(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년02월20일
(11) 등록번호 10-2079860
(24) 등록일자 2020년02월14일

(51) 국제특허분류(Int. Cl.)
G06F 16/00 (2019.01)
(52) CPC특허분류
G06F 16/00 (2019.01)
(21) 출원번호 10-2017-7025509
(22) 출원일자(국제) 2016년02월04일
심사청구일자 2018년10월30일
(85) 번역문제출일자 2017년09월11일
(65) 공개번호 10-2017-0117481
(43) 공개일자 2017년10월23일
(86) 국제출원번호 PCT/CN2016/073441
(87) 국제공개번호 WO 2016/127904
국제공개일자 2016년08월18일
(30) 우선권주장
201510079914.6 2015년02월13일 중국(CN)
(56) 선행기술조사문헌
KR101153030 B1*
KR1020120090131 A*
Peter Christen, Alan Willmore, and Tim Churches, "A Probabilistic Geocoding System Utilising a Parcel Based Address File", 2006*
JP2010539738 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
알리바바 그룹 홀딩 리미티드
케이만군도, 그랜드 케이만, 피오박스 847, 원 캐피탈 플레이스 4층
(72) 발명자
슝, 준
중국 311121 항저우시 위항 디스트릭트 웨스트 엔이 로드 넘버 969 빌딩 3 5층 알리바바 그룹 리갈 디파트먼트
(74) 대리인
특허법인 광장리앤고

전체 청구항 수 : 총 26 항 심사관 : 권현수

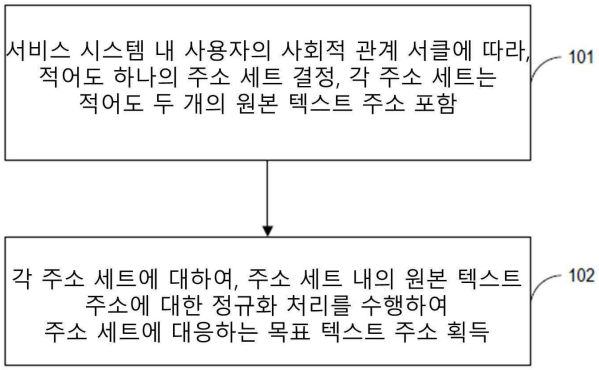
(54) 발명의 명칭 **텍스트 주소 처리 방법 및 장치**

(57) 요 약

본 출원은 텍스트 주소 처리 방법 및 장치를 제공한다. 일부 방법 실시예는: 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트 결정 -- 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함 --; 및, 각 주소 세트에 대하여, 주소 세트 내의 원본 텍스트 주소에 대해 정규화 처리를 수행하여 주소 세트에

(뒷면에 계속)

대 표 도 - 도1



대응하는 목표 텍스트 주소 획득을 포함한다. 본 출원은 사용자의 사회적 관계 서클에 따라 정규화 대상 원본 텍스트 주소를 분할하며, 이는 한편으로는 정규화 대상 원본 텍스트 주소의 범위를 줄이는 것에 해당하고, 다른 한편으로는 연관되는 텍스트 주소 사이로 텍스트 주소의 정규화를 고정하는 것에 해당한다. 따라서, 텍스트 주소 사이의 내결함성 경계 제어가 용이해질 수 있고, 정규화 결과의 정확도를 높이는 데 도움이 될 수 있다.

명 세 서

청구범위

청구항 1

서비스(業務) 시스템 내 사용자의 사회적 관계 서클에 따라 적어도 하나의 주소 세트를 결정하는 것 -- 상기 적어도 하나의 주소 세트의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함함 --; 및

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 각 주소 세트에 대하여 상기 주소 세트 내의 원본 텍스트 주소에 대한 정규화를 수행하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 2

제1항에 있어서, 상기 서비스 시스템 내 사용자의 사회적 관계 서클에 따라 적어도 하나의 주소 세트를 결정하는 것은:

상기 서비스 시스템 내 사용자의 사회적 관계 서클을 결정하는 것; 및

주소 세트를 구성하기(make up) 위하여 상기 사용자가 사용하는 텍스트 주소 및 상기 사회적 관계 서클 내의 사용자들이 사용하는 텍스트 주소를 획득하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 3

제1항에 있어서, 상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 상기 주소 세트 내의 원본 텍스트 주소에 대한 정규화 를 수행하는 것은:

상기 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것; 및

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 상기 유사도에 따라 상기 두 개의 원본 텍스트 주소가 상기 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지를 결정하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 4

제3항에 있어서, 상기 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것은:

상기 두 개의 원본 텍스트 주소 각각의 표준 단편 특징, 경도 및 위도 특징, 또는 영숫자 특징 중 적어도 하나를 추출하는 것; 및

각 추출된 특징에 따라 상기 추출된 특징에 대응하는 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 5

제4항에 있어서, 상기 각 추출된 특징에 따라 상기 특징에 대응하는 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것은:

상기 추출된 특징이 표준 단편 특징인 것에 대응하여, 심해시(SimHash) 알고리즘을 사용하여 표준 단편 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것;

상기 추출된 특징이 경도 및 위도 특징인 것에 대응하여, 경도 및 위도 거리 알고리즘을 사용하여 경도 및 위도 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것; 및

상기 추출된 특징이 영숫자 특징인 것에 대응하여, 자카드(Jaccard) 계수 알고리즘을 사용하여 영숫자 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 6

제1항에 있어서, 목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계에 따라 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 결정하는 것; 및

상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것을 더 포함하는 텍스트 주소 처리 방법.

청구항 7

제6항에 있어서, 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것은:

상기 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편의 단편 교집합을 획득하는 것; 및

상기 단편 교집합에 따라 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 8

제7항에 있어서, 상기 단편 교집합에 따라 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것은:

상기 단편 교집합이 상기 적어도 두 개의 목표 텍스트 주소 중 하나를 나타내는 것에 대응하여, 상기 적어도 두 개의 목표 텍스트 주소를 상기 단편 교집합이 나타내는 목표 텍스트 주소로 정규화하는 것을 포함하는 텍스트 주소 처리 방법.

청구항 9

제8항에 있어서,

상기 단편 교집합 및 상기 단편 교집합이 나타내는 목표 텍스트 주소를 특징 지식 베이스로 저장하는 것을 더 포함하는 텍스트 주소 처리 방법.

청구항 10

서비스 시스템 내 사용자의 사회적 관계 서클에 따라 적어도 하나의 주소 세트를 결정하도록 구성되는 결정 모듈 -- 상기 적어도 하나의 주소 세트의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함함 --; 및

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 각 주소 세트에 대하여 상기 주소 세트 내의 원본 텍스트 주소에 대한 정규화를 수행하도록 구성되는 정규화 모듈을 포함하는 텍스트 주소 처리 장치.

청구항 11

제10항에 있어서, 상기 결정 모듈은:

상기 서비스 시스템 내 사용자의 사회적 관계 서클을 결정하고;

주소 세트를 구성하기 위해, 상기 사용자가 사용하는 텍스트 주소 및 상기 사회적 관계 서클 내의 사용자들이 사용하는 텍스트 주소를 획득하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 12

제10항에 있어서, 상기 정규화 모듈은:

상기 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하고;

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 상기 유사도에 따라 상기 두 개의 원본 텍스트 주소가 상기 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지를 결정하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 13

제12항에 있어서, 상기 정규화 모듈은:

상기 두 개의 원본 텍스트 주소의 표준 단편 특징, 경도 및 위도 특징, 또는 영숫자 특징 중 적어도 하나를 추출하고;

각 추출된 특징에 따라 상기 추출된 특징에 대응하는 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 14

제13항에 있어서, 상기 정규화 모듈은:

상기 추출된 특징이 표준 단편 특징이면, 심해시 알고리즘을 사용하여 표준 단편 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하고;

상기 추출된 특징이 경도 및 위도 특징이면, 경도 및 위도 거리 알고리즘을 사용하여 경도 및 위도 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하며;

상기 특징이 영숫자 특징이면, 자카드 계수 알고리즘을 사용하여 영숫자 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 15

제10항에 있어서,

상기 결정 모듈은 목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계에 따라 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 결정하도록 더 구성되고;

상기 정규화 모듈은 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 16

제15항에 있어서, 상기 정규화 모듈은:

상기 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편의 단편 교집합을 획득하고;

상기 단편 교집합에 따라 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하도록 더 구성되는 텍스트 주소 처리 장치.

청구항 17

제16항에 있어서,

상기 단편 교집합이 상기 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 때, 상기 단편 교집합 및 상기 단편 교집합이 나타내는 목표 텍스트 주소를 저장하도록 구성되는 특징 지식 베이스를 더 포함하는 텍스트 주소 처리 장치.

청구항 18

텍스트 주소 처리 장치의 적어도 하나의 프로세서에 의해 실행 가능하고, 상기 텍스트 주소 처리 장치가 텍스트 주소 처리 방법을 수행하게 하는 명령어 세트를 저장하는 비일시적 컴퓨터 판독가능 매체로서, 상기 방법은:

서비스 시스템 내 사용자의 사회적 관계 서클에 따라 적어도 하나의 주소 세트를 결정하는 것 -- 상기 적어도 하나의 주소 세트의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함함 --; 및

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 각 주소 세트에 대하여 상기 주소 세트 내의 원본 텍스트 주소에 대한 정규화를 수행하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 19

제18항에 있어서, 상기 서비스 시스템 내 사용자의 사회적 관계 서클에 따라 적어도 하나의 주소 세트를 결정하

는 것은:

상기 서비스 시스템 내 사용자의 사회적 관계 서클을 결정하는 것; 및

주소 세트를 구성하기 위하여 상기 사용자가 사용하는 텍스트 주소 및 상기 사회적 관계 서클 내의 사용자들이 사용하는 텍스트 주소를 획득하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 20

제18항에 있어서, 상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 상기 주소 세트 내의 원본 텍스트 주소에 대한 정규화를 수행하는 것은:

상기 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것; 및

상기 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여, 상기 유사도에 따라 상기 두 개의 원본 텍스트 주소가 상기 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지를 결정하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 21

제20항에 있어서, 상기 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것은:

상기 두 개의 원본 텍스트 주소 각각의 표준 단편 특징, 경도 및 위도 특징, 및 영숫자 특징 중 적어도 하나를 추출하는 것; 및

각 추출된 특징에 따라, 상기 특징에 대응하는 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 22

제21항에 있어서, 상기 각 추출된 특징에 따라, 상기 특징에 대응하는 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것은:

상기 추출된 특징이 표준 단편 특징인 것에 대응하여, 심해시 알고리즘을 사용하여 표준 단편 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것;

상기 추출된 특징이 경도 및 위도 특징인 것에 대응하여, 경도 및 위도 거리 알고리즘을 사용하여 경도 및 위도 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것; 및

상기 추출된 특징이 영숫자 특징인 것에 대응하여, 자카드 계수 알고리즘을 사용하여 영숫자 특징 차원에서 상기 두 개의 원본 텍스트 주소 사이의 유사도를 결정하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 23

제18항에 있어서, 상기 텍스트 주소 처리 장치의 상기 적어도 하나의 프로세서에 의해 실행 가능한 상기 명령어 세트는 상기 텍스트 주소 처리 장치가:

목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계에 따라 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 결정하고;

상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것을 더 수행하게 하는 비일시적 컴퓨터 판독가능 매체.

청구항 24

제23항에 있어서, 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것은:

상기 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편의 단편 교집합을 획득하는 것; 및

상기 단편 교집합에 따라 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것을 포함하는 비일

시적 컴퓨터 판독가능 매체.

청구항 25

제24항에 있어서, 상기 단편 교집합에 따라 상기 적어도 두 개의 목표 텍스트 주소에 대해 정규화를 수행하는 것은:

상기 단편 교집합이 상기 적어도 두 개의 목표 텍스트 주소 중 하나를 나타내는 것에 대응하여, 상기 적어도 두 개의 목표 텍스트 주소를 상기 단편 교집합이 나타내는 목표 텍스트 주소로 정규화하는 것을 포함하는 비일시적 컴퓨터 판독가능 매체.

청구항 26

제25항에 있어서, 상기 텍스트 주소 처리 장치의 상기 적어도 하나의 프로세서에 의해 실행 가능한 상기 명령어 세트는 상기 텍스트 주소 처리 장치가:

상기 단편 교집합 및 상기 단편 교집합이 나타내는 목표 텍스트 주소를 특징 지식 베이스로 저장하는 것을 더 수행하게 하는 비일시적 컴퓨터 판독가능 매체.

발명의 설명

기 술 분 야

[0001] 본 출원은 일반적으로 통신 기술 분야에 관한 것이며, 구체적으로 텍스트 주소 처리 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 인터넷 기술의 발달과 함께 인터넷 기반 응용 프로그램이 점점 널리 사용되고 있다. 인터넷 응용 프로그램에서, 사용자는 종종 집 주소, 회사 주소 등과 같은 일부 주소 정보를 텍스트로 기입해야 한다. 사용자에 의한 주소 정보의 텍스트 설명을 텍스트 주소라고 지칭한다. 주소 정보를 텍스트로 기입할 때, 사용자는 일반적으로 자신의 설명 습관을 가지며, 상이한 사용자의 설명 습관은 일반적으로 상이하다. 결과적으로 동일한 주소 정보가 상이한 텍스트 주소에 대응한다. 예를 들면, "빌딩1, 유닛 1"과 같은 주소 정보에 대하여, 일부 사용자는 주소 정보를 "101"로 기술하고, 일부 사용자는 주소 정보를 "1-01"로 기술하며, 일부 사용자는 주소 정보를 "빌딩 1, 유닛 1"로 기술하는 등이다. 이로 인해 동일한 주소 정보가 여러 텍스트 주소를 가지게 될 수 있다.

[0003] 주소 정보의 관리 및 적용을 용이하게 하기 위하여(예를 들면, 주소 정보에 기반하여 애플리케이션 마이닝(application mining) 등이 수행될 수 있다), 텍스트 주소에 대한 정규화 처리를 수행할 필요가 있다. 즉, 동일한 주소 정보에 대응하는 상이한 텍스트 주소를 하나의 텍스트 주소로 통합할 필요가 있다. 텍스트 주소를 정규화하기 위한 기존의 생각은 주로 정규화될 모든 텍스트 주소를 식별하고, 텍스트 주소에 포함된 표준 단편을 추출한 다음, 텍스트 주소에 포함된 표준 단편을 기반으로 상관 정도를 쌍으로 계산하고, 두 텍스트 주소 사이의 상관 정도에 기반하여 두 텍스트 주소가 정규화되어야 하는지 판단한다.

[0004] 텍스트 주소가 다양화됨에 따라, 동일한 주소 정보를 기술하기 위한 상이한 텍스트 주소의 텍스트 내용은 크게 달라질 수 있는 한편, 상이한 주소 정보를 기술하기 위한 텍스트 주소의 텍스트 내용이 약간 다를 수도 있다. 이러한 이유로, 많은 양의 텍스트 주소를 정규화할 경우, 내결함성(fault-tolerant) 경계를 제어하는 것이 어려울 수 있으며, 이는 텍스트 주소 정규화 결과의 정확성을 상대적으로 낮춘다.

발명의 내용

[0005] 다양한 양상에서, 본 출원은 텍스트 주소의 정규화 결과의 정확성을 개선하기 위한 텍스트 주소 처리 방법 및 장치를 제공한다.

[0006] 일 양상에서, 본 출원은 텍스트 주소 처리 방법을 제공하며, 방법은:

[0007] 서비스(業務) 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트 결정 -- 상기 적어도 하나의 주소 세트 내의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함 --; 및

[0008] 각 주소 세트에 대하여, 주소 세트 내의 원본 텍스트 주소에 대한 정규화 처리를 수행하여, 주소 세트에 대응하

는 목표 텍스트 주소 획득을 포함한다.

[0009] 다른 양상에서, 본 출원은 텍스트 주소 처리 장치를 제공하며, 텍스트 주소 처리 장치는:

[0010] 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트를 결정하도록 구성되는 결정 모듈 -- 상기 적어도 하나의 주소 세트 내의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함 --; 및

[0011] 각 주소 세트에 대하여, 주소 세트에 대응하는 목표 텍스트 주소를 획득하기 위하여 주소 세트 내의 원본 텍스트 주소에 대한 정규화 처리를 수행하도록 구성되는 정규화 모듈을 포함한다.

[0012] 본 출원에서, 적어도 하나의 주소 세트는 서비스 시스템 내 사용자의 사회적 관계 서클에 따라 결정된다. 정규화 처리는 주소 세트를 단위로 하여 각 주소 세트 내의 원본 텍스트 주소에 대해 각각 수행하여, 각 주소 세트에 대응하는 목표 텍스트 주소를 획득하고, 따라서 텍스트 주소에 대한 정규화 처리를 달성한다. 정규화 대상 원본 텍스트 주소가 사용자의 사회적 관계 서클에 따라 분할되므로, 한편으로, 정규화 대상 원본 텍스트 주소의 범위가 사용자의 사회적 관계 서클로 제한되어, 정규화 대상 원본 텍스트 주소의 범위를 줄이는 것에 해당한다. 다른 한편으로, 사회적 관계 서클 외부의 사용자가 사용하는 텍스트 주소와 비교하여, 사회적 관계 서클 내의 사용자가 사용하는 텍스트 주소는 어느 정도의 연관을 가지므로, 텍스트 주소의 정규화를 연관성이 있는 텍스트 주소 사이로 고정하는(locking) 것에 해당한다. 이 방법으로, 텍스트 주소 사이의 내결함성(fault-tolerant) 경계가 더 용이하게 제어될 수 있으며, 텍스트 주소의 정규화 결과의 정확성을 개선하는 데 도움이 된다.

도면의 간단한 설명

[0013] 본 출원의 실시예에서의 기술적 해결책을 더욱 명확하게 설명하기 위해, 실시예 또는 종래 기술에 대한 설명에 사용되는 첨부 도면을 이하에서 간략하게 소개한다. 다음의 설명에서 첨부된 도면은 본 출원의 일부 실시예를 도시하는 것이 명백하다. 또한, 이 분야의 기술자는 창의적인 노력을 기울이지 않고 첨부 도면에 따라 다른 첨부 도면을 얻을 수 있다.

도 1은 본 출원의 실시예에 따른 텍스트 주소 처리 방법의 개략적인 흐름도이다.

도 2는 본 출원의 실시예에 따른 정규화 처리의 개략도이다.

도 3은 본 출원의 실시예에 따른 텍스트 주소 처리 장치의 개략적인 구조도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 본 출원의 실시예의 목적, 기술적 해결책 및 장점을 더욱 명확하게 하기 위하여, 이하에서 본 출원의 일부 실시예의 기술적 해결책이 첨부 도면을 참조하여 명확하고 완전하게 설명된다. 설명된 실시예가 본원의 모든 실시예가 아니라 일부인 것이 명백하다. 창조적인 노력이 없이 본 출원의 실시예에 기반하여 이 분야의 기술자가 얻은 다른 모든 실시예는 본 출원의 보호 범위에 속한다.

[0015] 도 1은 본 출원의 실시예에 따른 텍스트 주소 처리 방법의 개략적인 흐름도이다. 도 1에 나타난 바와 같이, 방법은 다음을 포함한다:

[0016] 101: 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트 결정, 적어도 하나의 주소 세트 내의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함.

[0017] 102: 각 주소 세트에 대하여, 주소 세트 내의 원본 텍스트 주소에 대해 정규화 처리를 수행하여, 주소 세트에 대응하는 목표 텍스트 주소 획득.

[0018] 이 실시예에 따른 텍스트 주소 처리 방법은 텍스트 주소 처리 장치에 의해 실행될 수 있다. 이 실시예에 따른 방법은 주로 텍스트 주소에 대한 정규화 처리를 수행하는 데 사용된다.

[0019] 우선, 이 실시예에서 텍스트 주소는 주소 정보에 대한 텍스트 설명을 지칭하는 점을 유의해야 한다. 상이한 텍스트 주소는 동일한 주소 정보에 대한 텍스트 설명일 수 있다. 또한, 설명 및 구별의 편이를 위하여, 이 실시예에서 정규화 이전의 텍스트 주소는 원본 텍스트 주소로 지칭되며, 정규화 이후에 얻어진 텍스트 주소는 목표 텍스트 주소로 지칭된다. 원본 텍스트 주소와 목표 텍스트 주소는 모두 주소 정보에 대한 텍스트 설명이다.

[0020] 일반적으로, 정규화에 대한 요구가 있을 때에만 텍스트 주소에 대해 정규화 처리가 수행된다. 텍스트 주소에 대한 정규화 처리 수행의 요구는 일반적으로 특정한 서비스 시스템 또는 일부 서비스 시스템에 특유하다. 요약하자면, 정규화된 텍스트 주소를 통해 새로운 서비스 또는 새로운 서비스 요구가 발굴되거나 관련 정보의 통계 분

석이 가능하게 되는 등을 위하여 특정한 서비스 시스템 또는 일부 서비스 시스템에 대해 그와 연관된 텍스트 주소를 정규화하는 것이 필요하다.

[0021] 이 실시예는 서비스 시스템을 제한하지 않는다는 점을 유의해야 한다. 서비스 시스템은 텍스트 주소에 관한 다양한 서비스 시스템 일 수 있으며, 예를 들면, 전자 상거래 시스템, 온라인 지불 시스템, 인스턴트 메시징 시스템, 전자 메일 시스템 등일 수 있다.

[0022] 텍스트 주소에 대해 정규화 처리가 수행되기 전에, 서비스 시스템과 연관된 정규화 대상 원본 텍스트 주소를 결정하는 것이 필요하다. 본 출원에서, 서비스 시스템과 연관된 원본 텍스트 주소는 서비스 시스템 내 사용자의 사회적 관계 서클에 따라 결정될 수 있다. 사용자에 대해, 자신의 사회적 관계 서클은 사용자와 연관 관계가 있는 다른 사용자를 주로 포함한다. 바람직하게는, 사용자의 사회적 관계 서클로서, 사용자와 긴밀한 연관이 있는 사용자들이, 사용자와 연관 관계가 있는 사용자들로부터 선택될 수 있다. 예를 들면, 사용자의 사회적 관계 서클은 다음 방법 중 적어도 하나로 획득될 수 있다.

[0023] 사용자와 금융 거래(예를 들면, 계좌 이체)가 있는 다른 사용자는 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다. 바람직하게는, 사용자와의 이체 빈도 또는 금액이 임계치를 넘는 다른 사용자가 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다.

[0024] 사용자의 주소록에 있는 다른 사용자는 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다. 일반적으로, 사용자의 승인에 따라, 각 응용 프로그램이 사용자의 주소록을 읽을 수 있다.

[0025] 사용자와 인스턴트 메시징 도구로 통신하는 다른 사용자는 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다. 인스턴트 메시징 도구는, 위챗(WeChat), QQ 등을 포함하지만 이에 제한되지 않는다. 바람직하게는, 사용자와의 상호작용 빈도 또는 통신 시간이 임계치를 넘는 다른 사용자가 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다.

[0026] 사용자와 동일한 장치를 사용하는 다른 사용자는 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다. 여기에서의 장치는 컴퓨터, 이동전화, WIFI 등을 포함할 수 있다. 바람직하게는, 사용자와 동일한 장치를 사용하는 빈도 또는 시간이 임계치를 넘는 다른 사용자가 사용자의 사회적 관계 서클 내의 사용자로 획득될 수 있다.

[0027] 특히, 텍스트 주소에 대해 정규화 처리 수행이 필요할 때, 텍스트 주소 처리 장치는, 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트를 결정한다. 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함한다. 처리에서, 사용자의 사회적 관계 서클에 따라 서비스 시스템에 연관된 원본 텍스트 주소의 결정에 더하여, 동시에, 서비스 시스템에 연관된 원본 텍스트 주소가 분할된다. 서비스 시스템에 연관된 원본 텍스트 주소는 상이한 주소 세트로 분할된다.

[0028] 선택적인 구현에서, 주소 세트의 수는 서비스 시스템 내의 사용자 수에 따라 결정되며, 예를 들면, 한 명의 사용자가 하나의 주소 세트에 대응될 수 있다.

[0029] 특히, 서비스 시스템 내의 각 사용자에 대하여, 텍스트 주소 처리 장치는 먼저 사용자의 사회적 관계 서클을 결정하여야 한다(구체적으로, 사용자의 사회적 관계 서클은 상술한 방법으로 결정될 수 있다). 다음으로, 사용자가 사용한 주소 정보 및 사용자의 사회적 관계 서클 내의 사용자들이 사용한 주소 정보가 주소 세트로 획득될 수 있다.

[0030] 각 주소 세트에 대하여, 텍스트 주소 처리 장치는 주소 세트 내의 원본 텍스트 주소에 대해 정규화 처리를 수행하여 주소 세트에 대응하는 목표 텍스트 주소를 얻는다. 이는 텍스트 주소에 대한 정규화 처리를 각 주소 세트로 제한하는 것에 해당한다. 한편으로, 이는 정규화 대상 원본 텍스트 주소의 범위를 줄이는 것에 해당한다. 다른 한편으로, 사회적 관계 서클 외부의 사용자들이 사용하는 텍스트 주소와 비교하여, 사회적 관계 서클 내의 사용자들이 사용하는 텍스트 주소는 어느 정도는 연관성이 있으며, 이는 연관성을 갖는 주소 정보 사이로 텍스트 주소의 정규화를 고정하는 것에 해당한다. 이 방법으로, 텍스트 주소 처리 장치는 텍스트 주소 사이의 내결함성 경계를 더 쉽게 제어할 수 있으며, 이는 텍스트 주소의 정규화 결과의 정확성을 개선하는 데 도움이 될 수 있다.

[0031] 선택적 구현에서, 각 주소 세트에 대하여, 텍스트 주소 처리 장치가 주소 세트 내의 원본 텍스트 주소에 대해 정규화 처리를 수행하여 주소 세트에 대응하는 목표 텍스트 주소를 획득하는 프로세스는:

[0032] 텍스트 주소 처리 장치에 의해, 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라, 상기 각 두 개의 원본 텍스트 주소 사이의 유사도 계산; 및 상기 각 두 개의 원본 텍스트 주소 사이의 유사도에 따라, 상기 각

두 개의 원본 텍스트 주소가 상기 각 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지 결정하여 주소 세트에 대응하는 목표 텍스트 주소 획득을 포함할 수 있다.

[0033] 주소 세트에 대응하는 하나 이상의 목표 텍스트 주소가 있음을 유의하여야 한다.

[0034] 특히, 하나의 주소 세트에 대하여, 텍스트 주소 처리 장치는 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징을 추출하여, 상기 각 두 개의 원본 텍스트 주소의 특징을 얻을 수 있다. 텍스트 주소 처리 장치는 이어서, 상기 각 두 개의 원본 텍스트 주소의 추출된 특징에 따라, 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산한다. 텍스트 주소 처리 장치는, 상기 각 두 개의 원본 텍스트 주소 사이의 유사도에 따라, 상기 각 두 개의 원본 텍스트 주소가 상기 각 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지 더 결정한다.

[0035] 선택적으로, 이 실시예에서 채택되는 원본 텍스트 주소의 특징은, 표준 단편 특징, 경도 및 위도 특징, 영숫자 특징 중 적어도 하나를 포함할 수 있다.

[0036] 위에 기반하여, 하나의 주소 세트에 대하여, 텍스트 주소 처리 장치는 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징을 추출하여, 상기 각 두 개의 원본 텍스트 주소의 표준 단편 특징, 경도 및 위도 특징, 및 영숫자 특징 중 적어도 하나를 얻는다. 적어도 하나의 특징 중 각 특징에 대하여, 텍스트 주소 처리 장치는, 특징에 따라, 특징에 대응하는 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산한다. 텍스트 주소 처리 장치는, 특징에 대응하는 상기 각 두 개의 원본 텍스트 주소 사이의 유사도에 따라, 상기 각 두 개의 원본 텍스트 주소가 상기 각 두 개의 원본 텍스트 주소 중 하나로 정규화되어야 하는지 더 결정한다.

[0037] 표준 단편 특징은 구체적으로 원본 텍스트 주소 내에 포함된 표준 주소 단편을 반영할 수 있다. 예를 들면, 원본 텍스트 주소에 대해 구조 해석(structure parsing)이 수행되어, 원본 텍스트 주소 내에 포함된 표준 단편을 획득할 수 있다. 이 실시예에서, 텍스트 주소는 24개의 표준 주소 단편으로 미리 분할될 수 있다. 예를 들면, 원본 텍스트 주소에 대해 구조 분석이 수행되어, 원본 텍스트 주소가 24개의 표준 단편 중 어떤 단편을 포함하는지 얻을 수 있다. 24개의 표준 단편은, 예를 들면, 도, 시, 구, 개발 구역, 도로 등과 같은 단편 정보를 포함할 수 있다.

[0038] 경도 및 위도 특징은 구체적으로 원본 텍스트 주소에 의해 설명된 주소 정보의 경도 및 위도 정보를 반영할 수 있다. 예를 들면, 원본 텍스트 주소의 경도 및 위도 특징은 오토내비(AutoNavi)의 지오코딩(Geocoding) 기술을 사용하여 추출될 수 있다. 지오코딩 기술은 지리적 정보 시스템(GIS)에서 사용될 수 있는 지리 좌표로 텍스트 주소를 변환하는 방식을 제공하는 공간 위치 지정 기술을 기반으로 한 인코딩 방법이다. 상세한 설명을 위해서는 종래 기술을 참조할 수 있다.

[0039] 영숫자 특징은 구체적으로 원본 텍스트 주소 내에 포함된 영문자 및/또는 숫자를 반영할 수 있다. 영숫자 특징은 원본 텍스트 주소로부터 직접 추출되고 획득될 수 있다.

[0040] 표준 단편 특징, 경도 및 위도 특징, 및 영숫자 특징 중 적어도 하나의 특징의 각 특징에 대하여:

[0041] 특징이 표준 단편 특징이면, 텍스트 주소 처리 장치는 심해시(SimHash) 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 표준 단편 특징을 처리하여, 표준 단편 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득할 수 있다.

[0042] 심해시 알고리즘의 주요 고려사항은 특징 차원 축소로서, 고차원 표준 단편 특징을 저차원 표준 단편 특징으로 매핑하고, 이어서 두 개의 저차원 표준 단편 특징 사이의 해밍 거리를 비교함으로써, 두 개의 저차원 표준 단편 특징에 의해 식별되는 두 개의 텍스트 주소가 반복되거나 또는 매우 유사한지 결정한다. 두 코드 단어 내의 상이한 대응하는 비트 값을 갖는 비트 수가 두 코드 단어 사이의 해밍 거리로 지칭된다. 유효 코드 세트에서, 임의의 두 코드 단어 사이의 해밍 거리의 최소값이 코드 세트의 해밍 거리로 지칭된다. 예를 들면, 코드 단어 10101과 코드 단어 00110에 대해, 첫번째 비트로부터 시작해서, 첫번째 비트, 네번째 비트 및 다섯번째 비트가 순차적으로 다르고, 해밍 거리는 3이다.

[0043] 상술한 특징이 경도 및 위도 특징이면, 텍스트 주소 처리 장치는 경도 및 위도 거리 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 경도 및 위도 특징을 처리하여, 경도 및 위도 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득할 수 있다.

[0044] 구체적으로, 텍스트 주소 처리 장치는, 두 원본 텍스트 주소의 경도 및 위도 특징에 따라, 두 원본 텍스트 주소에 의해 설명되는 주소 정보 사이의 거리를 계산할 수 있다. 텍스트 주소 처리 장치는 이어서, 거리에 따라 경

도 및 위도 특징 차원에서 두 원본 텍스트 주소 사이의 유사도를 결정할 수 있다.

[0045] 실제 응용에서, 일부 사용자에 의해 설명된 원본 텍스트 주소는 지도 상의 지점에 대해 정확할 수 있고, 일부 사용자에 의해 설명된 원본 텍스트 주소는 지도 상의 선에 대해서만 정확할 수 있으며, 일부 사용자에 의해 설명된 원본 텍스트 주소는 지도 상의 평면에 대해서만 정확할 수도 있다. 세분성(granularities)이 동일하지 않기 때문에, 원본 텍스트 주소가 지도의 관점에서 정규화되면, 가장 거친(coarsest) 세분성으로부터만 정규화가 수행될 수 있으므로 정규화 결과의 정확성이 불충분하게 된다. 그러나, 이 실시예에서, 원본 텍스트 주소는 경도 및 위도로 매핑될 수 있다. 모든 텍스트 주소가 경도 및 위도로 매핑될 수 있고 경도 및 위도의 세분성이 비교적 미세하기 때문에, 이는 정규화 결과의 정확도를 향상시킬 수 있는 비교적 미세한 세분화로 정규화 처리를 통합하는 데 도움이 된다.

[0046] 상술한 특징이 영숫자 특징이면, 텍스트 주소 처리 장치는 자카드(Jaccard) 계수 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 영숫자 특징을 처리하여, 영숫자 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득할 수 있다.

[0047] 자카드 계수는 표본 집합 내의 유사성과 다양성을 비교하는 데 주로 사용되는 확률이다. 자카드 계수는 표본 집합의 교집합과 표본 집합의 합집합의 비와 같다, 즉 J = |A∩B| / |A∪B|이다. 위의 원본 텍스트 주소 중 하나의 영숫자 특징을 표본 집합으로 사용할 수 있으며, 영숫자 특징의 영문자 및/또는 숫자가 표본 집합의 요소로 사용될 수 있다.

[0048] 표준 단편 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도, 경도 및 위도 특징 차원에서 그 유사도, 및 영숫자 특징 차원에서 그 유사도에 기반하여, 텍스트 주소 처리 장치는 구체적으로, 동시에 표준 단편 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도, 경도 및 위도 특징 차원에서 그 유사도, 및 영숫자 특징 차원에서 그 유사도에 따라, 두 원본 텍스트 주소가 그 중 하나의 원본 텍스트 주소로 정규화될 수 있는지 결정할 수 있다.

[0049] 예를 들면, 각 차원에서의 두 원본 텍스트 주소 사이의 유사도가 대응하는 임계치와 각각 비교될 수 있다. 각 차원에서의 두 원본 텍스트 주소 사이의 유사도가 대응하는 임계치보다 크면, 두 원본 텍스트 주소가 그 중 하나의 텍스트 주소로 정규화될 수 있다고 결정할 수 있다. 그렇지 않으면, 다른 상황에서, 두 원본 텍스트 주소가 그 중 하나의 텍스트 주소로 정규화될 수 없다고 결정할 수 있다.

[0050] 다른 예를 들면, 한 차원에서의 두 원본 텍스트 주소 사이의 유사도가 대응하는 임계치와 우선적으로 비교될 수 있다. 그 차원에서의 두 원본 텍스트 주소 사이의 유사도가 대응하는 임계치보다 크면, 두 원본 텍스트 주소가 그 중 하나의 텍스트 주소로 정규화될 수 있다고 바로 결정할 수 있다.

[0051] 다른 예를 들면, 각 차원의 유사도에 대해 미리 가중치가 구성될 수 있다. 각 차원에서의 두 원본 텍스트 주소 사이의 유사도 및 대응하는 가중치에 대해 수치 처리가 수행되어, 처리 결과를 얻을 수 있다. 처리 결과는 사전 설정된 임계치와 비교될 수 있다. 처리 결과가 임계치보다 크면, 두 원본 텍스트 주소가 그 중 하나의 원본 텍스트 주소로 정규화될 수 있다고 결정할 수 있다. 그렇지 않으면, 두 원본 텍스트 주소가 그 중 하나의 원본 텍스트 주소로 정규화될 수 없다고 결정할 수 있다.

[0052] 또한, 선택적 구현에서, 한 사용자가 서비스 시스템 내의 복수의 사용자와 동시에 사회적 관계를 갖기 쉽고 복수의 사용자의 사회적 관계 서클에 나타난다는 점을 고려하면, 이는 그 사용자에 의해 사용되는 원본 텍스트 주소가 상이한 주소 세트에 나타날 수 있다는 것을 의미한다. 이러한 상황을 위해, 각 주소 세트에 대응하는 목표 텍스트 주소 획득 후에, 주소 세트 사이의 정규화가 더 수행되어, 더 정확하고 단순화된 정규화 결과를 획득할 수 있다.

[0053] 각 주소 세트의 정규화 프로세스에서, 텍스트 주소 처리 장치가 목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계를 기록할 수 있음을 유의하여야 한다. 대응 관계는 어느 원본 텍스트 주소로부터 정규화를 통해 목표 텍스트 주소가 구체적으로 획득되었는지 나타낼 수 있다.

[0054] 위에 기반하여, 각 주소 세트에 대응하는 목표 텍스트 주소 획득 이후에, 텍스트 주소 처리 장치는, 정규화 처리 동안 형성된 목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계에 따라, 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 더 결정할 수 있다. 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소는 각각 다른 주소 세트에 대응한다. 텍스트 주소 처리 장치는 이어서 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 대해 정규화 처리를 수행할 수 있다.

[0055] 선택적 구현에서, 텍스트 주소 처리 장치는 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편을 획득할 수 있다. 텍스트 주소 처리 장치는 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편의 단편 교집합을 더 획득할 수 있다. 단편 교집합은 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 모든 원본 텍스트 주소에 포함되는 표준 주소 단편을 포함한다. 텍스트 주소 처리 장치는 이어서, 단편 교집합에 따라, 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 대해 정규화 처리를 수행할 수 있다.

[0056] 구체적인 정규화 처리 방식은, 텍스트 주소 처리 장치에 의해, 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 있는지 판단하는 것을 포함할 수 있다. 판단 결과가 예이면, 즉, 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 있으면, 방식은 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 단편 교집합이 나타낼 수 있는 목표 텍스트 주소로 정규화하는 것을 더 포함한다. 반대로, 판단 결과가 아니오이면, 즉, 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 없으면, 방식은 정규화 처리 수행을 포함하지 않는다.

[0057] 구체적으로, 목표 텍스트 주소를 나타내기 위해 필요한 단편 세트는 사전 설정될 수 있다. 단편 교집합은 사전 설정된 단편 세트와 비교될 수 있다. 단편 교집합이 사전 설정된 단편 세트와 일치하면, 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 있다고 판단될 수 있다. 그렇지 않으면, 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 어느 것도 나타낼 수 없다고 판단될 수 있다.

[0058] 또한, 위의 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 있으면, 단편 교집합과 단편 교집합이 나타낼 수 있는 목표 텍스트 주소가 특징 지식 베이스 내로 대응하여 저장될 수 있다. 이 방법으로, 이 특징 지식 베이스가 더 많은 원본 텍스트 주소를 정규화하는 데 사용될 수 있다.

[0059] 이하에서, 제1 사용자 및 제2 사용자를 포함하는 서비스 시스템을 예로 든다. 제1 사용자의 사회적 관계 서클은 사용자 A, 사용자 B, 및 사용자 C를 포함하고, 제2 사용자의 사회적 관계 서클은 사용자 D, 사용자 E, 및 사용자 F를 포함한다고 가정한다. 제1 사용자가 사용한 텍스트 주소 및 제1 사용자의 사회적 관계 서클의 사용자들이 사용한 텍스트 주소가 제1 주소 세트를 구성한다. 제1 주소 세트에는 텍스트 주소 X1, X2, 및 X3이 포함된다고 가정한다. 제1 사용자, 사용자 A, 사용자 B, 및 사용자 C와 텍스트 주소 X1, X2, 및 X3 사이에는 고정된 대응 관계가 없다. 한 사용자가 하나의 텍스트 주소를 사용하는 것이 가능하고, 복수의 사용자가 동일한 텍스트 주소를 사용하는 것 또한 가능하며, 하나의 사용자가 복수의 텍스트 주소를 사용하는 것도 가능하다. 예를 들면, 제1 사용자가 텍스트 주소 X1을 사용하였고, 사용자 A 및 사용자 B가 텍스트 주소 X2를 사용하였으며, 사용자 C가 텍스트 주소 X1 및 X3을 사용하였다.

[0060] 제2 사용자가 사용한 텍스트 주소 및 제2 사용자의 사회적 관계 서클의 사용자들이 사용한 텍스트 주소가 제2 주소 세트를 구성한다. 제2 주소 세트에는 텍스트 주소 X2, X4, 및 X5가 포함된다고 가정한다. 유사하게, 제2 사용자, 사용자 D, 사용자 E, 및 사용자 F와 텍스트 주소 X2, X4, 및 X5 사이에는 고정된 대응 관계가 없다. 한 사용자가 하나의 텍스트 주소를 사용하는 것이 가능하고, 복수의 사용자가 동일한 텍스트 주소를 사용하는 것 또한 가능하며, 하나의 사용자가 복수의 텍스트 주소를 사용하는 것도 가능하다. 예를 들면, 제2 사용자가 텍스트 주소 X2를 사용하였고, 사용자 D가 텍스트 주소 X2, X4, 및 X5를 사용하였으며, 사용자 E 가 텍스트 주소 X2 및 X5를 사용하였고, 사용자 F가 텍스트 주소 X5를 사용하였다.

[0061] 위에 기반하여, 전체 정규화 프로세스는 구체적으로 도 2에 나타난 바와 같을 수 있다.

[0062] 먼저, 제1 사용자의 사회적 관계 서클이 결정되고, 제1 주소 세트가 획득된다. 도 2에 나타난 바와 같이, 제1 주소 세트는 텍스트 주소 X1, X2, 및 X3을 포함한다. 제2 사용자의 사회적 관계 서클이 결정되어, 제2 주소 세트가 획득된다. 도 2에 나타난 바와 같이, 제2 주소 세트는 텍스트 주소 X2, X4, 및 X5를 포함한다.

[0063] 제1 주소 세트 내의 텍스트 주소 사이의 유사도가 쌍 단위로 계산되고, 유사도에 따라 정규화 처리가 수행된다. 텍스트 주소 X1 및 X2는 X1과 X2 중 하나로 정규화된다. 여기에서, 텍스트 주소 X1 및 X2는 X1로 정규화되고 텍스트 주소 X3은 텍스트 주소 X3으로 정규화된다고 가정한다. 즉 말하자면, 도 2에 나타난 바와 같이, 제1 주소 세트에 대응하는 두 목표 텍스트 주소가 각각 텍스트 주소 X1 및 X3이다. 유사하게, 제2 주소 세트 내의 텍스트

주소 사이의 유사도가 쌍 단위로 계산되고, 유사도에 따라 정규화 처리가 수행된다. 텍스트 주소 X2 및 X4는 X2와 X4 중 하나로 정규화된다. 여기에서 텍스트 주소 X2 및 X4는 X4로 정규화되고 텍스트 주소 X5는 텍스트 주소 X5로 정규화된다고 가정한다. 즉 말하자면, 도 2에 나타난 바와 같이, 제2 주소 세트에 대응하는 두 목표 텍스트 주소는 각각 텍스트 주소 X4 및 X5이다.

[0064] 또한, 제1 주소 세트에 대응하는 목표 텍스트 주소 X1과 제2 주소 세트에 대응하는 목표 텍스트 주소 X4가 모두 텍스트 주소 X2의 정규화를 통해 획득되므로, 정규화 처리는 두 목표 텍스트 주소에 대해 수행될 수 있다. 두 목표 텍스트 주소는 X1 및 X4 중 하나로 더 정규화될 수 있다. 여기에서, 도 2에 나타난 바와 같이 두 목표 텍스트 주소가 X1로 정규화되는 것으로 가정한다. 지금까지, 원본 텍스트 주소 X1, X2, X3, X4, 및 X5는 텍스트 주소 X1, X3, 및 X5로 정규화된다.

[0065] 위의 과정에서, 텍스트 주소 X1 및 X2가 X2로 정규화되고 텍스트 주소 X2 및 X4가 또한 X2로 정규화되면, 텍스트 주소 X1, X2, 및 X4가 동일한 텍스트 주소로 정규화되었으므로, 위의 두 주소 세트에 대응하는 목표 텍스트 주소의 정규화 처리에서, 두 개의 동일한 목표 텍스트 주소에 대해 정규화 처리가 수행되지 않을 수 있다는 점을 주목하여야 한다. 이는 정규화에 의해 소비되는 자원을 절약하고 정규화 처리의 효율성을 개선하는 데 도움이 된다.

[0066] 텍스트 주소에 대해 정규화 처리를 수행함으로써, 텍스트 주소의 수가 간소화되고, 텍스트 주소가 통합되어, 텍스트 주소의 관리 및 응용을 더 용이하게 한다는 것을 위로부터 알 수 있다. 또한, 정규화 대상 원본 텍스트 주소는 사용자의 사회적 관계 서클에 따라 분할된다. 한편으로, 정규화 대상 원본 텍스트 주소의 범위가 각 사용자의 사회적 관계 서클로 제한되어, 정규화 대상 원본 텍스트 주소의 범위를 줄이는 것에 해당한다. 다른 한편으로, 사회적 관계 서클 외부의 사용자가 사용하는 텍스트 주소와 비교하여, 사회적 관계 서클 내의 사용자가 사용하는 텍스트 주소는 어느 정도의 연관을 가지므로, 텍스트 주소의 정규화를 연관성이 있는 텍스트 주소 사이로 고정하는 것에 해당한다. 이 방법으로, 텍스트 주소 사이의 내결함성 경계가 더 용이하게 제어될 수 있으며, 텍스트 주소의 정규화 결과의 정확성을 개선하는 데 도움이 된다.

[0067] 간단한 설명을 용이하게 하기 위하여, 전술한 방법 실시예는 모두 일련의 동작 조합으로 표현된다는 점에 유의해야 한다. 그러나, 본 출원에 따르면, 일부 단계가 다른 순서로 또는 동시에 수행될 수 있기 때문에, 본 출원이 설명된 동작 순서에 의해 제한되지 않음을 이 분야의 기술자는 이해해야 한다. 둘째로, 이 분야의 기술자는 명세서에 기재된 실시예는 모두 바람직한 실시예이고, 관련된 동작 및 모듈이 본 출원에서 반드시 요구되는 것은 아니라는 것을 또한 이해해야 한다.

[0068] 위의 실시예에서, 각 실시예의 설명은 각각 그 자체의 초점을 갖고 있다. 어떤 실시예에서 상세히 설명되지 않은 내용에 대해서, 다른 실시예의 연관된 설명에 대한 참조가 이루어질 수 있다.

[0069] 도 3은 본 출원의 실시예에 따른 텍스트 주소 처리 장치의 개략적 구조도이다. 도 3에 나타난 바와 같이, 장치는 결정 모듈(31) 및 정규화 모듈(32)을 포함할 수 있다.

[0070] 결정 모듈(31)은 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트를 결정하도록 구성될 수 있으며, 적어도 하나의 주소 세트 내의 각 주소 세트는 적어도 두 개의 원본 텍스트 주소를 포함한다.

[0071] 정규화 모듈(32)은 결정 모듈(31)에 의해 결정된 각 주소 세트에 대하여, 주소 세트 내의 원본 텍스트 주소에 대해 정규화 처리를 수행하여, 주소 세트에 대응하는 목표 텍스트 주소를 획득하도록 구성될 수 있다.

[0072] 주소 세트에 대응하는 목표 텍스트 주소는 하나 이상일 수 있음을 유의해야 한다.

[0073] 선택적인 구현에서, 결정 모듈(31)은 구체적으로:

[0074] 서비스 시스템 내의 각 사용자의 사회적 관계 서클을 결정하고;

[0075] 각 사용자에 의해 사용되는 텍스트 주소 및 각 사용자의 사회적 관계 서클 내의 사용자에 의해 사용되는 텍스트 주소를 획득하여 주소 세트를 구성하도록 구성될 수 있다.

[0076] 선택적인 구현에서, 정규화 모듈(32)은 구체적으로:

[0077] 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라, 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산하고;

[0078] 상기 각 두 개의 원본 텍스트 주소 사이의 유사도에 따라, 상기 각 두 개의 원본 텍스트 주소가 상기 각 두 개의 원본 텍스트 주소 중 하나로 정규화될 수 있는지 결정하여 주소 세트에 대응하는 목표 텍스트 주소를 획득하도록 구성될 수 있다.

[0079] 정규화 모듈(32)은, 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징에 따라, 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산할 때, 더 구체적으로:

[0080] 주소 세트 내의 각 두 개의 원본 텍스트 주소의 특징을 추출하여, 상기 각 두 개의 원본 텍스트 주소의 표준 단편 특징, 경도 및 위도 특징, 및 영숫자 특징 중 적어도 하나의 특징을 획득하고;

[0081] 적어도 하나의 특징의 각 특징에 대하여, 특징에 따라, 특징에 대응하는 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산하도록 구성될 수 있다.

[0082] 정규화 모듈(32)은, 적어도 하나의 특징의 각 특징에 대하여, 특징에 따라, 특징에 대응하는 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 계산할 때, 더 구체적으로:

[0083] 특징이 표준 단편 특징이면, 심해시 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 표준 단편 특징을 처리하여, 표준 단편 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득하고;

[0084] 특징이 경도 및 위도 특징이면, 경도 및 위도 거리 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 경도 및 위도 특징을 처리하여, 경도 및 위도 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득하고;

[0085] 특징이 영숫자 특징이면, 자카드 계수 알고리즘을 사용하여 상기 각 두 개의 원본 텍스트 주소의 영숫자 특징을 처리하여, 영숫자 특징 차원에서 상기 각 두 개의 원본 텍스트 주소 사이의 유사도를 획득하도록 구성될 수 있다.

[0086] 선택적인 구현에서, 결정 모듈(31)은 정규화 모듈(32)이 각 주소 세트에 대응하는 목표 텍스트 주소를 획득한 후에, 정규화 처리 동안 형성된 목표 텍스트 주소와 원본 텍스트 주소 사이의 대응 관계에 따라, 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소를 결정하도록 더 구성될 수 있고;

[0087] 정규화 모듈(32)은 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 대해 정규화 처리를 수행하도록 더 구성될 수 있다.

[0088] 정규화 모듈(32)은, 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 대해 정규화 처리를 수행할 때, 구체적으로:

[0089] 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 각각 대응하는 원본 텍스트 주소에 포함된 표준 주소 단편의 단편 교집합을 획득하고;

[0090] 단편 교집합에 따라, 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소에 대해 정규화 처리를 수행하도록 구성될 수 있다.

[0091] 선택적인 구현에서, 이 실시예의 텍스트 주소 처리 장치는 단편 교집합이 동일한 원본 텍스트 주소에 대응하는 적어도 두 개의 목표 텍스트 주소 중 하나를 나타낼 수 있을 때, 단편 교집합 및 단편 교집합이 나타낼 수 있는 목표 텍스트 주소를 대응하여 저장하도록 구성되는 특징 지식 베이스를 더 포함할 수 있다.

[0092] 이 실시예에 따른 텍스트 주소 처리 장치는, 서비스 시스템 내 사용자의 사회적 관계 서클에 따라, 적어도 하나의 주소 세트를 결정할 수 있다. 텍스트 주소 처리 장치는 이어서 주소 세트를 단위로 하여 각 주소 세트 내의 원본 텍스트 주소에 대해 각각 정규화 처리를 수행하여 각 주소 세트에 대응하는 목표 텍스트 주소를 획득하고, 따라서 텍스트 주소에 대한 정규화 처리를 달성한다. 이 실시예에 따른 텍스트 주소 처리 장치는 사용자의 사회적 관계 서클에 따라 정규화 대상 원본 텍스트 주소를 분할한다. 한편으로, 정규화 대상 원본 텍스트 주소의 범위가 사용자의 사회적 관계 서클로 제한되어, 정규화 대상 원본 텍스트 주소의 범위를 줄이는 것에 해당한다. 다른 한편으로, 사회적 관계 서클 너머의 사용자가 사용하는 텍스트 주소와 비교하여, 사회적 관계 서클 내의 사용자가 사용하는 텍스트 주소는 어느 정도의 연관을 가지므로, 텍스트 주소의 정규화를 연관성이 있는 텍스트 주소 사이로 고정하는 것에 해당한다. 그러므로, 텍스트 주소 사이의 내결함성 경계가 더 용이하게 제어될 수 있으며, 텍스트 주소의 정규화 결과의 정확성을 개선하는 데 도움이 된다.

[0093] 이 분야의 기술자라면, 설명을 편리하고 간결하게 하기 위하여, 상술한 시스템, 장치 및 유닛의 특정 작업 프로

세스에 대한 전술한 방법 실시예에서의 대응하는 프로세스를 참조할 수 있음을 명확하게 이해할 수 있으며, 이들은 여기에서 반복되지 않는다.

[0094] 본 출원에 제공된 몇몇 실시예에서, 개시된 시스템, 장치 및 방법은 다른 방식으로 구현될 수 있음을 이해하여야 한다. 예를 들면, 설명된 장치 실시예는 단지 예시적인 것이다. 예를 들면, 유닛의 분할은 단지 논리적 기능의 분할일 뿐이며 실제 구현에서 다른 방식의 분할이 존재할 수 있다. 예를 들면, 복수의 유닛 또는 구성요소가 결합되거나 다른 시스템에 통합되거나, 일부 특징이 무시되거나 수행되지 않을 수 있다. 또한, 설명되거나 논의된 상호 결합 또는 직접 결합 또는 통신 접속은 몇몇 인터페이스, 장치 또는 유닛을 통한 간접 결합 또는 통신 접속일 수 있으며, 전기적으로, 기계적으로 또는 다른 형태로 구현될 수 있다.

[0095] 분리된 부품으로 기술된 유닛들은 물리적으로 분리되거나 그렇지 않을 수 있고, 유닛으로 설명되는 부품들은 물리적 유닛이거나 그렇지 않을 수 있으며, 이는 하나의 지점에 위치하거나 또는 복수의 네트워크 유닛 상에 분산될 수 있다. 유닛의 일부 또는 전부는 실시예의 해결책의 목적을 달성하기 위해 실제 요구에 따라 선택될 수 있다.

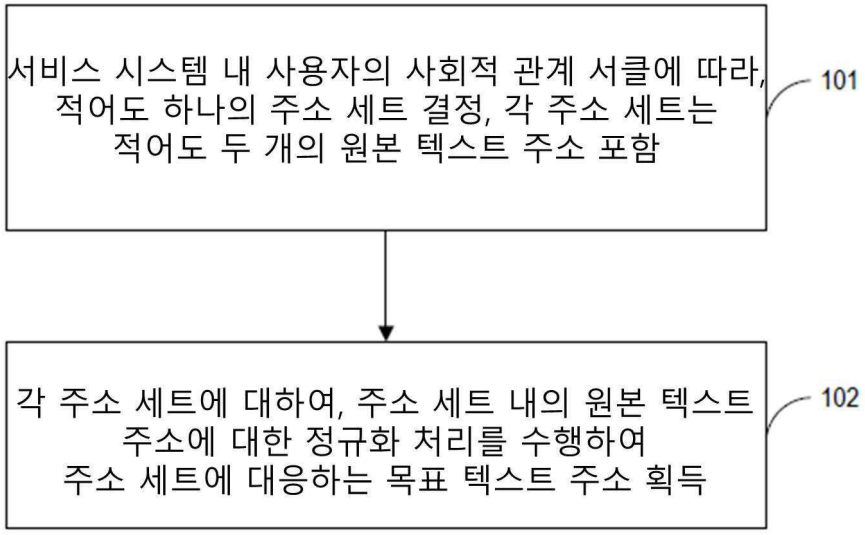
[0096] 또한, 본 출원의 실시예에서의 기능 유닛은 하나의 처리 유닛에 통합되거나, 또는 각 유닛이 물리적으로 단독으로 존재할 수도 있고, 또는 둘 이상의 유닛이 하나의 유닛으로 통합될 수도 있다. 통합 유닛은 하드웨어의 형태로 구현될 수 있거나 하드웨어 플러스 소프트웨어 기능 유닛의 형태로 구현될 수 있다.

[0097] 소프트웨어 기능 유닛의 형태로 구현된 통합 유닛은 컴퓨터 판독가능 저장 매체에 저장될 수 있다. 소프트웨어 기능 유닛은 저장 매체에 저장될 수 있고, 컴퓨터 장치(개인용 컴퓨터, 서버, 네트워크 장치 등일 수 있음) 또는 프로세서가 본 출원의 실시예에서 설명된 방법의 단계 중 일부를 수행하도록 지시하는 명령을 포함한다. 전술한 저장 매체는 USB 플래시 드라이브, 착탈식 하드 디스크, 읽기전용 메모리(ROM), 임의접근 메모리(RAM), 자기 디스크, 또는 광 디스크와 같은 프로그램 코드를 저장할 수 있는 임의의 매체를 포함할 수 있다.

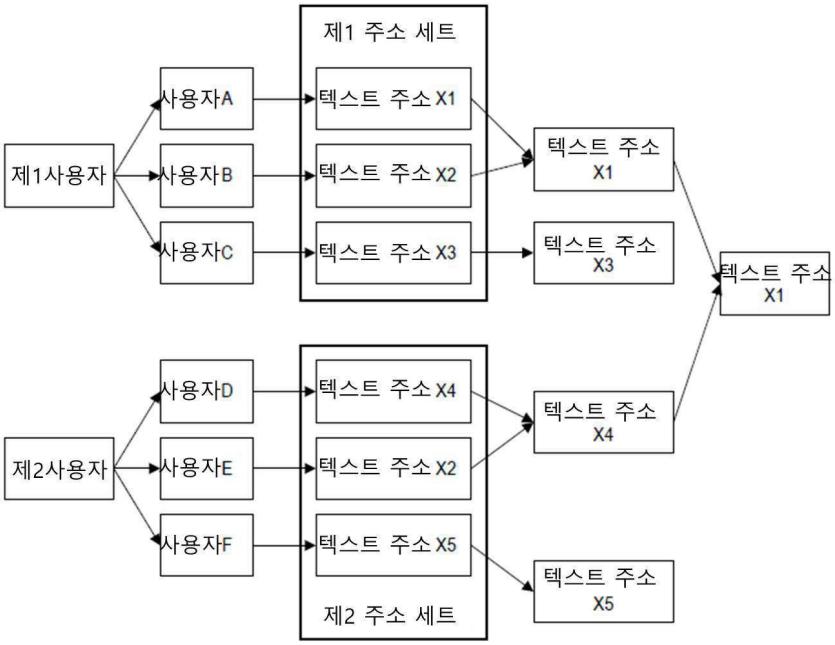
[0098] 마지막으로, 위의 실시예는 단지 본 출원의 기술적 해결책을 설명하기 위해 제공되는 것이며 본 출원을 제한하고자 의도하지 않는다는 점을 유의해야 한다. 본 출원은 전술한 실시예를 참조하여 상세하게 설명되었지만, 전술한 실시예에서 설명된 기술적 해결책에 여전히 변형이 가해질 수 있거나 또는 동등한 대체가 이루어질 수 있음을 이 분야의 기술자는 이해해야 한다. 이러한 변형 또는 대체는 대응하는 기술적 해결책의 본질이 본 출원의 실시예의 기술적 해결책의 사상 및 범위를 벗어나도록 하지 않는다.

도면

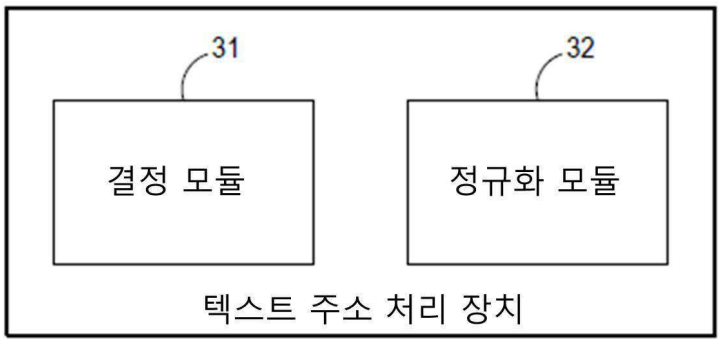
도면1



도면2


제1 주소 세트
사용자A
사용자B
사용자C
제1사용자
텍스트 주소 X1
텍스트 주소 X2
텍스트 주소 X3
텍스트 주소 X1
텍스트 주소 X3
텍스트 주소 X1
사용자D
사용자E
사용자F
제2사용자
텍스트 주소 X4
텍스트 주소 X2
텍스트 주소 X5
텍스트 주소 X4
텍스트 주소 X5
제2 주소 세트

도면3


31
32
결정 모듈
정규화 모듈
텍스트 주소 처리 장치