(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년05월02일
(11) 등록번호 10-1140187
(24) 등록일자 2012년04월19일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2006-7006282
(22) 출원일자(국제) 2004년09월13일
심사청구일자 2009년09월10일
(85) 번역문제출일자 2006년03월30일
(65) 공개번호 10-2006-0090689
(43) 공개일자 2006년08월14일
(86) 국제출원번호 PCT/US2004/029772
(87) 국제공개번호 WO 2005/033967
국제공개일자 2005년04월14일
(30) 우선권주장
10/676,724 2003년09월30일 미국(US)
(56) 선행기술조사문헌
US06360196 B1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
구글 잉크.
미국 캘리포니아 마운틴 뷰 앰피씨어터 파크웨이 1600 (우편번호 94043)
(72) 발명자
미탈 비부
미국 94087 캘리포니아주 서니베일 엘소나 드라이브 1327
폰테 제이 엠
미국 94043 캘리포니아주 마운틴 뷰 마조리 코트 2439
(뒷면에 계속)
(74) 대리인
특허법인코리아나

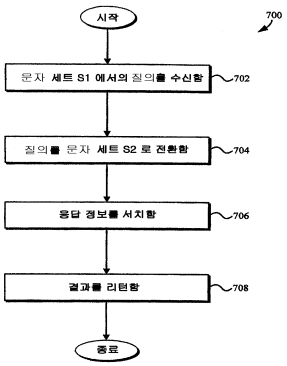
전체 청구항 수 : 총 52 항 심사관 : 이종익

(54) 발명의 명칭 **타깃 페이지와는 상이한 문자-세트 및/또는 언어로 기입된 질의를 이용하여 서치하는 시스템 및 방법**

(57) 요 약

본 발명과 부합하는 방법 및 장치는 사용자로 하여금 모호한 서치 질의를 제출하고 적절한 서치 결과를 수신하게 한다. 질의는, 서치될 데이터의 적어도 일부의 문자 세트 및/또는 언어와 상이한 문자 세트 및/또는 언어를 사용하여 표현될 수 있다. 이들 문자 세트 및/또는 언어 간의 전환은 정렬된 텍스트에서 용어의 사용을 검사함으로써 수행될 수 있다. 확률은 각각의 가능한 전환과 관련될 수 있다. 서치 결과와의 사용자 상호작용을 검사함으로써, 이들 확률에 대한 개선이 가능할 수 있다.

대 표 도 - 도7



(72) 발명자

사하미 메흐란

미국 94063 캘리포니아주 레드우드 시티 후버 스트리트 3238

게마와트 산제이

미국 94043 캘리포니아주 마운틴 뷰 노스 렌프스토프 애비뉴 111

바우어 존 에이

미국 94040 캘리포니아주 마운틴 뷰 델 메디오 애비뉴 415 넘버8

특허청구의 범위

청구항 1

삭제

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

사용자로부터, 제 1 포맷으로 기입된 질의를 획득하는 단계;

확률 딕셔너리를 이용하여, 상기 질의를 제 2 포맷으로 전환하는 단계로서, 상기 확률 딕셔너리는 용어를 상기 제 1 포맷으로부터 상기 제 2 포맷으로 매핑하는, 상기 전환 단계;

상기 전환된 질의에 응답하는 정보에 대한 데이터베이스를 서치하는 단계;

상기 제 2 포맷으로 기입된 서치 결과를 상기 사용자에게 리턴하는 단계;

상기 사용자로부터의 서치 결과 선택을 획득하는 단계; 및

상기 서치 결과 선택을 이용하여 용어 매핑의 상기 확률 딕셔너리를 수정하는 단계를 포함하는, 서치 방법.

청구항 15

삭제

청구항 16

제 14 항에 있어서,

상기 수정은, 상기 확률 딕셔너리에서의 하나 이상의 매핑과 관련된 하나 이상의 확률을 조정하는 단계를 포함하는, 서치 방법.

청구항 17

제 14 항에 있어서,

상기 질의를 제 2 포맷으로 전환하는 단계는 상기 질의를 확장하는 단계를 포함하는, 서치 방법.

청구항 18

제 17 항에 있어서,

상기 확장된 질의는 상기 질의 용어의 대안적인 인코딩을 포함하는, 서치 방법.

청구항 19

제 17 항에 있어서,

상기 확장된 질의는 상기 질의 용어의 대안적인 언어 전환을 포함하는, 서치 방법.

청구항 20

제 17 항에 있어서,

상기 확장된 질의는 상기 질의 용어의 대안적인 인코딩 및 대안적인 언어 전환을 포함하는, 서치 방법.

청구항 21

제 18 항에 있어서,

상기 확장된 질의는 상기 질의 용어의 상기 대안적인 인코딩의 동의어를 포함하는, 서치 방법.

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

제 1 포맷으로 기입된 하나 이상의 질의 용어를 포함하는 질의를 수신하는 단계;

상기 질의 용어를 제 2 포맷으로 기입된 복수의 변형체 (variants) 로 전환하는 단계; 및

상기 변형체 중 하나 이상을 이용하여, 상기 질의에 응답하는, 상기 제 2 포맷으로 기입된 정보를 서치하는 단계를 포함하는, 방법.

청구항 40

제 39 항에 있어서,

상기 제 1 포맷은 전화 키패드로부터 입력된 숫자의 시퀀스를 포함하며,

상기 제 2 포맷은 문자숫자 텍스트를 포함하는, 방법.

청구항 41

제 39 항에 있어서,

미리 정의된 어휘집의 일부가 아닌 상기 복수의 변형체에서의 변형체를 폐기함으로써 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 방법.

청구항 42

제 39 항에 있어서,

미리 정의된 낮은 확률의 문자 조합을 포함하는 상기 복수의 변형체에서의 변형체를 폐기함으로써 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 방법.

청구항 43

제 39 항에 있어서,

상기 제 1 포맷은, 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하며,

상기 제 2 포맷은, 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 방법.

청구항 44

전화 키패드로부터 입력된 숫자 질의를 수신하는 단계;

상기 숫자 질의를 제 1 포맷의 잠재적인 문자숫자 전환의 그룹으로 전환하는 단계;

미리 정의된 낮은 확률의 문자 조합을 포함하도록 결정되는 잠재적인 전환을 폐기하는 단계;

나머지 문자숫자 전환을 확률 딕셔너리를 이용하여, 상기 제 1 포맷으로부터 제 2 포맷으로 전환하는 단계; 및

상기 제 2 포맷의 상기 문자숫자 전환을 이용하여 서치를 수행하는 단계를 포함하는, 방법.

청구항 45

제 44 항에 있어서,

상기 제 1 포맷은, 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 텍스트를 포함하며,

상기 제 2 포맷은, 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 텍스트를 포함하는, 방법.

청구항 46

제 39 항에 있어서,

제 2 포맷으로 기입된 상기 복수의 변형체를 제 3 포맷으로 전환하는데 확률 딕셔너리가 사용되고,

상기 질의에 응답하는 정보를 서치하는데 상기 제 3 포맷으로 기입된 상기 복수의 변형체가 이용되는, 방법.

청구항 47

제 39 항에 있어서,

상기 전환하는 단계는 번호의 시퀀스 내의 각각의 번호를 상기 번호와 연관된 하나 이상의 문자숫자 문자(character)와 매핑하는 단계를 포함하고,

상기 번호에 대한 하나 이상의 문자숫자 균등물을 생성하는데 상기 하나 이상의 문자숫자 문자가 이용되는, 방법.

청구항 48

제 39 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 함께 나타나는 소정의 워드의 확률에 기초하여 결정되는, 방법.

청구항 49

제 48 항에 있어서,

상기 확률은 베이지안 방법, 히스토그램 평활화, 커넬 평활화, 및 축소 추정량 중 적어도 하나를 이용하여 획득되는, 방법.

청구항 50

제 39 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 상기 문자숫자 변형체에 포함된, 기 제출된 서치 질의 카운트에 기초하여 결정되는, 방법.

청구항 51

제 39 항에 있어서,

상기 제 1 포맷은 로마자로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택되는 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 방법.

청구항 52

제 39 항에 있어서,

상기 전환하는 단계는 확률 딕셔너리를 이용하여 수행되고,

상기 확률 딕셔너리는 용어를 상기 제 1 포맷으로부터 상기 제 2 포맷으로 매핑하는, 방법.

청구항 53

제 52 항에 있어서,

상기 제 2 포맷으로 기입된 서치 결과를 사용자에게 리턴하는 단계;

상기 사용자로부터의 서치 결과 선택을 획득하는 단계; 및

상기 서치 결과 선택을 이용하여 용어 매핑의 상기 확률 딕셔너리를 수정하는 단계를 더 포함하는, 방법.

청구항 54

프로그램 제품을 포함하는 컴퓨터 판독가능 매체; 및

상기 프로그램 제품을 실행하고, 다음의 동작을 수행하도록 구성된 하나 이상의 프로세서를 포함하고,

상기 동작은

제 1 포맷으로 기입된 적어도 하나의 질의 용어를 포함하는 질의를 수신하는 단계;

상기 질의 용어를 제 2 포맷으로 기입된 복수의 변형체로 전환하는 단계; 및

하나 이상의 상기 변형체를 이용하여 상기 질의에 응답하는, 상기 제 2 포맷으로 기입된 정보를 서치하는 단계

를 포함하는, 시스템.

청구항 55

제 54 항에 있어서,

상기 제 1 포맷은 전화 키패드로부터 입력된 번호의 시퀀스를 포함하고,

상기 제 2 포맷은 문자숫자 텍스트를 포함하는, 시스템.

청구항 56

제 54 항에 있어서,

미리 정의된 어휘집의 일부가 아닌 상기 복수의 변형체에서의 변형체를 폐기함으로써 상기 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 시스템.

청구항 57

제 54 항에 있어서,

미리 정의된 낮은 확률의 문자 조합을 포함하는 상기 복수의 변형체에서의 변형체를 폐기함으로써 상기 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 시스템.

청구항 58

제 54 항에 있어서,

상기 제 1 포맷은 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 시스템.

청구항 59

제 54 항에 있어서,

제 2 포맷으로 기입된 상기 복수의 변형체를 제 3 포맷으로 전환하는데 확률 딕셔너리가 사용되고,

상기 질의에 응답하는 정보를 서치하는데 상기 제 3 포맷으로 기입된 상기 복수의 변형체가 이용되는, 시스템.

청구항 60

제 54 항에 있어서,

상기 전환하는 단계는 번호의 시퀀스 내의 각각의 번호를 상기 번호와 연관된 하나 이상의 문자숫자 문자와 매핑하는 단계를 포함하고,

상기 번호에 대한 하나 이상의 문자숫자 균등물을 생성하는데 상기 하나 이상의 문자숫자 문자가 이용되는, 시스템.

청구항 61

제 54 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 함께 나타나는 소정의 워드의 확률에 기초하여 결정되는, 시스템.

청구항 62

제 61 항에 있어서,

상기 확률은 베이지안 방법, 히스토그램 평활화, 커넬 평활화, 및 축소 추정량 중 적어도 하나를 이용하여 획득되는, 시스템.

청구항 63

제 54 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 상기 문자숫자 변형체에 포함된, 기 제출된 서치 질의 카운트에 기초하여 결정되는, 시스템.

청구항 64

제 54 항에 있어서,

상기 제 1 포맷은 로마자로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택되는 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 시스템.

청구항 65

제 54 항에 있어서,

상기 전환하는 단계는 확률 딕셔너리를 이용하여 수행되고,

상기 확률 딕셔너리는 용어를 상기 제 1 포맷으로부터 상기 제 2 포맷으로 매핑하는, 시스템.

청구항 66

제 65 항에 있어서,

상기 제 2 포맷으로 기입된 서치 결과를 사용자에게 리턴하는 단계;

상기 사용자로부터의 서치 결과 선택을 획득하는 단계; 및

상기 서치 결과 선택을 이용하여 용어 매핑의 상기 확률 딕셔너리를 수정하는 단계를 더 포함하는, 시스템.

청구항 67

프로그램 제품을 포함하는 컴퓨터 판독가능 매체; 및

상기 프로그램 제품을 실행하고, 다음의 동작을 수행하도록 구성된 하나 이상의 프로세서를 포함하고,

상기 동작은

전화 키패드로부터 입력된 숫자 질의를 수신하는 단계;

상기 숫자 질의를 제 1 포맷의 잠재적인 문자숫자 전환의 그룹으로 전환하는 단계;

미리 정의된 낮은 확률의 문자 조합을 포함하도록 결정되는 잠재적인 전환을 폐기하는 단계;

나머지 문자숫자 전환을 확률 딕셔너리를 이용하여, 상기 제 1 포맷으로부터 제 2 포맷으로 전환하는 단계; 및

상기 제 2 포맷의 상기 문자숫자 전환을 이용하여 서치를 수행하는 단계를 포함하는, 시스템.

청구항 68

제 67 항에 있어서,

상기 제 1 포맷은 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터

선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 시스템.

청구항 69

컴퓨터 프로그램이 인코딩된 컴퓨터 저장 매체로서,

상기 프로그램은 데이터 프로세싱 장치에 의해 실행될 때 상기 데이터 프로세싱 장치가 다음의 동작을 수행하도록 야기하는 명령을 포함하고,

상기 동작은

제 1 포맷으로 기입된 적어도 하나의 질의 용어를 포함하는 질의를 수신하는 단계;

상기 질의 용어를 제 2 포맷으로 기입된 복수의 변형체로 전환하는 단계; 및

하나 이상의 상기 변형체를 이용하여 상기 질의에 응답하는, 상기 제 2 포맷으로 기입된 정보를 서치하는 단계를 포함하는, 컴퓨터 저장 매체.

청구항 70

제 69 항에 있어서,

상기 제 1 포맷은 전화 키패드로부터 입력된 번호의 시퀀스를 포함하고,

상기 제 2 포맷은 문자숫자 텍스트를 포함하는, 컴퓨터 저장 매체.

청구항 71

제 69 항에 있어서,

미리 정의된 어휘집의 일부가 아닌 상기 복수의 변형체에서의 변형체를 폐기함으로써 상기 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 컴퓨터 저장 매체.

청구항 72

제 69 항에 있어서,

미리 정의된 낮은 확률의 문자 조합을 포함하는 상기 복수의 변형체에서의 변형체를 폐기함으로써 상기 하나 이상의 변형체를 획득하는 단계를 더 포함하는, 컴퓨터 저장 매체.

청구항 73

제 69 항에 있어서,

상기 제 1 포맷은 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 컴퓨터 저장 매체.

청구항 74

제 69 항에 있어서,

제 2 포맷으로 기입된 상기 복수의 변형체를 제 3 포맷으로 전환하는데 확률 딕셔너리가 사용되고,

상기 질의에 응답하는 정보를 서치하는데 상기 제 3 포맷으로 기입된 상기 복수의 변형체가 이용되는, 컴퓨터 저장 매체.

청구항 75

제 69 항에 있어서,

상기 전환하는 단계는 번호의 시퀀스 내의 각각의 번호를 상기 번호와 연관된 하나 이상의 문자숫자 문자와 매핑하는 단계를 포함하고,

상기 번호에 대한 하나 이상의 문자숫자 균등물을 생성하는데 상기 하나 이상의 문자숫자 문자가 이용되는, 컴퓨터 저장 매체.

청구항 76

제 69 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 함께 나타나는 소정의 워드의 확률에 기초하여 결정되는, 컴퓨터 저장 매체.

청구항 77

제 76 항에 있어서,

상기 확률은 베이지안 방법, 히스토그램 평활화, 커넬 평활화, 및 축소 추정량 중 적어도 하나를 이용하여 획득되는, 컴퓨터 저장 매체.

청구항 78

제 69 항에 있어서,

상기 전환하는 단계는 전화 키패드로부터 입력된 번호의 시퀀스를 하나 이상의 문자숫자 변형체와 매핑하는 단계를 포함하고,

하나의 문자숫자 변형체는 상기 문자숫자 변형체에 포함된, 기 제출된 서치 질의 카운트에 기초하여 결정되는, 컴퓨터 저장 매체.

청구항 79

제 69 항에 있어서,

상기 제 1 포맷은 로마자로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택되는 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 컴퓨터 저장 매체.

청구항 80

제 69 항에 있어서,

상기 전환하는 단계는 확률 딕셔너리를 이용하여 수행되고,

상기 확률 딕셔너리는 용어를 상기 제 1 포맷으로부터 상기 제 2 포맷으로 매핑하는, 컴퓨터 저장 매체.

청구항 81

제 80 항에 있어서,

상기 제 2 포맷으로 기입된 서치 결과를 사용자에게 리턴하는 단계;

상기 사용자로부터의 서치 결과 선택을 획득하는 단계; 및

상기 서치 결과 선택을 이용하여 용어 매핑의 상기 확률 딕셔너리를 수정하는 단계를 더 포함하는, 컴퓨터 저장 매체.

청구항 82

컴퓨터 프로그램이 인코딩된 컴퓨터 저장 매체로서,

상기 프로그램은 데이터 프로세싱 장치에 의해 실행될 때 상기 데이터 프로세싱 장치가 다음의 동작을 수행하도록 야기하는 명령을 포함하고,

상기 동작은

전화 키패드로부터 입력된 숫자 질의를 수신하는 단계;

상기 숫자 질의를 제 1 포맷의 잠재적인 문자숫자 전환의 그룹으로 전환하는 단계;

미리 정의된 낮은 확률의 문자 조합을 포함하도록 결정되는 잠재적인 전환을 폐기하는 단계;

나머지 문자숫자 전환을 확률 딕셔너리를 이용하여, 상기 제 1 포맷으로부터 제 2 포맷으로 전환하는 단계; 및

상기 제 2 포맷의 상기 문자숫자 전환을 이용하여 서치를 수행하는 단계를 포함하는, 컴퓨터 저장 매체.

청구항 83

제 82 항에 있어서,

상기 제 1 포맷은 로마지, 로마자, 및 핀인으로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하고,

상기 제 2 포맷은 간지, 가타가나, 히라가나, 한글, 한자, 및 전통적인 중국어 문자로 이루어진 그룹으로부터 선택된 문자 세트로 기입된 문자숫자 텍스트를 포함하는, 컴퓨터 저장 매체.

명 세 서

[0001] **관련 출원에 대한 상호 참조**

[0002] 본 출원은, "DATA ENTRY AND SEARCH FOR HANDHELD DEVICES" 인 명칭으로 2000년 7월 6일자로 출원된 미국특허 가출원 제 60/216,530 호에 대해 35 U.S.C.§119(e) 에 따라 우선권 주장하는, "METHODS AND APPARATUS FOR PROVIDING SEARCH RESULTS IN RESPONSE TO AN AMBIGUOUS SEARCH QUERY" 인 명칭으로 2000년 12월 26일자로 출원된 미국특허 출원번호 제 09/748,431 호의 일부 계속출원이며, 이들 양 출원은 전부 여기에 참조로서 포함된다.

[0003] **발명의 배경**

[0004] **1. 발명의 기술분야**

[0005] 본 발명은 일반적으로 정보 서치 및 검색에 관한 것이다. 좀더 상세하게는, 서치되는 문헌의 적어도 일부의 문자 (character) 세트 또는 언어와 상이한 문자 세트 또는 언어로 기입된 질의 (query) 를 이용하여 서치를 수행하는 시스템 및 방법이 개시된다.

[0006] **2. 관련 기술의 설명**

[0007] 대부분의 서치 엔진은, 최종 사용자가 종래의 키보드와 같은 어떤 것을 이용하여 서치 질의 (search query) 를 입력하고 있다는 가정 하에 동작하며, 여기서, 문자숫자 (alphanumeric) 스트링의 입력은 어렵지 않다. 그러나, 소형 디바이스가 더 일반화됨에 따라, 이러한 가정이 언제나 유효한 것은 아니다. 예를 들어, 사용자들은, 무선 애플리케이션 프로토콜 (WAP) 표준을 지원하는 무선 전화를 이용하여 서치 엔진에 질의할 수도 있다. 통상적으로, 무선 전화와 같은 디바이스는, 사용자에 의한 특정 액션 (예를 들어, 키를 누르는 것) 이 2 개 이상의 문자숫자 문자에 대응할 수도 있는 데이터 입력 인터페이스를 가진다. WAP 구조의 상세한 설명은 http://www1.wapforum.org/tech/documents/SPEC-WAPArch-19980439.pdf ("WAP 100 무선 애플리케이션 프로토콜 구조 명세") 에서 입수가능하다.

[0008] 통상적인 경우, WAP 사용자는 서치 질의 페이지로 네비게이션하며, 자신의 서치 질의를 입력하는 양식을 제공받는다. 종래의 방법의 경우, 사용자는 특정 문자 (letter) 를 선택하기 위하여 다중의 키를 누르는 것이 요구될 수도 있다. 표준 전화 키패드 상에서, 사용자는 "2" 키를 2회 누름으로써 문자 "b" 를 선택하고, 또는, "7" 키를 4회 누름으로써 문자 "s" 를 선택한다. 따라서, "ben smith" 에 대한 질의를 입력하기 위하여, 통상적으로, 사용자는 다음과 같은 키 누름의 스트링, 즉, 223366077776444844 를 입력하는 것이 필요하며, 이는 다음과 같은 문자에 매핑한다.

22 → b

33 → e

66 → n

0 → 공백

7777 → s

6 → m

444 → i

8 → t

44 → h

[0009]

[0010] 사용자가 자신의 서치 요청을 입력한 후, 서치 엔진은 사용자로부터 워드 또는 워드들을 수신하며, 마치 사용자가 종래의 키보드를 사용한 데스크탑 브라우저로부터 요청을 수신한 것과 동일한 방식으로 대부분 진행한다.

[0011] 전술한 예로부터 알 수 있는 바와 같이, 이러한 형태의 데이터 입력은, "ben smith" 에 대응하는 (공백을 포함하여) 9 개의 문자숫자 문자를 입력하기 위하여 18회의 키 스트로크 (key stroke) 를 요구한다는 점에 있어서 비효율적이다.

[0012] 유사한 난점이, 비-타깃-언어 (non-target-language) 키보드를 이용하여 질의를 타이핑할 경우에 발생할 수도 있다. 예를 들어, 일본어 텍스트는 히라가나, 가타가나, 및 간지 (kanji) 를 포함하여 다양한 서로 다른 문자 세트를 이용하여 표현될 수 있지만, 이들 어떠한 것도 로마 알파벳에 기초한 통상의 ASCII 키보드를 이용하여 용이하게 입력되지 않는다. 그러한 상황에서, 사용자는, 종종, 일본국 Tokushima 시 소재의 JustSystem Corp. 에 의해 제조된 Ichitaro 와 같은 워드-프로세서를 이용하는데, 이는 로마지 (romaji; 일본어의 표음의 로마-알파벳 표현) 로 기입된 텍스트를 히라가나, 가타가나, 및 간지로 변환할 수 있다. 그 워드 프로세서를 사용하여, 사용자는 질의를 로마지로 타이핑한 후, 워드 프로세서의 스크린으로부터 전환된 (translated) 텍스트를 브라우저 상의 서치 박스에 컷-앤-페이스트 (cut-and-paste) 할 수 있다. 이러한 접근법의 단점은, 비교적 느리고 지루할 수 있고, 사용자가 워드 프로세서의 카피 (copy) 에 액세스할 것을 요구하며, 이는 비용 및/또는 메모리 제약으로 인해 실현 불가능할 수도 있다는 것이다.

[0013] 따라서, 모호한 서치 질의에 응답하여 적절한 서치 결과를 제공하는 방법 및 장치가 요구된다.

[0014] **발명의 개요**

[0015] 여기에서 구현되고 대략적으로 설명된 바와 같이, 본 발명과 부합하는 방법 및 장치는 모호한 서치 질의에 응답하여 적절한 서치 결과를 제공한다. 본 발명과 부합하여, 그러한 방법은 사용자로부터 모호한 정보 컴포넌트의 시퀀스를 수신하는 단계를 포함한다. 그 방법은, 모호한 정보 컴포넌트를 덜 모호한 정보 컴포넌트에 매핑하는 매핑 정보를 획득한다. 이러한 매핑 정보는, 모호한 정보 컴포넌트의 시퀀스를 덜 모호한 정보 컴포넌트의 하나 이상의 대응하는 시퀀스로 전환하는데 이용된다. 덜 모호한 정보의 이러한 시퀀스 중 하나 이상은 서치 엔진에 입력으로서 제공된다. 서치 엔진으로부터 서치 결과가 획득되어, 사용자에게 제공된다.

[0016] 또한, 서치되는 문헌 중 적어도 일부의 문자 세트 또는 언어와 상이한 문자 세트 또는 언어로 표현되는 질의를 사용하여 서치를 수행하는 시스템 및 방법이 개시된다. 본 발명의 실시형태들은 사용자로 하여금 표준 입력 디바이스 (예를 들어, ASCII 키보드) 를 이용하여 질의를 타이핑하게 하고, 그 질의를 서버에서 적절한 형태로 전환되게 하며 (예를 들어, 로마지로 기입된 질의를 가타가나, 히라가나, 및/또는 간지로 전환시킴), 변환된 형태에 기초하여 서치 결과를 수신하게 한다.

[0017] 본 발명은, 프로그램 명령이 광학 또는 전자 통신 회선을 통해 전송되는 컴퓨터 네트워크, 캐리어파, 또는 컴퓨터 판독가능 저장 매체와 같은 컴퓨터 판독가능 매체, 프로세스, 장치, 시스템, 디바이스, 또는 방법을 포함하는 다양한 방식으로 구현될 수 있음을 인식하여야 한다. 수개의 창의적인 실시형태들이 아래에서 설명된다.

[0018] 일 실시형태에서는, 하나의 언어 및/또는 문자 세트로부터 다른 것으로 질의 용어를 자동으로 전환하는 방법이 개시된다. 앵커 (anchor) 텍스트가 포인팅하는 문헌의 세트 (예를 들어, 웹 페이지) 인 것으로서, 소정의 질의 용어를 포함하는 제 1 세트의 앵커 텍스트가 식별된다. 그 후, 제 2 포맷으로 기입되고 동일한 문헌의 세트를 포인팅하는 제 2 세트의 앵커 텍스트가 식별된다. 그 후, 제 2 세트의 앵커 텍스트가 분석되어, 제

1 포맷으로의 소정 질의 용어의 표현이 제 2 포맷으로의 소정 질의 포맷의 표현에 대응하는 확률이 획득된다.

[0019] 다른 실시형태에서는, 제 1 포맷 (예를 들어, 일 언어 및/또는 문자 세트) 으로 기입된 용어를 제 2 포맷 (예를 들어, 또 다른 언어 및/또는 문자 세트) 에 매핑하는 확률 딕셔너리 (probabilistic dictionary) 가 생성된다. 확률 딕셔너리는 제 1 포맷으로 기입된 질의를 제 2 포맷으로 전환하는데 이용된다. 그 후, 전환된 질의는 서치를 수행하는데 이용되며, 그 결과가 사용자에게 리턴된다. 일부 실시형태에서, 사용자의 서치 결과와의 상호작용이 모니터링될 수 있으며, 확률 딕셔너리 내의 확률을 업데이트하는데 이용된다. 또한, 일부 실시형태에서는, 서치 이전에, 질의 자체가 다른 언어 및/또는 문자 세트 매핑을 포함하도록 확장될 수 있다.

[0020] 또 다른 실시형태에서는, 확률 딕셔너리를 생성하는 방법이 설명된다. 확률 딕셔너리는 제 1 포맷으로의 용어를 제 2 포맷으로 전환하는데 사용될 수 있다. 그 딕셔너리는, 그 용어를 포함하는 앵커 텍스트 또는 다른 데이터를 식별함으로써, 용어별로 생성되는 것이 바람직하다. 다음으로, 앵커 텍스트 또는 다른 데이터와 정렬되는 데이터가 분석되어, 제 1 포맷으로의 소정의 용어가 제 2 포맷으로의 하나 이상의 용어에 매핑하는 확률이 결정된다.

[0021] 또 다른 실시형태에서는, 질의 용어 중 하나 이상을 포함하고 제 1 언어 또는 문자 세트로 기입된 앵커 텍스트를, 그 제 1 앵커 텍스트에 대응하고 제 2 언어 또는 문자 세트로 기입된 앵커 텍스트와 비교함으로써, 제 1 언어 또는 문자 세트로 제공되는 질의가 제 2 언어 또는 문자 세트로 전환된다.

[0022] 또 다른 실시형태에서는, 제 1 포맷으로 기입된 용어를 제 2 포맷으로 전환시키는 컴퓨터 프로그램이 제공된다. 컴퓨터 프로그램 제품은 컴퓨터 시스템으로 하여금 정렬된 앵커 텍스트를 식별하게 하고, 제 1 포맷으로의 소정 용어의 표현이 제 2 포맷으로의 하나 이상의 용어에 대응하는 확률을 결정하게 하도록 동작가능하다.

[0023] 또 다른 실시형태에서는, 잠재적으로 모호한 질의를 이용하여 서치를 수행하는 방법이 제공된다. 사용자가 제 1 포맷으로 질의를 입력할 경우, 그것은 제 2 포맷으로 기입된 하나 이상의 변형체 (variant) 의 그룹으로 전환된다. 그 후, 전환된 변형체를 이용하여 서치가 수행되며, 응답 정보가 사용자에게 리턴된다. 예를 들어, 제 1 포맷은 전화 키패드를 사용하여 입력된 번호의 시퀀스를 포함할 수도 있으며, 제 2 포맷은 문자숫자 텍스트 (예를 들어, 영어, 로마지, 로마자 (romaja), 핀인 (pinyin; 병음) 등) 를 포함할 수도 있다. 일부 실시형태에서, 하나 이상의 변형체의 그룹은, 미리 정의된 어휘집 (lexicon) 에 나타나지 않고/않거나 미리 정의된 낮은 확률의 문자 조합을 포함하는 전환 변형체를 폐기함으로써 선택된다. 또한, 일부 실시형태에서, 확률 딕셔너리는, 서치가 수행되기 전에, 하나 이상의 변형체의 그룹을 제 3 포맷으로 전환하는데 사용된다. 예를 들어, 확률 딕셔너리는 로마지, 로마자, 또는 핀인으로부터 간지, 가타가나, 히라가나, 한글, 한자 또는 전통적인 중국어 문자로 하나 이상의 변형체의 그룹을 전환하는데 사용될 수 있으며, 그 후, 전환된 변형체를 이용하여 서치가 수행될 수 있다.

[0024] 본 발명의 이들 특성 및 이점, 그리고 다른 특성 및 이점은, 본 발명의 원리를 예로서 설명하는 다음의 상세한 설명 및 첨부 도면에서 더 상세히 제공될 것이다.

[0025] **도면의 간단한 설명**

[0026] 본 명세서에 포함되며 그 일부를 구성하는 첨부 도면은 본 발명의 실시형태들을 나타내며, 상세한 설명과 함께, 본 발명의 이점 및 원리를 설명하도록 제공된다.

[0027] 도 1 은, 본 발명과 부합하는 방법 및 장치가 구현될 수도 있는 시스템의 블록도를 나타낸 것이다.

[0028] 도 2 는 본 발명과 부합하는 클라이언트 디바이스의 블록도를 나타낸 것이다.

[0029] 도 3 은 3 개의 문헌을 도시한 도면을 나타낸 것이다.

[0030] 도 4a 는 종래의 문자숫자 인덱스를 나타낸 것이다.

[0031] 도 4b 는 종래의 문자숫자 서치 질의에 응답하여 서치 결과를 제공하는 흐름도를 나타낸 것이다.

[0032] 도 5a 는 모호한 서치 질의에 응답하여 서치 결과를 제공하기 위한, 본 발명과 부합하는 흐름도를 나타낸 것이다.

[0033] 도 5b 는 문자숫자 정보를 숫자 정보에 매핑하는 도면을 나타낸 것이다.

[0034] 도 6 은 모호한 서치 질의에 응답하여 서치 결과를 제공하기 위한, 본 발명과 부합하는 또 다른 흐름도를 나타낸 것이다.

[0035] 도 7 은 본 발명의 실시형태들에 따라 서치를 수행하는 방법을 나타낸 것이다.

[0036] 도 8 은 문자-세트 전환의 확률 딕셔너리를 나타낸 것이다.

[0037] 도 9 는 확률 딕셔너리를 형성하기 위한 병렬 앵커 텍스트의 이용을 나타낸 것이다.

[0038] 도 10 은 또 다른 텍스트를 이용하여 링크된 문헌의 수집을 나타낸 것이다.

[0039] 도 11a 및 도 11b 는, 도 10 에 도시된 앵커 텍스트에 기초하여 가능성있는 전환의 계산을 나타낸 것이다.

[0040] 도 12 는 예시적인 워드 전환과 관련된 확률 분포를 나타낸 것이다.

[0041] **특정 실시형태의 설명**

[0042] 다음으로, 첨부도면에서 설명된 바와 같은 본 발명의 실시형태들을 더 상세히 참조한다. 동일하거나 유사한 부분을 참조하기 위해 도면 및 상세한 설명 전반에 걸쳐 동일한 참조부호가 사용될 수도 있다. 다음의 설명은 당업자로 하여금 작업의 창의적인 실체를 제조 및 이용하게 하도록 제공된다. 특정 실시형태 및 애플리케이션의 설명은 오직 예로써 제공되며, 다양한 변형은 당업자에게 명백하다. 예를 들어, 비록 다수의 예가 인터넷 웹 페이지의 맥락에서 설명되지만, 본 발명의 실시형태들은 서적, 신문, 잡지 등과 같이 다른 타입의 문헌 및/또는 정보를 서치하는데 사용될 수 있음을 이해하여야 한다. 유사하게, 비록 설명을 위하여, 다수의 예들이 로마지로부터 히라가나, 가타가나, 및/또는 간지로의 일본어 텍스트의 전환을 설명하지만, 당업자는 본 발명의 시스템 및 방법이 임의의 적절한 전환에 적용될 수 있음을 알 수 있다. 예를 들어, 제한없이, 본 발명의 실시형태는, 일부 다른 포맷 (예를 들어, 핀인 또는 로마자) 으로 수신된 질의에 기초하여, 예를 들어, 전통적인 중국어 문자 또는 한국의 한글 또는 한자 문자로 기입된 텍스트를 서치하는데 이용될 수 있다. 여기에서 설명된 일반적인 원리는 본 발명의 사상 및 범위를 벗어나지 않고 다른 실시형태 및 애플리케이션에 적용될 수도 있다. 따라서, 본 발명은 여기에서 개시된 원리 및 특징과 부합하는 다수의 변경예, 변형예, 및 균등물을 포함하여 최광의 범위를 부여하려는 것이다. 명료화를 위해, 본 발명과 관련된 기술분야에서 공지된 기술적 자료에 관한 세부설명은 본 발명을 불필요하게 불명료하게 하지 않도록 상세히 설명하지 않는다.

[0043] **A. 개관**

[0044] 본 발명과 부합하는 방법 및 장치는 사용자로 하여금 모호한 서치 질의를 제출하게 하고, 잠재적으로 명료하게 된 서치 결과를 수신하게 한다. 일 실시형태에서, 표준 전화 키패드의 사용자로부터 수신된 번호의 시퀀스는 잠재적으로 대응하는 문자숫자 시퀀스의 세트로 전환된다. 이들 잠재적으로 대응하는 문자숫자 시퀀스는, 부울 (boolean) "OR" 표현을 사용하여, 종래의 서치 엔진에 입력으로서 제공된다. 이러한 방식으로, 서치 엔진은, 서치 결과를 사용자가 흥미있어 했을 가능성이 있는 것들로 제한하게 하는데 사용된다.

[0045] **B. 구조**

[0046] 도 1 은, 본 발명과 부합하는 방법 및 장치가 구현될 수도 있는 시스템 (100) 을 도시한 것이다. 시스템 (100) 은 네트워크 (140) 를 통하여 다중의 서버 (120 및 130) 에 접속된 다중의 클라이언트 디바이스 (110) 를 포함할 수도 있다. 네트워크 (140) 는 LAN (local area network), WAN (wide area network), PSTN (Public Switched Telephone Network) 과 같은 전화 네트워크, 인트라넷, 인터넷 또는 네트워크들의 조합을 포함할 수도 있다. 간략화를 위하여, 2 개의 클라이언트 디바이스 (110) 및 3 개의 서버 (120 및 130) 가 네트워크 (140) 에 접속된 것으로서 도시되어 있다. 실제로, 더 많거나 더 적은 클라이언트 디바이스 및 서버가 존재할 수도 있다. 또한, 어떤 경우, 클라이언트 디바이스는 서버의 기능을 수행할 수도 있고, 서버는 클라이언트 디바이스의 기능을 수행할 수도 있다.

[0047] 클라이언트 디바이스 (110) 는, 네트워크 (140) 에 접속할 수 있는 메인프레임 (mainframes), 미니컴퓨터, 퍼스널 컴퓨터, 랩탑, 개인휴대 정보 단말기 (PDA) 등과 같은 디바이스를 포함할 수도 있다. 클라이언트 디바이스 (110) 는 네트워크 (140) 를 통하여 데이터를 송신하거나, 유선, 무선, 또는 광학 접속을 통하여 네트워크 (140) 로부터 데이터를 수신할 수도 있다.

[0048] 도 2 는 본 발명과 부합하는 예시적인 클라이언트 디바이스 (110) 를 도시한 것이다. 클라이언트 디바이스 (110) 는 버스 (210), 프로세서 (220), 메인 메모리 (230), ROM (read only memory; 240), 저장 디바이스 (250), 입력 디바이스 (260), 출력 디바이스 (270), 및 통신 인터페이스 (280) 를 포함할 수도 있다.

[0049] 버스 (210) 는, 클라이언트 디바이스 (110) 의 컴포넌트 사이에서 통신을 허용하는 하나 이상의 종래의 버스를 포함할 수도 있다. 프로세서 (220) 는, 명령을 해석 및 실행하는 임의의 타입의 종래의 프로세서 또는 마이

크로프로세서를 포함할 수도 있다. 메인 메모리 (230) 는, 프로세서 (220) 에 의한 실행을 위해 정보 및 명령을 저장하는 RAM (random access memory) 또는 다른 타입의 동적 저장 디바이스를 포함할 수도 있다. ROM (240) 은, 프로세서 (220) 에 의한 사용을 위해 정적 정보 및 명령을 저장하는 종래의 ROM 디바이스 또는 다른 타입의 정적 저장 디바이스를 포함할 수도 있다. 저장 디바이스 (250) 는 자기 및/또는 광학 기록 매체 및 그 대응하는 디바이스를 포함할 수도 있다.

[0050] 입력 디바이스 (260) 는, 키보드, 마우스, 펜, 음성 인식 및/또는 바이오메트릭 (biometric) 메커니즘 등과 같이, 사용자가 클라이언트 디바이스 (110) 에 정보를 입력하는 것을 허용하는 하나 이상의 종래의 메커니즘을 포함할 수도 있다. 출력 디바이스 (270) 는 디스플레이, 프린터, 스피커 등을 포함하는, 사용자에게 정보를 출력하는 하나 이상의 종래의 메커니즘을 포함할 수도 있다. 통신 인터페이스 (280) 는, 클라이언트 디바이스 (110) 가 다른 디바이스 및/또는 시스템과 통신하도록 하는 임의의 트랜시버-형 메커니즘을 포함할 수도 있다. 예를 들어, 통신 인터페이스 (280) 는, 네트워크 (140) 와 같은, 네트워크를 통하여 다른 디바이스 또는 시스템과 통신하는 메커니즘을 포함할 수도 있다.

[0051] 아래에서 상세히 설명되는 바와 같이, 본 발명과 부합하는 클라이언트 디바이스 (110) 는 소정의 서칭-관련 동작을 수행한다. 클라이언트 디바이스 (110) 는, 메모리 (230) 와 같은 컴퓨터-판독가능 매체에 포함된 소프트웨어 명령을 실행하는 프로세서 (220) 에 응답하여 이들 동작을 수행할 수도 있다. 컴퓨터-판독가능 매체는 하나 이상의 메모리 디바이스 및/또는 캐리어파로서 정의될 수도 있다. 소프트웨어 명령은 데이터 저장 디바이스 (250) 와 같은 다른 컴퓨터-판독가능 매체로부터, 또는 통신 인터페이스 (280) 를 통하여 다른 디바이스로부터 메모리 (230) 내로 판독될 수도 있다. 메모리 (230) 에 포함된 소프트웨어 명령은 프로세서 (220) 로 하여금 하술되는 서치-관련 활동을 수행하게 한다. 다른 방법으로, 본 발명과 부합하는 프로세스를 구현하기 위하여, 하드와이어드 (hardwired) 회로가 소프트웨어 명령 대신에 또는 소프트웨어 명령과 조합하여 사용될 수도 있다. 따라서, 본 발명은 하드웨어 회로 및 소프트웨어의 임의의 특정 조합으로 제한되지 않는다.

[0052] 서버 (120 및 130) 는, 그 서버 (120 및 130) 가 클라이언트 디바이스 (110) 와 통신하게 할 수 있도록 네트워크 (140) 에 접속할 수 있는, 메인프레임, 미니컴퓨터, 또는 퍼스널 컴퓨터와 같은 하나 이상의 타입의 컴퓨터 시스템을 포함할 수도 있다. 또 다른 구현예에서, 서버 (120 및 130) 는 하나 이상의 클라이언트 디바이스 (110) 에 직접 접속하는 메커니즘을 포함할 수도 있다. 서버 (120 및 130) 는 네트워크 (140) 를 통하여 데이터를 송신하거나, 유선, 무선, 또는 광학 접속을 통하여 네트워크 (140) 로부터 데이터를 수신할 수도 있다.

[0053] 서버는, 클라이언트 디바이스 (110) 에 대해 도 2 를 참조하여 상술된 바와 유사한 방식으로 구성될 수도 있다. 본 발명과 부합하는 구현예에서, 서버 (120) 는 클라이언트 디바이스 (110) 에 의해 이용가능한 서치 엔진 (125) 을 포함할 수도 있다. 서버 (130) 는 클라이언트 디바이스 (110) 에 의해 액세스 가능한 문헌 (또는 웹 페이지) 을 저장할 수도 있다.

[0054] **C. 구조적 동작**

[0055] 도 3 은, 예를 들어, 서버 (130) 중 하나에 저장될 수도 있는 3 개의 문헌을 도시한 도면을 나타낸 것이다.

[0056] 제 1 문헌 (문헌 1) 은 2 개의 엔트리 (entry), 즉, "car repair" 및 "car rental" 을 포함하고, 그 저부에 "3" 으로 번호가 매겨져 있다. 제 2 문헌 (문헌 2) 은 엔트리 "video rental" 을 포함한다. 제 3 문헌 (문헌 3) 은 3 개의 엔트리, 즉, "wine", "champagne", 및 "bar items" 를 포함하고, 문헌 2 로의 링크 (또는 참조) 를 포함한다.

[0057] 설명의 간략화를 위하여, 도 3 에 도시된 문헌은 오직 정보의 문자숫자 스트링 (예를 들어, "car", "repair", "wine" 등) 만을 포함한다. 그러나, 당업자는, 다른 상황에서, 그 문헌들은 표음 정보 또는 시청각적 정보와 같은 다른 타입의 정보를 포함할 수도 있음을 알 수 있다.

[0058] 도 4a 는 도 3 에 도시된 문헌에 기초하여, 종래의 문자숫자 인덱스를 나타낸 것이다. 그 인덱스의 제 1 컬럼 (column) 은 문자숫자 용어의 리스트를 포함하며, 제 2 컬럼은 이들 용어에 대응하는 문헌의 리스트를 포함한다. 문자숫자 용어 "3" 과 같은 일부 용어는 오직 하나의 문헌 (이 경우, 문헌 1) 에 대응한다 (예를 들어, 하나의 문헌에 나타난다). "rental" 과 같은 다른 용어는 다중의 문헌 (이 경우, 문헌 1 및 2) 에 대응한다.

[0059] 도 4b 는, 서치 엔진 (125) 와 같은 종래의 서치 엔진이 도 4a 에 나타낸 인덱스를 어떻게 이용하여 문자숫자 서치 질의에 응답하여 서치 결과를 제공하는지를 나타낸 것이다. 문자숫자 질의는 임의의 종래의 기술을 이용하여 생성될 수도 있다. 예시를 위하여, 도 4b 는 2 개의 문자숫자 질의, 즉, "car" 및 "wine" 을 나타낸

다. 종래의 접근법에 의하면, 서치 엔진 (125) 은 "car" 와 같은 문자숫자 질의를 수신하고 (단계 410), 문자숫자 인덱스를 이용하여, 어떤 문헌이 그 질의에 대응하는지를 결정한다 (단계 420). 이 예에서, 종래의 서치 엔진 (125) 은 도 4a 에 도시된 인덱스를 이용하여, "car" 가 문헌 1 에 대응한다고 결정하고, 서치 결과로서 사용자에게 문헌 1 (또는 그에 대한 참조) 을 리턴한다. 유사하게, 종래의 서치 엔진은 "wine" 이 문헌 3 에 대응한다고 결정하고, 사용자에게 문헌 3 (또는 그에 대한 참조) 을 리턴한다 (단계 430).

[0060] 도 5a 는, 도 3 및 도 4a 에 각각 도시된 문헌 및 인덱스에 기초하여, 숫자 서치 질의에 응답하여 서치 결과를 제공하는 바람직한 기술의, 본 발명과 부합하는 흐름도를 나타낸 것이다. 설명의 용이를 위하여, 도 5a 는 표준 전화 핸드셋의 매핑에 기초하여 숫자 질의를 프로세싱하는 특정 기술을 나타낸 것이지만, 당업자는, 본 발명과 부합하는 다른 기술이 사용될 수도 있음을 알 수 있다.

[0061] 단계 510 에서, 시퀀스 "227" (숫자 컴포넌트 "2", "2", 및 "7" 로 이루어짐) 이 사용자로부터 수신된다. 단계 520 에서, 숫자 컴포넌트가 어떻게 문자에 매핑하는지에 관한 정보가 획득된다. 사용자가 표준 전화 키패드로부터 정보를 입력하였다고 가정하면, 이러한 매핑 정보는 도 5b 에 도시되어 있다. 도 5b 에 도시된 바와 같이, 문자 "a", "b", 및 "c" 는 각각 번호 "2" 에 매핑하고, 문자 "p", "q", "r", 및 "s" 는 각각 번호 "7" 에 매핑하는 등의 식이다.

[0062] 단계 530 에서, 이러한 매핑 정보를 이용하여, 시퀀스 "227" 은 그 잠재적인 문자숫자 균등물로 전환된다. 도 5b 에 도시된 정보에 기초하여, 다음의 aap, bap, cap, abp, bbp, ... bar ... car ... ccs 를 포함하여, 시퀀스 "227" 에 대응하는 문자의 36 개의 가능한 조합이 존재한다. 만약 번호가 그 가능한 조합에 포함되어 있으면 (예를 들어, "aa7"), 80 개의 가능한 조합이 존재한다. 모든 가능한 문자숫자 균등물을 생성하는 것보다는, 어떠한 어휘집에 기초하여, 생성된 균등물을 제한하는 것이 바람직할 수도 있다. 예를 들어, 딕셔너리, 이전 서치 질의의 서치 엔진 로그 등에 나타난 이들 문자숫자 균등물만을 생성하거나, 그렇지 않으면, 기지(旣知)의 통계 기술 (예를 들어, 함께 나타나는 소정 워드의 확률) 을 이용함으로써 문자숫자 균등물을 제한하는 것이 바람직할 수도 있다.

[0063] 단계 540 에서, 이들 문자숫자 균등물은, 논리 "OR" 연산을 이용하여, 도 4a 및 도 4b 를 참조하여 설명된 바와 같이, 종래의 서치 엔진에 입력으로서 제공된다. 예를 들어, 서치 엔진에 제공되는 서치 질의는 "aap OR bap OR cap OR abp ... OR bar ... OR car" 일 수 있다. 비록 모든 가능한 문자숫자 균등물이 서치 엔진에 제공될 수도 있지만, 의도될 가능성이 없는 균등물을 제거하기 위하여 종래의 기술을 사용함으로써 서브세트가 대신 사용될 수도 있다. 예를 들어, 문자 또는 워드의 사용에 관한 확률 정보를 이용하는 기술을 이용함으로써 가능한 조합의 더 협소한 리스트를 생성할 수 있으며, "qt" 로 시작하는 조합을 무시하지만, "qu" 로 시작하는 조합을 포함 (및 선호) 할 수 있다.

[0064] 단계 550 에서, 서치 결과가 서치 엔진으로부터 획득된다. "aap" 및 "abp" 와 같은 용어가 서치 엔진의 인덱스에 나타나지 않기 때문에, 그들은 사실상 무시된다. 실제로, 오직 도 4a 에 도시된 인덱스 내에 포함된 용어는 "car" 및 "bar" 이며, 따라서, 리턴되는 유일한 서치 결과는 문헌 1 및 3 을 참조하는 것들이다. 단계 560 에서, 이들 서치 결과가 사용자에게 제공된다. 서치 결과는 서치 엔진에 의해 제공된 동일한 순서로 제공될 수도 있거나, 사용자의 언어와 같은 고려사항에 기초하여 재정렬될 수도 있다. 용어 "car" 를 포함하는 문헌에만 사용자가 관심이 있었다고 가정하면, 사용자는 원하는 결과 (문헌 1) 에 부가하여 원하지 않은 결과 (문헌 3) 을 수신한다. 그러나, 이것은, 서치 질의를 공식화 (formulate) 하기 위해 오직 3 개의 키를 눌러야 하는 사용자의 이익을 위해서 지불할 허용가능한 댓가일 수도 있다.

[0065] 도 6 은, 도 3 및 도 4a 에 각각 도시된 문헌 및 인덱스에 기초하여, 숫자 서치 질의에 응답하여 서치 결과를 제공하는 바람직한 기술의, 본 발명과 부합하는 또 다른 흐름도를 나타낸 것이다. 이 흐름도는, 수신 시퀀스의 사이즈의 증가가 사용자에 의해 요구된 결과로 서치 결과를 어떻게 제한하게 할 수 있는지를 설명한다. 설명의 용이를 위하여, 도 6 또한 표준 전화 핸드셋의 매핑에 기초하여 숫자 질의를 프로세싱하는 특정 기술을 설명하지만, 당업자는, 본 발명과 부합하는 다른 기술이 사용될 수도 있음을 알 수 있다.

[0066] 단계 610 에서, 시퀀스 "227 48367" (숫자 컴포넌트 "2", "2", "7", "4", "8", "3", "6", "7" 로 이루어짐) 이 사용자로부터 수신된다. 설명을 위하여, 시퀀스 "227" 은 "번호 워드" 로 지칭하고, 전체 시퀀스 "227 48367" 는 "번호 어구 (number phrase)" 로 지칭한다. 번호 워드의 가능한 문자숫자 균등물은 "문자 워드" 로 지칭하고, 번호 어구의 가능한 문자숫자 균등물은 "문자 어구" 로 지칭한다.

[0067] 단계 620 에서, 숫자 컴포넌트가 어떻게 문자에 매핑하는지에 관한 정보가 획득된다. 도 5b 에 도시된 바와

동일한 매핑 정보가 사용된다고 가정하면, 단계 630 에서, 번호 어구 "227 48367" 은 잠재적으로 대응하는 문자 어구로 전환된다. 도 5b 에 도시된 정보에 기초하여, 시퀀스 "227 48367" 에 대응하는 11664 개의 가능한 문자 어구가 존재한다.

[0068] 단계 640 에서, 이들 문자 어구는, 논리 "OR" 연산을 이용하여, 도 4a 및 도 4b 를 참조하여 설명된 바와 같이, 종래의 서치 엔진에 입력으로서 제공된다. 예를 들어, 서치 엔진에 제공되는 서치 질의는 "'aap gtdmp' OR 'aap htdmp' ... OR 'bar items' ... OR 'car items'" 일 수 있다. 비록 모든 가능한 문자 어구가 서치 엔진에 제공될 수도 있지만, 의도될 가능성이 없는 문자 어구를 제거하기 위하여 종래의 기술을 채용함으로써 서브세트가 대신 사용될 수도 있다.

[0069] 단계 650 에서, 서치 결과가 서치 엔진으로부터 획득된다. 다수의 서치 엔진이, 검색된 정확한 어구를 포함하는 이들 문헌을 높게 랭크 (rank) 하도록 설계되기 때문에, 문헌 3 은 최고로 랭크된 서치 결과일 가능성이 있다 (즉, 정확한 어구 "bar items" 을 포함하기 때문에). 본 실시예에서 어떠한 다른 문헌도 단계 620 에서 생성된 다른 문자 어구 중 하나를 포함하지 않는다. 또한, 다수의 서치 엔진이, 전체 어구가 아니라 개별적인 어구 부분을 포함하는 서치 결과의 중요도를 떨어뜨린다 (downweight) (또는 제거한다). 예를 들어, 문헌 1 은, 문자 어구의 제 1 부분에 대응하는 문자 워드 "car" 를 포함하지만 그 문자 어구의 제 2 부분에 대응하는 어떠한 문자 워드도 포함하지 않기 때문에, 중요도가 떨어지거나 제거된다. 마지막으로, "aap htdmp" 와 같은 문자 어구는, 서치 엔진의 인덱스에 나타나는 어떠한 문자 워드도 포함하지 않기 때문에 사실상 무시된다.

[0070] 단계 660 에서, 서치 결과가 사용자에게 제공된다. 도시된 예에서, 사용자에게 나타낸 제 1 결과는, 사용자의 질의에 가장 적절할 가능성이 있는 문헌 3 이다. 문헌 1 은, 가능한 문자 어구 중 어떠한 것도 포함하지 않기 때문에 함께 제거될 수도 있다. 이러한 방식으로, 사용자는 가장 적절한 서치 결과를 제공받는다.

[0071] 도 5 및 도 6 을 참조한 상기 설명이 숫자 정보를 수신하는 단계 및 그 숫자 정보를 문자숫자 정보에 매핑하는 단계와 관련하여 수행되었지만, 당업자는, 다른 구현이 본 발명과 부합하여 가능함을 알 수 있다. 예를 들어, 사용자에 의해 눌러진 키에 대응하는 번호의 시퀀스를 수신하는 대신, 수신 시퀀스는, 사용자에 의해 눌러진 키에 대응하는 제 1 문자로 이루어질 수도 있다. 즉, "227" 을 수신하는 대신, 수신 시퀀스는 "aap" 일 수도 있다. 또한, 본 발명과 부합하여, 단계 530 또는 단계 630 에서 생성된 균등한 문자 시퀀스는, "aap" 에 대응하는 다른 문자 시퀀스 (예를 들어, "bar") 일 수 있다. 실제로, 수신 시퀀스는 표음적, 시청각적, 또는 다른 타입의 정보 컴포넌트를 포함할 수도 있다.

[0072] 시퀀스가 수신되는 형태에 무관하게, 일반적으로, 정보가 서치 엔진의 인덱스에 저장되는 포맷에 대응하는 시퀀스로 수신 시퀀스가 전환되는 것이 바람직하다. 예를 들어, 만약 서치 엔진의 인덱스가 문자숫자 포맷으로 저장되면, 수신 시퀀스는 문자숫자 시퀀스로 전환되어야 한다.

[0073] 또한, 정보 컴포넌트의 수신 시퀀스를 전환하는데 이용되는 매핑 기술은, 사용자의 디바이스에 의해 생성되는 정보에 사용자의 입력을 매핑하기 위해 사용자의 디바이스에서 채용되는 동일한 기술임이 바람직하다. 그러나, 사용자 입력용으로 사용되는 것과 상이한 매핑 기술을 사용하는 것이 바람직한 경우가 존재할 수도 있다.

[0074] 또한, 본 발명의 실시형태는 사용자로 하여금 비-타깃-언어 키보드를 사용하여 입력되는 서치를 수행하는 것을 가능하게 할 수 있다. 예를 들어, 일본어 텍스트를 포함하는 웹 페이지는 간지로 기입될 수도 있지만, 그 페이지의 서치를 시도하는 사용자는 로마 알파벳에 기초하여 표준 ASCII 키보드 (또는 핸드셋) 에 액세스할 수도 있다.

[0075] 도 7 은 그러한 서치를 수행하는 방법을 나타낸 것이다. 도 7 에 도시된 바와 같이, 사용자는 표준 입력 디바이스 (예를 들어, ASCII 키보드, 전화 핸드셋 등) 를 사용하여 질의를 타이핑하고, 그 질의를 서치 엔진에 전송한다. 그 질의는, 응답 문헌 중 일부가 기입되는 문자 세트 (예를 들어, 간지) 와 상이한 문자 세트 (예를 들어, 로마지) 로 기입될 수도 있다. 서치 엔진은 질의를 수신하고 (블록 702), 그 질의를 적절한 형태로 전환하고 (블록 704), 예를 들어, 종래의 서치 기술을 이용하여, 그 전환된 질의에 응답하는 문헌에 대한 서치를 수행한다 (블록 706). 그 후, 서치 엔진은 응답 문헌 (및/또는 문헌 자체의 사본) 의 리스트를 사용자에게 리턴한다 (블록 708). 예를 들어, 도 6 과 관련하여 상술된 유사한 방식으로 사용자에게 결과가 리턴될 수 있다.

[0076] 도 7 에 도시된 바와 같이, 사용자의 질의는, 클라이언트에 대향하는 서치 엔진의 서버에서 전환되는 것이 바람직하며, 따라서, 전환을 수행하기 위해 특수-목적 소프트웨어를 획득하기 위한 사용자의 필요성을 완화시킨다.

그러나, 다른 실시형태에서는, 전환의 일부 또는 전부가 클라이언트에서 수행될 수 있음을 알 수 있다. 또한, 일부 실시형태에서는, 전화 키패드와 같은 디바이스를 이용하여 질의가 입력될 수도 있다. 그러한 실시형태에서, 초기의 숫자 질의는, 예를 들어, 낮은 확률의 매핑 (예를 들어, 로마지에서 발생하지 않는 문자 조합을 포함하는 매핑) 을 폐기하기 위한 확률 기술 및/또는 어휘집의 애플리케이션을 포함하는, 도 5 및 도 6 과 관련하여 상술된 매핑 기술을 이용하여, 먼저, 문자숫자 형태 (예를 들어, 로마지) 로 전환될 수도 있다. 일단 질의의 문자숫자 전환이 획득되면, 도 7 에 도시된 단계의 나머지가 수행될 수 있다 (즉, 단계 704, 706, 및 708).

[0077] 하나의 문자 세트 또는 언어로부터 다른 문자 세트 또는 언어로의 질의의 전환 (즉, 도 7 의 블록 704) 은 다양한 방식으로 수행될 수 있다. 하나의 기술은, 워드 의미 또는 전환의 종래의 정적 딕셔너리를 이용하여, 질의에서의 각각의 용어를 타깃 언어 또는 문자 세트에서의 대응하는 용어에 매핑시키는 것이다. 그러나, 이러한 접근법의 문제는, 워드가 종종 모호하고, 질의가 종종 너무 짧아서 이러한 모호성을 해결하기 위한 적절한 문맥상의 단서를 제공하지 못하기 때문에, 종종, 부정확한 결과를 야기한다는 것이다. 예를 들어, 워드 "bank" 는 강의 제방, 금융 기관, 또는 비행기에 의한 기동을 지칭할 수 있기 때문에, 추상적으로 정확하게 전환하기 어렵다. 또한, 만약 딕셔너리가 비교적 크고/크거나 자주 업데이트되지 않으면, 흔히 사용되지 않는 워드, 속어, 관용어, 적절한 명칭 등과 같이, 서치 엔진이 조우할 수도 있는 모든 용어에 대한 엔트리를 포함하지 않을 수도 있다.

[0078] 본 발명의 실시형태들은, 질의 용어를 하나의 언어 또는 문자 세트 (예를 들어, ASCII) 로부터 다른 언어 또는 문자 세트 (예를 들어, 간지) 로 전환하기 위해 확률 딕셔너리를 이용함으로써 이들 문제의 전부 또는 그 일부를 극복하거나 개선시키는데 이용될 수 있다. 바람직한 실시형태에서, 확률 딕셔너리는 일 세트의 용어를 다른 세트의 용어로 매핑하고, 각각의 매핑과 확률을 관련시킨다. 편리를 위해, "용어 (term)" 또는 "토큰 (token)" 은, 워드, 어구, 및/또는 (더 일반적으로는) 공백을 포함할 수도 있는 하나 이상의 문자의 시퀀스를 지칭한다.

[0079] 도 8 은 상술된 바와 같은 확률 딕셔너리 (800) 의 일 예를 도시한 것이다. 도 8 에 도시된 예시적인 확률 딕셔너리 (800) 는 로마지 (일본어의 로마 알파벳 표현) 로 기입된 워드를 간지 (비-로마의 표의문자-기반 일본어 문자 세트) 로 기입된 워드에 매핑한다. 설명을 용이하게 하기 위해, 도 8 은 "<용어>$_{romaji}$" 로서의 로마지 용어, 및 "<용어>$_{kanji}$" 로서의 간지 용어를 도시한 것이다. 실제 로마지-간지 딕셔너리에서, 도 8 에 도시된 영어 전환 보다는 실제 로마지 및 간지 용어가 사용된다. 따라서, 도 8 은 본 발명의 실시형태에 대한 설명을 용이하게 하도록 제공되며, 일본어 텍스트의 실제 특성 및 의미를 나타내려는 것은 아님을 알 수 있을 것이다.

[0080] 딕셔너리 (800) 는 다양한 로마지 용어 (802) 에 대한 엔트리 (808, 810, 812, 814) 를 포함한다. 또한, 딕셔너리는 간지 (804) 에서의 이들 용어 각각의 잠재적인 표현과 함께, 각각의 그러한 표현이 정확하다는 대응 확률 (806) 을 포함한다. 예를 들어, 로마지 용어 "bank" 는, 0.3 의 확률로 "가파른 경사 (steep slope)" 를 의미하는 간지 용어로 매핑되고, 0.4 의 확률로 "금융 기관 (financial institution)" 을 의미하는 용어로 매핑되고, 0.2 의 확률로 "비행기 기동 (airplane maneuver)" 을 의미하는 용어로 매핑된다. 0.1 의 확률로, 그 용어는, 각각의 용어가 딕셔너리에 있지 않을 수도 있는 용어에 매핑되게 하는 간단한 일반 방법인 "기타 (other)" 에 매핑될 수도 있다.

[0081] 또한, 도 8 에 도시된 실시예는, 제 1 문자 세트 또는 언어로의 소정의 용어 (예를 들어, 워드 "bank") 가 다른 문자 세트 또는 언어로의 2 개 이상의 용어에 매핑될 수도 있음을 설명하기 위해 구성되었음을 알 수 있다. 그러나, 당업자는, 명료화를 위해, 도 8 의 특정 실시예가 영어 워드 및 의미를 이용하여 이러한 원리를 설명하지만, 예를 들어, 워드 "bank" 의 실제 로마지 표현은 그 영어 균등물과 동일한 방식으로 모호하지 않을 수도 있음을 알 수 있다 (예를 들어, 금융 기관에 대한 워드와 비행기 기동에 대한 워드 간의 로마지에서의 모호성이 존재하지 않을 수도 있음). 또한, 설명을 용이하게 하기 위해, 도 8 에 도시된 딕셔너리는 다른 점에서도 간략화될 수 있음을 알 수 있다. 예를 들어, 실제 확률 딕셔너리는 각각의 용어에 대한 다수의 더 잠재적인 매핑을 포함할 수도 있거나, 미리 정의된 확률 임계값을 초과하는 매핑만을 포함할 수도 있다.

[0082] 본 발명의 바람직한 실시형태는 그러한 확률 딕셔너리를 이용하여, 하나의 언어 및/또는 문자 세트로 표현된 질의를 다른 언어 또는 문자 세트로 전환시킴으로써, 사용자로 하여금 그 오리지널 질의와 상이한 문자 세트 및/또는 언어로 기입된 문헌을 찾게 한다. 예를 들어, 만약 사용자가 "cars" 에 대한 질의를 로마지로 입력하면, 확률 딕셔너리는 "cars" 에 대한 로마지 용어를, 예를 들어, "cars" 에 대한 간지 용어에 매핑하는데 이용

될 수 있다. 이러한 방식으로, 비록 질의의 문자 세트 (예를 들어, 로마지) 와 매칭 문헌의 문자 세트 (예를 들어, 간지) 가 동일하지 않더라도, 사용자는 그 질의에 관한 문헌을 찾을 수 있다. 이 특정 실시예에서, 질의의 실제 언어는 변경되지 않고 (로마지와 간지 모두가 일본어를 표현하는데 사용됨), 오직 문자 인코딩만이 변경된다.

[0083] 또 다른 예로서, ASCII 영어로 "tired" 라는 용어는, 문자 움라우트-유 (umlaut-u) 가 ASCII 에 존재하지 않기 때문에, 라틴 1 문자 인코딩을 사용하여 독일어로 용어 "müde" 에 매핑될 수 있다. 이 예에서, 딕셔너리는 다른 언어로의 전환 (영어-독일어) 및 다른 문자 인코딩으로의 전환 (ASCII-라틴 1) 모두를 제공한다.

[0084] 바람직한 실시형태에서, 상술된 매핑 딕셔너리는, 통계 기술과 함께 웹 상에서 이용가능한 정보를 이용하여, 자동으로 구축된다. 바람직한 실시형태는, 정확한 전환에 도달하기 위하여, 상이한 언어 및/또는 문자 세트로 기입된 앵커 텍스트와 같은 병렬 정렬된 양국어 코퍼스 (bilingual corpus) 를 이용한다. 이러한 데이터를 사용하여, 바람직한 실시형태는 잠재적인 워드 매핑의 딕셔너리를 구성할 수 있다. 이것은, 예를 들어, 언어 S_i (소스 언어) 로 된 토큰이 정렬된 텍스트 쌍 (예를 들어, 앵커, 문장 등) 에 있어서의 토큰 T_j (타깃 언어) 와 동시에 발생한 횟수를 간단히 카운트함으로써 달성될 수 있다. 그러나, 임의의 적절한 기술이 사용될 수 있음을 알 수 있을 것이다.

[0085] 충분히 많고 정확히 정렬된 데이터 세트가 부재하는 경우, 이러한 방법은 비교적 모호한 다대다 (many-to-many) 매핑을 생성할 수도 있다. 따라서, 예를 들어, 오직 S_1 이 T_2, T_3, T_7 및 T_8 에 어떠한 확률로 매핑될 수 있다고만 결정될 수도 있다. 그러나, 이것은 허용가능하며, 아래에서 더 상세히 설명되는 바와 같이, 일부 실시형태에서는, 예를 들어, 이전의 사용자 질의, 결과 페이지에 대한 아이템의 사용자 선택 등을 검사함으로써, 각각의 매핑의 각각의 가능성 (likelihood) 을 증가시키기 위해 추가적인 정제 (refinement) 가 수행될 수 있다.

[0086] 도 9 는 확률 딕셔너리를 형성하기 위한 병렬 앵커의 사용을 나타낸 것이다. 앵커 텍스트는, 웹 페이지들 (또는 소정 웹 페이지 내의 위치들) 간의 하이퍼링크와 관련된 텍스트를 포함한다. 예를 들어, 하이퍼텍스트 마크업 언어 (HTML)에서, 커맨드 "<A href="http://www.abc.com">Banks and Savings and Loans</A>" 는 텍스트 "Banks and Savings and Loans" 이 http://www.abc.com에서 찾아진 웹 페이지를 포인팅하는 하이퍼링크로서 디스플레이되게 한다. 텍스트 "Banks and Savings and Loans" 은 앵커 텍스트로 지칭되며, 통상적으로, 포인팅하는 웹 페이지 (예를 들어, www.abc.com) 의 짧은 설명을 제공한다. 실제로, 앵커 텍스트는 종종 웹 페이지 자체보다는 웹 페이지의 더 정확한 설명을 제공하며, 따라서, 포인팅하는 웹 페이지의 성질을 결정하는데 있어서 특히 유용할 수 있다. 또한, 앵커 텍스트에서의 워드 사용 및 분포는 종종 사용자 질의에서 찾아진 취지 및 길이에 더 근접하다. 또한, 그것은, 소정의 페이지에 포인팅하는 다수의 앵커는 동일하거나 매우 유사한 텍스트를 포함하는 경우이다. 예를 들어, www.google.com 을 포인팅하는 앵커는 종종 "Google" 이라고 간단히 칭하거나, 다른 텍스트와 함께 적어도 이러한 용어를 사용할 것이다. 따라서, www.google.com 을 포인팅하는 모든 앵커 (예를 들어, 가타가나) 를 검사함으로써, (가능하게는, "click here" 를 단순히 칭하는 것과 같이, 소정의 미리 정의된 낮은 정보-콘텐츠 앵커를 폐기한 이후) 최고 빈도로 나타나는 용어를 간단히 찾아서 비교적 높은 신뢰도로 "Google" 에 대한 가타가나 전환이 추정될 수 있다. 본 발명의 바람직한 실시형태는 정확한 전환을 제공하기 위하여 앵커 텍스트의 이러한 특성들을 이용한다.

[0087] 도 9를 참조하면, 제 1 문자 세트 (예를 들어, ASCII) 로 기입된 용어를 포함하는 질의를 수신할 때 (블록 902), 서버는, 용어가 나타나는 앵커 텍스트의 세트를 식별한다 (블록 904). 예를 들어, 서버는 모든 기지의 앵커의 인덱스를 검사하여, 그 용어를 포함하는 그 앵커들을 식별할 수도 있다. 다음으로, 그 앵커들이 포인팅하는 웹 페이지를 식별하며 (블록 906), 이들 페이지를 포인팅하는 타깃 언어 또는 타깃 문자 세트 (예를 들어, 히라가나, 가타가나, 및/또는 간지) 로 기입된 임의의 앵커를 식별한다 (블록 908). 다음으로, 시스템은 2 개의 문헌 세트 (여기서, 앵커 텍스트는 문헌의 형태인 것으로 고려됨) 를 가진다. 그 후, 일 문헌 세트 (예를 들어, 오리지널 ASCII 질의를 포함하는 앵커) 에 있어서의 질의 용어의 분포는, 다른 문헌 세트 (예를 들어, 병렬 앵커) 에 있어서의 전환된 어구에 대한 가장 가능성 있는 후보를 식별하는데 이용된다. 앵커 텍스트 용어가 나타나는 빈도에 관한 통계가 계산될 수 있으며, 이들 통계는, 오리지널 질의의 정확한 전환인 앵커 텍스트에서 찾아진 용어의 상대적인 빈도 또는 확률을 결정하는데 이용될 수 있다 (블록 910). 다수의 워드를 갖는 질의에 대하여, 상술된 프로세스는 각각의 워드에 대하여 반복될 수 있거나, 전체 질의는 단일 용어로서 간단히 처리될 수 있으며, 또는, 워드의 기타 다른 적절한 그룹화가 사용될 수 있다. 예를 들어, 만약 질의가 "big houses" 이면, 가능한 전환의 딕셔너리는, 그 어구 (또는 어구 내 워드 중 적어도 하나) 를 포

함하는 정렬된 앵커 텍스트를 찾음으로써 구성될 수 있다. 유사하게, 만약 질의가 3 개 이상의 용어를 포함하였으면, 질의 용어의 적절한 서브세트를 선택하고 그 용어에 대한 결과를 생성함으로써, 적절한 매핑을 결정하기 위한 시험이 구성될 수 있다.

[0088] 도 9 에 도시된 방식으로 전환을 수행하는 이점은, 전환 시스템이 하나의 언어 또는 문자 세트에서의 용어와 타깃 세트에서의 용어 간의 매핑의 사전 정보를 가질 필요가 없다는 것이다. 대신, 그 매핑은, 통계적 분석을 수행하기 위해 이용가능한 데이터의 보디 (body) 에 기초하여 동적으로 결정될 수 있다. 따라서, 예를 들어, 종래의 정적 딕셔너리를 유지하는 노력 또는 비용 (예를 들어, 언어 분석 또는 연구) 을 초래하지 않고 속어 용어, 관용어, 적절한 명칭 등에 대한 정확한 전환을 발견하는 것이 가능하다.

[0089] 다음으로, 전술한 전환 기술의 예시적인 실시형태를 도 10 내지 도 12 와 관련하여 설명한다. 이 예에서, 사용자는 질의 용어 "house" 를 입력하였으며, 스페인어로 기입된 서치 결과 (또는 간단히 질의 용어의 전환) 를 획득하길 원하는 것으로 가정한다. 따라서, 서버는 영어 용어 "house" 를 그 스페인어 균등물로 전환하려 한다.

[0090] 도 10 을 참조하면, 다양한 웹 페이지 (959, 961, 963, 965) 는 앵커 텍스트 (960, 962, 964, 966) 를 통해 페이지 (972 및 974) 에 링크된다. 페이지 중 일부 및 그 관련 앵커 텍스트는 영어로 기입되고 (즉, 페이지 (959a 내지 959e 및 963a 내지 963t)), 일부는 스페인어로 기입되어 있다 (즉, 페이지 (961a 내지 961e 및 965a 내지 965j)). 서버는, 먼저, 용어 "house" 를 이용하는 모든 앵커를 위치지정한다. 이들 앵커는, 예를 들어, 서버에 저장된 앵커 텍스트의 인덱스를 서치함으로써 위치지정될 수 있다. 그러한 인덱스를 이용하여, 서버는, 어구 "big house" 를 각각 사용하고 웹 페이지 (972) 에 포인팅하는 5 개의 앵커 (960) 를 먼저 찾는다. 다음으로, 서버는, 페이지 (972) 를 포인팅하는 5 개의 타깃-언어 (즉, 스페인어) 앵커 (962) 또한 존재한다고 결정한다. 도 10 에 도시된 예에서, 이들 앵커는 텍스트 "casa grande" 를 포함한다. (앵커 (960) 및 앵커 (962) 와 같은) 동일한 페이지 또는 미리 정의된 관계를 갖는 페이지를 포인팅하는 앵커는 "정렬됨" 으로 칭하며, 여기서, 더 일반적인 의미로, 통상적으로, 정렬은 정렬된 아이템의 균등 (또는 가능성있는 균등) 을 지칭한다.

[0091] 도 11a 는, 각각의 타깃-언어 용어가 타깃-언어 앵커 (962) 에 나타나는 빈도를 나타낸 것이다. 도 11a 에 도시된 바와 같이, 용어 "casa" 및 "grande" 는 각각 5회 나타난다 (즉, 각각의 앵커 (962) 에서 한번). 따라서, 타깃 앵커 (962) 에서 나타나는 총 10 개의 용어 중에서 (5 개의 앵커 각각에서 앵커 당 2 개의 용어), "casa" 가 절반을 차지하고, "grande" 가 다른 절반을 차지한다. 따라서, 도 11a 에 도시된 바와 같이, 이러한 점에서, 용어 "house" 는 동일한 확률로 "casa" 또는 "grande" 에 매핑될 수 있는데, 이는 두 용어가 동일한 빈도로 나타나기 때문이다.

[0092] 그러나, 도 10 에 도시된 바와 같이, 그 시스템은 또한 용어 "house" 를 포함하고 페이지 (974) 에 포인팅하는 20 개의 영어 앵커 (964), 및 용어 "casa" 를 포함하고 페이지 (974) 에 또한 포인팅하는 10 개의 스페인어 앵커 (966) 를 찾는다. 다음으로, 도 11b 에 도시된 바와 같이, 용어 "house" 는 0.75 (즉, 15/20) 의 확률로 "casa" 에 매핑되고, 0.25 (즉, 5/20) 의 확률로 "grande" 에 매핑된다. 이들 확률은, 타깃 언어 앵커에서의 각 용어의 총 발생 수 (즉, "casa" 의 경우, 15) 를 타깃 언어 앵커에서의 중복을 포함한 용어의 총수 (즉, 앵커 (962) 에 포함된 10 개 및 앵커 (964) 에 포함된 10 개로서, 20 개의 용어) 로 단순하게 나눔으로써 계산된다. 다르게는, 또는 부가하여, 소정의 전환 또는 매핑의 확률을 계산 및/또는 개선시키기 위하여 다른 기술이 사용될 수도 있다. 예를 들어, 당업자는, 베이지안 (Bayesian) 방법, 히스토그램 평활화 (histogram smoothing), 커넬 (kernel) 평활화, 축소 추정량 (shrinkage estimators), 및/또는 다른 추정 기술과 같이 확률 추정치의 분산 에러를 감소시키기 위한 임의의 다양한 공지의 기술이 사용될 수 있음을 알 수 있다.

[0093] 만약 더 많은 앵커 텍스트가 이용가능하면, 확률은 훨씬 더 개선될 수 있다. 예를 들어, 최종 확률 분포는, "house" 가 "casa" 및 그 지소 형태 (diminutive form) "casita" 에 비교적 높은 확률로 매핑하고, "casino" 및 "mansión" 와 같은 용어 (맨션에 대한 스페인어 워드) 에 약간 더 낮은 확률로 매핑하며, "grande" 와 같은 용어에 무시할 수 있는 확률로 매핑하는 도 12 에 도시된 바와 유사할 수도 있다. 따라서, 가능성있는 동의어의 식별은 물론, 정확한 전환은, 전환되는 언어 및/또는 문자 세트의 정보 없이도 획득될 수 있다.

[0094] 질의 용어를 전환하면, 다음으로, 서버는 그 전환을 이용하여 서치를 실행할 수 있다. 예를 들어, 만약 사용자가 "hotels in Kyoto" 에 대한 로마지 질의를 입력하였다면, 서버로 하여금 그 질의의 가타가나, 히라가나, 및 간지 형태를 추정하게 하고, 그러한 질의를 이용하여 서치를 수행하게 하며, 적절한 사용자 인터페이스 내에

서, 각각의 그러한 질의에 대한 조합된 결과를 사용자에게 제공할 수 있게 하기 위해, 상술된 기술이 사용될 수 있다.

[0095] 도 10 내지 도 12 와 관련된 예는 한정적 목적이 아니라 예시를 목적으로 제공되며, 그 내에서 나타낸 방법에 대한 다수의 변경이 수행될 수 있음을 알 수 있다. 예를 들어, 확률에 도달하기 위해, 상이한 통계 기술이 사용될 수 있으며/있거나, 상술된 기본 기술에 대해 변경이 가능할 수 있다. 유사하게, 상술된 전환은 사용자에 의해 입력된 워드 또는 어구의 전환을 수행하는데 간단히 사용될 수 있으며, 관련 인터넷 서치를 수행하거나 확률 딕셔너리를 생성하는데 사용될 필요는 없음을 알 수 있다. 또한, 비록 상기 실시예는 사용자의 질의의 수신 이후에 발생하는 것으로서 전환 기술을 설명하였지만, 다른 실시형태에서는, 매핑 프로세스가 사용자의 질의를 수신하기 전에 수행될 수 있음을 알 수 있다. 그러한 사전-계산된 매핑은, 수신될 때에 사용자 질의를 전환하도록 제공되는, 도 8 에 설명된 바와 같은 딕셔너리에 저장될 수 있다. 마지막으로, 정렬된 앵커 텍스트 이외의 텍스트는 전환을 수행하는데 사용될 수 있음을 알 수 있다. 예를 들어, 정렬된 문장 또는 다른 데이터가 유사한 방식으로 사용될 수 있다. 다수의 국가에서, 하나 이상의 공인 언어 또는 인정 언어가 존재하며, 신문 및 정기 간행물은 종종 각각의 이들 언어로 기입된 동일한 기사를 포함한다. 워드 전환의 확률 딕셔너리를 작성하기 위하여 전술된 앵커 텍스트와 동일한 방식으로, 이러한 병렬 전환이 사용될 수 있다.

[0096] 따라서, 바람직한 실시형태는 사용자로 하여금 서치 질의 및/또는 전환 요청을 편리한 방식으로 (예를 들어, ASCII 키보드를 이용하여) 입력하게 하고, 정확하고 자동적인 전환 및 서치를 제공하게 할 수 있는 것이 바람직하다. 일부 실시형태에서는, 상술된 기본 모델에 대해 추가적인 개선이 행해질 수 있다. 예를 들어, 일부 실시형태에서는, 오리지널 질의에서 및/또는 다른 정렬된 앵커에서의 용어 개수와 유사한 용어의 수를 포함하는 앵커에 대해 선호 (웨이팅) 가 제공될 있다. 예를 들어, 도 10 에 도시된 시스템에서는, 오리지널 질의와 같이, 각각 단일 용어를 포함하기 때문에, 페이지 (974) 를 포인팅하는 앵커들에 대한 선호가 제공될 수도 있다. 유사하게, 만약 텍스트 "la casa grande" 를 포함하는 앵커가 또한 페이지 (972) 에 포인팅되어 있으면, 정렬된 다른 앵커보다 더 많은 용어 (즉, 3) 를 포함하기 때문에, 그 웨이팅은 적절한 인자에 의해 디스카운트될 수 있다. 그러한 웨이팅 방식은, 이들 앵커의 용어와 관련된 빈도를 적절한 인자로 승산함으로써, 도 11b 에 도시된 확률 계산에 반영될 수 있다.

[0097] 또한, 상술된 전환 프로세스는 서치 자체의 효과를 개선시키는데 이용될 수 있다. 예를 들어, 오리지널 질의 용어의 다양한 전환 및 동의어를 포함하도록 질의를 급히 확장하기 위해, 예를 들어, 확률 딕셔너리가 사용될 수 있다. 문헌 검색 이전에 사용자 질의를 확장함으로써, 동일한 "개념" 에 대한 동시 서치가 수행될 수 있으며, 이에 따라, 사용자가 검색하고 있는 것을 서치 결과가 포함할 가능성을 증가시킨다. 다른 방법으로, 또는 부가적으로, 확률 딕셔너리는, 문헌 용어의 확장을 제공함으로써, 통상의 문헌 인덱싱 프로세스를 보충하는데 사용될 수 있다. 예를 들어, 문헌에서 찾아진 용어는 확률 딕셔너리로부터의 전환을 갖는 문헌 인덱스에 보충될 수 있으며, 따라서, 오리지널 문헌에서 찾아진 동일한 용어를 정확히 사용하지 않는 서치에 의해서도 문헌이 위치지정되는 확률을 증가시킨다.

[0098] 상술된 전환 기술을 이용할 경우에 발생할 수도 있는 문제는, 데이터 희소성 (data sparsity; 예를 들어, "casa" 가 "house" 에 매핑된다고 결론적으로 결정하기에 충분한 앵커가 아님) 또는 다양성의 부족 (예를 들어, 모든 앵커가 동일한 것을 칭함) 으로 인해, 시스템이 충분히 정확한 확률 매핑에 도달할 수 없을 수도 있다는 점이다. 따라서, 일부 실시형태에서는, 사용자 행동 (behavior) 을 검사함으로써, 확률 매핑이 더 개선될 수 있다. 수개의 예시적인 기술이 아래에서 설명된다.

[0099] 예를 들어, 서버가 "house" 에 대한 전환의 획득을 원한다고 다시 한번 가정한다. 그러나, 찾아질 수 있는 유일한 앵커 텍스트는 어구 "big house" 또는 어구 "casa grande" 를 포함한다고 가정한다. 앵커 텍스트에서의 이러한 다양성의 부족으로 인해, 확률 딕셔너리는 다음의 매핑에 도달할 수도 있다.

[0100] house → casa, 0.5 의 확률

[0101] house → grande, 0.5 의 확률

[0102] big → casa, 0.5 의 확률

[0103] big → grande, 0.5 의 확률

[0104] grande → house, 0.5 의 확률

[0105] grande → big, 0.5 의 확률

[0106] casa → house, 0.5 의 확률

[0107] casa → big, 0.5 의 확률

[0108] 다음으로, 사용자가 용어 "casa" 로 서치 엔진을 질의한다고 가상해보자. 이 시점에서, 서치 엔진은 용어 "casa" 를 포함하는 페이지를 리턴할 수 있으며, 또한, 용어 "house" 만을 포함하는 N 개의 결과와 용어 "big" 만을 포함하는 M 개의 결과를 혼합할 수 있다. 실제로, N 및 M 은 매핑의 하위 확률을 고려하여 조정될 수 있어서, 비교적 가능성없는 매핑이 더 적은 디스플레이 결과를 야기하게 한다. 만약 용어 "big" 만을 포함하는 결과에 클릭하는 것보다 10배 더 많이 용어 "house" 만을 포함하는 결과에 클릭하는 것을 사용자가 발견하였으면, 매핑의 확률은, 예를 들어, 다음과 같이 조정될 수 있다.

[0109] house → casa, 0.9 의 확률

[0110] house → grande, 0.1 의 확률

[0111] big → casa, 0.1 의 확률

[0112] big → grande, 0.9 의 확률

[0113] grande → house, 0.1 의 확률

[0114] grande → big, 0.9 의 확률

[0115] casa → house, 0.9 의 확률

[0116] casa → big, 0.1 의 확률

[0117] 실제 숫자는, 클릭이 고려되는 사용자의 수, 양자의 용어를 포함하는 페이지에 대한 클릭의 수, 결과 세트 중에서 당해 용어를 포함하는 결과의 배치 등과 같이 다양한 다른 인자에 의존할 수 있다. 또한, 이 예에서 주어진 조정 확률 (즉, 0.1 및 0.9) 은 예시를 목적으로 함을 알 수 있다. 당업자는, 상술된 바와 같은 사용자 피드백에 주어지는 실제 웨이팅은 임의의 적절한 방식으로 구현될 수 있음을 알 수 있다.

[0118] 또한, 상기 예는 사용자 피드백의 사용의 설명을 용이하게 하도록 간략화되었다. 예를 들어, 일부 시스템에서, 소정의 전환의 수행을 원조하기 위해 다른 전환으로부터 획득된 정보를 이용하는 것이 가능하다. 예를 들어, 직전에 제공된 예에서, 비록 용어 "house" 가 "big house" 를 칭하는 앵커 텍스트에서만 나타났지만, "house" 가 "grande" 보다 "casa" 에 더 적절히 매핑된다고 여전히 결정할 수도 있다. 예를 들어, 만약 "big" 이 매우 높은 확률로 그리고 충분히 큰 데이터 세트에 걸쳐 "grande" 에 매핑되었다고 이미 결정하였으면 (앵커 텍스트가 좀처럼 동의어의 리스트로 이루어지지 않는다고 가정하면), 비록 "house" 또는 "casa" 가 여전히 확정적이지 않더라도, house 대 casa 매핑은 house 대 grande 매핑에 비하여 여전히 선호된다.

[0119] 또한, 전환의 정확도 및/또는 서치 결과의 유용성은 사용자의 질의 세션 이력을 검사함으로써 개선될 수 있다. 예를 들어, 많은 경우, 시스템은, 사용자가 입력한 이전의 질의를 (예를 들어, 서버에서의 사용자의 계정에 저장된 쿠키 또는 정보를 통하여) 알 것이다. 이러한 이력 데이터는 그 사용자로부터의 질의의 가능한 의미를 랭크하는데 이용될 수 있으며, 따라서, 피싱 (fishing)-관련 질의에 대한 "bank" 를 플라잉 (flying) 에 관

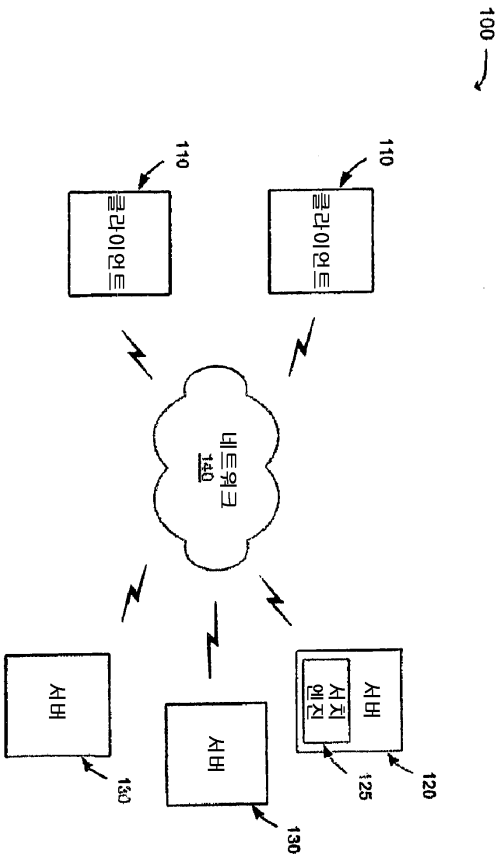
련된 것으로부터 잠재적으로 명확히 할 수 있다. 따라서, 이러한 프로세스는 가능한 전환 세트를 좁히는데 이용될 수 있다. 일부 실시형태에서, 시스템은 사용자 인터페이스에서 "Did you mean to search for X" 와 같은 메시지 (여기서, X 는 예상 전환 선호를 나타냄) 와 함께 디스플레이하면서, 또한, 각각의 가능한 재공식화 (reformulation) 로부터의 적은 수의 결과들은 결과들의 제 1 페이지에 잠재적으로 디스플레이함으로써 이들을 제안할 수도 있다. 사용자가 "did you mean ..." 디스플레이에 의해 제안된 대안 중 하나, 또는 결과 페이지에 제공된 결과 중 하나를 선택할 경우, 시스템은, 사용자의 가능성있는 서치 바이어스는 물론, 질의 워드 (들) 의 가능성있는 전환에 관한 추가적인 증거를 획득한다. 그 후, 이들 신호 양자는, 사용자-특정인 경우뿐 아니라 일반적인 경우 양자에서, (예를 들어, 확률 딕셔너리에서) 용어 매핑에 대한 가능성 스코어를 업데이트하기 위해 시스템에 의해 사용될 수 있다.

[0120] **D. 결론**

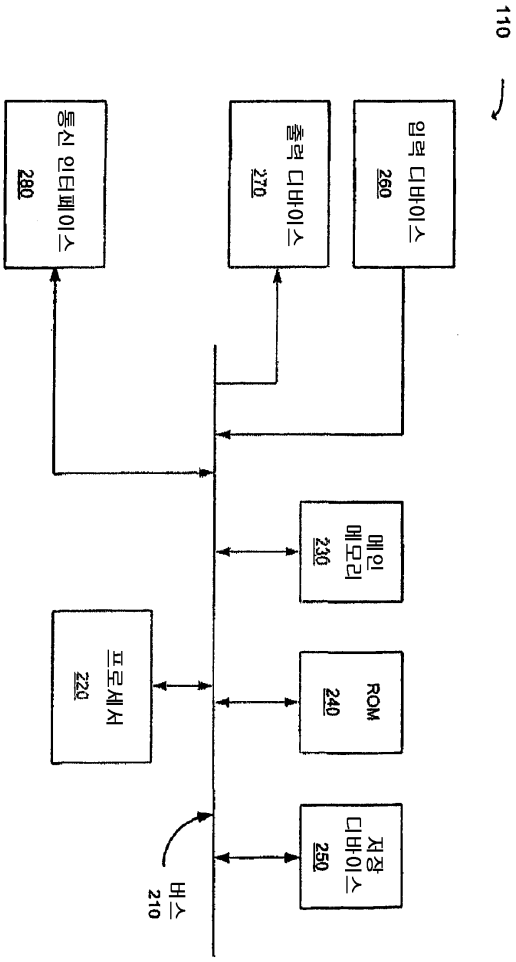
[0121] 상기에서 상세히 설명된 바와 같이, 본 발명과 부합하는 방법 및 시스템은 모호한 서치 질의에 응답하여 서치 결과를 제공하고/하거나 다른 문자 세트 및/또는 언어로 용어를 전환하는데 이용될 수 있다. 다양한 전환 및 서치 기술 및 시스템이 설명되었다. 그러나, 전술한 설명은 예시적으로 제공되었으며 상기 교시의 관점에서 또는 본 발명의 실시를 통하여 다양한 변형 및 변경이 가능함을 알 수 있다. 예를 들어, 비록 전술한 설명이 클라이언트-서버 구조에 기초하지만, 당업자는 피어-투-피어 (peer-to-peer) 구조가 본 발명과 부합하게 사용될 수도 있음을 알 수 있다. 또한, 비록 설명된 구현이 소프트웨어를 포함하지만, 본 발명은 하드웨어와 소프트웨어의 조합, 또는 하드웨어 단독으로 구현될 수도 있다. 또한, 비록 본 발명의 양태가 메모리에 저장되는 것으로서 설명되지만, 당업자는 이들 양태가 또한 하드 디스크, 플로피 디스크, 또는 CD-ROM 과 같은 2차 저장 디바이스; 인터넷으로부터의 캐리어파; 또는 다른 형태의 RAM 또는 ROM 과 같은 다른 타입의 컴퓨터-판독가능 매체에 저장될 수도 있다. 따라서, 본 발명의 범위는 특허청구범위 및 그 균등물에 의해 정의된다.

도면

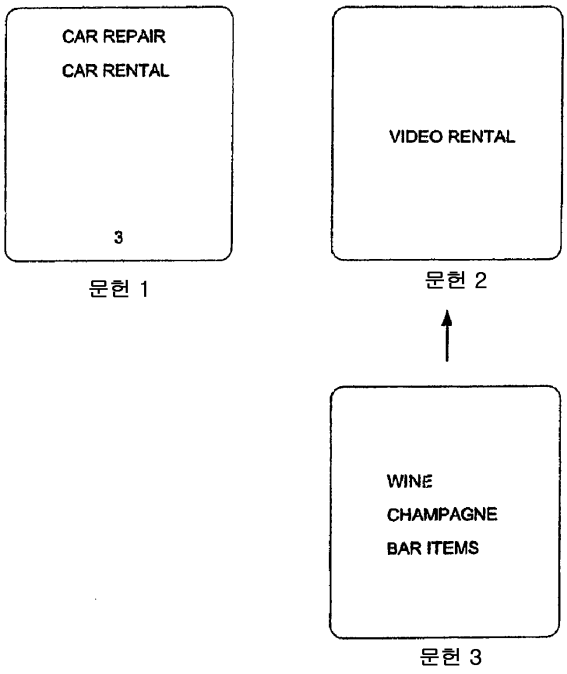
도면1



도면2


110
입력 디바이스
260
출력 디바이스
270
통신 인터페이스
280
메인 메모리
230
ROM
240
프로세서
220
저장 디바이스
250
버스
210

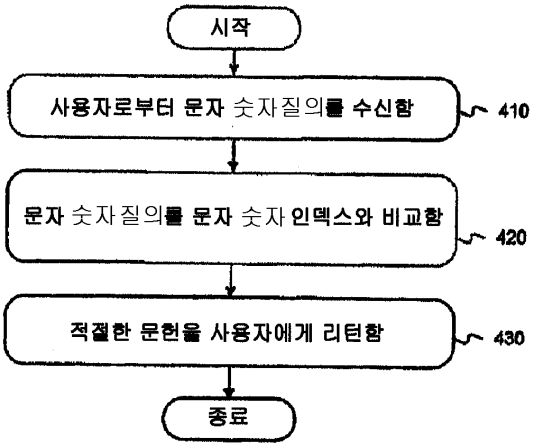
도면3



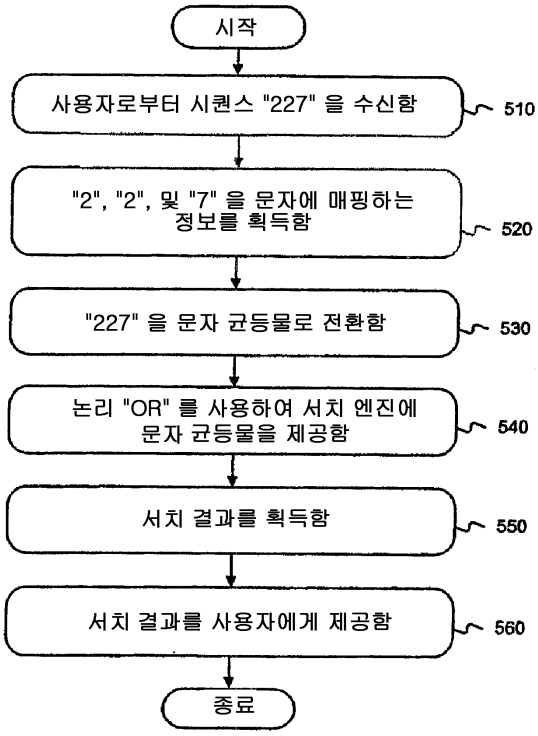
도면4a

용어	위치 (문헌)
3	문헌 1
BAR	문헌 3
CAR	문헌 1
CHAMPAGNE	문헌 3
ITEMS	문헌 3
RENTAL	문헌 1 및 2
REPAIR	문헌 1
VIDEO	문헌 2
WINE	문헌 3

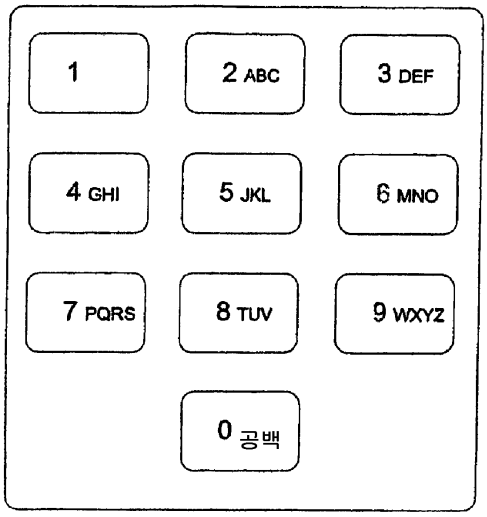
도면4b


시작
사용자로부터 문자 숫자질의를 수신함
410
문자 숫자질의를 문자 숫자 인덱스와 비교함
420
적절한 문헌을 사용자에게 리턴함
430
종료

도면5a


시작
사용자로부터 시퀀스 "227" 을 수신함
510
"2", "2", 및 "7" 을 문자에 매핑하는 정보를 획득함
520
"227" 을 문자 균등물로 전환함
530
논리 "OR" 를 사용하여 서치 엔진에 문자 균등물을 제공함
540
서치 결과를 획득함
550
서치 결과를 사용자에게 제공함
560
종료

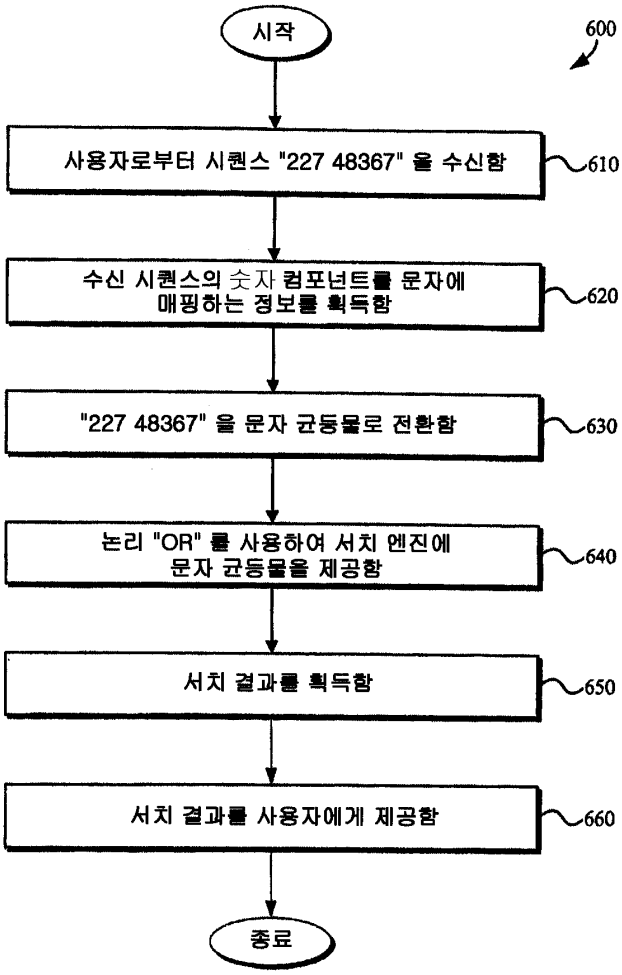
도면5b



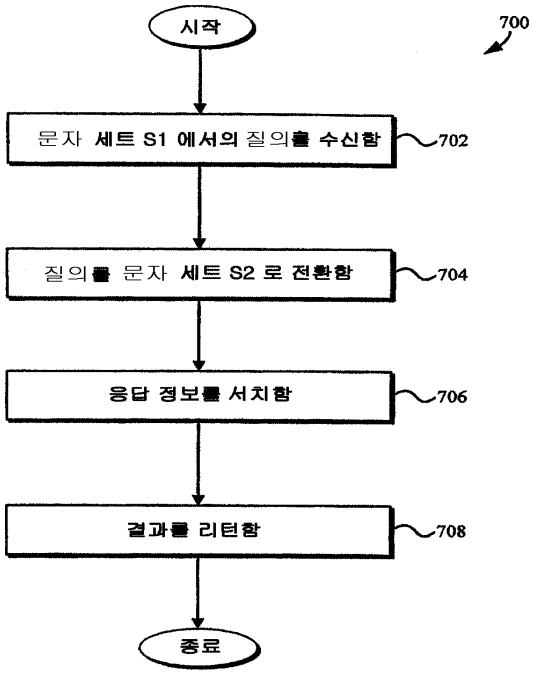
도면5c

용어	위치(문헌)
3	문헌 1
227	문헌 1 및 3
242672463	문헌 3
48367	문헌 3
736825	문헌 1 및 2
737247	문헌 1
84336	문헌 2
8463	문헌 3

도면6


시작
600
사용자로부터 시퀀스 "227 48367" 을 수신함
610
수신 시퀀스의 숫자 컴포넌트를 문자에 매핑하는 정보를 획득함
620
"227 48367" 을 문자 균등물로 전환함
630
논리 "OR" 를 사용하여 서치 엔진에 문자 균등물을 제공함
640
서치 결과를 획득함
650
서치 결과를 사용자에게 제공함
660
종료

도면7

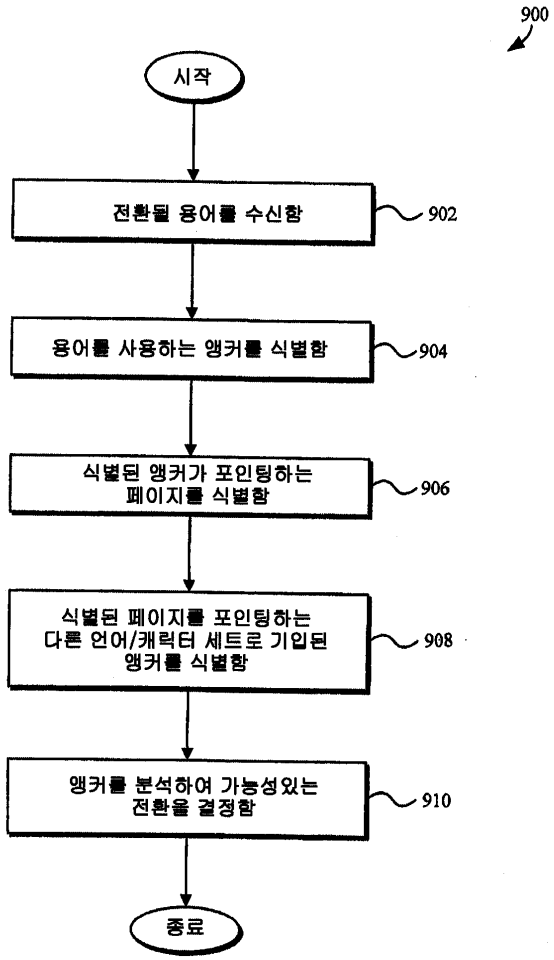


도면8

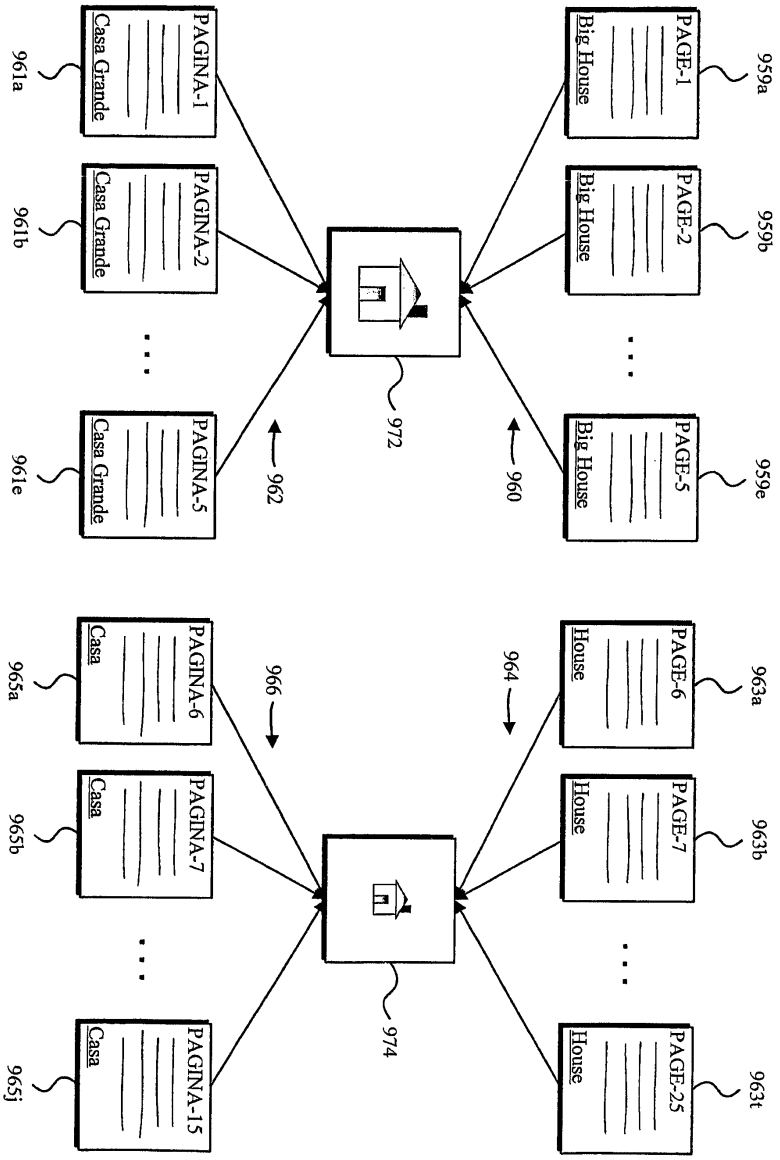
800

802 로마지 용어	804 간지 용어	806 확률 (%)
808 <Bank>romaji	<Financial institution>kanji	0.4
	<Steep slope>kanji	0.3
	<Airplane maneuver>kanji	0.2
	<Other>kanji	0.1
810 <Car>romaji	<Automobile>kanji	0.9
	<Other>kanji	0.1
812 <House>romaji	<A dwelling>kanji	0.7
	<To contain>kanji	0.25
814 <Plane>romaji	<Airplane>kanji	0.6
	<Flat surface>kanji	0.25
	<Carpenter's tool>kanji	0.1
	<Other>kanji	0.05

도면9


900
시작
전환될 용어를 수신함
902
용어를 사용하는 앵커를 식별함
904
식별된 앵커가 포인팅하는
페이지를 식별함
906
식별된 페이지를 포인팅하는
다른 언어/캐릭터 세트로 기입된
앵커를 식별함
908
앵커를 분석하여 가능성있는
전환을 결정함
910
종료

도면10



도면11a

House

1	Casa
2	Casa
3	Casa
4	Casa
5	Casa
6	Grande
7	Grande
8	Grande
9	Grande
10	Grande

확률 (house=casa): 5/10 = 0.5

확률 (house=grande): 5/10 = 0.5

도면11b

House

1	Casa
2	Casa
3	Casa
. . .	Casa
13	Casa
14	Casa
15	Casa
16	Grande
17	Grande
18	Grande
19	Grande
20	Grande

확률 (house=casa): 15/20 = 0.75

확률 (house=grande): 5/20 = 0.25

도면12