(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년11월17일
(11) 등록번호 10-2328234
(24) 등록일자 2021년11월15일

(51) 국제특허분류(Int. Cl.)
G06Q 50/10 (2012.01) *G06F 16/35* (2019.01)
G06F 16/387 (2019.01) *G06F 16/906* (2019.01)
G06F 40/268 (2020.01) *G06Q 50/00* (2018.01)
G06Q 50/30 (2012.01)
(52) CPC특허분류
G06Q 50/10 (2015.01)
G06F 16/358 (2019.01)
(21) 출원번호 10-2020-0033367
(22) 출원일자 2020년03월18일
심사청구일자 2020년03월18일
(65) 공개번호 10-2021-0117038
(43) 공개일자 2021년09월28일
(56) 선행기술조사문헌
KR1020110022627 A*
KR102086248 B1*
KR101122436 B1
KR101536520 B1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
충북대학교 산학협력단
충청북도 청주시 서원구 충대로 1 (개신동)
(72) 발명자
박수빈
충청북도 청주시 서원구 충대로14번길 65, 207호 (행복나무빌라)
최도진
경상북도 상주시 청리면 수선로 95-4
(뒷면에 계속)
(74) 대리인
김정현

전체 청구항 수 : 총 6 항 심사관 : 장우진

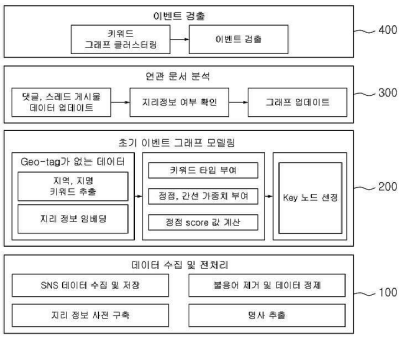
(54) 발명의 명칭 **소셜 네트워크에서 연관 문서 분석을 통한 지역 이벤트 검출 시스템 및 방법**

(57) 요 약

본 발명은 소셜 네트워크에서 지역 이벤트 검출 시스템에 관한 것으로서, 소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행하는 데이터 수집 모듈, 상기 데이터 수집 모듈에서 정제된 데이터를 이용하여 초기 그래프를 모델링하는 초기 그래프 모델링 모듈, 타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트하는 연관 데이터 반영 모듈 및 업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출하는 이벤트 검출 모듈을 포함한다.

본 발명에 의하면, 소셜 네트워크에서 연관 문서 분석을 통해 지역 이벤트를 검출함으로써, 보다 정확하고 빠르게 지역 이벤트를 검출할 수 있는 효과가 있다.

대 표 도



(52) CPC특허분류
G06F 16/387 (2019.01)
G06F 16/906 (2019.01)
G06F 40/268 (2020.01)
G06Q 50/01 (2013.01)
G06Q 50/30 (2015.01)

(72) 발명자
임종태
충청북도 청주시 서원구 창직로31번길 5, 203호(스위트빌)
복경수
세종특별자치시 시청대로 236, 302동 602호(새샘마을3단지)
유재수
충청북도 청주시 서원구 월평로 24, 805동 1001호(현대대우아파트)

이 발명을 지원한 국가연구개발사업
과제고유번호 2019R1A2C2084257
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견연구
연구과제명 실시간 그래프 스트림 데이터에 대한 분산 인메모리 기반 처리 및 분석
기 여 율 25/100
과제수행기관명 충북대학교
연구기간 2019.09.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 B0101-15-0266
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기술진흥센터
연구사업명 정보통신·방송 연구개발사업
연구과제명 실시간 대규모 영상 데이터 이해?예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발
기 여 율 25/100
과제수행기관명 한국전자통신연구원
연구기간 2014.04.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업
과제고유번호 NRF-2017M3C4A706943
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 차세대정보컴퓨팅기술개발사업
연구과제명 인텔리전트 DB를 위한 고성능 자율 기계학습 플랫폼
기 여 율 25/100
과제수행기관명 충북대학교
연구기간 2017.08.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 P0010202
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술진흥원
연구사업명 R&D재발견프로젝트
연구과제명 소셜 빅데이터 기반의 개인맞춤형 취업 콘텐츠 추천(큐레이션) 및 디지털 증명서 발급 시스템
기 여 율 25/100
과제수행기관명 (주)다인리더스
연구기간 2019.06.01 ~ 2020.05.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행하는 데이터 수집 모듈;

상기 데이터 수집 모듈에서 정제된 데이터를 이용하여 초기 그래프를 모델링하는 초기 그래프 모델링 모듈;

타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트하는 연관 데이터 반영 모듈; 및

업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출하는 이벤트 검출 모듈을 포함하고,

상기 초기 그래프 모델링 모듈은,

지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고,

지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고,

지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행하고,

상기 초기 그래프 모델링 모듈은,

각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고,

각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성하고,

상기 초기 그래프 모델링 모듈은, 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정하고,

상기 연관 데이터 반영 모듈은,

초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가 발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고,

초기 그래프에서 지역 키워드와 일치하는 키 노드가 있으면 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고,

초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 상기 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성하고,

상기 이벤트 검출 모듈은,

키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고,

생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생

성하고, 연관이 있는 클러스터들은 병합하고,

이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출하고,

상기 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정하며,

m은 전체 간선 수, n은 전체 노드 수, w_{ij} 는 정점 V_i 와 V_j 사이를 연결하는 간선의 가중치, k_i 는 V_i 와 연결된 모든 간선의 가중치 합이고, $\delta(c_i, c_j)$ 는 V_i 와 V_j 가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수라고 할 때,

네트워크 모듈성 NM을,

$$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (3)$$

의 수학식으로 나타낼 수 있는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 시스템.

청구항 2

청구항 1에 있어서,

상기 데이터 수집 모듈은, 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행하는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 시스템.

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

청구항 9

청구항 1에 있어서,

상기 이벤트 검출 모듈은,

하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때,

두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의

클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합하는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 시스템.

청구항 10

소셜 네트워크에서 지역 이벤트 검출 시스템에서의 지역 이벤트 검출 방법에서,

소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행하는 데이터 수집 단계;

상기 데이터 수집 단계에서 정제된 데이터를 이용하여 초기 그래프를 모델링하는 초기 그래프 모델링 단계;

타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트하는 연관 데이터 반영 단계; 및

업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출하는 이벤트 검출 단계를 포함하고,

상기 초기 그래프 모델링 단계에서,

지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고,

지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고,

지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행하고,

상기 초기 그래프 모델링 단계에서,

각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고,

각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성하고,

상기 초기 그래프 모델링 단계에서, 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정하고,

상기 연관 데이터 반영 단계에서,

초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가 발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고,

초기 그래프에서 지역 키워드와 일치하는 키 노드가 있으면 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고,

초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 상기 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성하고,

상기 이벤트 검출 단계에서,

키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고,

생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생

성하고, 연관이 있는 클러스터들은 병합하고,

이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출하고,

상기 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정하며,

m은 전체 간선 수, n은 전체 노드 수, w_{ij} 는 정점 V_i 와 V_j 사이를 연결하는 간선의 가중치, k_i 는 V_i 와 연결된 모든 간선의 가중치 합이고, $\delta(c_i, c_j)$ 는 V_i 와 V_j 가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수라고 할 때,

네트워크 모듈성 NM을,

$$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (3)$$

의 수학식으로 나타낼 수 있는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 방법.

청구항 11

청구항 10에 있어서,

상기 데이터 수집 단계에서,

수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행하는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 방법.

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

청구항 10에 있어서,

상기 이벤트 검출 단계에서,

하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때,

두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의 클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합하는 것을 특징으로 하는 소셜 네트워크에서 지역 이벤트 검출 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 소셜 네트워크에 관한 것으로서, 더욱 상세하게는 소셜 네트워크에서 연관 문서 분석을 통한 지역 이벤트 검출 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 유무선 인터넷의 발달과 모바일 스마트 기기의 대중화로 사용자 간의 의사소통과 정보를 공유하고 인맥을 관리하는 도구인 소셜 네트워크 서비스(Social Network Service, 이하 SNS) 사용자가 증가하고 있다. 사용자들은 모바일 기기를 이용해 페이스북, 인스타그램, 트위터와 같은 SNS에 접속하여 시간과 공간의 제약 없이 언제든지 자신의 의견이나 의도에 맞는 글을 남길 수 있다.

[0003] 2017년 기준 SNS 사용자는 전 세계적으로 24억 6천만 명이 있으며 사람들은 일반적으로 다른 사람들과 메시지를 주고받고, 사진을 게시하여 공유하며, 게시물에 댓글을 남기거나 공유하기 위해서 SNS를 사용한다. 또한, SNS를 통해 사회적 관계망을 생성하고, 유지하며 이와 같이 온라인에서 이루어지는 사용자들의 상호작용 행위들은 실생활에서의 인간관계와 매우 유사한 수준을 가지고 있다.

[0004] 대표적인 소셜 네트워크 서비스인 트위터는 게시글 단위 트윗을 통해 하루에 약 5억 개가 생성되고 사용자의 약 80%는 모바일 사용자이다. 또한, 트위터는 스마트 모바일 기기 위주의 인터페이스를 가지고 있기 때문에 사용자들은 자신이 휴대한 기기를 이용하여 언제 어디서든 소셜 네트워크 서비스에 접속 할 수 있다. 이러한 특성을 이용하여 SNS 사용자들은 실생활에서 자신이 경험한 일이나 사건들을 다른 사용자들과 온라인상에서 공유함으로써 정보를 제공한다. 그리고 사용자들은 정보를 생성하는 단계에서 나아가 SNS에서 형성된 친구 관계를 바탕으로 '댓글', '좋아요', '리트윗', '공유하기' 등과 같은 정보를 확산하는 역할도 한다. 정보 생성과 확산의 과정을 통해서 트위터의 상위 트렌드 토픽의 약 85%가 주요 뉴스나 이벤트와 연관되어 있다는 연구 결과를 통해 이벤트 발생시 SNS 데이터의 중요성을 알 수 있다.

[0005] SNS는 인맥 관계를 확장하는 역할 뿐만 아니라 지역 이벤트가 발생 했을 때 빠르게 정보를 전달하는 도구로 사용되고 있다. 지역 이벤트는 산불과 같은 자연재해, 정치적인 시위 또는 스포츠 게임과 같은 형태로 나타난다. 소셜 네트워크 사용자들은 특정 시간과 장소에서 이벤트가 발생했을 때 실시간으로 게시글을 업로드하고 현재 상황을 공유한다. 예를 들어, 2012년 허리케인 샌디가 미국 동부를 강타했을 때 당시 송전탑 피해로 인해 전력 공급이 어려워져 각 가구의 발전기를 가동하기 위해 기름 확보를 위한 주유대란이 발생했다. 이 때 사람들은 페이스 북, 트위터 등을 통해 주유소의 기름 보유 상태, 연락처, 대기 시간 등을 공유하는 등 SNS를 실시간으로 빠르게 정보를 확산 할 수 있는 도구로 이용하였다. 이처럼 재난재해와 같은 지역 이벤트 시에 소셜 네트워크 서비스의 특징인 신뢰성 높은 집단지성을 활용 할 수 있다는 장점이 있다.

[0006] 여러 사람들이 공통된 목적을 가지고 같은 시간과 장소에 모인 행사 또는 특정 지역에서 발생한 사회적으로 파급력이 큰 이슈 등을 '지역 이벤트'로 정의할 수 있다. 지역 이벤트는 통상 대규모 시위, 스포츠 게임과 같은 큰 규모의 행사에서부터 지역의 맥주 축제나 동네 슈퍼마켓의 할인 행사까지 주변에서 다양한 크기로 나타날 수 있다. 지역 이벤트를 빠르게 검출하여 해당 이벤트에 대한 정보를 얻음으로써 다양한 곳에 활용할 수 있다. 또한, 자연재해가 발생했을 때 실시간 이벤트 탐지를 통해 사람들에게 대피 알람을 보내 인명피해나 경제적인 피해를 줄일 수 있다. 따라서 SNS 데이터를 활용하여 지역 이벤트를 검출하는 연구가 필요하다.

[0007] 종래 소셜 네트워크 서비스 데이터 분석을 통해서 다양한 방법을 통해 이벤트를 검출하고자 하는 연구가 진행되었다. 이벤트를 검출하는 기존 발명들은 세 가지로 분류 할 수 있다. 텍스트 마이닝을 통한 텍스트 기반의 검출 방법, 그래프 기반 이벤트 검출 방법 그리고 지오 태그(Geo-Tag) 서비스를 사용한 지역 기반 이벤트 검출 방법이 있다.

[0008] 텍스트 기반 이벤트 검출 방법은 주로 TF-IDF 알고리즘을 이용하여 키워드를 추출하는 기법을 사용하였으나, 최

근에는 Word2Vec와 같은 기계학습 알고리즘을 사용하기도 한다.

[0009] 또한, 종래 소셜 네트워크 서비스 데이터에서 추출한 키워드를 그래프로 생성하여 다양한 방법으로 클러스터링하여 이벤트를 검출하는 방법이 있는데, 소셜 네트워크 서비스 데이터에서 추출한 키워드를 벡터 공간 모델에 적용하여 그래프를 생성고, 동시 출현 빈도를 고려하여 그래프에 가중치를 부여한 뒤 클러스터링한 뒤 이벤트를 검출하였다.

[0010] 그리고, 기존의 지오 태그(Geo-Tag)를 활용한 이벤트 검출 방법에서는 실제 소셜 네트워크 서비스 데이터의 대부분에서 지오 태그(Geo-Tag)가 없다는 문제점이 있다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 대한민국 공개특허 10-2018-0055580

발명의 내용

해결하려는 과제

[0012] 본 발명은 상기와 같은 문제점을 해결하기 위하여 안출된 것으로서, 소셜 네트워크에서 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제공하는데 그 목적이 있다.

[0013] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0014] 이와 같은 목적을 달성하기 위한 본 발명의 소셜 네트워크에서 지역 이벤트 검출 시스템은 소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행하는 데이터 수집 모듈, 상기 데이터 수집 모듈에서 정제된 데이터를 이용하여 초기 그래프를 모델링하는 초기 그래프 모델링 모듈, 타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트하는 연관 데이터 반영 모듈 및 업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출하는 이벤트 검출 모듈을 포함한다.

[0015] 상기 데이터 수집 모듈은, 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행할 수 있다.

[0016] 상기 초기 그래프 모델링 모듈은, 지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고, 지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고, 지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행할 수 있다.

[0017] 상기 초기 그래프 모델링 모듈은, 각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고, 각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성할 수 있다.

[0018] 상기 초기 그래프 모델링 모듈은, 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정할 수 있다.

[0019] 상기 연관 데이터 반영 모듈은, 초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가

발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 있으면 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 상기 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성할 수 있다.

[0020] 상기 이벤트 검출 모듈은, 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고, 생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생성하고, 연관이 있는 클러스터들은 병합하고, 이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출할 수 있다.

[0021] 본 발명의 일 실시예에서 상기 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정하며, m은 전체 간선 수, n은 전체 노드 수, w_{ij}는 정점 V_i와 V_j사이를 연결하는 간선의 가중치, k_i는 V_i와 연결된 모든 간선의 가중치 합이고, $\delta(c_i, c_j)$는 V_i와 V_j가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수라고 할 때, 네트워크 모듈성 MN을,

$$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (3)$$

의 수학식으로 나타낼 수 있다.

[0022] 상기 이벤트 검출 모듈은, 하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때, 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의 클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합할 수 있다.

[0023] 본 발명의 소셜 네트워크에서 지역 이벤트 검출 시스템에서의 지역 이벤트 검출 방법에서, 소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행하는 데이터 수집 단계, 상기 데이터 수집 단계에서 정제된 데이터를 이용하여 초기 그래프를 모델링하는 초기 그래프 모델링 단계, 타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트하는 연관 데이터 반영 단계 및 업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출하는 이벤트 검출 단계를 포함한다.

[0024] 상기 데이터 수집 단계에서, 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행할 수 있다.

[0025] 상기 초기 그래프 모델링 단계에서, 지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고, 지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고, 지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행할 수 있다.

[0026] 상기 초기 그래프 모델링 단계에서, 각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고, 각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성할 수 있다.

[0027] 상기 초기 그래프 모델링 단계에서, 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정할 수 있다.

[0028] 상기 연관 데이터 반영 단계에서, 초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가 발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고, 초기 그래프에서 지역 키워드와

일치하는 키 노드가 있으면 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 상기 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성할 수 있다.

[0029] 상기 이벤트 검출 단계에서, 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고, 생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생성하고, 연관이 있는 클러스터들은 병합하고, 이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출할 수 있다.

[0030] 상기 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정하며, m은 전체 간선 수, n은 전체 노드 수, w_{ij} 는 정점 V_i 와 V_j 사이를 연결하는 간선의 가중치, k_i 는 V_i 와 연결된 모든 간선의 가중치 합이고, $\delta(c_i, c_j)$ 는 V_i 와 V_j 가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수라고 할 때, 네트워크 모듈성 MN을,

$$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

(3)의 수학식으로 나타낼 수 있다.

[0031] 상기 이벤트 검출 단계에서, 하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때, 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의 클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합할 수 있다.

발명의 효과

[0032] 본 발명에 의하면, 소셜 네트워크에서 연관 문서 분석을 통해 지역 이벤트를 검출함으로써, 보다 정확하고 빠르게 지역 이벤트를 검출할 수 있는 효과가 있다. 따라서, 본 발명에 의하면, 화재, 지진 등의 재난재해와 같은 지역 이벤트 발생 시, 소셜 네트워크에서 지역 이벤트 검출을 통한 정보를 빠르게 확산시킴으로써, 인명 피해와 경제적 피해를 줄일 수 있는 효과가 있다.

[0033] 더 나아가서, 본 발명을 활용하면, 지도 API를 사용하여 건물의 이름과 같이 더욱 구체적이고 정확한 지역 이벤트 검출이 가능하다. 또한, 실시간 처리를 통한 이벤트 검출을 구현하여 소규모 재난 알림 시스템 등에 적용할 수 있다.

도면의 간단한 설명

[0034] 도 1은 본 발명에서 제안하는 지역 이벤트 검출 시스템의 전체 구조 및 세부 모듈을 나타낸 도면이다.

도 2는 수집한 소셜 네트워크 서비스 데이터의 일부를 이용하여 불용어 제거 및 명사 추출을 하는 과정을 나타낸다.

도 3은 지리정보 임베딩에 대한 예시를 나타낸다.

도 4는 가중치 부여 및 정점 점수 계산 과정을 도시한 것이다.

도 5는 키워드 그래프 생성 및 키 노드 선정 과정을 예시한 것이다.

도 6은 연관 문서를 예시한 것이다.

도 7은 본 발명의 일 실시예에 따른 연관 문서 분석 내용을 그래프에 추가하기 위한 과정을 나타낸 흐름도이다.

도 8은 본 발명의 일 실시예에 따른 이벤트 검출 과정을 예시한 것이다.

도 9는 본 발명의 일 실시예에 따른 소셜 네트워크에서 지역 이벤트 검출 시스템에서의 지역 이벤트 검출 방법을 보여주는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0035] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고

상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0036] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0037] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 갖는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0038] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0039] 본 발명에서는 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제안한다.

[0040] 본 발명의 연관 문서 분석을 통한 지역 이벤트 검출 기법은 다음과 같다.

[0041] 전체 데이터에 대해서 지오 태그(Geo-Tag)가 없는 데이터는 텍스트 마이닝 기법을 사용하여 지리적인 정보를 임베딩(embedding)하여 사용하고, 연관 문서가 지역 정보를 가지고 있으면 이를 활용하여 이벤트 검출에 사용한다. 여기서, 연관 문서란 소셜 네트워크 서비스에서 제공하는 댓글과 스레드(thread) 기능을 의미한다. 사용자들은 본인 또는 다른 사람들이 남긴 게시물에 댓글을 작성하여 자신의 의견을 표출한다. 그리고 소셜 네트워크 서비스 중 트위터의 경우 게시글인 트윗을 한 개 발행 할 때 140자 이내로 작성해야 하는 제한이 있다. 그래서 트위터를 이용하여 2개 이상의 트윗을 올려야 하는 경우, 스레드라는 기능을 사용하여 글을 작성할 수 있다. 스레드는 하나의 주제와 관련된 여러 개의 트윗을 연결하여 조회할 수 있고, 게시물에 대한 추가적인 정보나 업데이트 된 정보를 게시할 때 사용한다.

[0042] 도 1은 본 발명에서 제안하는 지역 이벤트 검출 시스템의 전체 구조 및 세부 모듈을 나타낸 도면이다.

[0043] 도 1을 참조하면, 본 발명의 소셜 네트워크에서 지역 이벤트 검출 시스템은 데이터 수집 모듈(100), 초기 그래프 모델링 모듈(200), 연관 데이터 반영 모듈(300), 이벤트 검출 모듈(400)을 포함한다.

[0044] 데이터 수집 모듈(100)은 소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행한다.

[0045] 초기 그래프 모델링 모듈(200)은 데이터 수집 모듈(100)에서 정제된 데이터를 이용하여 초기 그래프를 모델링한다.

[0046] 연관 데이터 반영 모듈(300)은 타임 윈도우(time window) 내의 연관 문서를 분석하여 그래프를 업데이트한다.

[0047] 이벤트 검출 모듈(400)은 업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출한다.

[0048] 본 발명에서 데이터 수집 모듈(100)은 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행할 수 있다.

[0049] 초기 그래프 모델링 모듈(200)은 지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고, 지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고, 지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행할 수 있다.

[0050] 초기 그래프 모델링 모듈(200)은 각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고, 각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성할 수 있다.

[0051] 초기 그래프 모델링 모듈(200)은 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정할 수 있다.

[0052] 상기 연관 데이터 반영 모듈(300)은 초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가 발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 있으면 초기 그래프에 연관 문서 내용을 추가하여 업데이트하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성할 수 있다.

[0053] 이벤트 검출 모듈(400)은 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고, 생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생성하고, 연관이 있는 클러스터들은 병합하고, 이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출할 수 있다.

[0054] 이벤트 검출 모듈(400)은 하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때, 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의 클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합할 수 있다.

[0055] 도 1을 참조하면, 데이터 수집 모듈(100)에서는 소셜 네트워크 서비스 데이터를 저장하고 행정구역별 지리 정보 사전을 구축한다. 그리고 데이터 전처리 과정을 통해서 검색용어로 사용하지 않는 단어인 불용어를 제거하고 명사를 추출한다. 소셜 네트워크 서비스 데이터에는 특수문자, 줄임말, 은어 등의 비격식체를 사용하며 비문 표현이 많기 때문에 전처리 과정을 통해서 데이터를 정제하는 과정이 필요하다.

[0056] 그리고, 초기 그래프 모델링 모듈(200)에서는 정제된 데이터를 이용하여 초기 그래프를 모델링하는 과정을 수행한다. 초기 그래프 모델링 모듈(200)에서는 위치 정보인 지오 태그(Geo-Tag)가 없는 데이터에 대해서 지리 정보 사전을 이용하여 지역에 대한 정보를 임베딩(embedding)하고, 키워드 그래프를 생성한 뒤, 각 정점과 간선에 가중치를 부여한 뒤, 중심 노드가 되는 키(Key) 노드(node)를 선정한다.

[0057] 연관 데이터 반영 모듈(300)에서는 타임 윈도우(time window) 내의 연관 문서를 분석하여 그래프를 업데이트 하는 과정을 수행한다.

[0058] 마지막으로, 이벤트 검출 모듈(400)에서는 그래프 클러스터링을 통해 이벤트를 검출한다.

[0059] 소셜 네트워크 서비스는 사용자들이 자신의 의견을 자유롭게 표현 할 수 있는 공간이기 때문에 소셜 네트워크 서비스 데이터를 수집하게 되면 필요한 정보와 불필요한 정보가 섞여 있을 수 있다. 이러한 데이터를 그대로 사용하게 된다면 이벤트를 검출하는데 있어서 이상치나 무작위의 오류가 발생하는 노이즈가 커질 수 있고 정확도가 떨어진다. 따라서 노이즈를 줄이고 정확도를 높이기 위해서 데이터를 전처리 해주는 과정이 필요하다. 즉, 수집한 소셜 네트워크 서비스 데이터는 특수 문자와 줄임말 사진, 링크 등이 포함 되어 있기 때문에 이를 제거해주는 전처리 과정이 한다. 그리고 데이터 전처리 과정에서 지오 태그(Geo-Tag)가 포함되어 있는 게시글의 경우 지오 태그(Geo-Tag) 정보를 텍스트와 별도로 저장하여 추후 키워드 그래프를 생성할 때 활용한다. 그리고, 본 발명에서 제안하는 기법에서는 키워드 그래프를 생성하여 이벤트 검출에 활용함으로써, 불용어 제거를 거친 데이터에서 형태소 분석을 실행하여 명사를 추출하여 사용한다.

[0060] 도 9는 본 발명의 일 실시예에 따른 소셜 네트워크에서 지역 이벤트 검출 시스템에서의 지역 이벤트 검출 방법을 보여주는 흐름도이다.

[0061] 도 9를 참조하면, 본 발명의 지역 이벤트 검출 방법은 데이터 수집 단계(S100), 초기 그래프 모델링 단계

(S200), 연관 데이터 반영 단계(S300), 이벤트 검출 단계(S400)를 포함한다.

[0062] 데이터 수집 단계(S100)에서는 소셜 네트워크 서비스 데이터를 수집하여 저장하고, 행정구역별 지리 정보 사전을 구축하고, 검색용어로 사용하지 않는 단어인 불용어를 제거하고 데이터를 정제하고, 명사를 추출하는 데이터 전처리 과정을 수행한다.

[0063] 초기 그래프 모델링 단계(S200)에서는 데이터 수집 단계(S100)에서 정제된 데이터를 이용하여 초기 그래프를 모델링한다.

[0064] 연관 데이터 반영 단계(S300)에서는 타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트한다.

[0065] 이벤트 검출 단계(S400)에서는 업데이트된 그래프에 대하여 그래프 클러스터링을 통해 지역 이벤트를 검출한다.

[0066] 본 발명에서 데이터 수집 단계(S100)에서 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 및 링크를 제거하고, 형태소 분석을 실행하여 명사를 추출하는 방식으로 데이터 전처리 과정을 수행할 수 있다.

[0067] 초기 그래프 모델링 단계(S200)에서 지리 정보가 포함되어 있는 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그 정보를 그대로 이용하여 지역 노드로 사용하고, 지리 정보가 없는 데이터에 대하여 지역, 지명 및 키워드를 추출하고, 사전에 구축된 지리 정보 사전을 바탕으로 일치하는 지리 정보가 있으면, 해당 지리 정보를 데이터에 임베딩하는 방식으로 데이터에 지리 정보를 부여하여 지역 노드로 사용하고, 지역 노드가 포함되어 있으며, 키워드를 기반으로 정점과 간선으로 이루어진 키워드 기반 그래프를 생성하고, 해당 소셜 네트워크 서비스의 특성을 반영하여 상기 키워드 기반 그래프의 각 정점과 간선에 가중치를 부여하고, 정점들 중에서 중심이 되는 키 노드를 선정하는 방식으로 초기 그래프 모델링을 수행할 수 있다.

[0068] 초기 그래프 모델링 단계(S200)에서 각 정점에 키워드 타입을 부여하되, 각 정점의 해당 키워드가 지역, 지명에 해당하거나 지오 태그가 있으면 지리정보 타입을 부여하고, 나머지 키워드에 일반 단어 타입을 부여하고, 각 정점에서 해당 키워드와 동시 출현한 키워드의 정점을 간선으로 연결하는 방식으로 키워드 기반 그래프를 생성할 수 있다.

[0069] 초기 그래프 모델링 단계(S200)에서 각 정점과 간선들에 대해 가중치를 부여하되, 정해진 타임 윈도우에서 해당 키워드의 출현 빈도에 따라 각 정점에 가중치를 부여하고, 해당하는 두 키워드의 동시 출현 빈도에 따라 각 간선에 가중치를 부여하고, 각 정점 가중치와 각 간선 가중치를 이용하여 각 정점에 부여되는 점수를 계산하고, 각 정점에 부여된 점수를 이용하여 핵심 단어를 의미하는 키 노드를 선정할 수 있다.

[0070] 연관 데이터 반영 단계(S300)에서 초기 그래프가 구축된 상태에서 댓글, 스레드 게시물을 포함하는 연관 문서가 발생하면, 데이터 전처리를 통해 상기 연관 문서에서 지역 키워드를 추출하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 있으면 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 초기 그래프에서 지역 키워드와 일치하는 키 노드가 없으면, 상기 연관 문서에서 나머지 키워드를 추출하고, 초기 그래프에서 나머지 키워드와 일치하는 키워드가 있으면, 초기 그래프에 상기 연관 문서 내용을 추가하여 업데이트하고, 일치하는 키워드가 없으면 별도의 그래프를 생성할 수 있다.

[0071] 이벤트 검출 단계(S400)에서 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 정점들을 하나의 클러스터로 묶고, 생성된 클러스터들 사이의 연관 관계를 분석하여, 연관이 없는 클러스터들은 간선을 끊어 독립된 클러스터로 생성하고, 연관이 있는 클러스터들은 병합하고, 이러한 방식으로 도출된 각 독립적인 클러스터를 지역 이벤트로 검출할 수 있다.

[0072] 이벤트 검출 단계(S400)에서 하나의 클러스터로 묶인 내부 정점 사이에 연결된 간선들의 평균 가중치를 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의할 때, 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교하여, 내부 가중치의 값이 더 크면 두 개의 클러스터를 연결하는 외부 간선을 제거하여 각각 독립된 클러스터로 생성하고, 외부 가중치의 값이 더 크면 두 클러스터를 하나로 병합할 수 있다.

[0074] 도 2는 수집한 소셜 네트워크 서비스 데이터의 일부를 이용하여 불용어 제거 및 명사 추출을 하는 과정을 나타낸다.

[0075] 도 2를 참조하면, 수집한 소셜 네트워크 서비스 데이터에서 특수문자, 초성 그리고 링크를 제거한 뒤, 한글 형태소 분석기를 사용하여 명사만을 추출하면 각각의 트윗에 대한 전처리 과정 수행 결과가 나타난다. 이때, 추출된 명사 키워드 전체를 사용하게 되면 키워드 그래프 생성 시 처리 시간이 오래 걸리고 노이즈가 증가하여 정확

도가 떨어지게 된다. 따라서 일정 횟수 이상 나타나는 단어를 선정하여 사용하는 것이 바람직하다.

[0076] 본 발명에서 소셜 네트워크 서비스 데이터 전처리 과정을 거치면 키워드 집합이 생성된다. 이러한 키워드 집합만으로는 현재 이슈가 되고 있는 이벤트를 알기 어렵고, 불필요한 정보들을 많이 포함하고 있다. 따라서 키워드 집합을 이용하여 키워드 기반 그래프를 생성하여 각 단어의 중요도와 언급량과 같은 화제성을 고려해야한다.

[0077] 초기 그래프 모델링 모듈(200)은 지리정보 임베딩, 키워드 그래프 생성 및 가중치 부여, Key 노드 선정의 총 3단계로 나눠진다. 본 발명에서 제안하는 기법에서는 우선 지오 태그(Geo-Tag)가 없는 정보에 대해서 사전에 구축한 지리 정보 사전을 바탕으로 일치하는 지리정보가 있으면 이를 임베딩하여 지리정보를 부여한다.

[0078] 그리고, 동시에 출현한 단어들을 키워드 기반 그래프로 생성하고, 소셜 네트워크 서비스의 특성을 반영하여 그래프의 각 정점과 간선에 가중치를 부여한다.

[0079] 그리고, 키워드 그래프에서 중심이 되는 Key 노드를 선정하여 초기 그래프 모델링을 수행한다.

[0080] 보통 사용자들의 지오 태그(Geo-Tag)에 대한 활용이 낮아 소셜 네트워크 서비스 데이터의 약 1% 정도만 지오 태그(Geo-Tag)에 대한 정보를 가지고 있다. 따라서 지오 태그(Geo-Tag)가 없는 데이터에 대해서 지리정보를 활용할 수 있는 방법이 필요하다.

[0081] 본 발명에서는 앞서 데이터 수집 모듈(100)에서 데이터 전처리 및 지리 정보 사전 구축 과정에서 미리 정의해 놓은 지리 정보 사전을 이용하여 지오 태그(Geo-Tag)가 없는 데이터에 대해서 지리정보를 임베딩한다.

[0082] 도 3은 지리정보 임베딩에 대한 예시를 나타낸다.

[0083] 도 3에서 (a)와 같이 지오 태그(Geo-Tag)가 있는 데이터에 대해서는 지오 태그(Geo-Tag)의 정보를 그대로 이용하여 키워드 그래프 생성 시 지역노드로 사용하고, (b)와 같이 지오 태그(Geo-Tag)가 없는 데이터에 대해서는 지리 정보 사전을 이용한 텍스트 마이닝을 통해서 추출한 지리정보를 지역 노드로 구분하는 방식의 임베딩을 사용한다.

[0084] 지역 이벤트를 검출하기 위해서 키워드 집합과 지리정보를 이용하여 키워드 그래프 G_i를 생성한다. G_i는 지리정보 임베딩을 마친 단어들에 대해서 키워드를 정점으로 생성하고, 각각의 정점에는 키워드 타입을 부여한 그래프이다.

[0085] 여기서, 키워드 타입은 지리정보와 일반 단어로 나뉜다. 단어가 지역, 지명에 해당하거나 지오 태그(Geo-Tag)가 있다면 지리정보 타입으로 정의될 수 있고, 나머지 단어들은 일반 단어에 해당한다. 지역 이벤트는 지역, 지명과 밀접한 연관을 가지고 있으므로 타입 분류를 통해 추후 그래프 클러스터링 또는 지역별 이벤트 검색에 활용하기 용이하다.

[0086] 그리고, 각 키워드 정점들에 대해서 동시 출현한 단어들에 대해서 간선으로 연결해준다.

[0087] 도 4는 가중치 부여 및 정점 점수 계산 과정을 도시한 것이다.

[0088] 도 4는 동시 출현빈도와 좋아요, 리트윗과 같은 소셜 네트워크에서 사용자들의 명시적인 관심을 반영하여 가중치를 부여하고, 정점 점수를 계산하는 과정을 나타낸다.

[0089] 본 발명에서 생성된 키워드 그래프를 이벤트 검출 목적에 맞게 사용하려면 각각의 정점과 간선들에 대해서 가중치를 부여한다. 동시 출현한 단어를 기반으로 연결된 간선의 가중치는 두 단어 간의 유사도를 나타낸다. 예를 들어, '고양이'와 '반려동물'이라는 단어는 두 단어 사이에 연관이 있기 때문에 가중치를 높게 주어야 하고, '고양이'와 '달력' 같은 단어는 가중치를 낮게 주어야 한다.

[0090] 따라서 간선에는 단어의 동시 출현 빈도를 사용하여 가중치를 부여한다. 동시 출현 빈도는 특정 단어를 기준으로 사용자가 설정한 윈도우(window)의 크기에 따라 달라진다. 예를 들어, 윈도우의 크기가 1이면 해당하는 단어의 바로 앞, 뒤 단어만 동시 출현 횟수에 포함시킨다. 그리고 동시 출현 횟수 값을 0에서 1사이의 값으로 정규화한다.

[0091] 타임 윈도우 t 시간 내에 발생한 키워드들의 정점에 대해서 모두 같은 가중치를 부여한다면 어떤 키워드가 이벤트와 관련 있는지 알기 어렵다. 따라서, 들이 많이 언급한 단어와 좋아요 또는 리트윗과 같이 관심을 명시적으로 표현한 게시글에 있는 단어일수록 중요한 단어일 가능성이 높으므로, 이를 키워드 정점에 반영한다.

[0092] 본 발명에서 제안하는 기법에서는 변형된 TF-IDF 알고리즘을 사용하여 각각의 정점에 가중치를 부여하여 점수를

계산한다.

[0093] $$S(V_i) = V_i * tf_i * \frac{idf_{i,t}}{idf_{i,t-1}} * \log(like + retweet) \quad (1)$$

[0094] 수식 1에서 $S(V_i)$ 는 단어의 중요도에 따라 계산된 키워드 i에 대한 정점 V_i 의 점수를 나타낸다. 초기 정점 V_i 는 모두 1로 초기화 한 뒤 시간 속성을 고려한 TF-IDF를 계산한다. 그리고 좋아요 수인 $like$와 리트윗 수인 $retweet$을 더한 뒤, log를 사용하여 곱해준다. 좋아요와 리트윗은 사람들이 클릭 한번으로 자신의 의견을 표현할 수 있는 수단이기 때문에 중요한 단어일수록 값이 커지게 된다. 따라서 이를 그대로 곱해주게 된다면 결과 값이 좋아요와 리트윗 수에 많은 영향을 받기 때문에 log를 사용하여 조절한다. tf_i 와 $idf_{i,t}$는 각각 단어 i의 출현 빈도(Term Frequency)와 역문서 빈도(Inverse Document Frequency)를 나타낸다. 역문서 빈도는 현재의 타임윈도우 t 시간 값과 바로 이전 시간의 값의 비율을 사용한다.

[0095] 본 발명에서 생성한 키워드 그래프 정점의 점수를 계산하면 각 키워드의 중요도를 알 수 있다. 이때, 점수가 높은 정점일수록 이벤트 검출에 있어서 핵심 단어가 된다. 각 정점에 시간 속성을 고려한 TF-IDF를 이용하여 가중치를 부여하고, TextRank 알고리즘을 사용하여 핵심 단어를 추출하기 위한 점수를 계산하여 부여한다.

[0096] $$TR(V_i) = (1-d) + d * \sum_{j \in V_i} \frac{w_{ij}}{\sum_{k \in V_j} w_{jk}} TR(V_j) \quad (2)$$

[0097] 수식 2에서 $TR(V_i)$ 는 TextRank를 이용하여 각 정점에 부여되는 점수를 나타낸다. 초기 계산에 사용되는 $TR(V_i)$ 는 수식 1의 $S(V_i)$ 결과를 이용한다. V_i 는 지역 타입 정점과 일반 타입 정점을 모두 포함하는 정점 집합 중 점수를 계산하고자 하는 정점이다. V_j 는 V_i 와 간선으로 연결된 키워드 정점을 나타낸다. w_{ij}는 V_i 의 간선들 중 V_j 와 연결된 간선의 가중치를 나타내며, w_{jk} 는 V_j 와 연결된 간선의 가중치를 나타낸다. 본 발명에서는 댐핑 팩터(Damping Factor)인 d를 이용하여 랜덤 확률 변수를 조정할 수 있으며, 예를 들어 일반적으로 사용하는 값인 0.85를 d로 채택하여 사용할 수 있다.

[0098] 핵심단어를 기준으로 그래프 클러스터링을 통해 이벤트 검출을 하면, 해당 이벤트와 연관 있는 단어들을 알기 쉽고, 전체 키워드 정점에 대해서 클러스터링 하는 것보다 처리 시간을 줄일 수 있다. 따라서 키워드 그래프 정점의 점수 값을 이용하여 핵심 단어를 의미하는 키(Key) 노드를 선정한다.

[0099] 키 노드는 정점의 점수 값을 내림차순으로 정렬하여 사용자 설정에 따라 Top-k개를 선정한다. 그리고 지역 이벤트 검출에 있어서 지역 정보는 중요한 역할을 하므로, 지오 태그(Geo-Tag) 또는 지리 정보 사전을 이용하여 지역 정점으로 라벨링한 정점도 키 노드에 포함시킨다.

[0100] 도 5는 키워드 그래프 생성 및 키 노드 선정 과정을 예시한 것이다.

[0101] 도 5에서 전처리한 데이터를 이용하여 명사집합들에 대해서 초기 그래프를 생성하면 (a)와 같은 결과가 나타난다. 여기서, '속초'와 '고성군'은 지역 정점으로 라벨링이 되어 있는 상태이다. 그리고 각 정점과 간선의 가중치에 따라서 (b)와 같이 정점 점수를 계산하여 내림차순으로 정렬하고 상위 3개의 결과 값인 '산불', '속초', '고성군'을 키(Key) 노드로 선정한다. 따라서 (c)에서 보는 바와 같이, 이러한 결과가 반영되어 Key 노드가 선정된 그래프가 생성된다.

[0102] 본 발명에서 소셜 네트워크 서비스 데이터를 이용한 이벤트 검출에 연관 문서를 분석하여 고려할 필요가 있다. 특히, 그 정보가 지역과 관련이 되어 있는 경우, 지역 이벤트 검출에 있어서 중요한 정보가 되기 때문에 이를 활용하여 정확도를 높일 수 있다.

[0103] 도 6은 연관 문서를 예시한 것이다.

[0104] 도 6을 참조하면, User1이 이벤트에 대한 정보를 트윗으로 게시하였지만, 해당 게시글에는 지리적인 정보가 포함되어 있지 않아서 지역 이벤트로 검출하기 어렵다.

[0105] 그러나 User2의 댓글에 대한 답변으로 User1이 댓글을 통해서 지리정보가 추가되었다. 이처럼, 기존의 기법에서

는 연관 문서에 포함되어 있는 정보들이 무시되었지만, 본 발명에서 제안하는 기법에서는 연관 문서 분석을 통해서 추가적으로 제공되는 지리정보를 그래프에 반영할 수 있다.

[0106] 도 7은 본 발명의 일 실시예에 따른 연관 문서 분석 내용을 그래프에 추가하기 위한 과정을 나타낸 흐름도이다.

[0107] 도 7을 참조하면, 초기 그래프가 구축된 상태에서 연관 문서가 발생하면(S701), 데이터 전처리를 통해서 해당 연관 문서에 지역 키워드가 있는지 검사한다(S703).

[0108] 본 발명에서 제안하는 기법에서는 지역 정보를 가지고 있는 연관 문서만을 사용한다. 이벤트는 주로 특정 지역, 장소와 밀접한 연관이 있기 때문에 지역 이벤트 검출에서 지역 정보는 중요한 역할을 가지고 있다. 모든 연관 문서에 대해서 그래프에 추가하려면 처리 시간이 늘어나고 필요 없는 정보가 다수 포함되어 정확도가 떨어질 수 있기 때문에 지역 키워드 여부를 통해서 해당하는 연관 문서만 분석 과정을 거친다.

[0109] 만약 지역 키워드가 존재하면 해당 키워드와 이미 구축되어 있는 키(Key) 노드를 비교한다(S705). 이전 단계인 키 노드 선정에서 키 노드는 지역 정점을 반드시 포함하므로 일치하는 키 노드가 있으면, 기존에 존재하는 그래프에 연관 문서 내용을 추가하여 부족한 지리정보를 보충한다(S711).

[0110] 그러나, 해당 연관 문서에서 추출된 지역 키워드가 기존 그래프에 존재하지 않는 정점이라면 나머지 키워드를 기존의 그래프와 비교하여 일치하는 정보가 있는지 여부를 판단해야한다(S707, S709). 만약 있다면 기존에 존재하는 그래프에는 지역 정보가 없었지만 연관 문서 정보를 반영하여 정보를 추가한다(S711). 하지만 없다면 새롭게 검출된 이벤트일 가능성이 높기 때문에 별도의 그래프를 생성하여 연관 문서를 분석한 정보를 추가하고, 전술한 과정을 수행하여 추가된 정점과 간선에도 가중치를 부여한다(S713).

[0111] 연관 문서 분석을 마친 키워드 그래프를 통해서는 그래프의 연결 관계들 때문에 이벤트를 쉽게 파악하기 힘들다. 따라서 생성한 그래프를 클러스터링하여 이벤트를 검출해야한다. 본 발명에서 제안하는 기법에서는 키 노드와 간선의 가중치를 이용하여 그래프를 클러스터링 한다. 그래프 클러스터링을 통해 각 클러스터 간의 관계를 기반으로 유사한 이벤트를 병합하거나 별도의 지역 이벤트로 검출할 수 있다. 키워드 그래프의 각각 간선에는 키워드 간의 유사도에 따라 가중치가 부여되어 있으며, 가중치가 높을수록 연관성이 높은 단어를 나타낸다. 이를 반영하여 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 노드를 하나의 클러스터로 묶는다. 이때, 기준이 되는 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정한다. 다음 수식 3은 네트워크 모듈성 NM을 계산하는 수식을 나타낸다.

[0112] $$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (3)$$

[0113] 여기서, m은 전체 간선 수, n은 전체 노드 수를 의미하며, w_{ij} 는 정점 V_i 와 V_j 사이를 연결하는 간선의 가중치를 나타낸다. k_i 는 V_i 와 연결된 모든 간선의 가중치 합이고, $\delta(c_i, c_j)$ 는 V_i 와 V_j 가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수이다.

[0114] 사용자가 원하는 최종 k개의 이벤트를 검출하기 위해서는 클러스터 사이의 연관 관계를 통하여, 서로 관련이 없는 클러스터라면 간선을 끊어 독립된 클러스터로 생성하거나, 두 개의 클러스터가 연관이 있다면 병합하여 하나의 클러스터로 만드는 과정이 필요하다.

[0115] 본 발명에서는 간선 가중치를 이용하여 결과로 도출된 각각의 독립적인 클러스터를 하나의 이벤트로 검출하여 제공한다. 클러스터로 묶인 내부 정점 사이에 연결된 간선의 평균 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의한다. 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교한다. 만약 외부 가중치의 값이 더 크다면 두 클러스터는 서로 밀접한 연관이 있기 때문에 클러스터를 하나로 병합한다. 하지만 내부 가중치의 합이 더 크다면 연관성이 떨어지는 각각의 독립된 이벤트일 가능성이 높으므로 두 개의 클러스터를 연결하는 외부 간선을 제거한다.

[0116] 최종적으로 도출된 클러스터의 정점들의 점수 값을 더하여 사용자가 원하는 Top k 개의 이벤트를 검출할 수 있다.

[0117] 도 8은 본 발명의 일 실시예에 따른 이벤트 검출 과정을 예시한 것이다.

[0118] 도 8에서 (a)는 키워드 그래프가 클러스터링 과정을 거쳐 3개의 클러스터로 나누어진 것을 알 수 있다. (b)는 내부 가중치와 외부 가중치를 비교하여 병합 클러스터를 서로 병합하거나, 외부 간선이 제거된 클러스터를 나타

내고 있으며, 이러한 과정을 통해 최종적으로 2개의 이벤트가 검출 된 것을 확인할 수 있다.

[0120] 본 발명에서는 소셜 네트워크 환경에서 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제안한다. 소셜 네트워크 서비스 데이터에서 키워드를 추출하여 키워드 기반 그래프를 생성한다. 그리고 키워드로 구성된 그래프의 정점과 간선에 소셜 네트워크 특성을 반영하여 가중치를 부여한다. 이와 같이 키워드 기반 그래프를 사용하면 검출되는 이벤트와 관련된 단어들을 쉽게 파악할 수 있는 장점이 있다. 지오 태그(Geo-Tag) 정보와 더불어 지리 정보 사전을 생성하여 생성된 키워드 그래프의 정점 중 지역 정보를 가지고 있는 정점을 지역 노드로 분류한다. 기존의 지오 태그(Geo-Tag)를 활용한 이벤트 검출 방법이 가지고 있는 실제 소셜 네트워크 서비스 데이터의 대부분은 지오 태그(Geo-Tag)가 없다는 한계점을 해결하기 위하여 지리 정보 사전을 사용한다. 지리 정보 사전이란, 지리를 대표하는 명사에 맵핑되는 위치 정보를 가지고 있는 데이터베이스를 의미한다. 지리 정보 사전을 사용하면 부족한 지리적 정보를 보충할 수 있다. 키워드 그래프를 가중치에 따라 클러스터링 한 후 연관문서 분석 과정을 통해 제안하는 기법의 정확도를 높인다. 클러스터 내부와 외부 간선 가중치를 이용하여 클러스터를 병합 또는 분리함으로써 지역 이벤트를 추출하여 결과를 도출한다.

[0121] 본 발명을 활용하면, 지도 API를 사용하여 건물의 이름과 같이 더욱 구체적이고 정확한 지역 이벤트 검출이 가능하다. 또한, 실시간 처리를 통한 이벤트 검출을 구현하여 소규모 재난 알림 시스템 등에 적용할 수 있다.

[0122] 이상 본 발명을 몇 가지 바람직한 실시예를 사용하여 설명하였으나, 이들 실시예는 예시적인 것이며 한정적인 것이 아니다. 본 발명이 속하는 기술분야에서 통상의 지식을 지닌 자라면 본 발명의 사상과 첨부된 특허청구범위에 제시된 권리범위에서 벗어나지 않으면서 다양한 변화와 수정을 가할 수 있음을 이해할 것이다.

부호의 설명

[0123] 100 데이터 수집 모듈

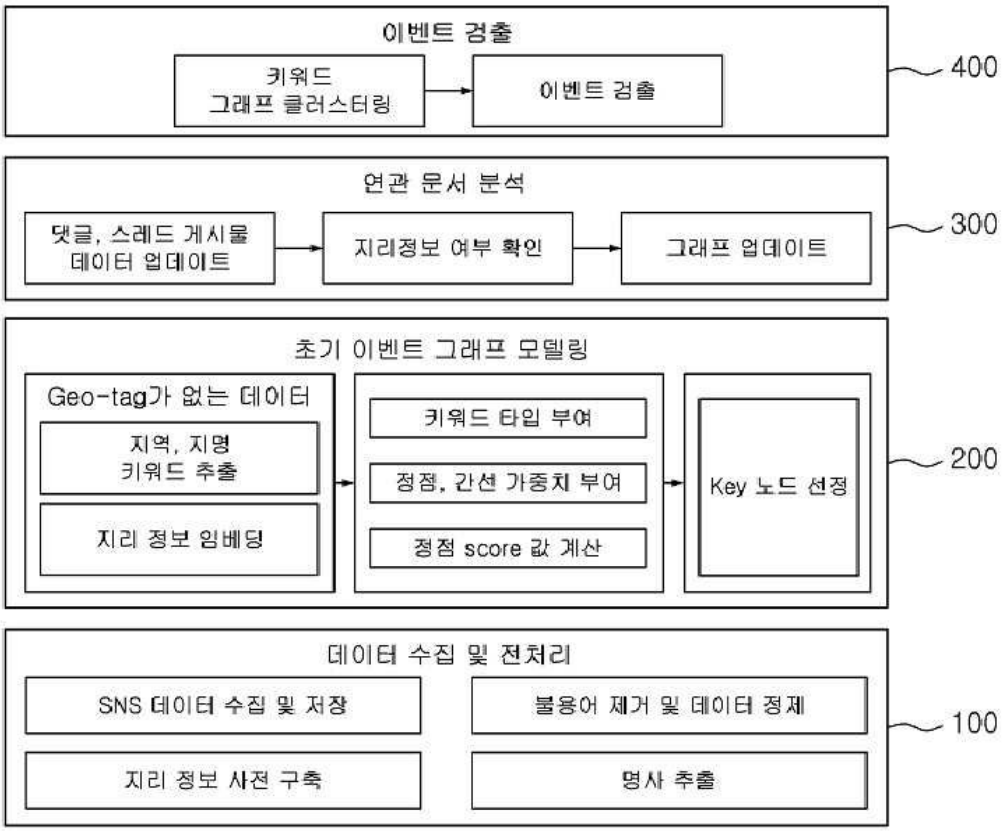
200 초기 그래프 모델링 모듈

300 연관 데이터 반영 모듈

400 이벤트 검출 모듈

도면

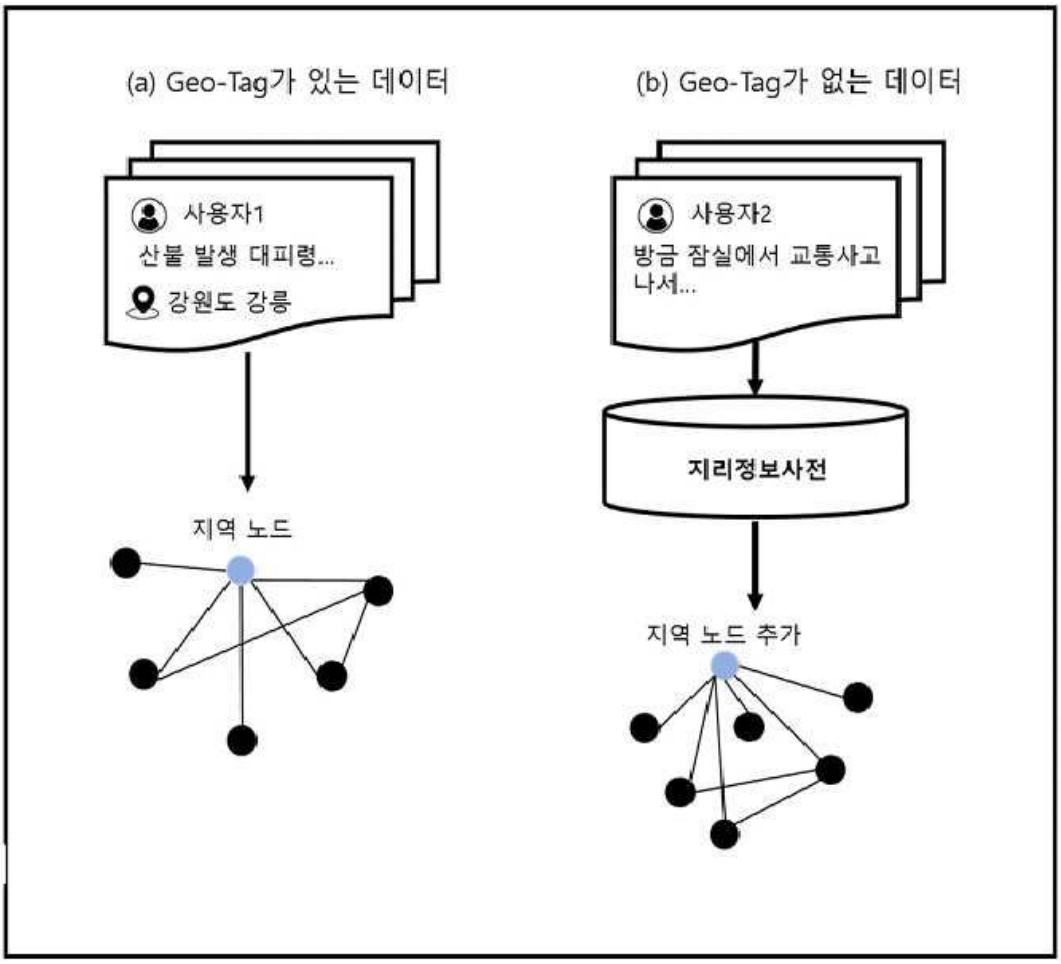
도면1


이벤트 검출
키워드
그래프 클러스터링
이벤트 검출
400
연관 문서 분석
댓글, 스레드 게시물
데이터 업데이트
지리정보 여부 확인
그래프 업데이트
300
초기 이벤트 그래프 모델링
Geo-tag가 없는 데이터
지역, 지명
키워드 추출
지리 정보 임베딩
키워드 타입 부여
정점, 간선 가중치 부여
정점 score 값 계산
Key 노드 선정
200
데이터 수집 및 전처리
SNS 데이터 수집 및 저장
불용어 제거 및 데이터 정제
지리 정보 사전 구축
명사 추출
100

도면2


User1 @Auser
지금 속초고등학교 기숙사에 불 붙었어 ㅠㅠㅜ
기숙사에 경보음 들려서 무섭다....
User2 @Buser
헐 개오바 ㄷㄷ 우리오빠 삼척 사는데 삼척캠퍼스 앞까지 번졌다고 함!!!
#삼척산불
User3 @Cuser
"30명 고립", "폭발", "기숙사 위험" ..고성산불 '심각한 재난'
news.v.daum.net
불용어 제거 및 명사 추출
지금, 속초 고등학교, 기숙사, 불, 기숙사, 경보음
우리, 오빠 ,삼척, 삼척, 캠퍼스, 앞, 삼척, 산불
30명, 고립, 폭발, 기숙사, 위험, 고성, 산불, 재난

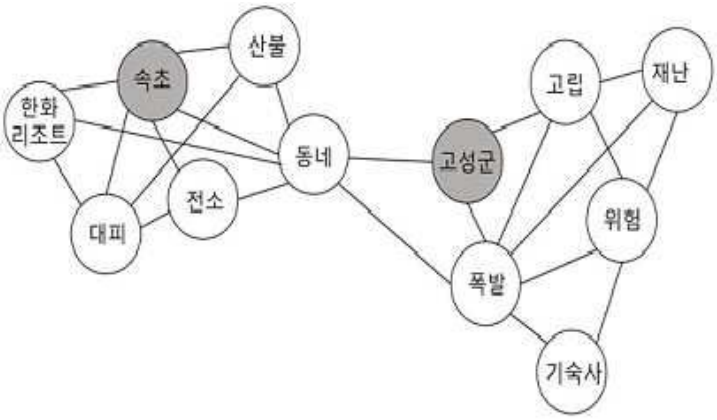
도면3


(a) Geo-Tag가 있는 데이터
(b) Geo-Tag가 없는 데이터
사용자1
산불 발생 대피령...
강원도 강릉
사용자2
방금 잠실에서 교통사고 나서...
지리정보사전
지역 노드
지역 노드 추가

도면4


동시 출현 빈도
출현 빈도
좋아요
리트윗
간선 가중치
정점 가중치
정점 점수 계산

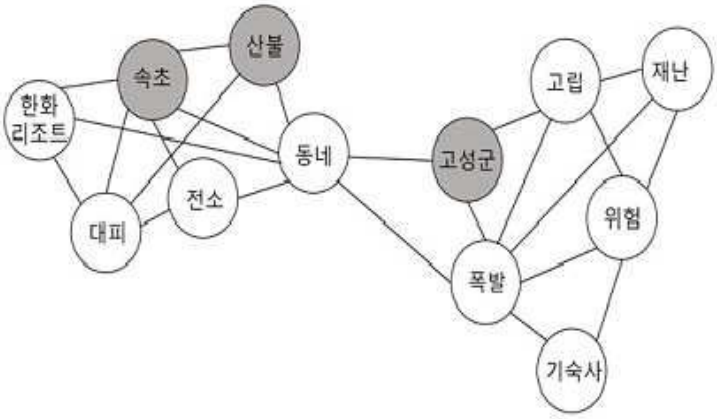
도면5



(a) 키워드 그래프 생성 및 지리 정보 임베딩

순위	키워드	점수
1	산불	31.36
2	속초	17.67
3	고성군	17.10
4	대피	15.32
5	위험	14.84
...		
12	고립	9.91

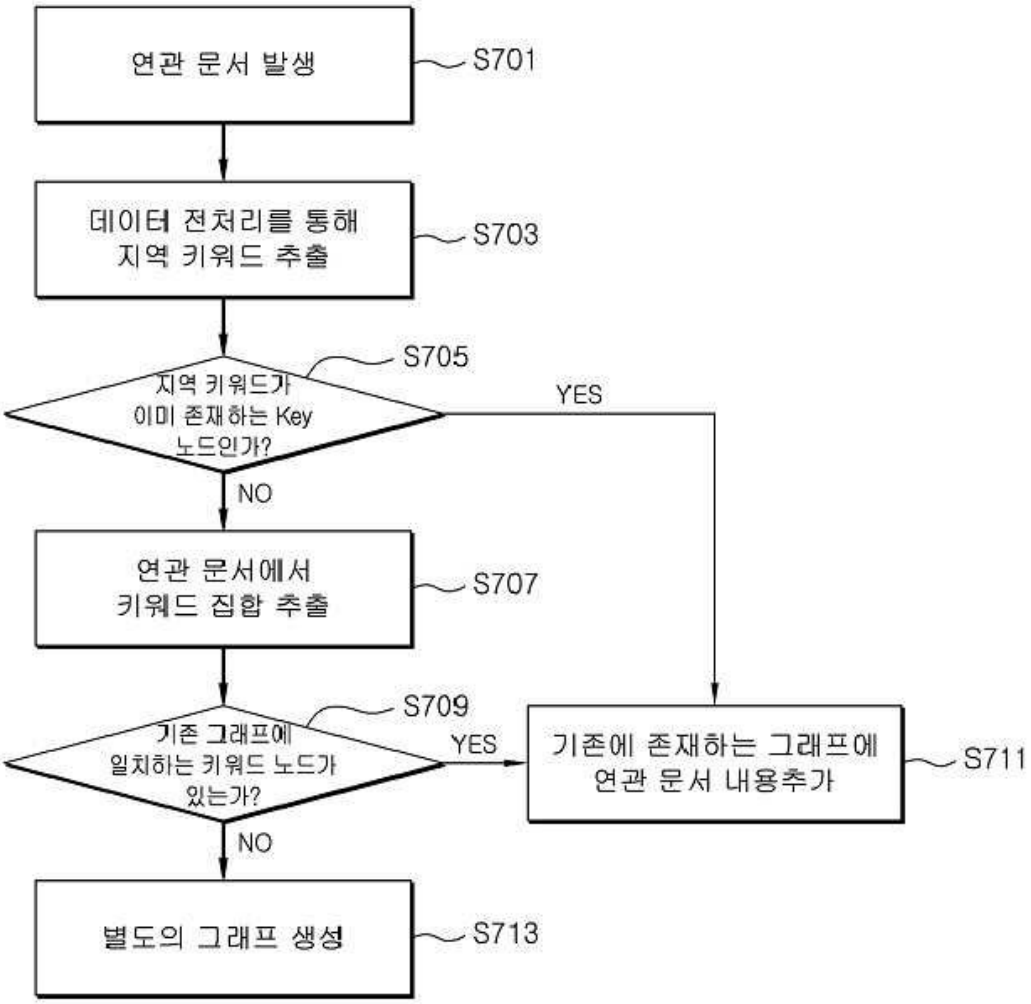
(b) 정점 점수 계산



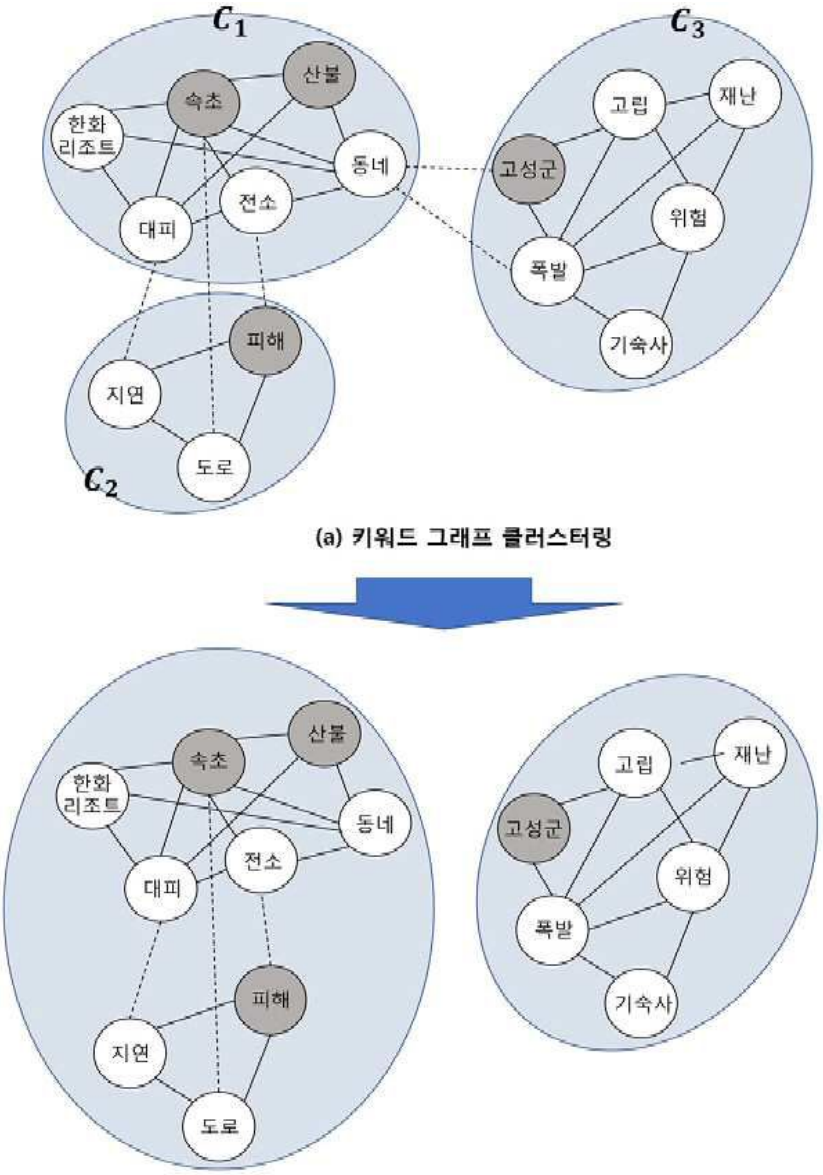
(c) Key 노드 선정

도면6

도면7


연관 문서 발생
S701
데이터 전처리를 통해 지역 키워드 추출
S703
지역 키워드가 이미 존재하는 Key 노드인가?
S705
YES
NO
연관 문서에서 키워드 집합 추출
S707
기존 그래프에 일치하는 키워드 노드가 있는가?
S709
YES
NO
기존에 존재하는 그래프에 연관 문서 내용추가
S711
별도의 그래프 생성
S713

도면8



(a) 키워드 그래프 클러스터링

(b) 이벤트 검출

도면9