



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0022230
(43) 공개일자 2017년03월02일

(51) 국제특허분류(Int. Cl.)
G06Q 50/10 (2012.01) G06F 17/30 (2006.01)
(52) CPC특허분류
G06Q 50/10 (2013.01)
G06F 17/3061 (2013.01)
(21) 출원번호 10-2015-0116966
(22) 출원일자 2015년08월19일
심사청구일자 2015년08월19일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
윤지원
서울특별시 용산구 한강대로43길 8, 102동 1401호 (한강로2가, 벽산메가트리움)
정지수
대구광역시 수성구 달구벌대로522길 33, 104동 803호 (만촌동, 수성2차 e 편한세상아파트)
(74) 대리인
특허법인엠에이피에스

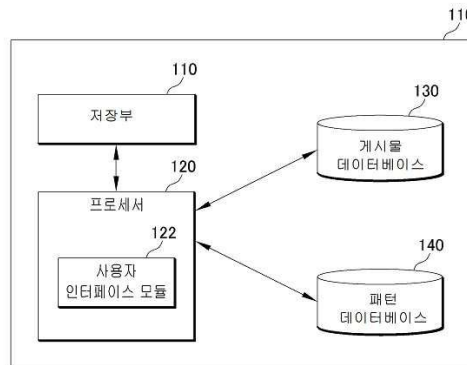
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템 및 이를 이용한 범죄탐지 방법

(57) 요약

본 발명은 온라인상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템을 개시한다. 본 발명의 일 실시예에 따르면, 비정형데이터를 구성하는 게시글 및 해당 게시글에 대한 댓글이 각각 구분되어 저장된 게시물 데이터베이스; 상기 게시글에 대한 패턴정보를 저장하는 패턴정보 데이터베이스; 비정형데이터로부터 범죄탐지를 수행하는 프로그램을 저장하는 저장부 및 프로그램을 실행하는 프로세서를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06Q 50/00 (2013.01)

(72) 발명자

김성훈

서울특별시 성동구 마장로42길 16, 9동 701호 (마장동, 세림아파트)

한병우

서울특별시 송파구 올림픽로 203, 528동 802호 (잠실동, 주공5단지아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711026813

부처명 미래창조과학부

연구관리전문기관 정보통신기술진흥센터

연구사업명 정보통신기술인력양성

연구과제명 제어시스템 보안 연구

기 여 율 1/1

주관기관 고려대학교 산학협력단

연구기간 2015.01.01 ~ 2015.12.31

명세서

청구범위

청구항 1

온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템에 있어서,
비정형데이터를 구성하는 게시물 및 댓글이 저장되되, 게시물과 해당 게시물에 대한 댓글을 포함하는 게시물 그룹이 각각 구분되어 저장된 게시물 데이터베이스,
상기 게시물 그룹에 대한 패턴정보를 저장하는 패턴정보 데이터베이스
상기 비정형데이터로부터 범죄탐지를 수행하는 프로그램을 저장하는 저장부; 및
상기 프로그램을 실행하는 프로세서를 포함하되,
상기 프로세서는 상기 프로그램의 수행에 따라 범죄탐지를 위한 대상 비정형데이터를 가공처리하고, 가공처리된 데이터 중 대상 텍스트를 추출하고, 추출된 대상 텍스트와 상기 패턴정보의 비교를 통해, 상기 대상 텍스트의 이상점을 판단하고, 상기 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하고, 상기 추출된 키워드를 기초로 상기 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시물 또는 이상점 댓글을 추출하고, 상기 대상 텍스트와 상기 이상점 게시물 또는 이상점 댓글을 군집화하는 것인 범죄탐지를 수행하는 시스템.

청구항 2

제 1항에 있어서,
상기 프로세서는 상기 게시물 그룹에 포함된 문장을 구성하는 형태소 쌍을 추출하고, 추출된 형태소 쌍의 출현 빈도 및 인접 정도에 기초하여 각 형태소 쌍에 대한 누적 연관성 점수를 산출하여, 상기 게시물 그룹에 대한 패턴정보 정보를 생성하는 것인 범죄탐지를 수행하는 시스템.

청구항 3

제 2항에 있어서,
상기 프로세서는 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수 보다 높게 설정하여 상기 누적 연관성 점수를 산출하는 것인 범죄탐지를 수행하는 시스템.

청구항 4

제 2항에 있어서,
상기 프로세서는 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 제 1 점수로 설정하고, 하나의 문장에 각 형태소 쌍이 이격되어 배치된 경우의 연관성 점수를 제 2 점수로 설정하고, 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수를 제 3 점수로 설정하되,
제 1 점수가 제 2 점수보다 높고, 제 2 점수가 제 3 점수보다 높도록 설정하여 상기 누적 연관성 점수를 산출하는 것인 범죄탐지를 수행하는 시스템.

청구항 5

제 1항에 있어서,
상기 프로세서는 상기 대상 텍스트의 산출된 연관성 점수가 임계값 미만일 경우, 이상점으로 판단하며, 해당 대상 텍스트로부터 키워드를 추출하되,
상기 추출된 키워드는 상기 이상점 텍스트에 대한 형태소 정보이거나 또는 상기 이상점 텍스트에 대한 명사 정보인 것인 범죄탐지를 수행하는 시스템.

청구항 6

제 1항에 있어서,

상기 프로세서는 상기 상기 대상 텍스트의 산출된 연관성 점수가 임계값 이상일 경우, 이상점이 아니라고 판단하며, 해당 대상 텍스트의 형태소 쌍의 연관성 점수를 상기 패턴정보 데이터베이스에 갱신하는 것인 범죄탐지를 수행하는 시스템.

청구항 7

제 1항에 있어서,

상기 프로세서는 상기 추출된 키워드를 기초로

상기 대상 텍스트가 포함된 게시글 그룹으로부터 상기 이상점 댓글을 추출하는 것인, 범죄탐지를 수행하는 시스템.

청구항 8

제 1항에 있어서,

상기 프로세서는 상기 추출된 키워드를 기초로

상기 대상 텍스트가 포함되지 않은 다른 게시글 그룹으로부터 상기 이상점 게시글 또는 이상점 댓글을 추출하는 것인 범죄탐지를 수행하는 시스템.

청구항 9

제1항에 있어서

상기 프로세서는 상기 게시글 그룹을 각각 가공처리하여 게시물 데이터베이스에 저장하되, 상기 가공처리는 상기 대상 비정형데이터에 대하여 특수문자를 제거한 텍스트에 대하여 형태소 분석한 정보를 저장하는 것인 범죄탐지를 수행하는 시스템.

청구항 10

제 1항에 있어서,

상기 프로세서는 상기 군집화된 이상점 게시글 또는 이상점 댓글을 그 작성 시점에 따라 시계열적으로 시각화하여 표시하되, 작성 날짜를 기준으로 하는 X 축과 작성시간을 기준으로 하는 Y축 상에 2차원 그래프 상으로 표시하는 것인 범죄탐지를 수행하는 시스템.

청구항 11

범죄탐지를 수행하는 시스템의 범죄탐지 방법에 있어서,

비정형데이터를 구성하는 게시글 및 댓글이 저장되되, 게시글과 해당 게시글에 대한 댓글을 포함하는 게시글 그룹이 각각 구분되어 저장된 게시물 데이터베이스에 대하여 범죄탐지를 위한 대상 비정형데이터를 가공처리하는 단계;

상기 가공처리된 데이터 중 대상 텍스트를 추출하는 단계;

상기 추출된 대상 텍스트와 상기 게시글 그룹에 대한 패턴정보의 비교를 통해 이상점 여부를 판단하는 단계;

상기 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하는 단계;

상기 추출된 키워드를 기초로 상기 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시글 또는 이상점 댓글을 추출하는 단계; 및

상기 대상 텍스트와 상기 이상점 게시글 또는 이상점 댓글을 군집화하는 단계를 포함하는 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 12

제 11항에 있어서,

상기 게시글 그룹에 대한 패턴정보 정보를 패턴정보 데이터베이스에 저장하는 단계를 더 포함하되,

상기 저장하는 단계는

상기 게시글 그룹에 포함된 문장을 구성하는 형태소 쌍을 추출하고;

추출된 형태소 쌍의 출현 빈도 및 인접 정도에 기초하여 각 형태소 쌍에 대한 누적 연관성 점수를 저장하는 것인 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 13

제 12항에 있어서,

상기 저장하는 단계상기 누적 연관성 점수의 산출하는 단계를 더 포함하고,

상기 산출하는 단계는

하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수 보다 높게 설정하여 상기 누적 연관성 점수를 산출하는 것인 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 14

제 13항에 있어서,

상기 산출하는 단계는,

하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 제 1 점수로 설정하고, 하나의 문장에 각 형태소 쌍이 이격되어 배치된 경우의 연관성 점수를 제 2 점수로 설정하고, 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수를 제 3 점수로 설정하되,

상기 제 1 점수가 제 2 점수보다 높고, 상기 제 2 점수가 제 3 점수보다 높도록 설정하여 상기 누적 연관성 점수를 산출하는 것인 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 15

제 11항에 있어서,

상기 대상 비정형데이터를 가공처리하는 단계는

상기 대상 비정형데이터에 대하여 특수문자를 제거하는 단계; 및

상기 제거된 텍스트에 대하여 형태소를 분류하는 단계를 더 포함하는 것인 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 16

제 11항에 있어서,

키워드를 추출하는 단계는

상기 대상 텍스트가 이상점에 해당할 경우 키워드 추출 알고리즘을 통해 상기 키워드를 추출하되,

상기 추출된 키워드는 상기 이상점 텍스트에 대한 형태소 정보이거나 또는 상기 이상점 텍스트에 대한 명사 정보인 것인 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

청구항 17

제 11항에 있어서,

상기 군집화하는 단계는

상기 군집화된 이상점 게시글 또는 이상점 댓글을 그 작성 시점에 따라 시계열적으로 시각화하는 단계를 더 포

함하고,

상기 시각화는 작성 날짜를 기준으로 하는 X 축과 작성시간을 준으로 하는 Y축 상에 2차원 그래프 상에서 이루어지는 것인, 범죄탐지를 수행하는 시스템의 범죄탐지 방법.

발명의 설명

기술 분야

[0001] 본 발명은 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템 및 이를 이용한 범죄 탐지 방법에 관한 것이다.

배경 기술

[0002] 최근 웹 파싱(Web parsing)이나 웹 크롤링(Web crawling) 등 데이터 수집 기법을 통해 수집된 비정형데이터에 대하여 유사도 측정이 가능하도록 가공하고, 유사도가 높은 문서와 비교함으로써, 가공된 비정형데이터의 이상점 여부를 판단하는 비정형데이터의 이상점 탐지 기법에 대한 연구가 이루어지고 있다.

[0003] 하지만, 유사도가 높은 문서와의 비교를 통한 이상점 여부 판단은 방대한 양의 비정형데이터를 처리하는데 있어 처리속도의 지연 및 정확도가 떨어질 수 있다는 문제점이 있다.

[0004] 따라서, 웹에 등록된 비정형데이터에 대하여 좀더 빠르고 정확한 이상점 탐지가 이루어지도록하는 새로운 이상점 탐지 방법이 필요한 실정이다.

[0005] 한편, 대한민국공개특허 제2010-0045539호(발명의 명칭: 패턴정보분석을 이용한 댓글 관리 방법, 그 시스템, 및 웹 서버)는 댓글을 자동적으로 분석하여 스팸 또는 광고성 댓글여부를 판단하는 패턴정보데이터 생성을 통하여 댓글을 관리할 수 있는 방법 및 시스템에 대하여 개시하고 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명의 일 실시예는 온라인상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템 및 이를 이용한 범죄탐지 방법을 제공하는데 그 목적이 있다.

[0007] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 더 존재할 수 있다.

과제의 해결 수단

[0008] 기술적 과제를 달성하기 위한 본 발명의 일 실시예에 따른 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템은, 상기 비정형데이터를 구성하는 게시글 및 해당 게시글에 대한 댓글이 각각 구분되어 저장된 게시물 데이터베이스; 상기 게시물에 대한 패턴정보가 저장된 패턴정보 데이터베이스; 상기 비정형데이터로부터 범죄탐지를 수행하는 프로그램을 저장하는 저장부 및 프로그램을 실행하는 프로세서를 포함한다.

[0009] 이때, 프로세서는 상기 프로그램의 수행에 따라 범죄탐지를 위한 대상 비정형데이터를 가공처리하고, 가공처리된 데이터 중 대상 텍스트를 추출하고, 상기 추출된 대상 텍스트와 상기 패턴정보를 비교하여 이상점을 판단하고, 상기 산출된 연관성 점수가 임계값 미만일 경우 상기 대상 텍스트를 이상점이라 판단하고, 상기 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하고, 상기 추출된 키워드를 기초로 상기 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시글 또는 이상점 댓글을 추출하고, 상기 대상 텍스트와 상기 이상점 게시글 또는 이상점 댓글을 군집화한다.

[0010] 또한, 본 발명의 일 실시예에 따른 범죄탐지를 수행하는 시스템의 범죄탐지 방법은 비정형데이터를 구성하는 게시글 및 해당 게시글에 대한 댓글이 각각 구분되어 저장된 게시물 중 범죄탐지를 위한 대상 비정형데이터를 가공처리하는 단계; 상기 가공처리된 데이터 중 대상 텍스트를 추출하는 단계; 상기 추출된 대상 텍스트와 상기 패턴정보를 비교하여 이상점 여부를 판단하는 단계; 상기 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하는 단계; 상기 추출된 키워드를 기초로 상기 게시물 데이터베이스로부터 이상점에 해

당하는 이상점 게시글 또는 이상점 댓글을 추출하는 단계; 및 상기 대상 텍스트와 상기 이상점 게시글 또는 이상점 댓글을 군집화하는 단계를 포함한다.

발명의 효과

- [0011] 전술한 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 게시글에 대한 패턴정보를 이상점 탐지를 위한 비교 데이터로 사용함으로써, 방대한 양의 비정형데이터의 이상점 탐지 수행시 처리속도를 향상시킬 수 있다.
- [0012] 또한, 사용자 인터페이스를 통해 탐지된 이상점 데이터에 대하여 군집화 또는 시계열적으로 분석한 정보를 시각화하여 제공함으로써, 사용자는 이상점에 해당하는 글의 유포 경로를 탐지할 수 있다.
- [0013] 또한, 이상점이 탐지된 이상점 데이터에 대하여 키워드가 추출되고 해당 키워드를 이상점 탐지에 사용함으로써, 상기 이상점과 연관된 관련 범죄를 추적할 수 있으며, 더 나아가 연관범죄 자동 추적 시스템을 구축할 수 있다.

도면의 간단한 설명

- [0014] 도 1은 본 발명의 일 실시예에 따른 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 범죄탐지 시스템에 대한 간략한 구성도이다.
- 도 2는 일예에 따른 패턴정보 분석방법을 설명하기 위한 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 범죄탐지 시스템의 범죄탐지 방법을 설명하기 위한 순서도이다.
- 도 4a 내지 도 4b는 일예에 따른 사용자 인터페이스를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0015] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0016] 또한, 도면을 참고하여 설명하면서, 같은 명칭으로 나타낸 구성일지라도 도면에 따라 도면 번호가 달라질 수 있고, 도면 번호는 설명의 편의를 위해 기재된 것에 불과하고 해당 도면 번호에 의해 각 구성의 개념, 특징, 기능 또는 효과가 제한 해석되는 것은 아니다.
- [0017] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0018] 이하에서는, 도면을 참조하여 본 발명의 일 실시예인 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 범죄탐지 시스템 및 범죄탐지 방법에 대해 구체적으로 설명하기로 한다.
- [0019] 도 1은 본 발명의 일 실시예에 따른 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 범죄탐지 시스템에 대한 간략한 구성도이다.
- [0020] 이하에서는 도 1을 통해 본 발명의 일 실시예에 따른 범죄탐지를 수행하는 시스템(이하에서는, '범죄탐지 시스템'이라 지칭함)에 대하여 살펴보도록 한다. 도 1을 참고하면, 범죄탐지 시스템은 저장부(110), 프로세서(120), 사용자 인터페이스 모듈(122), 게시물 데이터베이스(130), 패턴정보 데이터베이스(140)를 포함한다. 다만, 이러한 도 1의 범죄탐지 시스템은 본 발명의 일 실시예에 불과하므로 도 1을 통해 본 발명이 한정 해석되는 것은 아니다. 즉, 본 발명의 다양한 실시예들에 따르면, 범죄탐지 시스템은 도 1과 다르게 구성될 수도 있다.
- [0021] 본 발명의 일 실시예에 따른 범죄탐지 시스템은 종류, 성능, 형태 등에 의해 특별히 제한되지 않고, 휴대용 단말기 또는 컴퓨터로 구현될 수 있다. 여기서, 컴퓨터는 예를 들어, 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop), 태블릿 PC, 슬레이트 PC 등을 포함하고, 사용자에게 의해 휴대되는 형태이거나 자동차 등에 미리 장착된 형태일 수 있다.

- [0022] 또한, 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, PCS(Personal Communication System), GSM(Global System for Mobile communications), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access)-2000, W-CDMA(W-Code Division Multiple Access), WiBro(Wireless Broadband Internet) 단말, 스마트폰(Smartphone), 태블릿 PC, 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치가 포함될 수 있다.
- [0023] 저장부(110)는 비정형데이터로부터 범죄탐지를 수행하는 프로그램 (이하에서 ‘범죄탐지 프로그램’이라 지칭함)을 저장한다.
- [0024] 저장부(110)는 콤팩트 플래시(CF: Compact flash;) 카드, SD(SD: Secure digital) 카드, 메모리 스틱(Memory stick), 솔리드 스테이트 드라이브(SSD: Solid-state drive), 마이크로SD 카드(Micro SD)등과 같은 낸드 플래시 메모리(NAND flash memory), 하드 디스크 드라이브(HDD: hard disk drive) 등과 같은 마그네틱 컴퓨터 기억 장치, CD-ROM, DVD-ROM 등과 같은 광학 디스크 드라이브(optical disc drive) 등을 포함할 수 있다. 또한, 저장부(110)에 저장된 범죄탐지 프로그램은 소프트웨어 또는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)와 같은 하드웨어 형태로 구현될 수 있으며, 소정의 역할들을 수행할 수 있다.
- [0025] 프로세서(120)는 저장부(110)에 저장된 범죄탐지 프로그램을 실행한다.
- [0026] 프로세서(120)는 프로그램의 수행에 따라 범죄탐지를 위한 대상 비정형데이터를 가공처리하고, 가공처리된 데이터 중 대상 텍스트를 추출하고, 추출된 대상 텍스트와 패턴정보의 비교를 통해, 대상 텍스트의 이상점을 판단한다.
- [0027] 프로세서(120)는 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하고, 추출된 키워드를 기초로 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시물 또는 이상점 댓글을 추출한다.
- [0028] 프로세서(120)는 대상 텍스트와 이상점 게시물 또는 이상점 댓글을 군집화한다.
- [0029] 프로세서(120)는 게시물 그룹을 각각 가공처리하여 게시물 데이터베이스에 저장할 수 있다.
- [0030] 프로세서(120)는 데이터 마이닝 기법을 사용하여 대상 비정형데이터를 수집하고, 대상 비정형데이터로부터 대상 텍스트를 획득하기 위한 가공처리 과정을 수행한다.
- [0031] 예를들어, 텍스트 마이닝 기법을 사용하여 수집된 대상 비정형데이터는 SNS(Social Network Service), 포털 사이트, 블로그 등 사이버 상에 게시물 및 게시물에 해당하는 댓글에 포함된 정보(제목, 글, 작성자 등)가 있을 수 있다. 또한, 가공처리를 통해 획득되는 정보는 게시물의 제목, 게시물 텍스트 및 게시물 텍스트를 구성하는 명사 또는 동사, 작성자 ID 등일 수 있다.
- [0032] 프로세서(120)는 대상 비정형데이터로부터 가공처리를 통해 대상 텍스트를 추출하고, 추출된 대상 텍스트와 패턴정보의 비교를 통해 연관성 정도를 파악하여 이상점 여부를 판단할 수 있다.
- [0033] 프로세서(120)는 게시물 데이터베이스에 저장된 게시물 및 해당 게시물에 대한 댓글을 하나의 게시물 그룹으로 설정하고, 각 게시물 그룹 단위로 패턴정보를 파악하여 패턴정보 데이터베이스에 저장하는 별도의 과정을 수행할 수 있다.
- [0034] 프로세서(120)는 게시물 그룹에 포함된 문장을 구성하는 형태소 쌍을 추출하고, 추출된 형태소 쌍의 출현빈도 및 인접정도에 기초하여 각 형태소 쌍에 대한 누적 연관성 점수를 산출하여, 게시물 그룹에 대한 패턴정보를 생성할 수 있다.
- [0035] 프로세서(120)는 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수 보다 높게 설정하여 누적 연관성 점수를 산출할 수 있으나 이에 한정되는 것은 아니다.
- [0036] 프로세서(120)는 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 제 1 점수로 설정하고, 하나의 문장에 각 형태소 쌍이 이격되어 배치된 경우의 연관성 점수를 제 2 점수로 설정하고, 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수를 제 3 점수로 설정하되, 제 1 점수가 제 2 점수보다 높고, 제 2 점수가 제 3 점수보다 높도록 설정하여 누적 연관성 점수를 산출할 수 있으나 이에 한정되는 것은 아니다.

해당 내용은 도 2를 통해 좀더 자세히 설명하도록 한다.

- [0037] 프로세서(120)는 게시물 데이터베이스 내 모든 데이터에 대하여 패턴정보를 분석할 수 있다. 이때, 프로세서(120)는 게시물에 포함된 게시글별, 댓글별 또는 게시글 및 해당 게시글에 대한 댓글을 포함하는 게시글 그룹별로 구분하여 패턴정보를 분석할 수 있다.
- [0038] 프로세서(120)는 코사인 유사도 측정(Cosine), 자카드 유사도 측정(Jaccard), 오버랩(Overlap) 유사도 측정 중 하나의 알고리즘을 이용하여 패턴정보를 분석할 수 있다.
- [0039] 프로세서(120)는 프로그램의 수행에 따라 범죄탐지를 위한 대상 비정형데이터를 가공처리하고, 가공처리된 데이터 중 대상 텍스트의 형태소 쌍을 추출하고, 추출된 형태소 쌍과 패턴정보의 형태소 쌍 및 그에 해당하는 누적 연관성 점수를 매칭하여 연관성 점수를 산출하되, 산출된 연관성 점수가 임계값 미만일 경우 대상 텍스트를 이상점이라 판단할 수 있다. 이는 후술할 도 2를 통해 자세히 설명하도록 한다.
- [0040] 프로세서(120)는 대상 텍스트의 산출된 연관성 점수가 임계값 미만일 경우, 이상점으로 판단하며, 해당 대상 텍스트로부터 키워드를 추출하되, 추출된 키워드는 이상점 텍스트에 대한 형태소 정보이거나 또는 이상점 텍스트에 대한 명사 정보일 수 있다.
- [0041] 프로세서(120)는 대상 텍스트의 산출된 연관성 점수가 임계값 이상일 경우, 이상점이 아니라고 판단하며, 해당 대상 텍스트의 형태소 쌍의 연관성 점수를 패턴정보 데이터베이스의 패턴정보를 갱신할 수도 있다.
- [0042] 프로세서(120)는 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출하고, 추출된 키워드를 기초로 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시글 또는 이상점 댓글을 추출할 수 있다. 이때 이상점 댓글 및 대상 텍스트와 이상점 게시글 또는 이상점 댓글을 군집화할 수 있다.
- [0043] 프로세서(120)는 키워드 추출 알고리즘을 통해 추출된 키워드를 기초로 대상 텍스트가 포함된 게시글 그룹으로부터 이상점 댓글을 추출할 수 있다.
- [0044] 프로세서(120)는 키워드 추출 알고리즘을 통해 추출된 키워드를 기초로 대상 텍스트가 포함되지 않은 다른 게시글 그룹으로부터 이상점 게시글 또는 이상점 댓글을 추출할 수 있다. 이는, 이상점 댓글 및 이상점 게시글로부터 추출된 키워드를 사용함으로써, 패턴정보 비교를 통한 이상점 탐지보다 더 넓은 범위의 이상점(연관 범죄)을 탐지할 수 있다.
- [0045] 프로세서(120)는 군집화된 이상점 게시글 또는 이상점 댓글을 그 작성 시점에 따라 시계열적으로 시각화하여 표시하되, 작성 날짜를 기준으로 하는 X 축과 작성시간을 기준으로 하는 Y축 상에 2차원 그래프 상으로 표시할 수 있다. 이는 후술할 도 6a 및 도 6b를 통해 자세히 설명하도록 한다.
- [0046] 게시물 데이터베이스(130)는 비정형데이터를 저장한다.
- [0047] 게시물 데이터베이스(130)는 SNS(Social Network Service), 포털 사이트, 블로그 등 사이버 상에 존재하는 모든 비정형데이터에 대하여 게시글 및 해당 게시글에 대한 댓글을 하나의 그룹 단위로 구분하여 저장할 수 있다.
- [0048] 패턴정보 데이터베이스(140)는 게시물 그룹에 대한 패턴정보가 저장된다.
- [0049] 여기서, 패턴정보는 형태소의 조합에 따라 생성된 각각의 형태소 쌍 및 형태소 쌍 별로 누적합산된 누적 연관성 점수를 포함한다.
- [0050] 형태소 쌍은 게시물에 존재하는 형태소를 조합하여 구성할 수 있다.
- [0051] 누적합산된 누적 연관성 점수는 게시물을 구성하는 복수의 문장 중 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우 또는 하나의 문장에 각 형태소 쌍이 이격되어 배치된 경우 또는 서로 다른 문장에 각 형태소가 배치될 경우에 따라 상이하게 설정된 연관성 점수를 매칭하여 계산할 수 있다.
- [0052] 도 2는 일예에 따른 패턴정보 분석방법을 설명하기 위한 도면이다.
- [0053] 도 2를 살펴보면, 게시물 중 댓글별 패턴정보를 분석하고 있다. 구체적으로, 댓글 1(303), 댓글2(304), 댓글 3(305)의 문장을 각각의 형태소 단위로 분리하고, 댓글 1(303), 댓글2(304), 댓글 3(305)의 댓글간의 형태소 쌍(301)에 따른 누적 연관성 점수(302)를 계산하고 있다.
- [0054] 이때, 형태소 쌍은 댓글 1(303), 댓글2(304), 댓글 3(305)에 존재하는 형태소를 조합하여 구성된 것이다.
- [0055] 또한, 댓글 1(303), 댓글2(304), 댓글 3(305)에 대한 누적 연관성 점수는 하나의 댓글 안에 각 형태소 쌍이 서

로 인접하여 배치된 경우 또는 하나의 댓글에 각 형태소 쌍이 이격되어 배치된 경우 또는 서로 다른 댓글에 각 형태소가 배치될 경우에 따라 상이하게 설정된 연관성 점수를 반영하여 계산할 수 있다. 예를 들어, 형태소 쌍이 동일한 댓글 내 존재할 경우 3점으로, 다른 댓글과의 조합을 통해 형태소 A 및 형태소 B가 조합될 경우 1점으로 적용하여 누적 연관성 점수를 계산할 수 있으나, 이에 한정되는 것은 아니다. 즉, 경우의 수에 따른 점수를 모두 합한 점수가 누적 연관성 점수이다.

[0056] 도 2를 살펴보면, 형태소 A 및 형태소 B의 조합으로 이루어진 형태소 쌍의 경우, 댓글 1(303)에 해당 형태소 쌍이 존재하므로 3점, 댓글 1(303)과 댓글 2(304)의 조합이 이루어지는 경우 1점, 댓글 2(303)와 댓글 1(304)를 통해 해당 조합이 이루어지는 경우 1점, 댓글 2(303)에 해당 형태소 쌍이 존재하므로 3점이고, 모든 경우의 수에 따라 누적 합산된 누적 연관성 점수는 $3+1+1+3=8$ 일 수 있다.

[0057] 즉, 각각의 형태소 쌍 및 각각의 형태소 쌍에 대응하는 누적 연관성 점수를 패턴정보라고 할 수 있다. 이외의 다른 형태소 쌍에 대한 누적 연관성 점수의 계산은 동일함으로 설명을 생략한다.

[0058] 패턴정보는 타 대상 텍스트의 이상점 여부를 판단하는데 비교 데이터로 사용될 수 있다.

[0059] 구체적으로, 타 대상 텍스트로부터 패턴정보와 매칭되는 형태소 쌍을 추출하고, 매칭되는 형태소 쌍에 대한 누적 연관성 점수를 모두 더하여 타 대상 텍스트의 연관성 점수를 산출함으로써 이상점을 탐지할 수 있다.

[0060] 대상 텍스트가 타댓글일 경우, 타댓글을 형태소 단위로 분리하고, 패턴정보와 매칭되는 형태소 쌍으로 조합하되, 해당 형태소 쌍의 누적 연관성 점수를 더하여 타댓글의 연관성 점수를 산출함으로써 이상점 여부를 판단할 수 있다. 즉, 패턴정보의 누적 연관성 점수를 참고로하여, 대상 텍스트의 연관성 점수를 모두 합산하고, 합산된 연관성 점수가, 기설정된 임계값보다 그 수치가 낮을 경우 해당 분석대상을 이상점이라고 판단할 수 있다.

[0061] 예를들어, 대상 텍스트의 형태소가 B, D, F로 이루어졌을 경우, 도 2의 패턴정보를 통해 이상점을 탐지하는 방법은 다음과 같다.

[0062] 먼저, 대상 텍스트를 형태소 단위로 분리하고, 패턴정보의 형태소 쌍과 동일하게 생성한다. 도 2를 참조하면 대상 텍스트로부터 생성되는 형태소 쌍은 형태소 B와 D의 조합, 형태소 D와 F의 조합일 수 있으며, 이때 각각의 조합에 대한 누적 연관성 점수는 1점이므로 모두 합산하면 2점이다.

[0063] 즉, 대상 텍스트의 연관성 점수는 2점이며, 해당 값이 기설정된 임계값보다 낮을 경우 해당 분석대상을 이상점이라 판단할 수 있다. 대상 텍스트의 연관성 점수가 임계값 이상일 경우에는 해당 대상 텍스트의 대한 연관성 점수를 패턴정보 데이터베이스에 반영하여 누적 연관성 점수를 갱신할 수도 있다.

[0064] 도 3은 본 발명의 일 실시예에 따른 범죄탐지 시스템의 범죄탐지 방법을 설명하기 위한 순서도이다.

[0065] 본 발명의 일 실시예에 따른 범죄탐지를 수행하는 시스템의 범죄탐지 방법은 비정형데이터를 가공처리하는 단계(S310), 대상 텍스트를 추출하는 단계(S320), 이상점 여부를 판단하는 단계(S330), 키워드를 추출하는 단계(S340), 이상점을 추출하는 단계(S350), 이상점을 군집화하는 단계(S360)를 포함한다.

[0066] 먼저, 비정형데이터를 구성하는 게시물 및 해당 게시물에 대한 댓글이 각각 구분되어 저장된 게시물 중 범죄탐지를 위한 대상 비정형데이터를 가공처리한다(S310).

[0067] 이때, 비정형데이터를 가공처리하는 단계(S310)는 대상 비정형데이터에 대하여 특수문자를 제거하는 단계 및 제거된 텍스트에 대하여 형태소를 분류하는 단계를 더 포함할 수 있다.

[0068] 다음으로, 가공처리된 데이터 중 대상 텍스트를 추출한다(S320).

[0069] 다음으로, 추출된 대상 텍스트와 패턴정보의 비교를 통해 이상점 여부를 판단한다(S330).

[0070] 다음으로, 대상 텍스트가 이상점에 해당할 경우, 해당 대상 텍스트로부터 키워드를 추출한다(S340).

[0071] 키워드를 추출하는 단계(S340)는 키워드 추출 알고리즘을 통해 키워드를 획득한다. 이때, 키워드 추출 알고리즘은 대상 텍스트가 이상점에 해당할 경우, 이상점 텍스트에 대한 형태소 정보이거나 또는 이상점 텍스트에 대한 명사 정보일 수 있다.

[0072] 여기서, 이상점 텍스트로부터 추출되는 키워드는 이상점과 관련된 연관 범죄를 탐색하기 위한 데이터로 활용될 수 있다.

- [0073] 다음으로, 추출된 키워드를 기초로 게시물 데이터베이스로부터 이상점에 해당하는 이상점 게시글 또는 이상점 댓글을 추출한다(S350).
- [0074] 다음으로, 대상 텍스트와 이상점 게시글 또는 이상점 댓글을 군집화한다(S360).
- [0075] 군집화하는 단계(S360)는 군집화된 이상점 게시글 또는 이상점 댓글을 그 작성 시점에 따라 시계열적으로 시각화하는 단계를 더 포함할 수 있다. 이때, 시각화는 작성 날짜를 기준으로 하는 X 축과 작성시간을 기준으로 하는 Y축 상에 2차원 그래프의 형태로 제공될 수 있다. 구체적으로, Y축은 해당 날짜에 대한 시간을 분으로 표현할 수 있다. 즉, 24시간을 1440분으로 나타내어 표시할 수 있다.
- [0076] 한편, 본 발명의 일 실시예에 따른 범죄탐지를 수행하는 시스템의 범죄탐지 방법은 게시글 그룹에 대한 패턴 정보를 패턴 데이터베이스에 저장하는 단계를 더 포함할 수 있다. 이때, 패턴정보 데이터베이스에는 게시글 그룹에 포함된 문장을 구성하는 형태소 쌍을 추출하고, 추출된 형태소 쌍의 출현 빈도 및 인접 정도에 기초하여 각 형태소 쌍에 대한 누적 연관성 점수가 저장될 수 있다.
- [0077] 이때, 누적 연관성 점수를 산출하는 단계를 더 포함할 수 있다. 누적 연관성 점수를 산출하는 단계는 하나의 문장에 각 형태소 쌍이 서로 인접하여 배치된 경우의 연관성 점수를 제 1 점수로 설정하고, 하나의 문장에 각 형태소 쌍이 이격되어 배치된 경우의 연관성 점수를 제 2 점수로 설정하고, 서로 다른 문장에 각 형태소가 배치된 경우의 연관성 점수를 제 3 점수로 설정하되, 제 1 점수가 제 2 점수보다 높고, 제 2 점수가 제 3 점수보다 높도록 설정하여 누적 연관성 점수를 산출할 수 있으나 이에 한정되는 것은 아니다.
- [0078] 도 4a 내지 도 4b는 본 발명의 일예에 따른 사용자 인터페이스를 구체적으로 설명하기 위한 도면이다.
- [0079] 도 4a는 글의 패턴정보를 분석하여 패턴정보와 유사하지 않은 이상점을 탐지하여 시각화한 화면이다.
- [0080] 도 4a를 살펴보면, 사용자가 설정한 기간 범위인 2015-01-01부터 2015-05-31까지 해당 날짜 내의 글의 패턴정보를 비교하여 이상점을 탐지하여 출력하고 있다. 여기서, 가로축은 해당 기간을 나타내고, 세로축은 작성시간을 나타낸다. 구체적으로, 세로축은 해당 날짜에 대한 시간을 분으로 표현한 것이다. 즉, 24시간을 1440분으로 바꾸어 표시할 수 있다.
- [0081] 이상점 탐지구간(401)의 경우, 해당 구간에서 동일한 내용을 유포하는 이상 행위가 탐지됨을 확인할 수 있다.
- [0082] 도 4b는 패턴정보 분석을 통해 획득한 이상점으로부터 키워드를 추출하고, 해당 키워드를 이상점을 탐지하여 시각화한 화면이다.
- [0083] 도 4b를 살펴보면, 사용자가 설정한 기간 범위인 2015-01-01부터 2015-05-31까지 해당 날짜 내의 패턴정보 분석을 통해 이상점을 탐지하고, 탐지된 이상점에 대하여 키워드를 추출하여 해당 키워드와 유사도가 높은 문서를 탐지하여 시각화한 화면이다.
- [0084] 이상점 탐지구간(402)의 경우, 해당 구간에서 동일한 내용을 유포하는 이상 행위가 탐지됨을 확인할 수 있다.
- [0085] 지금까지 대상 텍스트와 패턴정보 비교를 통한 이상점 탐지가 가능한 범죄탐지 시스템 및 이상점 대상 텍스트로부터 키워드 추출을 통해 연관범죄 탐지를 수행하는 범죄탐지 시스템을 설명하였다.
- [0086] 본 발명은 대상 텍스트와 패턴정보 비교를 통해 획득한 이상점 탐지정보를 군집화되거나 또는 시계열적으로 시각화하여 사용자에게 제공할 수 있다.
- [0087] 또한, 본 발명은 패턴정보 비교를 통해 획득한 이상점 탐지정보로부터 키워드를 추출하여 해당 키워드를 기초로 탐지한 이상점 탐지정보를 군집화하거나 또는 시계열적으로 시각화하여 사용자에게 제공할 수 있다.
- [0088] 본 발명은 포털사이트의 게시판 (예를들어, 다음의 아고라 게시판 등)에서 사용자들에 의해 발생하는 각종루머, 악성댓글, 악성광고문구, 도배댓글 등의 각종 범죄글에 대하여 최초유포자와 또는 최초 유포글이 발생된 곳을 확인할 수 있으며, 해당 범죄글로부터 키워드를 추출하고 해당 키워드를 기초로하여 연관범죄 수색도 가능할 수 있다.
- [0089] 추가적으로 본 발명은 패턴정보 및 키워드를 활용하여 한 개의 포털 사이트만이 아닌 타 포털 사이트에서도 범죄탐지가 가능하도록 시스템을 구축하여 보다 효율적으로 기술의 효과를 증가시킬 수 있다.
- [0090] 본 발명에서의 온라인 상의 사용자들에 의하여 작성된 비정형데이터로부터 범죄탐지를 수행하는 시스템은 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능 한 명령어를 포함하는 기록 매체의 형태로도

구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 매커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다. 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

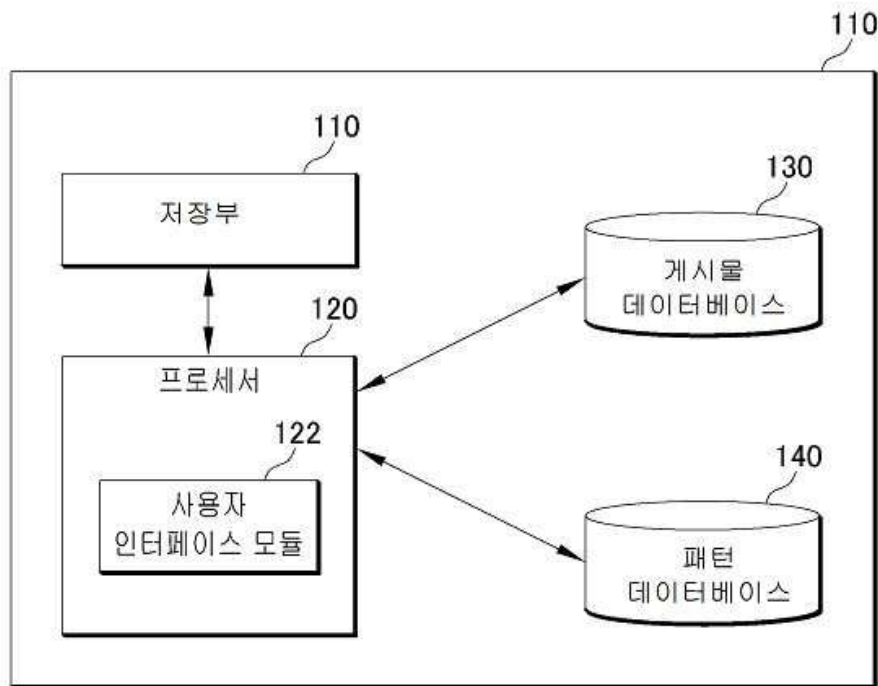
[0091] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다. 본 발명의 범위는 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

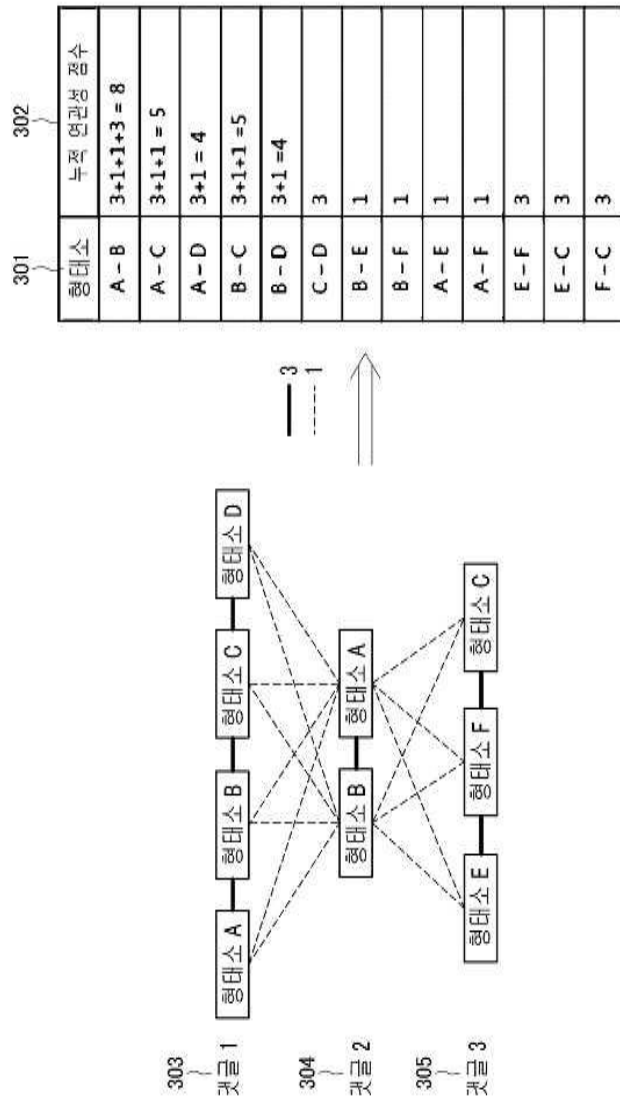
[0092] 100: 범죄탐지수행 시스템
110: 메모리
120: 프로세서
122: 사용자 인터페이스 모듈
130: 게시물 데이터베이스
140: 패턴정보 데이터베이스

도면

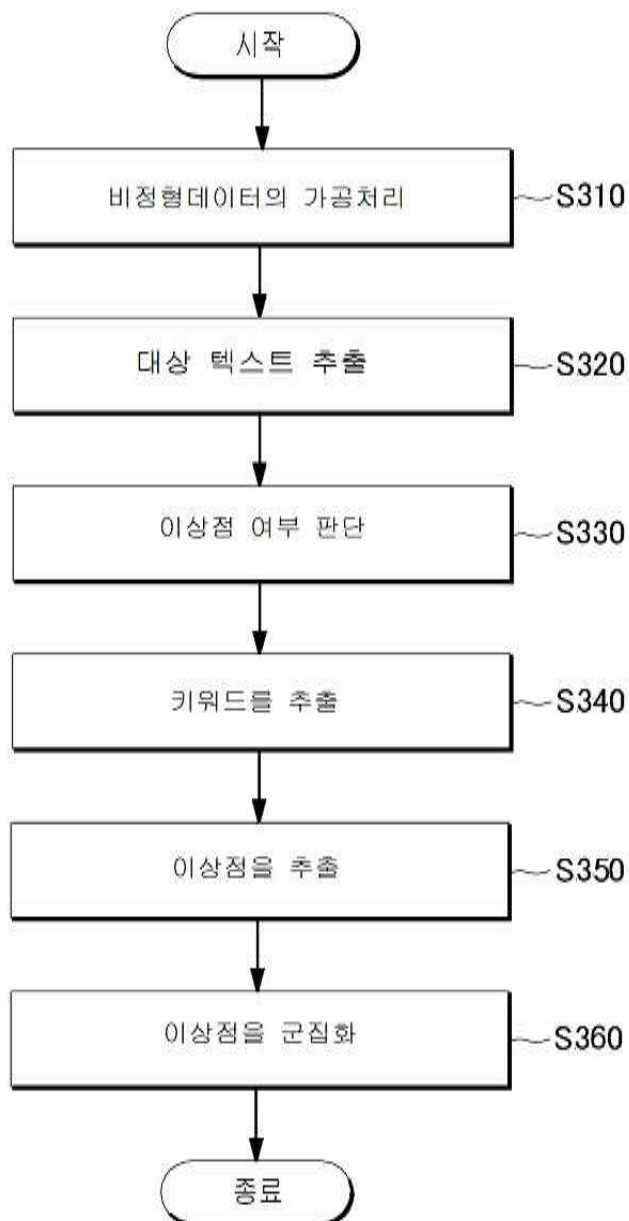
도면1



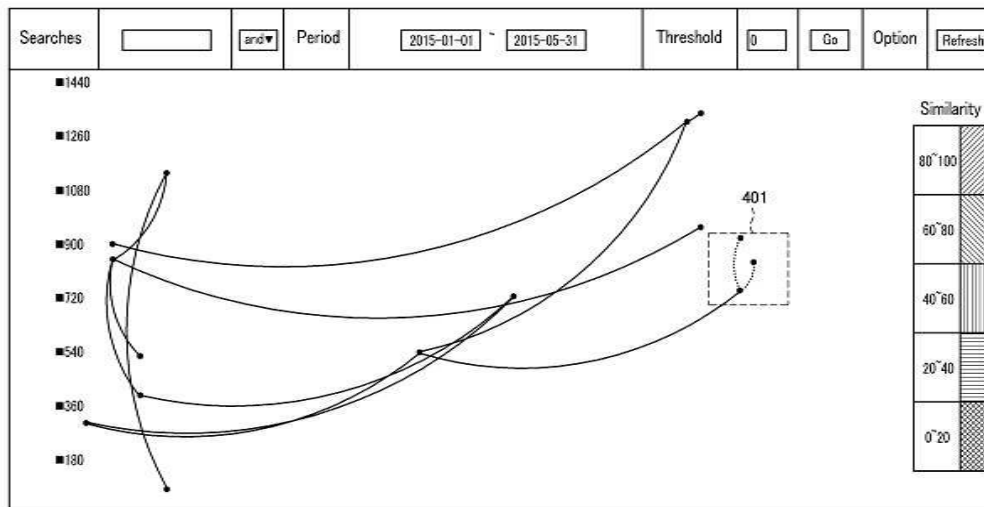
도면2



도면3



도면4a



도면4b

