



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0024575
(43) 공개일자 2020년03월09일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01) G06F 16/00 (2019.01)
(52) CPC특허분류
G06F 40/258 (2020.01)
G06F 16/313 (2019.01)
(21) 출원번호 10-2018-0101429
(22) 출원일자 2018년08월28일
심사청구일자 2018년08월28일

(71) 출원인
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
박정희
대전광역시 유성구 은구비남로 56, 908동 1803호
최혁준
대전광역시 유성구 궁동로18번길 17, 202호
(74) 대리인
홍성욱, 심경식

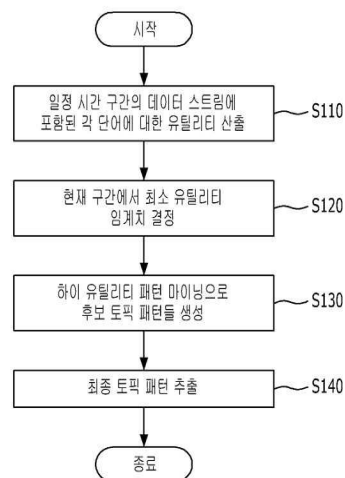
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법 및 그 장치

(57) 요약

소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법 및 그 장치가 개시된다. 본 발명의 일 실시예에 따른 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법은 상기 데이터 스트림에 포함된 각 단어에 대한 유틸리티(utility)를 산출하는 단계, 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 결정하는 단계, 상기 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하는 단계, 상기 후보 토픽 패턴들 중에서 최종 토픽 패턴을 추출하는 단계를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06F 16/3335 (2019.01)

G06F 16/3344 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 NRF-2015R1D1A1A01056622

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 이공분야기초연구-기본연구

연구과제명 스트리밍 데이터 마이닝을 위한 시스템 개발

기 여 율 1/1

주관기관 충남대학교 산학협력단

연구기간 2015.11.01 ~ 2018.10.31

명세서

청구범위

청구항 1

토픽 탐지 장치가 소셜 네트워크 서비스의 데이터 스트림에서 토픽을 탐지하는 방법에 있어서,
 상기 데이터 스트림에 포함된 각 단어에 대한 유틸리티(utility)를 산출하는 단계;
 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 결정하는 단계;
 상기 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하는 단계; 및
 상기 후보 토픽 패턴들중에서 최종 토픽 패턴을 추출하는 단계
 를 포함하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 2

제1항에 있어서,
 상기 유틸리티(utility)를 산출하는 단계 이전에,
 일정 시간 간격으로 상기 소셜 네트워크 서비스의 데이터 스트림을 수집하는 단계를 더 포함하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 3

제1항에 있어서,
 상기 유틸리티를 산출하는 단계는,
 상기 데이터 스트림에 슬라이딩 윈도우 기법을 적용하여 적어도 하나 이상의 트랜잭션을 포함하는 복수의 배치(batch)로 구분하는 단계;
 상기 복수의 배치 중 가장 최근 배치에 포함된 각 트랜잭션별로 단어 및 각 단어의 빈도수를 획득하여, 트랜잭션 테이블을 생성하는 단계; 및
 상기 가장 최근 배치에 포함된 각 단어의 빈도수에 기초하여 각 단어에 대한 가중치를 산출하여 유틸리티 테이블을 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 4

제3항에 있어서,
 상기 유틸리티 테이블을 생성하는 단계는,
 상기 가장 최근 배치에 포함된 각 단어의 최근 빈도수와 이전 배치에서의 과거 빈도수를 상기 트랜잭션 테이블로부터 획득하는 단계;
 상기 가장 최근 배치에 포함된 각 단어에 대해, 최근 빈도수와 과거 빈도수의 빈도수 차이 및 빈도수 증가율을 산출하는 단계;
 상기 빈도수 차이 및 빈도수 증가율에 기초하여 상기 가장 최근 배치에 포함된 각 단어에 대한 가중치를 각각 산출하는 단계; 및
 상기 각 단어에 대한 가중치를 이용하여 유틸리티 테이블을 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 5

제4항에 있어서,

상기 각 단어에 대한 가중치는 아래 수학적식으로 산출되는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

[수학적식]

$$W(i) = \begin{cases} diff(i) \times \log(rate(i)) & , if\ diff(i) > 0 \\ 0 & , else \end{cases}$$

여기서, $W(i)$ 는 단어 i 에 대한 가중치, $diff(i)$ 는 빈도수 차이, $rate(i)$ 는 빈도수 증가율을 의미함.

청구항 6

제3항에 있어서,

상기 최소 유틸리티 임계치를 결정하는 단계는,

상기 가장 최근 배치에 포함된 단어들 중에서 가중치가 높은 순으로 일정 개수의 단어들을 선택하는 단계;

상기 선택된 단어 각각에 대해, 항목집합을 포함하는 트랜잭션에 속한 항목들의 유틸리티 평균(α), 상기 항목집합을 포함하는 트랜잭션의 평균 길이(β), 상기 항목집합을 포함하는 트랜잭션 개수(γ)를 각각 산출하는 단계; 및

상기 선택된 단어들에 대한 유틸리티 평균(α), 트랜잭션의 평균 길이(β), 트랜잭션 개수(γ)의 평균을 각각 산출하여, 최소 유틸리티 임계치를 산출하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 7

제6항에 있어서,

상기 최소 유틸리티 임계치(min-util)는 아래 수학적식으로 산출되는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

[수학적식]

$$min-util = avg(\alpha) \times avg(\beta) \times avg(\gamma)$$

여기서, $avg(\alpha)$ 는 α 의 평균, $avg(\beta)$ 는 β 의 평균, $avg(\gamma)$ 는 γ 의 평균을 의미함.

청구항 8

제6항에 있어서,

상기 후보 토픽 패턴을 생성하는 단계는,

상기 가장 최근 배치에서 선택된 단어를 포함하는 단어 집합에 대한 트랜잭션-가중치 유틸리티(TWU)를 산출하는 단계; 및

상기 트랜잭션-가중치 유틸리티가 최소 유틸리티 임계치 이상인 단어 집합을 후보 토픽 패턴으로 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 9

제1항에 있어서,

상기 최종 토픽 패턴을 추출하는 단계는

상기 후보 토픽 패턴들로부터 트리를 구축하는 단계;

상기 구축된 트리의 루트노드에서 리프노드까지 경로를 탐색하여 패턴을 추출하는 단계;

상기 추출된 각 패턴들의 유틸리티를 각각 산출하는 단계; 및

상기 패턴 유틸리티 값을 내림차순으로 정렬하여 상위 일정 개수를 최종 토픽 패턴으로 추출하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 10

제9항에 있어서,

상기 후보 토픽 패턴들로부터 트리를 구축하는 단계는,

제1 후보 토픽 패턴의 제1 프리픽스 노드가 헤더 테이블에 존재하는지를 판단하는 단계;

상기 제1 프리픽스 노드가 헤더 테이블에 존재하지 않은 경우, 상기 제1 후보 토픽 패턴을 루트 노드의 서브트리로 추가하고, 상기 제1 프리픽스 노드가 헤더 테이블에 존재하는 경우, 상기 제1 프리픽스 노드의 헤더 테이블로부터 연결된 경로의 서브트리에 나머지 노드의 존재 여부를 판단하고, 나머지 노드가 존재할 경우 삽입을 종료하고, 나머지 경로가 존재하지 않을 경우 경로 마지막 노드의 서브 트리로 나머지 노드를 추가하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

청구항 11

제9항에 있어서,

상기 패턴 유틸리티(PU)는 아래 수학적식을 이용하여 산출하는 것을 특징으로 하는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법.

[수학적식]

$$PU(p) = \sum_{i \in p} eu(i)$$

여기서, p는 패턴, eu(i)는 p패턴에 속한 단어 i에 대한 가중치를 의미함.

청구항 12

소셜 네트워크 서비스의 데이터 스트림에서 토픽을 탐지하는 토픽 탐지 장치에 있어서,

통신망을 통해 소셜 네트워크 서비스를 제공하는 서버와의 통신을 위한 통신부;

일정 시간 간격으로 소셜 네트워크 서비스의 데이터 스트림을 수집하고, 상기 데이터 스트림에 포함된 각 단어에 대한 유틸리티를 산출하며, 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 동적으로 결정하고, 상기 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하며, 상기 후보 토픽 패턴들중에서 최종 토픽 패턴을 추출하는 토픽 패턴 추출부

를 포함하는 토픽 탐지 장치.

발명의 설명

기술 분야

[0001] 본 발명은 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법 및 그 장치에 관한 것으로, 특히 소셜 네트워크 서비스의 데이터 스트림의 전체 트랜잭션에서 항목집합(item set)의 빈도수와 가치(utility)를 고려한 하이 유틸리티 패턴 마이닝(High-utility pattern mining)을 이용하여 토픽 패턴을 추출하는 방법 및 장치에 관한 것이다.

배경 기술

- [0002] 최근 인터넷의 발전과 스마트 기기의 광범위한 보급과 함께 트위터, 페이스북과 같은 소셜 미디어가 급성장하기 시작했다. 특히 트위터의 경우 한 포스팅 당 작성 가능한 글자 수가 제한되어 있다는 특징과 단순한 인터페이스를 갖는 장점으로 인해 현재는 가장 대표적인 소셜 미디어 중의 하나로써 성장하였다. 2016년 기준 월간 3억 명 이상의 사용자들이 이용하고 있으며, 매달 5억 건 이상의 트윗들이 새롭게 생성되고 있다. 트위터 사용자의 대부분은 스마트 기기를 통해 트위터에 접근하기 때문에 세계 각지에서 발생하는 사건들을 온라인상에서 실시간으로 전파하고 있다. 학계에서는 이러한 특징에 주목하여 트위터에서 가치 있는 정보를 발굴하기 위한 다양한 연구를 진행하고 있으며, 특히 트위터에서 “현재 화제가 되는 사건은 무엇인가?”를 탐지하기 위한 연구들이 활발히 진행되고 있다.
- [0003] 문서 집합으로부터 화제가 되는 사건을 추출하여 유용한 정보의 형태로 요약하는 과정을 토픽 탐지(Topic detection)라고 하며, 전통적으로는 토픽 탐지 및 추적(Topic Detection and Tracking, TDT)이라는 주제로 연구가 진행되어 왔다. 전통적인 방법들의 대부분은 토픽을 탐지하기 위해 뉴스 기사처럼 비교적 정형화된 틀을 갖추고 있고, 긴 문장으로 이루어진 문서 집합에 대한 토픽 탐지 방법을 제안해왔다. 하지만 이러한 전통적인 접근 방법은 트위터에서 생성되는 문서들에 대해서는 적용하기 힘들다는 문제점이 있다. 트위터에서 생성되는 문서들은 매우 짧은 문장으로 구성되어 있으며, 맞춤법이 지켜지지 않은 채로 작성되는 경우가 많다. 더군다나, 문서 내에 문장이 아닌 특수문자, URL과 같은 정보를 포함하는 경우가 많기 때문에 생성되는 전체 문서에 비해 유용한 정보를 지닌 문서의 개수가 매우 적은 비율을 차지하고 있다. 또한, 발생하는 사건들이 다양하여 문서 간의 연관성을 찾기 힘들고, 이로 인해 토픽이 무엇인지 쉽게 도출하기 힘들다는 문제점을 갖고 있다. 마지막으로 스트림에서 대용량의 문서들이 생성되고 있기 때문에 실시간으로 토픽을 탐지하기 위한 접근 방법이 필요하다.
- [0004] 일반적으로 트위터에서 토픽을 탐지하기 위해 제안된 방법들은 크게 3가지로 분류된다. 먼저, 단어 중심(Feature-pivot) 토픽 탐지 방법은 문서 집합에서 동시에 등장하는 단어 그룹을 찾아내어 토픽을 탐지하는 방법이며, 문서 중심(Document-pivot) 토픽 탐지 방법은 문서들 간의 유사도를 비교하고, 문서들을 군집화함으로써 토픽을 탐지하는 방법이다. 확률 토픽 모델(Probabilistic topic model)은 토픽에 대한 단어의 분포와 문서에 대한 토픽의 확률 분포를 추론하여 토픽을 탐지하는 방법이다.
- [0005] 그러나, 이러한 방법들은 매개 변수 설정이 어렵고, 토픽 탐지를 위한 계산 과정이 오래 걸리기 때문에 스트림에서 토픽을 탐지하기에 적합하지 않은 문제가 있다.

선행기술문헌

특허문헌

- [0006] (특허문헌 0001) 한국 등록특허공보 제10-1702559호(2017.02.03.)

발명의 내용

해결하려는 과제

- [0007] 본 발명이 해결하고자 하는 과제는 트위터와 같은 소셜 네트워크 서비스의 데이터 스트림에서 실시간으로 토픽을 탐지할 수 있는 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법 및 그 장치를 제공하는 것이다.
- [0008] 본 발명이 해결하고자 하는 과제는 이상에서 언급한 과제(들)로 제한되지 않으며, 언급되지 않은 또 다른 과제(들)는 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

- [0009] 상기한 목적을 달성하기 위하여, 본 발명의 일 실시예에 따른 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법은, 토픽 탐지 장치가 소셜 네트워크 서비스의 데이터 스트림에서 토픽을 탐지하는 방법에 있어서, 상

기 데이터 스트림에 포함된 각 단어에 대한 유틸리티(utility)를 산출하는 단계, 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 결정하는 단계, 상기 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하는 단계, 상기 후보 토픽 패턴들중에서 최종 토픽 패턴을 추출하는 단계를 포함한다.

[0010] 바람직하게는, 상기 유틸리티(utility)를 산출하는 단계 이전에, 일정 시간 간격으로 상기 소셜 네트워크 서비스의 데이터 스트림을 수집하는 단계를 더 포함할 수 있다.

[0011] 바람직하게는, 상기 유틸리티를 산출하는 단계는, 상기 데이터 스트림에 슬라이딩 윈도우 기법을 적용하여 적어도 하나 이상의 트랜잭션을 포함하는 복수의 배치(batch)로 구분하는 단계, 상기 복수의 배치 중 가장 최근 배치에 포함된 각 트랜잭션별로 단어 및 각 단어의 빈도수를 획득하여, 트랜잭션 테이블을 생성하는 단계, 상기 가장 최근 배치에 포함된 각 단어의 빈도수에 기초하여 각 단어에 대한 가중치를 산출하여 유틸리티 테이블을 생성하는 단계를 포함할 수 있다.

[0012] 바람직하게는, 상기 유틸리티 테이블을 생성하는 단계는, 상기 가장 최근 배치에 포함된 각 단어의 최근 빈도수와 이전 배치에서의 과거 빈도수를 상기 트랜잭션 테이블로부터 획득하는 단계, 상기 가장 최근 배치에 포함된 각 단어에 대해, 최근 빈도수와 과거 빈도수의 빈도수 차이 및 빈도수 증가율을 산출하는 단계, 상기 빈도수 차이 및 빈도수 증가율에 기초하여 상기 가장 최근 배치에 포함된 각 단어에 대한 가중치를 각각 산출하는 단계, 상기 각 단어에 대한 가중치를 이용하여 유틸리티 테이블을 생성하는 단계를 포함할 수 있다.

[0013] 바람직하게는, 상기 각 단어에 대한 가중치는 아래 수학식으로 산출될 수 있다.

[0014] [수학식]

$$W(i) = \begin{cases} diff(i) \times \log(rate(i)) & , if \ diff(i) > 0 \\ 0 & , else \end{cases}$$

[0015]

[0016] 여기서, $W(i)$ 는 단어 i 에 대한 가중치, $diff(i)$ 는 빈도수 차이, $rate(i)$ 는 빈도수 증가율을 의미함.

[0017] 바람직하게는, 상기 최소 유틸리티 임계치를 결정하는 단계는, 상기 가장 최근 배치에 포함된 단어들 중에서 가중치가 높은 순으로 일정 개수의 단어들을 선택하는 단계, 상기 선택된 단어 각각에 대해, 항목집합을 포함하는 트랜잭션에 속한 항목들의 유틸리티 평균(α), 상기 항목집합을 포함하는 트랜잭션의 평균 길이(β), 상기 항목집합을 포함하는 트랜잭션 개수(γ)를 각각 산출하는 단계, 상기 선택된 단어들에 대한 유틸리티 평균(α), 트랜잭션의 평균 길이(β), 트랜잭션 개수(γ)의 평균을 각각 산출하여, 최소 유틸리티 임계치를 산출하는 단계를 포함할 수 있다.

[0018] 바람직하게는, 상기 최소 유틸리티 임계치(min-util)는 아래 수학식으로 산출될 수 있다.

[0019] [수학식]

$$min-util = avg(\alpha) \times avg(\beta) \times avg(\gamma)$$

[0020]

[0021] 여기서, $avg(\alpha)$ 는 α 의 평균, $avg(\beta)$ 는 β 의 평균, $avg(\gamma)$ 는 γ 의 평균을 의미함.

[0022] 바람직하게는, 상기 후보 토픽 패턴을 생성하는 단계는, 상기 가장 최근 배치에서 선택된 단어를 포함하는 단어 집합에 대한 트랜잭션-가중치 유틸리티(TWU)를 산출하는 단계, 상기 트랜잭션-가중치 유틸리티가 최소 유틸리티 임계치 이상인 단어 집합을 후보 토픽 패턴으로 생성하는 단계를 포함할 수 있다.

[0023] 바람직하게는, 상기 최종 토픽 패턴을 추출하는 단계는 상기 후보 토픽 패턴들로부터 트리를 구축하는 단계, 상기 구축된 트리의 루트노드에서 리프노드까지 경로를 탐색하여 패턴을 추출하는 단계, 상기 추출된 각 패턴들의 유틸리티를 각각 산출하는 단계, 상기 패턴 유틸리티 값을 내림차순으로 정렬하여 상위 일정 개수를 최종 토픽 패턴으로 추출하는 단계를 포함할 수 있다.

[0024] 바람직하게는, 상기 후보 토픽 패턴들로부터 트리를 구축하는 단계는, 제1 후보 토픽 패턴의 제1 프리픽스 노드가 헤더 테이블에 존재하는지를 판단하는 단계, 상기 제1 프리픽스 노드가 헤더 테이블에 존재하지 않은 경우, 상기 제1 후보 토픽 패턴을 루트 노드의 서브트리로 추가하고, 상기 제1 프리픽스 노드가 헤더 테이블에 존재하

는 경우, 상기 제1 프리픽스 노드의 헤더 테이블로부터 연결된 경로의 서브트리에 나머지 노드의 존재 여부를 판단하고, 나머지 노드가 존재할 경우 삽입을 종료하고, 나머지 경로가 존재하지 않을 경우 경로 마지막 노드의 서브 트리로 나머지 노드를 추가하는 단계를 포함할 수 있다.

[0025] 바람직하게는, 상기 패턴 유틸리티(PU)는 아래 수학적식을 이용하여 산출할 수 있다.

[0026] [수학적식]

$$PU(p) = \sum_{i \in p} eu(i)$$

[0027]

[0028] 여기서, p는 패턴, eu(i)는 p패턴에 속한 단어 i에 대한 가중치를 의미함.

[0029] 상기한 목적을 달성하기 위하여, 본 발명의 다른 실시예에 따른 토픽 탐지 장치는 소셜 네트워크 서비스의 데이터 스트림에서 토픽을 탐지하는 토픽 탐지 장치에 있어서, 통신망을 통해 소셜 네트워크 서비스를 제공하는 서버와의 통신을 위한 통신부, 일정 시간 간격으로 소셜 네트워크 서비스의 데이터 스트림을 수집하고, 상기 데이터 스트림에 포함된 각 단어에 대한 유틸리티를 산출하며, 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 동적으로 결정하고, 상기 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하며, 상기 후보 토픽 패턴들중에서 최종 토픽 패턴을 추출하는 토픽 패턴 추출부를 포함한다.

발명의 효과

[0030] 본 발명에 따르면, 항목집합(item set)의 빈도수와 가치(Utility)를 고려한 하이 유틸리티 패턴 마이닝(High-utility pattern mining)을 이용함으로써, 트위터와 같은 소셜 네트워크 서비스의 데이터 스트림에서 실시간으로 토픽을 탐지할 수 있다.

[0031] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 일 실시예에 따른 데이터 스트림에서 토픽을 탐지하는 방법을 설명하기 위한 흐름도이다.

도 2는 본 발명의 일 실시예에 따른 각 단어에 대한 유틸리티를 산출하는 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 트위터 스트림과 트랜잭션 테이블을 설명하기 위한 예시도이다.

도 4는 본 발명의 일 실시예에 따른 유틸리티 테이블을 구성하는 방법을 설명하기 위한 흐름도이다.

도 5는 본 발명의 일 실시예에 따른 최소 유틸리티 임계치를 결정하는 방법을 설명하기 위한 흐름도이다.

도 6은 본 발명의 일 실시예에 따른 후보 토픽 패턴 생성 방법을 설명하기 위한 도면이다.

도 7은 본 발명의 일 실시예에 따른 토픽 패턴 추출 방법을 설명하기 위한 흐름도이다.

도 8은 본 발명의 일 실시예에 따른 트리를 구축하는 방법을 설명하기 위한 예시도이다.

도 9는 본 발명의 일 실시예에 따른 트위터 스트림에서 토픽을 탐지하는 토픽 탐지 장치의 블록도이다.

도 10은 매개변수 선택에 따른 성능 비교 결과를 나타낸다.

도 11은 토픽 탐지에 소요되는 평균 시간을 측정한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0033] 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 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설

명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

- [0034] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0035] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [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 명세서상에 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 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0038] 본 발명은 블로그, 트위터, 페이스북 등과 같은 SNS, 전자 신문 등 실시간으로 업데이트되는 데이터 스트림에서 빈번하게 등장하면서 가치(유틸리티)가 높은 항목집합(패턴)을 찾는 기술에 관한 것이다. 이때, 데이터 스트림은 SNS, 블로그, 트위터, 신문 등 실시간으로 업데이트되는 모든 정보를 포함하나, 이하에서는 설명의 편의를 위해 트위터 스트림을 예로하여 설명하기로 한다. 트위터에서의 토픽 탐지 관점에서 보았을 때, 하나의 트윗은 단어를 항목으로 갖는 트랜잭션으로 볼 수 있으며, 하이 유틸리티 패턴 마이닝을 통해 생성되는 패턴들은 가치 있는 단어들로 구성된 집합을 도출하는 과정이라고 볼 수 있다. 이때, 가치 있는 단어들이란 토픽으로 구성된 만한 핵심단어(Keyword)들을 의미하고, 주요한 핵심단어들로 구성되는 패턴은 결과적으로 토픽을 의미하게 된다.
- [0039] 이하에서는 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0040] 도 1은 본 발명의 일 실시예에 따른 소셜 네트워크 서비스의 데이터 스트림에서 토픽 탐지 방법을 설명하기 위한 흐름도이다.
- [0041] 도 1을 참조하면, 토픽 탐지 장치는 일정 시간 구간의 데이터 스트림에 포함된 각 단어에 대한 유틸리티를 산출한다(S110). 즉, 토픽 탐지 장치는 트위터 스트림에서 슬라이딩 윈도우 기법을 적용하여 최초에 발생한 트윗들의 윈도우에서 트랜잭션 테이블을 구성하고, 트랜잭션 테이블에 포함된 각 단어의 가중치(가치, 유틸리티, utility)를 계산하여 유틸리티 테이블을 구성한다. 다시 말하면, 토픽 탐지 장치는 슬라이딩 윈도우 기법과 단어 가중치 계산을 통한 트랜잭션 데이터베이스를 구성한다. 여기서, 트랜잭션 데이터베이스는 트랜잭션 테이블과 유틸리티 테이블을 포함할 수 있다.
- [0042] 토픽 탐지 장치가 각 단어에 대한 유틸리티를 산출하는 방법에 대한 상세한 설명은 도 2를 참조하기로 한다.
- [0043] 단계 S110이 수행되면, 토픽 탐지 장치는 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치(Minimum utility threshold, min-unit)를 결정한다(S120). 즉, 토픽 탐지 장치는 변화하는 트랜잭션 데이터베이스에 따라서 최소 유틸리티 임계치를 동적으로 결정한다. 다시 말하면, 토픽 탐지 장치는 각 단어의 가중치에 기초하여 선택된 핵심단어들의 트랜잭션-가중치 유틸리티를 이용하여 최소 유틸리티 임계치를 동적으로 결정한다.
- [0044] 토픽 탐지 장치가 최소 유틸리티 임계치를 결정하는 방법에 대한 상세한 설명은 도 5를 참조하기로 한다.
- [0045] 단계 S120이 수행되면, 토픽 탐지 장치는 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성하고(S130), 후보 토픽 패턴들중에서 최종 토픽 패턴을 추출한다(S140). 즉, 토픽 탐지 장치는 각 단어 집합에 대한 트랜잭션-가중치 유틸리티를 계산하고, 트랜잭션-가중치 유틸리티의 값이 최소 유틸리

티 임계치 이상인 단어 집합을 후보 토픽 패턴으로 생성할 수 있다. 그런 후, 토픽 탐지 장치는 후보 토픽 패턴에 대해 TP-Tree를 구축하여 실제 토픽이 되는 최종 토픽 패턴들을 추출한다.

[0046] 토픽 탐지 장치가 후보 토픽 패턴을 생성하는 방법에 대한 상세한 설명은 도 6을 참조하고, 최종 토픽 패턴을 추출하는 방법에 대한 상세한 설명은 도 7을 참조하기로 한다.

[0047] 도 2는 본 발명의 일 실시예에 따른 각 단어에 대한 유틸리티를 산출하는 방법을 설명하기 위한 흐름도, 도 3은 본 발명의 일 실시예에 따른 트위터 스트림과 트랜잭션 테이블을 설명하기 위한 예시도, 도 4는 본 발명의 일 실시예에 따른 유틸리티 테이블을 구성하는 방법을 설명하기 위한 흐름도이다.

[0048] 도 2를 참조하면, 토픽 탐지 장치는 트위터 스트림에 일정한 시간 간격으로 겹치지 않는 윈도우를 사용하여 복수의 배치(batch)를 생성한다(S210).

[0049] 예를 들어, 시간 순서에 따라 발생하는 트윗을 T_i 라 하고, 트위터 스트림을 $TS = \{T_1, T_2, T_3, \dots\}$ 라고 가정하여 설명하기로 한다. 이 경우, 트위터 스트림 TS에 일정한 시간 간격으로 겹치지 않는 윈도우를 사용하여 슬라이딩 윈도우 기법을 적용하면 트위터 스트림은 다수의 배치들로 구성된다. 즉, 트위터 스트림은 일정한 시간 간격을 갖는, 겹치지 않은 다수의 배치들($B_1, B_2, B_3, \dots, B_L$)로 구성되고, 각 배치는 적어도 하나 이상의 트윗을 포함한다. 예컨대, 도 3의 (a)를 참조하면, 트위터 스트림은 다수의 배치들($B_1, B_2, B_3, \dots, B_L$)로 구성되고, B_L 의 경우 T_i, T_{i+1}, T_{i+2} 의 트윗을 포함한다.

[0050] 단계 S210이 수행되면, 토픽 탐지 장치는 현재(가장 최근)의 배치에 포함된 각 트윗별로 단어 및 각 단어의 빈도수를 획득하여, 트랜잭션 테이블을 구성한다(S220). 예를 들어, 도 3의 (b)에서와 같이 현재 배치 B_L 지점에 있다고 가정한 경우, B_L 에 포함된 각각의 트윗은 항목 집합으로 갖는 트랜잭션 T로 볼 수 있다. 따라서, 배치 B_L 에 속하는, 모든 트윗에 나타난 단어들과 각 단어의 빈도수를 이용하면 트랜잭션 테이블을 구성할 수 있다. 이때, 트랜잭션에서 발생한 각 단어의 빈도수를 내부 유틸리티라고 칭할 수 있고, 항목 i에 대한 내부 유틸리티(Internal utility)는 트랜잭션에서 항목에 대한 수량을 나타내며, $iu(i, T)$ 로 표기할 수 있다. 즉, 항목(item)은 배치 B_L 에서 발생한 단어를 의미하고, 내부 유틸리티는 각 트랜잭션에서 발생한 단어 i의 빈도수일 수 있다.

[0051] 트랜잭션 테이블 구성 시, 토픽 탐지 장치는 최근 배치인 [B_{L-1}, B_L]을 유지하여 사용하며, B_L 에서 발생한 트윗들에 대해 트랜잭션 테이블을 구성할 수 있다.

[0052] 트랜잭션 테이블은 아래 표 1과 같을 수 있다.

[0053] [표 1]

Tid	Transaction	Count
T_1	{a, b, c, d}	{2, 1, 3, 1}
T_2	{c, e, f}	{2, 1, 1}
T_3	{a, b, d, e}	{5, 1, 1, 2}
T_4	{a, b, c, f}	{3, 1, 1, 2}
T_5	{d, e, g}	{1, 1, 5}
T_6	{b, c, d, g}	{1, 1, 1, 2}
T_7	{a, c, d}	{4, 2, 1}

[0054]

[0055] 표 1을 살펴보면, 트랜잭션 테이블은 트랜잭션 식별번호(Tid), 트랜잭션의 항목(Transaction), 각 항목에 대한 수량(Count)를 포함할 수 있다.

[0056] 단계 S220이 수행되면, 토픽 탐지 장치는 가장 최근 배치에 포함된 각 단어의 빈도수에 기초하여 각 단어에 대

한 가중치를 산출하여 유틸리티 테이블을 구성한다(S230).

[0057] 유틸리티 테이블은 트랜잭션 테이블에서 발생한 항목들의 외부 유틸리티 정보를 포함하고 있다. 외부 유틸리티는 하이 유틸리티 패턴 마이닝에서 각 항목에 대한 가치를 나타내는 수치로, 외부 유틸리티 값에 따라 생성된 하이 유틸리티 패턴의 결과가 크게 달라질 수 있다. 토픽 탐지 관점에서 유틸리티 테이블의 구성은 단어들에 대한 특정한 가중치를 부여하는 것과 동일한 단계로 볼 수 있다. 슬라이딩 윈도우 기법을 적용하여 트랜잭션 테이블을 구성하게 되면, 트랜잭션 테이블에서 발생한 단어들과 빈도수 정보가 배치마다 변화하게 되므로, 유틸리티 테이블 또한 새롭게 구성해주어야 한다. 트위터는 최대 작성 가능한 글자 수가 제한되어있기 때문에 화제가 되는 사건이 발생할 경우 특정한 단어들의 빈도수가 급격히 증가하는 특징을 갖고 있다. 따라서 단어들에 대한 과거 빈도수 정보와 최근 빈도수 정보를 활용하면 각각의 단어들에 대한 가중치를 계산할 수 있다. 단어에 대한 과거 빈도수 정보와 최근 빈도수 정보를 활용하기 위해, 가장 최근의 배치인 $[B_{L-1}, B_L]$ 을 유지하여 사용한다. 단어 가중치에 대한 계산은 가장 최근 배치인 B_L 에서 나타난 단어들을 기준으로 계산하며, 계산 과정에서는 과거 빈도수와 최근 빈도수에 대한 차이와 증가율을 이용한다.

[0058] 유틸리티 테이블을 구성하는 방법에 대해 도 4를 참조하면, 토픽 탐지 장치는 가장 최근 배치 B_L 에 포함된 각 단어의 최근 빈도수와 이전 배치 B_{L-1} 에서의 과거 빈도수를 트랜잭션 테이블로부터 획득한다(S232).

[0059] 그런 후, 토픽 탐지 장치는 가장 최근 배치 B_L 에 포함된 각 단어에 대해, 최근 빈도수와 과거 빈도수의 빈도수 차이 및 빈도수 증가율을 산출한다(S234).

[0060] 배치 B_L 에 대한 단어 i 의 빈도수를 $f(B_L, i)$, 배치 B_{L-1} 에 대한 단어 i 의 빈도수를 $f(B_{L-1}, i)$ 라고 할 때, 빈도수 차이 $diff(i)$ 는 아래 수학적 식 1과 같이 산출할 수 있다.

[0061] [수학적 식 1]

$$diff(i) = f(B_L, i) - f(B_{L-1}, i)$$

[0062]

$diff(i)$ 는 가장 최근 배치인 B_L 과 그 이전 배치인 B_{L-1} 에서 발생한 단어 i 의 빈도수 차이를 의미한다.

$diff(i) > 0$ 일 경우 단어 i 의 빈도수가 현재 증가상태에 있음을 의미하며, $diff(i) < 0$ 일 경우 단어의 빈도수가 감소상태에 있음을 의미한다.

[0064] 빈도수 증가율은 가장 최근 배치인 B_L 과 그 이전 배치인 B_{L-1} 에서 단어 i 에 대한 증가율을 의미하며, 빈도수 증가율 $rate(i)$ 은 아래 수학적 식 2와 같이 산출할 수 있다.

[0065] [수학적 식 2]

$$rate(i) = \frac{f(B_L, i) + 1}{f(B_{L-1}, i) + 1}$$

[0066]

수학적 식 2에서 분모가 ;0'이 되는 것을 방지하기 위해 분자와 분모에 각각 1을 더할 수 있다. $rate(i) > 1$

일 경우 단어 i 의 빈도수가 증가상태에 있음을 의미하며, $rate(i) < 1$ 일 경우 단어 i 의 빈도수가 감소상태에 있음을 의미한다.

[0068] 단계 S234이 수행되면, 토픽 탐지 장치는 빈도수 차이와 빈도수 증가율을 이용하여 가장 최근 배치인 B_L 에서 발생한 각 단어에 대한 가중치를 각각 산출한다(S236). 즉, 토픽 탐지 장치는 아래 수학적 식 3을 이용하여 가장 최근 배치인 B_L 에서 발생한 단어 i 의 가중치($W(i)$)를 산출할 수 있다.

[0069] [수학식 3]

[0070]

[0071] $diff \leq 0$ 인 단어는 빈도수가 변화하지 않았거나 또는 감소상태에 있으므로 토픽으로 구성되는 핵심단어가 되기 어렵다. 따라서 해당 단어에 대해서는 가중치를 '0'으로 설정할 수 있다. $rate(i)$ 의 경우 단어 빈도수에 대한 비율을 계산하기 때문에 값이 지나치게 커질 수 있다. 따라서 $W(i)$ 의 계산에서 $rate(i)$ 값에 로그를 취하여 사용할 수 있다. 수학식 5를 사용하여 B_L 에서 발생한 단어 i 의 가중치를 계산할 수 있다.

[0072] 단계 S236에서 각 단어에 대한 가중치가 산출되면, 토픽 탐지 장치는 각 단어에 대한 가중치를 이용하여 유틸리티 테이블을 생성한다(S238).

[0073] 유틸리티 테이블은 아래 표 2와 같을 수 있다.

[0074] [표 2]

[0075]

[0076] 표 2는 각 단어에 대한 유틸리티(가중치)를 나타내는 유틸리티 테이블일 수 있다.

[0077] 다시 도 2를 참조하면, 단계 S230에서 유틸리티 테이블이 구성되면, 토픽 탐지 장치는 내부 유틸리티와 외부 유틸리티를 이용하여 트랜잭션 유틸리티 테이블을 구성한다(S240). 즉, 토픽 탐지 장치는 항목 i 에 대한 외부 유틸리티와 내부 유틸리티를 이용하여 트랜잭션 유틸리티(TU)를 산출할 수 있고, 산출된 트랜잭션 유틸리티가 포함된 트랜잭션 유틸리티 테이블을 구성할 수 있다.

[0078] 트랜잭션 유틸리티 테이블은 아래 표 3과 같을 수 있다.

[0079] [표 3]

[0080]

[0081] 표 3을 살펴보면, 트랜잭션 유틸리티 테이블은 트랜잭션 식별번호(Tid), 트랜잭션의 항목(Transaction), 각 항목에 대한 수량(Count), 트랜잭션 유틸리티(TU)를 포함할 수 있다.

[0082] 예를 들어, $I = \{i_1, i_2, \dots, i_m\}$ 를 항목 집합이라 하고, I 의 부분집합으로 구성되는 트랜잭션들의 집합을? $D = \{T_1, T_2, \dots, T_n\}$ 이라 하자. 항목 i 에 대한 외부 유틸리티(External utility)는 항목의 가치를 의미하며, $eu(i)$ 로 표기하고, 항목 i 에 대한 내부 유틸리티(Internal utility)는 트랜잭션에서 항목에 대한 수량을 나타내며, $iu(i, T)$ 로 표기할 수 있다.

[0083] $eu(i)$ 와 $iu(i, T)$ 를 사용하여 트랜잭션 T에서 항목 i에 대한 유틸리티 $u(i, T)$ 를 아래 수학적 식 4와 같이 정의할 수 있다.

[0084] [수학적 식 4]

$$u(i, T) = iu(i, T) \times eu(i)$$

[0086] 수학적 식 4를 통해 트랜잭션 T에서 항목 i에 대한 유틸리티 $u(i, T)$ 가 산출되면, 토픽 탐지 장치는 아래 수학적 식 5를 이용하여 트랜잭션 유틸리티($tu(T)$)를 산출할 수 있다.

[0087] [수학적 식 5]

$$tu(T) = \sum_{i \in T} u(i, T)$$

[0089] 여기서, $tu(T)$ 는 트랜잭션 T에 대한 유틸리티, $u(i, T)$ 는 트랜잭션 T에서 항목 i에 대한 유틸리티를 의미할 수 있다.

[0090] 예를 들어, 트랜잭션 $T_2 = \{c, e, f\}$ 에 대한 트랜잭션 유틸리티는 $tu(T_2) = 2 \times 3 + 1 \times 2 + 1 \times 3 = 11$ 이다

[0091] 도 5는 본 발명의 일 실시예에 따른 최소 유틸리티 임계치를 결정하는 방법을 설명하기 위한 흐름도이다.

[0092] 도 5를 참조하면, 토픽 탐지 장치는 가장 최근 배치에 포함된 단어들 중에서 가중치가 높은 순으로 일정 개수의 단어들을 선택한다(S510). 즉, 토픽 탐지 장치는 가장 최근 배치에 포함된 모든 단어들 중에서 높은 가중치를 가지는 $r\%$ 개의 단어를 선택한다. 단어 가중치 계산에 의한 유틸리티가 높은 단어들은 토픽으로 구성되는 핵심 단어가 될 가능성이 높은 것들이다. 이때 배치의 크기에 따라 나타나는 단어들의 수가 크게 차이가 날 수 있고, 이에 따라 선택되는 단어의 수가 너무 적거나 많을 수 있으므로, 선택되는 단어의 개수에 대한 하한(Lower-bound)값 ω 와 상한(Upper-bound)값 θ 를 설정하여 단어를 선택한다. 배치 B_L 에서 나타난 모든 단어의 수의 $r\%$ 를 S라고 하면, 선택단어수는 아래 수학적 식 6과 같이 정의될 수 있다.

[0093] [수학적 식 6]

$$\text{선택 단어 수} = \begin{cases} \omega, & \text{if } S < \omega \\ S, & \text{if } \omega \leq S \leq \theta \\ \theta, & \text{if } S > \theta \end{cases}$$

[0095] 단계 S510의 수행 후, 토픽 탐지 장치는 선택된 단어 각각에 대해 α , β , γ 를 각각 산출한다(S520). 이때, α 는 항목집합 X를 포함하는 트랜잭션에 속한 항목들의 유틸리티 평균, β 는 항목집합 X를 포함하는 트랜잭션의 평균 길이, γ 는 항목집합 X를 포함하는 트랜잭션 개수를 의미할 수 있다.

[0096] 하이 유틸리티 패턴 마이닝은 항목 집합 X에 대한 트랜잭션-가중치 유틸리티(TWU)를 계산하여 $twu(X) \geq \text{min-util}$ 일 경우 X를 하이 유틸리티 패턴으로 생성한다. 이때, min-util이 너무 작을 경우 지나치게 많은 패턴들이 생성될 수 있으며, 반대로 값이 너무 클 경우 생성되는 패턴들이 거의 존재하지 않을 수 있다. 또한, min-util은 생성되는 패턴의 길이를 결정하는 데에도 중요한 역할을 한다. 하이 유틸리티 패턴 마이닝을 통해 생성되는 패턴들이 토픽 패턴이 되기 위해서는 적정한 개수의 단어들이 패턴 내에 포함되어 있어야 하며, 트위터 스트림에 슬라이딩 윈도우 기법을 적용하여 배치를 구분할 경우 배치에 포함되는 트랜잭션의 정보가 변화하기 때문에 고정된 min-util을 사용할 수 없다.

[0097] 결과적으로 항목 집합 X 에는 토픽 구성을 위한 적절한 개수의 단어들이 포함되어 있어야 하며, 최소 유틸리티 임계치(min-util)는 이러한 X 에 대해서 $twu(X) \geq min-util$ 을 만족해야 한다. min-util을 결정하기 위한 방법들은 주로 상위-k 하이 유틸리티 패턴 마이닝(Top-k High-utility pattern mining)이라는 주제로 연구가 되고 있으나, 이러한 방법들의 경우 min-util을 결정하기 위한 시간이 오래 걸리기 때문에 데이터 스트림에 대해서 실시간으로 적용하기 어렵다.

[0098] 따라서, 본 발명은 핵심단어들에 대한 TWU를 활용하여 min-util을 추정한다. 항목 집합 X 가 포함된 모든 트랜잭션의 개수를 $s(X)$ 라고 하고, 트랜잭션 T 에 속한 항목들의 내부 유틸리티의 합 $\sum_{i \in T} iu(i, T)$ 를 트랜잭션의 길이 $l(T)$ 라고 나타내자.

[0099] $s(X)$ 와 $l(T)$ 를 사용하면 항목 집합 X 에 대한 TWU는 아래 수학적식 7을 이용하여 산출할 수 있다.

[0100] [수학적식 7]

$$twu(X) = \frac{\sum_{T \in D \wedge X \subseteq T} tu(T)}{\sum_{T \in D \wedge X \subseteq T} l(T)} \times \frac{\sum_{T \in D \wedge X \subseteq T} l(T)}{s(X)} \times s(X)$$

[0101]

[0102] 수학적식 7에서 오른쪽의 세개의 항을 각각 α , β , γ 로 나타내면, $twu(X) = \alpha \times \beta \times \gamma$ 로 나타낼 수 있다.

[0103] α 는 X 를 포함하는 트랜잭션에 속한 항목들의 유틸리티 평균을 의미하는 것으로, 아래 수학적식 8과 같이 정의할 수 있다.

[0104] [수학적식 8]

$$\alpha = \frac{\sum_{T \in D \wedge X \subseteq T} tu(T)}{\sum_{T \in D \wedge X \subseteq T} l(T)}$$

[0105]

[0106] β 는 X 를 포함하는 트랜잭션의 평균 길이를 의미하는 것으로, 아래 수학적식 9와 같이 정의할 수 있다.

[0107] [수학적식 9]

$$\beta = \left(\sum_{T \in D \wedge X \subseteq T} l(T) \right) / s(X)$$

[0108]

[0109] γ 는 X 를 포함하는 트랜잭션 개수를 의미하는 것으로, 아래 수학적식 10과 같이 정의할 수 있다.

[0110] [수학적식 10]

$$\gamma = s(X)$$

[0111]

[0112] 단계 S520이 수행되면, 토픽 탐지 장치는 선택된 단어들에 대한 α , β , γ 의 평균을 각각 산출하고(S530), 산출된 α , β , γ 의 평균을 이용하여 최소 유틸리티 임계치를 산출한다(S540). 이때, 선택된 단어들에 대한 α , β , γ 의 평균을 각각 $avg(\alpha)$, $avg(\beta)$, $avg(\gamma)$ 로 표기한 경우, 최소 유틸리티 임계치는 아래 수학적식 11과 같이 정의될 수 있다.

[0113] [수학적식 11]

$$min-util = avg(\alpha) \times avg(\beta) \times avg(\gamma)$$

[0114]

[0115] 수학적식 11에서 계산된 최소 유틸리티 임계치(min-util)는 선택된 단어(핵심 단어)가 포함된 항목 집합 X 에 대한 TWU를 반영하며, 이를 하이 유틸리티 패턴 마이닝에 사용하면 적절한 개수의 패턴들을 생성할 수 있다. 또한

트위터 스트림에서 배치가 변화할 때마다 계산이 이루어지게 되므로 min-util을 동적으로 결정할 수 있게 된다.

[0116] 도 6은 본 발명의 일 실시예에 따른 후보 토픽 패턴 생성 방법을 설명하기 위한 도면이다.

[0117] 도 6을 참조하면, 토픽 탐지 장치는 가장 최근 배치에서 핵심 단어를 포함하는 단어 집합에 대한 트랜잭션-가중치 유틸리티(TWU)를 산출하고(S610), 그 트랜잭션-가중치 유틸리티에 따라 단어 집합을 오름차순으로 정렬한다(S620).

[0118] 단계 S620의 수행 후, 토픽 탐지 장치는 각 단어집합의 트랜잭션-가중치 유틸리티를 최소 유틸리티 임계치와 비교하여(S630), 트랜잭션-가중치 유틸리티가 최소 유틸리티 임계치 이상인 단어 집합을 후보 토픽 패턴으로 생성한다(S640).

[0119] 즉, 하이 유틸리티 패턴 마이닝의 경우 표 4와 같이 처음 1개 항목으로 구성된 항목 집합의 TWU를 계산하고, TWU에 따라 항목들을 오름차순으로 정렬한다.

[0120] [표 4]

Itemset	{a}	{b}	{c}	{d}	{e}	{f}	{g}
TWU	108	93	100	111	56	34	26

[0121]

[0122] TWU는 하향 폐쇄성(Downward closure property)을 만족하는 중요한 성질이며, 하이 유틸리티 패턴 마이닝에서 패턴의 생성 및 가지치기(Pruning)에 사용된다.

[0123] 따라서, 전체 트랜잭션에서 TWU가 min-util보다 작은 모든 항목들을 트랜잭션에서 제거하고 남은 항목들을 TWU의 오름차순으로 정렬하여 전체 트랜잭션을 재구성한다. 이후, 재구성된 트랜잭션에 대해 유틸리티 리스트라는 자료구조를 사용하여 패턴을 생성하는데, 이때 TWU값이 작은 항목부터 깊이 우선 탐색(Depth-First Search, DFS)을 사용하여 패턴을 생성하게 된다. 이렇게 생성되는 패턴들은 대부분 중복되는 단어들을 포함하고 있게 되며, 따라서 실제 토픽이 되는 패턴을 추출하기 위해서는 후처리 과정을 통해 중복되는 패턴을 제거해야 한다. 중복되는 패턴을 제거하는 후처리 과정에 대해 도 7을 참조하여 설명하기로 한다.

[0124] 도 7은 본 발명의 일 실시예에 따른 토픽 패턴 추출 방법을 설명하기 위한 도면, 도 8은 본 발명의 일 실시예에 따른 트리를 구축하는 방법을 설명하기 위한 예시도이다.

[0125] 도 7을 참조하면, 토픽 탐지 장치는 후보 토픽 패턴들로부터 트리를 구축한다(S710).

[0126] 하이 유틸리티 패턴 마이닝에 의해 생성되는 후보 토픽 패턴들은 단어를 중복해서 포함하는 경향이 있다. 따라서, 토픽 탐지 장치는 후보 토픽 패턴들에서 나타나는 중복성을 효과적으로 제거하기 위해 후보 토픽 패턴들을 압축해서 나타내는 트리를 구성한다. 트리의 구성은 빈발 패턴 마이닝의 한 방법인 FP-Tree를 응용한 TP-Tree(Tpoic-Tree)를 이용한다. 트리의 노드를 가리키는 헤더 테이블을 유지하여 새로운 후보 패턴을 삽입할 때 기존에 삽입된 패턴들과의 중복성을 조사하기 쉽게 하였다.

[0127] 트리를 구축하는 방법에 대해 도 8을 참조하여 설명하기로 한다. 도 8은 후보 토픽 패턴 {e, c, b, d}, {e, c, d}, {c, d, f}, {c, b, d, f}를 순서대로 삽입한다고 가정한 것이다.

[0128] 도 8의 (a)는 첫 번째 패턴 {e, c, b, d}를 트리에 추가하는 과정을 보여준다. 초기 헤더 테이블에 프리픽스(Prefix) 노드 {e}에 대한 링크가 존재하지 않으므로 {e, c, b, d}는 루트노드의 서브트리로 추가된다.

[0129] 도 8의 (b)는 두 번째 패턴 {e, c, d}를 트리에 추가하는 과정을 보여준다. 프리픽스 노드 {e}에 대한 링크가 존재하므로 {e}의 서브트리에 {c, d}가 존재하는지 확인하게 되며, {e}에 대한 첫 번째 링크의 서브트리에 {c, d}가 존재하므로 {e, c, d}는 추가되지 않는다.

[0130] 도 8의 (c)는 세 번째 패턴 {c, d, f}를 추가하는 과정을 보여준다. 프리픽스 노드 {c}에 대한 헤더 테이블의 링크가 존재하므로 {c}의 서브트리에 {d, e}가 존재하는지 확인하게 되며, {c}에 대한 첫 번째 링크의 서브트리에 {d}는 존재하지만 {f}는 존재하지 않는다. {c}에 대한 다음 링크가 공백이므로 {c, d, f}는 루트노드의 자식으로 추가된다.

- [0131] 도 8의 (d), (e), (f)는 네 번째 패턴 {c, b, d, f}를 트리에 추가하는 과정을 보여준다. {c}에 대한 첫 번째 링크의 서브트리에서 {b, d}는 존재하지만 {f}는 존재하지 않는다. {c}에 대한 두 번째 링크의 서브트리에서 {d, f}는 존재하지만 {b}는 존재하지 않는다. {c}에 대한 다음 링크가 공백이므로 {c, b, d, f}는 루트 노드의 서브트리로 추가된다.
- [0132] 상술한 후보 토픽 패턴들로부터 트리를 구축하는 방법에 대해 정리하면, 먼저, 추가하려는 후보 토픽 패턴 $CP = \{i_1, i_2, \dots, i_n\}$ 에 대해, 프리픽스 노드 $\{i_1\}$ 가 헤더테이블에 존재하는지를 판단한다. 그 판단결과, 프리픽스 노드 $\{i_1\}$ 이 헤더테이블에 존재하지 않을 경우 CP를 루트노드의 서브트리로 추가한다.
- [0133] 만약, 프리픽스 노드 $\{i_1\}$ 이 헤더테이블에 존재할 경우 $\{i_1\}$ 의 헤더테이블로부터 연결된 경로를 따라서 공백(NULL)이 아닐 때까지 다음을 수행한다. 헤더테이블로부터 연결된 경로의 서브트리에 $\{i_2, \dots, i_n\}$ 이 존재할 경우 삽입을 종료하고, 헤더테이블로부터 연결된 경로의 모든 서브트리에서 $\{i_2, \dots, i_n\}$ 이 존재하지 않을 경우 경로 루트노드의 서브트리로 $\{i_1, i_2, \dots, i_n\}$ 을 추가한다.
- [0134] 상술한 바와 같이 트리가 구축되면, 토픽 탐지 장치는 구축된 트리의 루트노드에서 리프노드까지 경로를 탐색하여 패턴을 추출한다(S720).
- [0135] 그런 후, 토픽 탐지 장치는 최종 패턴들에 대한 우선순위를 주기 위해, 추출한 각 패턴들의 유틸리티를 각각 산출한다(S730). 이때, 토픽 탐지 장치는 패턴 p의 유틸리티 PU를 아래 수학적 식 12를 이용하여 산출할 수 있다.
- [0136] [수학적 식 12]
- $$PU(p) = \sum_{i \in p} eu(i)$$
- [0137]
- [0138] 즉, 패턴 유틸리티 PU는 패턴 p에 속한 단어들에 대해 단어 가중치의 합을 의미한다.
- [0139] 단계 S730이 수행되면, 토픽 탐지 장치는 패턴 유틸리티 값을 내림차순으로 정렬한 후(S740), 상위 일정 개수의 패턴을 최종 토픽 패턴으로 추출한다(S750). 최종 토픽 패턴들은 하나 이상의 단어들의 집합일 수 있다.
- [0140] 도 9는 본 발명의 일 실시예에 따른 소셜 네트워크 서비스의 데이터 스트림에서 토픽을 탐지하는 토픽 탐지 장치의 블록도이다.
- [0141] 도 9를 참조하면, 본 발명의 일 실시예에 따른 소셜 네트워크 서비스의 트위터 스트림에서 토픽을 탐지하는 토픽 탐지 장치(900)는 통신부(910), 저장부(920), 출력부(930), 제어부(940), 토픽 패턴 추출부(950)를 포함한다.
- [0142] 통신부(910)는 통신망을 통해 다양한 전자 장치들과 데이터를 송수신하기 위한 구성이다. 특히, 통신부(910)는 통신망을 통해 소셜 네트워크 서비스를 제공하는 서버(또는 장치)로부터 데이터 스트림을 수집할 수 있다.
- [0143] 저장부(920)는 토픽 탐지 장치(900)의 동작과 관련된 데이터들을 저장하는 구성이다. 여기서 저장부(920)는 공지된 저장매체를 이용할 수 있으며, 예를 들어, ROM, PROM, EPROM, EEPROM, RAM 등과 같이 공지된 저장매체 중 어느 하나 이상을 이용할 수 있다.
- [0144] 특히, 저장부(920)에는 트위터 스트림에서 토픽 탐지 처리를 위한 프로그램 또는 애플리케이션이 저장될 수 있다. 또한, 저장부(920)에는 토픽 패턴 추출과 관련된 다양한 알고리즘(또는 수학적 식)을 저장할 수 있다. 이러한 경우, 제어부(940)는 저장부(920)를 호출하여 필요한 알고리즘을 획득할 수 있다.
- [0145] 출력부(930)는 토픽 탐지 장치(900)의 동작과 관련된 다양한 정보를 표시하기 위한 구성이다. 특히, 출력부(930)는 토픽 패턴 추출부(950)에서 추출된 최종 토픽 패턴 등의 다양한 정보를 표시할 수 있다. 이러한 출력부

(930)는 LCD, LED 등을 포함하는 다양한 디스플레이 장치를 통해 구현될 수 있다.

- [0146] 제어부(940)는 토픽 탐지 장치(900)의 다양한 구성부들의 동작을 제어하는 구성으로, 적어도 하나의 연산 장치를 포함할 수 있는데, 여기서 상기 연산 장치는 범용적인 중앙연산장치(CPU), 특정 목적에 적합하게 구현된 프로그래머블 디바이스 소자(CPLD, FPGA), 주문형 반도체 연산장치(ASIC) 또는 마이크로 컨트롤러 칩일 수 있다.
- [0147] 토픽 패턴 추출부(950)는 트위터 스트림에서 하이 유틸리티 패턴 마이닝을 이용하여 토픽을 탐지한다. 이러한 토픽 패턴 추출부(950)는 수집모듈(951), 유틸리티 산출모듈(952), 최소 유틸리티 임계치 결정모듈(953), 후보 토픽 패턴 생성모듈(954), 최종 토픽 추출모듈(955)을 포함한다.
- [0148] 수집모듈(951)은 실시간으로 트위터 스트림을 수집한다. 즉, 수집모듈(951)은 트위터의 Streaming API를 사용하여 트위터 데이터를 수집한다. Streaming API는 트위터에서 실시간으로 발생하는 트윗의 1%를 무작위로 제공하는 API이다. Streaming API를 통해 수집되는 트윗들은 트윗을 작성한 사용자와 트윗에 대한 정보를 JSON 파일의 형태로 반환하게 되며 여기서 필요한 정보만을 파싱하여 사용할 수 있다.
- [0149] 유틸리티 산출모듈(952)은 일정 시간 구간의 데이터 스트림에 포함된 각 단어에 대한 유틸리티를 산출한다. 즉, 유틸리티 산출모듈(952)은 데이터 스트림에 슬라이딩 윈도우 기법을 적용하여 적어도 하나 이상의 트랜잭션을 포함하는 복수의 배치(batch)로 구분한다. 그런 후, 유틸리티 산출모듈(952)은 복수의 배치 중 가장 최근 배치에 포함된 각 트랜잭션별로 단어 및 각 단어의 빈도수를 획득하여, 트랜잭션 테이블을 생성한다. 그런 후, 유틸리티 산출모듈(952)은 가장 최근 배치에 포함된 각 단어의 빈도수에 기초하여 각 단어에 대한 가중치를 산출하여 유틸리티 테이블을 생성한다.
- [0150] 최소 유틸리티 임계치 결정모듈(953)은 각 단어에 대한 유틸리티를 이용하여 최소 유틸리티 임계치를 결정한다. 즉, 최소 유틸리티 임계치 결정모듈(953)은 가장 최근 배치에 포함된 단어들 중에서 가중치가 높은 순으로 일정 개수의 단어들을 선택하고, 선택된 단어 각각에 대해, 항목집합을 포함하는 트랜잭션에 속한 항목들의 유틸리티 평균(α), 상기 항목집합을 포함하는 트랜잭션의 평균 길이(β), 상기 항목집합을 포함하는 트랜잭션 개수(γ)를 각각 산출한다. 그런 후, 최소 유틸리티 임계치 결정모듈(953)은 선택된 단어들에 대한 유틸리티 평균(α), 트랜잭션의 평균 길이(β), 트랜잭션 개수(γ)의 평균을 각각 산출하여, 최소 유틸리티 임계치를 산출한다.
- [0151] 후보 토픽 패턴 생성모듈(954)은 최소 유틸리티 임계치에 기초한 하이 유틸리티 패턴 마이닝으로 후보 토픽 패턴들을 생성한다. 즉, 후보 토픽 패턴 생성모듈(954)은 가장 최근 배치에서 선택된 단어를 포함하는 단어 집합에 대한 트랜잭션-가중치 유틸리티(TWU)를 산출하고, 트랜잭션-가중치 유틸리티가 최소 유틸리티 임계치 이상인 단어 집합을 후보 토픽 패턴으로 생성한다.
- [0152] 최종 토픽 패턴 추출모듈(955)은 후보 토픽 패턴에 대해 TP-Tree를 구축하여 실제 토픽이 되는 최종 토픽 패턴들을 추출한다. 즉, 최종 토픽 패턴 추출모듈(955)은 후보 토픽 패턴들로부터 트리를 구축하고, 구축된 트리의 루트노드에서 리프노드까지 경로를 탐색하여 패턴을 추출한다. 그런 후, 최종 토픽 패턴 추출모듈(955)은 추출된 각 패턴들의 유틸리티를 각각 산출하고, 패턴 유틸리티 값을 내림차순으로 정렬하여 상위 일정 개수를 최종 토픽 패턴으로 추출한다.
- [0153] 본 발명의 일 실시예에 따르면, 토픽 패턴 추출부(950)는 외부 단말 장치나 외부 서버 등과 통신하는 프로그램 모듈들일 수 있다. 이러한 프로그램 모듈들은 운영 시스템, 응용 프로그램 모듈 및 기타 프로그램 모듈로서 토픽 탐지 장치(900)에 포함될 수 있으며, 물리적으로는 여러 가지 공지의 기억 장치상에 저장될 수 있다. 또한, 이러한 프로그램 모듈들은 토픽 탐지 장치(900)와 통신 가능한 원격 기억 장치에 저장될 수도 있다. 이러한 프로그램 모듈들은 본 발명에 따라 전술한 특정 동작을 수행하거나 특정 추상 데이터 유형을 실행하는 루틴, 서브루틴, 프로그램, 오브젝트, 컴포넌트, 데이터 구조 등을 포괄하지만, 이에 제한되지는 않는다.
- [0154] 한편, 토픽 탐지 장치(900)는 컴퓨터, 노트북, 넷북, PDA 등의 통신단말기일 수 있으며, 스마트폰, 스마트 노트, 태블릿 PC, 스마트 TV 등의 스마트 기기일 수도 있다. 또한, 토픽 탐지 장치(900)는 단일의 연산 장치로 구현되거나 둘 이상의 연산 장치가 서로 연결된 집합 장치의 형태로 구현될 수도 있다. 예를 들어, 토픽 탐지 장치(900)는 단일의 서버로 구현되거나 둘 이상의 서버가 연결된 형태로 구현될 수 있다.
- [0155] 이하, 트위터에서 토픽 탐지 성능 평가를 위한 실험 과정과 결과에 대해 설명하기로 한다.
- [0156] 먼저, 사용한 트위터 데이터는 스포츠 및 정치 분야에 대한 3가지 트위터 데이터셋을 구성하였으며, 각 데이터셋에 포함된 트윗들의 아이디를 웹페이지를 통해 제공하고 있다. 제공된 트윗들의 아이디와 트위터 Streaming

API를 사용하여 데이터셋을 구성하여 실험을 진행하였다. 전체 데이터셋은 2012년에 화제가 되었던 FA컵 결승(FA Cup Final), 슈퍼 화요일(Super Tuesday), 미국 대통령 선거(US Elections)에 대한 트윗들을 포함하고 있다. 탐지된 토픽의 평가를 위해 제공되는 Ground-truth 데이터셋은 실제 발생했던 토픽의 정보들을 포함하고 있으며, 하나의 Ground-truth 데이터에 포함된 각 토픽은 [필수적 단어, 선택적 단어, 금지된 단어]의 형태로 토픽 구성을 위해 필요한 단어들로 구성되어 있다.

[0157] 실험에서 사용되는 모든 트윗들은 전처리 과정을 거치게 되며 전처리 과정에서는 1)모든 문자를 소문자로 바꾸고, 2)불용어(Stopwords)를 제거하며, 3)모든 특수문자를 제거하고, 4)트윗에 포함된 URL을 제거한다. FA 데이터셋에 대해서는 배치 구성을 위한 윈도우 크기를 1분으로 설정하며, ST 데이터셋에 대해서는 60분, US 데이터셋에 대해서는 10분으로 설정하여 실험을 진행한다. 제안하는 방법의 경우 min-util을 결정하기 위해 r 값과 ω 의 값, 상한 값을 설정해주어야 한다. 실험에서는 $r=3$, $\omega=50$, $\Theta=300$ 으로 설정하여 성능을 측정하였다.

[0158] 탐지된 토픽의 성능 평가를 위한 평가 스크립트를 사용하여 성능을 측정하였으며, 성능평가 척도에는 토픽 재현율, 토픽 연관성, 핵심단어 정확도, 핵심단어 재현율이 사용된다. 추가적으로 핵심단어 정확도와 재현율에 대한 조화평균을 사용하여 성능을 평가하였다. 토픽 재현율(Topic Recall)은 Ground-truth 토픽들 중에서 성공적으로 탐지된 토픽들의 비율을 의미한다. 토픽 연관성(Topic Relevance)은 탐지된 토픽들 중에서 Ground-truth 토픽과 매칭되는 토픽들의 비율을 의미한다. 핵심단어 정확도(Keyword Precision)는 Ground-truth 토픽과 일치하는 탐지된 토픽에서 나타난 단어들이 Ground-truth 토픽의 단어와 일치하는 비율을 의미한다. 핵심단어 재현율(Keyword Recall)은 Ground-truth 토픽의 단어가 탐지된 전체 토픽의 단어들에서 나타난 비율을 의미한다. 핵심단어 조화평균(F-measure)은 핵심단어 정확도와 핵심단어 재현율에 대한 조화평균을 의미한다.

[0159] 성능 평가를 위한 비교 알고리즘에는 단어-중심 기반의 BNgram, SFPM, Graph-Based와 문서-중심 기반의 Doc-Pivot, 확률 토픽 모델 기반의 LDA에 대해 토픽 탐지 성능을 비교하며, 비교 알고리즘에 대한 매개변수는 평가 스크립트에 설정되어있는 기본 값을 사용하였다. 토픽 탐지를 위해 추출되는 토픽의 개수는 FA 데이터셋의 경우 구간마다 2개, ST 데이터셋과 US 데이터셋의 경우 구간마다 10개를 추출하여 평가를 진행한다. 성능 비교에서는 토픽에 대한 탐지 성능을 평가하는 것이 목적이기 때문에 토픽 재현율을 주요 기준으로 하여 설명하며, 필요에 따라 조화평균을 같이 비교한다.

[0160] 표 5는 각 데이터셋에 대해 제안하는 방법과 비교 알고리즘의 토픽 재현율에 대한 성능 평가를 나타낸다.

[0161] [표 5]

FA Cup					
Method	Topic related		Keyword related		
	Recall	Relevance	Precision	Recall	F-measure
Proposed	0.923	0.692	0.320	0.600	0.418
BNgram	0.846	0.423	0.341	0.569	0.426
SFPM	0.538	0.538	0.370	0.606	0.460
G-Based	0.538	0.269	0.211	0.516	0.299
D-Pivot	0.538	0.269	0.283	0.515	0.366
LDA	0.385	0.192	0.260	0.565	0.356
Super tuesday					
Method	Topic related		Keyword related		
	Recall	Relevance	Precision	Recall	F-measure
Proposed	0.364	0.313	0.486	0.708	0.576
BNgram	0.318	0.100	0.650	0.605	0.627
SFPM	0.182	0.288	0.600	0.577	0.588
G-Based	0.000	0.000	0.000	0.000	0.000
D-Pivot	0.136	0.038	0.300	0.450	0.360
LDA	0.273	0.075	0.350	0.600	0.442
US elections					
Method	Topic related		Keyword related		
	Recall	Relevance	Precision	Recall	F-measure
Proposed	0.453	0.242	0.352	0.565	0.434
BNgram	0.328	0.085	0.413	0.529	0.464
SFPM	0.313	0.385	0.380	0.522	0.440
G-Based	0.031	0.008	0.182	0.667	0.286
D-Pivot	0.047	0.012	0.259	0.467	0.333
LDA	0.391	0.104	0.324	0.579	0.415

[0162]

[0163] 표 5에서 FA 데이터셋에 대한 실험 결과를 보면 제안하는 방법의 토픽 재현율이 0.923으로 가장 높은 토픽 재현율을 보여주고 있다. 비교 알고리즘의 경우 BNgram의 토픽 재현율이 0.846으로 가장 높은 수치를 보여주고 있으나, 제안하는 방법보다 약 8% 낮은 수치를 보여주고 있다. 핵심단어에 대한 정확도와 재현율은 SFPM이 가장 높았지만, 핵심단어에 대한 조화평균은 SFPM이 0.460, 제안하는 방법은 0.418로 약 4%의 차이만 나타내고 있으며, 특히 SFPM은 제안하는 방법보다 토픽 재현율이 약 40%낮은 수치를 보여주고 있다. ST 데이터셋에 대한 실험 결과를 보면 제안하는 방법의 토픽 재현율이 0.364로 가장 높은 토픽 재현율을 보이고 있다. 비교 알고리즘의 경우 BNgram이 0.318로 가장 높은 수치를 보여주고 있으나, 제안하는 방법보다 약 5% 낮은 수치를 보여주고 있다. 핵심단어에 대한 정확도는 BNgram이 0.486으로 가장 높았지만, 핵심단어에 대한 재현율은 제안하는 방법이 0.708로 가장 높은 수치를 보여주었다. US 데이터셋에 대한 실험 결과를 보면 제안하는 방법의 토픽 재현율이 0.453으로 가장 높은 수치를 보여주고 있다. 비교 알고리즘의 경우 LDA, BNgram순으로 높은 수치를 보여주었다. LDA의 토픽 재현율은 0.391로 제안하는 방법과 비교하여 약 6% 낮은 수치를 보여주고 있으며, BNgram의 토픽 재현율은 0.328로 제안하는 방법과 비교하여 약 13% 낮은 수치를 보여주고 있다. 핵심단어 정확도와 재현율은 각각 BNgram, G-Based가 0.413, 0.667로 가장 높게 나왔지만, 핵심단어 조화평균은 제안하는 방법과 큰 차이를 보이지 않고 있다. 전체적으로 제안하는 방법의 토픽 재현율이 모든 경우에서 가장 높은 수치를 나타내었고, 핵심단어에 대한 성능 평가 또한 비교 알고리즘과 큰 차이를 보이지 않고 있다. 이를 통해 제안하는 방법의 토픽 탐지 성능이 우수함을 확인할 수 있다.

[0164] 또한, 본 발명에서 제안하는 방법에서는 min-util을 결정하기 위한 매개변수로 r 값, 하한 값 ω , 상한 값 Θ 를 설정하여 사용한다. 실제 토픽 탐지에서 알고리즘이 매개변수에 민감하게 반응할 경우 매개변수 값 설정이 어려

우므로, 매개변수에 따른 알고리즘의 민감성(Sensitivity)은 중요한 요소이다.

[0165] 도 10은 매개변수 선택에 따른 성능 비교 결과를 나타낸다. 도 10은 p 를 {0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04}로 변화시키고, ω 를 {10, 30, 50, 70}으로 변화시키며, θ 를 {30 0, 50 0, 70 0, 90 0}으로 변화시켰을 때의 토픽 재현율을 나타낸다. FA 데이터 셋의 경우, p 의 값이 0.025 이상일 때, 토픽 재현율이 일정하고, p 값이 0.025보다 작을 때, 토픽 재현율이 높은 ω 값에 대해 증가하는 경향이 있음을 확인할 수 있다. ST 데이터 셋의 경우, θ 값이 200에서 400 범위에 있을 때 토픽 재현율이 높다는 것을 알 수 있다. US 데이터 셋의 경우 p 값이 0.015 이상이고 θ 값이 300 인 경우 성능이 가장 좋다는 것을 확인할 수 있다. 또한, p 값이 증가하여도 θ 값에 민감하지 않다는 것을 알 수 있다. 결론적으로, 3 개의 데이터 셋에 대해, p 값이 0.025 이상, θ 값이 200 내지 400의 범위, ω 값이 70 내지 90의 범위인 경우, 토픽 재현율이 좋다는 것을 확인할 수 있다.

[0166] 한편, 트위터 스트림에서의 토픽 탐지는 실시간 처리성이 중요하다. 따라서 토픽 탐지에 소요되는 시간을 평가하기로 한다.

[0167] 도 11은 토픽 탐지에 소요되는 평균 시간을 측정한 그래프이다. 도 11을 참조하면, FA, ST, US 데이터셋에 대한 Ground-truth 토픽 탐지에 소요된 평균 시간을 나타낸다. FA 데이터셋의 Ground-truth 토픽을 탐지하는데 소요된 평균 시간을 보면, 제안하는 방법의 경우 토픽 탐지를 위해 평균 0.31초가 소요되었으며, 이를 통해 가장 빠르게 토픽을 탐지한 것을 알 수 있다. 특히 비교 알고리즘에서 가장 오랜 시간이 걸린 BNgram과 비교하여 약 50 배 빠른 결과를 나타내고 있다. ST 데이터셋에서는 비교 알고리즘 중에서 SFPM이 평균 7.88초로 가장 빠르게 토픽을 탐지하였으나, 제안하는 방법과 비교하여 각각 약 1초의 미세한 차이를 나타내고 있다. US 데이터셋에서는 제안하는 방법의 경우 토픽 탐지를 위해 소요된 시간은 평균 3.83초로 가장 빠르게 토픽을 탐지하였으며, 비교 알고리즘 중에서 시간이 가장 많이 소요된 BNgram과 비교하여 약 9배의 차이를 나타내고 있다. 이러한 결과를 통해 제안하는 방법이 실시간으로 토픽을 탐지하기에 적합함을 알 수 있다.

[0168] 한편, 본 발명의 실시예들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광매체(magneto-optical), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 본 발명의 일 실시예들의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0169] 이제까지 본 발명에 대하여 그 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

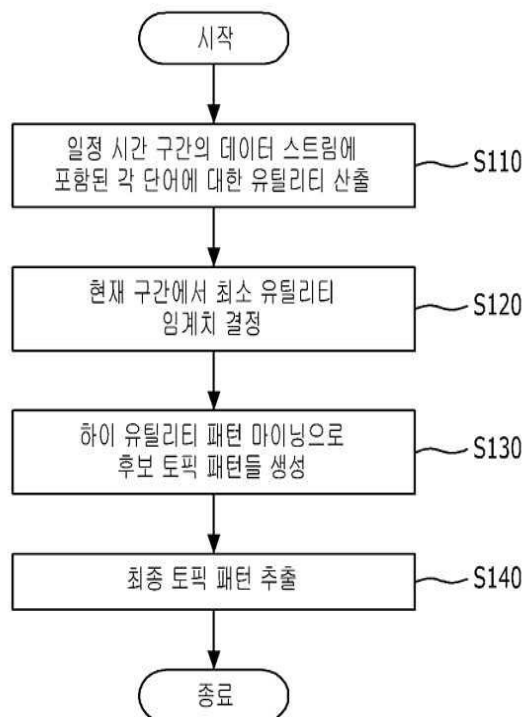
부호의 설명

[0170] 900 : 토픽 탐지 장치
910 : 통신부
920 : 저장부
930 : 출력부
940 : 제어부

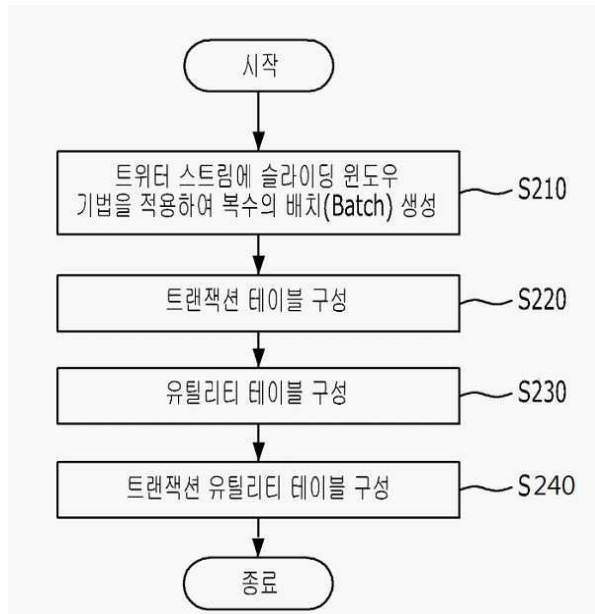
- 950 : 토픽 패턴 추출부
- 951 : 수집모듈
- 952 : 유틸리티 산출모듈
- 953 : 최소 유틸리티 임계치 결정모듈
- 954 : 후보 토픽 패턴 생성모듈
- 955 : 최종 토픽 추출모듈

도면

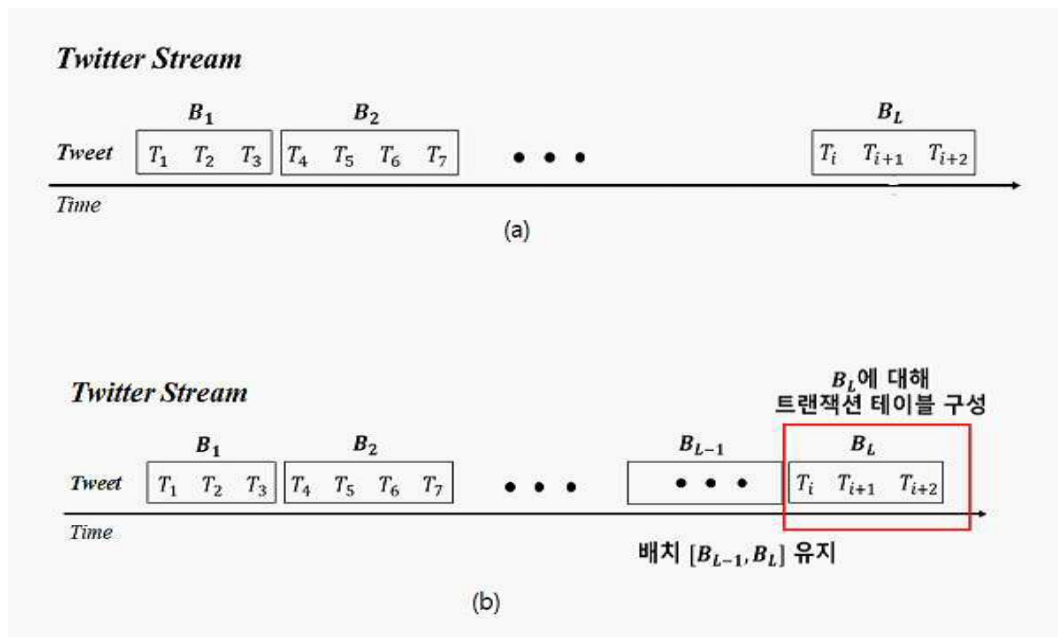
도면1



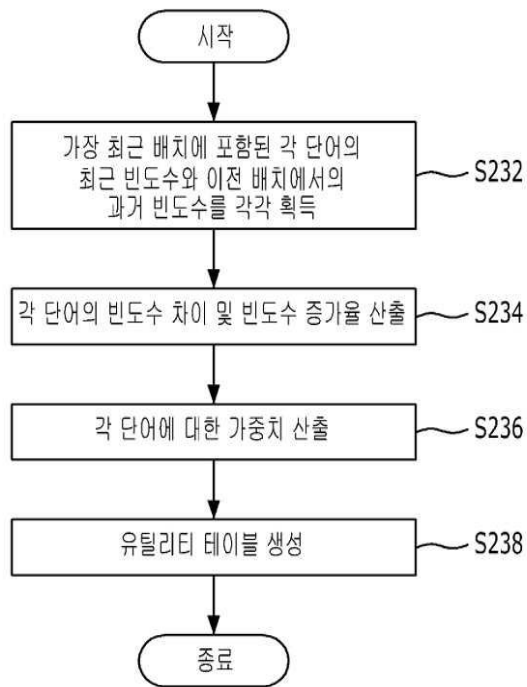
도면2



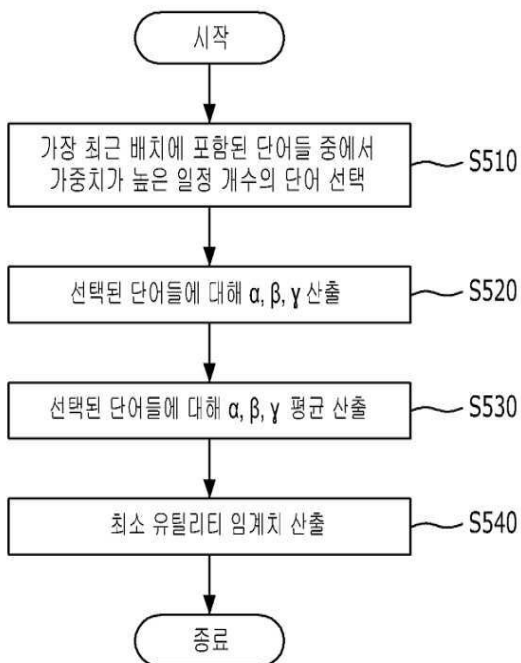
도면3



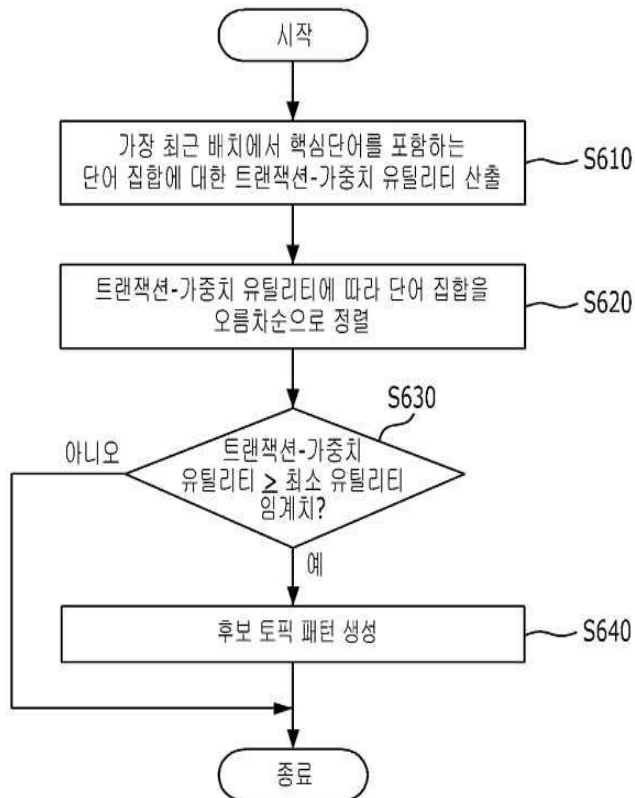
도면4



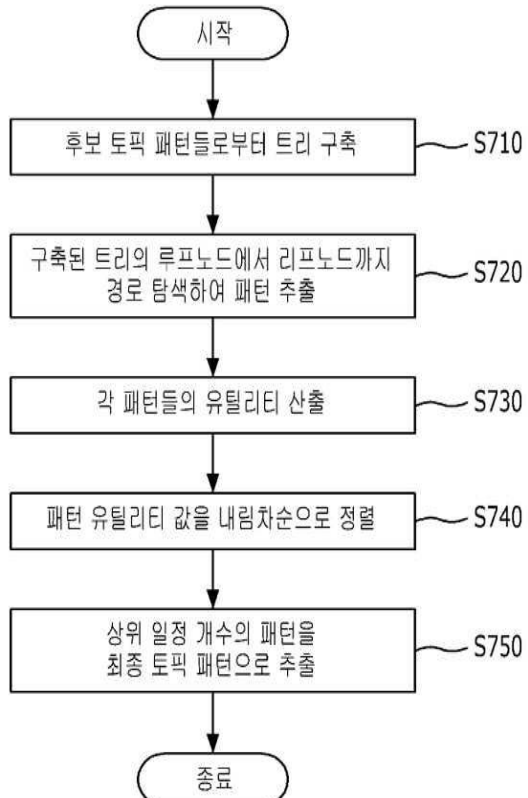
도면5



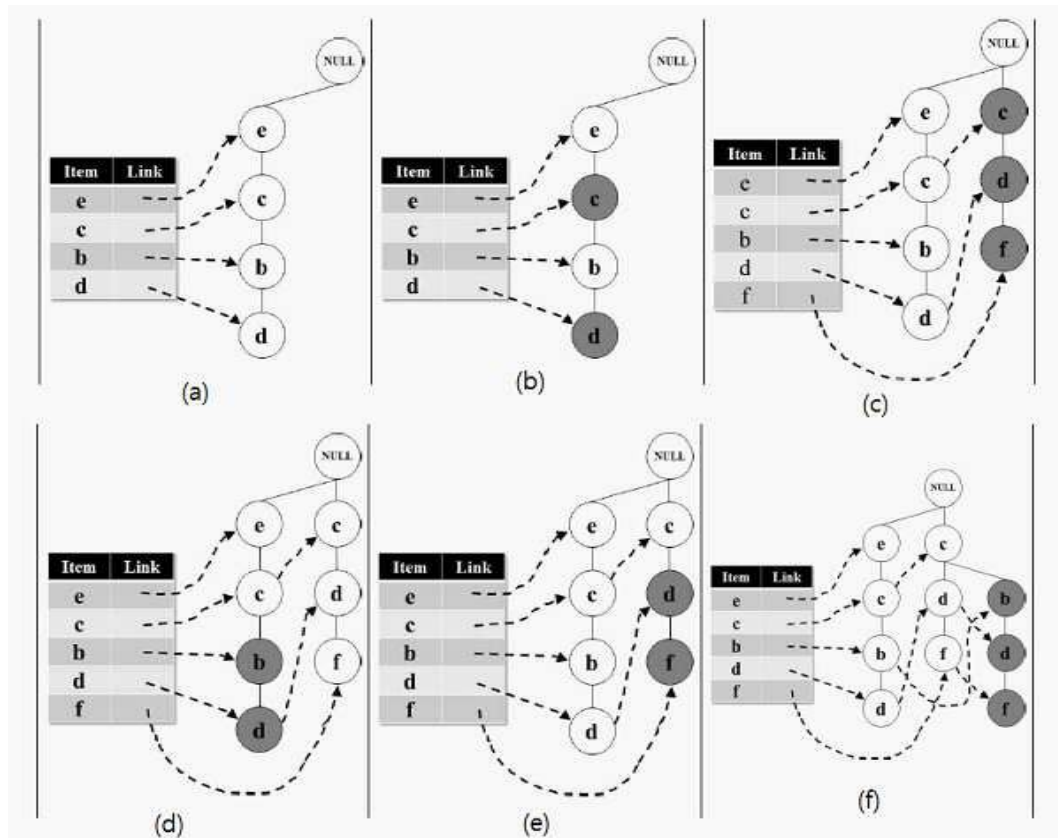
도면6



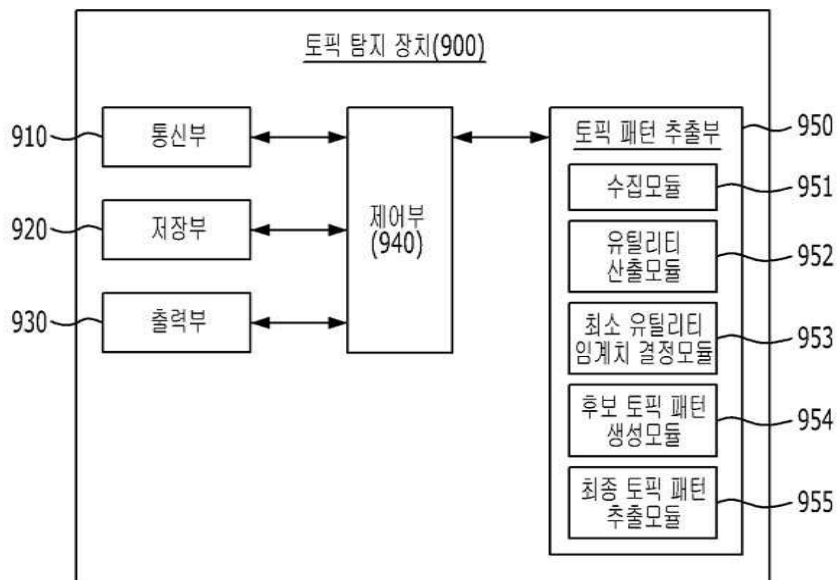
도면7



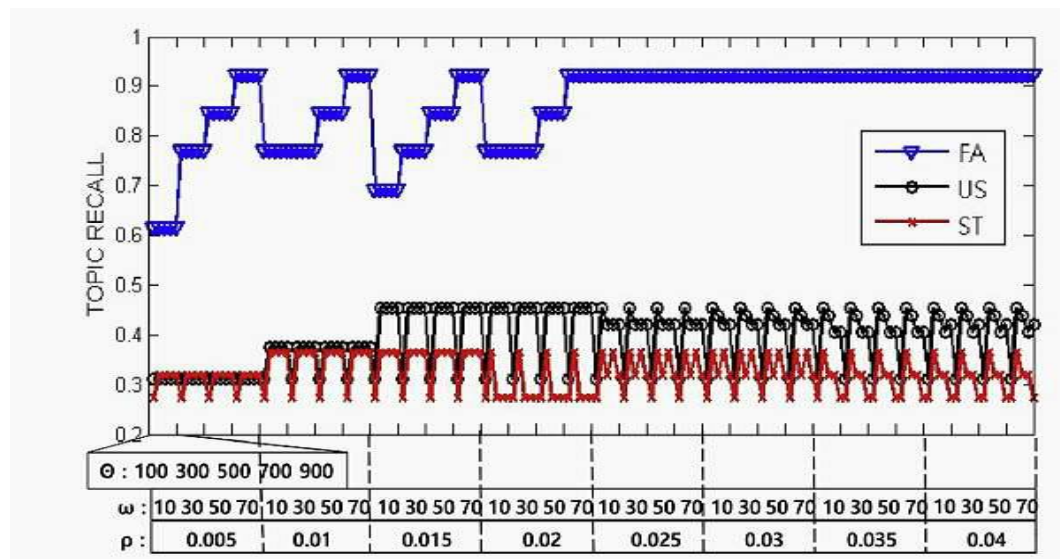
도면8



도면9



도면10



도면11

