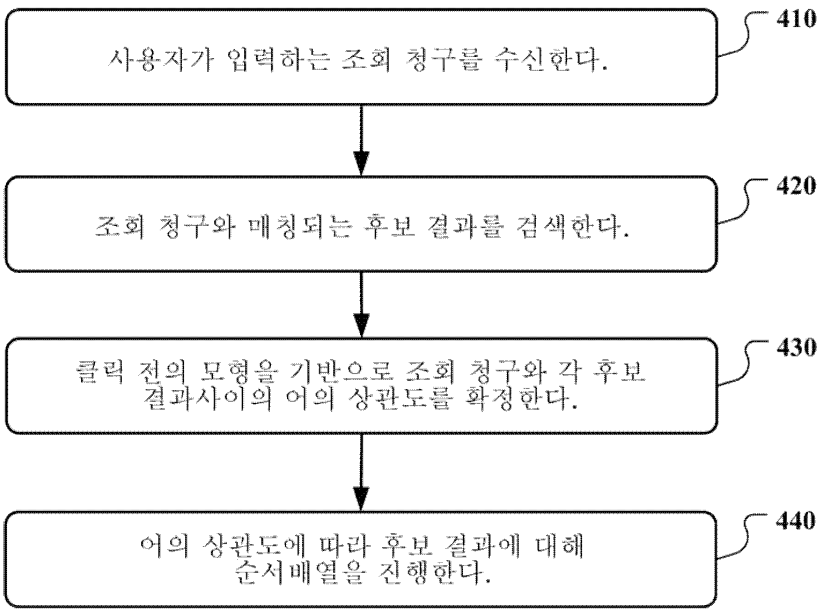
(54) 발명의 명칭 **검색 엔진 및 그의 구현 방법**

(57) 요 약

본 발명은 검색 엔진 및 그의 구현 방법을 개시한다. 검색 엔진의 구현 방법은, 사용자가 입력한 조회 청구를 수신하는 단계; 조회 청구와 매칭되는 후보 결과를 획득하는 단계; 클릭 전의 모형을 기반으로 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 단계; 및 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하는 단

(뒷면에 계속)

대 표 도 - 도4



계;를 포함하되, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하고, 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함한다. 본 발명의 기술적 방안에 의하면, 어의 상관도에 따라 검색된 후보 결과에 대해 순서배열을 진행여 검색 결과의 순서배열 효과를 향상 시키고 검색 결과 리스트의 전열에 사용자의 검색 수요에 부함되지 않는 검색 결과가 발생하는 것을 피면할 수 있으므로 사용자로 하여금 양호한 사용 체험을 가지도록 확보한다.

(52) CPC특허분류
G06F 17/30991 (2013.01)

(56) 선행기술조사문헌
KR1020130116330 A
KR1020100101621 A
JP2014512600 A
KR1020150010740 A*
*는 심사관에 의하여 인용된 문헌

명 세 서

청구범위

청구항 1

사용자가 입력한 조회 청구를 수신하는 단계;

상기 조회 청구와 매칭되는 후보 결과를 검색하는 단계;

클릭 전의 모형을 기반으로 상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 단계; 및

상기 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하는 단계;를 포함하되,

상기 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 상기 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하며, 상기 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하되,

상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 단계는,

각 후보 결과에 대하여, 상기 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정하는 단계; 및

확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 상기 조회 청구와 상기 후보 결과사이의 어의 상관도를 확정하는 단계;를 포함하되, 상기 어구는 후보 결과의 타이틀, 앵커 문구 및 본문 중의 핵심 문장 중 적어도 하나를 포함하는 검색 엔진의 구현 방법.

청구항 2

제 1 항에 있어서,

상기 조회 청구와 후보 결과의 어구사이의 어의 상관도를 확정하는 단계는,

상기 클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 상기 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 단계;

상기 조회 청구와 후보 결과의 어구사이의 매칭 정황에 따라 전의 인자를 확정하는 단계; 및

상기 전의 인자와 상기 주제 매칭 유사도를 기반으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출하는 단계;를 포함하는 검색 엔진의 구현 방법.

청구항 3

제 2 항에 있어서,

상기 클릭 전의 모형을 기반으로 상기 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 단계는,

단어 정렬을 이용하여 후보 결과의 어구로부터 상기 조회 청구 중의 단어와 정렬되는 인접한 상문과 하문을 확정하는 단계;

상기 전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구 중의 상응한 상문과 하문의 유사도 가중치를 조정하는 단계; 및

조정된 유사도 가중치에 따라 문장사이의 문구 주제 매칭 모형을 이용하여 상기 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 단계;를 포함하는 검색 엔진의 구현 방법.

청구항 4

제 3 항에 있어서,

전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구 중의 상응한 상문과 하문의 유사도 가중치를 조정하는

단계는,

비전의 사전에 후보 결과의 어구 중의 상응한 단어 및 그의 상문 또는 하문이 포함될 경우, 상기 상문 또는 하문의 유사도 가중치를 낮추는 단계; 및

전의 사전에 후보 결과의 어구 중의 상응한 단어 및 그의 상문 또는 하문이 포함될 경우, 상기 상문 또는 하문의 유사도 가중치를 높히는 단계;를 포함하는 검색 엔진의 구현 방법.

청구항 5

제 3 항에 있어서,

상기 문장사이의 문구 주제 매칭 모형은 벡터 공간 모형

$$Sim(Q,S)=\frac{\sum_{w_{1k}=w_{2l}}(Wgt(w_{1k})*Wgt(w_{2l}))}{\sqrt{\sum_{k=1\cdots M}Wgt(w_{1k})^2}\sqrt{\sum_{l=1\cdots N}Wgt(w_{2l})^2}}*SentType(Q,S)$$ 이고,

여기서, $Sim(Q,S)$ 은 Q와 S사이의 주제 매칭 유사도를 표시하고, Q는 조회 청구를 표시하고, S는 후보 결과의 어구를 표시하고, $SentType(Q,S)$ 는 두개의 문장 유형이 매칭되는 가중치 계수를 표시하고, $Wgt(w_{1k})$ 는 조회 청구로부터 획득한 단어 w_{1k} 의 유사도 가중치를 표시하고, M는 단어 w_{1k}의 수량이고, $Wgt(w_{2l})$는 후보 결과의 어구로부터 획득한 단어 w_{2l} 의 유사도 가중치이고, N는 단어 w_{2l}의 수량인 검색 엔진의 구현 방법.

청구항 6

제 2 항에 있어서,

상기 조회 청구와 후보 결과의 어구사이의 매칭 정황에 따라 전의 인자를 확정하는 단계는,

매칭 정황이 조회 청구 중 제일 중요한 단어가 후보 결과의 어구에 나타나지 않은 정황일 경우, 전의 인자를 제1값으로 확정하는 단계;

매칭 정황이 상하문의 매칭이 존재하는 정황일 경우, 전의 인자를 제2값으로 확정하는 단계; 및

매칭 정황이 상하문의 완전 매칭이 존재하지 않는 정황일 경우, 전의 인자를 제3값으로 확정하는 단계;를 포함하되,

상기 제1값은 제2값보다 작고, 상기 제2값은 제3값보다 작은 검색 엔진의 구현 방법.

청구항 7

제 2 항에 있어서,

조회 청구와 후보 결과의 어구사이의 어의 상관도는 하기 등식:

$Rele(Q,S)=\beta(Q,S)Sim(Q,S)$ 에 따라 산출하되,

$Rele(Q,S)$ 는 Q와 S사이의 어의 상관도를 표시하고, $\beta(Q,S)$ 는 Q와 S사이의 전의 인자를 표시하고, $Sim(Q,S)$ 는 Q와 S사이의 주제 매칭 유사도를 표시하고, Q는 조회 청구를 표시하고, S는 후보 결과의 어구를 표시하는 검색 엔진의 구현 방법.

청구항 8

제 1 항에 있어서,

상기 클릭 전의 모형 중의 전의 사전와 비전의 사전은 조회 청구와 검색 결과Query-Title 쌍의 클릭 회수를 학습하여 구축하는 검색 엔진의 구현 방법.

청구항 9

제 8 항에 있어서,

상기 전의 사전과 비전의 사전은,

Query-Title 쌍의 클릭 표시 비율을 획득하고, 단어 정렬을 이용하여 검색 결과에서 조회 어구 중의 단어와 정렬되는 인접한 상하문를 획득하고, 클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 전의 사전에 첨가하며, 클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 비전의 사전에 첨가하는 방법으로 구축되는 원생 전의 사전과 원생 비전의 사전를 포함하되,

상기 클릭 표시 비율은 클릭 회수와 표시 회수의 비율이고, 표시 회수는 검색 결과가 조회 청구에 응하여 표시되는 회수를 지시하고, 클릭 회수는 검색 결과가 조회 청구에 응하여 표시될 때 사용자에 의해 클릭되는 회수를 지시하는 검색 엔진의 구현 방법.

청구항 10

제 9 항에 있어서,

상기 전의 사전과 비전의 사전은,

조회 청구 중의 단어에 대해 어의 유형을 표기하고, 표기된 어의 유형을 이용하여 원생 전의 사전과 원생 비전의 사전에 대응되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 구축하는 방법으로 구축되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 더 포함하는 검색 엔진의 구현 방법.

청구항 11

제 3 항 또는 제 9 항에 있어서,

상기 단어 정렬은 동의어 정렬을 포함하는 검색 엔진의 구현 방법.

청구항 12

사용자가 입력한 조회 청구를 수신하는 수신 유닛;

상기 조회 청구와 매칭되는 후보 결과를 검색하는 검색 유닛;

클릭 전의 모형을 기반으로 상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 어의 상관도 확정 유닛;

상기 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하는 순서배열 유닛;

후보 결과에 대하여 상기 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정하되, 상기 어구는 후보 결과의 타이틀, 앵커 문구 및 본문 중의 핵심 문장 중 적어도 하나를 포함하는 산출 유닛; 및

확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 상기 조회 청구와 상기 후보 결과사이의 어의 상관도를 확정하는 확정 유닛;을 포함하되,

상기 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 상기 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하고, 상기 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하는 검색 엔진.

청구항 13

제 12 항에 있어서,

상기 산출 유닛은,

상기 클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 상기 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 주제 매칭 유사도 모듈;

상기 조회 청구와 후보 결과의 어구사이의 매칭 정황에 따라 전의 인자를 확정하는 전의 인자 모듈; 및

상기 전의 인자와 상기 주제 매칭 유사도를 기반으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출하는 합성 모듈;을 포함하는 검색 엔진.

청구항 14

제 13 항에 있어서,

상기 주제 매칭 유사도 모듈은,

단어 정렬을 이용하여 후보 결과의 어구로부터 상기 조회 청구 중의 단어와 정렬되는 인접한 상문과 하문을 확정하고,

상기 전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구 중의 상응한 상문과 하문의 유사도 가중치를 조정하며,

조정된 유사도 가중치에 따라 어구사이의 문구 주제 매칭 모형을 이용하여 상기 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 검색 엔진.

청구항 15

제 13 항에 있어서,

상기 전의 인자 모듈은,

매칭 정황이 조회 청구 중 제일 중요한 단어가 후보 결과의 어구에 나타나지 않은 정황일 경우, 전의 인자를 제1값으로 확정하고,

매칭 정황이 상하문의 매칭이 존재하는 정황일 경우, 전의 인자를 제2값으로 확정하며,

매칭 정황이 상하문의 완전 매칭이 존재하지 않는 정황일 경우, 전의 인자를 제3값으로 확정하되,

상기 제1값은 제2값보다 작고, 상기 제2값은 제3값보다 작은 검색 엔진.

청구항 16

제 12 항에 있어서,

상기 클릭 전의 모형 중의 전의 사전과 비전의 사전은 조회 청구와 검색 결과Query-Title 쌍의 클릭 회수를 학습하여 구축하는 검색 엔진.

청구항 17

제 16 항에 있어서,

상기 전의 사전과 비전의 사전은,

Query-Title 쌍의 클릭 표시 비율을 획득하고, 단어 정렬을 이용하여 검색 결과에서 조회 어구 중의 단어와 정렬되는 인접한 상하문를 획득하고, 클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 전의 사전에 첨가하며, 클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 비전의 사전에 첨가하는 방법으로 구축되는 원생 전의 사전과 원생 비전의 사전를 포함하되,

상기 클릭 표시 비율은 클릭 회수와 표시 회수의 비율이고, 표시 회수는 검색 결과가 조회 청구에 응하여 표시되는 회수를 지시하고, 클릭 회수는 검색 결과가 조회 청구에 응하여 표시될 때 사용자에 의해 클릭되는 회수를 지시하는 검색 엔진.

청구항 18

제 17 항에 있어서,

상기 전의 사전과 비전의 사전은,

조회 청구 중의 단어에 대해 어의 유형을 표기하고, 표기된 어의 유형을 이용하여 원생 전의 사전과 원생 비전의 사전에 대응되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 구축하는 방법으로 구축되는 추상화 된 전

의 사전과 추상화 된 비전의 사전을 더 포함하는 검색 엔진.

청구항 19

프로세서; 및

메모리 장치;를 포함하되,

상기 메모리 장치는 컴퓨터 판독 가능한 명령을 저장하고, 상기 프로세서로 상기 컴퓨터 판독 가능한 명령을 실행할 경우, 상기 프로세서는,

사용자가 입력한 조회 청구를 수신하고,

상기 조회 청구와 매칭되는 후보 결과를 검색하고,

클릭 전의 모형을 기반으로 상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하고,

상기 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하되,

상기 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 상기 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하며, 상기 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하되,

상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 것은,

각 후보 결과에 대하여, 상기 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정하고,

확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 상기 조회 청구와 상기 후보 결과사이의 어의 상관도를 확정하는 것을 포함하되, 상기 어구는 후보 결과의 타이틀, 앵커 문구 및 본문 중의 핵심 문장 중 적어도 하나를 포함하는 검색 엔진을 구현하기 위한 시스템.

청구항 20

컴퓨터 판독 가능한 명령을 저장하는 비휘발성 컴퓨터 판독 가능한 기록 매체에 있어서,

프로세서로 상기 컴퓨터 판독 가능한 명령을 실행할 경우, 상기 프로세서는,

사용자가 입력한 조회 청구를 수신하고,

상기 조회 청구와 매칭되는 후보 결과를 검색하고,

클릭 전의 모형을 기반으로 상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하고,

상기 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하되,

상기 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 상기 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하며, 상기 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하되,

상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 것은,

각 후보 결과에 대하여, 상기 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정하고,

확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 상기 조회 청구와 상기 후보 결과사이의 어의 상관도를 확정하는 것을 포함하되, 상기 어구는 후보 결과의 타이틀, 앵커 문구 및 본문 중의 핵심 문장 중 적어도 하나를 포함하는 것을 특징으로 하는 비휘발성 컴퓨터 판독 가능한 기록 매체.

청구항 21

삭제

청구항 22

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 컴퓨터 기술 분야에 관한 것으로, 구체적으로 정보 검색 분야에 관한 것이며, 특히는 검색 엔진 및 그의 구현 방법에 관한 것이다.

배 경 기 술

[0002] 인터넷은 여러가지 자원을 에세스할 입구를 제공한다. 이러한 자원은 예를 들면, 화상 파일, 음성 파일, 영상 파일 및 웹페이지 등을 포함한다. 사용자는 검색 시스템 또는 검색 엔진을 통하여 방문하려고 하는 자원을 검색할 수 있다.

[0003] 검색 과정에서는, 통상적으로 사용자가 하나의 조회(Query)를 입력하고, 검색 엔진이 조회와 매칭되는 결과를 피드백한다. 조회는 문서 조회일 수 있으며, 하나 또는 다수의 조회 단어(Term) 또는 문구를 포함한다. 검색 엔진은 예를 들면, 문서 상관의 매칭 방법을 통하여 검색 조회와 상응한검색 결과를 피드백할 수 있다.

[0004] 실제 검색 과정에서, 문서 상관의 매칭 방법을 통하여 피드백한 결과는 종종 사용자의 조회 수요와 매칭되지 않으며, 전의가 발생한다. 예를 들면, 사용자가 모 스타 A를 검색할 경우, 검색 결과에는 "A의 자가용"을 포함하는 상관 문서가 포함될 수 있으며, “中国国旗”를 검색할 경우, “海里有挂满中国国旗的渔船”의 결과가 발생할 수 있다.

[0005] 기존의 문서 매칭 방안은, 주로 검색 결과 문서의 공동 부분이 조회 및 검색 결과에서 차지하는 비례를 조회하는 방식, BM25의 상관성 방식 등이 존재한다. 그러나, 이러한 매칭 방안은 상술한 전의 문제를 해결하지 못한다.

발명의 내용

해결하려는 과제

[0006] 상기와 같은 문제점들을 감안하여, 본 발명은 검색 결과 전의 문제를 효과적으로 해결할 수 있는 방안을 제공한다.

과제의 해결 수단

[0007] 제1 방면에 있어서, 본 발명의 실시예는 검색 엔진의 구현 방법을 제공한다. 상기 방법은, 사용자가 입력한 조회 청구를 수신하는 단계; 조회 청구와 매칭되는 후보 결과를 획득하는 단계; 클릭 전의 모형을 기반으로 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 단계; 및 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하는 단계;를 포함하되, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하며, 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함한다.

[0008] 제2 방면에 있어서, 본 발명의 실시예는 검색 엔진을 더 제공한다. 상기 검색 엔진은, 사용자가 입력한 조회 청구를 수신하는 수신 유닛; 상기 조회 청구와 매칭되는 후보 결과를 검색하는 검색 유닛; 클릭 전의 모형을 기반으로 상기 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하는 어의 상관도 확정 유닛; 및 상기 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행하는 순서배열 유닛;을 포함한다. 여기서, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 상기 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하고, 상기 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함한다.

발명의 효과

[0009] 본 발명의 실시예가 제공하는 검색 엔진 및 그의 구현 방법은, 클릭을 기반으로 URL에 관련된 HTTP 청구 링크를 획득함으로써, 비교적 전면적인 URL 관련 웹페이지 내용을 획득할 수 있으며, 악성 웹주소에 대해 정확한 검출을 진행할 수 있다. 본 발명의 실시예의 기술적 방안에 의하면, 어의 상관도에 따라 검색된 후보 결과에 대해 순서배열을 진행함으로써, 검색 결과의 순서배열 효과를 향상시킬 수 있으며 검색 결과 리스트의 전열에 사용자

의 검색 수요에 부합되지 않는 결과(즉, 전의 결과)가 발생하는 것을 피면할 수 있으므로 사용자로 하여금 양호한 사용 체험을 가지도록 확보한다.

도면의 간단한 설명

[0010] 본 발명의 기타 특징, 목적 및 장점들은 하기 도면을 결합하여 진행하는 비제한적 실시예들에 대한 구체적인 설명을 통하여 더욱 명확해 질 것이다.

도1은 본 발명의 실시예를 적용할 수 있는 예시적 시스템 구조(100)를 보여준다.

도2는 본 발명의 실시예에 따른 클릭 전의(转义) 모형을 구성하는 방법의 예시적 흐름도를 보여준다.

도3은 본 발명의 실시예에 따른 단어 정렬을 이용하여 인접한 상하문을 획득하는 일 실시예의 구현을 보여준다.

도4는 본 발명의 실시예에 따른 검색 엔진의 구현 방법의 예시적 흐름도를 보여준다.

도5는 본 발명의 실시예에 따른 클릭 전의 모형을 기반으로 조회 청구와 후보 결과사이의 어의 상관도를 확정하는 방법의 예시적 흐름도를 보여준다.

도6은 본 발명의 실시예에 따른 문구에 대해 처리를 지행한 결과의 예시도를 보여준다.

도7은 본 발명의 실시예에 따른 클릭 전의 모형을 기반으로 단어 분리 유사도의 가중치를 조정하는 방법의 일 예시적 흐름도를 보여준다.

도8은 본 발명의 실시예에 따른 검색 엔진의 예시적 구조 블록도를 보여준다.

도9는 본 발명의 실시예의 서버를 실현하기 위한 컴퓨터 시스템의 예시적 구조도를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0011] 이하, 첨부된 도면 및 실시예들을 결합하여 본 발명을 상세히 설명하기로 한다. 본 명세서에 설명된 구체적인 실시예들은 오직 해당 발명을 설명하기 위한 것일 뿐, 해당 발명을 한정하기 위한 것이 아님을 자명하여야 할 것이다. 또한, 설명의 편의를 위하여, 도면에는 오직 본 발명에 관련된 부분만이 도시되어 있다.

[0012] 본 발명의 실시예 및 실시예의 특징들은 서로 모순되지 않는한 상호 조합할 수 있다. 이하, 첨부된 도면을 참조하여 본 발명의 실시예들을 상세히 설명하기로 한다.

[0013] 배경 기술에 기재된 바와 같이, 문서 검색에서 통상적으로 문서의 국부적 매칭으로 인하여 전의(转义)문제를 초래하게 된다. 예를 들면, 모기향을 검색할 경우, 결과는 모기향 곽을 포함하고; 휴대폰을 검색할 경우, 결과는 휴대폰 케이스를 포함하며; 상산 (常山)을 검색할 경우, 결과는 상산 배추를 포함한다. 문서을 이용하여 픽처를 검색할 경우 이러한 문제들은 더욱 뚜렷하다. 예를 들면, "스타 A"의 픽처를 검색할 경우, 결과는 스타A의 촬영도, 스타A의 고화질 초상도, 스타A의 콘서트, 스타A의 자가용 등을 포함한다. 이러한 결과중, 스타A의 자가용은 전의(转义)된 결과이고 사용자가 원하는 결과가 아니다.

[0014] 기존 기술의 상기 결함을 감안하여, 본 발명의 실시예는 상기 전의(转义) 문제를 해결하도록 어의 전의(转义) 정도에 따라 검색 결과에 대해 순서배열을 진행하는 방안을 제공한다. 통상적으로, 검색 과정에 표시된 결과 중 클릭되는 회수가 높은 결과가 흔히 사용자가 원하는 결과임을 이해하여야 한다. 즉, 클릭되는 회수가 높은 결과가 사용자가 조회하는 Query에 대하여 전의가 발생하지 않는 확률이 매우 높다. 반면, 여러번 표시되었으나 클릭되는 회수가 낮거나 클릭된 적이 없는 결과는 통상적으로 사용자가 원하지 않는 것이다. 즉, 이러한 결과는 사용자가 조회하는 Query에 대하여 전의가 발생하지 확률이 매우 높다. 또한, 전의된 데이터에 대해 분석을 진행할 때, 대다수의 전의가 모두 인전합 상하문중에 발생하며 거리가 비교적 먼 상하문에는 기본상 아무런 영향을 미치지 않음 발견하였다. 따라서, 상기 분석을 기반으로 본 발명의 여러 실시예에 따른 검색 엔진의 구현 방법을 개시한다.

[0015] 도1은 본 발명의 실시예를 적용할 수 있는 예시적 시스템 구조(100)를 보여준다.

[0016] 도1에 도시된 바와 같이, 시스템 구조(100)는 단말기 장치(101, 102), 네트워크(103) 및 서버(104)를 포함할 수

있다. 네트워크(103)는 단말기 장치(101, 102)와 서버(104)사이에서 통신링크의 매체를 제공한다. 네트워크(103)는 예를 들면, 유선, 무선 통신 링크 또는 광섬유 케이블 등 각종 연결 유형을 포함할 수 있다.

[0017] 사용자(110)는 단말기 장치(101, 102)를 사용하여 네트워크(103)를 통하여 서버(104)와 교호하여 예를 들면, 정보 검색, 웹 페이지 열람, 데이터 다운로드 등 각종 서비스를 방문할 수 있다. 단말기 장치(101, 102)에는 예를 들면, 유니폼 리소스 로케이터URL 클라우드 서비스를 어세스할 수 있는 어플리케이션과 같은 각종 클라이언트 어플리케이션이 설치될 수 있으며, 브라우저, 안전 어플리케이션 등을 포함하나 이에 한정된 것은 아니다.

[0018] 단말기 장치(101, 102)는 각종 전자 장치일 수 있으며, 예를 들면, 예를 들면, 스마트폰, 태블릿 PC, PDA, 전자책 열람기 등과 같은 각종 이동가능한 휴대용 장치, 및 예를 들면, 개인용 컴퓨터, 스마트 TV, 조회 서비스 단말기 등과 같은 고정형 단말기 장치를 포함할 수 있으나 이에 한정된 것은 아니다.

[0019] 서버(104)는 각종 서비스를 제공하는 서버일 수 있다. 서버는 서비스 청구애 응하여 서비스를 제공할 수 있다. 하나의 서버가 하나 또는 다수의 서비스를 제공하거나, 다수의 서버가 동일한 서비스를 제공할 수 있음을 이해하여야 한다. 본 발명의 실시예에서, 관련된 서버(104)는 검색 서버일 수 있다.

[0020] 도1 중의 단말기 장치, 네트워크 및 서버의 수량은 오직 예시적이다. 구현의 수요에 따라, 임의의 수량의 단말기 장치, 네트워크 및 서버를 구비할 수 있다.

[0021] 본 발명의 실시예의 검색 엔진의 구현 방법을 설명하기 위하여, 먼저 본 발명의 실시예에 개시된 클릭 전의 모형의 구축을 설명한다. 앞서 분석한 바와 같이, 클릭되는 회수가 높은 검색 결과가 상응한 죄회 Query에 대하여 전의가 발생하지 않는 확률이 높고, 클릭되는 회수가 낮거나 클릭된 적이 없는 검색 결과가 상응한 Query에 대하여 전의가 발생하는 확률이 높다. 또한, 대다수의 전의는 모두 인접한 상하문중에 발생하며 거리가 비교적 먼 상하문에는 기본상 아무런 영향을 미치지 않는다. 따라서, 본 발명의 실시예에서, 조회 청구와 검색 결과 (예를 들면, 웹페이지 타이틀 표시)Query-Title 쌍의 클릭 회수를 학습하면서 전의가 발생하는 상하문을 고려하여 클릭 전의 모형을 구축한다. 구체적으로, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함할 수 있다. 여기서, 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하고, 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함한다.

[0022] 도2는 본 발명의 실시예에 따른 클릭 전의 모형을 구축하는 방법의 예시적 흐름도를 보여준다.

[0023] 도2에 도시된 바와 같이, 단계(210)에서, Query-Title 쌍의 클릭 표시 비율을 획득한다.

[0024] 클릭 전의 모형은 기왕 Query-Title 쌍을 학습하여 구축할 수 있다. 이러한 기왕 Query-Title 쌍은 Query 일지에 저장될 수 있다. Query 일지는 예를 들면, 매번 사용자 조회 대화에서 사용되는 조회 청구Query, 표시된 검색 결과 및 사용자가 검색 결과에 대한 클릭 조작 등을 기록한다. 이러한 검색 결과는 예를 들면, 웹페이지 타이틀Title를 이용하여 표시되고, 따라서, Query-Title 쌍이 가리키는 것은 조회-검색 결과 쌍이다.

[0025] 각 Query-Title 쌍의 표시 정황 및 클릭 정황을 통계하여 Query-Title 쌍의 클릭 표시 비율을 획득할 수 있다. 여기서, 클릭 표시 비율은 클릭 회수와 표시 회수사이의 비율이고, 표시 회수는 검색 결과Title가 조회 청구Query에 응하여 표시되는 회수를 지시하고, 클릭 회수는 검색 결과Title가 조회 청구Query에 응하여 표시될 때 사용자에 의해 클릭되는 회수를 지시한다.

[0026] 앞서 분석한 바와 같이, 클릭되는 회수가 높은 검색 결과가 상응한 죄회 Query에 대하여 전의가 발생하지 않는 확률이 높고, 클릭되는 회수가 낮거나 클릭된 적이 없는 검색 결과가 상응한 Query에 대하여 전의가 발생하는 확률이 높다. 따라서, Query-Title 쌍의 클릭 표시 비율은 Title이 Query에 대한 전의도 (转义度)또는 전의 확률을 비교적 잘 표현할 수 있다. 예를 들면, 표시 클릭 비율 또는 클릭 회수를 기반으로 하는 기타 파라미터를 사용하여 전의도 또는 전의 확률을 표현할 수도 있음을 해당 기술 분야에서 통상 지식을 가진 자가 자명할 것이다.

[0027] 다음, 단계(220)에서, 단어 정렬을 이용하여 검색 결과Title에서 조회 Query 어구 중 단어와 정렬된 인접한 상하문을 획득한다.

[0028] 각 Query-Title 쌍에 있어서, 먼저 Query와 Title에 대해 각각 단어 분리를 진행할 수 있다. 다음, 단어 정렬을 이용하여 Query 중의 각 단어에 대해 그가 Title에서의 상응한 위치를 조회한다. 여기의 단어 정렬은 동의 정렬 (同义对齐)도 포함한다. 예를 들면, 완전히 대응되는 단어가 존재하지 않을 경우, 그의 동의어를 고려한다. 마지막으로, Title에서 Query 중 첫 단어 정렬과 끝 단어 정렬의 인접한 상하문을 획득한다.

[0029] 도3은 본 발명의 실시예에 따른 단어 정렬을 이용하여 인접한 상하문을 획득하는 일 실시예의 구현을 보여준다. 도3의 예에서, Query는 “中国国旗”이고, Title은 “海里有挂满中国国旗的渔船”이다.

[0030] 도3에 도시된 바와 같이, Query와 Title에 대해 각각 단어 분리를 진행한다. 구체적으로, Query는 “中国”과 “国旗”로 분리되고, Title은 “海里”, “有”, “挂满”, “中国”, “国旗”, “的”와 “渔船”으로 분리되며, 도면에서는 블록으로 각 단어를 분리한다.

[0031] 다음, 단어 정렬을 이용하여 Query 중의 각 단어에 대해 그가 Title에서의 상응한 위치를 조회한다. 도3의 예에서, 화살표가 지시하는 바와 같이, Query 중의 각 단어 “中国”과 “国旗”는 모두 Title에서 완전히 대응되는 단어를 조회할 수 있다.

[0032] 마지막으로, Title에서 첫 단어 정렬 및 끝 단어 정렬된 인접한 상하문을 획득한다. 더욱, 구체적으로, 첫 단어 정렬된 인접한 상 문장과 끝 단어 정렬된 인접한 하문을 획득한다. 본 예시에서, 첫 단어 “中国”의 인접한 상문은 “挂满”” 이고, 끝 단어 “国旗”의 인접한 하문은 “的” 이고, 정지어 “的” 를 필터링하고 그뒤의 비정지어를 계속하여 검색하여 하문으로 이용한다. 즉, “国旗”의 인접한 하문은 “渔船”이다.

[0033] 인류 언어는 수많은 기능어들을 포함한다. 기타 단어와 비교시, 기능어는 실질적인 함의를 구비하지 않는다. 제일 보편적인 기능어는 한정어(“这”, “这个”, “那”, “那些”, “the”, “a”, “an”, “that” 및 “those”)이고, 이러한 단어는 문장중에서 지점 또는 수량과 같은 명사를 묘사하고 개념을 표달하는 것을 돕는다. “在..上”, “在..下”, “over”, “under”, “above” 등과 같은 개사는 두 단어의 상대적 위치를 표시한다. 이러한 기능어들은 매우 보편적이며, 이러한 단어들은 각 문서어세의 수량을 기록하는 데는 매우 큰 디스크 공간을 차지한다. 또한, 이들의 보편성과 기능에 인하여 이러한 단어들이 문서 상관 정도의 정보를 단독적으로 표현하는 경우가 흔치 않다. 검색 과정에서 문구를 고려하지 않고 각 단어만 고려할 경우, 이러한 기능어들은 기본상 도움되지 않는다.

[0034] 정보 검색에서, 이러한 기능어의 다른 명칭은 정지어 (stopword) 이다. 이들을 정지어라 지칭하는 이유는 문서처리 과정에서 이들을 마주치게 되는 경우 즉시 처리를 정지하고 이를 버리기 때문이다. 이러한 단어를 버림으로써 인덱스 양을 감소하고 검색 효율을 증가하며 통상적으로 검색 효과도 향상시킨다. 정지어는 주로 영어 문자 부호, 수자, 수학 문자 부호, 문장 부호 및 사용 빈도가 매우 높은 홑 한자 등을 포함한다.

[0035] 다시 도2를 참조하면, 단계(230)에서, 클릭 표시 비율을 기반으로 상응하게 전의 사전 및/또는 비전의 사전을 구축한다. 구체적으로, 클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 전의 사전에 추가하고; 및/또는 클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 비전의 사전에 추가한다. 제1 역치는 제2 역치와 동일하거나 부동할 수 있다.

[0036] 기왕 Query-Title 쌍 중의 각 Query-Title 쌍에 대하여 도2에 도시된 처리를 실행하여, 클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 모든 단어를 주적하고 상응한 상하문을 합병하여 전의 사전을 생성할 수 있으며; 클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 모든 단어를 누적하고 상응한 상하문을 합병하여 비전의 사전을 생성할 수 있다. 상기 전의 사전과 비전의 사전의 생성과정에서 Query 중의 단어를 확장하지 않았으므로, 여기서 생성된 전의 사전은 원생 전의 사전이라 지칭할 수 있으며, 상응한 비전의 사전은 원생 비전의 사전이라 지칭할 수 있다.

[0037] 선택적 또는 부가적으로, 일부 실시예에서, 통계된 상하문을 더욱 큰 범위로 보급하기 위하여, Query 중의 단어의 어의 유형을 추상화하여 추상화 된 전의 사전 및/또는 추상화 된 비전의 사전을 생성할 수 있다.

[0038] 이러한 실시예에서, Query 중의 단어에 대해 어의 유형 표기를 진행하고 단어의 어의 유형을 통하여 추상화를 진행할 수있다. 예를 들면, 단어가 모 스타 A의 이름일 경우, 그의 어의 유형을 스타로 표기할 수 있고; 단어가 주자이거우(九寨沟)일 경우, 그의 어의 유형을 명승지로 표기할 수 있다. 어의 유형 표기를 통하여 일수 실체의 단어를 어의 유형으로 대체할 수 있다.

[0039] 여러가지 방식을 이용하여 단어에 대해 어의 유형 표기를 진행할 수 있으며, 예를 들면, 범용의 최대 엔트로피

분류기를 이용하여 단어에 대해 분류 식별을 진행할 수 있다. 어의 유형은 예를 들면, 예능 스타, 스포츠 스타, 과학 기술 인물, 명승지, 영상, 자동차, 애니메이션, 동물, 식물 등을 포함할 수 있으나 이에 한정된 것은 아니다.

[0040] 표기된 어의 유형을 이용하여 원생 전의 사전과 원생 비전의 사전에 대응되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 구축할 수 있다. 일 구현에서, 원생 전의 사전/원생 비전의 사전 중의 원 단어를 추상화 된 어의 유형으로 간단하게 대체하여 추상화 된 전의 사전/추상화 된 비전의 사전을 생성할 수 있다.

[0041] 이상에서 본 발명의 실시예의 클릭 어의 모형의 구축을 설명하였고, 아래에는 흐름도를 결합하여 클릭 어의 모형을 기반으로 검색 엔진의 검색 결과를 개선하는 방안을 설명하기로 한다.

[0042] 도4는 본 발명의 일 실시예에 따른 검색 엔진의 구현 방법의 예시적 흐름도를 보여준다. 도4에 도시된 방법은 검색 엔진이 위치한 서버(예를 들면, 도1의 서버(104))로 실행할 수 있다.

[0043] 도4에 도시된 바와 같이, 단계(410)에서, 사용자가 입력한 조회 청구를 수신한다.

[0044] 사용자는 각종 단말기 장치(예를 들면, 도1에 도시된 단말기 장치(101, 102))를 통하여 검색 조회를 진행할 수 있다. 이러한 단말기 장치는 사용자에게 사용자 인터페이스(예를 들면, 브라우저 인터페이스)를 표시하여 조회 청구를 입력하도록 한다. 사용자는 예를 들면, 터치 스크린, 스타일러스, 키보드, 마이크 등과 같은 각종 입력 공구를 통하여 조회 청구를 입력할 수 있다. 조회 청구는 문서 조회, 음성 조회 또는 기타 유형의 조회일 수 있다. 조회 청구가 비 문서 조회일 경우, 광학 문자 판독 OCR, 음성 식별 등과 같은 각종 적합한 기술을 이용하여 비문서 조회를 문서 조회로 전환할 수 있다. 나아가, 단말기 장치는 수신된 원 조회 청구 또는 전환된 조회 청구를 검색 서버(예를 들면, 도1의 서버(104))에 발송할 수 있다.

[0045] 다음, 단계(420)에서, 수신된 조회 청구와 매칭되는 후보 결과를 검색한다.

[0046] 조회 청구와 매칭되는 후보 결과는 여러가지 방식을 이용하여 검색할 수 있다. 일부 구현에서, 조회 청구와 매칭되는 후보 결과는 예를 들면 단어 매칭과 같은 문서 매칭 방법을 사용하여 검색할 수 있다. 단어 매칭 방법의 일부 범용의 연산법은 예를 들면, BM25 (Best Match, 최적 매칭) 연산법, proximity(Term proximity scoring, 단어 인접 스코어링) 연산법 등을 포함할 수 있다. 단어 매칭 연산법을 통하여 검새하는 문서과 조회 청구의 매칭 정도를 산출하고, 나아가 매칭 정도를 기반으로 조회 청구와 매칭되는 후보 결과를 제공할 수 있다. 상기 검색 방법은 현재 이미 알려진 각종 연산법을 사용하여 실현할 수 있으므로 불필요한 설명은 생략하기로 한다.

[0047] 나아가, 단계(430)에서, 클릭 전의 모형을 기반으로 조회 청구와 각 후보 결과사이의 어의 상관도를 확정한다.

[0048] 실제 검색에서, 조회 청구와 매칭되는 후보 결과에 대해 통상적으로 일정한 수량의 후보 결과를 선별하고 세분화 처리를 진행한다. 예를 들면, 2000개 후보 결과를 선별하고 이러한 결과 중 각 후보 결과와 조회 청구의 어의 상관도를 분석할 수 있다.

[0049] 앞서 도2 및 도3을 결합하여 설명한 바와 같이, 클릭 전의 모형은 조회 청구와 검색 결과 Query-Title 쌍의 클릭 회수를 학습하면서 전의 발생의 상하문을 고려하여 구축한다. 구체적으로, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함할 수 있으며, 전의 사전은 전의가 발생함을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함하고, 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문을 포함한다.

[0050] 따라서, 클릭 전의 모형을 기반으로 확정한 어의 상관도는 Query-Title 쌍의 클릭 회수를 고려하였을 뿐만 아니라 전의 발생의 상하문고 고려하였으므로, 확정된 어의 상관도는 조회 청구에 대한 후보 결과의 전의 확률을 정확히 표시할 수 있다. 아래에 클릭 전의 모형을 기반으로 어의 상관도를 확정하는 상세한 방법을 설명하기로 한다.

[0051] 마지막으로, 단계(440)에서, 어의 상관도에 따라 후보 결과에 대해 순서배열을 진행한다.

[0052] 본 단계에서, 각 후보 결과와 조회 청구의 어의 상관도가 낮아지는 순서에 따라 선택된 후보 결과에 대해 순서배열을 진행하고 표시하여 조회 청구와 비교적 상관된 검색 결과가 앞에 표시되도록 하여 사용자로 하여금 표시된 검색 결과로부터 원하는 상관된 문서을 신속히 획득하도록 하여 자신의 검색 수요를 만족 시키고 검색 효율을 향상시킬 수 있다. 본 단계는 수요에 따라 기타 순서를 이용하여 순서배열 처리를 진행할 수 있음을 이해하

여야 한다.

[0053] 도5는 본 발명의 실시예에 따른 클릭 전의 모형을 기반으로 조회 청구와 후보 결과사이의 어의 상관도를 확정하는 방법의 예시적 흐름도를 보여준다. 즉, 도5는 도4의 단계(430)의 일 예시적 구현을 보여준다.

[0054] 도5에 도시된 바와 같이, 단계(510)에서, 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정한다.

[0055] 후보 결과는 각종 웹 페이지 정보로서, 문서 (document) 를 사용하여 표시할 수 있다. 통상적으로, 문서는 다수의 어구로 조성되고, 그의 구조상 예를 들면 타이틀(Title), 앵커 문구 (Anchor text) 및 본문 등을 포함할 수 있다. 타이틀은 문서의 주제를 간단하고 정련하게 설명한다. 앵커 문구는 앵커 문구 링크라고도 지칭하며, 링크의 일 형식이며, 하이퍼 링크와 유사하게, 키워드를 하나의 링크로 하고 다른 웹페이지에 지향한다. 이러한 형식의 링크를 앵커 문구라 지칭한다. 앵커 문구는 실질상 문서 키워드와 URL 링크의 관계를 생성한다. 본문은 통상적으로 비교적 많은 내용을 포함한다.

[0056] 후보 결과가 통상적으로 비교적 많은 어구를 포함하므로, 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 각각 확정할 수 있다. 이러한 어구는 예를 들면, 타이틀, 앵커 문구, 본문 중의 핵심 문장 등으로부터 선택될 수 있다. 본문 중의 핵심 문장은 기존 기술 중에 이미 알려진 또는 미래에 개발될 여러가지 방식을 이용하여 확정할 수 있다. 일 구현에서, 본문 중의 첫 구절을 그의 핵심 문장으로 인정할 수 있다.

[0057] 다음, 단계(520)에서, 확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 조회 청구와 해당 후보 결과사이의 어의 상관도를 확정한다.

[0058] 조회 청구와 후보 결과사이의 최종 어의 상관도는 여러가지 방식을 통하여 확정할 수 있다. 일 구현에서, 확정된 다수의 어의 상관도로부터 그의 최대치를 선택하여 조회 청구와 해당 후보 결과사이의 어의 상관도로 할 수 있다. 다른 일 구현에서, 확정된 다수의 어의 상관도의 평균치를 조회 청구와 해당 후보 결과사이의 어의 상관도로 할 수 있다. 기타 함수 관계를 사용하여, 확정된 다수의 어의 상관도를 기반으로 조회 청구와 해당 후보 결과의 최종 어의 상관도를 확정할 수 있으며 본 발명은 이에 한정되지 않음을 해당 기술 분야에서 통상 지식을 가진 자가 자명할 것이다.

[0059] 단계(510)는 본 발명의 실시예에 따른 조회 청구와 후보 결과의 모 어구사이의 어의 상관도를 확정하는 방법의 예시적 구현을 진일보로 보여준다. 이러한 구현에서, 어의 상관도는 주로 어구사이의 주제 매칭 유사도 및 어구사이의 전의 인자 이 두개의 부분으로 조성된다.

[0060] 구체적으로, 단계(511)에서, 미리 구축된 클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출한다.

[0061] 두 어구사이의 주제 매칭 유사도는 여러가지 도량 방식을 이용하여 표시할 수 있다. 일부 구현에서, 통합된 프레임의 벡터 공간 모형 유사도 산출 방법을 이용하여 어구사이의 주제 매칭 유사도를 산출할 수 있다.

[0062] 예를 들면, 두 문장을 각각 S_1, S_2로 표시하면,

[0063] $$S_1 = \{(w_{11_1}, w_{11_2} \cdots w_{11_k}), (w_{12_1}, w_{12_2} \cdots w_{12_k}), \cdots, (w_{1m_1}, w_{1m_2} \cdots w_{1m_k})\} \quad (1)$$

[0064] $$S_2 = \{(w_{21_1}, w_{21_2} \cdots w_{21_k}), (w_{22_1}, w_{22_2} \cdots w_{22_k}), \cdots, (w_{2n_1}, w_{2n_2} \cdots w_{2n_k})\} \quad (2)$$이다.

[0065] 상기 공식에서, 문장을 단어 분리한다. 예를 들면, 제1 문장S_1을 m개의 단어로 분리하고, 제2 문장S_2을 n개의 단어로 분리한다. 분리된 단어에 대해 품사 표기를 진행하여 각 단어 분리위치에서 하나의 단어 집합을 획득한다. 예를 들면, 제1 문장S_1의 단어 분리 위치 W_{1i} 상의 단어 집합은 $(w_{1i_1}, w_{1i_2} \cdots w_{1i_k})$ 이다. 해당 단어 집합은 단어 분리 위치 W_{1i} 에 대응되는 원 단어, 상관된 단어 및 작은 입도로 조성된 부분을 포함한다.

[0066] 여기서, 산관된 단어는 원 단어의 어의와 동인한 단어(또는 동의어라고도 지칭함) 또는 어의가 유사한 단어를 가리키고, 이를 상관된 단어로 통칭한다. 원 단어와 상관된 단어는 여러가지 방식을 이용하여, 예를 들면, Query-Title 클릭 쌍을 기반으로 발굴할 수 있다. 상기 상관된 단어를 발굴하는 방법은 현재 이미 알려진 각종 방안을 사용하여 실현할 수 있으므로 불필요한 설명은 생략하기로 한다.

[0067] 문장을 공간 벡터로 표시한 후, 각종 도량 방식을 이용하여 두 벡터사이의 유사도, 즉, 어구사이의 주제 매칭

유사도를 산출할 수 있다. 이러한 도량 방식은 코사인 거리(또는 코사인 유사도라고도 지칭함), 유클리드 거리, 피어슨Pearson 상관 계수법 또는 수정된 Pearson 상관 계수법을 포함하나, 이에 한정된 것은 아니다. 이러한 유사도 또는 상관성을 산출하는 방법은 해당 기술 분야에서 이미 알려진 것이다. 이하 코사인 거리를 예로 들어 설명하기로 한다.

[0068] 코사인 거리는 벡터 공간 중 두개의 벡터 협각의 코사인 값을 두개의 개체사이의 차이의 크기를 판단하는 도량으로 한다. 예를 들면, 하기 등식으로 두개의 어구사이의 주제 매칭 유사도를 산출할 수 있다.

$$Sim(S_1,S_2)=\frac{\sum_{w_{1k_i}=w_{2k_j}}(Wgt(w_{1k_i})*Wgt(w_{2k_j}))}{\sqrt{\sum_{t=1\cdots m}Wgt(w_{1k_t})^2}\sqrt{\sum_{j=1\cdots n}Wgt(w_{2k_j})^2}}*SentType(S_1,S_2)$$

[0069] (3)

[0070] $Wgt(w_{1k_i})$는 단어 w_{1k_i}의 유사도 가중치를 표시하고, $SentType(S_1,S_2)$는 두 문장이 상응한 가중치 계수에 매칭되는지 표시하고, 다 문장S_1, S_2의 의문문 유형이 매칭되면 상응한 가중치 계수가 제1값, 예를 들면 1이고, 아니면 제2값, 예를 들면 0.8이다.

[0071] 이하, 구체적인 실시예를 결합하여 어떻게 두 어구사이의 주제 매칭 유사도를 산출하는지 설명하기로 한다. 제1 문장S_1을 "华中科技大学在湖北武汉哪个地方"로 가정할 경우, 제2 문장 S_2은 "华科大在武汉市什么位置"이다.

[0072] 먼저, 이 두 문장에 대해 각각 단어 분리 처리와 품사 표기를 진행한다. 간단 명료함을 위하여, 본 실시예에는 품사 표기를 도시하지 않았다. S_1이 획득한 단어 분리 결과는 "华中科技大学", "在", "湖北", "武汉", "哪个地方"이다. 여기서, "华中科技大学"가 대응되는 더 작은 단어 분리 입도의 단어는 "华中", "科技", "大学"이고, "哪个地方"가 대응되는 더 작은 단어 분리 입도의 단어는 "哪个", "地方"이다. S_2가 획득한 단어 분리 결과는 "华科大", "在", "武汉市", "什么位置"이다. 여기서, "什么位置"가 대응되는 더 작은 단어 분리 입도의 단어는 "什么", "位置"이다.

[0073] 단어 분리를 진행하여 획득한 각 단어에 가중치를 부여한다. 선택적 또는 부가적으로, 어구 중 어의 잉여 단어를 식별하고 잉여 단어의 가중치를 낮춘다. 어의 잉여 단어 식별은 기존 기술 중에 이미 알려진 또는 미래에 개발될 여러가지 기술을 이용하여 식별할 수 있으며, 본 발명은 이를 한정하지 않는다. 어의 잉여 단어의 식별을 진행한 후, 예를 들면, 제1 문장 중의 "湖北"를 어의 잉여 단어로 확정하고 그의 가중치를 낮춘다.

[0074] 다음, 어의 매핑이 존재하는 어의를 통합된 표현으로 매핑한다. 구체적으로 저1 문장S_1에서 "华中科技大学"를 "华中科技大学"으로 매핑하고, "武汉"을 으로 "武汉"매핑하며, "哪个地方"을 "哪里"로 매핑한다. 제2 문장S_2에서 "华科大"를 "华中科技大学"로 매핑하고, "武汉市"를 "武汉"으로 매핑하며, "什么位置"를 "哪里"로 매핑한다.

[0075] 또한, 두 문장의 의문문 유형에 대해 매핑을 진행한다. 의문어 "哪个"와 그의 상하문에 나타난 명사 "地方"에 대응되는 의문문 유형이 "地点(지점)"이고, 의문어 "什么"와 그의 상하문에 나타난 명사 "位置"에 대응되는 의문문 유형이 "地点(지점)"이므로, 의문문 S_1과 S_2가 동일한 의문문 유형에 속함을 식별할 수 있다. 따라서, 가중치 계수$SentType(S_1,S_2)$가 제1값, 예를 들면 1을 취함을 확정할 수 있다.

[0076] 도6은 어구에 대해 상기 처리를 진행한 결과의 예시도를 보여준다.

[0077] 도6에 도시된 바와 같이, 저1 문장S_1에서, “华中科技大学”, “华中”, “科技”, “大学”는 제1 어의 매핑위치에 대응되고, “湖北”는 제2 어의 매핑위치에 대응되고, “武汉”은 제3 어의 매핑위치에 대응되며, “哪个地方”, “哪个”, “地方”는 제4 어의 매핑 위치에 대응된다. 저2 문장S_2에서, “华科大”는 제1 어의 매핑위치에 대응되고, “武汉市”는 제2 어의 매핑위치에 대응되고, “什么位置”, “什么”, “位置”는 제3 어의 매핑위치에 대응된다.

[0078] “华中科技大学”과 “华科大”가 동일한 통합된 표현에 매핑됨으로 “华中科技大学”과 “华科大”는 매핑 성공된 단어이다. “在”는 정지어임으로 이를 그냥 지나치고 산출에 참여하지 않는다. “武汉”과 “武汉市”는 동일한 통합된 표현에 매핑됨으로 “武汉”과 “武汉市”는 매핑 성공된 단어이다. “哪个地方”과 “什么位置”는 동일한 통합된 표현에 매핑됨으로 “哪个地方”과 “什么位置”는 매핑 성공된 단어이다.

[0079] 앞서 기재된 공식(3)으로 두 문장사이의 주제 매칭 유사도를 산출할 수 있다.

$$Sim(S_1,S_2)=1*[Wgt(华中科技大学)*Wgt(华科大)+Wgt(武汉)*Wgt(武汉市)+Wgt(哪个地方)*Wgt(什么位置)]/\{[Wgt(华中科技大学)^2+Wgt(湖北)^2+Wgt(武汉)^2+Wgt(哪个地方)^2]^{\frac{1}{2}}*[Wgt(华科大)^2+Wgt(武汉市)^2+Wgt(什么位置)^2]^{\frac{1}{2}}\}。$$

[0080]

[0081] 본 발명의 실시예에서, 미리 구축된 클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 산출한 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도는, 클릭 전의 모형을 이용하여 후보 결과의 어구 중의 일부 단어의 유사도 가중치를 조정하여 표현할 수 있다.

[0082] 단어의 초기 유사도 가중치는 문서 발굴 분야에 이미 알려진 기술로 분배할 수 있다. 다중 가중치 분배 방식이 존재할 수 있으며, 범용의 예로 TF-IDF (term frequency-inverse document frequency) 가 포함한다.

[0083] TF-IDF는 정보 검색과 정보 발굴에 범용의 가중치 분배 기술이며, 검색, 문서 분류 및 기타 상응한 분야에서 광범위하게 응용되고 있다. TF-IDF의 주요 사상은, 모 단어 또는 문구가 한편의 문장에서 나타나는 빈도TF가 높으나 기타 문장에서는 아주 적게 나타나면, 해당 단어 또는 문구가 아주 좋은 유형 구별 능력을 구비하고 분류에 접합한 것으로 인정한다. TF 단어 빈도(Term Frequency)는 모 주어진 단어가 문서중에 나타나는 회수를 가리킨다. IDF 반 문서 빈도(Inverse Document Frequency)의 주요 사상은, 수록어를 포함하는 문서가 적고 IDF가 크면, 수록어가 아주 좋은 유형 구별 능력을 구비하는 것을 설명한다. TF와 IDF를 이용하여 모 키워드가 모 문장에서의 중요성을 산출할 수 있고, TF와 IDF를 기반으로 각 함수 관계를 이용하여 수록어의 가중치를 구성할 수 있다.

[0084] 일부 구현에서, 단어의 초기 가중치는 하기 등식을 이용하여 산출할 수 있다.

[0085] $$Wgt_{imi}(w_{1k_i})=(\log TF(w_{1k_i})+1)*\log IDF(w_{1k_i})=(\log TF(w_{1k_i})+1)*\log(N/DF(w_{1k_i})) \quad (4)$$

[0086] 여기서, $TF(w_{1k_i})$는 분리된 단어w_{1k_i}의 단어 빈도이고, 분리된 단어 w_{1k_i}가 해당 문서에서 나타나는 회수와 해당 문서 분리된 단어의 총수사이의 비례로 표시할 수 있다. $IDF(w_{1k_i})$는 분리된 단어 w_{1k_i}의 반문서 빈도이고, N은 총 문서수이며, $DF(w_{1k_i})$는 분리된 단어w_{1k_i}가 나타난 문서수이다.

[0087] 본 출원의 일부 실시예에서, 각 어구 중의 분리된 단어에 대해 초기 가중치를 확정한 후, 클릭 전의 모형을 기

반으로 후보 결과의 어구 중의 일부 분리된 단어의 유사도를 조정할 수 있다.

[0088] 도7은 본 발명의 실시예에 따른 클릭 전의 모형을 기반으로 단어 분리 유사도의 가중치를 조정하는 방법의 일 예시적 흐름도를 보여준다.

[0089] 도7에 도시된 바와 같이, 단계(710)에서, 단어 정렬을 이용하여 후보 결과의 어구로부터 조회 청구 중의 단어와 정렬되는 인접한 상문과 하문을 확정한다. 해당 단계는 앞서 도2를 결합하여 설명한 클릭 전의 모형을 구축하는 단계(220)와 유사하므로 중복된 설명은 생략하기로 한다.

[0090] 다음, 단계(720)에서, 전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구 중의 상응한 상문과 하문의 유사도 가중치를 조정한다.

[0091] 해당 단계에서, 식별된 인접한 상문과 인접한 하문에 대하여, 전의 사전 및 비전의 사전을 조회하여 이러한 인접한 상문과 인접한 하문의 유사도 가중치를 조정할 수 있다.

[0092] 구체적으로, 비전의 사전에 후보 결과의 어구 중의 상응한 단어 및 그의 인접한 상문 또는 인접한 하문이 포함될 경우, 해당 인접한 상문 또는 인접한 하문의 유사도 가중치를 낮춘다. 전의 사전에 후보 결과의 어구 중의 상응한 단어 및 그의 인접한 상문 또는 인접한 하문이 포함될 경우, 해당 인접한 상문 또는 인접한 하문의 유사도 가중치를 높인다. 비전의 사전과 전의 사전에서 모두 상응한 단어 및 그의 인접한 상문 또는 인접한 하문을 찾아내지 못할 경우, 그의 유사도 가중치를 조정하지 않을 수 있다.

[0093] 예를 들면, 조회 어구가 “中国国旗”이고, 후보 결과가 “海里有挂满中国国旗的渔船”이며, 인접한 상문이 “挂满”이고, 인접한 하문이 “渔船”이다. 단어 “中国”과 인접한 상문 “挂满”에 대하여, 먼저 원생 전의 사전과 비전의 사전에서 조회를 진행할 수 있다. 원생 비전의 사전에 “中国，挂满”이 존재할 경우, “挂满”의 유사도 가중치를 낮추어 주제 매칭 유사도를 향상할 수 있다. 원생 전의 사전과 비전의 사전에 모두 “中国，挂满”이 존재하지 않을 경우, 추상화 된 전의 사전과 비전의 사전에서 계속하여 조회할 수 있다. 추상화 된 비전의 사전에서 “【地名】，挂满”이 조회될 경우에도 “挂满”의 가중치를 낮출수 있다. 단어 “国旗”와 인접한 하문“渔船”에 대하여, 동일한 사로를 기반으로 처리할 수 있으므로 중복된 설명은 생략하기로 한다.

[0094] 클릭 전의 모형을 기반으로 단어의 유사도 가중치를 조정한 후, 앞서 설명한 문장사이의 문구 주제 매칭 모형을 이용하여 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출할 수 있다.

[0095] 예를 들면, 하기 공식에 따라 조회 청구와 후보 결과의 어구 사이의 주제 매칭 유사도를 산출할 수 있다.

$$Sim(Q,S)=\frac{\sum_{w_{1k}=w_{2l}}(Wgt(w_{1k})*Wgt(w_{2l}))}{\sqrt{\sum_{k=1\cdots M}Wgt(w_{1k})^2}\sqrt{\sum_{l=1\cdots N}Wgt(w_{2l})^2}}*SentType(Q,S)$$

[0096] (5)

[0097] 여기서, $Sim(Q,S)$는 Q와 S사이의 주제 매칭 유사도를 표시하고, Q는 조회 청구를 표시하며, S는 후보 결과의 어구를 표시하고, $SentType(Q,S)$는 두 문장 유형 매칭의 가중치 계수를 표시하며, $Wgt(w_{1k})$는 조회 청구로부터 획득한 단어 w_{1k}의 유사도 가중치를 표시하고, M는 단어 w_{1k}의 수량을 표시하며, $Wgt(w_{2l})$는 후보 결과의 어구로부터 획득한 단어 w_{2l}의 유사도 가중치를 표시하고, N는 단어 w_{2l}의 수량이다. 여기서, 후보 결과의 어구 중의 일부 단어(예를 들면, 인접한 상문 및/또는 인접한 하문)는 클릭 전의 모형을 기반으로 조정을 진행한다.

[0098] 다시 도5를 참조하면, 단계(512)에서, 조회 청구와 후보 결과의 어구사이의 매칭 정황을 기반으로 전의 인자를 확정한다.

[0099] 단계(511)에서, 문장사이의 주제 매칭 유사도를 확정하고, 미시적으로 감안하여, 클릭 전의 모형을 기반으로 구체적 단어의 유사도 가중치를 조정한다. 이러한 단계(512)에서, 조회 청구와 후보 결과의 어구사이의 매칭 정황

에 다라, 즉, 거시적으로 감안하여, 하나의 전의 인자를 확정한다.

[0100] 조회 청구와 후보 결과의 어구사이의 매칭 정황은 예를 들면, 조회 청구 중 제일 중요한 단어가 후보 결과의 어구에 나타나지 않은 정황, 상하문의 매칭이 존재하는 정황, 및 상하문의 완전 매칭이 존재하지 않는 정황을 포함할 수 있다.

[0101] 조회 청구 중 제일 중요한 단어가 후보 결과의 어구에 나타나지 않을 경우, 이는 통상적으로 양자사이의 상관성이 비교적 낮고, 전의의 가능성이 비교적 높음을 표시한다. 이때, 전의 인자를 제1값(예를 들면, 0.7)으로 확정할 수 있다. 조회 청구 중의 단어의 중요성은 앞서 확정한 유사도 가중치를 기반으로 확정할 수 있다. 예를 들면, 직접 TF-IDF로 확정한 가중치에 따라 확정할 수 있다.

[0102] 상하문의 매칭이 존재한다는 것은 단어의 문자상의 매칭이외에 후보 결과에 해당 단어의 인접한 상문 또는 인접한 하문이 더 존재함을 가리킨다. 즉, 이때 후보 결과에도 전의의 가능성이 존재한다. 따라서, 전의 인자를 제2값으로 확정할 수 있다. 여기서, 제2값은 제1값보다 크다. 예를 들면, 제2값은 0.95이다.

[0103] 상하문의 완전 매칭이 존재하지 않는다는 것은 단어의 문자상의 매칭이외에 후보 결과에 해당 단어의 인접한 상문과 인접한 하문이 존재하지 않음을 가리킨다. 즉, 이때 후보 결과에는 기본상 전의의 가능성이 존재하지 않는다. 따라서, 전의 인자를 제3값으로 확정할 수 있다. 여기서, 제3값은 제2값보다 크다. 예를 들면, 제3값은 1이다.

[0104] 마지막으로, 단계(513)에서, 전의 인자 및 주제 매칭 유사도를 기반으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출한다.

[0105] 전의 인자 및 주제 매칭 유사도를 기반으로, 여러가지 함수 관계에 따라 어의 상관도를 구축할 수 있다. 일 구현에서, 하기 등식으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출할 수 있다.

[0106] $$Rele(Q,S)=\beta(Q,S)Sim(Q,S)$$

[0107] 여기서, $Rele(Q,S)$는 Q와 S사이의 어의 상관도를 표시하고, $\beta(Q,S)$는 Q와 S사이의 전의 인자를 표시하며, $Sim(Q,S)$는 Q와 S사이의 주제 매칭 유사도를 표시하고, Q는 조회 청구를 표시하며, S는 후보 결과의 어구를 표시한다.

[0108] 도면에서는 특정의 순서로 본 발명의 방법의 조작을 설명하였으나, 상기 특정 순서로 이러한 조작을 진행하여야 한다고 요구하거나 암시하는 것이 아니며 또는 도시된 모든 조작을 실행하여야만 기대하는 결과를 실현할 수 있는 것이 아님을 응당 주의하여야 한다. 반대로, 흐름도에 도시된 단계의 실행 순서는 바뀔수 있다. 부가적 또는 대안으로, 일부 단계를 생략할 수 있으며 다수의 단계를 한 단계로 합병하여 실행할 수 있으며, 및/또는 한 단계를 다수의 단계로 분할하여 실핼할 수 있다.

[0109] 도8은 본 발명의 실시예에 따른 검색 엔진의 예시적 구조 블록도를 보여준다.

[0110] 도8에 도시된 바와 같이, 검색 엔진(800)은 수신 유닛(810), 검색 유닛(820), 어의 상관도 확정 유닛(830) 및 순서배열 유닛(840)을 포함한다.

[0111] 수신 유닛(810)은 사용자가 입력한 조회 청구를 수신하도록 배치될 수 있다. 검색 유닛(820)은 조회 청구와 매칭되는 후보 결과를 조회하도록 배치될 수 있다. 어의 상관도 확정 유닛(830)는 클릭 전의 모형을 기반으로 조회 청구와 각 후보 결과사이의 어의 상관도를 확정하도록 배치될 수 있다. 순서배열 유닛(840)은 어의 상관도에 따라 후보 결과에 대해 순서 배열을 진행하도록 배치될 수 있다. 여기서, 클릭 전의 모형은 전의 사전 및/또는 비전의 사전을 포함하고, 전의 사전은 전의가 발생함을 확정한 검색 결과와 상응한 단어 및 그의 상하문을 포함하고, 비전의 사전은 전의가 발생하지 않음을 확정한 검색 결과의 상응한 단어 및 그의 상하문이 존재한다.

[0112] 일부 실시예에서, 어의 상관도 확정 유닛(830)은 각 후보 결과에 대하여 조회 청구와 후보 결과의 하나 또는 다수의 어의사이의 어의 상관도를 확정하는 산출 유닛(831)을 포함할 수 있고, 어구는 후보 결과의 타이틀, 앵커 문구 및 본문 중의 핵심 문장 중 적어도 하나를 포함한다. 어의 상관도 확정 유닛(830)은 확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 조회 청구와 후보 결과사이의 어의 상관도를 확정하는 확정 유닛(832)을 포함할 수 있다.

[0113] 일부 구현에서, 산출 유닛(831)은 클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 조회

청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하는 주제 매칭 유사도 모듈(미도시)를 포함할 수 있다.

[0114] 주제 매칭 유사도 모듈은, 구체적으로 단어 정렬을 이용하여 후보 결과의 어구로부터 조회 청구 중의 단어와 정렬된 인접한 상문과 하문을 확정하고, 전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구의 상응한 상문과 하문의 유사도 가중치를 조정하고, 조정된 유사도 가중치에 따라 문장사이의 문구 주제 매칭 모형을 이용하여 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출하도록 이용될 수 있다.

[0115] 산출 유닛(831)은 조회 청구와 후보 결과의 어구사이의 매칭 정황에 따라 전의 인자를 확정하는 전의 인자 모듈(미도시)을 더 포함할 수 있다.

[0116] 전의 인자 모듈은, 구체적으로 매칭 정황이 조회 청구 중 제일 중요한 단어가 후보 결과의 어구에 나타나지 않을 경우 전의 인자를 제1값으로 확정하고, 매칭 정황이 상하문의 매칭이 존재할 경우 전의 인자를 제2값으로 확정하며, 매칭 정황이 상하문의 완전 매칭이 존재하지 않을 경우 전의 인자를 제3값으로 확정하도록 이용될 수 있으며, 여기서, 제1값은 제2값보다 작고, 제2값은 제3값보다 작다.

[0117] 산출 유닛(831)은 전의 인자와 주제 매칭 유사도를 기반으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출하는 합성 모듈(미도시)을 더 포함할 수 있다.

[0118] 일부 실시예에서, 클릭 전의 모형 중의 전의 사전 및 비전의 사전은 조회 청구와 검색 결과Query-Title 쌍의 클릭 회수를 학습하여 구축한다.

[0119] 일부 구현에서, 전의 사전과 비전의 사전은, Query-Title 쌍의 클릭 표시 비율을 획득하고, 단어 정렬을 이용하여 검색 결과에서 조회 어구 중의 단어와 정렬되는 인접한 상하문를 획득하고, 클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 전의 사전에 첨가하며, 클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 원생 비전의 사전에 첨가하는 방법으로 구축되는 원생 전의 사전과 원생 비전의 사전를 포함하고, 상기 클릭 표시 비율은 클릭 회수와 표시 회수의 비율이고, 표시 회수은 검색 결과가 조회 청구에 응하여 표시되는 회수를 지시하고, 클릭 회수는 검색 결과가 조회 청구에 응하여 표시될 때 사용자에 의해 클릭되는 회수를 지시한다.

[0120] 선택적 또는 부가적으로, 전의 사전 및 비전의 사전은, 조회 청구 중의 단어에 대해 어의 유형을 표기하고, 표기된 어의 유형을 이용하여 원생 전의 사전과 원생 비전의 사전에 대응되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 구축하는 방법으로 구축되는 추상화 된 전의 사전과 추상화 된 비전의 사전을 더 포함한다.

[0121] 검색 엔진(800) 중의 기재된 여러 유닛 또는 서브 유닛은 앞서 흐름도를 참조하여 설명한 방법중의 각 단계에 대응되는 것을 응당 자명하여야 할 것이다. 따라서, 상기에 방법에 대해 설명한 조작과 특징은 검색 엔진(800) 및 이에 포함되는 유닛에도 적용될 수 있으며 이에 대한 중복된 설명은 생략한다.

[0122] 도9는 본 발명의 실시예를 실현하기 위한 컴퓨터 시스템(900)를 보여준다.

[0123] 도9에 도시된 바와 같이, 컴퓨터 시스템(900)은 중앙 처리 유닛(CPU)(901)을 포함하고, 이는 읽기 전용 메모리 장치(ROM)(902)에 저장된 프로그램 또는 저장부(908)로부터 랜덤 액세스 메모리 장치(RAM)(903)에 로딩되는 프로그램에 의하여 각종 적당한 동작 및 처리를 실행할 수 있다. RAM(903)에는 시스템(900) 조작에 필요한 각종 프로그램 및 데이터들이 더 포함되어 있다. CPU(901), ROM(902) 및 RAM(903)은 버스라인(904)을 통하여 서로 연결된다. 입력/출력(I/O) 인터페이스(905)도 버스라인(904)에 연결된다.

[0124] 키보드, 마우스 등을 포함하는 입력부(906); 음극선관(CRT), 액정 표시 장치(LCD) 등 및 스피커 등을 포함하는 출력부(909); 하드 디스크 등을 포함하는 저장부(908); 및 LAN카드, 변복조 장치 등과 같은 네트워크 에세스 카드를 포함하는 통신부(909);를 포함하는 구성요소는 I/O 인터페이스(905)에 연결된다. 통신부(909)는 인터넷과 같은 네트워크를 통하여 통신 처리를 실행한다. 구동부(910)는 수요에 따라 I/O 인터페이스(905)에 연결된다. 구동부(910)에서 판독된 컴퓨터 프로그램이 수요에 따라 저장부(908)에 설치되도록 구동부(910)에는 수요에 따라 디스크, 콤팩트디스크, 광자기 디스크, 반도체 메모리 장치 등과 같은 착탈 가능한 매질(911)이 설치된다.

[0125] 특히, 본 발명의 실시예에 의하면, 도2 내지 도7을 참조하여 설명한 프로세스는 컴퓨터 소프트웨어 프로그램으로 실현할 수 있다. 예를 들면, 본 발명의 실시예는 일 컴퓨터 프로그램 제품을 포함한다. 상기 컴퓨터 프로그램 제품은 유형적으로 컴퓨터 판독 가능한 매질에 포함되는 컴퓨터 프로그램을 포함하되, 컴퓨터 프로그램은 도2 내지 도7의 방법을 실행하기 위한 프로그램 코드를 포함한다. 이러한 실시예에서, 상기 컴퓨터 프로그램은 통

신부(909)를 통하여 네트워크로부터 다운로드되어 설치되고, 및/또는 착탈 가능한 매질(911)로부터 설치될수 있다.

[0126] 첨부한 도면중의 흐름도 및 블록도는 본 발명의 여려 실시예에 따른 시스템, 방법, 컴퓨터 프로그램 제품의 실시 가능한 체계구조, 기능 및 동작을 도시하였다. 이러한 방면에 있어서, 흐름도 또는 블록도 중의 각 블록은 하나의 모듈, 프로그램 세그먼트, 또는 코드의 일부분을 대표하고, 상기 모듈, 프로그램 세그먼트, 또는 코드의 일부분은 소정의 로직 기능을 실현하기 위한 하나이상의 실행가능한 명령을 포함한다. 일부 대체 실시예에서, 블록에 표기된 기능은 도면에 표기된 순서와 다른 순서로 진행될 수 있음을 자명하여야 할 것이다. 예를 들면, 연속되게 표시된 두개의 블록은 사실상 관련된 기능에 의하여 기본적으로 병렬되게 진행될 수 있으며, 반대된 순서로 진행될 수도 있다. 블록도 및/또는 흐름도의 각 블록 및 블록도 및/또는 흐름도의 블록의 조합은 소정의 기능 또는 동작을 진행하는 하드웨어를 기반으로하는 전용의 시스템으로 실현하거나, 전용의 하드웨어 및 컴퓨터 명령의 조합으로 실현할 수 있다.

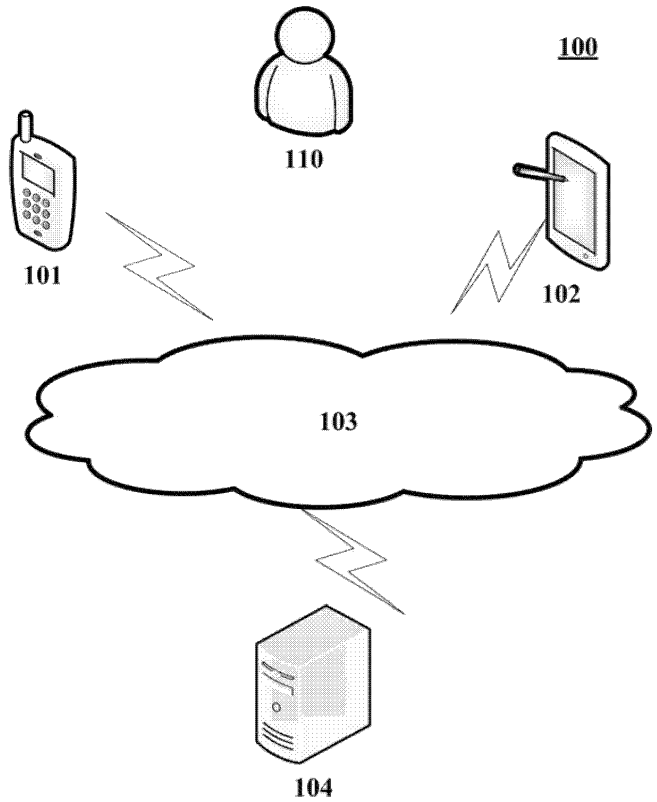
[0127] 본 발명에 설명된 관련된 유닛 또는 모듈은 소프트 웨어 방식으로 실현할 수 있으며, 하드 웨어 방식으로 실현할 수도 있다. 설명된 유닛 또는 모듈은 프로세서에 설치될 수 있다. 이러한 유닛 또는 모듈의 명칭은 일부 경우에 해당 유닛 또는 모듈 자체를 한정하지 않는다.

[0128] 한편, 본 발명은 또한 컴퓨터 판독 가능한 기록 매체를 제공한다. 이러한 컴퓨터 판독 가능한 기록 매체는 상기 실시예중 상기 장치에 포함되는 컴퓨터 판독 가능한 기록 매체이거나, 장치에 설치되지 않은 독립적으로 존재하는 컴퓨터 판독 가능한 기록 매체일 수 있다. 컴퓨터 판독 가능한 기록 매체에는 하나이상의 프로그램이 저장되어 있을수 있고, 하나이상의 프로세서는 이러한 프로그램으로 본 발명에 설명된 공식 입력 방법을 진행한다.

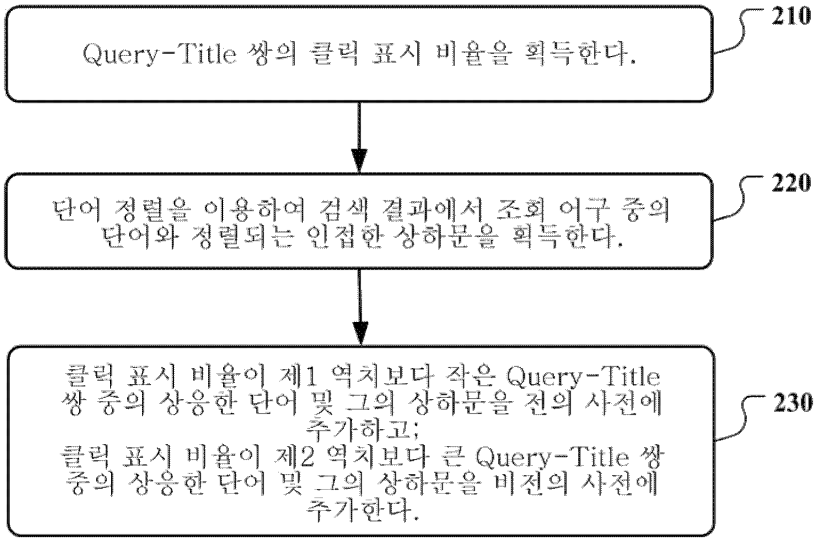
[0129] 이상의 설명은 오직 본 발명의 바람직한 실시예 및 이용하는 기술 원리에 대한 설명일 뿐이다. 본 발명의 청구범위는 상기 기술적 특징의 특정 조합으로 이루어진 기술적 방안에 한정되는 것이 아니라, 본 발명의 사상을 벗어나지 않는 한 상기 기술적 특징 또는 그의 등가 특징들의 임의의 조합으로 이루어진 기타 기술적 방안도 포함하는 것을 본 분야에서 통상 지식을 가진자는 자명할 것이다. 상기 특징과 본 발명에 개시된 유사한 기능을 구비한 기술적 특징을 서로 교체하여 형성된 기술적방안을 예로 들수 있으나, 이에 한정된 것은 아니다.

도면

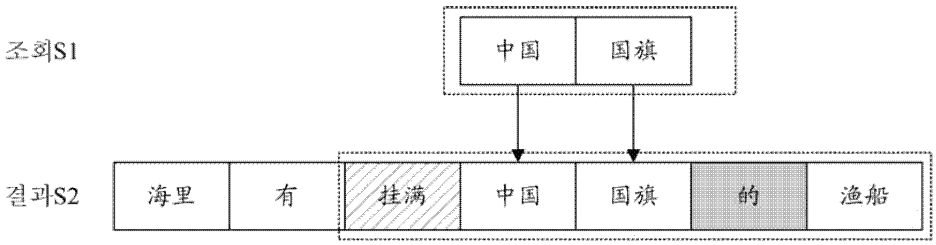
도면1



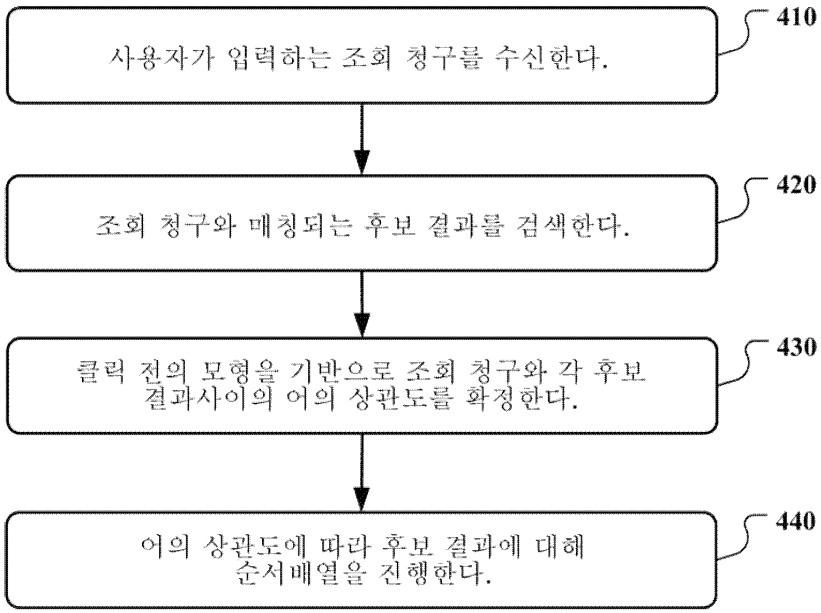
도면2


Query-Title 쌍의 클릭 표시 비율을 획득한다.
210
단어 정렬을 이용하여 검색 결과에서 조회 어구 중의 단어와 정렬되는 인접한 상하문을 획득한다.
220
클릭 표시 비율이 제1 역치보다 작은 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 전의 사전에 추가하고;
클릭 표시 비율이 제2 역치보다 큰 Query-Title 쌍 중의 상응한 단어 및 그의 상하문을 비전의 사전에 추가한다.
230

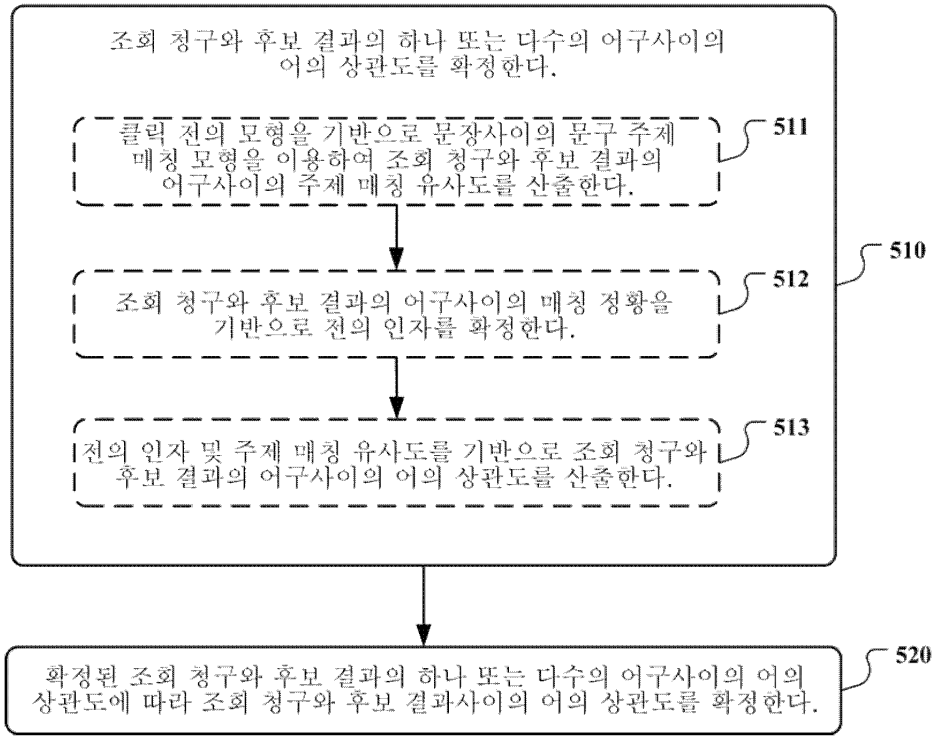
도면3


조회S1
中国
国旗
결과S2
海里
有
挂满
中国
国旗
的
渔船

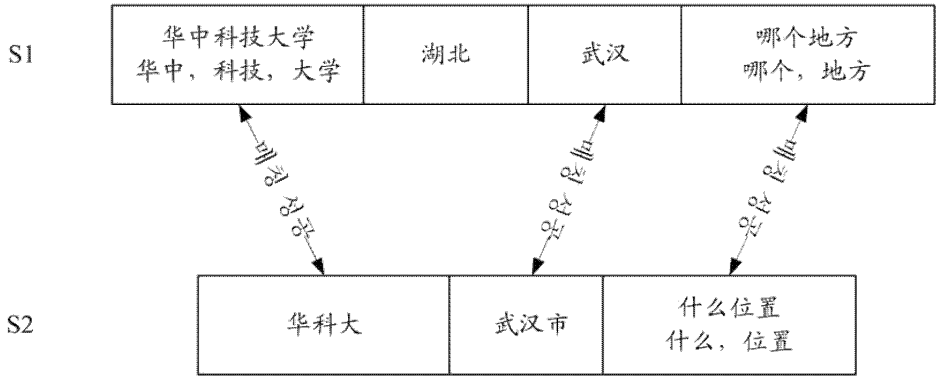
도면4


사용자가 입력하는 조회 청구를 수신한다.
410
조회 청구와 매칭되는 후보 결과를 검색한다.
420
클릭 전의 모형을 기반으로 조회 청구와 각 후보 결과사이의 어의 상관도를 확정한다.
430
어의 상관도에 따라 후보 결과에 대해 순서배열을 진행한다.
440

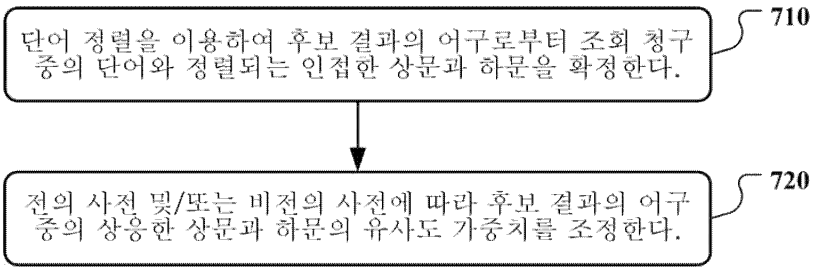
도면5


조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도를 확정한다.
클릭 전의 모형을 기반으로 문장사이의 문구 주제 매칭 모형을 이용하여 조회 청구와 후보 결과의 어구사이의 주제 매칭 유사도를 산출한다.
511
조회 청구와 후보 결과의 어구사이의 매칭 정황을 기반으로 전의 인자를 확정한다.
512
전의 인자 및 주제 매칭 유사도를 기반으로 조회 청구와 후보 결과의 어구사이의 어의 상관도를 산출한다.
513
510
확정된 조회 청구와 후보 결과의 하나 또는 다수의 어구사이의 어의 상관도에 따라 조회 청구와 후보 결과사이의 어의 상관도를 확정한다.
520

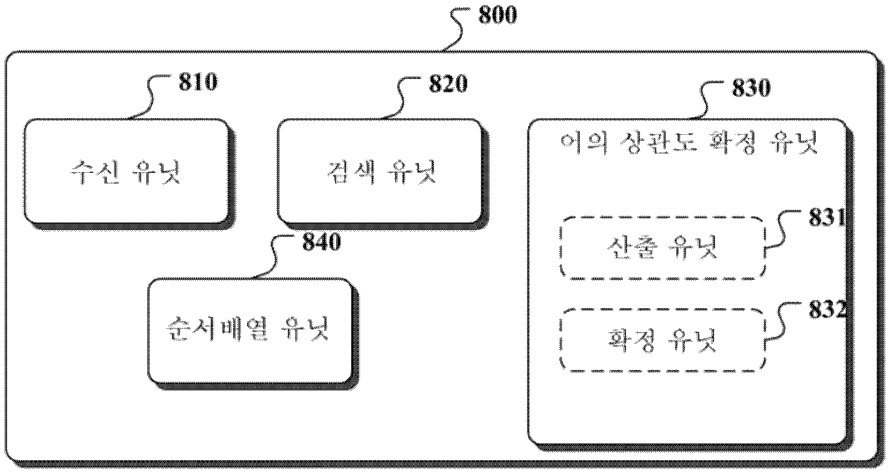
도면6


S1
华中科技大学
华中，科技，大学
湖北
武汉
哪个地方
哪个，地方
매칭 정황
S2
华科大
武汉市
什么位置
什么，位置

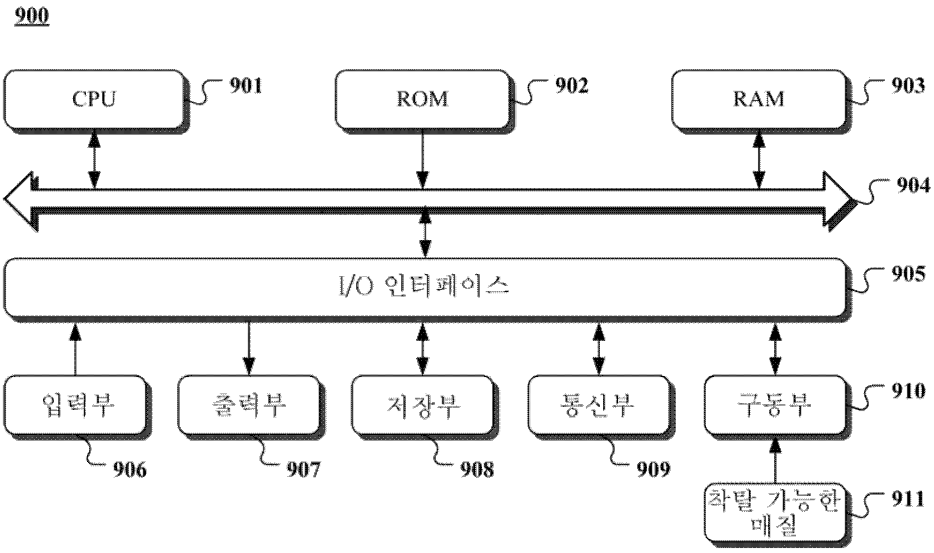
도면7


단어 정렬을 이용하여 후보 결과의 어구로부터 조회 청구 중의 단어와 정렬되는 인접한 상문과 하문을 확정한다.
710
전의 사전 및/또는 비전의 사전에 따라 후보 결과의 어구 중의 상응한 상문과 하문의 유사도 가중치를 조정한다.
720

도면8


800
810
수신 유닛
820
검색 유닛
830
어의 상관도 확정 유닛
831
산출 유닛
832
확정 유닛
840
순서배열 유닛

도면9


900
CPU
901
ROM
902
RAM
903
904
I/O 인터페이스
905
입력부
906
출력부
907
저장부
908
통신부
909
구동부
910
착탈 가능한 매질
911