(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0118452
(43) 공개일자 2021년09월30일

(51) 국제특허분류(Int. Cl.)
G06F 16/906 (2019.01) *G06F 16/9536* (2019.01)
G06F 16/9537 (2019.01) *G06F 16/9538* (2019.01)
(52) CPC특허분류
G06F 16/906 (2019.01)
G06F 16/9536 (2019.01)
(21) 출원번호 10-2021-7027045
(22) 출원일자(국제) 2020년01월29일
심사청구일자 2021년08월25일
(85) 번역문제출일자 2021년08월25일
(86) 국제출원번호 PCT/US2020/015732
(87) 국제공개번호 WO 2020/160186
국제공개일자 2020년08월06일
(30) 우선권주장
62/798,388 2019년01월29일 미국(US)

(71) 출원인
트위터, 인크.
미국, 캘리포니아 94103, 샌프란시스코, 스윗 900, 마켓 스트리트 1355
(72) 발명자
페도리즈작, 마테우즈
미국 94103 캘리포니아 샌 프라시스코 마켓 스트리트 1355 스윗 900 트위터, 인크.
프레더릭, 브렌트
미국 94103 캘리포니아 샌 프라시스코 마켓 스트리트 1355 스윗 900 트위터, 인크.
(뒷면에 계속)
(74) 대리인
특허법인 남앤남

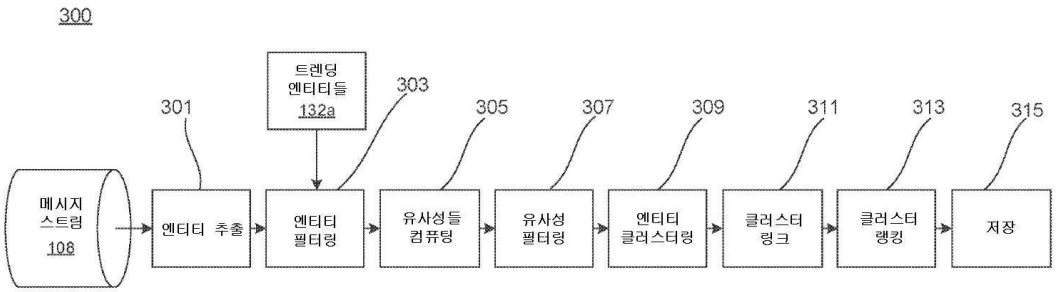
전체 청구항 수 : 총 25 항

(54) 발명의 명칭 **소셜 데이터 스트림에 대한 실시간 이벤트 검출**

(57) 요 약

일 양상에 따르면, 소셜 데이터 스트림들에 대한 이벤트 검출을 위한 방법은 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하는 단계, 및 메시지들의 스트림으로부터 이벤트를 검출하는 단계를 포함하고, 이는, 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하는 것, 제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 그룹을 검출하는 것, 및 제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하는 것을 포함할 수 있고, 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현한다. 방법은 이벤트를 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 저장하는 단계를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06F 16/9537 (2019.01)
G06F 16/9538 (2019.01)
H04L 51/16 (2013.01)
H04L 51/32 (2013.01)
(72) 발명자
라자람, 비제이
미국 94103 캘리포니아 샌 프라시스코 마켓 스트리트 1355 스윗 900 트위터, 인크.
종, 창타오
미국 94103 캘리포니아 샌 프라시스코 마켓 스트리트 1355 스윗 900 트위터, 인크.

명 세 서

청구범위

청구항 1

소셜 데이터 스트림들에 대한 이벤트 검출을 위한 방법으로서,

메시징 플랫폼에 의해, 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하는 단계;

상기 메시징 플랫폼에 의해, 상기 메시지들의 스트림으로부터 이벤트를 검출하는 단계로서, 상기 검출하는 단계는,

제1 시간 기간에 걸쳐 트렌딩(trending) 엔티티들의 제1 클러스터 그룹을 검출하는 단계 - 상기 제1 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함함 -;

제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하는 단계 - 상기 제2 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함함 -; 및

상기 제2 클러스터 그룹을 상기 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하는 단계 - 상기 클러스터 체인은 상기 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함 - 를 포함하는, 상기 검출하는 단계; 및

상기 메시징 플랫폼에 의해, 상기 이벤트를 상기 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 저장하는 단계를 포함하는, 방법.

청구항 2

제1 항에 있어서,

상기 메시징 플랫폼에 의해, 클라이언트 애플리케이션의 사용자 인터페이스에서 상기 이벤트에 관한 정보를 렌더링하기 위해 상기 클라이언트 애플리케이션에 디지털 데이터를 송신하는 단계를 더 포함하고, 상기 이벤트에 관한 정보는 상기 클러스터 체인으로부터의 정보를 포함하는, 방법.

청구항 3

제2 항에 있어서,

상기 이벤트에 관한 정보는 상기 제1 클러스터 그룹으로부터의 제1 트렌딩 엔티티 및 상기 제2 클러스터 그룹으로부터의 제2 트렌딩 엔티티를 식별하고, 상기 제2 트렌딩 엔티티는 상기 제1 트렌딩 엔티티와는 상이한, 방법.

청구항 4

제1 항에 있어서,

각각의 개개의 클러스터 그룹과 연관된 트렌딩 엔티티들의 인기도에 기초하여 상기 제1 클러스터 그룹 및 상기 제2 클러스터 그룹을 랭킹하는 단계를 더 포함하고, 상기 클러스터 체인은 랭킹된 클러스터 그룹들의 리스트를 포함하는, 방법.

청구항 5

제1 항에 있어서,

상기 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 식별하는 단계 - 상기 제1 클러스터 그룹은 상기 제1 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들로부터 검출됨 -; 및

상기 제2 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 식별하는 단계를 더 포함하고, 상기 제2 클러스터 그룹은 상기 제2 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들로부터 검출되는, 방법.

청구항 6

제1 항에 있어서,

상기 제1 클러스터 그룹에 클러스터 식별자를 할당하는 단계; 및

상기 제2 클러스터 그룹이 상기 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 상기 제1 클러스터 그룹의 상기 클러스터 식별자를 상기 제2 클러스터 그룹에 할당하는 단계를 더 포함하는, 방법.

청구항 7

제1 항에 있어서,

상기 제1 클러스터 그룹을 검출하는 단계는,

상기 제1 시간 기간의 복수의 트렌딩 엔티티들과 연관된 유사성 값들에 기초하여 유사성 그래프를 생성하는 단계 - 상기 유사성 그래프는 상기 복수의 트렌딩 엔티티들을 표현하는 노드들 및 상기 유사성 값들로 어노테이트된 에지들을 포함함 -; 및

상기 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 상기 유사성 그래프를 파티셔닝하는 단계를 포함하는, 방법.

청구항 8

제7 항에 있어서,

시간 윈도우에 걸친 상기 복수의 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 단계; 및

유사성 임계치 미만의 유사성 값들을 갖는 에지들이 상기 유사성 그래프로부터 제거되도록 상기 유사성 임계치에 기초하여 상기 유사성 그래프를 필터링하는 단계를 더 포함하고, 상기 필터링된 유사성 그래프는 상기 제1 클러스터 그룹을 검출하기 위해 상기 클러스터링 알고리즘에 따라 파티셔닝되는, 방법.

청구항 9

제1 항에 있어서,

상기 제2 클러스터 그룹은 최대 가중된 이분(bipartite) 매칭에 기초하여 상기 제1 클러스터 그룹에 링크되는, 방법.

청구항 10

실시간 이벤트를 검출하기 위한 메시징 시스템으로서,

상기 메시징 시스템은,

네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼; 및

메시지들을 전송 및 수신하기 위해 상기 메시징 플랫폼과 통신하도록 구성된 클라이언트 애플리케이션을 포함하고,

상기 메시징 플랫폼은,

제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하고 - 상기 제1 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함함 -;

제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하고 - 상기 제2 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함함 -;

상기 제1 클러스터 그룹과 상기 제2 클러스터 그룹 사이에서 공유되는 트렌딩 엔티티들의 수에 기초하여 상기 제2 클러스터 그룹을 상기 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하고 - 상기 클러스터 체인은 상기 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함 -;

상기 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 이벤트를 저장하도록 구성되고, 상기 클러스터 체인은 장래의 클러스터 링크를 위해 리트리브가능한, 메시징 시스템.

청구항 11

제10 항에 있어서,

상기 메시징 플랫폼은, 상기 클라이언트 애플리케이션의 사용자 인터페이스에서 상기 이벤트에 관한 정보를 렌더링하기 위해 디지털 데이터를 상기 클라이언트 애플리케이션에 송신하도록 구성되고, 상기 이벤트에 관한 정보는 상기 클러스터 체인으로부터의 정보를 포함하고, 상기 클러스터 체인으로부터의 정보는 트렌드 섹션, 타임라인, 또는 상기 클라이언트 애플리케이션에 리턴되는 검색 결과들의 일부에서 렌더링되는, 메시징 시스템.

청구항 12

제10 항에 있어서,

상기 메시징 플랫폼은 각각의 클러스터 그룹과 연관된 집계 인기도 메트릭에 기초하여 상기 제1 클러스터 그룹 및 상기 제2 클러스터 그룹을 랭킹하도록 구성되는, 메시징 시스템.

청구항 13

제10 항에 있어서,

상기 메시징 플랫폼은,

트렌드 검출기 서비스로부터 상기 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 리스트를 획득하고;

상기 메시징 플랫폼 상에서 교환되는 메시지들의 스트림으로부터 엔티티들을 추출하고;

상기 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 획득하기 위해 상기 트렌딩 엔티티들의 리스트를 사용하여 상기 추출된 엔티티들을 필터링하도록 구성되고,

상기 제1 클러스터 그룹은 상기 제1 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들을 사용하여 검출되는, 메시징 시스템.

청구항 14

제10 항에 있어서,

상기 메시징 플랫폼은 단일 클러스터 체인의 클러스터 그룹들에 동일한 클러스터 식별자를 할당하도록 구성되는, 메시징 시스템.

청구항 15

제10 항에 있어서,

상기 메시징 플랫폼은 시간 윈도우에 걸쳐 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하고, 상기 유사성 값들에 기초하여 유사성 그래프를 생성하도록 구성되고, 상기 유사성 그래프는 상기 트렌딩 엔티티들을 표현하는 노드들 및 상기 유사성 값들로 어노테이트된 에지들을 포함하는, 메시징 시스템.

청구항 16

실행가능 명령들을 저장하는 비일시적 컴퓨터 판독가능 매체로서,

상기 명령들은 적어도 하나의 프로세서에 의해 실행될 때, 상기 적어도 하나의 프로세서로 하여금,

메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하게 하고;

상기 메시지들의 스트림으로부터 이벤트를 검출하게 하고, 상기 검출하게 하는 것은,

제1 시간 기간에 걸쳐 상기 메시지들의 스트림으로부터 복수의 트렌딩 엔티티들을 식별하게 하고;

상기 제1 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들로부터 제1 클러스터 그룹을 검출하게 하고;

제2 시간 기간에 걸쳐 상기 메시지들의 스트림으로부터 복수의 트렌딩 엔티티들을 식별하게 하고;

상기 제2 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들로부터 제2 클러스터 그룹을 검출하게 하고;

상기 제2 클러스터 그룹을 상기 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하게 하는 것 - 상기 클러스터 체인은 상기 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함 - 을 포함하고;

클라이언트 애플리케이션의 사용자 인터페이스에서 상기 이벤트에 관한 정보를 렌더링하기 위해 상기 클라이언트 애플리케이션에 디지털 데이터를 송신하게 하도록 구성되고, 상기 이벤트에 관한 정보는 상기 클러스터 체인으로부터의 정보를 포함하고, 상기 클러스터 체인으로부터의 정보는 상기 제1 클러스터 그룹으로부터의 제1 트렌딩 엔티티 및 상기 제2 클러스터 그룹으로부터의 제2 트렌딩 엔티티를 식별하는, 비일시적 컴퓨터 판독가능 매체.

청구항 17

제16 항에 있어서,

각각의 개개의 클러스터 그룹과 연관된 인기도 메트릭에 기초하여 상기 제1 클러스터 그룹 및 상기 제2 클러스터 그룹을 랭킹하는 것을 더 포함하는, 비일시적 컴퓨터 판독가능 매체.

청구항 18

제16 항에 있어서,

상기 메시지들의 스트림으로부터 엔티티들을 추출하는 것 - 상기 엔티티들은 명명된 엔티티들 또는 해시 태그들 중 적어도 하나를 포함함 -;

서버 통신 인터페이스를 통해 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트를 획득하는 것; 및

비-트렌딩 엔티티들이 상기 추출된 엔티티들로부터 필터링되도록, 상기 트렌딩 엔티티들의 리스트에 기초하여 상기 추출된 엔티티들로부터 상기 제1 시간 기간에 걸쳐 상기 복수의 트렌딩 엔티티들을 식별하는 것을 더 포함하는, 비일시적 컴퓨터 판독가능 매체.

청구항 19

제16 항에 있어서,

상기 제1 클러스터 그룹에 클러스터 식별자를 할당하는 것; 및

상기 제2 클러스터 그룹이 상기 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 상기 제1 클러스터 그룹의 상기 클러스터 식별자를 상기 제2 클러스터 그룹에 할당하는 것을 더 포함하는, 비일시적 컴퓨터 판독가능 매체.

청구항 20

제16 항에 있어서,

시간 윈도우에 걸쳐 상기 복수의 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 것 - 각각의 유사성 값은 2개의 트렌딩 엔티티들 사이의 유사성의 레벨을 표시함 -;

상기 유사성 값들에 기초하여 유사성 그래프를 생성하는 것 - 상기 유사성 그래프는 상기 복수의 트렌딩 엔티티들을 표현하는 노드들 및 상기 유사성 값들로 어노테이트된 에지들을 포함함 -;

상기 유사성 임계치 미만의 유사성 값들을 갖는 에지들이 상기 유사성 그래프로부터 제거되도록 유사성 임계치 값에 기초하여 상기 유사성 그래프를 필터링하는 것; 및

상기 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 상기 필터링된 유사성 그래프를 파티셔닝하는 것을 더 포함하고, 상기 클러스터링 알고리즘은 루뱅(Louvain) 알고리즘을 포함하는, 비일시적 컴퓨터 판독가능 매체.

청구항 21

제16 항에 있어서,

상기 제1 시간 기간 및 상기 제2 시간 기간에 걸친 상기 복수의 트렌딩 엔티티들은 버스트 검출기에 의해 상기

메시지들의 스트림으로부터 식별되고, 클러스터 체인 검출기에 의해 상기 제1 및 제2 클러스터 그룹들이 검출되고 상기 클러스터 체인이 생성되고,

상기 버스트 검출기의 컴퓨터 리소스들을 조정하는 것; 및

상기 버스트 검출기의 상기 컴퓨터 리소스들의 조정과 독립적으로 상기 클러스터 체인 검출기의 컴퓨터 리소스들을 조정하는 것을 더 포함하는, 비일시적 컴퓨터 판독가능 매체.

청구항 22

제21 항에 있어서,

상기 버스트 검출기의 하나 이상의 동작들은 상기 클러스터 체인 검출기의 하나 이상의 동작들과 병렬로 수행되는, 비일시적 컴퓨터 판독가능 매체.

청구항 23

소셜 미디어 스트림 상에서 실시간 이벤트들을 검출하기 위한 메시징 시스템으로서, 상기 메시징 시스템은,

네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼을 포함하고, 상기 메시징 플랫폼은 오프라인 모드 및 온라인 모드에서 실행되도록 구성된 이벤트 검출기를 포함하고,

상기 이벤트 검출기는 상기 오프라인 모드에서,

제어 파라미터의 가변 값들에 대한 하나 이상의 제1 클러스터 체인들을 생성하기 위해 평가 데이터세트 스트림에 대해 이벤트 검출 알고리즘을 실행하고;

상기 제어 파라미터의 상기 가변 값들에 대한 상기 이벤트 검출 알고리즘의 실행에 관한 성능 메트릭을 컴퓨팅하도록 구성되고, 상기 제어 파라미터의 값은 상기 성능 메트릭에 기초하여 선택되고,

상기 이벤트 검출기는 상기 온라인 모드에서,

상기 메시징 플랫폼 상에서 실시간으로 교환되는 메시지들에 대한 메시지 스트림을 수신하고;

하나 이상의 제2 클러스터 체인들을 생성하기 위해 상기 제어 파라미터의 선택된 값에 따라 상기 메시지 스트림에 대해 상기 이벤트 검출 알고리즘을 실행하도록 구성되는, 메시징 시스템.

청구항 24

제23 항에 있어서,

상기 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 상기 제어 파라미터는 유사성 임계치를 포함하는, 메시징 시스템.

청구항 25

제23 항에 있어서,

상기 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 상기 제어 파라미터는 클러스터링 알고리즘의 분해능을 포함하는, 메시징 시스템.

발명의 설명

기 술 분 야

[0001] [0001] 본 출원은, 2019년 1월 29일에 출원되고 발명의 명칭이 "Real-Time Event Detection on Social Data Streams"인 미국 가출원 제62/798,388호를 우선권으로 주장하며, 이 가출원의 개시는 그 전체가 본원에 통합된다.

배 경 기 술

[0002] [0002] 소셜 네트워크들은 신속하게 실세계 이벤트들 주위에서 일어나고 있는 것을 논의하기 위한 주요 매체가 되고 있다. 일부 종래의 접근법들은 주어진 시점, 예컨대 시점 분석에서 소셜 미디어 플랫폼 상에서 어떤 이벤

트들이 논의되는지를 결정하는 것을 수반할 수 있다. 그러나, 이러한 종래의 접근법들은 이벤트의 진화를 고려하는 데 실패할 수 있고, 대규모 소셜 미디어 네트워크에서 종종 생성되는 높은 볼륨 데이터세트들 또는 토픽 카운트들을 갖는 메모리 문제들에 취약할 수 있다.

발명의 내용

[0003] [0003] 시간에 걸친 이벤트들을 추적하는 이벤트 검출을 위한 기법들, 방법들 및 시스템들이 본원에서 개시되며, 이는, 클러스터 체인으로서 이벤트, 예컨대 시간에 걸쳐 링크된 클러스터 그룹들의 리스트를 모델링하는 것을 포함한다. 예컨대, 메시징 시스템은 메시징 플랫폼 상에서 교환되는 메시지들의 메시지 스트림으로부터 트렌딩 엔티티들을 주기적으로 식별할 수 있다. 메시지 스트림으로부터의 트렌딩 엔티티들의 식별은 버스트 검출로 지칭될 수 있다. 일부 예들에서, 메시지 스트림은 비교적 크다(예컨대, 초당 5K 메시지들). 트렌딩 엔티티는 비정상적으로 높은 레이트로 또는 임계 조건을 초과하는 레이트로 메시지 스트림에 나타나는 엔티티일 수 있다. 일부 예들에서, 트렌딩 엔티티는 단어(또는 단어들), 어구, 해시 태그, 식별자(예컨대, 사용자 식별자, 메시지 식별자 등), 웹 리소스(예컨대, URL), 및/또는 특정 객체를 참조하는 임의의 콘텐츠일 수 있다. 메시징 시스템은 하나 이상의 클러스터 그룹들을 검출하기 위해 트렌딩 엔티티들에 대해 유사성-기반 클러스터링 동작들을 주기적으로 수행할 수 있으며, 여기서 각각의 클러스터 그룹은 서로 유사한(예컨대, 임계치 레벨 초과의 유사성 값과 연관된) 것으로 결정되는 2개 이상의 트렌딩 엔티티들을 포함한다.

[0004] [0004] 특히, 메시징 시스템은 제1 시간 기간으로부터 검출된 트렌딩 엔티티들을 수신하고, 하나 이상의 클러스터 그룹들을 검출하기 위해 제1 시간 기간으로부터의 트렌딩 엔티티들에 대해 유사성-기반 클러스터링 동작들을 수행할 수 있다. 그 다음, 메시징 시스템은 제2 시간 기간으로부터 검출된 트렌딩 엔티티들을 수신하고, 하나 이상의 후속 클러스터 그룹들을 검출하기 위해 제2 시간 기간으로부터의 트렌딩 엔티티들에 대해 유사성-기반 클러스터링 동작들을 수행할 수 있다. 제2 시간 기간으로부터의 클러스터 그룹의 주제가 제1 시간 기간으로부터의 클러스터 그룹의 주제와 유사한 것으로 결정되면, 이들 클러스터 그룹들은 (상이한 시간 기간들에 걸쳐) 함께 링크되어, 클러스터 체인을 형성한다. 시간에 걸쳐 (그리고 실시간으로 또는 실질적으로 실시간으로) 클러스터 그룹들을 연속적으로(예컨대, 주기적으로) 검출 및 링크시킴으로써, 이벤트의 수명 동안 상이한 지점들에서 메시징 시스템 상의 이벤트를 설명하기 위해 사용되는 용어들은 클러스터 체인에 의해 캡처될 수 있다. 따라서, 정적 이벤트 검출(예컨대, 데이터의 스냅샷으로부터 이벤트들을 검출)을 사용하는 일부 종래의 접근법들과는 대조적으로, 메시징 시스템에 의해 제공되는 이벤트 검출은 메시징 시스템에 대해 논의되는 것의 동적 성질을 설명하기 위해 동적이다.

[0005] [0005] 일부 예들에서, 버스트 검출 동작들은 클러스터링 동작들과 독립적으로(예컨대, 별개의 CPU들 및 메모리 디바이스들에 의해 실행가능함) 실행된다. 버스트 검출 동작들을 클러스터링 동작들로부터 분리함으로써, 클러스터링 동작들 중 일부(또는 이들 모두)는 버스트 검출 동작들과 병렬로 실행될 수 있으며, 이는 (특히, 많은 양의 데이터를 처리하는 소셜 미디어 플랫폼들에 대해) 이벤트 검출의 속도를 증가시킬 수 있다. 또한, 이러한 컴포넌트들은 독립적으로 스케일링가능하고, 이는 이벤트 검출의 속도를 증가시키고, 버스트 검출 및 엔티티 클러스터링의 다양한 프로세싱 부하들에 적응하기 위해 그리고 실시간으로 이벤트들의 추적을 가능하게 하기 위해 메시징 시스템의 유연성을 증가시킬 수 있다.

[0006] [0006] 메시징 플랫폼은, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트(들)에 관한 정보를 렌더링하기 위해 디지털 데이터를 클라이언트 애플리케이션에 송신할 수 있다. 이벤트(들)에 관한 정보는 클러스터 체인들로부터의 정보를 포함할 수 있다. 클러스터 체인 정보는 하나 이상의 이벤트들을 식별하고 클러스터 체인으로부터 하나 이상의 트렌딩 엔티티들을 식별할 수 있다. 일부 예들에서, 클러스터 체인 정보는 클라이언트 애플리케이션의 트렌드 섹션에 포함된다. 예를 들어, 트렌드 섹션은 트렌딩 엔티티들(또는 토픽들)의 리스트를 식별할 수 있다. 하나의 특정 예에서, 해시 태그 "#bucks"는 트렌딩일 수 있고 클러스터 체인에 의해 표현되는 이벤트의 일부로서 포함될 수 있다. 클러스터 체인은 또한, 해시 태그 "#bucks"를 포함하는 클러스터 그룹에 링크된 다른 클러스터 그룹 내의 엔티티 "Giannis"를 식별할 수 있다. 트렌드 섹션은 트렌딩 엔티티 "#bucks"와 함께 관련 용어 "Giannis"(및 클러스터 체인으로부터의 다른 관련 용어들)를 식별할 수 있다.

[0007] [0007] 메시징 플랫폼은 메시지들의 타임라인을 렌더링하기 위해 클라이언트 애플리케이션에 디지털 데이터를 송신할 수 있다. 타임라인은 연결 그래프에서 사용자의 계정과 관계들을 갖는 계정들로부터의 메시지들의 스트림을 포함할 수 있다. 일부 예들에서, 타임라인은 랭킹되고, 메시지들의 랭킹은 검출된 이벤트에 (부분적으로) 기초할 수 있다. 예를 들어, 이벤트는 이벤트의 검출된 시작 시간 및 이벤트의 검출된 종료 시간을 식별하는 이벤트 메타데이터를 포함할 수 있다. 타임라인 관리자는 그 이벤트(또는 클러스터 체인)에 속하는 상이한 클

러스터 그룹들에 걸쳐 트렌딩 엔티티들을 식별하는 이벤트를 수신할 수 있다. 타임라인 관리자는 사용자의 타임라인의 일부로서 렌더링될 메시지가 이벤트의 지속기간 동안(예컨대, 검출된 시작 시간과 검출된 종료 시간 사이에) 클러스터 체인으로부터 트렌딩 엔티티를 포함하는지 여부를 결정할 수 있다. 그 메시지가 이벤트의 지속기간 동안 클러스터 체인의 일부인 트렌딩 엔티티를 포함하면, 타임라인 관리자는 사용자의 타임라인의 랭킹 내에서 그 메시지를 부스팅(또는 업-랭크)할 수 있다.

[0008] [0008] 일부 예들에서, 타임라인은 광고 메시지들을 포함할 수 있는 촉진된 콘텐츠를 포함한다. 연결 그래프에 따라 전달될 메시지들과 유사하게, 그 촉진된 메시지가 이벤트의 지속기간 동안 클러스터 체인으로부터의 트렌딩 엔티티를 포함하면, 촉진된 메시지는 타임라인의 랭킹에서 부스팅될 수 있다. 일부 예들에서, 메시징 플랫폼은 촉진된 메시지들에 대한 가격 책정을 결정하도록 구성된 광고 스택 엔진을 포함한다. 일부 예들에서, 촉진된 콘텐츠가 이벤트의 지속기간 동안 하나 이상의 트렌딩 엔티티들을 포함하면, 광고 스택 엔진은 촉진된 콘텐츠에 대한 자신의 가격 책정을 증가시킬 수 있다. 일부 예들에서, 클러스터 체인 정보는 사용자에게 리턴된 검색 결과들의 일부로서 포함된다. 예컨대, 사용자는 질의 검색을 제출할 수 있고, 검색 관리자는 다른 관련된 엔티티들을 포함하도록 검색 결과들을 확장시키기 위해 클러스터 체인들을 사용할 수 있다.

[0009] [0009] 일 양상에 따르면, 소셜 데이터 스트림들에 대한 이벤트 검출을 위한 방법은 메시징 플랫폼에 의해, 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하는 단계, 및 메시징 플랫폼에 의해, 메시지들의 스트림으로부터 이벤트를 검출하는 단계를 포함한다. 검출하는 단계는 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하는 단계, 제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하는 단계, 및 제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하는 단계를 포함하고, 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현한다. 제1 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다. 제2 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다. 방법은 메시징 플랫폼에 의해, 이벤트를 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 저장하는 단계를 포함한다. 일부 예들에서, 시스템 또는 비일시적 컴퓨터 판독가능 매체에는 이러한 동작들이 제공될 수 있다.

[0010] [0010] 일부 양상들에 따르면, 방법, 시스템, 또는 비일시적 컴퓨터 판독가능 매체는 다음의 특징들(또는 이들의 임의의 조합) 중 하나 이상을 포함할 수 있다. 방법은 메시징 플랫폼에 의해, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 클라이언트 애플리케이션에 디지털 데이터를 송신하는 단계를 포함할 수 있고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함한다. 이벤트에 관한 정보는 제1 클러스터 그룹으로부터의 제1 트렌딩 엔티티 및 제2 클러스터 그룹으로부터의 제2 트렌딩 엔티티를 식별하고, 제2 트렌딩 엔티티는 제1 트렌딩 엔티티와는 상이하다. 방법은 각각의 개개의 클러스터 그룹과 연관된 엔티티들의 인기도에 기초하여 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하는 것을 포함하여, 클러스터 체인 내의 클러스터 그룹들을 랭킹하는 단계를 포함할 수 있다. 방법은 메시지들의 스트림으로부터 엔티티들을 추출하는 단계, 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트를 획득하는 단계, 및 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트에 기초하여 추출된 엔티티들로부터 트렌딩 엔티티들을 식별하는 단계를 포함할 수 있다. 방법은 제1 클러스터 그룹에 클러스터 식별자를 할당하는 단계, 및 제2 클러스터 그룹이 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 제1 클러스터 그룹의 클러스터 식별자를 제2 클러스터 그룹에 할당하는 단계를 포함할 수 있다. 제1 클러스터 그룹을 검출하는 것은 제1 시간 기간의 트렌딩 엔티티들과 연관된 유사성 값들에 기초하여 유사성 그래프를 생성하는 것 - 유사성 그래프는 제1 시간 기간에 걸친 트렌딩 엔티티들을 표현하는 노드들 및 유사성 값들을 표현하는 에지들을 포함함 -, 및 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 유사성 그래프를 파티셔닝하는 것을 포함할 수 있다. 방법은 시간 윈도우에 걸친 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 단계, 및 유사성 임계치 미만의 유사성 값들을 갖는 에지들이 유사성 그래프로부터 제거되도록 유사성 임계치에 기초하여 유사성 그래프를 필터링하는 단계를 포함할 수 있고, 필터링된 유사성 그래프는 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 파티셔닝된다. 제2 클러스터 그룹은 최대 가중된 이분(bipartite) 매칭에 기초하여 제1 클러스터 그룹에 링크될 수 있다.

[0011] [0011] 일 양상에 따르면, 실시간 이벤트를 검출하기 위한 메시징 시스템은 네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼, 및 메시지들을 전송 및 수신하기 위해 메시징 플랫폼과 통신하도록 구성된 클라이언트 애플리케이션을 포함한다. 메시징 플랫폼은, 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하고, 제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하고, 제1 클러스터 그룹과 제2 클러스터 그룹 사이에서 공유되는 트렌딩 엔티티들의 수에 기초하여 제2 클러스터 그룹

을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하고 – 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함 –, 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 이벤트를 저장하도록 구성되고, 클러스터 체인은 장래의 클러스터 링크를 위해 리트리브가능하다. 일부 예들에서, 방법 또는 비일시적 컴퓨터 판독가능 매체에는 이러한 동작들이 제공될 수 있다.

[0012] [0012] 일부 양상들에 따르면, 방법, 시스템, 또는 비일시적 컴퓨터 판독가능 매체는 위의/아래의 특징들(또는 이들의 임의의 조합) 중 하나 이상을 포함할 수 있다. 메시징 플랫폼은, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 디지털 데이터를 클라이언트 애플리케이션에 송신하도록 구성되고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함하고, 클러스터 체인으로부터의 정보는 트렌드 섹션, 타임라인, 또는 클라이언트 애플리케이션에 리턴되는 검색 결과들의 일부에서 렌더링된다. 메시징 플랫폼은 각각의 클러스터 그룹과 연관된 집계 인기도 메트릭에 기초하여 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하도록 구성된다. 복수의 엔티티들은 트렌딩 엔티티들을 포함하고, 메시징 플랫폼은 메시지들의 스트림으로부터 엔티티들을 추출하고, 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트를 획득하고, 트렌딩 엔티티들의 리스트에 기초하여 추출된 엔티티들로부터의 트렌딩 엔티티들을 식별하도록 구성된다. 메시징 플랫폼은 단일 클러스터 체인의 클러스터 그룹들에 동일한 클러스터 식별자를 할당하도록 구성된다. 메시징 플랫폼은 시간 윈도우에 걸쳐 복수의 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하고, 유사성 값들에 기초하여 유사성 그래프를 생성하도록 구성되고, 여기서 유사성 그래프는 복수의 엔티티들을 표현하는 노드들 및 유사성 값들을 표현하거나 그에 의해 어노테이트된 에지들을 포함한다.

[0013] [0013] 일 양상에서, 비일시적 컴퓨터 판독가능 매체는 실행가능 명령들을 저장하고, 명령들은 적어도 하나의 프로세서에 의해 실행될 때, 적어도 하나의 프로세서로 하여금, 메시징 플랫폼에 의해, 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하게 하고, 메시징 플랫폼에 의해, 메시지들의 스트림으로부터 이벤트를 검출하게 하고, 이는 메시지들의 스트림으로부터 복수의 엔티티들을 식별하는 것, 제1 시간 기간에 걸쳐 복수의 엔티티들로부터 제1 클러스터 그룹을 검출하는 것, 제2 시간 기간에 걸쳐 복수의 엔티티들로부터 제2 클러스터 그룹을 검출하는 것, 제1 클러스터 그룹과 제2 클러스터 그룹 사이에서 공유되는 엔티티들의 수에 기초하여 제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하는 것 – 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함 –, 메시징 플랫폼에 의해, 이벤트를 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 저장하는 것을 포함하고, 메시징 플랫폼에 의해, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 클라이언트 애플리케이션에 디지털 데이터를 송신하게 하도록 구성되고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함하고, 클러스터 체인으로부터의 정보는 제1 클러스터 그룹으로부터의 제1 엔티티 및 제2 클러스터 그룹으로부터의 제2 엔티티를 식별한다. 일부 예들에서, 방법 또는 시스템에는 이러한 동작들이 제공될 수 있다.

[0014] [0014] 일부 양상들에 따르면, 방법, 시스템, 또는 비일시적 컴퓨터 판독가능 매체는 위의/아래의 특징들(또는 이들의 임의의 조합) 중 하나 이상을 포함할 수 있다. 동작들은 각각의 개개의 클러스터 그룹과 연관된 인기도 메트릭에 기초하여 상기 제1 클러스터 그룹 및 상기 제2 클러스터 그룹을 랭킹할 수 있다. 복수의 엔티티들은 트렌딩 엔티티들을 포함하고, 동작들은, 메시지들의 스트림으로부터 엔티티들을 추출하는 것 – 엔티티들은 명명된 엔티티들 또는 해시 태그들 중 적어도 하나를 포함함 –, 서버 통신 인터페이스를 통해 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트를 획득하는 것, 및 비-트렌딩 엔티티들이 추출된 엔티티들로부터 필터링되도록, 트렌딩 엔티티들의 리스트에 기초하여 추출된 엔티티들로부터 트렌딩 엔티티들을 식별하는 것을 포함할 수 있다. 동작들은 제1 클러스터 그룹에 클러스터 식별자를 할당하는 것, 및 제2 클러스터 그룹이 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 제1 클러스터 그룹의 클러스터 식별자를 제2 클러스터 그룹에 할당하는 것을 포함할 수 있다. 동작들은 시간 윈도우에 걸쳐 복수의 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 것 – 각각의 유사성 값은 2개의 엔티티들 사이의 유사성의 레벨을 표시함 –, 유사성 값들에 기초하여 유사성 그래프를 생성하는 것 – 유사성 그래프는 복수의 엔티티들을 표현하는 노드들 및 유사성 값들을 표현하는 에지들을 포함함 –, 유사성 임계치 미만의 유사성 값들을 갖는 에지들이 유사성 그래프로부터 제거되도록 유사성 임계치 값에 기초하여 유사성 그래프를 필터링하는 것, 및 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 필터링된 유사성 그래프를 파티셔닝하는 것을 포함할 수 있고, 클러스터링 알고리즘은 루뱅(Louvain) 알고리즘을 포함한다.

[0015] [0015] 일 양상에 따르면, 소셜 미디어 스트림 상에서 실시간 이벤트를 검출하기 위한 메시징 시스템은 네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼을 포함한다. 메시징 플랫폼은 오프라인 모드 및 온라인 모드에서 실행되도록 구성된 이벤트 검출기를 포함한다. 이벤트 검출기는 오프라인 모

드에서, 제어 파라미터의 가변 값들에 대한 하나 이상의 제1 클러스터 체인들을 생성하기 위해 평가 데이터세트 스트림에 대해 이벤트 검출 알고리즘을 실행하고, 제어 파라미터의 가변 값들에 대한 이벤트 검출 알고리즘의 실행에 관한 성능 메트릭을 컴퓨팅하도록 구성되고, 제어 파라미터의 값은 성능 메트릭에 기초하여 선택된다. 이벤트 검출기는 온라인 모드에서, 메시징 플랫폼 상에서 실시간으로 교환되는 메시지들에 대한 메시지 스트림을 수신하고, 하나 이상의 제2 클러스터 체인들을 생성하기 위해 제어 파라미터의 선택된 값에 따라 메시지 스트림에 대해 이벤트 검출 알고리즘을 실행하도록 구성된다. 일부 예들에서, 방법 또는 비일시적 컴퓨터 판독가능 매체에는 이러한 동작들이 제공될 수 있다.

[0016] [0016] 일부 양상들에 따르면, 방법, 시스템, 또는 비일시적 컴퓨터 판독가능 매체는 위의/아래의 특징들(또는 이들의 임의의 조합) 중 하나 이상을 포함할 수 있다. 일부 예들에서, 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 제어 파라미터는 유사성 임계치를 포함한다. 일부 예들에서, 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 제어 파라미터는 클러스터링 알고리즘의 분해능을 포함한다. 일부 예들에서, 성능 메트릭은 클러스터링 스코어를 포함한다. 일부 예들에서, 성능 메트릭은 이벤트 검출 비율을 포함한다. 일부 예들에서, 성능 메트릭은 병합된 이벤트 비율을 포함한다. 일부 예들에서, 성능 메트릭은 복제 이벤트 비율을 포함한다. 일부 예들에서, 제어 파라미터는 시간 윈도우를 포함한다.

도면의 간단한 설명

[0017] [0017] 도 1a는 일 양상에 따른, 시간에 걸쳐 클러스터 체인으로서 이벤트를 검출 및 추적하도록 구성된 이벤트 검출기를 갖는 메시징 시스템을 예시한다.

[0018] [0018] 도 1b는 일 양상에 따른, 시간에 걸쳐 복수의 클러스터 그룹들을 갖는 클러스터 체인에 의해 표현되는 이벤트를 예시한다.

[0019] [0019] 도 1c는 일 양상에 따른, 이벤트 검출기에 의해 추출되는 엔티티들의 다양한 예들을 예시한다.

[0020] [0020] 도 1d는 일 양상에 따른, 트렌딩 엔티티들을 표현하는 노드들 및 유사성 값들을 표현하는 에지들을 갖는 유사성 그래프를 예시한다.

[0021] [0021] 도 1e는 일 양상에 따른, 복수의 클러스터 그룹들로 파티셔닝된 필터링된 유사성 그래프를 예시한다.

[0022] [0022] 도 2는 일 양상에 따른, 클러스터 체인으로부터의 정보를 렌더링하는 클라이언트 애플리케이션의 예시적인 스크린샷을 예시한다.

[0023] [0023] 도 3은 일 양상에 따른, 이벤트 검출기의 예시적인 동작을 도시하는 프로세스 흐름을 예시한다.

[0024] [0024] 도 4는 일 양상에 따른, 오프라인 분석 모드 내의 이벤트 검출기를 예시한다.

[0025] [0025] 도 5a 내지 도 5f는 일 양상에 따른, 오프라인 분석 모드에서 이벤트 검출기의 성능 메트릭들 및 파라미터들을 도시하는 다양한 그래프들을 예시한다.

[0026] [0026] 도 6a 내지 도 6c는 일 양상에 따른, 온라인 분석 모드에서 이벤트 검출기의 성능을 도시하는 다양한 그래프들을 예시한다.

[0027] [0027] 도 7은 일 양상에 따른, 온라인 분석 모드에서 이벤트 검출기의 성능을 도시하는 그래프들를 예시한다.

[0028] [0028] 도 8은 일 양상에 따른, 실세계 이벤트에 대한 복수의 클러스터 체인들의 예를 예시한다.

[0029] [0029] 도 9는 일 양상에 따른, 시간에 걸친 클러스터 체인의 세부 사항들을 예시한다.

[0030] [0030] 도 10은 일 양상에 따른, 메시징 시스템의 예시적인 동작들을 도시하는 흐름도를 예시한다.

발명을 실시하기 위한 구체적인 내용

[0018] [0031] 도 1a 내지 도 1e는 일 양상에 따라 메시지 스트림(108)으로부터 시간에 걸쳐 이벤트들(112)을 검출 및 추적하기 위한 메시징 시스템(100)을 예시한다. 예컨대, 시점 분석을 사용하여 특정 시간에 메시징 시스템(100) 상에서 어떤 이벤트들(112)이 논의되는지를 결정하는 대신에, 메시징 시스템(100)은 클러스터 그룹들의 링크된 리스트(144)를 갖는 클러스터 체인(114)으로서 이벤트(112)를 모델링함으로써 시간에 걸친 이벤트들(112)을 표현한다.

[0019] [0032] 메시징 시스템(100)은 메시지 스트림(108)으로부터 트렌딩 엔티티들(132b)을 주기적으로 식별할 수

있다. 트렌딩 엔티티(132b)는 비정상적으로 높은 레이트로 또는 임계 조건을 초과하는 레이트로 메시지 스트림(108)에 나타나는 엔티티일 수 있으며, 여기서 트렌딩 엔티티(132b)는 메시지 스트림(108)의 메시지 내의 콘텐츠에 대한 태그이다(예컨대, 명명된 엔티티들, 해시 태그들, URL들, 사용자 식별자들 및/또는 메시지 식별자들 등). 메시징 시스템(100)은 하나 이상의 클러스터 그룹들(144)을 검출하기 위해 트렌딩 엔티티들(132)에 대해 유사성-기반 클러스터링 동작들을 주기적으로 수행할 수 있다. 예를 들어, 메시징 시스템(100)은 제1 시간 기간(예컨대, 주어진 시간 인터벌)으로부터 검출된 트렌딩 엔티티들(132b)을 수신하고, 하나 이상의 클러스터 그룹들(144)을 검출하기 위해 제1 시간 기간으로부터의 트렌딩 엔티티들(132b)에 대해 유사성-기반 클러스터링 동작들을 수행할 수 있다. 특정 클러스터 그룹(144)은 서로 유사하게 관련되는 것으로 결정된 2개 이상의 트렌딩 엔티티들(132b)을 포함할 수 있다(예컨대, 동일한 클러스터 그룹(144)으로부터의 트렌딩 엔티티들(132b)은 의미론적으로 동일하거나 유사한 주제를 지칭하는 것으로 고려될 수 있음). 그 다음, 메시징 시스템(100)은 제2 시간 기간(예컨대, 다음 시간 인터벌)으로부터 검출된 트렌딩 엔티티들(132b)을 수신하고, 하나 이상의 후속 클러스터 그룹들(144)을 검출하기 위해 제2 시간 기간으로부터의 트렌딩 엔티티들(132b)에 대해 유사성-기반 클러스터링 동작들을 수행할 수 있다. 제2 시간 기간으로부터의 클러스터 그룹(144)의 주제가 제1 시간 기간으로부터의 클러스터 그룹(144)의 주제와 유사한 것으로 결정되면, 이들 클러스터 그룹들(144)은 (상이한 시간 기간들에 걸쳐) 함께 링크되어, 클러스터 체인(114)을 형성한다.

[0020] [0033] 예를 들어, 도 1b를 참조하면, 클러스터 체인(114)은 트렌딩 엔티티들(132b)의 제1 클러스터 그룹(144-1)(예컨대, 제1 시간 기간(196-1) 동안 검출됨), 트렌딩 엔티티들(132b)의 제2 클러스터 그룹(144-2)(예컨대, 제2 시간 기간(196-2) 동안 검출됨), 및 트렌딩 엔티티들(132b)의 제3 클러스터 그룹(144-3)(예컨대, 제3 시간 기간(196-3) 동안 검출됨)을 포함할 수 있다. 도 1b에 도시된 바와 같이, 제2 클러스터 그룹(144-2)은 제1 클러스터 그룹(144-1)에 링크되고, 제3 클러스터 그룹(144-3)은 제2 클러스터 그룹(144-2)에 링크되며, 이는 제1 클러스터 그룹(144-1), 제2 클러스터 그룹(144-2), 제3 클러스터 그룹(144-3)이 동일한 이벤트(112)와 관련됨을 표시한다. 시간에 걸쳐 (그리고 실시간으로 또는 실질적으로 실시간으로) 클러스터 그룹들(144)을 연속적으로(예컨대, 주기적으로) 검출 및 링크시킴으로써, 이벤트(112)의 수명 동안 상이한 지점들에서 메시징 시스템(100) 상의 이벤트(112)를 설명하기 위해 사용되는 용어들은 클러스터 체인(114)에 의해 캡처될 수 있다.

[0021] [0034] 예컨대, 2011년 3월에 발생한 일본의 쓰나미의 예와 관련하여, 초기에, 이 이벤트는 지진 및 쓰나미와 같은 키워드들에 의해 지배되었지만, 나중에 핵 및 방사선과 같은 이벤트 단어들이 도입되었다. 본원에서 논의된 기술들에 따르면, 메시징 시스템(100)은 이벤트들(112)의 진화를 추적함으로써 메시징 시스템(100) 상에서 발생하는 이벤트-기반 대화들의 동적 성질을 설명할 수 있으며, 이는 이벤트(112)에 대한 지진 및 쓰나미의 초기 용어들(예컨대, 지진 및 쓰나미는 제1 클러스터 그룹(144-1)의 일부일 수 있음)뿐만 아니라 동일한 이벤트(112)와 관련하여 추후에 도입된 핵 및 방사선 용어들(예컨대, 핵 및 방사선은 제3 클러스터 그룹(144-3)의 일부일 수 있음)을 식별할 수 있을 것이다. 따라서, 정적 이벤트 검출(예컨대, 데이터의 스냅샷으로부터 이벤트들을 검출)을 사용하는 일부 종래의 접근법들과는 대조적으로, 메시징 시스템(100)에 의해 제공되는 이벤트 검출은 메시징 시스템(100)에 대해 논의되는 것의 동적 성질을 설명하기 위해 동적이다.

[0022] [0035] 메시징 시스템(100)은 서버 컴퓨터(102)에 의해 실행가능한 메시징 플랫폼(104), 및 컴퓨팅 디바이스(174)에 의해 실행가능한 클라이언트 애플리케이션(176)을 포함한다. 클라이언트 애플리케이션(176)은 네트워크(172)를 통해 메시징 플랫폼(104)의 다른 사용자들에게(그리고 그로부터) 메시지들을 전송(및 수신)하기 위해 메시징 플랫폼(104)과 통신한다. 클라이언트 애플리케이션(176)은 사용자들이 메시지들을 포스팅하고 그와 상호작용하는 소셜 미디어 메시징 애플리케이션일 수 있다. 일부 예들에서, 클라이언트 애플리케이션(176)은 컴퓨팅 디바이스(174)의 운영 시스템 상에서 실행되는 고유 애플리케이션이거나, 또는 컴퓨팅 디바이스(174)의 브라우저-기반 애플리케이션과 함께 서버 컴퓨터(102)(또는 다른 서버) 상에서 실행되는 웹-기반 애플리케이션일 수 있다. 컴퓨팅 디바이스(174)는 클라이언트 애플리케이션(176) 및 메시징 플랫폼(104)이 서로 통신할 수 있게 하는 방식으로 임의의 타입의 네트워크 연결들 및/또는 API(application programming interface)들을 사용하여 네트워크(172)를 통해 메시징 플랫폼(104)에 액세스할 수 있다.

[0023] [0036] 컴퓨팅 디바이스(174)는 모바일 컴퓨팅 디바이스(예컨대, 스마트 폰, PDA, 태블릿 또는 랩톱 컴퓨터) 또는 비-모바일 컴퓨팅 디바이스(예컨대, 데스크톱 컴퓨팅 디바이스)일 수 있다. 컴퓨팅 디바이스(174)는 또한 다양한 네트워크 인터페이스 회로, 이를 테면, 예컨대, 컴퓨팅 디바이스(174)가 셀룰러 네트워크와 통신할 수 있게 하는 모바일 네트워크 인터페이스, 컴퓨팅 디바이스(174)가 Wi-Fi 기지국과 통신할 수 있게 하는 Wi-Fi 네트워크 인터페이스, 컴퓨팅 디바이스(174)가 다른 블루투스 디바이스들과 통신할 수 있게 하는 블루투스 네트워크 인터페이스, 및/또는 컴퓨팅 디바이스(174)가 네트워크(172)에 액세스할 수 있게 하는 이더넷 연결 또는 다

른 유선 연결을 포함한다.

[0024] [0037] 서버 컴퓨터(102)는 단일 컴퓨팅 디바이스일 수 있거나, 또는 작업 부하 및 리소스들을 공유하도록 통신 가능하게 연결된 2개 이상의 분산형 컴퓨팅 디바이스들의 표현일 수 있다. 서버 컴퓨터(102)는 적어도 하나의 프로세서, 및 적어도 하나의 프로세서에 의해 실행될 때 적어도 하나의 프로세서로 하여금 본원에서 논의된 동작들을 수행하게 하는 실행가능 명령들을 저장하는 비일시적 컴퓨터 판독가능 매체를 포함할 수 있다.

[0025] [0038] 메시징 플랫폼(104)은 사용자 디바이스들(이들 중 하나가 컴퓨팅 디바이스(174)로서 도시됨) 사이의 통신(예컨대, 실시간 통신)을 용이하게 하기 위한 컴퓨팅 플랫폼이다. 메시징 플랫폼(104)은 개인들, 기업들 및/또는 엔티티들(예컨대, 가명 계정들, 신규 계정들 등)의 수백만 개의 계정들(118)을 저장할 수 있다. 각각의 계정(118)의 하나 이상의 사용자들은 메시징 플랫폼(104) 내부 및/또는 외부의 다른 계정들(118)에 메시지들을 전송하기 위해 메시징 플랫폼(104)을 사용할 수 있다. 일부 예들에서, 메시징 플랫폼(104)은 사용자들이 "실시간"으로 통신할 수 있게, 예컨대, 최소의 지연으로 다른 사용자들과 대화하고 동시 세션들 동안 하나 이상의 다른 사용자들과 대화를 수행하게 할 수 있다. 다시 말해서, 메시징 플랫폼(104)은 사용자가 메시지들을 브로드캐스트하게 할 수 있고, 사용자들 사이의 라이브 대화를 용이하게 하기 위해 합리적인 시간 프레임(예컨대, 2초 미만) 내에 하나 이상의 다른 사용자들에게 메시지들을 디스플레이할 수 있다. 일부 예들에서, 메시지의 수신자들은 메시지를 브로드캐스트하는 사용자의 계정과 연결 그래프(116)에서 미리 정의된 그래프 관계를 가질 수 있다.

[0026] [0039] 연결 그래프(116)는, 메시징 플랫폼(104) 내의 어느 계정들(118)이 특정 계정(118)과 연관되는지(예컨대, 팔로우하는지, 친구인지, 가입되는지 등) 그리고 그에 따라 특정 계정(118)으로부터 메시지들을 수신하도록 가입되는지를 표시하는 데이터 구조를 포함한다. 예를 들어, 연결 그래프(116)는 제1 계정을 제2 계정과 링크시킬 수 있으며, 이는 제1 계정이 제2 계정과 관계에 있음을 표시한다. 제2 계정의 사용자는 제1 계정의 사용자에 의해 메시징 플랫폼(104) 상에 포스팅된 메시지들을 볼 수 있다(그리고/또는 그 반대의 경우도 마찬가지임). 관계들은 단방향(예컨대, 팔로어/팔로위(followee)) 및/또는 양방향(예컨대, 친구)을 포함할 수 있다. 일부 예들에서, 연결 그래프(116)는 클라이언트 애플리케이션(176)을 설치하고 클라이언트 애플리케이션(176)을 통해 사용자 계정을 셋업한 사용자를 표현할 수 있다. 메시지들은 특정 메시징 시스템 또는 프로토콜에 의해 제한될 수 있는 다양한 길이들 중 임의의 것일 수 있다.

[0027] [0040] 일부 예들에서, 특정 사용자에 의해 저작된 메시지들을 보는 데 관심이있는 사용자들은 특정 사용자를 팔로우하도록 선택할 수 있다. 제1 사용자는 제2 사용자를 제1 사용자가 팔로우하기를 원하는 사용자로서 식별함으로써 제2 사용자를 팔로우할 수 있다. 제1 사용자가 제2 사용자를 팔로우하기를 원한다고 표시한 후에, 연결 그래프(116)는 관계를 반영하도록 업데이트되고, 제1 사용자는 제2 사용자에 의해 저작된 메시지들을 제공받을 것이다. 사용자들은 다수의 사용자들을 팔로우하도록 선택할 수 있다. 사용자들은 또한 메시지들에 응답할 수 있고, 이로써 서로 대화할 수 있다. 또한, 사용자들은 메시지들에 참여할 수 있는데, 이를 테면, 그들의 팔로워들과 메시지를 공유하거나(예를 들어, 재공유) 또는 그들의 팔로워들과 참여가 공유되는 메시지를 선호(또는 "좋아요")할 수 있다.

[0028] [0041] 메시징 플랫폼(104)은 클라이언트 애플리케이션(176)의 사용자 인터페이스(178) 상에서 소셜 콘텐츠의 타임라인(180)을 렌더링하기 위해 (네트워크(172)를 통해) 데이터를 생성 및 송신하도록 구성된 타임라인 관리자(163)를 포함한다. 일부 예들에서, 타임라인(180)은 시간순(또는 역-시간순)으로 메시지들의 리스트(예컨대, 랭킹된 리스트)를 포함한다. 메시지들의 리스트는 연결 그래프(116)에서 클라이언트 애플리케이션(176)의 사용자와의 관계를 갖는 메시징 플랫폼(104) 상에 사용자들에 의해 포스팅된 메시지들을 포함할 수 있다.

[0029] [0042] 메시징 플랫폼(104)은 클라이언트 애플리케이션(176)의 사용자 인터페이스(178)에서 하나 이상의 트렌드들(182)(예컨대, 트렌드 섹션으로 또한 지칭됨)을 렌더링하기 위해 (네트워크(172)를 통해) 데이터를 생성 및 송신하도록 구성된 트렌드 관리자(106)를 포함한다. 일부 예들에서, 트렌드 관리자(106)는 트렌딩 엔티티들(132a)을 식별하기 위해 트렌드 검출기 서비스(128)와 통신한다. 트렌드 검출기 서비스(128) 및 트렌딩 엔티티들(132a)은 본 개시에서 나중에 추가로 설명된다. 그러나, 간략하게, 트렌드 검출기 서비스(128)는 메시징 플랫폼(104) 상의 트렌딩 엔티티들(132a)을 식별하기 위한 알고리즘을 실행할 수 있고, 트렌드 관리자(106)는 트렌딩 엔티티들(132)의 리스트를 획득하고 리스트(또는 이들의 일부)를 클라이언트 애플리케이션(176)에 제공할 수 있고, 이는 사용자 인터페이스(178)의 섹션 상에서 트렌드(182)로서 렌더링된다. 일부 예들에서, 트렌드들(182)은 클라이언트 애플리케이션(176)의 사용자에 대해 맞춤화되며, 이는 사용자가 팔로우하는 계정들(118), 사용자의 관심사들 및/또는 사용자의 위치에 기초할 수 있다. 일부 예들에서, 트렌드 관리자(106)는 사용자가

메시징 플랫폼(104) 상에서 새로운 논의 토픽들을 발견하는 것을 돕기 위해 현재 대중적인 토픽들을 식별할 수 있다.

[0030] [0043] 메시징 플랫폼(104)은, 검색 질의를 실행하고 클라이언트 애플리케이션(176)이 클라이언트 애플리케이션(176)의 사용자 인터페이스(178) 상에서 검색 결과들(186)을 렌더링할 수 있게 하는 정보를 (네트워크(172)를 통해) 송신할 수 있는 검색 관리자(161)를 포함한다. 예를 들어, 검색 관리자(161)는 메시지들, 사람들, 해시태그들, 토픽들, 사진들, 비디오들 등을 찾기 위해 메시지 플랫폼을 검색하기 위해 클라이언트 애플리케이션(176)으로부터 질의 요청을 수신할 수 있다. 질의 요청에 대한 응답으로, 검색 관리자(161)는 클라이언트 애플리케이션(176)의 사용자 인터페이스(178) 상에서 (질의 요청의 하나 이상의 용어들에 맞춤화된) 검색 결과들(186)을 제공할 수 있다.

[0031] [0044] 메시징 플랫폼(104)은 메시지 스트림(108)을 수신하고 메시지 스트림(108) 내의 메시지들에 기초하여 하나 이상의 이벤트들(112)을 검출하도록 구성된 이벤트 검출기(120)를 포함한다. 일부 예들에서, 이벤트(112)는 메시징 플랫폼(104) 상에서 논의되고 있는 중요한 것을 반영할 수 있고, 사람들의 그룹이 문제에 대해 정상 대화 레벨과는 상이한 크기로 그에 대해 얘기하고 있을 때 중요한 것이 발생한다(예를 들어, 그것이 트렌딩이다). 일부 예들에서, 이벤트(112)는 메시징 플랫폼(104) 상에서 논의되는 실세계 이벤트를 표현한다. 일부 예들에서, 이벤트 검출기(120)는 시간에 걸쳐 특정 이벤트(112)를 링크된 클러스터 그룹들(144)의 클러스터 체인(114)으로서 검출하기 위해 유사성-기반 시간 이벤트 검출 알고리즘(예컨대, 아래에서 제공되는 알고리즘 1 참조)을 실행한다. 이벤트 검출기(120)는 새로운 클러스터 그룹들(144)로 클러스터 체인(114)을 업데이트함으로써 시간에 걸쳐 이벤트(112)를 추적할 수 있다. 따라서, 이벤트(112)는 시간에 걸쳐 상이하게 표현될 수 있고, 클러스터 체인(114)으로서의 이벤트(112)의 모델링은 사건이 많은 대화가 시간에 걸쳐 변할 수 있다는 사실을 반영한다. 검출된 이벤트(112)는 클러스터 체인(114)에 대응하고, 특정 시점에 트렌딩 엔티티들(132b)의 클러스터 그룹(144)에 의해 특성화된다.

[0032] [0045] 일부 예들에서, 메시지 스트림(108)은 메시징 플랫폼(104) 상에서 생성된 모든 메시지들의 큰(예컨대, 매우 큰) 스트림이다. 일부 예들에서, 사용자가 메시징 플랫폼(104)에 메시지를 포스팅함에 따라, 그 메시지는 메시지 스트림(108)에 추가된다. 일부 예들에서, 메시지 스트림(108)은 초당 5k개의 메시지들에 걸친 레이트로 전달된 메시지를 포함한다. 하루의 과정에 걸쳐, 일부 예들에서, 메시지 스트림(108)은 5억개 초과의 메시지들을 포함할 수 있다. 일부 예들에서, 메시지 스트림(108)은 초당 10k개의 메시지들에 걸친 레이트로 전달된 메시지들을 포함한다. 일부 예들에서, 메시지 스트림(108)은 초당 25k개의 메시지들에 걸친 레이트로 전달된 메시지들을 포함한다. 일부 예들에서, 메시지 스트림(108)은 초당 50k개의 메시지들에 걸친 레이트로 전달된 메시지들을 포함한다. 일부 예들에서, 메시지 스트림(108)은 메시징 플랫폼(104) 상에 생성 및 포스팅된 메시지들에 대한 메시지 생성 이벤트들, 메시징 플랫폼(104) 상에서 재-공유되는 기존 메시지들에 대한 메시지 재-공유 이벤트들, 및/또는 메시징 플랫폼(104) 상에서 선호되거나 링크되는 기존 메시지들에 대한 참여 이벤트들을 포함한다. 일부 예들에서, 메시징 플랫폼(104) 상에서 교환되는 특정 메시지는 문자 제한을 갖는다. 일부 예들에서, 문자 제한은 280자의 문자들이다.

[0033] [0046] 소셜 메시징 플랫폼들의 다양한 팩터들 및 특징들은 (사건들이 발생하고 있을 때) 이벤트들의 검출을 비교적 어렵게 할 수 있다. 예컨대, 메시지 스트림(108)의 스케일은 비교적 크다(예컨대, 메시지 스트림(108) 내의 메시지들은 초당 5k개의 메시지들을 초과할 수 있다). 또한, 메시지들은 비교적 짧을 수 있으며, 이는 메시지들이 때때로 소셜 미디어 플랫폼의 간결성에 특정한 고유 대화 스타일로 기록된다는 점을 고려할 때, 근본적인 텍스트의 의미론적 이해를 비교적 어렵게 할 수 있다. 추가로, 메시지 스트림(108)은 비교적 높은 정도의 잡음을 포함할 수 있는데, 예컨대, 이벤트들과 관련되지 않은 많은 메시지들이 존재할 수 있고, 심지어 관련된 것들조차 무관한 용어들을 포함할 수 있다. 또한, 위에서 표시된 바와 같이, 메시징 플랫폼(104) 상에서 논의된 이벤트들(112)은 본질적으로 동적일 수 있고, 따라서 시간에 걸쳐 변할 수 있다.

[0034] [0047] 위에서 식별된 난제들을 해결하기 위해, 이벤트 검출기(120)는 메시지 스트림(108)의 메시지들에 대해 클러스터링 동작들을 주기적으로 (예컨대, 분 단위로 또는 다른 주기적인 시간 프레임으로) 적용하고, 이벤트들의 동적으로 업데이트된 리스트(112)(예컨대, 클러스터 체인들(114)의 동적 업데이트된 리스트)를 생성할 수 있다. 예를 들어, 이벤트 검출기(120)는 전체 메시지 스트림(108)을 수신하고 주기적으로(예컨대, 분 단위 시간 인터벌 또는 다른 주기적인 시간 프레임으로) 트렌딩 엔티티들(132b)의 클러스터 그룹들(144)을 식별하는 실시간 이벤트 검출 시스템으로 고려될 수 있다. 이벤트 검출기(120)는, 이벤트(112)가 시간에 걸쳐 진행됨에 따라 이러한 클러스터 그룹들(144)을 클러스터 체인들(114)에 링크시킴으로써, 클러스터 체인들(114)을 실시간으로 추적할 수 있다. 예컨대, 이벤트 검출기(120)는 클러스터 체인(114)으로서 이벤트(112)를 모델링할 수 있고,

여기서 클러스터 체인(114)은 함께 링크된 2개 이상의 클러스터 그룹들(144)을 포함하고, 각각의 클러스터 그룹(144)은 2개 이상의 트렌딩 엔티티들(132b)을 포함한다.

[0035] [0048] 일반적으로, 이벤트(112)는 사람들이 메시징 시스템(100) 상의 이벤트(112)를 논의하기 위해 사용하는 엔티티들의 그룹에 의해 표현될 수 있다. 예컨대, 영화 시상식 쇼에 대한 이벤트(112)는, 논의되고 있는 지명된 배우들, 여배우들, 및 영화들과 같은 엔티티들에 의해 표현될 수 있다. 그러나, 본원에서 논의된 기술들에 따르면, 이벤트-기반 대화가 시간에 걸쳐 변할 수 있기 때문에, 이벤트(112)는 클러스터 체인(114), 예컨대, 트렌딩 엔티티들(132b)의 클러스터 그룹들(144)의 리스트(예컨대, 시간 순서로 인덱싱됨)로서 모델링된다. 검출된 이벤트(112)는 클러스터 체인(114)에 대응하고, 특정 시점에 트렌딩 엔티티들(132b)의 클러스터 그룹(144)에 의해 특성화된다.

[0036] [0049] 예를 들어, 도 1b를 참조하면, 이벤트 검출기(120)는 메시지 스트림(108)을 수신하고, 제1 시간 기간(196-1)으로부터 트렌딩 엔티티들(132b)의 제1 클러스터 그룹(144-1)을 검출할 수 있고, 제2 시간 기간(196-2)으로부터 트렌딩 엔티티들(132b)의 제2 클러스터 그룹(144-2)을 검출할 수 있다. 그 다음, 제1 및 제2 클러스터 그룹들(144-1, 144-2)이 동일한 이벤트(112)와 관련된다고 이벤트 검출기(120)가 결정하면, 메시징 시스템(100)은 클러스터 체인(114)을 형성하기 위해 이들 2개의 클러스터 그룹들(144)을 링크시킬 수 있다. 제1 시간 기간(196-1) 내의 특정 시간에, 이벤트(112)는 (위의 예에서 계속하여) 지진 및 쓰나미를 포함할 수 있는 제1 클러스터 그룹(144-1)의 트렌딩 엔티티들(132b)에 의해 특성화될 수 있다. 제2 시간 기간(196-2) 내의 특정 시간에, 이벤트(112)는 (위의 예에서 계속하여) 핵 및 방사선을 포함할 수 있는 제2 클러스터 그룹(144-2)의 트렌딩 엔티티들(132b)에 의해 특성화될 수 있다. 그러나, 제1 클러스터 그룹(144-1) 및 제2 클러스터 그룹(144-2)이 클러스터 체인(114) 내에서 링크되기 때문에, 이러한 용어들(예컨대, 엔티티들)은 이벤트-기반 대화의 변화하는 성질에도 불구하고 동일한 이벤트(112)와 관련된다.

[0037] [0050] 이벤트 검출기(120)는 버스트 검출기(126) 및 클러스터 체인 검출기(136)를 포함한다. 버스트 검출기(126)는 메시지 스트림(108)을 수신하고, 메시지 스트림(108)으로부터 트렌딩 엔티티들(132b)을 식별하기 위해 메시지 스트림(108)의 메시지들에 대해 버스트 검출 동작들을 실행할 수 있다. 클러스터 체인 검출기(136)는 트렌딩 엔티티들(132b)에 대해 클러스터링 동작들을 실행하여 클러스터 그룹들(144)을 검출하고, 메모리 디바이스(110)에 저장된 클러스터 체인들(114)을 형성할 수 있다.

[0038] [0051] 일부 예들에서, 버스트 검출기(126)는 클러스터 체인 검출기(136)와 별개인(또는 분리된) 컴포넌트(또는 모듈)이며, 여기서 버스트 검출기(126) 및 클러스터 체인 검출기(136)는 독립적으로 스케일링될 수 있다. 예들에서, 버스트 검출기(126)는 트렌드 검출기 서비스(128) 및 엔티티 검출기(130)를 포함한다. 일부 예들에서, 버스트 검출기(126)는 트렌드 검출기 서비스(128)를 포함하고, 엔티티 검출기(130)는 클러스터 체인 검출기(136)의 일부로서 포함된다.

[0039] [0052] 일부 예들에서, 버스트 검출기(126) 및 클러스터 체인 검출기(136)는 별개의 CPU들 및 메모리 디바이스들에 의해 실행가능하다. 또한, 버스트 검출(예컨대, 버스트 검출기(126)에 의해 수행됨)을 클러스터링 동작들(예컨대, 클러스터 체인 검출기(136)에 의해 수행됨)로부터 분리함으로써, 버스트 검출기(126)는 자신의 기능들 중 하나 이상을 클러스터 체인 검출기(136)의 기능들과 병렬로 실행할 수 있고, 이는 이벤트 검출 속도를 증가시킬 수 있다(예컨대, 특히 많은 양의 데이터를 처리하는 소셜 미디어 플랫폼들의 경우). 예컨대, 버스트 검출 후에 클러스터링 동작들을 순차적으로 수행하는 대신에, 이벤트 검출기(120)의 구성은 버스트 검출기(126)가 클러스터 체인 검출기(136)의 실행과 병렬로(예컨대, 적어도 부분적으로 병렬로) 실행되게 할 수 있다.

[0040] [0053] 일부 예들에서, 메시징 플랫폼(104)은 버스트 검출기(126) 및 클러스터 체인 검출기(136)에 할당된 컴퓨팅 리소스들을 모니터링하고 비교적 큰 메시지 스트림(108)에 대해 (예컨대, 실시간 또는 거의 실시간으로) 비교적 빠른 이벤트 검출을 가능하게 하기 위해 버스트 검출기(126) 및 클러스터 체인 검출기(136)에 할당된 컴퓨팅 리소스들의 양을 동적으로 증가(또는 감소)시키도록 구성된 컴퓨팅 리소스 모니터(165)를 포함한다. 일부 예들에서, 컴퓨팅 리소스 모니터(165)는 버스트 검출기(126)의 CPU 활용도 및/또는 메모리 사용량을 모니터링할 수 있다. 버스트 검출기(126)의 CPU 활용도 및/또는 메모리 사용량이 (예컨대, 메시지 스트림(108)의 크기의 갑작스러운 증가에 의해 야기될 수 있는) 임계 레벨을 초과하면, 컴퓨팅 리소스 모니터(165)는 버스트 검출기(126)에 대한 추가 컴퓨팅 리소스들의 할당을 초래할 수 있다. 또한, 컴퓨팅 리소스 모니터(165)는 클러스터 체인 검출기(136)의 CPU 활용도 및/또는 메모리 사용량을 별개로 모니터링할 수 있다. 클러스터 체인 검출기(136)의 CPU 활용도 및/또는 메모리 사용량이 임계 레벨을 초과하면, 컴퓨팅 리소스 모니터(165)는 클러스터 체인 검출기(136)에 대한 추가 컴퓨팅 리소스들의 할당을 초래할 수 있다. 버스트 검출기(126) 및 클러스터 체인

검출기(136)에 대해 독립적으로 스케일링가능한 컴포넌트들을 사용함으로써, 메시징 플랫폼(104)은 이벤트 검출의 속도를 증가시키고, 버스트 검출 및 엔티티 클러스터링의 다양한 프로세싱 부하들에 적응하기 위해 그리고 실시간으로 이벤트들(112)의 추적을 가능하게 하기 위해 메시징 플랫폼(104)의 유연성을 증가시킬 수 있다.

[0041] [0054] 트렌드 검출기 서비스(128)는 메시지 스트림(108)을 수신하고 트렌딩 엔티티들(132a)을 실시간으로 컴퓨팅할 수 있다. 트렌딩 엔티티(132a)는 비정상적으로 높은 레이트로 또는 임계 조건을 초과하는 레이트로 메시지 스트림(108)에 나타나는 엔티티일 수 있으며, 여기서 엔티티는 메시지 스트림(108)의 메시지 내의 콘텐츠에 대한 태그이다. 일부 예들에서, 트렌딩 엔티티(132a)는 특정 객체를 참조하는 메시지 내의 콘텐츠이다. 일부 예들에서, 트렌딩 엔티티(132a)는 단어(또는 단어들), 어구, 해시 태그, 식별자(예컨대, 사용자 식별자, 메시지 식별자 등), 웹 리소스(예컨대, URL), 및/또는 특정 객체를 참조하는 임의의 콘텐츠일 수 있다. 일부 예들에서, 트렌딩 엔티티들(132a)은 명명된 엔티티들, 해시 태그들, 식별자들(예컨대, 사용자 식별자들, 메시지 식별자들) 및 URL들을 포함한다. 일부 예들에서, 트렌드 검출기 서비스(128)는 메시징 플랫폼(104)의 다양한 컴포넌트들에 의해 사용될 수 있는 별개의 서비스이다. 일부 예들에서, 트렌드 검출기 서비스(128)는 실시간 지리적 구역들에 걸쳐 트렌딩 엔티티들(132a)을 식별한다.

[0042] [0055] 일부 예들에서, 트렌드 검출기 서비스(128)는 하나 이상의 도메인들을 정의할 수 있고, 여기서 각각의 도메인은 메시지 스트림(108)의 서브세트이고, 트렌드 검출기 서비스(128)는 개개의 도메인(예컨대, x 축 상의 메시지들의 수 및 y 축 상의 시간)에 대한 속도 그래프를 정의할 수 있다. 트렌드 검출기 서비스(128)는 트렌딩 엔티티들(132a)을 검출하기 위해 일정 시간 기간(예컨대, 스파이크)에 걸쳐 메시지들의 임계 수를 정의할 수 있다.

[0043] [0056] 일부 예들에서, 트렌드 검출기 서비스(128)는 데이터 준비, 엔티티 추출, 도메인 추출, 카운팅, 스코어링 및 랭킹을 포함할 수 있는 트렌딩 엔티티들(132a)을 식별하기 위해 일련의 프로세스 동작들을 실행할 수 있으며, 여기서 도메인 당 최상위 스코어링 트렌드들은 트렌드 검출기 서비스(128)에서 메모리에 지속되고 트렌드 관리자(106), 엔티티 검출기(130) 및/또는 클러스터 체인 검출기(136)와 같은 메시징 플랫폼(104)의 다른 서비스들에 의해 질의되도록 이용가능하다. 데이터 준비 동작은 낮은 텍스트 품질 또는 민감한 콘텐츠를 갖는 메시지들을 제거하는 것, 및 단일 사용자로부터의 트렌드에 대한 기여가 제한되는 것을 실질적으로 보장하기 위해 유사한 메시지들을 제거하는 것을 포함할 수 있다. 엔티티, 도메인 추출 및 카운팅 동작들은, 주어진 메시지에 대해, 이용가능한 엔티티들 및 지리적 도메인들을 추출하는 것을 포함할 수 있고, 모든 도메인 및 엔티티에 대해, 트렌드 검출기 서비스(128)는 <엔티티, 도메인, 1>의 튜플을 갖는 카운트를 방출하고 시간에 걸쳐 이 정보를 집계할 수 있다. 스코어링 동작은 이상 검출에 기초한 스코어링을 포함할 수 있으며, 여기서 트렌드 검출기 서비스(128)는 예상 <엔티티, 도메인> 카운트들을 컴퓨팅하고, 예상 카운트들을 관찰된 카운트들과 비교하여 각각의 쌍에 대한 스코어를 생성한다. 도메인 및 엔티티 쌍에 대한 예상 카운트를 계산하기 위해, 트렌드 검출기 서비스(128)는 다음 수학식을 사용할 수 있다:

[0044] 수학식 (1): $$E(d,e) = \frac{N_s(d)}{N_l(d)} \cdot N_1(d,e)$$

[0045] [0057] 수학식 (1)에서, $E(d,e)$는 도메인 d 및 엔티티 e에 대한 예상 카운트이고, N_l는 긴 시간 윈도우에 걸쳐 카운팅되고, N_S는 짧은 윈도우에 걸쳐 카운팅된다. 랭킹 동작은 도메인 당 최상위 스코어링 트렌드들을 결정하는 것을 포함할 수 있고, 도메인 당 최상위 스코어링 트렌드들은 트렌드 검출기 서비스(128)의 메모리에 지속되고, 질의되도록 이용가능하다.

[0046] [0058] 엔티티 검출기(130)는 트렌딩 엔티티들(132a)을 획득하기 위해 트렌드 검출기 서비스(128)에 주기적으로 질의할 수 있고, 그 다음 트렌딩 엔티티들(132a)을 엔티티 검출기(130)에 저장(예를 들어, 캐시)할 수 있다. 일부 예들에서, 엔티티 검출기(130)는 분 단위로 또는 다른 기간 시간 인터벌로 트렌드 검출기 서비스(128)에 주기적으로 질의한다. 일부 예들에서, 엔티티 검출기(130)는 매 30초마다 또는 15초마다와 같이 분 단위보다 더 빠르게, 또는 매 2분마다 또는 매 5분마다와 같이 분 단위보다 더 느리게 주기적으로 트렌드 검출기 서비스(128)에 질의할 수 있다.

[0047] [0059] 일부 예들에서, 엔티티 검출기(130)는 서버 통신 인터페이스를 통해 트렌드 검출기 서비스(128)와 통신할 수 있다. 일부 예들에서, 엔티티 검출기(130)는 하나 이상의 애플리케이션 프로그래밍 인터페이스(들)를 통해 트렌드 검출기 서비스(128)로부터 트렌딩 엔티티들(132a)을 획득할 수 있다. 일부 예들에서, 엔티티 검출기(130)는 원격 절차-호출(RPC)(예컨대, 절약 호출)을 트렌드 검출기 서비스(128)에 송신하고, 그 다음, 트렌드

검출기 서비스(128)로부터 트렌딩 엔티티들(132a)을 수신할 수 있다. 일부 예들에서, 엔티티 검출기(130)는 데이터베이스로부터 트렌딩 엔티티들(132a)을 획득하고, 데이터베이스는 트렌드 검출기 서비스(128)에 의해 기록된다. 일부 예들에서, 엔티티 검출기(130)는 트렌드 검출기 서비스(128)에 표현 상태 전달(REST) 요청을 송신하고, 그 다음, 트렌드 검출기 서비스(128)로부터 트렌딩 엔티티들(132a)을 수신할 수 있다. 일부 예들에서, 엔티티 검출기(130)는 GraphQL 요청을 통해 트렌드 검출기 서비스(128)와 통신한다.

[0048] [0060] 엔티티 검출기(130)는 메시지 스트림(108)을 수신하고, 메시지 스트림(108)으로부터 엔티티들(132)을 추출한다. 예를 들어, 엔티티 검출기(130)는 메시지 스트림(108)의 메시지들에 포함된 특정 타입들의 용어들을 추출할 수 있고, 이러한 타입들의 용어들은 엔티티들(132)로 지칭될 수 있다. 일부 예들에서, 엔티티 검출기(130)에 의해 추출된 엔티티(132)는 특정 객체를 지칭하는 메시지 내의 콘텐츠이다. 일부 예들에서, 엔티티 검출기에 의해 추출된 엔티티(132)는 단어(또는 단어들), 어구, 해시 태그, 식별자(예컨대, 사용자 식별자, 메시지 식별자 등), 웹 리소스(예컨대, URL), 및/또는 특정 객체를 참조하는 임의의 콘텐츠이다.

[0049] [0061] 일부 예들에서, 도 1c를 참조하면, 엔티티들(132)은 명명된 엔티티들(111)(예를 들어, "Oprah Winfrey"), 해시 태그들(113)(예를 들어, "#UsOpen"), 그래프 엔티티들(115)(예를 들어, "entity 123"), URL들(117) 및/또는 사용자 식별자들(119)을 포함한다. 정보 추출에서, 명명된 엔티티(111)는 사람들, 위치들, 조직들, 제품들 등과 같은 실세계 객체들일 수 있다. 명명된 엔티티(111)는 추상적이거나 물리적 존재를 가질 수 있다. 명명된 엔티티들(111)의 예들은 Barack Obama 또는 New York City 또는 명명될 수 있는 다른 것을 포함한다. 예컨대, "Barack Obama was the president of the United States"라는 다음의 문장을 고려하면, "Barack Obama"및 "United States" 둘 모두는 특정 객체들을 지칭하기 때문에 명명된 엔티티들(111)이지만, "president"는 많은 상이한 객체들을 지칭하는 데 사용될 수 있기 때문에 명명된 엔티티(111)로 고려되지 않을 수 있다. 해시 태그(113)는 옥토토프 심볼(octothorpe symbol)(#)에 의해 도입된 메시지 내의 단어 또는 비간격 어구이다. 그래프 엔티티들(115)은 메시징 플랫폼(104)의 지식 그래프 엔티티들을 표현한다. 일부 예들에서, 그래프 엔티티들(115)은 지식 그래프를 통해 캡처된 엔티티들의 내부 표현들을 포함한다. 사용자 식별자들(119)은 계정들(118)을 식별하는 사용자 또는 계정 식별자들일 수 있다.

[0050] [0062] 엔티티 검출기(130)는 트렌딩 엔티티들(132b)을 획득하기 위해, 트렌딩 엔티티들(132a)(트렌딩 검출기 서비스(128)로부터 수신됨)을 사용하여 엔티티들(132)로부터 비-트렌딩 엔티티들을 필터링 아웃할 수 있다. 일부 예들에서, 트렌딩 엔티티들(132b)은 도 1c에서 식별된 엔티티들의 타입(132) 중 하나 이상(또는 모든 타입들)을 포함하지만, 비-트렌딩 엔티티들이 필터링 아웃되기 때문에, 트렌딩 엔티티들(132b)은 트렌딩인 엔티티들(132)을 포함한다. 일부 예들에서, 트렌드 검출기 서비스(128)에 의해 식별된 트렌딩 엔티티들(132a)은 엔티티 검출기(130)에 의해 식별된 것들보다 더 큰 태그들의 집합을 포함한다. 예컨대, 트렌딩 엔티티들(132a)은 명명된 엔티티들(111), 해시 태그들(113), 그래프 엔티티들(115), URL들(117) 및 사용자 식별자들(119)을 포함할 수 있는 한편, 엔티티 검출기(130)에 의해 식별된 엔티티들(132)은 명명된 엔티티들(111), 해시 태그들(113) 및 그래프 엔티티들(115)을 포함한다. 따라서, 특정 사용자 식별자(119)가 트렌드 검출기 서비스(128)에 의해 트렌딩으로서 식별된 경우, 엔티티 검출기(130)가 오직 명명된 엔티티들(111), 해시 태그들(113) 및 그래프 엔티티들(115)만을 추출하기 때문에 엔티티 검출기(130)는 그 사용자 식별자(119)를 추출하지 않을 것이다. 일부 예들에서, 엔티티들(132)(및 결과적으로 트렌딩 엔티티들(132b))은 URL들(117) 및 사용자 식별자들(119)을 또한 포함하도록 확장된다.

[0051] [0063] 클러스터 체인 검출기(136)는 버스트 검출기(126)로부터 트렌딩 엔티티들(132b)을 수신하고, 시간에 걸쳐 링크된 클러스터 그룹들(144)의 클러스터 체인(114)을 생성할 수 있으며, 여기서 클러스터 체인(114)은 시간에 걸쳐 단일 이벤트(112)를 표현한다. 클러스터 체인 검출기(136)에 의해 생성된 클러스터 체인(114)은 메모리 디바이스(110)에 저장될 수 있다. 예를 들어, 클러스터 체인 검출기(136)는 시간에 걸쳐 트렌딩 엔티티들(132b)의 클러스터 그룹들(144)을 계속해서(예를 들어, 주기적으로) 검출할 수 있다. 예를 들어, 클러스터 체인 검출기(136)는 제1 시간 기간 동안 하나 이상의 클러스터 그룹들(144)을 검출하고, 제2 시간 기간 동안 하나 이상의 클러스터 그룹들(144)을 검출하고, 제3 시간 기간 동안 하나 이상의 클러스터 그룹들(144)을 검출할 수 있는 등등이다. 일부 예들에서, 클러스터 체인 검출기(136)는 주기적으로(예컨대, 1분마다) 하나 이상의 클러스터 그룹들(144)을 검출한다. 특정 클러스터 그룹(144)은 특정 시간 인터벌로 서로 유사하게 관련되는 것으로 결정된 2개 이상의 트렌딩 엔티티들(132b)을 포함할 수 있다(예컨대, 동일한 클러스터 그룹(144)으로부터의 트렌딩 엔티티들(132b)은 의미론적으로 동일한 주제를 지칭하는 것으로 고려될 수 있음). 일부 예들에서, 특정 클러스터 그룹(144)의 트렌딩 엔티티들(132b)은 공통 특성을 공유할 수 있다.

[0052] [0064] 클러스터 체인 검출기(136)는 클러스터 그룹들(144)을 검출하기 위해 임의의 타입의 커뮤니티-기반 클러

스터링 알고리즘을 실행할 수 있으며, 이는 루뱅 알고리즘 및 유사성 그래프들을 포함할 수 있거나 또는 포함하지 않을 수 있다. 그 다음, 클러스터 체인 검출기(136)는 클러스터 체인(114)을 생성하기 위해 현재 시간 기간으로부터의 하나 이상의 클러스터 그룹들(144)을 이전 시간 기간으로부터의 하나 이상의 클러스터 그룹들(144)에 링크시킬 수 있다.

[0053] [0065] 더 상세하게는, 클러스터 체인 검출기(136)는 엔티티 클러스터링 엔진(138), 유사성 그래프 생성기(148), 유사성 계산기(152), 유사성 그래프 필터(162), 클러스터 링커(168), 및 클러스터 랭커(170)를 포함한다. 일부 예들에서, 클러스터 체인 검출기(136)는 유사한 그래프 생성기(148)를 포함하지 않는다(예컨대, 유사성 그래프(146)가 생성되지 않음). 유사성 계산기(152)는 트렌딩 엔티티들(132b)을 수신하고, 시간 윈도우(158)에 걸친 빈도 카운트(154) 및 동시 발생들(156)에 기초하여 트렌딩 엔티티들(132b)에 대한 유사성 값들(150)을 컴퓨팅할 수 있다. 예를 들어, 유사성 계산기(152)는 시간 윈도우(158)에 걸쳐 트렌딩 엔티티들의 빈도 카운트(154) 및 동시 발생들(156)을 추적할 수 있다. 아래에서 추가로 설명되는 바와 같이, 시간 윈도우(158)는 동시 발생들(156) 및 빈도 카운트(154)의 집계를 위해 사용된다. 일부 예들에서, 시간 윈도우(158)는 슬라이딩 시간 윈도우이다. 일부 예들에서, 시간 윈도우(158)는 메모리 감소를 위해 조정될 수 있다.

[0054] [0066] 유사성 계산기(152)는 트렌딩 엔티티들(132b) 사이의 유사성들을 컴퓨팅하기 위해 시간 윈도우(158)에 걸쳐 빈도 카운트(154) 및 동시 발생들(156)을 사용할 수 있으며, 여기서 개개의 유사성 값(150)은 2개의 트렌딩 엔티티들(132b) 사이의 유사성의 레벨을 표현한다. 일부 예들에서, 유사성 값(150)은 에지 가중치에 의해 표현되거나 또는 에지 가중치로 간주된다. 표 1은 메시지 스트림(108)으로부터의 3개의 메시지들의 예를 예시한다.

[0055] 표 1

[0056]

메시지 ID	텍스트
1	iphone™ released during #appleevent
2	Tim Cook presents the new iphone™ #appleevent
3	Tim Cook unveiled the iphone™

[0057] [0067] 유사성 계산기(152)는 표 2에 나타난 바와 같은 동시 발생들(156)을 표현할 수 있다. 이 예에서, 트렌딩 엔티티들(132b)은 iphone™(예컨대, 명명된 엔티티(111)) 및 #appleevent(예컨대, 해시 태그(113))를 포함한다.

[0058] 표 2

[0059]

	메시지 1	메시지 2	메시지 3
iphone™	1	1	1
#appleevent	1	1	0

[0060] [0068] 유사성 계산기(152)는 iphone™ = Image#appleevent = Image와 같이 트렌딩 엔티티들(132b)(예컨대, iphone™ 및 #appleevent)에 대한 엔티티 벡터들을 컴퓨팅할 수 있다. 일부 예들에서, 트렌딩 엔티티들(132b)의 쌍에 대한 유사성 값(150)을 컴퓨팅할 때, 유사성 계산기(152)는 코사인 유사성에 기초하여 유사성 값을 컴퓨팅할 수 있다. 2개의 엔티티들 X 및 Y에 대한 코사인 유사성은 수학식 (2)에 나타난다.

[0061] [0069] 수학식 (2):

$$\cos(X,Y) = \frac{X \cdot Y}{\|X\|\|Y\|}$$

[0062] [0070] 위의 예에서, iphone™ 및 #appleevent의 트렌딩 엔티티들(132b)에 대해, 유사성 계산기(152)는 다음과 같이 유사성 값(150)을 컴퓨팅할 수 있다: Image 이 예가 코사인 유사성을 사용하지만, 유사성 계산기(152)는 시간 윈도우(158)에 걸쳐 빈도 카운트(154) 및 동시 발생들(156)을 사용하는 임의의 타입의 유사성 분석 및 2개의 엔티티들 사이의 유사성 거리(또는 유사성의 레벨)을 컴퓨팅할 수 있는 임의의 유사성 분석을 사용할 수 있

다.

[0063] [0071] 유사성 그래프 생성기(148)는 트렌딩 엔티티들(132b) 및 이들의 대응하는 유사성 값들(150)에 기초하여 유사성 그래프(146)를 생성할 수 있다. 도 1d는 일 양상에 따른 유사성 그래프(146)를 예시한다. 유사성 그래프(146)는 트렌딩 엔티티들(132b)을 표현하는 노드들(101)(예컨대, 노드들 1 내지 12) 및 노드들(101) 사이의 링크들을 표현하는 에지들(103)을 포함한다. 각각의 에지(103)는 유사성 계산기(152)에 의해 컴퓨팅된 유사성 값(150)인 에지 가중치와 연관된다.

[0064] [0072] 유사성 그래프 필터(162)는 유사성 임계치(160) 미만의 유사성 값들(150)을 갖는 에지들(103)이 유사성 그래프(146)로부터 제거되도록 유사성 임계치(160)에 기초하여 유사성 그래프(146)를 필터링할 수 있다. 일부 예들에서, 유사성 임계치(160)는 동일한 클러스터에 속하는 2개의 데이터 레코드들의 유사성에 대한 하한이다. 예컨대, 유사성 임계치(160)가 0.25로 설정되면, 25% 미만으로 유사한 필드 값들을 갖는 데이터 레코드들은 동일한 클러스터에 할당될 가능성이 낮다. 일부 예들에서, 유사성 임계치(160)는 0 내지 1의 범위를 갖는다. 유사성 임계치(160)는 유사성 그래프(146)의 에지 가중치들(예컨대, 유사성 값들(150))에 적용하기 위해 사용되는 최소 임계치일 수 있다(예컨대, 유사성 임계치(160) 미만의 유사성 값들(150)을 갖는 에지들(103)이 제거됨). 예를 들어, 유사성 그래프 필터(162)는 트렌딩 엔티티들(132b) 사이의 잡음이 있는 연결들을 제거하기 위해 유사성 임계치(160)(예컨대, 최소 유사성 임계치)를 사용하여 유사성 그래프(146)를 필터링할 수 있다. 에지(103)의 유사성 값(150)이 유사성 임계치(160) 미만이면, 유사성 그래프 필터(162)는 유사성 그래프(146)로부터 그 에지(103)를 제거할 수 있다.

[0065] [0073] 엔티티 클러스터링 엔진(138)은 클러스터 그룹들(144)을 검출하기 위해 유사성 그래프(146) 상에 클러스터링 알고리즘(140)을 적용할 수 있다. 엔티티 클러스터링 엔진(138)은 클러스터 그룹들(144)을 검출하기 위해 유사성 그래프(146)를 파티셔닝하는 클러스터링 알고리즘(140)을 실행할 수 있다. 일부 예들에서, 클러스터링 알고리즘(140)은 임의의 타입의 커뮤니티-기반 클러스터링 알고리즘을 포함한다. 일부 예들에서, 클러스터링 알고리즘(140)은 엔티티들을 비교 및 대조하여 국소적으로 관련된 엔티티들의 그룹들을 식별하도록 구성된 시뮬레이션 클러스터링 알고리즘을 포함한다. 일부 예들에서, 클러스터링 알고리즘(140)은 모듈성-기반 그래프 파티셔닝 알고리즘을 포함한다. 일부 예들에서, 클러스터링 알고리즘(140)은 루뱅 클러스터링 알고리즘을 포함한다. 일부 예들에서, 클러스터링 알고리즘(140)은 중립 네트워크를 정의하는 머신-러닝-기반 알고리즘이다. 일부 예들에서, 클러스터링 알고리즘(140)은 임의의 타입의 연결-기반 클러스터링 알고리즘(예컨대, 계층적 클러스터링 알고리즘), 중심-기반 클러스터링 알고리즘, 분포-기반 클러스터링 알고리즘, 및/또는 밀도-기반 클러스터링 알고리즘을 포함한다. 비교적 큰 네트워크들의 경우, 일부 예들에서, 루뱅 클러스터링 알고리즘은 클러스터 그룹들(144)을 결정하는 데 효율적일 수 있다. 일부 예들에서, 클러스터링 알고리즘(140)은 분해능(142)과 연관된다. 분해능(142)은 클러스터링 알고리즘(140)의 파라미터, 예컨대 루뱅 클러스터링 알고리즘일 수 있다. 분해능(142)은 복원된 클러스터들의 크기에 영향을 미치는 파라미터일 수 있다. 분해능(142)의 더 큰 값은 많은 더 작은 커뮤니티들을 초래할 수 있고, 더 작은 해상도(142)의 값은 몇몇 더 큰 커뮤니티들을 초래할 수 있다.

[0066] [0074] 도 1e는 일 양상에 따른 특정 시간 기간 동안 필터링 및 클러스터링 이후의 유사성 그래프(146)를 도시할 수 있다. 예를 들어, 유사성 그래프(146)는 도 1b의 제1 클러스터 그룹(144-1)에 대한 클러스터 그룹들(144)을 도시할 수 있다. 도 1d 및 도 1e를 참조하면, 노드(4)와 노드(6) 사이의 에지(103), 노드(5)와 노드(7) 사이의 에지(103), 노드(6)와 노드(7) 사이의 에지(103) 및 노드(10)와 노드(11) 사이의 에지(103)는 유사성 그래프 필터(162)에 의해 제거된다. 이어서, 클러스터링 알고리즘(140)의 적용에 대한 응답으로, 엔티티 클러스터링 엔진(138)은 클러스터 그룹들(144)을 검출할 수 있으며, 여기서 각각의 클러스터 그룹(144)은 2개 이상의 트렌딩 엔티티들(132b)(노드들(101)로 표현됨)을 포함한다. 일부 예들에서, 도 1e에 도시된 클러스터 그룹들(144)은 특정 시간 기간, 예컨대 제1 시간 기간(196-1), 제2 기간(196-2), 또는 제3 기간(196-3)에 대한 클러스터 그룹들(144)을 예시한다. 그 후, 후속 시간 기간(예컨대, 다음 1분) 동안, 엔티티 클러스터링 엔진(138)은 그 후속 시간 기간 동안 메시징 플랫폼(104) 상에서 교환되는 트렌딩 엔티티들(132b)을 수신하고 하나 이상의 클러스터 그룹들(144)을 검출할 수 있다.

[0067] [0075] 클러스터 링커(168)는 클러스터 체인(114)을 전개하기 위해 하나의 시간 기간으로부터의 하나 이상의 클러스터 그룹들(144)을 이전 시간 기간으로부터의 하나 이상의 클러스터 그룹들(144)과 링크시킬 수 있다. 예를 들어, 엔티티 클러스터링 엔진(138)이 주어진 시간 기간 C_T(예를 들어, 주어진 분)에 대한 클러스터 그룹들(144)을 생성하기 위해 클러스터링 알고리즘(140)을 적용한 후, 클러스터 링커(168)는 이전 시간 기간 C_{T-1}(예를

들어, 이전 분)로부터의 클러스터 그룹들(144)에 링크할 수 있다. 일부 예들에서, 클러스터 링커(168)는 하나의 시간 기간으로부터의 클러스터 그룹(144)과 이전의 시간 기간으로부터의 클러스터 그룹(144) 사이에 공유되는 다수의 트렌딩 엔티티들(132b)에 기초하여 클러스터 그룹들(144)을 링크시킬 수 있다.

[0068] [0076] 도 1b를 다시 참조하면, 이벤트(112)는 제1 클러스터 그룹(144-1), 제2 클러스터 그룹(144-1) 및 제3 클러스터 그룹(144-3)을 갖는 클러스터 체인(114)으로서 표현될 수 있다. 이벤트(112)(또는 클러스터 체인(114))는 이벤트 메타데이터(192)를 포함하거나 그와 연관될 수 있다. 이벤트 메타데이터(192)는 이벤트(112)의 검출된 시작 시간 및 종료 시간을 포함할 수 있다. 일부 예들에서, 각각의 클러스터 그룹(144)은 클러스터 메타데이터(141)를 포함하거나 그와 연관될 수 있다. 일부 예들에서, 클러스터 메타데이터(141)는 개개의 클러스터 그룹(144)에 의해 정의된 트렌딩 엔티티들(132b)의 엔티티 빈도 카운트를 포함할 수 있다. 일부 예들에서, 클러스터 메타데이터(141)는 개개의 클러스터 그룹(144)의 트렌딩 엔티티들(132b) 각각에 대한 빈도 카운트(154)를 포함할 수 있다. 일부 예들에서, 클러스터 메타데이터(141)는 개개의 클러스터 그룹(144)의 모든 트렌딩 엔티티들(143)에 걸친 빈도 카운트(154)의 총 수를 포함할 수 있다. 일부 예들에서, 클러스터 메타데이터(141)는 각각의 트렌딩 엔티티(132b)에 대한 또는 개개의 클러스터 그룹(144)의 트렌딩 엔티티들(132b) 전부에 걸쳐 그 수의 동시 발생들(156)을 포함할 수 있다. 일부 예들에서, 클러스터 메타데이터(141)는 개개의 클러스터 그룹(144)의 클러스터 식별자(166)를 포함할 수 있다.

[0069] [0077] 제1 클러스터 그룹(144-1)은 제1 시간 기간(196-1)과 연관되고, 제2 클러스터 그룹(144-2)은 제2 시간 기간(196-2)과 연관되며, 제3 클러스터 그룹(144-3)은 제3 시간 기간(196-3)과 연관된다. 제2 시간 기간(196-2)은 제1 시간 기간(196-1) 이후에 발생할 수 있고, 제3 시간 기간(196-3)은 제2 시간 기간(196-2) 이후에 발생할 수 있다. 시간 기간들 각각이 하나의 클러스터 그룹(144)을 도시하지만, 제1 시간 기간(196-1), 제2 시간 기간(196-2), 및 제3 시간 기간(196-3)은 (예컨대, 도 1e에 도시된 바와 같이) 다수의 클러스터 그룹들(144)을 포함할 수 있다. 일부 예들에서, 제1 시간 기간(196-1), 제2 시간 기간(196-2), 및 제3 시간 기간(196-3) 각각은 동일한 시간 길이를 갖는다(예컨대, 각각은 분의 시간 길이를 가짐). 그러나, 실시예들은 클러스터 그룹들(144)을 링크시키기 위한 임의의 주기적 시간 인터벌을 포함할 수 있다. 제1 클러스터 그룹(144-1)은 제1 시간 기간(196-1)으로부터의 트렌딩 엔티티들(132b)을 포함하고(예컨대, 엔티티 A, 엔티티 B, 엔티티 C), 제2 클러스터 그룹(144-2)은 제2 시간 기간(196-2)으로부터의 트렌딩 엔티티들(132b)을 포함하고(예컨대, 엔티티 A, 엔티티 D, 엔티티 C), 제3 클러스터 그룹(144-3)은 제3 시간 기간(196-3)으로부터의 트렌딩 엔티티들(132b)을 포함한다(예컨대, 엔티티 A, 엔티티 D, 엔티티 E).

[0070] [0078] 제2 클러스터 그룹(144-2)이 제1 클러스터 그룹(144-2)과 관련된다고 클러스터 링커(168)가 결정하면, 클러스터 링커(168)는 제1 클러스터 그룹(144-2)과 제2 클러스터 그룹(144-2) 사이에 링크(194)를 생성한다. 유사하게, 제3 클러스터 그룹(144-3)이 제2 클러스터 그룹(144-2)과 관련된다고 클러스터 링커(168)가 결정하면, 클러스터 링커(168)는 제2 클러스터 그룹(144-2)과 제3 클러스터 그룹(144-3) 사이에 링크(194)를 생성한다. 도 1b에 도시된 바와 같이, 이벤트(112)는 상이한 시간 기간들로부터의 클러스터 그룹들(144)을 링크시킴으로써 시간에 걸쳐(예컨대, 제1 시간 기간(196-1), 제2 시간 기간(196-2) 및 제3 시간 기간(196-3)에 걸쳐) 표현된다. 도 1b에 도시된 바와 같이, 제2 시간 기간(196-2) 동안, "엔티티 D"의 트렌딩 엔티티(132b)가 도입되었지만(그리고 제3 시간 기간(196-3) 동안, "엔티티 E"의 트렌딩 엔티티(132)가 도입되었지만), 본원에 설명된 기술들을 사용하여, 이벤트 검출기(120)는 제1 클러스터 그룹(144-1)의 트렌딩 엔티티들(132b)과 동일한 이벤트(112)와 관련된 것으로 나중에 도입되는 용어들을 검출할 수 있는데, 이는, 이벤트(112)가 시간 경과에 따른 클러스터 그룹들(144)의 리스트로서 모델링되기 때문이다.

[0071] [0079] 제2 클러스터 그룹(144-2)과 제1 클러스터 그룹(144-1) 사이의 링크(194)는 제1 클러스터 그룹(144-1) 및 제2 클러스터 그룹(144-2)이 동일한 이벤트(112)와 관련됨을 표시한다. 또한, 도 1b에 도시된 바와 같이, 클러스터 체인(114)은 제3 클러스터 그룹(144-3)과 제2 클러스터 그룹(144-2) 사이에 링크(194)를 가지며, 이는 제1 클러스터 그룹(144-1), 제2 클러스터 그룹(144-2), 및 제3 클러스터 그룹(144-3)이 동일한 이벤트(112)와 관련됨을 표시한다.

[0072] [0080] 일부 예들에서, 링크(194)는 제1 클러스터 그룹(144-1)과 제2 클러스터 그룹(144-2) 사이의 유사성의 레벨을 표현하는 에지 가중치이다. 일부 예들에서, 에지 가중치는 제1 클러스터 그룹(144-1)과 제2 클러스터 그룹(144-2) 사이에 공유되는 트렌딩 엔티티들(132b)의 수의 척도이다. 일부 예들에서, 에지 가중치는 유사성 그래프 생성기(148)와 관련하여 에지 가중치(예컨대, 유사성 값(150))와 유사하거나 동일하다. 일부 예들에서, 링크(194)와 연관된 에지 가중치가 임계치 값 미만이면, 링크(194)는 제거된다. 일부 예들에서, 클러스터 링커(168)는 최대 가중된 이분 매칭에 기초하여 제1 클러스터 그룹(144-1)과 제2 클러스터 그룹(144-2) 사이의 링크

(194)를 결정한다. 일부 예들에서, 최대 가중된 이분 매칭은 최대-가중치 매칭을 찾기 위한 최적화 문제를 포함한다.

[0073] [0081] 일부 예들에서, 클러스터 링커(168)는 이분 그래프를 생성하도록 구성되며, 여기서 주어진 시간 기간 C_T로부터의 클러스터 그룹들(144)은 이전 시간 기간 C_{T-1}로부터의 클러스터 그룹들(144)에 연결된다. 이분 그래프(예를 들어, G =(U, V, E))는, 각각의 에지 (u_i, v_j) ∈ E가 정점 u_i ∈ U 및 하나의 v_j ∈ V를 연결하도록, 정점들이 2개의 분리된 세트들 U 및 V로 분할될 수 있는 그래프이다. 그래프 G의 각각의 에지가 연관된 가중치 w_{ij}를 갖는 경우, 그래프 G는 가중된 이분 그래프로 지칭된다. 이분 그래프에서, 그래프 G의 매칭하는 M은, M의 어떠한 2개의 에지들도 공통 정점을 공유하지 않도록 E의 서브세트이다. 그래프 G가 가중된 이분 그래프이면, 최대 가중된 이분 매칭은 에지들의 가중치들의 합이 최대인 매칭이다. 이들 사이의 에지 가중치는 이전에 설명된 코사인 유사성과 유사하게, 이러한 클러스터 그룹들(144)이 얼마나 많은 엔티티들을 공유하는지의 척도일 수 있다. 클러스터 링커(168)는, 가중치가 임계치 미만으로 떨어지는 임의의 에지들을 필터링하고, 클러스터 링크들을 찾기 위해 최대 가중된 이분 매칭을 수행할 수 있다.

[0074] [0082] 일부 예들에서, 클러스터 그룹(144)이 성공적으로 링크될 때, 클러스터 링커(168)는 이전 시간 기간의 클러스터 그룹(144)으로부터 주어진 시간 기간의 클러스터 그룹(144)으로 클러스터 식별자(166)를 복사할 수 있다. 링크되지 않은 임의의 클러스터 그룹들(144)에 대해, 클러스터 링커(168)는 새로운 고유 클러스터 식별자(166)를 생성할 수 있다. 예를 들어, 제1 클러스터 그룹(144-1)이 이전 시간 기간으로부터의 클러스터 그룹(144)에 링크되지 않으면, 클러스터 링커(168)는 클러스터 식별자(166)(예컨대, 새로운 고유 식별자)를 제1 클러스터 그룹(144-1)에 할당할 수 있다. 제2 클러스터 그룹(144-2)이 (링크(194)를 통해) 제1 클러스터 그룹(144-1)에 링크되기 때문에, 클러스터 링커(168)는 제1 클러스터 그룹(144-1)에 할당된 동일한 클러스터 식별자(166)를 제2 클러스터 그룹(144-2)에 할당할 수 있다. 유사하게, 제3 클러스터 그룹(144-3)이 제2 클러스터 그룹(144-2)에 링크되기 때문에, 클러스터 링커(168)는 제2 클러스터 그룹(144-2)에 할당된 동일한 클러스터 식별자(166)를 제3 클러스터 그룹(144-3)에 할당할 수 있다. 따라서, 일부 예들에서, 클러스터 식별자(166)는 특정 클러스터 체인(114) 내의 클러스터 그룹들(144) 각각에 대해 동일할 수 있다.

[0075] [0083] 클러스터 랭커(170)는 클러스터 체인(114) 내에서 클러스터 그룹들(144)을 랭킹할 수 있다. 예를 들어, 도 1b의 클러스터 체인(114)과 관련하여, 클러스터 랭커(170)는 제1 클러스터 그룹(144-1), 제2 클러스터 그룹(144-2) 및 제3 클러스터 그룹(144-3)의 랭크된 리스트를 결정할 수 있다. 일부 예들에서, 클러스터 랭커(170)는 개개의 클러스터 그룹(144) 내에 포함된 트렌딩 엔티티들(132b)의 집계 인기도 메트릭에 기초하여 클러스터 그룹들(144)을 랭킹할 수 있다. 예를 들어, 제1 클러스터 그룹(144-1)과 관련하여, 클러스터 랭커(170)는 제1 클러스터 그룹(144-1)과 연관된 트렌딩 엔티티들(132b) 각각과 연관된 인기도 메트릭을 식별하고, 그 다음, 전체(집계 인기도 메트릭)를 획득하기 위해 인기도 메트릭들을 집계할 수 있다. 클러스터 랭커(170)는 제2 클러스터 그룹(144-2) 및 제3 클러스터 그룹(144-3)에 대해 동일한 동작들을 수행할 수 있다. 그런 다음, 클러스터 랭커(170)는 제1 클러스터 그룹(144-1), 제2 클러스터 그룹(144-2) 및 제3 클러스터 그룹(144-3)을 이들의 전체 인기도 스코어에 따라 랭킹할 수 있다.

[0076] [0084] 이벤트 검출기(120)는 클러스터 체인(114)을 메모리 디바이스(110)에 저장할 수 있다. 예컨대, 이벤트 검출기(120)는, 클러스터 체인(114)(또는 이들의 부분들)이 미래의 링크 단계들을 위해 클러스터 체인 검출기(136)에 의해 리트리브될 수 있거나 또는 메시징 플랫폼(104)에 의해 제공되는 하나 이상의 다른 서비스들에 의해 리트리브될 수 있도록 메모리 디바이스(110)에 클러스터 그룹들(144)의 랭킹된 리스트를 저장할 수 있다.

[0077] [0085] 메시징 플랫폼(104)은, 클라이언트 애플리케이션(176)의 사용자 인터페이스(178)에서 이벤트(들)(112)에 관한 정보를 렌더링하기 위해 디지털 데이터(190)를 (네트워크(172)를 통해) 클라이언트 애플리케이션(176)에 송신할 수 있다. 이벤트(들)(112)에 관한 정보는 클러스터 체인 정보(184)를 포함할 수 있다. 클러스터 체인 정보(184)는 하나 이상의 이벤트들(112)을 식별하고, 시간에 걸쳐 하나 이상의 클러스터 그룹들(144)로부터 하나 이상의 트렌딩 엔티티들(132b)을 식별할 수 있다. 일부 예들에서, 클러스터 체인 정보(184)는 트렌드들(182)에 통합된다. 예컨대, 트렌드들(182)은 트렌딩 해시 태그들을 포함하는 트렌딩 토픽들의 리스트를 식별할 수 있다. 하나의 특정 예에서, 해시 태그 "#bucks"는 트렌딩일 수 있고 이벤트(112)의 일부로서 포함될 수 있다. 일부 예들에서, 트렌드 관리자(106)는 클러스터 체인들(114)을 수신하고, 이벤트(112)와 연관된 하나 이상의 관련 용어들을 식별할지 여부를 결정할 수 있다. 해시 태그 "#bucks"는 제1 클러스터 그룹(144-1)의 일부로서 포함될 수 있고, 다른 용어("Giannis")는 제2 클러스터 그룹(144-2)의 일부로서 포함될 수 있다. 트렌드 관리자(106)는 엔티티(Giannis)를 식별하고 엔티티(Giannis)를 "#bucks"의 트렌딩 해시 태그에 대한 관련 용어로

서 표시하기 위해 그 이벤트(112)에 대한 클러스터 체인(114)을 사용할 수 있다.

[0078] [0086] 일부 예들에서, 클러스터 체인 정보(184)는 사용자의 타임라인(180)의 일부로서 통합된다. 예를 들어, 타임라인 관리자(163)는 클러스터 체인들(114)을 수신하고, 사용자의 타임라인(180)에서 렌더링될 메시지들 중 임의의 메시지가 클러스터 체인들(114) 내에 트렌딩 엔티티들(132b)을 포함하는지 여부를 결정할 수 있다. 만약 그렇다면, 타임라인 관리자(163)는 동일한 클러스터 그룹(144)으로부터의 다른 트렌딩 엔티티들(132b) 또는 동일한 클러스터 체인(114)의 일부인 다른 클러스터 그룹들(144)로부터의 트렌딩 엔티티들(132b)을 식별할 수 있다.

[0079] [0087] 타임라인(180)은 연결 그래프(116)의 클라이언트 애플리케이션(176)의 사용자의 계정(118)과 관계를 갖는 계정들(118)에 의해 포스팅된 메시지들의 리스트를 포함하는 메시지들의 스트림을 포함할 수 있다. 일부 예들에서, 타임라인(180)의 일부로서 포함된 메시지들의 스트림이 랭킹되고, 메시지들의 랭킹은 검출된 이벤트(112)에 (부분적으로) 기초할 수 있다. 예를 들어, 도 1b를 참조하여 논의된 바와 같이, 이벤트(112)는 이벤트(112)의 검출된 시작 시간 및 이벤트(112)의 검출된 종료 시간을 식별하는 이벤트 메타데이터(192)를 포함할 수 있다. 검출된 시작 시간은 클러스터 체인(114)의 제1 클러스터 그룹(144)과 연관된 시간에 기초할 수 있고, 검출된 종료 시간은 클러스터 체인(114)의 마지막 클러스터 그룹(144)과 연관된 시간에 기초할 수 있다. 타임라인 관리자(163)는 이벤트(112)(또는 클러스터 체인(114))에 속하는 상이한 클러스터 그룹들(144)에 걸쳐 트렌딩 엔티티들(132b)을 식별하는 이벤트(112)를 수신할 수 있다. 타임라인 관리자(163)는 사용자의 타임라인(180)의 일부로서 렌더링될 메시지가 이벤트(112)의 지속기간 동안(예컨대, 검출된 시작 시간과 검출된 종료 시간 사이에) 클러스터 체인(114)으로부터 트렌딩 엔티티(132b)를 포함하는지 여부를 결정할 수 있다. 그 메시지가 이벤트(112)의 지속기간 동안 클러스터 체인(114)의 일부인 트렌딩 엔티티(132b)를 포함하면, 타임라인 관리자(163)는 사용자의 타임라인(180)의 랭킹 내에서 그 메시지를 부스팅(또는 업-랭크)할 수 있다.

[0080] [0088] 일부 예들에서, 타임라인(180)은 광고 메시지들을 포함할 수 있는 촉진된 콘텐츠를 포함한다. 연결 그래프(116)에 따라 전달될 메시지들과 유사하게, 그 촉진된 메시지가 이벤트의 지속기간 동안 클러스터 체인(114)으로부터의 트렌딩 엔티티(132)를 포함하면, 촉진된 메시지는 타임라인의 랭킹에서 부스팅될 수 있다. 일부 예들에서, 메시징 플랫폼(104)은 촉진된 메시지들에 대한 가격 책정을 결정하도록 구성된 광고 스택 엔진을 포함한다. 일부 예들에서, 촉진된 콘텐츠가 이벤트(112)의 지속기간 동안 하나 이상의 트렌딩 엔티티들(132)을 포함하면, 광고 스택 엔진은 촉진된 콘텐츠에 대한 자신의 가격 책정을 증가시킬 수 있다.

[0081] [0089] 일부 예들에서, 클러스터 체인 정보(184)는 검색 결과들(186)의 일부로서 통합된다. 예컨대, 사용자는 질의 검색을 제출할 수 있고, 검색 관리자(161)는 다른 관련된 엔티티들을 포함하도록 검색 결과들(186)을 확장시키기 위해 클러스터 체인들(114)을 사용할 수 있다. 예를 들어, #bucks에 대한 질의에 대한 응답으로, 검색 관리자(161)는 이벤트들(112)의 리스트를 획득할 수 있고, #bucks라는 용어가 이벤트(112)와 연관되면, 검색 관리자(161)는 #bucks라는 용어를 포함하는 메시지들을 리턴할 수 있고, 제안된 용어(예컨대, 동일한 클러스터 체인(114)로부터의 다른 트렌딩 엔티티(132b))를 식별하거나 또는 #bucks라는 용어와 연관된 다른 트렌딩 엔티티들(132b)을 갖는 메시지들을 포함하는 검색 결과들을 리턴할 수 있다.

[0082] [0090] 도 2는 클라이언트 애플리케이션(276)의 트렌드 섹션(282) 내의 클러스터 체인 정보(284)를 도시하는 클라이언트 애플리케이션(276)의 사용자 인터페이스(278)의 예를 예시한다. 트렌드 섹션(282)은 트렌딩 엔티티들, 예컨대, #WorldRefugeeDay, 및 #NBADraft의 리스트를 디스플레이한다. 트렌드 섹션(282) 상에서 식별된 하나 이상의 엔티티들에 대해, 클라이언트 애플리케이션(276)은 (예컨대, 클러스터 체인들(114)로부터 결정된) 하나 이상의 관련 엔티티들을 식별하는 관련 라인(예컨대, 클러스터 체인 정보(284))을 렌더링한다. 예를 들어, 클러스터 체인 정보(284)는 동일한 클러스터 체인(114)의 일부로서 포함되는 몇몇 트렌딩 엔티티들(예컨대, 트렌딩 엔티티들(132b))(예컨대, "#WithRefugees", "Today is World Refugee Day")을 식별한다.

[0083] [0091] 도 3은 일 양상에 따른 유사성-기반 시간 이벤트 검출을 위한 프로세스 흐름(300)을 예시한다. 프로세스 흐름(300)은 도 1a 내지 도 1e의 메시징 시스템(100)과 관련하여 설명되며, 그러한 도면들을 참조하여 논의된 세부 사항들 중 임의의 것을 포함할 수 있다. 도 3의 흐름도(300)는 순차적인 순서로 동작들을 예시하지만, 이는 단지 예일 뿐이며, 추가적인 또는 대안적인 동작들이 포함될 수 있다는 것이 인식될 것이다. 추가로, 도 3의 동작들 및 관련된 동작들은 도시된 것과 상이한 순서로, 또는 병렬로 또는 중첩되는 방식으로 실행될 수 있다.

[0084] [0092] 동작(301)에서, 엔티티 추출은 이벤트 검출기(120)에 의해 수행된다. 예를 들어, 엔티티 검출기(130)는 메시지 스트림(108)을 수신하고, 메시지 스트림(108)의 메시지들로부터 엔티티들(132)을 추출한다. 일부 예들

에서, 엔티티들(132)은 명명된 엔티티들(111) 및 해시 태그들(113)을 포함한다. 일부 예들에서, 엔티티들(132)은 명명된 엔티티들(111), 해시 태그들(113), 그래프 엔티티들(115), URL들(117) 및/또는 사용자 식별자들(119)을 포함한다.

[0085] [0093] 동작(303)에서, 엔티티 필터링은 이벤트 검출기(120)에 의해 수행된다. 예컨대, 엔티티 검출기(130)는 트렌딩 엔티티들(132a)을 획득하기 위해 트렌드 검출기 서비스(128)와 통신한다. 엔티티 검출기(130)는 트렌딩 엔티티들(132b)의 리스트를 획득하기 위해 엔티티들(132)로부터 임의의 비-트렌딩 엔티티들을 필터링 아웃하기 위해 트렌드 검출기 서비스(128)로부터 수신된 트렌딩 엔티티들(132a)을 사용할 수 있다.

[0086] [0094] 동작(305)에서, 유사성들은 이벤트 검출기(120)에 의해 컴퓨팅된다. 예를 들어, 유사성 계산기(152)는 트렌딩 엔티티들(132b)의 쌍들에 대한 유사성 값(150)을 컴퓨팅하도록 구성된다. 예를 들어, 유사성 계산기(152)는 트렌딩 엔티티들(132b)을 수신하고, 시간 윈도우(158)에 걸쳐 이들의 빈도 카운트(154) 및 그들 사이의 동시 발생들을 추적한다. 유사성 계산기(152)는 트렌딩 엔티티들(132b) 사이의 유사성 값들(150)을 컴퓨팅하기 위해 빈도 카운트들(154) 및 동시 발생들(156)을 사용한다. 일부 예들에서, 유사성 계산기(152)는 2개의 트렌딩 엔티티들에 대한 코사인 유사성을 컴퓨팅한다. 일부 예들에서, 유사성 그래프 생성기(148)는 트렌딩 엔티티들(132b)을 노드들(101)로서 표현하고 유사성 값들(150)을 2개의 노드들(101)을 연결하는 에지들(103)에 대한 에지 가중치들로서 표현하는 유사성 그래프(146)를 구성할 수 있다.

[0087] [0095] 동작(307)에서, 유사성 필터링은 이벤트 검출기(120)에 의해 수행된다. 예를 들어, 유사성 임계치(160)는 유사성 그래프(146)에서 잡음있는 연결들을 필터링하는 데 사용된다. 유사성 그래프 필터(162)는 유사성 임계치(160)를 사용하여 유사성 그래프(146)를 필터링할 수 있고, 여기서 유사성 임계치(160) 미만의 유사성 값들(150)을 갖는 에지들(103)이 유사성 그래프(146)로부터 제거된다.

[0088] [0096] 동작(309)에서, 엔티티 클러스터링은 이벤트 검출기(120)에 의해 수행된다. 일부 예들에서, 이 스테이지에서, 엔티티 클러스터링 엔진(138)은 트렌딩 엔티티들(132b)의 클러스터 그룹들(144)을 검출하기 위해 유사성 그래프(146)를 파티셔닝하는 클러스터링 알고리즘(140)을 실행할 수 있다. 일부 예들에서, 클러스터링 알고리즘(140)은 분해능(142)을 갖는 루뱅 클러스터링 알고리즘을 포함한다.

[0089] [0097] 동작(311)에서, 클러스터 링크는 이벤트 검출기(120)에 의해 수행된다. 예를 들어, 클러스터 링커(168)는 동일한 이벤트(112)에 속하는 클러스터 그룹들(144)을 링크시키도록 구성된다. 일정 시간 기간 동안 클러스터 그룹들(144)을 생성하기 위해 커뮤니티 검출(예컨대, 엔티티 클러스터링)이 적용되면, 클러스터 링커(168)는 이전 시간 기간의 클러스터 그룹들(144)에 링크될 수 있다. 일부 예들에서, 클러스터 링커(168)는 현재 시간 기간의 클러스터 그룹들(144)이 제공되고 이전 시간 기간의 클러스터 그룹들(144)이 제공되는 이분 그래프를 구성하도록 구성된다. 이들 사이의 에지 가중치는 이전에 설명된 코사인 유사성과 유사하게, 이러한 클러스터 그룹들(144)이 얼마나 많은 엔티티들을 공유하는지의 척도이다. 클러스터 링커(168)는, 가중치가 임계치 미만으로 떨어지는 임의의 에지들을 필터링하고, 클러스터 링크들을 찾기 위해 가중된 이분 매칭을 수행할 수 있다. 클러스터 그룹(144)이 성공적으로 링크될 때, 클러스터 링커(168)는 이전 시간 기간의 클러스터 그룹(144)으로부터 현재 시간 기간의 클러스터 그룹(144)으로 클러스터 식별자(166)를 복사한다. 링크되지 않은 임의의 클러스터들에 대해, 클러스터 링커(168)는 새로운 고유 클러스터 식별자(166)를 생성할 수 있다.

[0090] [0098] 동작(313)에서, 클러스터 랭킹은 이벤트 검출기(120)에 의해 수행된다. 예를 들어, 클러스터 랭커(170)는 클러스터 그룹들(144)을 랭킹할 수 있다. 일부 예들에서, 클러스터 랭커(170)는 개개의 클러스터 그룹(144) 내에 포함된 트렌딩 엔티티들(132b)의 집계 인기도에 기초하여 클러스터 그룹들(144)을 랭킹할 수 있다.

[0091] [0099] 동작(315)에서, 클러스터 체인(114)은 메모리 디바이스(110)에 저장된다. 예컨대, 클러스터 그룹들(144)의 링크된, 랭킹된 리스트는 메모리 디바이스(110)에 지속되며, 그에 따라, 이들은 미래의 클러스터 링크 단계들을 위해 클러스터 체인 검출기(136) 내에서 리트리브되거나 또는 다른 서비스들(예컨대, 타임라인 관리자(163), 트렌드 관리자(106), 검색 관리자(161) 등)에 의해 리트리브될 수 있다.

[0092] [00100] 일부 예들에서, 이벤트 검출기(120)는 유사성-기반 시간 이벤트 검출 알고리즘(예컨대, 알고리즘 1)을 실행하여 클러스터 체인(114)을 생성하도록 구성된다. 예를 들어, 알고리즘 1은 도 3의 프로세스 흐름(300)의 예일 수 있다. 알고리즘 1에 대한 입력은 메시지 스트림(108), 유사성 임계치(160)(S), 분해능(142)(R), 및 시간 윈도우(158)(W)일 수 있다. 알고리즘 1의 출력은 시간 인터벌(예를 들어, 분 T)에 대한 클러스터 그룹들(144)의 리스트일 수 있다. 알고리즘 1에 대한 의사 코드가 제공되며, 여기서 그의 동작들은 도 3의 동작들에 맵핑된다. 아래에 표시된 바와 같이, 알고리즘 1은 일부 예들에서(본 개시에서 추후에 추가로 설명되는 바와

같이) 하나 이상의 클러스터링 이점들을 제공하는 루뱅 알고리즘을 사용한다. 그러나, 알고리즘 1은 임의의 타입의 커뮤니티 검출 알고리즘을 사용할 수 있다. 또한, 아래에 표시된 바와 같이, 알고리즘 1은 유사성 그래프들을 구축한다. 그러나, 일부 예들에서, 유사성 그래프들은 관련된 엔티티들을 그룹화하는 데 사용되지 않는다.

[0093] 알고리즘 1 (의사 코드)

[0094] M ←empty coOccurrence matrix

[0095] Trends ← {set of trending entities}

[0096] /* 백그라운드 스레드들에서 실행됨 */

[0097] for each message in the message stream do

[0098] E ←extract each entity e from a message (*operation 301*)

[0099] Filtered ← filter (E, e ∈ Trends) (*operation 303*)

[0100] for each entity Ef ∈ Filtered do (*operation 303*)

[0101] update Count(M, E_f) (*operation 303*)

[0102] end

[0103] end

[0104] /* 타이머 스레드를 통한 각각의 분 T */

[0105] remove(M,W) /* remove out-of-window updates */

[0106] G ←buildSimilarityGraph(M, S) (*operations 305, 307*)

[0107] C_T ← Louvain(G, R) (*operation 309*)

[0108] C_{T-1} ← fetch clusters for T - 1 (*operation 311*)

[0109] Links ←maxWeightedBipartiteMatching(C_T ,C_{T-1}) (*operation 311*)

[0110] For each c_t ∈ CT do (*operation 311)

[0111] if (c_t , c_{t-1}) ∈ Links then (*operation 311*)

[0112] copy ID from c_{t-1} to c_t (*operation 311*)

[0113] L ← L + c_t (*operation 311*)

[0114] end

[0115] Sort L (*operation 313*)

[0116] Return L

[0117] [00101] 도 1a를 다시 참조하면, 이벤트 검출기(120)는 (그의 성능을 평가하기 위해 하나 이상의 성능 메트릭들을 사용하여) 이벤트 검출기(120)의 결과들의 품질을 평가하기 위해 오프라인 분석 모드(122)에서 실행하고, 이어서, 분해능(142), 유사성 임계치(160) 및/또는 시간 윈도우(158)와 같은 이벤트 검출기(120)의 하나 이상의 파라미터들을 변경 또는 조정할 수 있다.

[0118] [00102] 도 4는 일 양상에 따른 오프라인 분석 모드(122)에서의 이벤트 검출기(120)의 예를 예시한다. 오프라인 분석 모드(122)에서, 이벤트 검출기(120)는 평가 데이터세트 스트림(105)을 수신하고 클러스터 체인들(114)을 컴퓨팅할 수 있다. 일부 예들에서, 평가 데이터세트 스트림(105)은 특정 지리적 영역(또는 다수의 지리적 영역들 또는 전세계)에서 특정 시간 기간(예컨대, 하루, 여러 날들, 1주일 등)에 걸친 메시지들을 포함한다. 일부 예들에서, 평가 데이터세트 스트림(105)은 하루(예컨대, 24시간 기간)에 걸쳐 메시징 플랫폼(104) 상에서

교환되는 메시지들을 포함한다.

[0119] [00103] 이벤트 검출기(120)는 엔티티들(132)(예컨대, 해시 태그들(113), 명명된 엔티티들(111), 및 그래프 기반 엔티티들(115) 등)을 추출할 수 있다. 그 다음, 이벤트 검출기(120)는, 임의의 유사성 필터링 없이(예컨대, 유사성 임계치(160)를 사용하여 유사성 그래프(146)를 필터링하지 않고) 위에서 논의된 것과 동일한 동작들을 수행할 수 있으며, 이는 원시 클러스터 체인들(114)의 세트를 생성한다. 그 다음, 클러스터 품질을 최적화하기 위해 엔티티 필터링 프로세스들이 튜닝될 수 있다. 각각의 클러스터 체인(114)에 대해, 이벤트 검출기(120)는 모든 시점으로부터 모든 트렌딩 엔티티들(132b)을 획득하고, 클러스터 체인(114) 당 하나의 중복 제거된 세트의 트렌딩 엔티티들(132b)을 생성할 수 있다. 각각의 클러스터 체인(114)에 대해, 대표적인 메시지들의 세트(예컨대, 10개의 가장 재-공유된 메시지들 및 10개의 랜덤 메시지들을 포함할 수 있는 20개의 메시지들)가 클러스터 체인(114)으로부터 적어도 2개의 동시 발생 엔티티들을 포함한다.

[0120] [00104] 일부 예들에서, 오프라인 분석의 일부로서, 대표적인 메시지들의 세트가 수동으로 검사되고, 클러스터 체인(114)이 이벤트(112)에 대응하면, 아래의 표 3에 도시된 바와 같이 클러스터 식별자(166) 및 타이틀이 할당된다.

[0121] 표 3

트렌딩 엔티티 132b	클러스터 식별자 166	타이틀	관련?
#madisonkeys	1	US Open Women's Quarterfinals	Y
#usopen	1	US Open Women's Quarterfinals	Y
Minnesota	1	US Open Women's Quarterfinals	N

[0122]

[0123] [00105] 클러스터 체인(114)이 다수의 이벤트들(112)을 포함하면, 상이한 클러스터 식별자들(166) 및 타이틀들이 이들 각각에 대해 생성된다. 그 다음, 모든 타이틀들이 체크되고, 복제물들은 단일 클러스터 식별자(166)로 병합된다. 또한, 무관 한 엔티티들이 마킹되고 거짓 포지티브 예들로서 저장된다(예컨대, 표 3의 Minnesota 참조). 데이터세트는 신뢰성을 보장하기 위해 별개의 개인에 의해 교차-검증된다. 일례에서, 평가 코퍼스(corpus)는 2695개의 트렌딩 엔티티들(132b) 및 460개의 이벤트들(112)(예컨대, 상이한 클러스터 식별자들(166))을 포함한다.

[0124] [00106] 일부 예들에서, 오프라인 분석 모드(122)에서, 이벤트 검출기(120)는 평가 데이터세트 스트림(105)과 동일한 세트의 메시지들에 대해 상이한 설정들로 실행될 수 있고, 이벤트 검출기(120)의 성능은 다음의 파라미터들 및 성능 메트릭들: 이벤트 검출된 프랙션(121), 통합(123), 구별(125), 분해능(142), 클러스터링 스코어(127), 병합된 이벤트 프랙션(129), 복제 이벤트 프랙션(131), 및 유사성 임계치(160) 중 하나 이상으로 평가된다.

[0125] [00107] 이벤트 검출된 프랙션(121)은 이벤트 검출 프로세스의 커버리지를 반영할 수 있다. 예를 들어, 이벤트 검출기(120)는 평가 데이터세트 스트림(105)으로부터 이벤트들의 프랙션을 컴퓨팅할 수 있다. 트렌딩 엔티티(132b)는 하나의 이벤트(112)에만 관련되는 경우 고유한 것으로 정의된다. 이벤트(112)의 적어도 하나의 고유 트렌딩 엔티티(132b)를 포함하는 하나보다 더 큰 크기의 클러스터 그룹(144)이 존재하면, 그 이벤트(112)가 검출된다. 일부 예들에서, 클러스터링 품질은 이러한 메트릭의 관심사가 아니다. 따라서, 몇몇 이벤트들(112)의 고유 트렌딩 엔티티들(132b)을 포함하는 단일 클러스터 그룹(144)이 존재하면, 그러한 이벤트들(112) 모두가 검출되는 것으로 간주된다.

[0126] [00108] 일부 예들에서, 유사성 임계치(160)는 검출된 이벤트들의 프랙션에 영향을 미치는 1차 필터로 간주될 수 있다. 도 5a는 일 양상에 따른, 유사성 임계치(160)의 증가하는 값에 대한 이벤트 검출된 프랙션(121)을 도시하는 그래프(510)를 예시한다. 도 5a에 도시된 바와 같이, 유사성 임계치(160)가 증가될 때, 이벤트 검출된 프랙션(121)은 감소되며, 이는 더 많은 에지들(103)이 필터링 아웃되기 때문이다. 다시 말해서, 더 많은 노드들(101)이 네트워크의 나머지로부터 격리되고(즉, 어떠한 에지도 없음) 클러스터 그룹들(144)로 그룹화될 수 없

다. 유사성 임계치(160)가 제로인 경우에도, 이벤트 검출된 프랙션(121)은 100% 미만임을 주목한다. 클러스터 그룹(144)이 적어도 2개의 노드들(101)을 포함한다고 가정하면, 네트워크 내의 격리된 노드들로부터의 이벤트들은 포함되지 않는다.

[0127] [00109] 일부 예들에서, 성능 메트릭들은 시스템의 품질을 평가하기 위한 2개의 주요 메트릭들, 예컨대 통합(123) 및 구별(125)을 포함한다. 일부 예들에서, 통합(123) 및 구별(125)은, 평가 데이터세트 스트림(105)보다 더 많은 이벤트들(112)을 검출하는 것에 대해서도 그리고 이벤트(112)에 대해 더 많은 트렌딩 엔티티들(132b)을 검출하는 것에 대해서도 페널티를 주지 않는 방식으로 설계된다. 통합(123) 및 구별(125)은, 단일 이벤트(112)를 표현하는 엔티티들을 병합하고 상이한 이벤트들(112)의 엔티티들을 각각 분리할 때의 유효성의 레벨을 측정하는 성능 메트릭들일 수 있다.

[0128] [00110] 2개의 트렌딩 엔티티들(132b)은, 이들이 실측 자료에서 단일 이벤트(112)의 일부이고 둘 모두가 관련된 것으로 마킹되면 관련있는 것으로 마킹된다. 2개의 트렌딩 엔티티들(132b)은, 이들이 실측 자료에서 단일 이벤트(112)의 일부이고 이들 중 정확히 하나가 무관한 것으로 마킹되면 관련되지 않을 것으로 지칭된다. 일부 예들에서, 본원에서 논의된 기술들은, 상이한 이벤트들(112)에 속하는 대부분의 엔티티 쌍들이 구별하기 쉽기 때문에, 이러한 명시적으로 마킹된 쌍들을 고려한다.

[0129] [00111] 통합(123)은 다음과 같이 정의될 수 있다:

[0130] 수학식 (3): $$C = \frac{\sum_{t \in t} a_t}{\sum_{t \in T} A_t}.$$

[0131] [00112] 구별(125)은 다음과 같이 정의될 수 있다:

[0132] 수학식 (4): $$D = \frac{\sum_{t \in t} b_t}{\sum_{t \in T} B_t}.$$

[0133] [00113] 다음의 파라미터들은 t: 타임스탬프, T: 시스템 출력의 모든 타임스탬프들의 세트, A_t: 타임스탬프 t에서 시스템 출력의 일부인 관련된 엔티티 쌍들의 수, a_t: 타임스탬프 t에서 시스템 출력의 공통 클러스터를 공유하는 관련된 엔티티 쌍들의 수, B_t: 타임스탬프 t에서 시스템 출력의 일부인 관련없는 엔티티 쌍들의 수, b_t: 타임스탬프 t에서 시스템 출력의 공통 클러스터에 있지 않은 관련없는 엔티티 쌍들의 수를 포함한다.

[0134] [00114] 직관적으로, 알고리즘은 100% 통합(123)을 달성하지만 0% 구별(125)을 갖는 것으로 단일 클러스터 그룹(144) 내의 모든 엔티티들을 배치하는 것으로 간주될 수 있다. 다른 한편으로, 각각의 엔티티에 대해 클러스터 그룹(144)을 생성하는 것은 0% 통합(123) 및 100% 구별(125)을 산출할 것이다. 일부 예들에서, 오프라인 분석 모드(122)에서의 이벤트 검출기(120)의 실행은 통합(123) 및 구별(125)을 최적화하기 위해 파라미터들(예컨대, 유사성 임계치(160), 분해능(142) 등)에 대해 어떤 값들을 사용할지를 결정할 수 있다.

[0135] [00115] 일부 예들에서, 통합(123) 및 구별(125)은 단일 메트릭, 예컨대, 다음과 같이 정의되는 클러스터링 스코어(127)로 조합된다:

[0136] 수학식 (5): $$CS = \left(\frac{C^{-1} + D^{-1}}{2}\right)^{-1} = \frac{2CD}{C+D}$$

[0137] [00116] 일부 예들에서, 오프라인 분석은 유사성 임계치(160)가 네트워크 구조에 어떻게 영향을 미치는지를 이해하기 위해 이러한 메트릭들을 레버리지할 수 있다. 도 5b는 이벤트 검출기(120)가 일 양상에 따른 클러스터링 알고리즘(140)으로서 연결된 컴포넌트 검출 방법을 사용할 때 유사성 임계치(160)의 값들을 증가시키기 위한 스코어들(예컨대, 클러스터링 스코어(127), 구별(125), 통합(123))을 도시하는 그래프(520)를 예시한다. 라인(501)은 구별(125)을 도시하고, 라인(503)은 클러스터링 스코어(127)를 도시하고, 라인(505)은 통합(123)을 도시한다. 도 5b에서, 유사성 임계치(160)가 0과 동일할 때, 모든 노드들(101)이 연결되기 때문에(즉, 완전한 유사성 그래프(146)) 통합 c = 1이고 식별 d = 0이다. 유사성 임계치(160)의 증가에 따라, 더 많은 에지들(103)이 유사성 그래프(146)로부터 제거되고, 따라서 구별(125)은 0에서 1로 증가한다.

[0138] [00117] 일부 예들에서, 유사성 임계치(160)가 특정 값 미만일 때(예컨대, S < 0.4), 더 많은 노드들(101)이 연결해제되기 때문에, 유사성 임계치(160)의 증가는 더 낮은 통합(123)과 관련된다. 그러나 유사성 임계치(160)

가 특정 값보다 클 때(예컨대, S > 0.4), 대부분의 에지들(103)이 제거되어, 많은 노드들(101)이 격리되고 최종 출력에 포함되지 않게 한다. 나머지 노드들(101)은 헤비 에지들(103)과 연결되며, 결과적으로 비교적 높은 통합(123)이 달성될 수 있다. 즉, 클러스터 그룹들(144) 및 이벤트 검출된 프랙션(121)의 크기는 더 작은 경향이 있다.

[0139] [00118] 그러나, 일부 예들에 따르면, 클러스터링 알고리즘(140)으로서 접속된 컴포넌트들에 의존하는 대신에, 클러스터링 알고리즘(140)은 유사성 임계치(160)(예컨대, S < 0.4)에 대한 증가된 클러스터링 성능을 달성하기 위해 루뱅 커뮤니티 검출 알고리즘을 포함할 수 있다. 도 5c는 일 양상에 따라 연결된 컴포넌트 알고리즘, 루뱅 알고리즘 및 이벤트 검출된 프랙션(121)에 대한 유사성 임계치(160)의 값들을 증가시키기 위한 클러스터링 스코어(127)를 도시하는 그래프(530)를 예시한다. 라인(507)은 루뱅 알고리즘을 도시하고, 라인(509)은 이벤트 검출된 프랙션(121)을 도시하고, 라인(511)은 연결된 컴포넌트 알고리즘을 도시한다. 일부 예들에서, 유사성 임계치(160)가 특정 값(예컨대, S < 0.2) 미만일 때, 루뱅 알고리즘은 컴포넌트들을 상이한 클러스터 그룹들(144)로 성공적으로 분할하기 때문에 더 양호한 성능을 달성한다. 유사성 임계치(160)가 특정 값을 초과할 때(예컨대, S > 0.2), 결과적인 컴포넌트들이 너무 작아서 분리될 수 없기 때문에, 루뱅 알고리즘은 연결된 컴포넌트 알고리즘과 동일한(또는 유사한) 결과들을 달성한다.

[0140] [00119] 도 5d는 이벤트 검출기(120)가 일 양상에 따른 클러스터링 알고리즘(140)으로서 루뱅 알고리즘을 사용할 때 분해능(142)의 값들을 증가시키기 위한 스코어들(예컨대, 클러스터링 스코어(127), 구별(125), 통합(123))을 도시하는 그래프(540)를 도시한다. 라인(513)은 통합(123)을 도시하고, 라인(515)은 클러스터링 스코어(127)를 도시하고, 라인(517)은 구별(125)을 도시한다. 유사성 임계치(160)가 S = 0.1인 경우, 일부 예들에서, 분해능(142)이 1과 동일한 것은 최적의 클러스터링 스코어를 초래할 수 있다.

[0141] [00120] 일부 예들에서, 성능 메트릭들은 병합된 이벤트 프랙션(129)으로 지칭되는 메트릭을 포함한다. 예컨대, 클러스터 품질 및 커버리지를 체크하는 것에 추가하여, 클러스터 체인들(114)이 평가된다. 구체적으로, 상이한 이벤트들(112)로부터의 트렌딩 엔티티들(132b)을 병합하는 클러스터 체인들(114)의 프랙션이 체크된다. 일부 예들에서, 병합된 이벤트 프랙션(129)은 클러스터링 품질 뿐만 아니라 시간 경과에 따른 클러스터 링크의 품질에도 민감할 수 있다. 도 5e는 유사성 임계치(160)의 상이한 값들에 대한 병합된 이벤트 프랙션(129)을 도시하는 그래프(550)를 예시한다. 도 5e에서, 병합된 이벤트 프랙션(129)은 유사성 임계치(160)의 값들을 증가시키기 위해 비교된다. 라인(519)은 연결된 컴포넌트 알고리즘을 도시하고, 라인(521)은 루뱅 알고리즘을 도시한다.

[0142] [00121] 일부 예들에서, 성능 메트릭들은 복제 이벤트 프랙션(131)으로 지칭되는 메트릭을 포함한다. 일부 예들에서, 복제 이벤트 프랙션(131)은 하나 초과의 클러스터 체인(114)에서 식별된 그들의 트렌딩 엔티티들(132b)을 갖는 평가 데이터세트 스트림(105) 내의 이벤트들의 프랙션으로서 정의된다. 도 5f는 일 양상에 따른 유사성 임계치(160)의 상이한 값들에 대한 복제 이벤트 프랙션(131)을 도시하는 그래프(560)를 예시한다. 라인(523)은 연결된 컴포넌트 알고리즘을 도시하고, 라인(525)은 루뱅 알고리즘을 도시한다. 예들에서, 복제 이벤트 프랙션(131)은 비교적 높을 수 있다. 일부 예들에서, 유사성 임계치(160)가 0.1일 때, 이는 복제 클러스터 체인(114)을 갖는 이벤트들의 약 35%를 초래할 수 있다. 일부 예들에서, 높은 값은 평가 데이터세트 스트림(105)이 서브-이벤트들을 포함하지 않는다는 사실로 인한 것일 수 있다. 그러나, 본 명세서에서 나중에 설명되는 바와 같이, 비교적 큰 이벤트(112)는 다수의 서브-이벤트들을 가질 수 있고, 일부 예들에서, 이들을 상이한 클러스터 체인들(114)에서 표현하는 것이 더 정확할 수 있다.

[0143] [00122] 도 1a를 다시 참조하면, 이벤트 검출기(120)는 실시간으로 클러스터 체인들(114)을 생성하기 위해 온라인 분석 모드(124)에서 실행될 수 있다. 예를 들어, 메시징 플랫폼(104)이 사용자들 사이에서 메시지들을 능동적으로 변경하고 있는 동안, 이벤트 검출기(120)는 이벤트들(112)이 언폴딩할 때 이들을 식별하기 위해 메시지 스트림(108)을 수신하고 시간에 걸쳐 클러스터 체인들(114)을 생성(및 업데이트)할 수 있다. 일부 예들에서, 이벤트 검출기(120)는 JVM(Java Virtual Machine) 상에 배치되고, 메모리 관리를 위해 GC(Garbage Collection)에 의존한다. 도 6a 내지 도 6c는 온라인 분석 모드(124)에서의 이벤트 검출기(120)의 성능의 다양한 프로파일들을 예시한다. 도 6a는 일 양상에 따른 초당 프로세싱된 엔티티들의 양을 도시하는 그래프(610)를 예시한다. 도 6b는 일 양상에 따른 CPU 활용 프로파일을 도시하는 그래프(620)를 예시한다. 도 6c는 일 양상에 따른 메모리 사용량 프로파일을 도시하는 그래프(630)를 예시한다. 도 6a 내지 도 6c에 도시된 성능 데이터는 비교적 연장된 시간 범위에 걸쳐 있으며, 이벤트 검출기(120)가 분당 수백만 개의 엔티티들을 스케일링 및 프로세싱할 수 있음을 입증할 수 있다.

[0144] [00123] 도 6b와 관련하여, CPU 사용량은 통상적으로 낮고(<10%) 트래픽의 정상일에 걸쳐 일관적이다. 유사하게, 도 6c에서, 메모리 사용량은 비교적 일관적이다. 그래프(630)의 시작 근처의 메모리 사용량의 감소는, 수명이 긴 객체들의 안정성으로 인해 1-2일마다 한 번만 발생하는 메이저 집합을 묘사하고, 마이너 집합들은 1-2분마다 지속적으로 발생한다. 도 6a에서, 중앙에 있는 하나의 특정 인스턴스에 대해, 시스템은 최대 50K PSEC(processed entities per second)의 스파이크들을 핸들링할 수 있다. 짧은 시간 기간에 PSEC의 배가를 표현하는 이러한 부하 스파이크 동안, 아래의 도면들에서 확인되는 바와 같이 대응하는 CPU 증가 및 메모리 영향은 비교적 무시가능하다. 따라서, 시스템(100)은 비정상적인 부하들에 직면할 때에도 안정적인 CPU 및 메모리 사용량을 나타낼 수 있다. 도 7은 일 양상에 따른 부하 셰딩(shedding)을 도시하는 그래프(710)를 예시한다. 도 7에서, 라인(701)은 초당 프로세싱된 메시지들을 표현하고, 라인(703)은 초당 드롭된 메시지들을 표현한다. 부하 셰딩은 일시적이고(예컨대, 주어진 분 동안만 발생함), 메시징 시스템(100)은 나중에 정상으로 재개된다는 것이 주목된다.

[0145] [00124] 도 8은 일 양상에 따른 실제 이벤트에 대한 최상위 클러스터 체인들(114)을 도시하는 그래프(810)를 예시한다. 도 8에 도시된 이벤트는, 01:00(UTC)에 시작하여, 2019년 1월 6일에 열린 골든 글로브 어워드 쇼였다. 이 이벤트는 메시징 플랫폼(104) 상에서 상당한 정도의 대화를 생성하였다. 아래의 설명은 그 이벤트 동안 이벤트 검출기(120)가 어떻게 수행되었는지 및 그것이 실제 작업 대화들을 얼마나 정확하게 반영했는지에 대한 예를 제공한다. 이벤트의 시간적 양상은 엔티티 클러스터들이 어떻게 나타나고, 진화하고, 사라지는지에 의해 나타난다.

[0146] [00125] 표 4는 도 8에 맵핑되는 최상위 클러스터 체인들(114)(및 연관된 최상위 엔티티들) 및 클러스터 체인들(114)을 예시한다. 예를 들어, 클러스터 체인들은 체인(801)(일반 변환), 체인(803)(호스트의 오프닝 스피치), 체인(805)(그린 북), 체인(807)(Christian Bale이 "Vice"로 코메디 음악 부분에서 남우 주연상을 수상함), 체인(809)(일반 대화), 체인(811)(Christian Bale이 그의 수락 스피치에서 Satan에게 감사를 표현함), 체인(813)(일반 대화), 체인(815)(Rami Malek이 "Bohemian Rhapsody"로 드라마 부분에서 남우 주연상을 수상함), 체인(817)(Glenn Close가 "The Wife"로 드라마 부분에서 여우 주연상을 수상함), 및 체인(819)(Green Book)을 포함한다.

[0147] 표 4

타이틀	최상위 엔티티들
801: 일반 변환	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Lady Gaga, Sandra Oh, Spider-Man: Into the Spider-Verse, Gaga
803: 호스트의 오프닝 스피치	Andy Samberg, Black Panther, Sandra Oh, #blackpanther, Jim Carrey, Michael B. Jordan
805: Green Book	Green Book, Maheshala Ali, Regina King, #greenbook
807: Christian Bale이 "Vice"로 코메디 음악 부분에서 남우 주연상을 수상함	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Christian Bale, Sandra Oh, Lady Gaga, Darren Criss, Vice
809: 일반 대화	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Lady Gaga, Jeff Bridges, Darren Criss
811: Christian Bale이 그의 수락 스피치에서 Satan에게 감사를 표현함	Christian Bale, The 76th Annual Golden Globe Awards 2019, Vice, Mitch McConnell, Satan
813: 일반 대화	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Sandra Oh, Alfonso Cuaron, Rami Malek, Roma, Olivia Colman
815: Rami Malek이 "Bohemian Rhapsody"로 드라마 부분에서 남우 주연상을 수상함	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Rami Malek, Bohemian Rhapsody, Lady Gaga, Sandra Oh
817: Glenn Close가 “The Wife”로 드라마 부분에서 여우 주연상을 수상함	Glenn Close, Taylor Swift, Lady Gaga, best actress, Glenn, Bradley Cooper
819: Green Book	Green Book, Mahershali Ali, Regina King, #greenbook

[0148]

[0149] [00126] 도 8은, 1월 6일과 7일 사이의 임의의 시간에 "*golden*globes*"에 매칭하는 엔티티를 포함하는 (총 메시지 카운트 관점에서) 10개의 가장 큰 클러스터 체인들을 포함하는 이벤트의 개요를 도시하며, 여기서 "*"는 임의의, 가능하게는 비어 있는 문자들의 시퀀스를 표현한다. 도 8에서, 이벤트 구조는 메인 행사 시간 외부에서, 즉, 01:00(UTC) 이전 및 4시 30분(UTC) 이후에 안정적이다. 행사 전에, 특정 주제가 아직 등장하지 않았기 때문에 모든 관련 엔티티들이 단일의 큰 체인으로 클러스터링된다. 행사 후에, 사용자들은 행사 동안 인기 있었던 토픽들(예컨대, Green Book, Glenn Close, 및 Rami Malek)에 대해 계속해서 이야기한다. 그러나, 가장 흥미로운 기간은 행사 자체 동안이어서: 시스템(100)은 빠르게 진화하는 토픽들을 캡처하고 이들에 대해 상이한 체인들을 생성할 수 있다.

[0150] [00127] 그린 북은 가장 호평을 받은 영화로, 최우수 조연상, 최우수 각본상, 최우수 작품상을 수상했다. 도 9는 일 양상에 따른 Green Book 클러스터 진화 및 대응하는 어워드 제시 시간들을 도시하는 그래프(910)를 도시한다. 도 9에서, Green Book 체인(도 8에서 체인(805)으로서 또한 표현됨)은 시간이 지남에 따라 진화한다. 처음에, Green Book 체인은 영화 및 영화의 총괄 프로듀서(Olivia Spencer)를 표현하는 엔티티들을 포함한다. 얼마 후, Mahershala Ali가 남우 주연상을 수상했으며, 적절한 엔티티가 클러스터에 추가된다. 유사하게, "각본상" 엔티티는 그 후에 추가된다. 작품상이 제시될 때, 클러스터는 완전히 개발된 대화를 표현한다: 관련된 해시 태그들이 사용되고 영화 제작진들이 언급된다. 05:09(UTC)에, 새로운 체인(도 8에서 체인(819)으로 표현됨)이 오리지널 Green Book 체인(도 8에서 체인(805))에서 나왔다. 후속적으로, 05:12(UTC)에, 새로운 체인은 새로운 클러스터 식별자를 유지하면서 오래된 체인을 흡수하였다. 이러한 차트들은 실시간으로 생성될 수 있기 때문에, 실세계 이벤트들이 펼쳐질 때 이를 추적하기 위해 사용될 수 있다.

[0151] [00128] 도 10은 일 양상에 따른 소셜 데이터 스트림들에 대한 이벤트 검출의 예시적인 동작들을 도시하는 흐름도(1000)를 예시한다. 흐름도(1000)는 도 1a 내지 도 1e의 메시징 시스템(100)을 참조하여 설명되지만, 동작들은 본원에서 논의된 실시예들 중 임의의 실시예에 의해 실행될 수 있다. 도 10의 흐름도(1000)는 순차적인 순

서로 동작들을 예시하지만, 이는 단지 예일 뿐이며, 추가적인 또는 대안적인 동작들이 포함될 수 있다는 것이 인식될 것이다. 추가로, 도 10의 동작들 및 관련된 동작들은 도시된 것과 상이한 순서로, 또는 병렬로 또는 중첩되는 방식으로 실행될 수 있다.

[0152] [00129] 동작(1002)은 메시징 플랫폼(104) 상에서 교환되는 메시지들의 스트림을 수신하는 것을 포함한다. 메시지들의 스트림은 메시지 스트림(108)의 일부일 수 있다. 동작(1004)은 제1 시간 기간(196-1)에 걸쳐 트렌딩 엔티티들(132b)의 제1 클러스터 그룹(144-1)을 검출하는 것을 포함한다. 제1 클러스터 그룹(144-1)은 서로 유사한 것으로 결정되는 2개 이상의 트렌딩 엔티티들(132b)을 포함한다. 동작(1006)은 제2 시간 기간(196-2)에 걸쳐 트렌딩 엔티티들(132b)의 제2 클러스터 그룹(144-2)을 검출하는 것을 포함한다. 제2 클러스터 그룹(144-2)은 서로 유사한 것으로 결정되는 2개 이상의 트렌딩 엔티티들(132b)을 포함한다. 동작(1008)은 제2 클러스터 그룹(144-2)을 제1 클러스터 그룹(144-1)과 링크시킴으로써 클러스터 체인(114)을 생성하는 것을 포함하고, 여기서 클러스터 체인(114)은 제1 및 제2 시간 기간들(196-1, 196-2)에 걸쳐 검출된 이벤트(112)를 표현한다. 일부 예들에서, 클러스터 그룹들(144)은 제1 클러스터 그룹(144-1)과 제2 클러스터 그룹(144-2) 사이에서 공유되는 다수의 트렌딩 엔티티들에 기초하여 링크된다. 동작(1010)은 메시징 플랫폼(104) 상의 메모리 디바이스(110)에 클러스터 체인(114)으로서 이벤트(112)를 저장하는 것을 포함한다.

[0153] [00130] 개시된 발명의 개념들이 첨부된 청구항들에서 정의된 것들을 포함하지만, 본 발명의 개념들은 또한 다음의 실시예들에 따라 정의될 수 있다는 것이 이해되어야 한다:

[0154] [00131] 실시예 1은, 소셜 데이터 스트림들에 대한 이벤트 검출을 위한 방법이고, 이 방법은 메시징 플랫폼에 의해, 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하는 단계, 및 메시징 플랫폼에 의해, 메시지들의 스트림으로부터 이벤트를 검출하는 단계를 포함한다.

[0155] [00132] 실시예 2는 실시예 1의 방법이고, 검출하는 단계는 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하는 단계를 포함하고, 제1 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다.

[0156] [00133] 실시예 3은 실시예 1 및 실시예 2 중 어느 하나의 방법이고, 검출하는 단계는 제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하는 단계를 포함하고, 제2 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다.

[0157] [00134] 실시예 4는 실시예 1 내지 실시예 3 중 어느 하나의 방법이고, 검출하는 단계는 제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인을 생성하는 단계를 포함하고, 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현한다.

[0158] [00135] 실시예 5는 실시예 1 내지 실시예 4 중 어느 하나의 방법이고, 메시징 플랫폼에 의해, 이벤트를 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 저장하는 단계를 더 포함한다.

[0159] [00136] 실시예 6은 실시예 1 내지 실시예 5 중 어느 하나의 방법이고, 메시징 플랫폼에 의해, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 클라이언트 애플리케이션에 디지털 데이터를 송신하는 단계를 더 포함하고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함한다.

[0160] [00137] 실시예 7은 실시예 1 내지 실시예 6 중 어느 하나의 방법이고, 이벤트에 관한 정보는 제1 클러스터 그룹으로부터의 제1 트렌딩 엔티티 및 제2 클러스터 그룹으로부터의 제2 트렌딩 엔티티를 식별하고, 제2 트렌딩 엔티티는 제1 트렌딩 엔티티와는 상이하다.

[0161] [00138] 실시예 8은 실시예 1 내지 실시예 7 중 어느 하나의 방법이고, 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하는 단계를 더 포함한다.

[0162] [00139] 실시예 9는 실시예 1 내지 실시예 8 중 어느 하나의 방법이고, 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하는 단계는 각각의 개개의 클러스터 그룹과 연관된 트렌딩 엔티티들의 인기도에 기초하고, 클러스터 체인은 랭킹된 클러스터 그룹들의 리스트를 포함한다.

[0163] [00140] 실시예 10은 실시예 1 내지 실시예 9 중 어느 하나의 방법이고, 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 검출하는 단계를 더 포함한다.

[0164] [00141] 실시예 11은 실시예 1 내지 실시예 10 중 어느 하나의 방법이고, 제1 클러스터 그룹은 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들로부터 검출된다.

[0165] [00142] 실시예 12는 실시예 1 내지 실시예 11 중 어느 하나의 방법이고, 제2 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 검출하는 단계를 더 포함한다.

[0166] [00143] 실시예 13은 실시예 1 내지 실시예 12 중 어느 하나의 방법이고, 제2 클러스터 그룹은 제2 시간 기간에 걸쳐 복수의 트렌딩 엔티티들로부터 검출된다.

[0167] [00144] 실시예 14는 실시예 1 내지 실시예 13 중 어느 하나의 방법이고, 제1 클러스터 그룹에 클러스터 식별자를 할당하는 단계, 및 제2 클러스터 그룹이 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 제1 클러스터 그룹의 클러스터 식별자를 제2 클러스터 그룹에 할당하는 단계를 더 포함한다.

[0168] [00145] 실시예 15는 실시예 1 내지 실시예 14 중 어느 하나의 방법이고, 제1 클러스터 그룹을 검출하는 단계는 제1 시간 기간의 복수의 트렌딩 엔티티들과 연관된 유사성 값들에 기초하여 유사성 그래프를 생성하는 단계를 포함한다.

[0169] [00146] 실시예 16은 실시예 1 내지 실시예 15 중 어느 하나의 방법이고, 유사성 그래프는 복수의 트렌딩 엔티티들을 표현하는 노드들 및 유사성 값들로 어노테이트된 에지들을 포함한다.

[0170] [00147] 실시예 17은 실시예 1 내지 실시예 16 중 어느 하나의 방법이고, 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 유사성 그래프를 파티셔닝하는 단계를 더 포함한다.

[0171] [00148] 실시예 18은 실시예 1 내지 실시예 17 중 어느 하나의 방법이고, 시간 윈도우에 걸친 복수의 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 단계를 더 포함한다.

[0172] [00149] 실시예 19는 실시예 1 내지 실시예 18 중 어느 하나의 방법이고, 유사성 임계치 미만의 유사성 값들을 갖는 에지들이 유사성 그래프로부터 제거되도록 유사성 임계치에 기초하여 유사성 그래프를 필터링하는 단계를 더 포함한다.

[0173] [00150] 실시예 20은 실시예 1 내지 실시예 19 중 어느 하나의 방법이고, 필터링된 유사성 그래프는 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 파티셔닝된다.

[0174] [00151] 실시예 21은 실시예 1 내지 실시예 20 중 어느 하나의 방법이고, 제2 클러스터 그룹은 최대 가중된 이분 매칭에 기초하여 제1 클러스터 그룹에 링크된다.

[0175] [00152] 실시예 22는 하나 이상의 컴퓨터들, 및 하나 이상의 컴퓨터들에 의해 실행될 때, 하나 이상의 컴퓨터들로 하여금 실시예 1 내지 실시예 21 중 어느 하나의 방법을 수행하게 하도록 동작가능한 명령들을 저장하는 하나 이상의 저장 디바이스들을 포함하는 시스템이다.

[0176] [00153] 실시예 23은 컴퓨터 프로그램으로 인코딩된 컴퓨터 저장 매체이며, 프로그램은, 데이터 프로세싱 장치에 의해 실행될 때, 데이터 프로세싱 장치로 하여금, 실시예 1 내지 실시예 21 중 어느 하나의 방법을 수행하게 하도록 동작가능한 명령들을 포함한다.

[0177] [00154] 실시예 24는 실시간 이벤트를 검출하기 위한 메시징 시스템이고, 메시징 시스템은, 네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼, 및 메시지들을 전송 및 수신하기 위해 메시징 플랫폼과 통신하도록 구성된 클라이언트 애플리케이션을 포함하고, 메시징 플랫폼은 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 제1 클러스터 그룹을 검출하도록 구성되고, 제1 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다.

[0178] [00155] 실시예 25는 실시예 24의 메시징 시스템이고, 메시징 플랫폼은 제2 시간 기간에 걸쳐 트렌딩 엔티티들의 제2 클러스터 그룹을 검출하도록 구성되고, 제2 클러스터 그룹은 서로 유사한 것으로 식별된 적어도 2개의 트렌딩 엔티티들을 포함한다.

[0179] [00156] 실시예 26은 실시예 24 및 실시예 25 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 제1 클러스터 그룹과 제2 클러스터 그룹 사이에서 공유되는 트렌딩 엔티티들의 수에 기초하여 제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써 클러스터 체인으로 구성된다.

[0180] [00157] 실시예 27은 실시예 24 내지 실시예 26 중 어느 하나의 메시징 시스템이고, 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현한다.

[0181] [00158] 실시예 28은 실시예 24 내지 실시예 27 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 메시징 플랫폼 상의 메모리 디바이스에 클러스터 체인으로서 이벤트를 저장하도록 구성된다.

[0182] [00159] 실시예 29는 실시예 24 내지 실시예 28 중 어느 하나의 메시징 시스템이고, 클러스터 체인은 장래의 클러스터 링크를 위해 리트리브가능하다.

[0183] [00160] 실시예 30은 실시예 24 내지 실시예 29 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 디지털 데이터를 클라이언트 애플리케이션에 송신하도록 구성되고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함하고, 클러스터 체인으로부터의 정보는 트렌드 섹션, 타임라인, 또는 클라이언트 애플리케이션에 리턴되는 검색 결과들의 일부에서 렌더링된다.

[0184] [00161] 실시예 31은 실시예 24 내지 실시예 30 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 각각의 클러스터 그룹과 연관된 집계 인기도 메트릭에 기초하여 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하도록 구성된다.

[0185] [00162] 실시예 32는 실시예 24 내지 실시예 31 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 트렌드 검출기 서비스로부터 상기 제1 시간 기간에 걸쳐 트렌딩 엔티티들의 리스트를 획득하고, 메시징 플랫폼 상에서 교환되는 메시지들의 스트림으로부터 엔티티들을 추출하도록 구성된다.

[0186] [00163] 실시예 33은 실시예 24 내지 실시예 32 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 획득하기 위해 트렌딩 엔티티들의 리스트를 사용하여 추출된 엔티티들을 필터링하도록 구성되고, 제1 클러스터 그룹은 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 사용하여 검출된다.

[0187] [00164] 실시예 34는 실시예 24 내지 실시예 33 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 단일 클러스터 체인의 클러스터 그룹들에 동일한 클러스터 식별자를 할당하도록 구성된다.

[0188] [00165] 실시예 35는 실시예 24 내지 실시예 34 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 시간 윈도우에 걸친 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하도록 구성된다.

[0189] [00166] 실시예 36은 실시예 24 내지 실시예 35 중 어느 하나의 메시징 시스템이고, 메시징 플랫폼은 유사성 값들에 기초하여 유사성 그래프를 생성하도록 구성된다.

[0190] [00167] 실시예 37은 실시예 24 내지 실시예 36 중 어느 하나의 메시징 시스템이고, 유사성 그래프는 트렌딩 엔티티들을 표현하는 노드들 및 유사성 값들로 어노테이트된 에지들을 포함한다.

[0191] [00168] 실시예 38은 실시예 24 내지 실시예 37 중 어느 하나의 메시징 시스템의 동작들을 포함하는 방법이다.

[0192] [00169] 실시예 39는 컴퓨터 프로그램으로 인코딩된 컴퓨터 저장 매체이며, 프로그램은, 데이터 프로세싱 장치에 의해 실행될 때, 데이터 프로세싱 장치로 하여금, 실시예 24 내지 실시예 37 중 어느 하나의 메시징 시스템의 동작들을 수행하게 하도록 동작가능한 명령들을 포함한다.

[0193] [00170] 실시예 40은, 적어도 하나의 프로세서에 의해 실행될 때, 적어도 하나의 프로세서로 하여금 메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신하게 하고 메시지들의 스트림으로부터 이벤트를 검출하게 하도록 구성되는 실행가능 명령들을 저장하는 비일시적 컴퓨터 판독가능 매체이다.

[0194] [00171] 실시예 41은 실시예 40의 비일시적 컴퓨터 판독가능 매체이고, 제1 시간 기간에 걸쳐 메시지들의 스트림으로부터 복수의 트렌딩 엔티티들을 식별하는 것을 더 포함한다.

[0195] [00172] 실시예 42는 실시예 40 및 실시예 41 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들로부터 제1 클러스터 그룹을 검출하는 것을 더 포함한다.

[0196] [00173] 실시예 43은 실시예 40 내지 실시예 42 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제2 시간 기간에 걸쳐 메시지들의 스트림으로부터 복수의 트렌딩 엔티티들을 식별하는 것을 더 포함한다.

[0197] [00174] 실시예 44는 실시예 40 내지 실시예 43 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제2 시간 기간에 걸쳐 복수의 트렌딩 엔티티들로부터 제2 클러스터 그룹을 검출하는 것을 더 포함한다.

[0198] [00175] 실시예 45는 실시예 40 내지 실시예 44 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제2 클러스터 그룹을 제1 클러스터 그룹과 링크함으로써 클러스터 체인을 생성하는 것을 더 포함한다.

[0199] [00176] 실시예 46은 실시예 40 내지 실시예 45 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 클러스터 체인은 제1 및 제2 시간 기간들에 걸쳐 검출된 이벤트를 표현한다.

[0200] [00177] 실시예 47은 실시예 40 내지 실시예 46 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 클라이언트 애플리케이션의 사용자 인터페이스에서 이벤트에 관한 정보를 렌더링하기 위해 디지털 데이터를 클라이언트 애플리케이션에 송신하는 것을 더 포함한다.

[0201] [00178] 실시예 48은 실시예 40 내지 실시예 47 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 이벤트에 관한 정보는 클러스터 체인으로부터의 정보를 포함하고, 클러스터 체인으로부터의 정보는 제1 클러스터 그룹으로부터의 제1 트렌딩 엔티티 및 제2 클러스터 그룹으로부터의 제2 트렌딩 엔티티를 식별한다.

[0202] [00179] 실시예 49는 실시예 40 내지 실시예 48 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 각각의 개개의 클러스터 그룹과 연관된 인기도 메트릭에 기초하여 제1 클러스터 그룹 및 제2 클러스터 그룹을 랭킹하는 것을 더 포함한다.

[0203] [00180] 실시예 50은 실시예 40 내지 실시예 49 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 메시지들의 스트림으로부터 엔티티들을 추출하는 것을 더 포함하고, 엔티티들은 명명된 엔티티들 또는 해시 태그들 중 적어도 하나를 포함한다.

[0204] [00181] 실시예 51은 실시예 40 내지 실시예 50 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 서버 통신 인터페이스를 통해 트렌드 검출기 서비스로부터 유도된 트렌딩 엔티티들의 리스트를 획득하는 것을 더 포함한다.

[0205] [00182] 실시예 52는 실시예 40 내지 실시예 51 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 비-트렌딩 엔티티들이 추출된 엔티티들로부터 필터링되도록, 트렌딩 엔티티들의 리스트에 기초하여 추출된 엔티티들로부터 제1 시간 기간에 걸쳐 복수의 트렌딩 엔티티들을 식별하는 것을 더 포함한다.

[0206] [00183] 실시예 53은 실시예 40 내지 실시예 52 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제1 클러스터 그룹에 클러스터 식별자를 할당하는 것, 및 제2 클러스터 그룹이 제1 클러스터 그룹에 링크되는 것에 대한 응답으로 제1 클러스터 그룹의 클러스터 식별자를 제2 클러스터 그룹에 할당하는 것을 더 포함한다.

[0207] [00184] 실시예 54는 실시예 40 내지 실시예 53 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 시간 윈도우에 걸친 복수의 트렌딩 엔티티들 사이의 동시 발생들 및 빈도 카운트에 기초하여 유사성 값들을 컴퓨팅하는 것을 더 포함한다.

[0208] [00185] 실시예 55는 실시예 40 내지 실시예 54 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 각각의 유사성 값은 2개의 트렌딩 엔티티들 사이의 유사성의 레벨을 표시한다.

[0209] [00186] 실시예 56은 실시예 40 내지 실시예 55 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 유사성 값들에 기초하여 유사성 그래프를 생성하는 것을 더 포함한다.

[0210] [00187] 실시예 57은 실시예 40 내지 실시예 56 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 유사성 그래프는 복수의 트렌딩 엔티티들을 표현하는 노드들 및 유사성 값들로 어노테이트된 에지들을 포함한다.

[0211] [00188] 실시예 58은 실시예 40 내지 실시예 57 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 유사성 임계치 미만의 유사성 값들을 갖는 에지들이 유사성 그래프로부터 제거되도록 유사성 임계치 값에 기초하여 유사성 그래프를 필터링하는 것을 더 포함한다.

[0212] [00189] 실시예 59는 실시예 40 내지 실시예 58 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제1 클러스터 그룹을 검출하기 위해 클러스터링 알고리즘에 따라 필터링된 유사성 그래프를 파티셔닝하는 것을 더 포함한다.

[0213] [00190] 실시예 60은 실시예 40 내지 실시예 59 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 클러스터링 알고리즘은 루뱅 알고리즘을 포함한다.

[0214] [00191] 실시예 61은 실시예 40 내지 실시예 60 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 제1 시간 기간 및 제2 시간 기간에 걸친 복수의 트렌딩 엔티티들은 버스트 검출기에 의해 메시지들의 스트림으로부터 식별되고, 클러스터 체인 검출기에 의해 제1 및 제2 클러스터 그룹들이 검출되고 클러스터 체인이 생성되고, 버스트 검출기의 컴퓨터 리소스들을 조정하는 것을 더 포함한다.

[0215] [00192] 실시예 62는 실시예 40 내지 실시예 61 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 버스트 검출기의 컴퓨터 리소스들의 조정과 독립적으로 클러스터 체인 검출기의 컴퓨터 리소스들을 조정하는 것을 더 포함한다.

[0216] [00193] 실시예 63은 실시예 40 내지 실시예 62 중 어느 하나의 비일시적 컴퓨터 판독가능 매체이고, 버스트 검출기의 하나 이상의 동작들은 클러스터 체인 검출기의 하나 이상의 동작들과 병렬로 수행된다.

[0217] [00194] 실시예 64는 하나 이상의 컴퓨터들, 및 하나 이상의 컴퓨터들에 의해 실행될 때, 하나 이상의 컴퓨터들로 하여금 실시예 40 내지 실시예 63 중 어느 하나의 비일시적 컴퓨터 판독가능 매체의 동작들을 수행하게 하도록 동작가능한 명령들을 저장하는 하나 이상의 저장 디바이스들을 포함하는 시스템이다.

[0218] [00195] 실시예 65는 실시예 40 내지 실시예 63의 비일시적 컴퓨터 판독가능 매체의 동작들의 단계들을 갖는 방법이다.

[0219] [00196] 실시예 66은 소셜 미디어 스트림 상에서 실시간 이벤트들을 검출하기 위한 메시징 시스템이고, 메시징 시스템은 네트워크를 통해 컴퓨팅 디바이스들에 메시지들을 교환하도록 구성된 메시징 플랫폼을 포함하고, 메시징 플랫폼은 오프라인 모드 및 온라인 모드에서 실행되도록 구성된 이벤트 검출기를 포함한다.

[0220] [00197] 실시예 67은 실시예 66의 메시징 시스템이고, 이벤트 검출기는 오프라인 모드에서, 제어 파라미터의 가변 값들에 대한 하나 이상의 제1 클러스터 체인들을 생성하기 위해 평가 데이터세트 스트림에 대해 이벤트 검출 알고리즘을 실행하고, 제어 파라미터의 가변 값들에 대한 이벤트 검출 알고리즘의 실행에 관한 성능 메트릭을 컴퓨팅하도록 구성되고, 제어 파라미터의 값은 성능 메트릭에 기초하여 선택된다.

[0221] [00198] 실시예 68은 실시예 66 및 실시예 67 중 어느 하나의 메시징 시스템이고, 이벤트 검출기는 온라인 모드에서, 메시징 플랫폼 상에서 실시간으로 교환되는 메시지들에 대한 메시지 스트림을 수신하고, 하나 이상의 제2 클러스터 체인들을 생성하기 위해 제어 파라미터의 선택된 값에 따라 메시지 스트림에 대해 이벤트 검출 알고리즘을 실행하도록 구성된다.

[0222] [00199] 실시예 69는 실시예 66 내지 실시예 68 중 어느 하나의 메시징 시스템이고, 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 제어 파라미터는 유사성 임계치를 포함한다.

[0223] [00200] 실시예 70은 실시예 66 내지 실시예 69 중 어느 하나의 메시징 시스템이고, 성능 메트릭은 구별 또는 통합 중 적어도 하나를 포함하고, 제어 파라미터는 클러스터링 알고리즘의 분해능을 포함한다.

[0224] [00201] 실시예 71은 실시예 66 내지 실시예 70 중 어느 하나의 메시징 시스템의 동작들을 포함하는 방법이다.

[0225] [00202] 실시예 72는 컴퓨터 프로그램으로 인코딩된 컴퓨터 저장 매체이며, 프로그램은, 데이터 프로세싱 장치에 의해 실행될 때, 데이터 프로세싱 장치로 하여금, 실시예 66 내지 실시예 70 중 어느 하나의 메시징 시스템의 동작들을 수행하게 하도록 동작가능한 명령들을 포함한다.

[0226] [00203] 위의 설명에서, 다수의 세부 사항들이 제시된다. 그러나, 본 개시의 이익을 갖는 당업자에게, 본 개시의 구현들이 이러한 특정 세부 사항들 없이 실시될 수 있다는 것은 명백할 것이다. 일부 경우들에서, 설명을 모호하게 하는 것을 피하기 위해, 잘-알려진 구조들 및 디바이스들이 상세하게보다는 블록도 형태로 도시된다.

[0227] [00204] 상세한 설명들의 일부 부분들은 컴퓨터 메모리 내의 데이터 비트들에 대한 동작들의 알고리즘들 및 심볼 표현들의 관점에서 제시된다. 이러한 알고리즘 설명들 및 표현들은 데이터 프로세싱 기술들의 당업자들이 자신의 작업의 본질을 다른 당업자들에게 가장 효과적으로 전달하기 위해 사용하는 수단이다. 알고리즘은 여기서 및 일반적으로는, 원하는 결과를 유도하는 단계들의 자체-일관성있는(self-consistent) 시퀀스인 것으로 고려된다. 단계들은 물리적 수량들의 물리적 조작들을 요구하는 것들이다. 통상적으로, 반드시 필요한 것은 아니지만, 이러한 수량들은 저장, 전달, 결합, 비교 또는 그렇지 않으면 조작될 수 있는 전기 및 자기 신호들의 형태를 취한다. 주로 일반적인 사용의 이유들 때문에, 비트들, 값들, 엘리먼트들, 심볼들, 문자들, 용어들, 숫자들 등으로서 이러한 신호들을 지칭하는 것이 종종 편리한 것으로 입증되었다.

[0228] [00205] 그러나, 이러한 및 유사한 용어들 모두는 적절한 물리 양들과 연관될 것이며, 단지 이러한 수량들에 적용되는 편리한 라벨들일 뿐임을 명심해야 한다. 위의 논의에서 명백한 바와 같이 달리 구체적으로 언급되지 않는 한, 설명 전반에 걸쳐 "식별", "결정", "계산", "검출", "송신", "수신", "생성", "저장", "랭킹", "추출", "획득", " 할당", "파티셔닝", "컴퓨팅", "필터링", "변경" 등과 같은 용어들을 활용하는 논의들은, 컴퓨터 시스템의 레지스터들 및 메모리들 내에서 물리적(예를 들어, 전자) 수량들로 표현된 데이터를 조작하고, 컴퓨터

시스템 메모리들 또는 레지스터들 또는 다른 이러한 정보 저장, 송신 또는 디스플레이 디바이스들 내의 물리적 수량들로서 유사하게 표현된 다른 데이터로 변환하는, 컴퓨터 시스템 또는 유사한 전자 컴퓨팅 디바이스의 액션들 및 프로세스들을 지칭하는 것이 인식된다.

[0229] [00206] 또한, 도면들에 도시된 논리 흐름들은, 바람직한 결과들을 달성하기 위해, 도시된 특정 순서 또는 순차적인 순서를 요구하지 않는다. 또한, 다른 단계들이 제공될 수 있거나, 설명된 흐름들로부터 단계들이 제거될 수 있고, 다른 컴포넌트들이 설명된 시스템들에 추가되거나 그로부터 제거될 수 있다. 따라서, 다른 실시예들은 다음의 청구항들의 범위 내에 존재한다.

[0230] [00207] 본 개시의 구현들은 또한, 본원의 동작들을 수행하기 위한 장치에 관한 것이다. 이러한 장치는 요구되는 목적을 위해 특별히 구성될 수 있거나, 또는 컴퓨터에 저장된 컴퓨터 프로그램에 의해 선택적으로 활성화되거나 재구성되는 범용 컴퓨터를 포함할 수 있다. 이러한 컴퓨터 프로그램은 플로피 디스크들, 광학 디스크들, CD-ROM들 및 자기 광학 디스크들, ROM들(read-only memories), RAM들(random access memories), EPROM들, EEPROM들, 자기 또는 광학 카드들, 플래시 메모리 또는 전자 명령들을 저장하기에 적합한 임의의 타입의 매체들과 같은(그러나 이에 제한되지 않음) 비일시적 컴퓨터 판독가능 저장 매체에 저장될 수 있다.

[0231] [00208] "예" 및/또는 "예시적인"이라는 단어들은, 예, 예증 또는 예시로서 기능하는 것을 의미하도록 본원에서 사용된다. "예" 또는 "예시적인" 것으로서 본원에 설명된 임의의 양상 또는 설계는 반드시 다른 양상들 또는 설계들에 비해 바람직하거나 유리한 것으로 해석될 필요는 없다. 오히려, "예" 또는 "예시적인"이라는 단어들의 사용은 구체적인 방식으로 개념들을 제시하도록 의도된다. 본 출원에서 사용되는 바와 같이, "또는"이라는 용어는 배타적인 "또는"보다는 포괄적인 "또는"을 의미하도록 의도된다. 즉, 달리 명시되지 않는 한 또는 문맥으로부터 명확하지 않은 경우, "X는 A 또는 B를 포함한다"는 자연적인 포괄적 치환들 중 임의의 것을 의미하는 것으로 의도된다. 즉, X가 A를 포함하면; X는 B를 포함하거나; 또는 X는 A 및 B 둘 모두를 포함하고, 이어서, "X는 A 또는 B를 포함한다"는 전술한 경우들 중 임의의 것 하에서 충족된다. 또한, 본 출원 및 첨부된 청구항들에서 사용되는 단수 표현들은 달리 명시되지 않거나 단수 형태로 지시되는 것으로 문맥상 명확하지 않다면, 일반적으로 "하나 이상"을 의미하는 것으로 해석되어야 한다. 더욱이, 전반에 걸쳐 "일 구현" 또는 "일 실시예" 또는 "구현"또는 "하나의 구현"이라는 용어의 사용은, 그렇게 설명되지 않는 한 동일한 실시예 또는 구현을 의미하는 것으로 의도되지 않는다. 또한, 본원에서 사용되는 바와 같은 "제1", "제2", "제3", "제4" 등의 용어들은 상이한 엘리먼트들 사이에서 구별하기 위한 라벨들로서 의도되며, 이들의 숫자 지정에 따라 반드시 서수적 의미를 가질 필요는 없다.

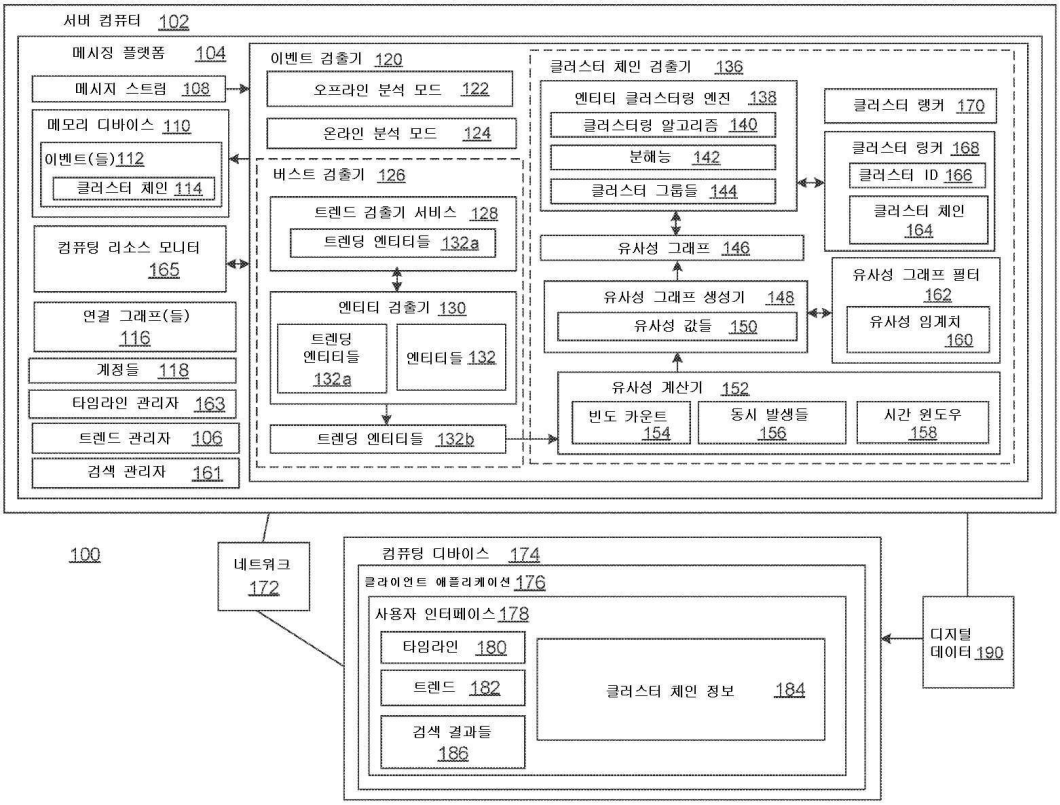
[0232] [00209] 본원에 제시된 알고리즘들 및 디스플레이들은 본질적으로 임의의 특정 컴퓨터 또는 다른 장치에 관한 것이 아니다. 다양한 범용 시스템들이 본원의 교시들에 따른 프로그램들과 함께 사용될 수 있거나, 또는 요구되는 방법 단계들을 수행하기 위해 보다 특수화된 장치를 구성하는 것이 편리한 것으로 입증될 수 있다. 이러한 다양한 시스템들에 요구되는 구조는 아래의 설명에서 나타날 것이다. 또한, 본 개시는 임의의 특정 프로그래밍 언어를 참조하여 설명되지 않는다. 본원에 설명된 바와 같이 본 개시의 교시들을 구현하기 위해 다양한 프로그래밍 언어들이 사용될 수 있다는 것이 인식될 것이다.

[0233] [00210] 위의 설명은 본 개시의 몇몇 구현들의 양호한 이해를 제공하기 위해, 특정 시스템들, 컴포넌트들, 방법들 등의 예들과 같은 다수의 특정 세부 사항들을 제시한다. 그러나, 본 개시의 적어도 일부 구현들은 이러한 특정 세부 사항들 없이도 실시될 수 있다는 것이 당업자에게 명백할 것이다. 다른 경우들에서, 잘-알려진 컴포넌트들 또는 방법들은 본 개시를 불필요하게 모호하게하는 것을 피하기 위해, 상세히 설명되지 않거나 간단한 블록도 포맷으로 제시된다. 따라서, 위에서 제시된 특정 세부 사항들은 단지 예들일 뿐이다. 특정 구현들은 이러한 예시적인 세부 사항들과 다를 수 있으며, 여전히 본 개시의 범위 내에 있는 것으로 고려된다.

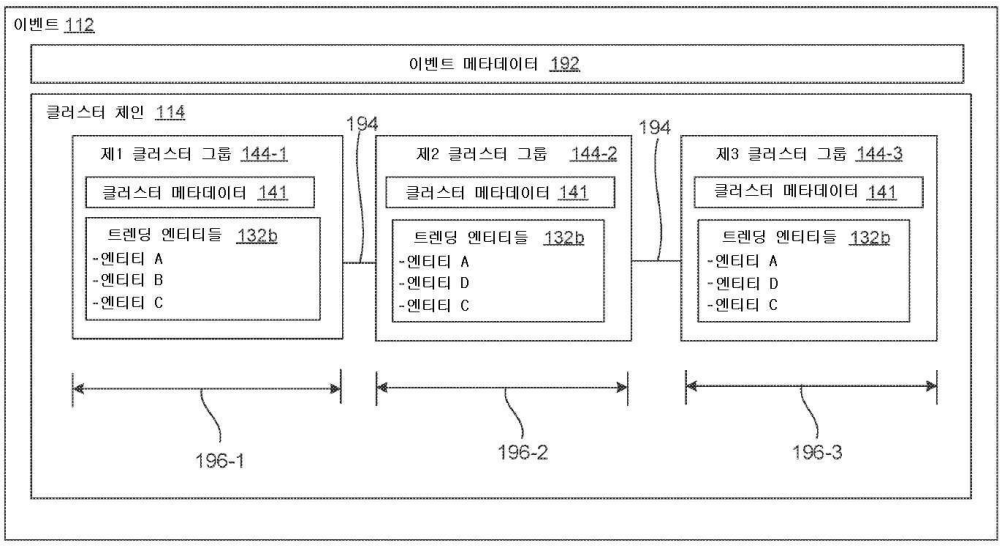
[0234] [00211] 위의 설명은 제한적인 것이 아니라 예시적인 것으로 의도된다는 것이 이해되어야 한다. 위의 설명을 읽고 이해하면 많은 다른 구현들이 당업자들에게 명백할 것이다. 따라서, 본 개시의 범위는, 그러한 청구항들에 부여된 등가물들의 전체 범위와 함께, 첨부된 청구항들을 참조하여 결정되어야 한다.

도면

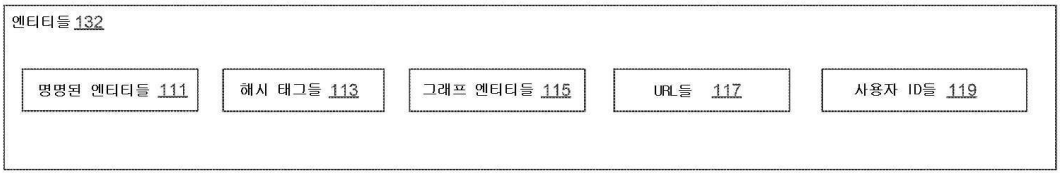
도면1a



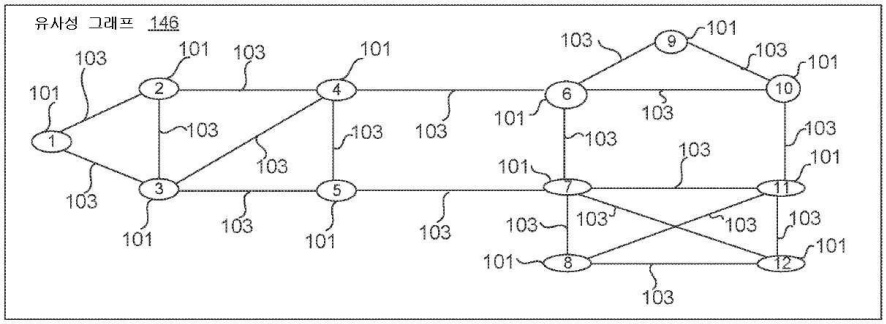
도면1b



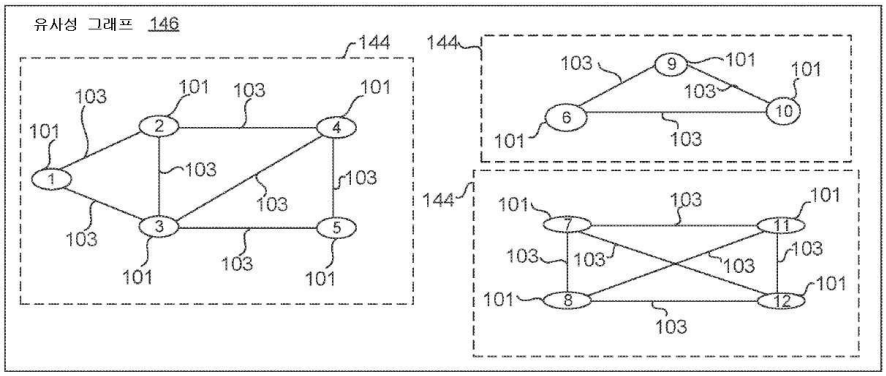
도면1c



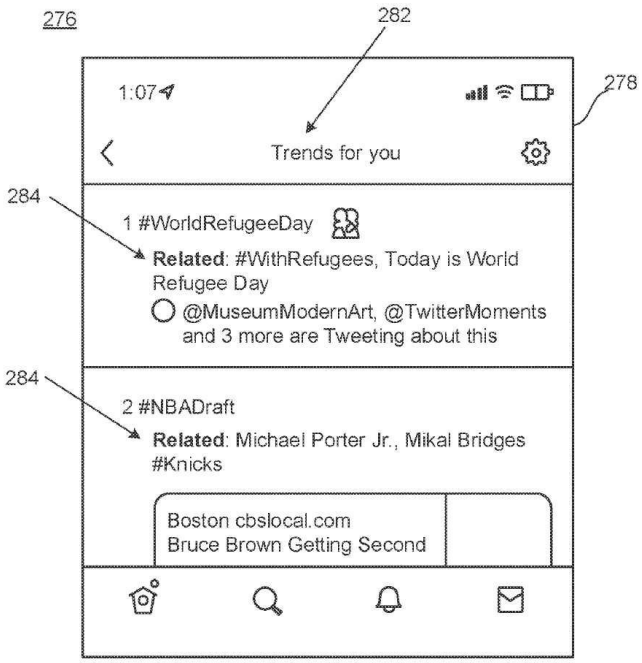
도면1d



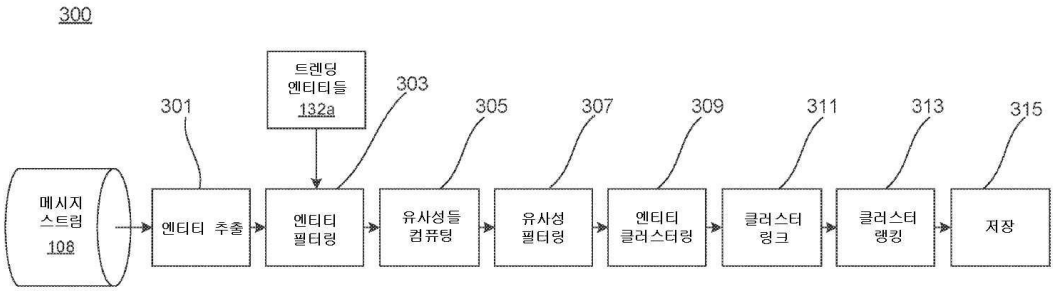
도면1e



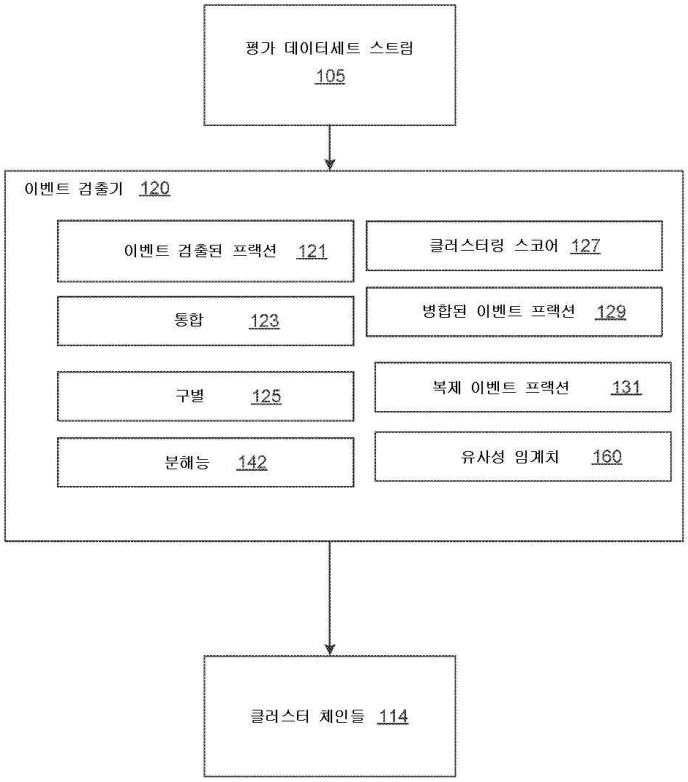
도면2



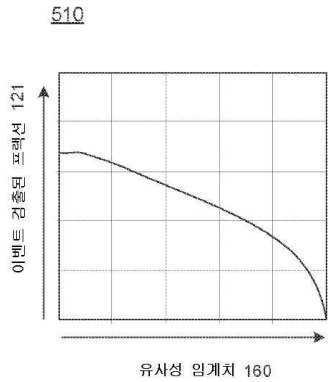
도면3


300
301
트렌딩 엔티티들 132a
303
305
307
309
311
313
315
메시지 스트림 108
엔티티 추출
엔티티 필터링
유사성들 컴퓨팅
유사성 필터링
엔티티 클러스터링
클러스터 링크
클러스터 랭킹
저장

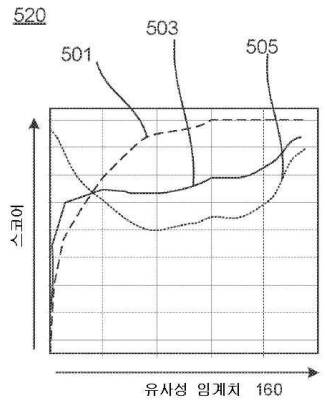
도면4


평가 데이터세트 스트림 105
이벤트 검출기 120
이벤트 검출된 프랙션 121
클러스터링 스코어 127
통합 123
병합된 이벤트 프랙션 129
구별 125
복제 이벤트 프랙션 131
분해능 142
유사성 임계치 160
클러스터 체인들 114

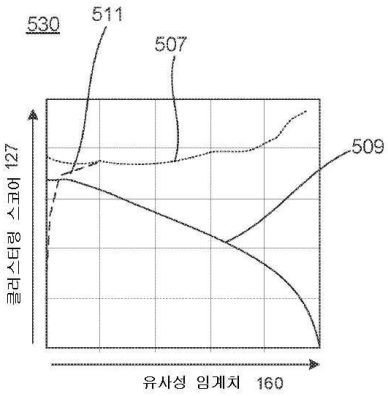
도면5a


510
이벤트 검출된 프랙션 121
유사성 임계치 160

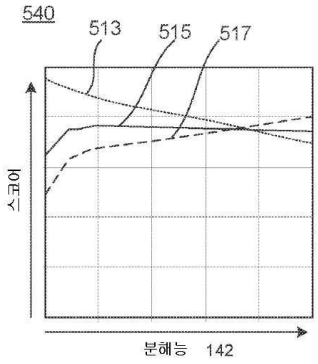
도면5b



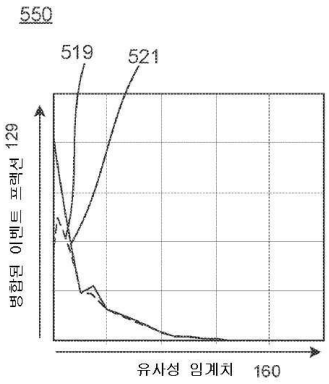
도면5c



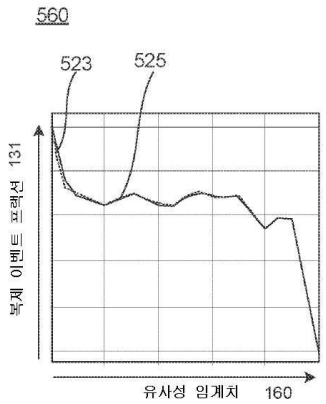
도면5d



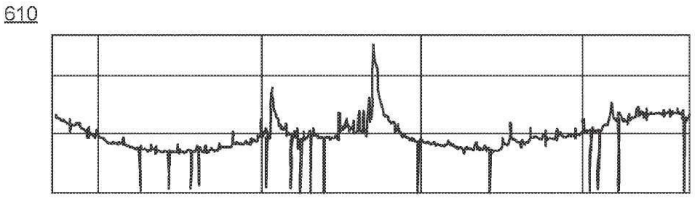
도면5e



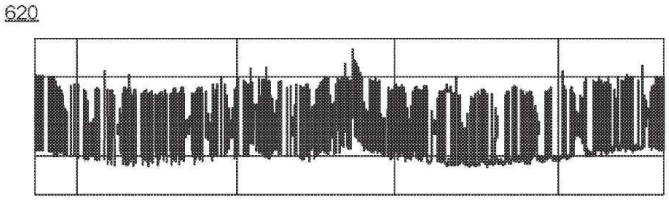
도면5f



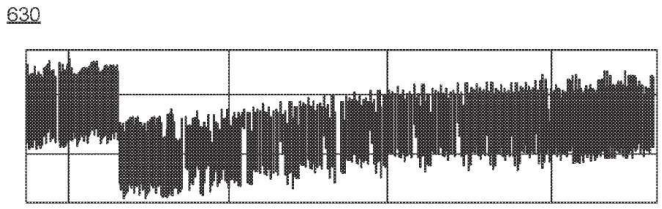
도면6a



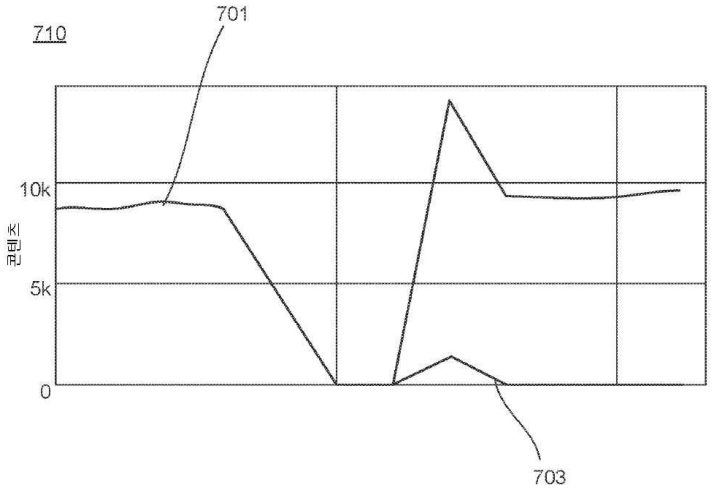
도면6b



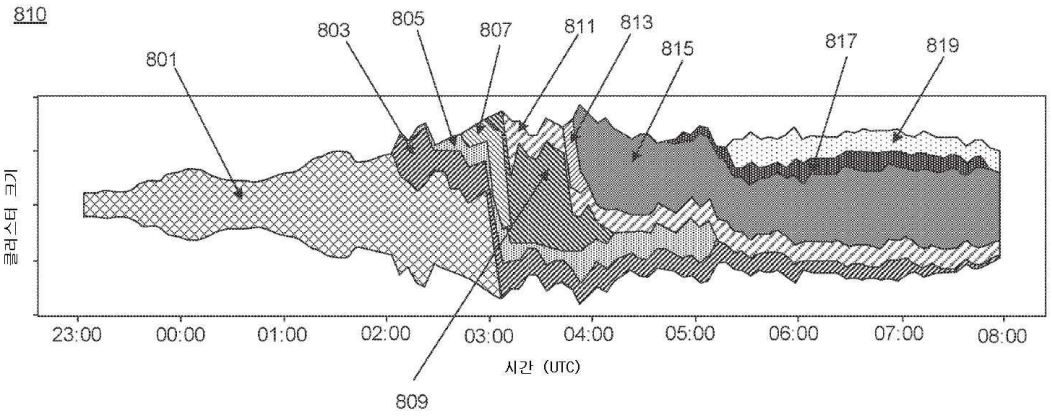
도면6c


630

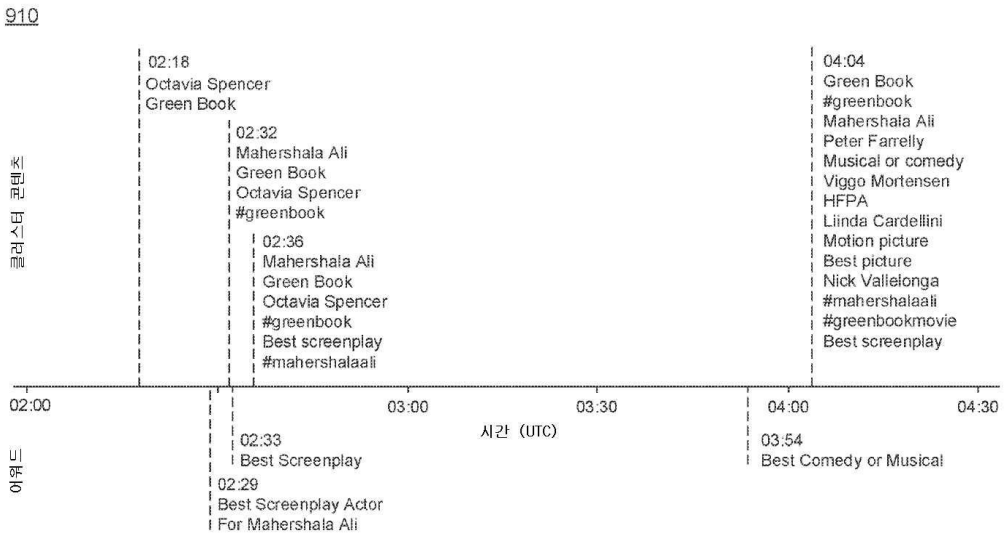
도면7


710
701
10k
5k
0
콘텐츠
703

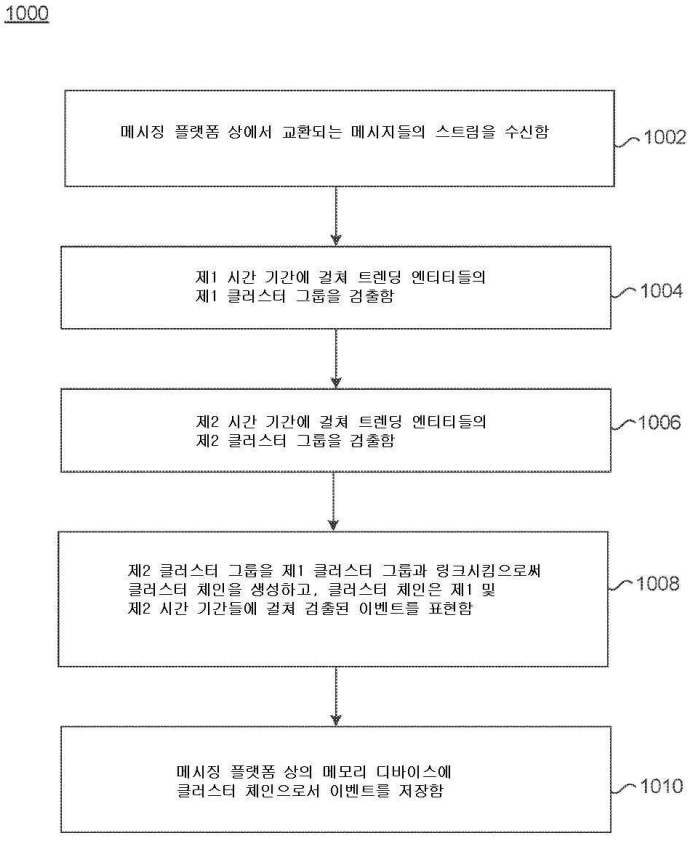
도면8


810
801
803
805
807
811
813
815
817
819
클러스터 크기
23:00
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
시간 (UTC)
809

도면9


910
클러스터 콘텐츠
02:18
Octavia Spencer
Green Book
02:32
Mahershala Ali
Green Book
Octavia Spencer
#greenbook
02:36
Mahershala Ali
Green Book
Octavia Spencer
#greenbook
Best screenplay
#mahershalaali
04:04
Green Book
#greenbook
Mahershala Ali
Peter Farrelly
Musical or comedy
Viggo Mortensen
HFPA
Liinda Cardellini
Motion picture
Best picture
Nick Vallelonga
#mahershalaali
#greenbookmovie
Best screenplay
02:00
03:00
03:30
04:00
04:30
시간 (UTC)
이벤트
02:33
Best Screenplay
02:29
Best Screenplay Actor
For Mahershala Ali
03:54
Best Comedy or Musical

도면10


1000
메시징 플랫폼 상에서 교환되는 메시지들의 스트림을 수신함
1002
제1 시간 기간에 걸쳐 트렌딩 엔티티들의
제1 클러스터 그룹을 검출함
1004
제2 시간 기간에 걸쳐 트렌딩 엔티티들의
제2 클러스터 그룹을 검출함
1006
제2 클러스터 그룹을 제1 클러스터 그룹과 링크시킴으로써
클러스터 체인을 생성하고, 클러스터 체인은 제1 및
제2 시간 기간들에 걸쳐 검출된 이벤트를 표현함
1008
메시징 플랫폼 상의 메모리 디바이스에
클러스터 체인으로서 이벤트를 저장함
1010